Name: Anooshka Bajaj

1 a.

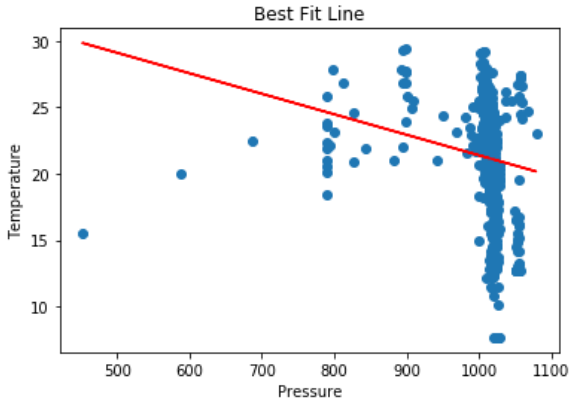

Figure 5 Pressure vs. temperature best fit line on the training data

Inferences:

1. No, the best fit line does not fit the training data perfectly.
2. In linear regression we are trying to find the line which minimizes the squared error. Since the training points are not collinear there can't be a line which passes through all of them. Thus, some other curve is required.
3. Bias is high as the best fit line underfits the data. Variance is low as the bias is high.

b. Prediction accuracy for training data - RMSE: 4.279790433682601

c. Prediction accuracy for test data - RMSE: 4.286985483129509

Inferences:

1. Amongst training and testing accuracy, training accuracy is higher (its RMSE value is lower).
2. Training accuracy is higher because the model is made on the training data and so its RMSE value is slightly lower.

d.

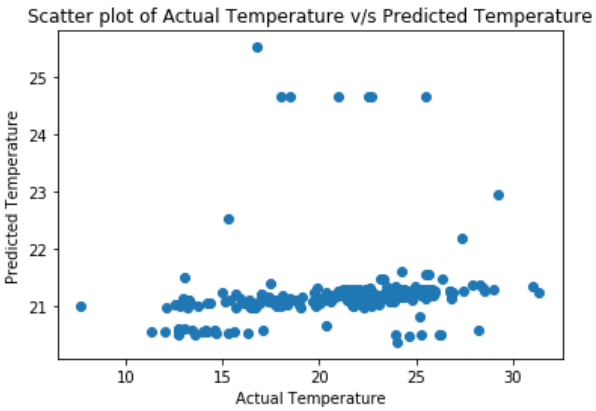

Figure 6 Scatter plot of predicted temperature from linear regression model vs. actual temperature on test data

Inferences:

1. The predicted temperature is not much accurate.
2. The linear model is not a good generalization of the data. The data is underfit, leading to high error.

2 a.

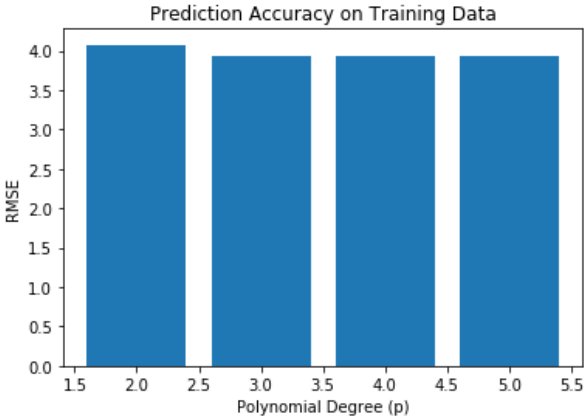

Figure 7 RMSE vs. different values of degree of polynomial (p = 2, 3, 4, 5) on the training data

Inferences:

1. RMSE value decreases with respect to increase in degree of polynomial (p = 2, 3, 4, 5).
2. After p = 3 the decrease becomes gradual.
3. As the degree increases the curve fits the data better, so the RMSE decreases. But after p= 3, it doesn't affect the fit much and RMSE value changes very little.
4. From the RMSE value, degree 3 curve will approximate the data best as after p =3 the decrease in RMSE is gradual and not that significant.
5. The bias decreases and variance increases with the increase in degree of the polynomial as the curve starts fitting the data better. The RMSE decreases significantly from p =2 to p = 3. After that the decrease is more gradual. So, using higher powers for modelling the data can cause the problem of overfitting.

b.

