Project 2 Report

Live Abalone at American Farms in Davenport, California
Photo Credit: Charles Russo [1]

David Nnaji
Thursday, May 14, 2020
ENGR 571 – Analytics is Systems Engineering
Colorado State University

Problem Statement

The following data analysis was based on project six listed in the project guidelines. It suggested building a "Classification model of Module 10 applied to your own data set."

Module 10 provided a walkthrough of the training, validation, testing, and deployment of a small classification algorithm with three features, three classes, and thirty samples for each development phase. It also provided a brief analysis at the end of the system performance.

With the available information from Module 10, an analogous system was built in Python and applied to a marine biology dataset published by the University of California Irvine online machine learning repository [2]. The dataset description states that the Marine Resources Division of Tasmania, Australia originally used the data in a biology study before eventually donating it to UCI in 1995. Interestingly, the documentation also reports that two papers have been published that used the data to test novel neural networks [2].

Abalone are marine snails that belong to the family *Haliotidae*. To determine the age of an individual abalone, it must be cut and stained before the number of rings can be counted through a microscope [2]. Since this is process is often time-consuming and tedious, other approaches involving much easier physical measurements have been proposed to estimate their age.

Thus, this project aims to develop a classification algorithm that will predict the number of rings (and by extension the age) of the abalone given the features and measurements in Table 1.

Feature Name	Data Type	Meas.	Description
Sex	nominal		M, F, and I (infant)
Length	continuous	mm	Longest shell measurement
Diameter	continuous	mm	perpendicular to length
Height	continuous	mm	with meat in shell
Whole weight	continuous	grams	whole abalone
Shucked weight	continuous	grams	weight of meat
Viscera weight	continuous	grams	gut weight (after bleeding)
Shell weight	continuous	grams	after being dried
Rings	integer		+1.5 gives the age in years

Table 1: Feature Table

Since the number of rings if not a continuous value, this problem can be considered a classification problem. For simplicity, the classes were set as r<4, 5, 6, …, 21, and r>21 (20 classes in total).

Experimental Approach

The overall approach to this project involved gathering data, conducting a broad analysis of the data, sniffing the data with ordinary least squares regression for meaningful feature combinations, training, and testing. The python code can be reviewed in the referenced Jupyter notebook [3].

First, the data was gathered from its source which initially included 4177 samples with 9 values for each feature per sample. Three new categories were immediately derived from these features including Age (rings+1.5), Coded Sex (where M, F, I are coded as 1,2,3 respectively), and Classes (r<4 = 3 and r>21=22). Additionally, the data was purged of some problematic outliers and anomalous groups of consecutive samples discovered during coding. This resulted in a final dataset with 2873 samples.

Four groups of plots were created to get a better sense of the properties of the dataset and its features. The scatter plots shown in Fig. 1 was the first group made. For brevity, no comments will be made aside from the observation that each sample feature is plotted on its particular measurement scale and that different abalone age groups appear to have been measured throughout sampling.

