IC 272: DATA SCIENCE - III
LAB ASSIGNMENT – V
Data classification using Bayes classifier with Gaussian mixture model (GMM); regression using linear regression and polynomial curve fitting

Student's Name: Vishal Sharma

Mobile No: 9540140310

Roll Number: B20239

Branch: Data Science and Engineering

PART - A

1 a.

		Prediction Outcome
True Label	106	12
	4	215

Figure 1 Bayes GMM Confusion Matrix for Q = 2

		Prediction Outcome
True Label	106	12
	3	216

Figure 2 Bayes GMM Confusion Matrix for Q = 4

		Prediction Outcome
True Label	103	15
	5	214

Figure 3 Bayes GMM Confusion Matrix for Q = 8

		Prediction Outcome
True Label	98	20
	3	216

Figure 4 Bayes GMM Confusion Matrix for Q = 16

b.

Table 1 Bayes GMM Classification Accuracy for Q = 2, 4, 8 & 16

Q	Classification Accuracy (in %)
2	95.3
4	95.5
8	94.1
16	93.2

Inferences:

1. The highest classification accuracy is obtained with Q =.4
2. Increasing the value of Q, increases the prediction accuracy first, but then it starts to decrease.
3. This is because nodes were added with less wright. Thus, it causes the model to fir for the training data more.

4. As the prediction accuracy increases, there is an increase in the values of diagonal elements.
5. Increases in the values of diagonal elements is the reason behind the increase of prediction accuracy.
6. With the increase in value of Q, the value of non-diagonal elements decreases.
7. The value of non-diagonal elements decreases because the value of accuracy increases.

2

Table 2 Comparison between Classifiers based upon Classification Accuracy

S. No.	Classifier	Accuracy (in %)
1.	KNN	89.6
2.	KNN on normalized data	97.3
3.	Bayes using unimodal Gaussian density	94.3
4.	Bayes using GMM	95.5

Inferences:

1. The highest accuracy is observed for KNN classifier on normalized data (97.3%) while the lowest is with KNN classifier (89.6%)
2. KNN classifier < Bayes unimodal Gaussian density classifier < Bayes classifier using GMM < KNN on normalized data.
3. KNN works better with normalized data, as attributes with much larger values cannot influence the others. Also, as only two clusters are involved, KNN Classifier will give better results than Bayes classifier with GMM.

PART – B

1

a.

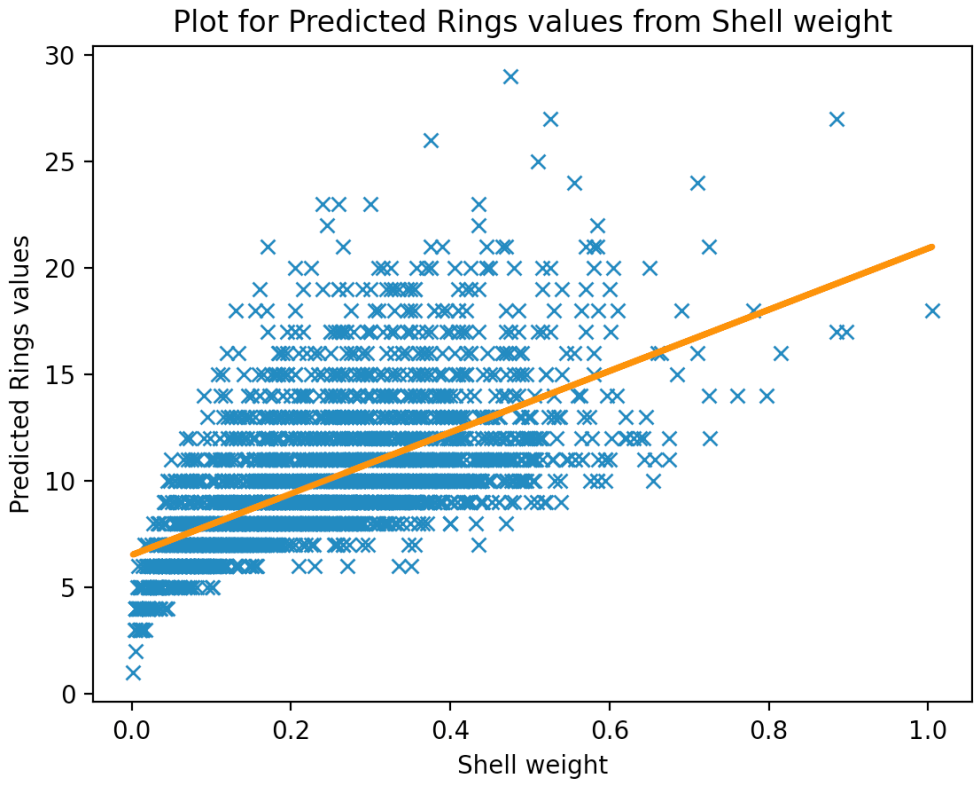

Figure 5 Univariate linear regression model: Rings vs. the chosen attribute name (replace) best fit line on the training data

Inferences:

1. The attribute with highest correlation is used for predicting the data because the target attribute is more dependent on the attribute having highest correlation, thus providing accurate results.
2. No, the line of best fit does not fit the training data perfectly.
3. The bias for the best line of fit is quite high, while the variance is quite low.

b.

