Spring
2017

Abalone Estimation

SWEN 5931 RESEARCH TOPICS IN SOFTWARE ENGINEERING

RUSHIKESH MANGRULKAR

Estimation model to predict the abalone age

Problem Statement:

We plan to build a Rapid Miner model for predicting the age of abalone from physical measurements. We have been provided 4177 instances of the following input attributes:

Sex, Length, Diameter, Height, Whole length, Shucked weight, Viscera weight, Shell weight and Class.

Overview:

We follow the CRISP – DM methodology for achieving the prediction, the steps involved in the CRISP – DM methodology are:

a. Business Understanding
b. Data Understanding
c. Data Preparation
d. Modeling
e. Evaluation
f. Deployment

We will primarily focus only on steps 2 – 5 for this project.

a. **Data Understanding:** On observing the data at first instance, the good thing was none of the attributes had any missing values. On further investigating the data, it was founded that ‘Height’ column holds ‘0’ values for some instances. This makes those rows invalid. Therefore, we further eliminate such entries. [1]

b. **Data Preparation:** The data had a column ‘Sex’ which included nominal entries of ‘M, F and I’. In order to have this attribute to be used as dependent variable(attribute), the attribute need to be converted to numerical. Thus, the sex attribute is split into three columns. Furthermore, the problem statement asked to eliminate rows with certain classes 1,2,24,25,26,27,28 and 29. Therefore we used ‘Filter Example’ operator to eliminate those entries.

The next operator introduced is ‘Select Attribute’ which is used to eliminate the ‘Sex’ with nominal data as discussed earlier. The last stage of the data preprocessing is setting the role of the attribute. The Class here is set to be the label, which will work as dependent variable and will be used for prediction.

c. **Modeling:** The first operator used for modeling is the 'Split Data', where we split data into 70 percent for training and 30 percent for testing. The default 'shuffled sampling' be selected in order to allow system to pick the data parts randomly. The machine learner operator we use is 'Linear Regression', for which the input received is the training data from the 'Split Data' operator.

The next operator in the flow is the 'Apply model' operator. This operator receives the data for testing also known as unlabeled data from the 'Split Data', and the model data from the 'Linear Regression' operator.

Lastly, we use 'Performance Regression' operator to check the performance of our prediction. This operator accepts input from the 'Apply Model' operator which is labeled data from the 'Apply Model' in this case. We are using two criterions of the 'Performance Regression' to check monitor the performance namely, 'root mean square error' and 'square correlation'.

The performance output from the 'Performance Regression' operator connects to the result. In order to have a prediction versus actual data overview from the test dataset, we connect the examples from the performance operator to the result.

