Aged Rings: The Use of Machine Learning in Determining Abalone Age from Morphological Traits

Abalones, a warm-seas marine snail commonly found in the intertidal habitat, are a commercially important organism with many endangered species. This study applies various machine learning models in analysis of abalone physical traits in order to determine a potential specimen's approximate age, which can serve as an extremely useful tool in conservation and commercial purposes.

Introduction

Abalones are known for a variety of commercial products harvested both from their shell and other parts of the body. One widespread example is that of food in many forms of cuisine across the world, both as traditional fare and also as a delicacy. Similar to other seafood such as shark fin soup or fugu, it may even be considered a luxury food. In this respect, it is highly sought after and expensive. Other abalone products include materials harvested from the shell itself. Mother-of-pearl, also known as nacre, is one example of a decorative material used in widespread forms of jewelry, furniture, and musical instruments. Lastly, both abalone meat and shells are used in many indigenous cultures across the world in methods such as currency or ornaments.

Potentially because of this array of commercial prospects, the majority of abalone species are facing extinction as a result of human activity. One factor of this is overfishing and overharvesting, particularly in the West Coast of the United States. According to the National Atmospheric and Oceanic Administration, commercial fishing has severely reduced abalone populations far below their historic numbers. In addition to this pressure, environmental effects have also played a large role in their survival. Increased emissions of carbon dioxide interactions with the oceans have increased the acidification of ocean water, at which the reduced pH will erode their calcium carbonate shells. This erosion will weaken the structural support provided by those shells as well as their defensive abilities from the environment and from predators. As such, many species have strict restrictions placed upon their harvesting.

One reason for their threatened status is their unique biology and methods of reproduction. As with many gastropods, their locomotion is severely limited by their shell weighing them down as with their grazing lifestyle. In addition, abalones only mate via broadcast spawning where these gametes are 'broadcasted' into the environment. This means fertilization succeeds more often when groups of adult male and female

abalone are close to each other when they spawn. Females are able to release hundreds of thousands of eggs, but without a male nearby, the spawn will fail. Due to overfishing and low population densities, males and females find it difficult to find one another in order. As a result, abalone aquaculture farms have proven to be one avenue of lessening the stress on these populations along with supplying those same commercial products.

One problem facing these aquaculture farms is that of appropriate abalone age dating. These gastropods have relatively late sexual maturity ages compared to other invertebrates, ranging from 4 to 6 years with a maximum lifespan of 35 to 40 years. As a result, it is necessary to maintain careful observation of these abalone's ages in order to appropriately encourage reproduction or to set restrictions on when to sell them. However, the traditional method of dating an abalone involves cutting through the shell in order to stain it and then counting the number of internal rings with a microscope. As this is a time-consuming and inefficient process, other methods of measurement have been suggested to find this age.

The method this project will involve is that of studying physical traits of the abalones themselves. Tabulating continuous variables such as length and width of the shell, abalone sex, total weight, and shell weight are all variables easily sorted and predicted using machine learning to vastly streamline the process as a whole. Even the act of measuring the length and weight of an abalone is far quicker than the older method of shell staining and examination, and could prove invaluable to the aquaculture industry and to the commercial world beyond.

Method

This project utilizes the Abalone dataset sourced from UC Irvine Machine Learning Archive, available from Abalone - UCI Machine Learning Repository. This data consists of a collection of categorical data in the form of abalone sex as well as continuous data in length, diameter, and width of the abalones. In addition, many relevant weights were measured, including total weight, shell weight, shucked weight or weight without the shell, and viscera weight. The target value of this dataset was measured by the number of rings measured, as rings correlate to abalone age.

In inputting the dataset for the machine learning models to use, firstly the categorical variable 'Sex' was converted into a numerical variable using one-hot encoding. As the others were all continuous, there was no need to modify them. Next, in preparation of using the different machine learning models, the dataset was split into training and test sets. These sets were all standardized to conclude their preparation.

With the data prepared, I decided to use four different machine learning models to compute this data. These models were logistic regression, random forest, support vector machine and decision tree. Logistic regression was expected to be the simplest model to work with the data, as many of these variables could be more or less linearly related to abalone age as with abalone size or weight being positively correlated with age in most circumstances. Decision tree was expected to be the most reliable model as the many relationships with these variables in relation to ring amount might be too complex for logistic regression to handle.