Report the prediction accuracy on training data, using RMSE is 2.528

c.

Report the prediction accuracy on testing data, using RMSE is 2.468

Inferences:

1. The accuracy of testing data is greater.
2. It is due to the fact that testing data fits better in the model.

d.

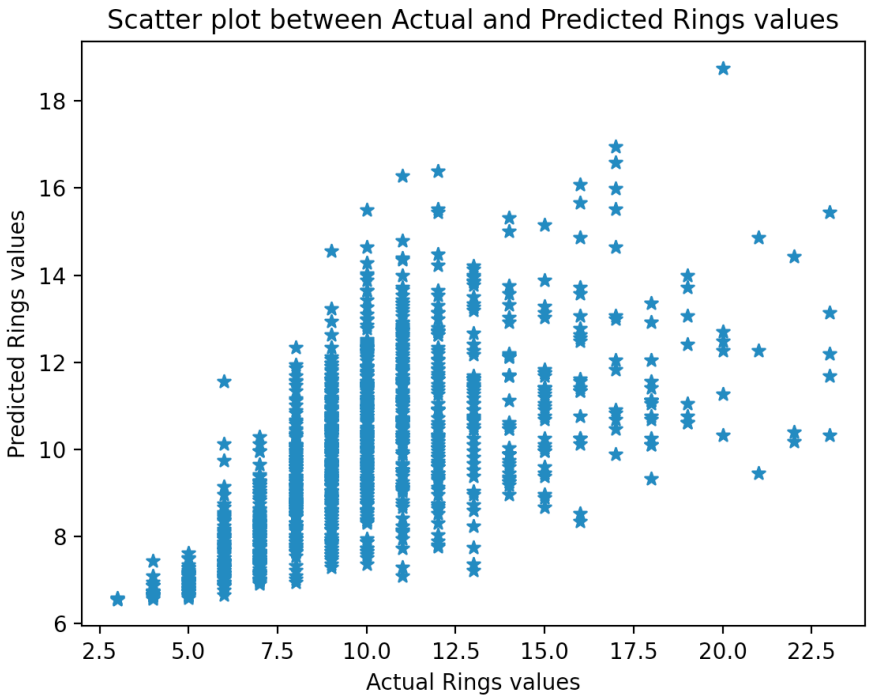

Figure 6 Univariate linear regression model: Scatter plot of predicted rings from linear regression model vs. actual rings on test data

Inferences:

1. Based on the spread of the data, the predicted values of Rings is not very accurate.
2. Because the spread of actual data in around 2 – 23, while that of predicted values is 6 -20.

2

a.

Report the prediction accuracy on training data is 2.216

b.

Report the prediction accuracy on testing data is 2.205

Inferences:

1. The prediction accuracy for both training and testing dataset is almost same.
2. It is because the model was trained for training data.

c.

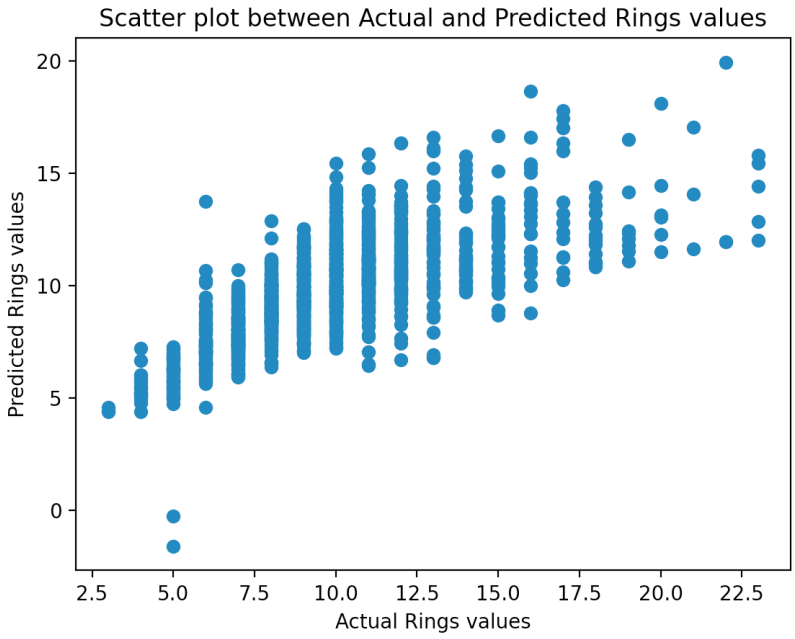

Figure 7 Multivariate linear regression model: Scatter plot of predicted rings from linear regression model vs. actual rings on test data

Inferences:

1. Based on the spread of the data, it can be inferred that the accuracy for predicted values is quite high.
2. The spread of predicted data is 4.8 – 22, while that from actual dataset is 5 – 23.
3. Univariate linear regression model is not accurate as the Multivariate linear regression.