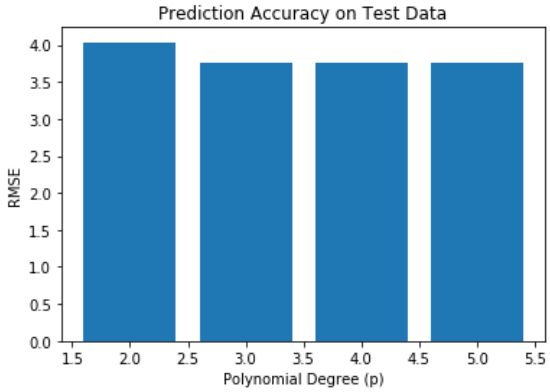

Figure 8 RMSE vs. different values of degree of polynomial (p = 2, 3, 4, 5) on the test data

Inferences:

1. RMSE value decreases with respect to increase in degree of polynomial (p = 2, 3, 4, 5).
2. After p = 3 the decrease becomes gradual.
3. It is similar to the training data. As the degree increases the curve fits the data better, so the RMSE decreases. But after p= 3, it doesn't affect the fit much and RMSE value changes very little.
4. From the RMSE value, degree 3 curve will approximate the data best.
5. The bias decreases and variance increases with the increase in degree of the polynomial as the curve starts fitting the data better. The RMSE decreases significantly from p =2 to p = 3. After that the decrease is more gradual. So, using higher powers for modelling the data can cause the problem of overfitting.

c.

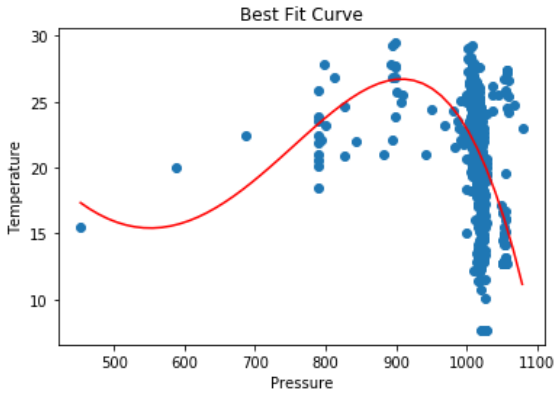

Figure 9 Pressure vs. temperature best fit curve using best fit model on the training data

Inferences:

1. p-value 3 is corresponding to the best fit model.
2. Degree 3 curve gives best fit as the RMSE value is low and variance is high. Moreover, curve with p>3 might overfit the data.
3. As the degree increases the bias decreases and variance increases as the model starts becoming more complex and starts fitting the data better.

d.

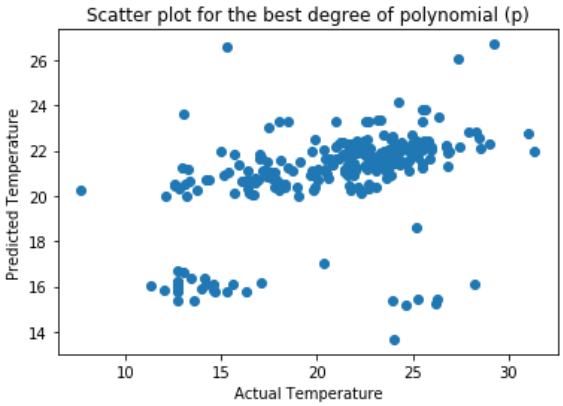

Figure 10 Scatter plot of predicted temperature from non- linear regression model vs. actual temperature on test data

Inferences:

1. Based upon the spread of the points accuracy, we can see that accuracy of predicted temperature is better than that in simple linear regression.
2. The accuracy has increased when using polynomial regression instead of linear regression because the data is not distributed in a linear fashion and this curve fits the data better.
3. Prediction accuracy of non-linear regression is better as the RMSE value is lower for it. Also, from the spread of data we can see that the non-linear regression is better than linear regression.
4. The above inference implies that the distribution of the data is not linear. Higher order polynomial terms are a better representation of the data.
5. In case of the linear regression the bias was high and variance was less when compared to the non-linear models. The data was underfit. Using polynomial regression decreased the bias and increased variance.