**The Weight of Crabs**

Joey Guthrie, [jguthrie@bellarmine.edu](mailto:jguthrie@bellarmine.edu)

I am working with a dataset with 3893 rows and 9 columns that pertains to the weight, size, sex, and age of crabs. I unfortunately do not know where the data was gathered from, but I found the CSV file on Kaggle.

Crabs are a type of amphibious animal belonging to the phylum Arthropoda and the infraorder Brachyura. They live in oceans, land, and some live in freshwater too. They have 10 limbs, with 2 claws, 6 legs for walking, and 2 legs specifically adapted for swimming. Like all arthropods, they have segmented bodies, encased by an armored exoskeleton of chitin, though most crabs have no visible abdomen, because it is almost entirely within the animal’s thorax. I chose this dataset because I have always been fond of arthropods, and this was the only dataset I could find on Kaggle related to a specific arthropod. The dataset has 3893 rows and 9 columns. All except ‘Sex’ and ‘Age’ are Float64.

**Table 1: Data Types**

|  |  |
| --- | --- |
| *Variable Name* | *Data Type* |
| Sex | Object |
| Length | Float64 |
| Diameter | Float64 |
| Height | Float64 |
| Weight | Float64 |
| Shucked Weight | Float64 |
| Viscera Weight | Float64 |
| Shell Weight | Float64 |
| Age | Int64 |

**CHART 1: Crab age as it relates to weight**A graph of a crab age

Description automatically generated

A screenshot of a graph

Description automatically generated

After analyzing the data, I concluded that there was nothing significant to drop besides the “sex” column, because it was just not very insightful or interesting when its variables were converted into binary numbers for analysis. Otherwise, the data was kept largely intact, and split into test and training set. As can be ascertained from Charts 1 and 2, as well as Table 3, there is a positive correlation with the weight, age, and size of a crab. However, Height has the least strong correlation with Weight OR Age. The Weight also seems to plateau for many crabs early on, while certain crabs kept growing as they aged.

Table 2: Experiment Parameters

|  |  |
| --- | --- |
| **Experiment Number** | **Parameters** |
| 1 | All variables from the training and test set, tested with 80/20 |
| 2 | All variables from the training and test set, tested with 80/20 |
| 3 | All variables from the training and test set, tested with 80/20 |
| 4 | All variables from the training and test set, tested with 80/20 |

I used Python from Jupiter Notebook (Anaconda 3), and the Python libraries of Pandas, Numpy, Matplotlib, and SKLearn. I used these tools because they were the tools that I’ve been taught to use, and the tools that I have grown to accept to utilize when necessary.

**TABLE 3: Predicted Values for Weight vs Actual Values for Weight**

|  |  |  |
| --- | --- | --- |
| Test # | Predicted Value | Actual Value |
| 1 | 25.96420046 | 24.635715 |
| 2 | 5.24713593 | 5.400580 |
| 3 | 7.78188847 | 7.952035 |
| 4 | 13.30178082 | 13.480187 |

I think the Linear Regression using the test and training sets offered the best results. Predicted Values seem to be more precise when the actual numbers for weight were smaller, but it overall appears quite accurate.

This project was riddled with problems. I was extremely busy with studying and taking finals recently, and the scope of this project was certainly overwhelming. Technical difficulties with the assignment submission on Moodle also greatly increased my anxiety and confusion. I also had not thought to check when the presentation date was, and it ended up being at the same time as my statistics final, which due to a bug, I could only schedule an 8:00 AM appointment for in the testing center, coincidentally coinciding with the same time and day as the presentation. However, luckily I was able to present the day prior to the initially scheduled day instead. I have also really been struggling to grasp many of the specific concepts for the linear regression models, so working around my confusion was a challenge. Thankfully, the coding part went relatively smoothly with no true errors that I was unable to fix within a minute or so of examination of the code. The biggest challenges were ultimately time limitations, schedule conflicts, confusion, and burnout.

My model is rather limited, because its not terribly fancy. I used much of the classwork over these models for reference, and thus, my code is nothing terribly innovative or new. I think the biggest limitations come not necessarily from its possible inferiority to other models, but the application of the data and prediction models themselves. Predicting the weight of a crab using a relatively large amount of very specific pieces in information is a very niche and specific task, where I cannot fathom its practical application.

For future work, I would definitely use more advanced models, and a dataset that is possibly more useful or applicable to real scenarios. I would also put much more time and resources into actually having a thorough understanding of the processes themselves, since a majority of my problems were due to time constraints. Using parts from several models and melding them together to create a new prediction model would be a very interesting project to undertake in the future, to see whether it would produce more accurate predictions or not.

Overall, I think I did alright on the project. My initial questions of any correlation between crab size, age, and weight were answered, I found my model to be satisfactory, (though not particularly spectacular) and the predicted results were generally very close to the actual ones. Ideally, I would’ve had more of an understanding of the matter in the project, but much of the content and concepts were only introduced to me within the past few weeks, and I was unable to fully grasp them before the final project.