3

a.

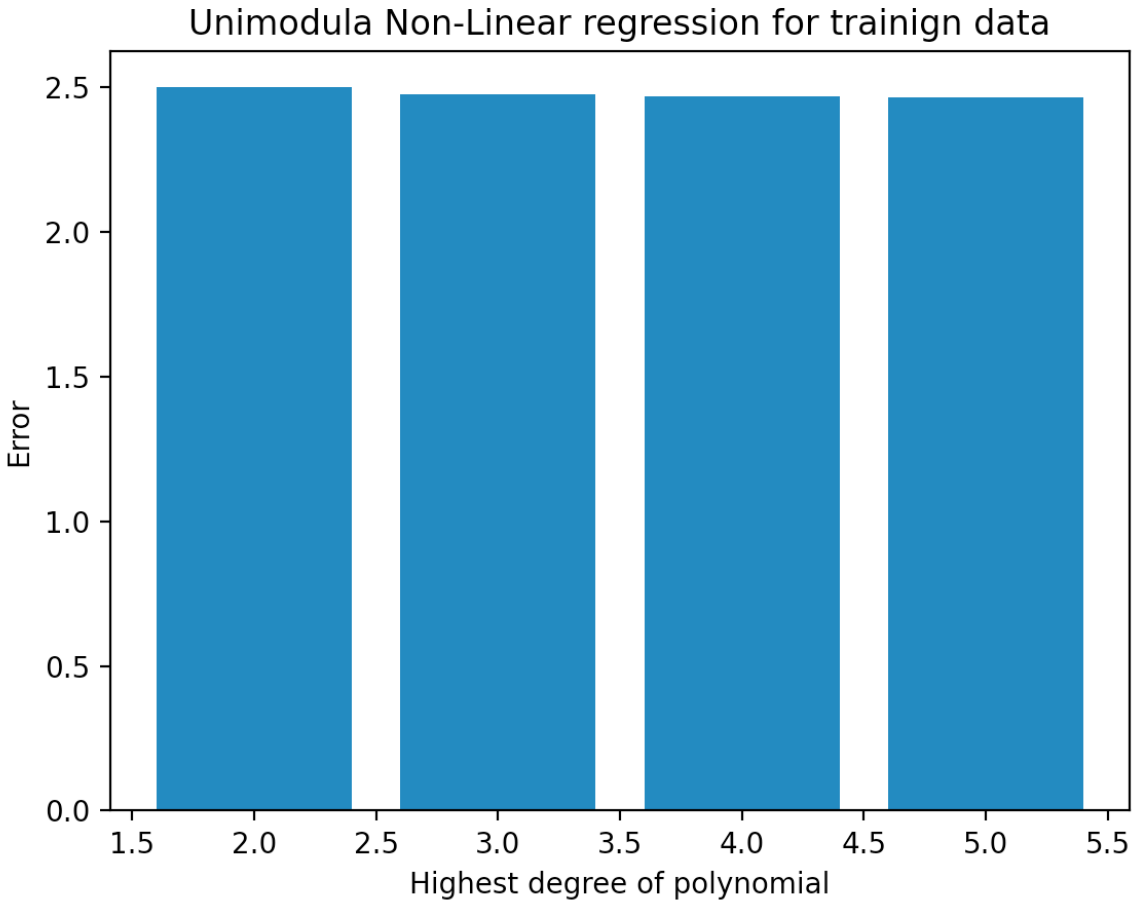

Figure 8 Univariate non-linear regression model: RMSE vs. different values of degree of polynomial (p = 2, 3, 4, 5) on the training data

Inferences:

1. The value of RMSE value decreases as the value of degree of polynomial increases.
2. The decrease from p = 2 to p = 3 is more, and for rest, decrease is gradual.
3. As degree of polynomial increases, the curve fits the data more accurately.
4. From the RMSE values, it can be inferred that p = 5 will approximate the data best.
5. As the degree of polynomial increases, the bias decreases and variance increases.

b.