After these 4 models were set up, they were evaluated to determine their accuracy of prediction.

```
Logistic Regression Accuracy: 0.2811004784688995
Random Forest Accuracy: 0.2452153110047847
SVM Accuracy: 0.31100478468899523
Decision Tree Accuracy: 0.21291866028708134
```

Figure 1. Listing of training model accuracy.

As the above shows, support vector machine proved to be the most reliable model with the dataset, followed by logistic regression. Even so however, the low percentage indicates these models as unreliable for close predictions, to which I decided to test in more detail to see the source of the error.

Results

Once the models had been trained, I selected a random single set of features of one individual abalone from the dataset to compare with the predictions of the models. The target value of Rings was noted before feeding the dataset's features into the models to see what target value would be predicted.

```
Actual 'Rings' Value: 11

Logistic Regression Prediction for 'Rings': 10
Random Forest Prediction for 'Rings': 10
SVM Prediction for 'Rings': 10
Decision Tree Prediction for 'Rings': 10
```

Figure 2. Results of test running of models with randomly selected data. Perhaps due to the low number of target value, the results are similar across the machine learning models.

As the above shows, the models proved to be very close to the original dataset's target value of 11 rings. The difference in accuracy is enough to sway the end result substantially even with these low values to work with.

Next, and to give more evidence of reliability or lack thereof with the predicted results, I generated five random features within each variable that the dataset was using. These were random in each variable of the dataset, ranging across the board of abalone size and width to total weight and shell weight, as shown in the code below.

```
# Generate sample data
sample_data = pd.DataFrame({
    "Length": np.random.uniform(0, 1, size=5),
    "Diameter": np.random.uniform(0, 1, size=5),
    "Height": np.random.uniform(0, 1, size=5),
    "WholeWeight": np.random.uniform(0, 1, size=5),
    "ShuckedWeight": np.random.uniform(0, 1, size=5),
    "VisceraWeight": np.random.uniform(0, 1, size=5),
    "ShellWeight": np.random.uniform(0, 1, size=5),
})

# Since 'Sex' was one-hot encoded
sample_data["Sex_I"] = np.random.choice([0, 1], size=5)
sample_data["Sex_M"] = np.random.choice([0, 1], size=5)
```

Figure 3. Random sampling generation. The np.random formula was used with each feature.

As the feature ‘Sex’ had been one-hot encoded earlier as a categorical variable with only two possible values, the prediction models were unable to use it with the same np.random.uniform formula, hence the np.random.choice. This was sufficient to reach the nearest extent I could with randomization in this case. I chose to generate five random samples so as to provide a wide range of predicted results in comparison with the actual results.

Finally, I transferred the results of the above random generation to graphically represent the variance of the predicted data, as well as compare to actual target values from the dataset. After the random samples had been fed into the predictive models, I plotted them onto a bar chart with each of the models. I also included five random target values from the actual dataset to roughly compare the differences.