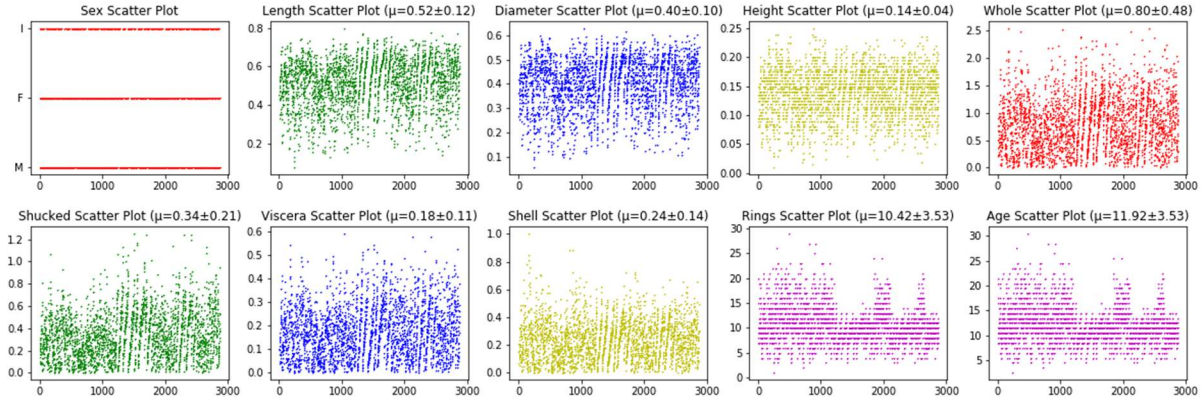


Figure 1: Run-Order Plots

Histograms were the second group of plots made and helped understand the center, spread, and skew of the data. These are shown in Fig. 2. For example, this plot indicates that the mean number of rings is roughly 10 but ranges from 1 to nearly 30. An interesting observation is that the length and diameter histograms as well as the whole, shucked, viscera histograms are notably similar. Note that the ring and age histograms appear aliased only because these values are integers with a constant increment. The same is observed in all the other plots as well.

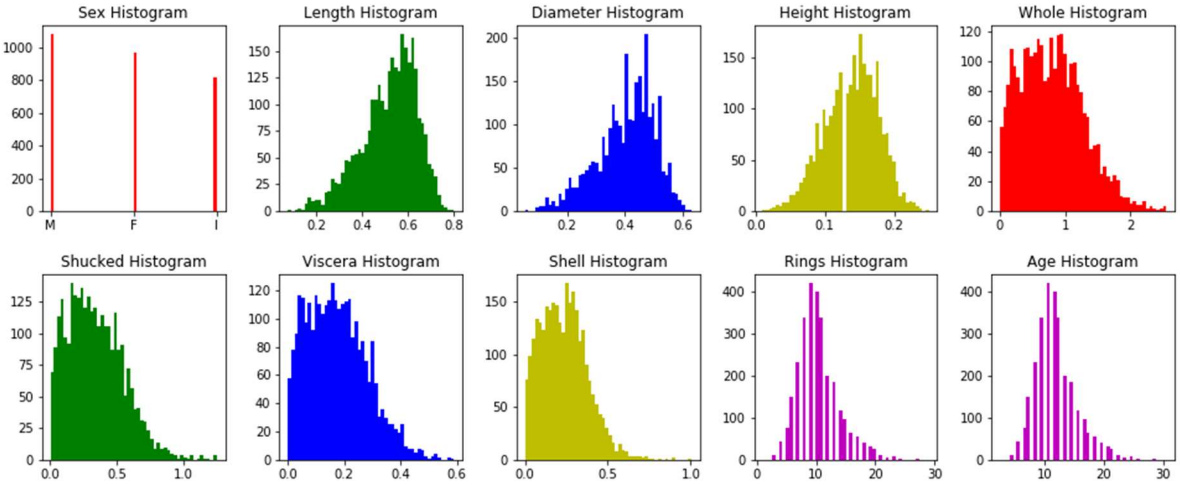


Figure 2:Histograms

Box plots were the third group of visuals (Fig. 3). All eight plots share the same vertical axis indicating the age. The sex box plot shows the average age of the male, female, and infant abalone. The remaining box plots convey the location and variation of each feature depending on whether it is above or below the feature mean. For example, the length box plot indicates the somewhat obvious statistic that

samples with a larger than average length tend to be older than those below the mean. Larger gaps between the means of each pair of box plots indicate that this trend is more severe.