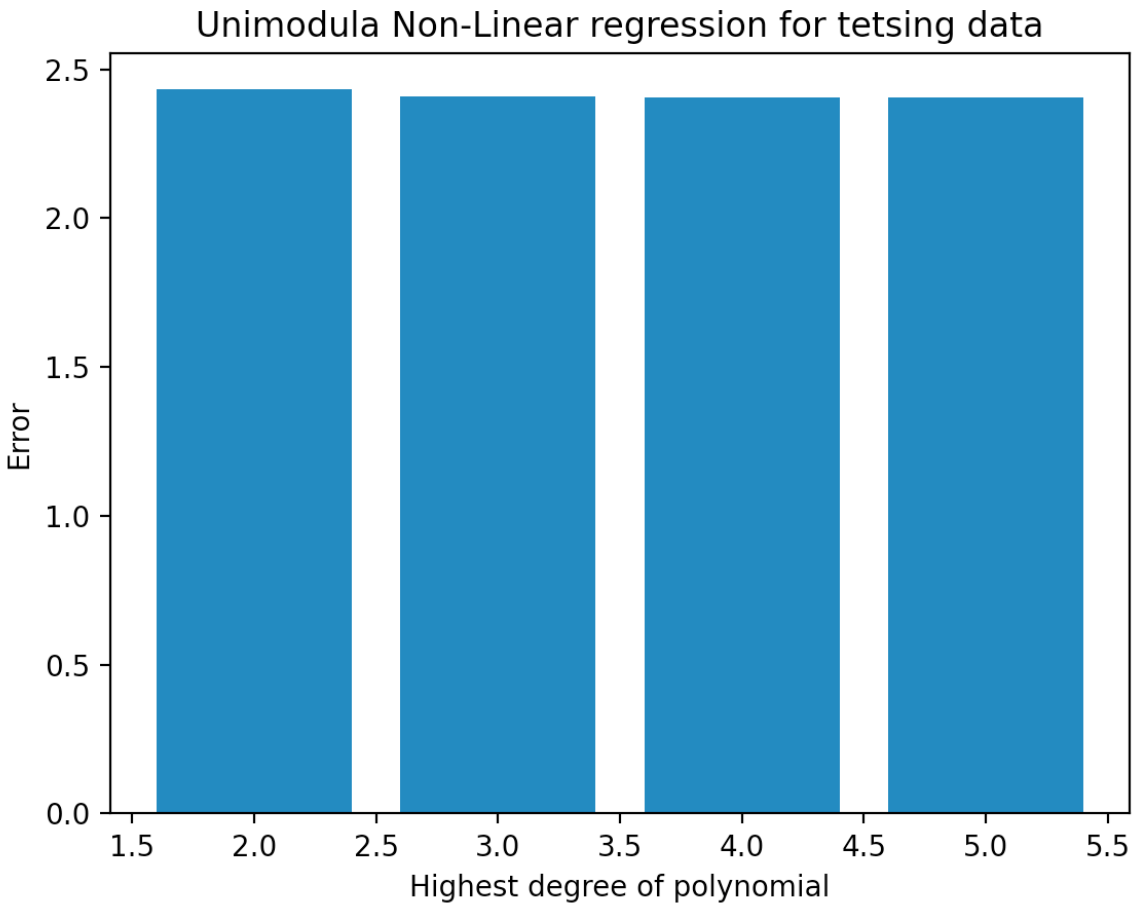

Figure 9 Univariate non-linear regression model: RMSE vs. different values of degree of polynomial (p = 2, 3, 4, 5) on the test data

Inferences:

1. The value of RMSE decreases as the degree of polynomial increases.
2. The decrease in value of RMSE is more from p = 2 to p =3, but after that, decrease in the values becomes more gradual.
3. As degree of polynomial increases, the curve fits the data more accurately.
4. From the RMSE values, it can be inferred that p = 5 will approximate the data best.
5. As the degree of polynomial increases, the bias decreases and variance increases.

c.