d. **Data Evaluation:**

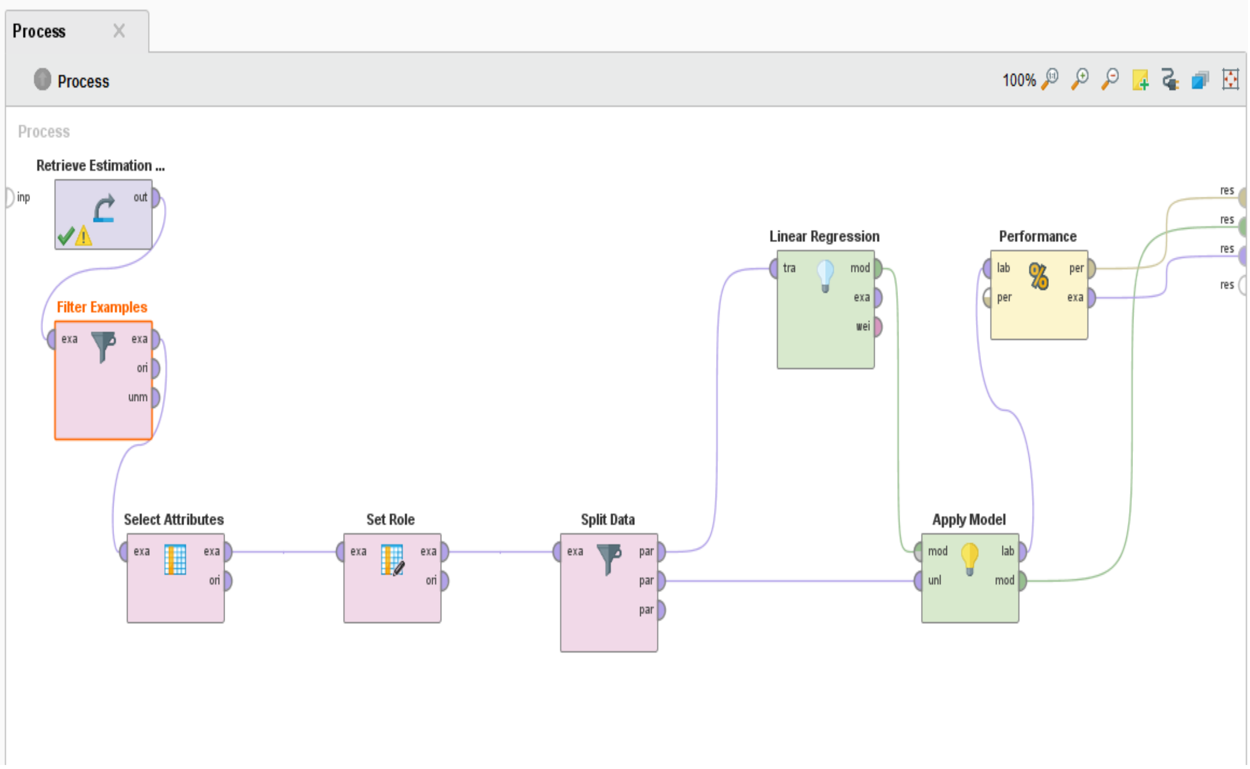
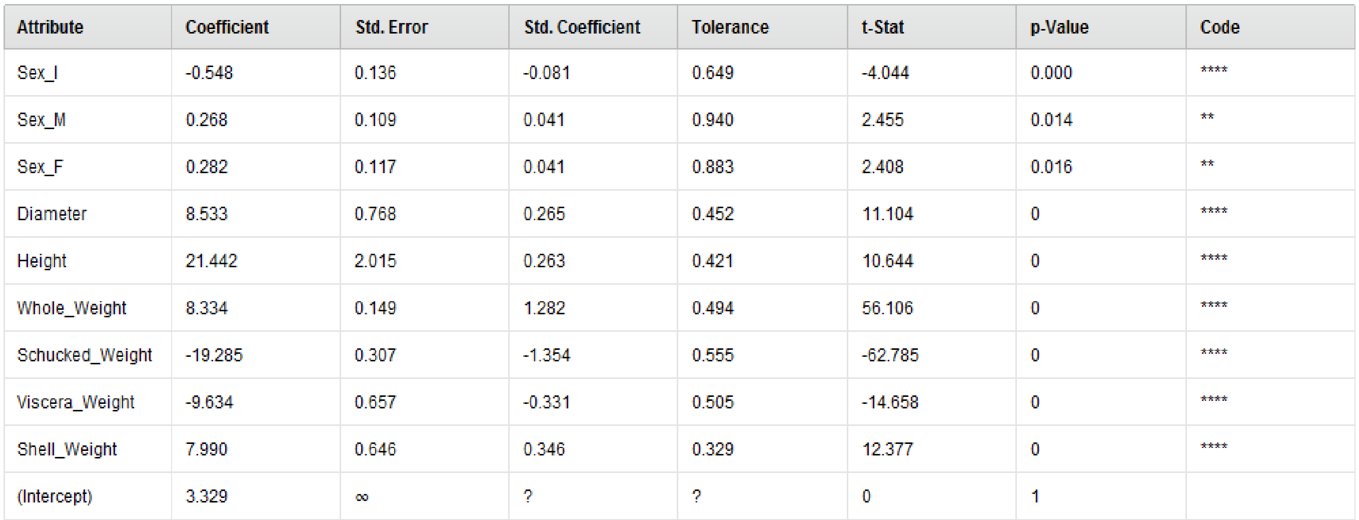