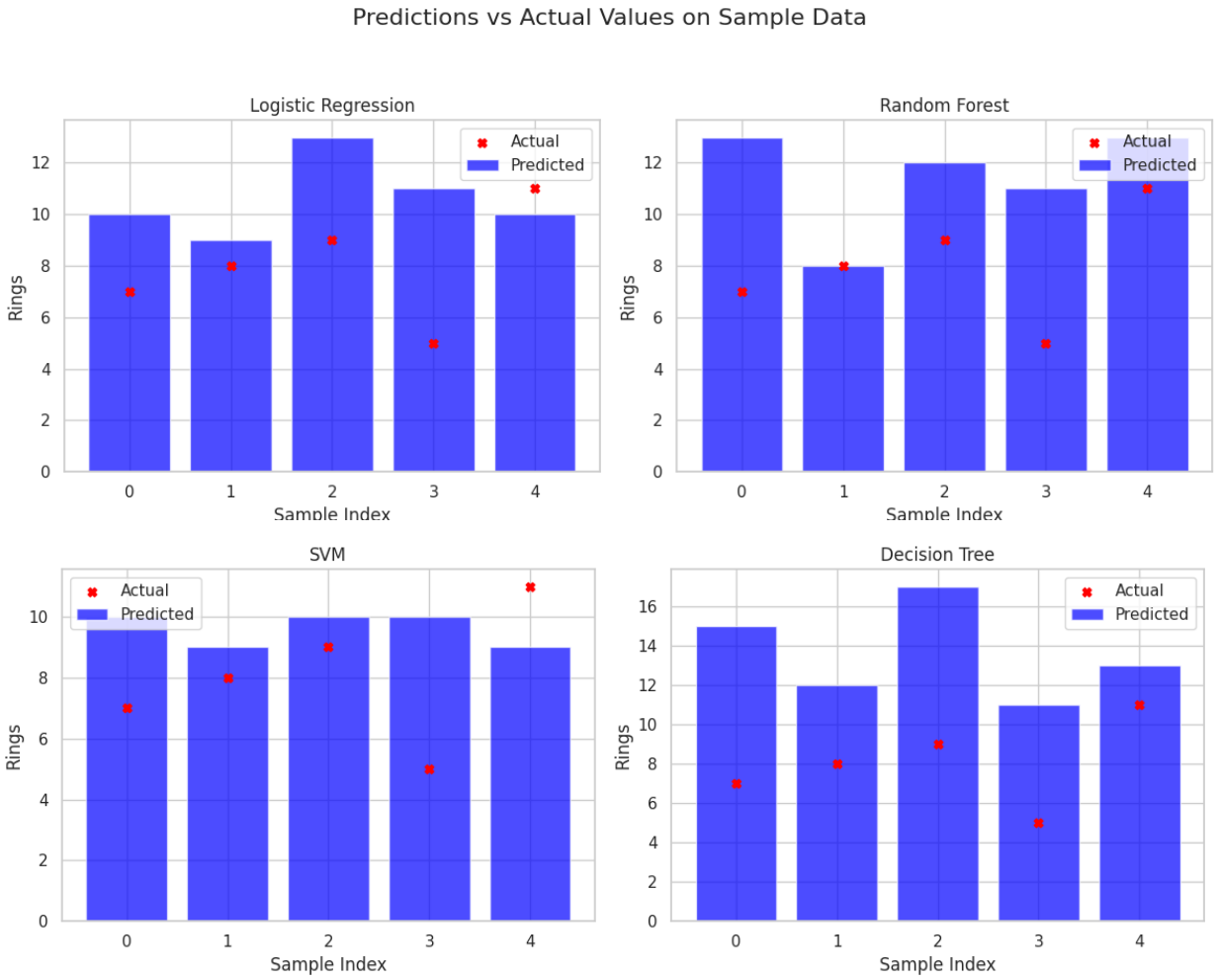

Figure 4. Bar Charts of the four prediction models.

The results above show a substantial amount of variance between the actual and predicted results. Each of the models are rather unreliable with most given data, with the two outliers being random forest and decision tree according to this sample of random data. Logistic regression follows the two in the order of accuracy, with support vector machine seeming to be the most accurate. This trend would follow that given by the evaluations of the models near the beginning of this project. However, at this level of inspection and with the low values of all the target values, any results obtained here are largely subjective.

Discussion

In conclusion of the project, there may be a variety of reasons as to the inaccuracy of these models. One may be that they were the incorrect choice of machine learning models to use with this dataset, as strictly continuous values of many diverse orders such as weight and length may be too complex to use with a model such as logistic

regression, which is most effective with linear relationships. Another may be the nature of these organisms and their morphological traits themselves. Further information, such as weather patterns and location (translating into food availability and proper nutrition, as well as abiotic factors of temperature) may be required as well for a more complete dataset. It may even be due to genetic or specific species variations, in the case that the sourced dataset mistakenly left out the studied specimen's species. In nature, many variables can be at play in determining traits such as physical size. In many cases, these variables all have undiscovered and unseen effects on each other too. These effects may cause a compounding or neutralizing effect on their development of size and other traits.

Even so, the ability of these models to gain a faint accuracy to the prediction of these rings and ages of abalone is not to be ignored. If given more data or tested with other forms of machine learning, this process of streamlining measurements of abalones may still be of great use to the aquaculture industry. As showcased in the final figures, already the support vector machine model was narrowing down on the actual target values. This project shows the potential for machine learning to accurately predict the relationship between even these few variables that are easily measured to vastly increase the efficiency of one monotonous task.

The work of this project may be expanded on in the future by testing more machine learning models for one with more applications towards this sort of scientific study. In the case of measuring living organisms and as stated earlier, continuous or categorical variables in the terms of things like behavior or weather may be vital to a further extension of this project. Indeed, including more variables such as location of collection may also increase the accuracy of any training model, as it reduces the amount of variance present in nature. As long as those essential details that play a role in the abalones' lives are excluded, that will be reflected in the model's ability to predict the phenomena that can throw off any scientist's measurements.