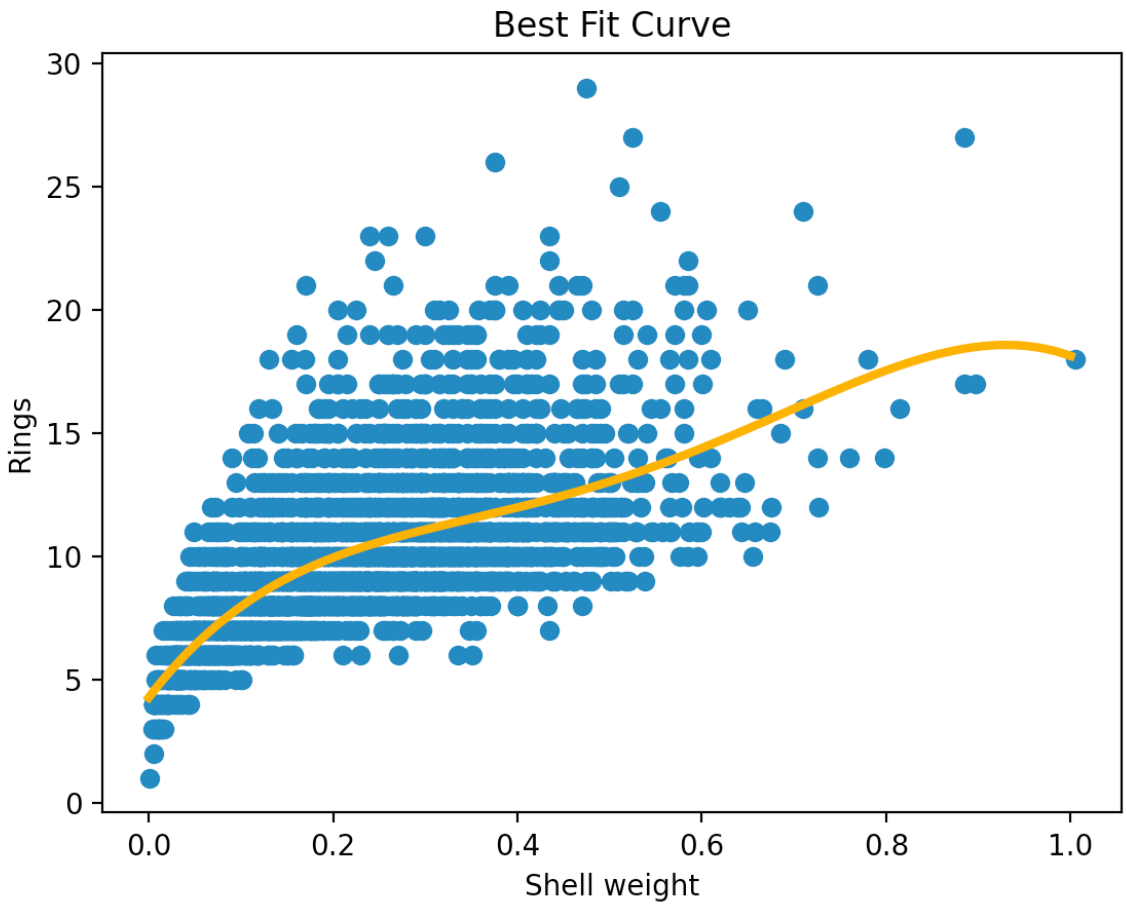

Figure 10 Univariate non-linear regression model: Rings vs. chosen attribute(replace) best fit curve using best fit model on the training data

Inferences:

1. The value of p for the best fit model is 4.
2. The value of p = 4 is chosen because it fits the data more and has more variance.
3. The bias decreases and variance increase with increasing values of p.

d.

