Arvind Maurya – AIMLCEP-Batch03

Assignment report for Q1:

==========

1.a.i. [C, R] Report the value of beta, RMSE and R2 values for train data set.

BETA:

Following are the beta value obtained from formula and sklearn library.

Parameters	**From Formula:**	**From SKlearn library**
intercept:	[-740.77633006]	[-740.77633006]
coefficients:	[[2.4173551] [27.8581628] [2.67475542] [13.06363395]]	[[2.4173551 27.8581628 2.67475542 13.06363395 0.]]

RMSE and R^2:

Parameters	**From Formula:**	**From SKlearn library**
R^2 Score	0.86381538	0.8638153791855696
RMSE	130.27007649	130.2700764891798

==========

1.a. ii. [R, C] Justify with appropriate reasons if the response variable Weight g in the training data set given in Q1 train.xlsx file can (or) cannot be modeled using a linear regression model identified by beta_0. Using beta_0, predict the Weight g value for the test data set in Q1 test.xlsx file and report the predictions.

Below is the prediction and statistics obtained from test dataset. We can see from the model that we can very well predict weight_g using linear regression model

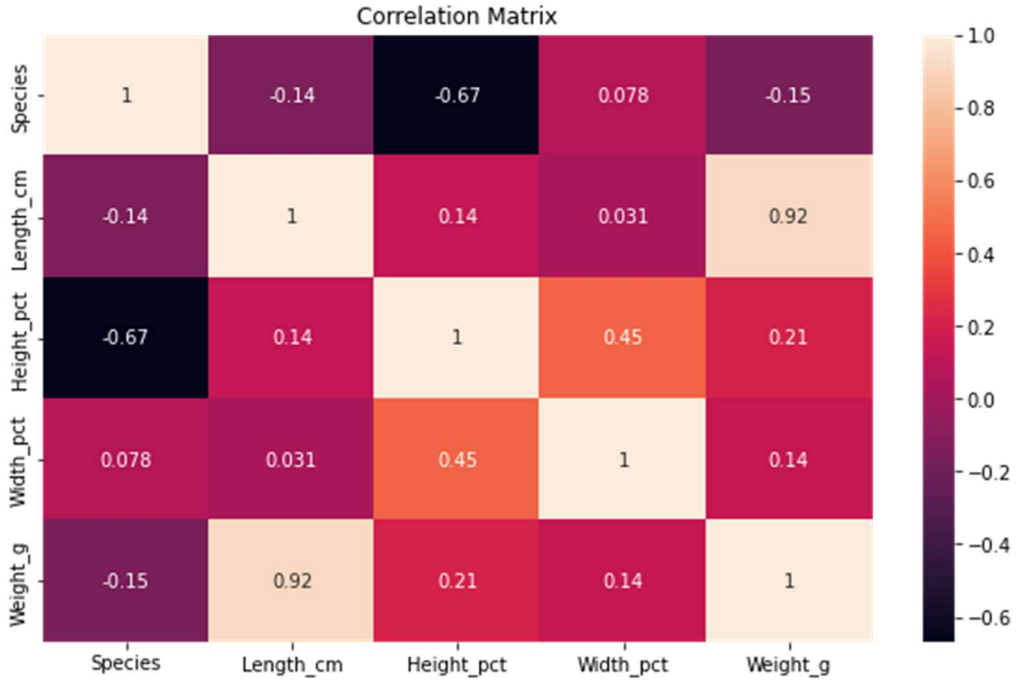
R^2 Score	R^2 score for test: [1.]
RMSE Value	Root Mean Squarred Error for test (RMSE): [0.]
Y Predicted Value	[518.67328896] [648.67882715] [837.39061103] [708.40728589] [526.46342929] [547.88632204] [207.17897006] [362.42661677] [282.50572739] [106.38663939] [215.56548102] [96.87962494] [-202.43397682] [-190.5206425] [-202.27791305] [1222.18243728] [708.17481742] [741.4151994] [765.35407437] [418.18557856] [98.60146593] [183.56236428] [188.6907456] [234.38373268] [212.06618812] [208.98421728] [353.13360178] [862.13526411] [383.2596555] [189.77831477] [749.44259551] [756.02210161] [-15.63034889]

==========

<u>1.a.iii:</u> [C, R] From the components of beta_0, can you compare the relative significance of the predictor variables and identify two most significant attributes? Using only the two most significant attributes, solve the OLSR problem and report the RMSE and R2 values for train data set. Explain using your results if the less significant attributes can be removed from the data set altogether.