Figure 1.0: Process to predict abalone age

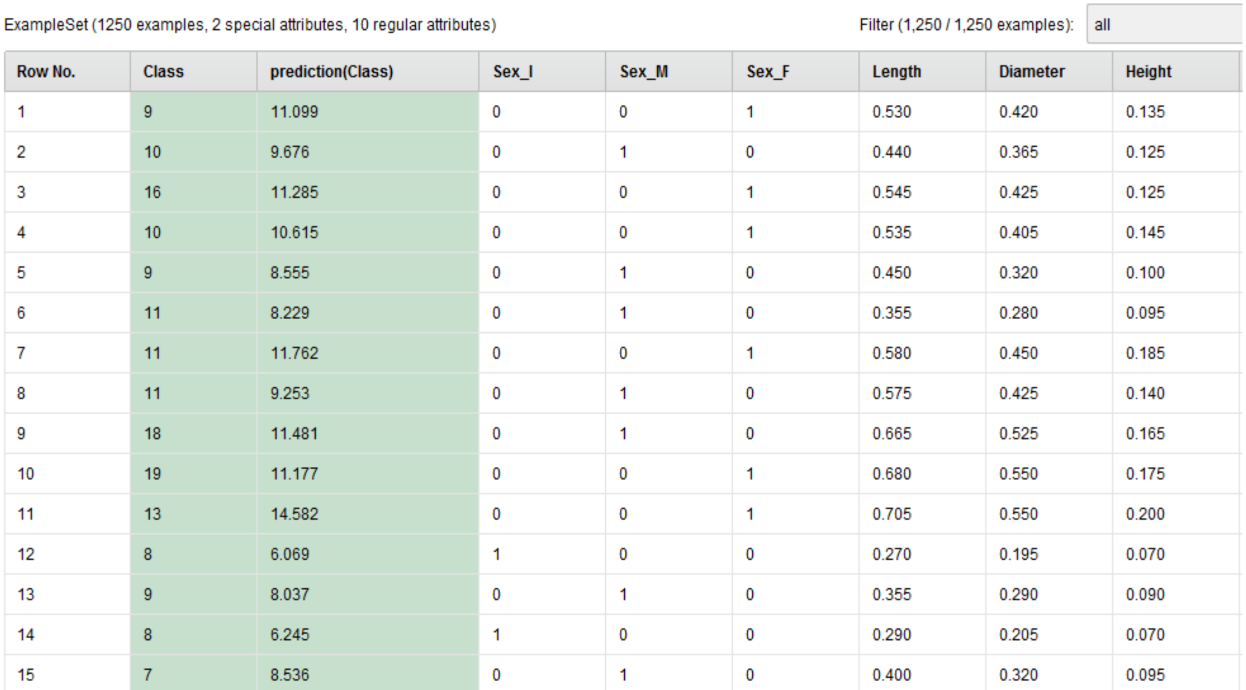
Results:



Attribute	Coefficient	Std. Error	Std. Coefficient	Tolerance	t-Stat	p-Value	Code
Sex_I	-0.548	0.136	-0.081	0.649	-4.044	0.000	****
Sex_M	0.268	0.109	0.041	0.940	2.455	0.014	**
Sex_F	0.282	0.117	0.041	0.883	2.408	0.016	**
Diameter	8.533	0.768	0.265	0.452	11.104	0	****
Height	21.442	2.015	0.263	0.421	10.644	0	****
Whole_Weight	8.334	0.149	1.282	0.494	56.106	0	****
Schucked_Weight	-19.285	0.307	-1.354	0.555	-62.785	0	****
Viscera_Weight	-9.634	0.657	-0.331	0.505	-14.658	0	****
Shell_Weight	7.990	0.646	0.346	0.329	12.377	0	****
(Intercept)	3.329	∞	?	?	0	1	

Table 01

The Coefficient column explains if the attributes has a negative or positive influence on the predictive variable i.e Class in this case. As seen above, the Sex_I, Whole_weight and Viscera_weight have negative influence over the Class(Age) of the car.



ExampleSet (1250 examples, 2 special attributes, 10 regular attributes)

Filter (1,250 / 1,250 examples): all

Row No.	Class	prediction(Class)	Sex_I	Sex_M	Sex_F	Length	Diameter	Height
1	9	11.099	0	0	1	0.530	0.420	0.135
2	10	9.676	0	1	0	0.440	0.365	0.125
3	16	11.285	0	0	1	0.545	0.425	0.125
4	10	10.615	0	0	1	0.535	0.405	0.145
5	9	8.555	0	1	0	0.450	0.320	0.100
6	11	8.229	0	1	0	0.355	0.280	0.095
7	11	11.762	0	0	1	0.580	0.450	0.185
8	11	9.253	0	1	0	0.575	0.425	0.140
9	18	11.481	0	1	0	0.665	0.525	0.165
10	19	11.177	0	0	1	0.680	0.550	0.175
11	13	14.582	0	0	1	0.705	0.550	0.200
12	8	6.069	1	0	0	0.270	0.195	0.070
13	9	8.037	0	1	0	0.355	0.290	0.090
14	8	6.245	1	0	0	0.290	0.205	0.070
15	7	8.536	0	1	0	0.400	0.320	0.095

Table 02 Test Data Set (30 percent of the data)

References:

[1] Linear Regression in R: Abalone Dataset, San Diego State University http://scg.sdsu.edu/linear-regression-in-r-abalone-dataset/

[2] Building Linear Regression Models using Rapid Miner Studio. Author: Pallab Sanyal https://www.youtube.com/watch?v=U9kwqBDIiZ4

[3] Predicting the age of abalone. Author: Joseph Janovsky http://www.slideshare.net/hyperak/predicting-the-age-of-abalone