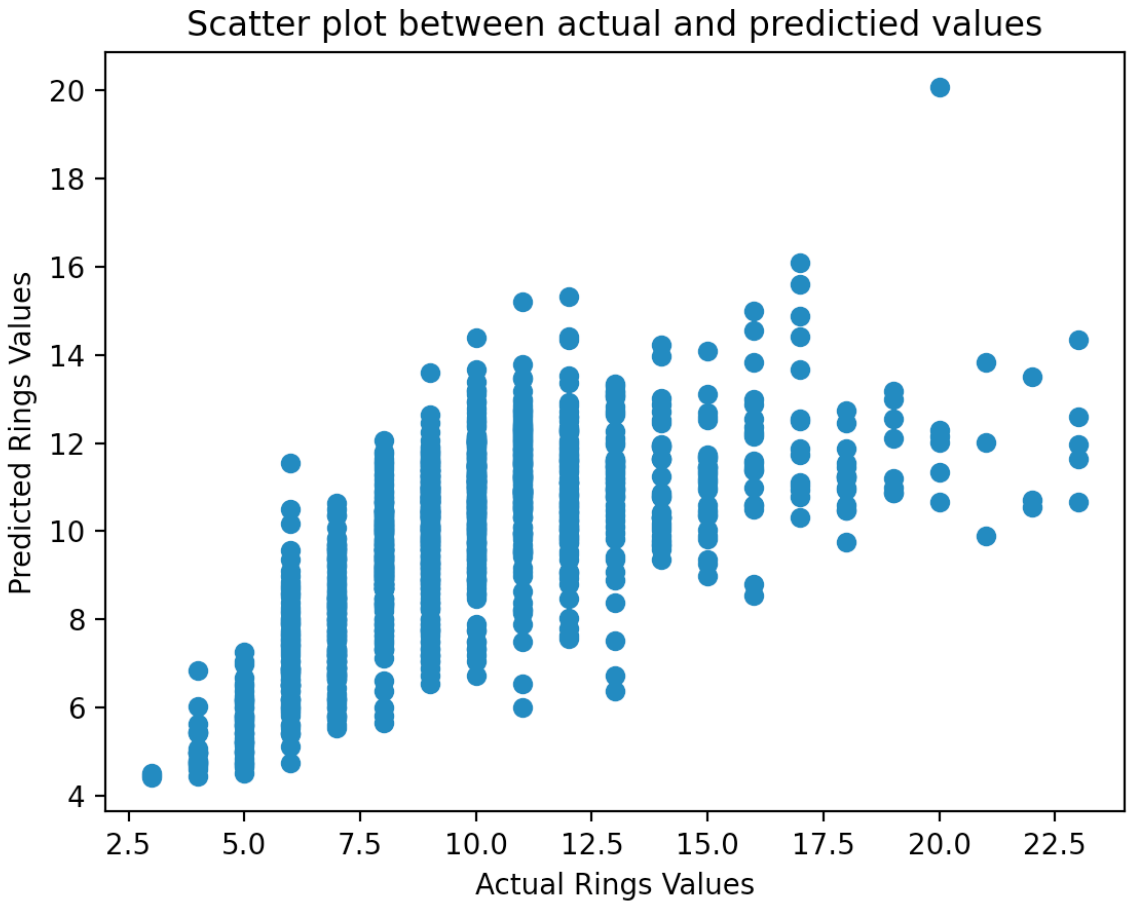

Figure 11 Univariate non-linear regression model: Scatter plot of predicted rings vs. actual rings on test data

Inferences:

1. Based on the spread of points, the predicted values of rings is quite accurate.
2. From the above plot, it can be stated that spread of predicted values is from 4 – 20, while that for actual value is 3 - 23
3. From the scatter plot it can be inferred that the accuracy of Univariate non – linear model > Multivariate Linear Model > Univariate Linear Model.
4. The value of RMSE error for non-linear model is less than that of linear model.
5. For linear regression models, the bias is high, and variance is low, but for non-linear regression model, the bias is low while the variance is high.

4 a.