Correlation Check: Correlation helps us investigate and establish relationships between variables Note that high amount of correlation between independent variables suggest that linear regression estimation will be unreliable

Text(0.5, 1.0, 'Correlation Matrix')



Looking at the beta value and correlation matrix, i see that the most significant attributes is ***length_cm*** and ***width_pct.***

Train Dataset beta, RMSE and R^2 Score by removing less significant column:

Parameters	**Values**
Beta	Beta from matrix multiplication: [[28.02625976] [17.87051592] [-726.86845722]] Beta from numpy linalg: [[28.02625976] [17.87051592] [-726.86845722]] Beta from scipy linalg: [[28.02625976] [17.87051592] [-726.86845722]]

RMSE	[131.24188644]
R^2 Score	[0.86177594]

Test Dataset beta, RMSE and R^2 Score by removing less significant column is as below:

Parameters	**Values**
RMSE	[0.]
R^2 Score	[1.]

Looking at the result, we can conclude that removing the less significant column increase the prediction for test data set.

==========

1.(b) ii. [C, R] Let the optimal beta corresponding to a particular value of lambda be called beta*_lambda. Report the value of beta_lambda and the RMSE and R2 values for train data set.

Lambda	Beta	RMSE	R^2 Score
0.001	[[2.40478713] [27.85678005] [2.67147441] [13.05370527] [-740.43836035]]	[130.27008424]	[0.86381536]
0.01	[[2.29219272] [27.84439174] [2.64208025] [12.96475214] [-737.41047467]]	[130.27084538]	[0.86381377]
0.1	[[1.21522313] [27.72585857] [2.36091552] [12.11355834] [-708.44204906]]	[130.34101504]	[0.86366702]
1.0	[[-6.19203216e+00] [2.69078413e+01] [4.26483569e-01] [6.23390156e+00] [-5.08730302e+02]]	[133.87466915]	[0.85617461]
10.0	[[-19.65894758] [25.35066218] [-3.10719032] [-5.08182166] [-133.87937211]]	[153.24500915]	[0.81154338]
100.0	[[-20.30951741] [24.61182352] [-3.51646317] [-10.51468709] [-16.87109099]]	[162.15502071]	[0.7889917]
1000.0	[[-11.27933917] [22.49551675] [-3.03448488] [-10.29326243] [-2.27867061]]	[167.89762127]	[0.77378166]

Conclusion: We observed that beta, RMSE, R^2 Score value obtained for lambda = 0.001 is optimal β

1.(b) iii. [C, R] Using beta_lambda, predict the Weight_g value for the test data set in Q1 test.xlsx file and report the predictions.

We observed that beta value obtained for lambda = 0.001 is optimal β

Lambda	Beta	RMSE	R^2 Score
0.001	[[2.40478713] [27.85678005] [2.67147441] [13.05370527] [-740.43836035]]	[130.27008424]	[0.86381536]

Predicted y value is:	[[518.68110615] [648.68395288] [837.39134134] [708.40472459] [526.47236814] [547.85683864] [207.21110456] [362.43386168] [282.53750383] [106.43527726] [215.54226635] [96.8831972] [-202.32905408] [-190.41637688] [-202.16471162] [1222.21197227] [708.22764526] [741.44745265] [765.27488915] [418.16288197] [98.57685156] [183.53639915] [188.66344985] [234.368867] [212.05307629] [208.9835037] [353.11368766] [862.05601392] [383.23244533] [189.76654179] [749.37477252] [755.95398494] [-15.63181114]]

1.(c) [R] Justify with proper reasons which value of λ can be considered to be the best value for the OLSR model.

Looking at the above, we found that for λ = 0.001 we are getting best β and subsequently getting best value for prediction of Y response variable