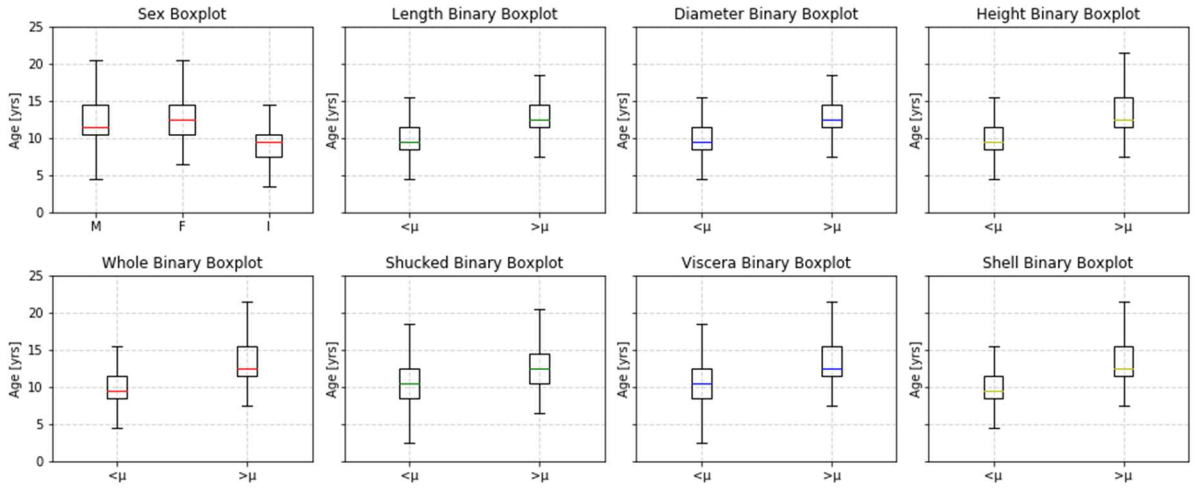


Figure 3: Box Plots

Finally, each factor was plotted against age in Fig. 4. Aside from the sex plot, there are two general trends observed in the data. The length, diameter, and height plots appear exponentially proportional to age. The remaining plots appear logarithmically proportional to age. All of these plots are heteroscedastic, that is, the variance tends to increase as the factor magnitude increases.

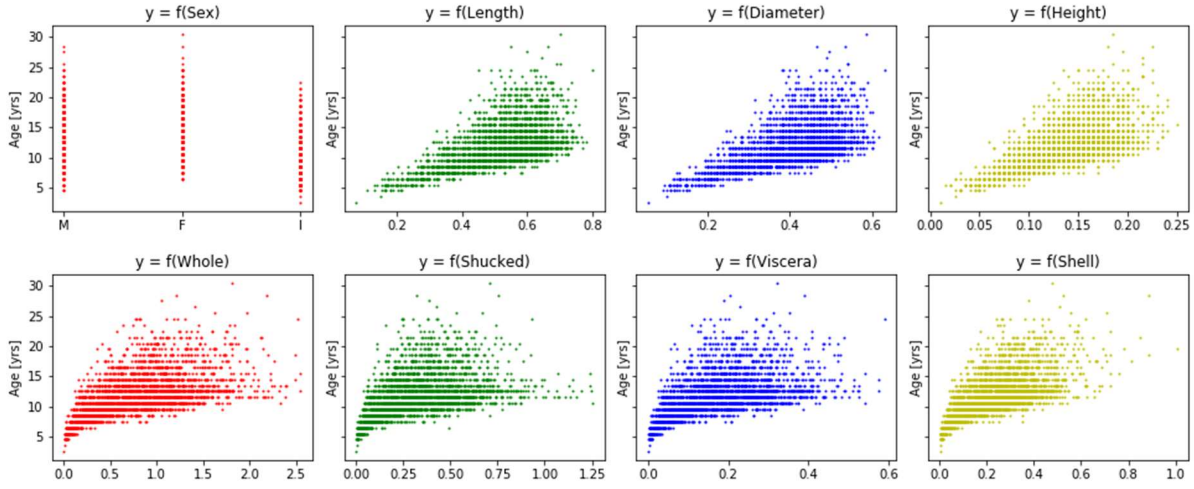


Figure 4: Scatter Plots of Factor vs. Age

An ordinary least squares regression (OLS model) was generated to provide a hint as to which factors were more correlated to the number of rings. Although the R2 value was roughly 65.2%, the results summary was difficult to interpret. It suggested high order interactions were the best. This summary can be found in the Jupiter notebook.