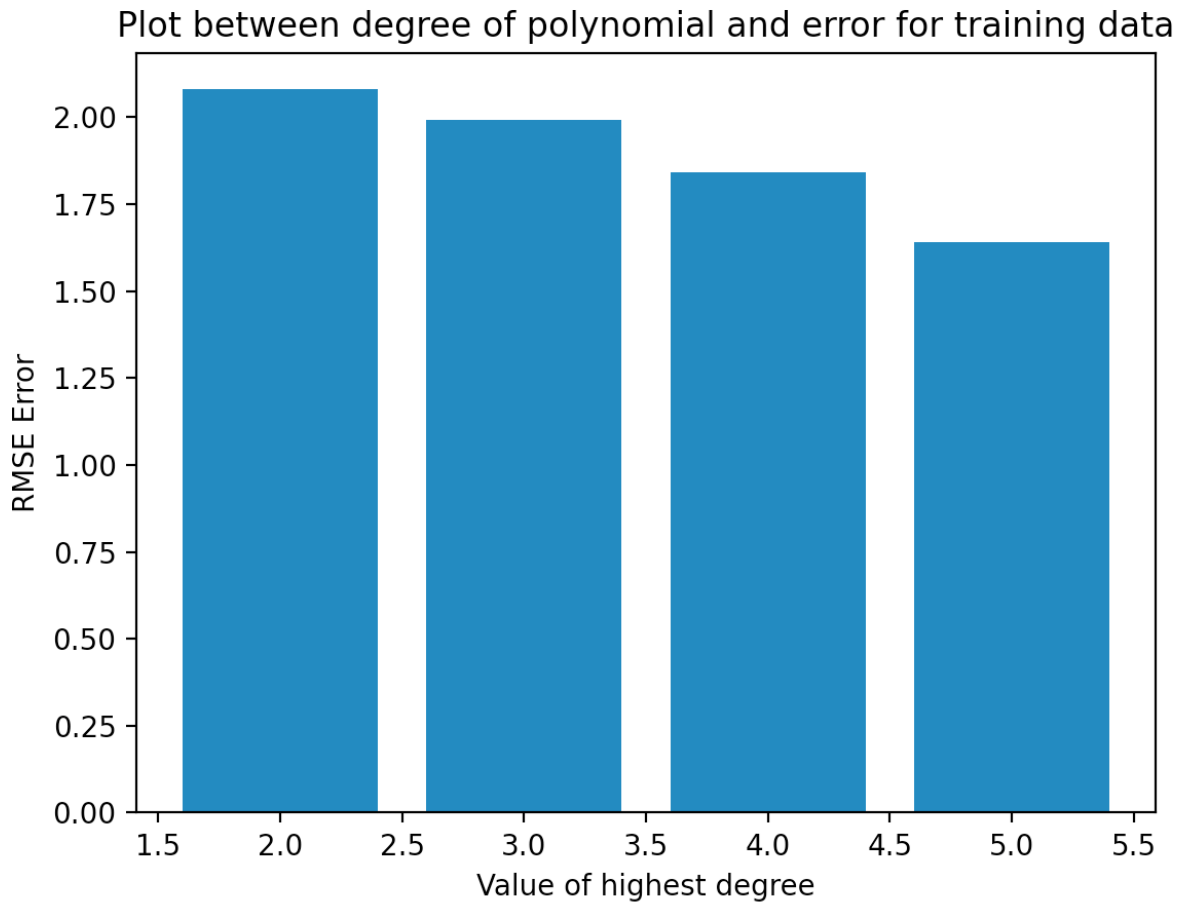

Figure 12 Multivariate non-linear regression model: RMSE vs. different values of degree of polynomial (p = 2, 3, 4, 5) on the training data

Inferences:

1. The value of RMSE decreases as the degree of polynomial increases.
2. The decrease in error is gradual from p = 2 to p = 3 to p = 4. But after p = 4, the decrease in value increases.
3. As the degree of polynomial increases, the curve fits more accurately in the data.
4. From the RMSE values, p = 5 will fit the data best as it has the minimum RMSE value
5. With increase in the value of p, the bias decrease and the variance increases.

b.

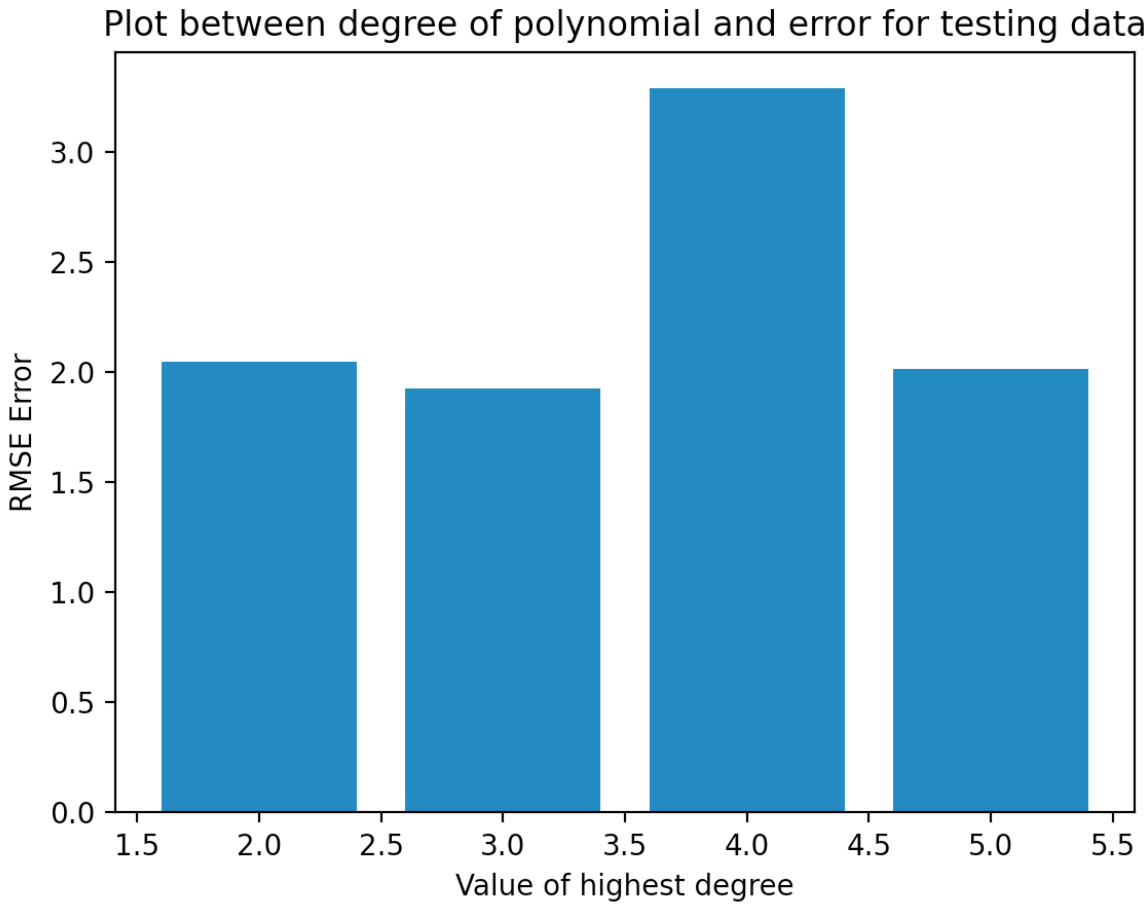

Figure 13 Multivariate non-linear regression model: RMSE vs. different values of degree of polynomial (p = 2, 3, 4, 5) on the test data

Inferences:

1. The value of RMSE error decreases from p = 2 to p = 3 but increases for p = 4.
2. The decrease in error in gradual from p = 2 to p =3, but the increase is much more for p = 4
3. As the degree of polynomial increases, the curve becomes over fitted.
4. From the plot, p = 3 has the minimum value for RMSE value.
5. The bias gradually decreases for p = 3, but suddenly increases afterwards due to over fitting of the curve on data.

c.

