Project 3 Proposal – Predict Abalone Age from Physical Measurements

Group NoFloods team members

- Amit Patel
- Daniah Al Bayati
- Roopa Patel
- Samuel Parks
- Steven Lee

GitHub Link

https://github.com/Roopa-16/Abalone-Project

Introduction

What is an abalone? From Wikipedia, an abalone is a common name for any of a group of small to very large sea snails, marine gastropod molluscs in the family Haliotidae. Determining the age of abalone is typically time consuming - requiring cutting the shell, staining it, then counting the number of rings with a microscope. We want to see if physical measurements can be used to predict the age.

Dataset Information

The data for the project comes from the UCI Machine Learning Repository: https://archive.ics.uci.edu/ml/datasets/Abalone

There are 4,177 rows in the dataset with 8 attributes:

Name	Data Type	Measurement	Description
Sex	Categorical		(M)ale, (F)emale, (I)nfant
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
Viscera weight	Continuous	grams	Gut weight (after bleeding)
Shell weight	Continuous	grams	After being dried
Rings	Integer		Rings + 1.5 gives the age in years

The variable to be predicted is Rings, for which Age = Rings + 1.5

Other notes about the dataset:

- Missing values have already been removed
- The ranges of the continuous variables have already been scaled by a factor of 200 for use with a neural network

Sample rows from the dataset:

Sex	Length	Diameter	Height	Whole_weight	Shucked_weight	Viscera_weight	Shell_weight	Rings
M	0.455	0.365	0.095	0.514	0.2245	0.101	0.15	15
M	0.35	0.265	0.09	0.2255	0.0995	0.0485	0.07	7
F	0.53	0.42	0.135	0.677	0.2565	0.1415	0.21	9
M	0.44	0.365	0.125	0.516	0.2155	0.114	0.155	10
I	0.33	0.255	0.08	0.205	0.0895	0.0395	0.055	7
I	0.425	0.3	0.095	0.3515	0.141	0.0775	0.12	8
F	0.53	0.415	0.15	0.7775	0.237	0.1415	0.33	20
F	0.545	0.425	0.125	0.768	0.294	0.1495	0.26	16
M	0.475	0.37	0.125	0.5095	0.2165	0.1125	0.165	9
F	0.55	0.44	0.15	0.8945	0.3145	0.151	0.32	19
F	0.525	0.38	0.14	0.6065	0.194	0.1475	0.21	14
M	0.43	0.35	0.11	0.406	0.1675	0.081	0.135	10
M	0.49	0.38	0.135	0.5415	0.2175	0.095	0.19	11
F	0.535	0.405	0.145	0.6845	0.2725	0.171	0.205	10
F	0.47	0.355	0.1	0.4755	0.1675	0.0805	0.185	10
M	0.5	0.4	0.13	0.6645	0.258	0.133	0.24	12
I	0.355	0.28	0.085	0.2905	0.095	0.0395	0.115	7
F	0.44	0.34	0.1	0.451	0.188	0.087	0.13	10
M	0.365	0.295	0.08	0.2555	0.097	0.043	0.1	7
M	0.45	0.32	0.1	0.381	0.1705	0.075	0.115	9
M	0.355	0.28	0.095	0.2455	0.0955	0.062	0.075	11
I	0.38	0.275	0.1	0.2255	0.08	0.049	0.085	10
F	0.565	0.44	0.155	0.9395	0.4275	0.214	0.27	12
F	0.55	0.415	0.135	0.7635	0.318	0.21	0.2	9
F	0.615	0.48	0.165	1.1615	0.513	0.301	0.305	10
F	0.56	0.44	0.14	0.9285	0.3825	0.188	0.3	11
F	0.58	0.45	0.185	0.9955	0.3945	0.272	0.285	11
M	0.59	0.445	0.14	0.931	0.356	0.234	0.28	12
M	0.605	0.475	0.18	0.9365	0.394	0.219	0.295	15
M	0.575	0.425	0.14	0.8635	0.393	0.227	0.2	11
M	0.58	0.47	0.165	0.9975	0.3935	0.242	0.33	10
F	0.68	0.56	0.165	1.639	0.6055	0.2805	0.46	15
M	0.665	0.525	0.165	1.338	0.5515	0.3575	0.35	18
F	0.68	0.55	0.175	1.798	0.815	0.3925	0.455	19
F	0.705	0.55	0.2	1.7095	0.633	0.4115	0.49	13
M	0.465	0.355	0.105	0.4795	0.227	0.124	0.125	8
F	0.54	0.475	0.155	1.217	0.5305	0.3075	0.34	16
F	0.45	0.355	0.105	0.5225	0.237	0.1165	0.145	8

Project Goal

The goal of this project is to find the best ML model to predict Age. This is a Supervised Learning ML challenge where Age can be treated as a discrete (Classification) or continuous (Regression) problem.

In addition, we will attempt to create a neural network to predict Age.

ML considerations:

- Age as a regression problem
    - Multivariate linear regression
    - Lasso
    - Ridge
    - ElasticNet
- Age as a classification problem
    - Logistic regression
    - KNN
- Applies to regression and classification problems:
    - SVM
    - RandomForest
- Try H2O AutoML

Other considerations

- Build a simple front-end interface to allow users to enter in physical attributes of an abalone and output the predicted age