The data was split into 70% training and 30% testing. To ensure the populations of each factor were sufficiently similar a two-sample t-test was run. All factors fell below 1.330. In total 287 total features

and derived features were created which included: all multiple combinations of the original 8 features, all divisible combination pairs of the original 8 features, and four hybrid features. These were defined as follows.

$$\boldsymbol{Hybrid\ 1}: CS\ +\ (W/Sl) + (L/H)$$
$$\boldsymbol{Hybrid\ 2}: (W/Sl) + (L/H)$$
$$\boldsymbol{Hybrid\ 3}: (H/W) + (Sk/W)$$
$$\boldsymbol{Hybrid\ 4}: [(L/H) + (D/H)]/[(W/Sk) + (W/V) + (W/Sl)]$$

Although the t-statistics and p-values were calculated using the ranked means and standard deviations, the top 8 features were selected using the highest nSTD-CPts values. The weights of these features were then determined and deployed for the testing phase.

The testing procedure was conducted by computing the z-values of the eight selected features for each sample. Similar to Module 10 the z-function was set as the inverse square of the z-value for each sample's features. Each f(z) was multiplied by its corresponding deployed weight before being summed. The highest weighted values were then used to determine the class of the sample. The accuracy of the system was determined and no additional proofing was conducted after the initial experiment. A run order plot is provided below of the classifier performance.

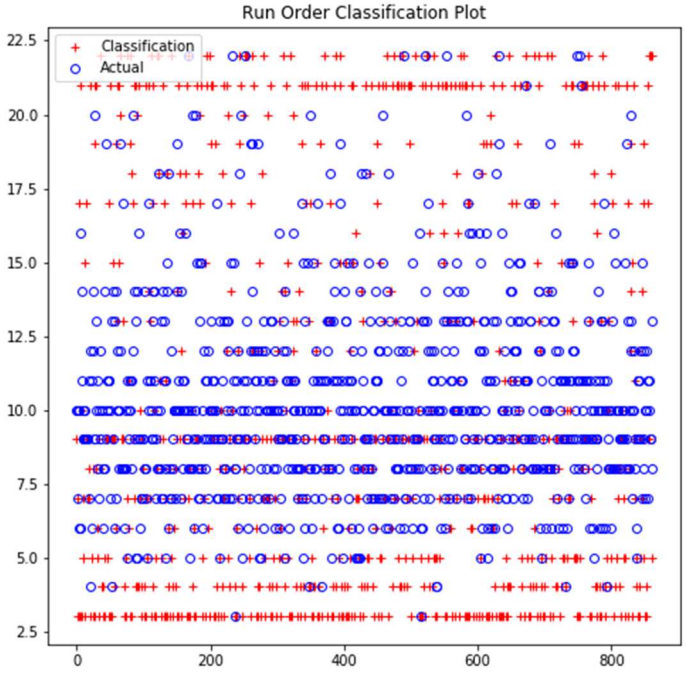


Figure 5: Run Order Classification Plot

Results

1. The summed nSTD-CPt metric and summed p-values did not rank the same top 8 features. Only one feature was shared between the two of them (Length/Shell). The nSTD-Cpt features were ultimately selected for testing. The summed t-values were also calculated but not used.