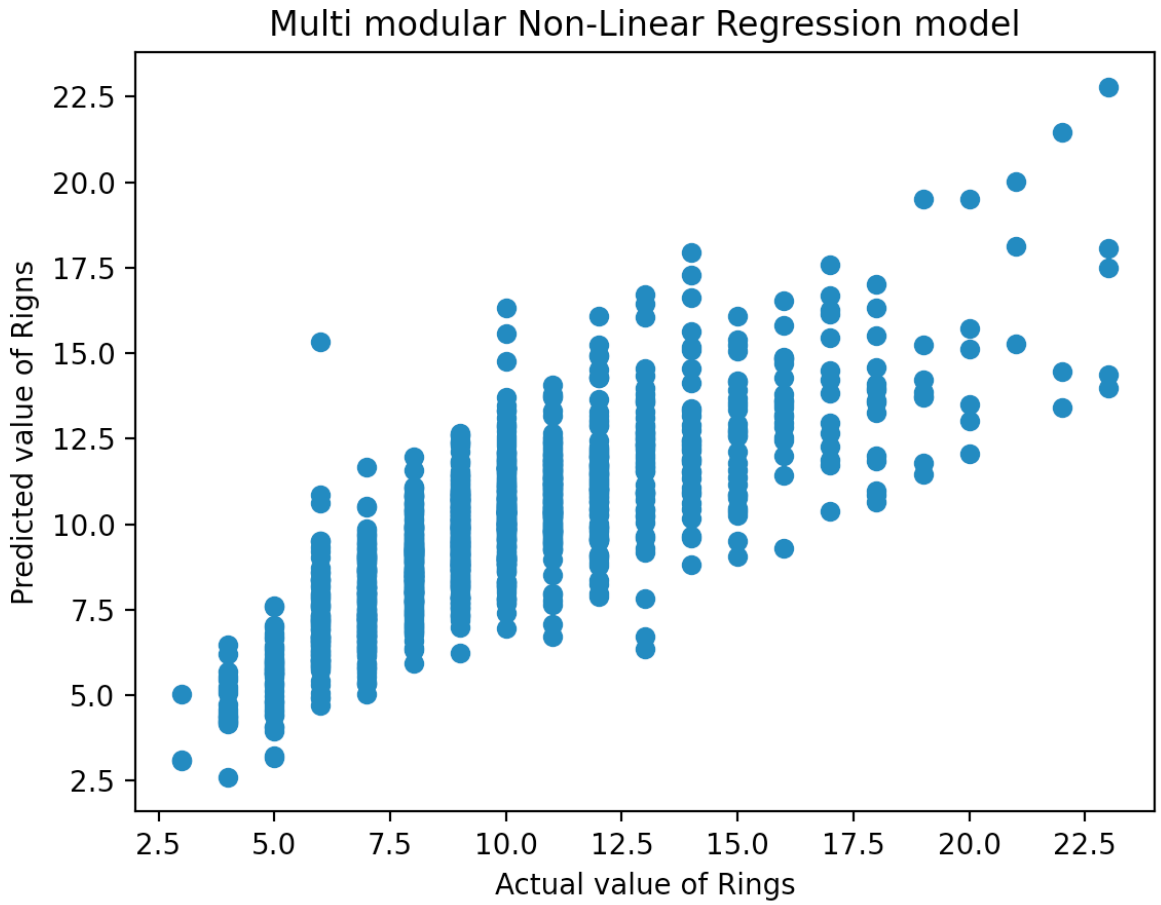

Figure 14 Multivariate non-linear regression model: Scatter plot of predicted rings vs. actual rings on test data

Inferences:

1. Based on the scatter plot, for p =3, the predicted rings are quite accurate.
2. The spread for the predicted values of rings is from 2.5 – 22.5m while that for actual value is from, 2.7 – 22.77
3. Non-linear regression models are more accurate than linear regression models. And multi-variate models give more accurate prediction than univariate models.
4. RMSE values for non-linear models is less than linear models.
5. For linear regression models, the bias is high, and variance is low, but for non-linear regression model, the bias is low while the variance is high.