Features	nSTD-Cpts	p	t
H1	169.871275	10.148371	-19.586718
H2	133.756239	10.933558	-13.506284
L/V	124.641953	10.443273	-30.089511
D/V	93.679837	10.705593	-29.453491
W/V	91.797481	13.116778	-8.394378
L/SI	79.393247	6.789942	-48.131532
L/H	71.860856	10.750272	-15.254836
L/Sk	61.832010	9.872302	-33.489670

Features	nSTD-Cpts	p	t
H3	13.745092	5.702444	-44.040860
D/SI	59.561624	6.696112	-48.102531
L/SI	79.393247	6.789942	-48.131532
SI	5.046722	7.032526	-50.240498
H/SI	19.846747	7.052305	-42.117822
H*SI	0.854099	7.208598	-47.444898
L*SI	3.019811	7.243114	-47.337961
D*SI	2.388435	7.301768	-47.431891

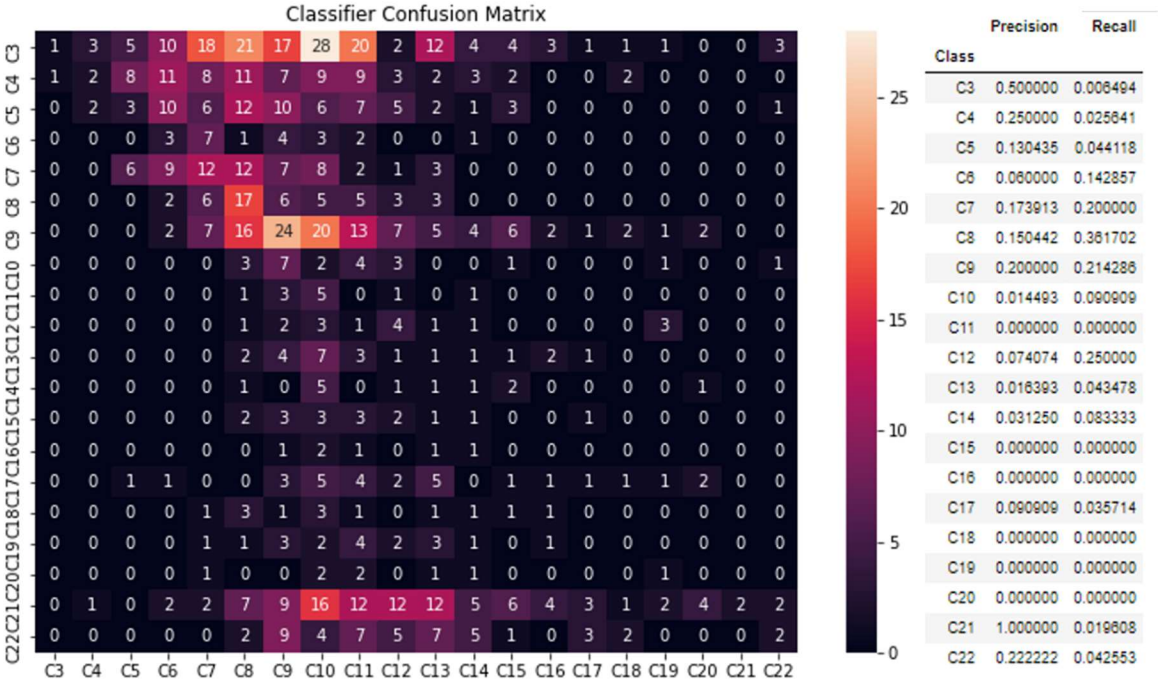
Table 2: Top 8 Features Based on 20 summed nSTD-CPt (Left) and p-values (Right)

2. The final weights were deployed in testing.

H1	0.190838
H2	0.117821
L/V	0.290070
D/V	0.155789
W/V	0.055505
L/Sl	0.098063
L/H	0.034137
L/Sk	0.057777

Table 3: Deployed Feature Weights

3. The final value counts after testing was 75 correctly classified and 787 misclassified. Thus, the final accuracy is 8.70%.



Class	Precision	Recall
C3	0.500000	0.006494
C4	0.250000	0.025641
C5	0.130435	0.044118
C6	0.060000	0.142857
C7	0.173913	0.200000
C8	0.150442	0.361702
C9	0.200000	0.214286
C10	0.014493	0.090909
C11	0.000000	0.000000
C12	0.074074	0.250000
C13	0.016393	0.043478
C14	0.031250	0.083333
C15	0.000000	0.000000
C16	0.000000	0.000000
C17	0.090909	0.035714
C18	0.000000	0.000000
C19	0.000000	0.000000
C20	0.000000	0.000000
C21	1.000000	0.019608
C22	0.222222	0.042553

Figure 6: Confusion Matrix of the Testing Results (Right) Table 4: Precision and Recall for Each Class in CM

4. A plot of the final confusion matrix shows three significant clusters centered at class 10 and class 9 in Fig. 6. Overall, this indicates that much of the poor performance occurs when the classifier is given a very young or very old abalone. The overall accuracy of 8.70% is determined by averaging the main diagonal.

5. The precision and recall are provided in Table 4.

Discussion

Given the poor performance, many enticing modifications that can be made to this algorithm to improve the overall accuracy since only one run was conducted. This includes the generation of new derived features and ideally hybrid features. Hybrid features tended to perform well compared to the multiplicative and divisive features based on nSTD-Cpts and p-value rankings (top performers in both were hybrids).

Perhaps the best improvement could stem from more informed methods of deriving these hybrid features. The four that were created used only factor relationships from the OLS regression summary. A better understanding of how to combine these features should result in better accuracy given that it is selected by either of the two rankings.

For simplicity, only the top 8 features from ranked by the nSTD-CPt metric were selected. However, the p-value and t-statistics could be a better indicator of factor fitness. Changes in f(z) may also yield minor improvements to the system. In fact, a run with f(z) = z-3 increased system accuracy by roughly 1.2%.

Regardless of the method, the selected results show that the current configuration of the system tends to misclassify samples as 9 and 10. This can more easily be seen with a probability of classification plot shown in Fig. 7.

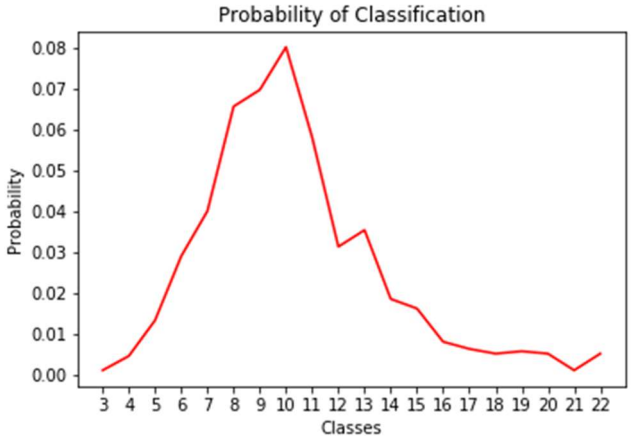


Figure 7: Probability of Classification Plot

This plot was generated by taking the sum of the columns of the confusion matrix and dividing it by the total. Again, the plot shows that class 10 has the highest probability.

Learning and Follow Up

Overall, this project was an excellent exercise of many of the topics discussed in class, and only a small portion of the analysis tools was utilized. As noted in the discussion the next activity is to run more tests with newer factors and more hybridized factors. Additionally, a comparison of the two performance metric rankings on system accuracy would also help in the optimization. Another follow up activity would be to incrementally analyze the results from testing using binary confusion matrices.

A less pressing activity will be to tidy up the code and generalize it more so it can be deployed for another other datasets.

References

[1] Kadvany, Elena. “The Coastside's Best-Kept Secret Is an Abalone Farm with a Seafood Market and a View.” Medium, THE SIX FIFTY, 3 Jan. 2020.
[2] UCI Machine Learning Repository: Abalone Data Set, mlr.cs.umass.edu/ml/datasets/Abalone.
[3] ENGR 571 Project 2 Github Repository. github.com/dcn1470/ENGR-501-Project-2