Evaluate linear-Regression and Non-Linear Regression model using OverFitting

Norina Akhtar
19643

Table of Content

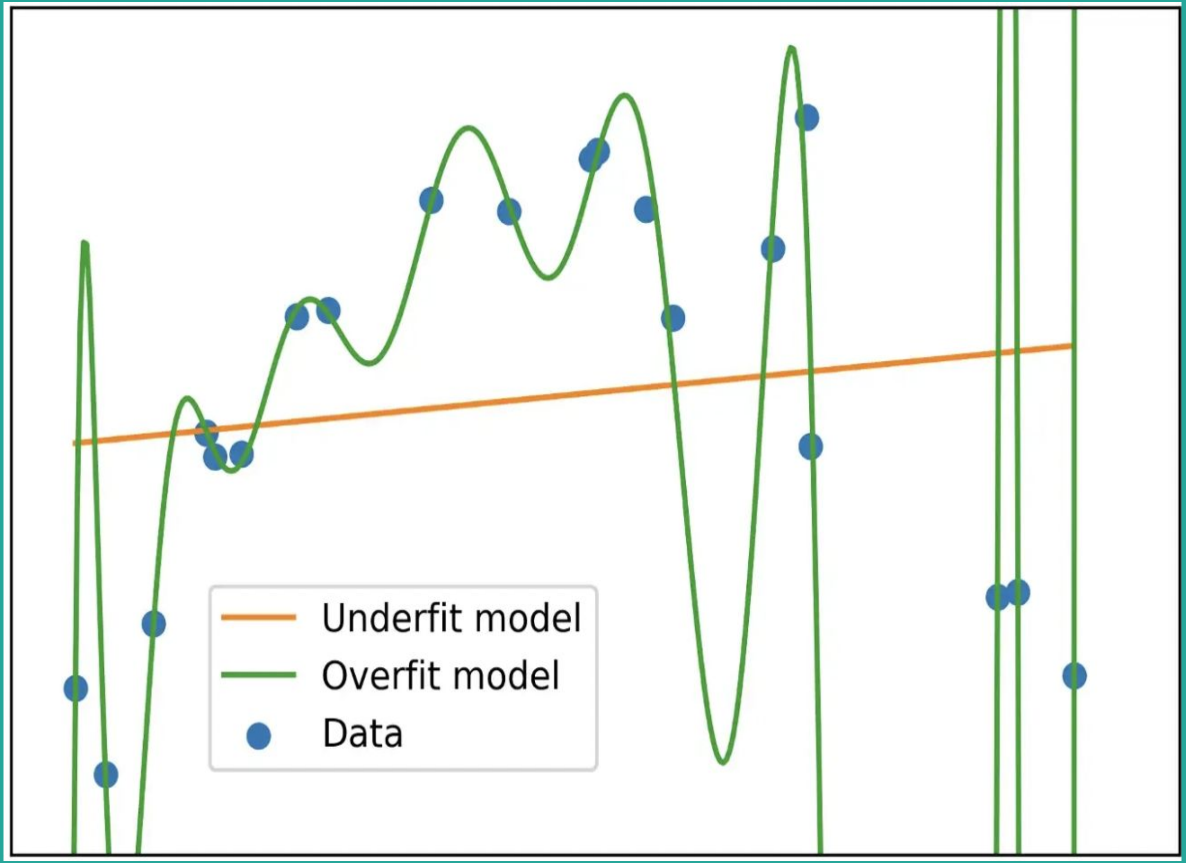
- Introduction
- Design
- Implementation
- Test
- Enhancement Ideas
- Conclusion
- References

Introduction

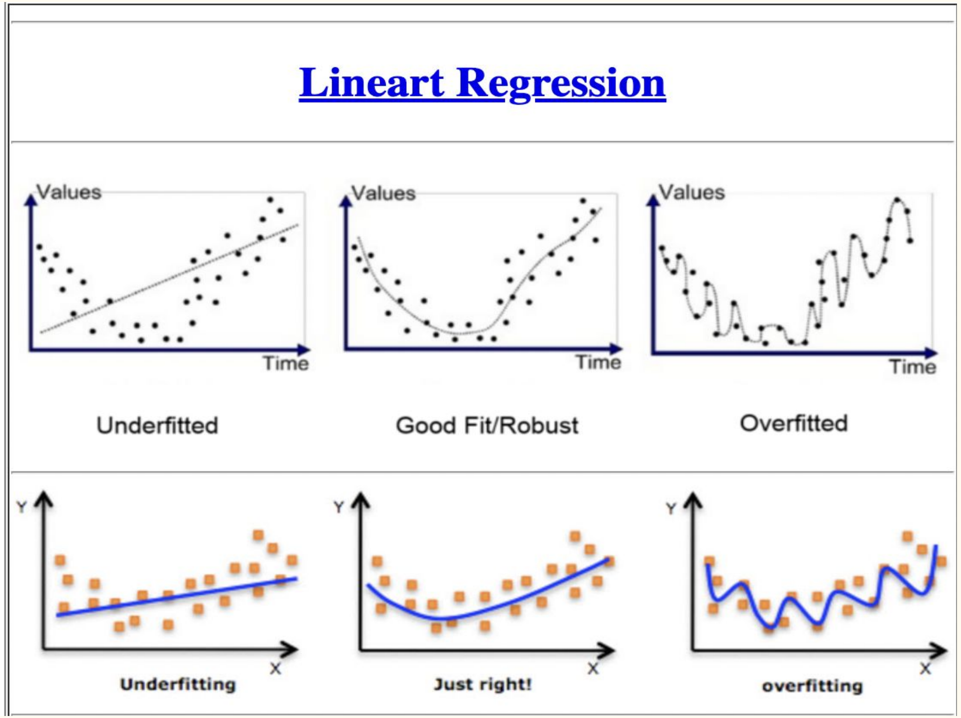
User overfitting to evaluate different models

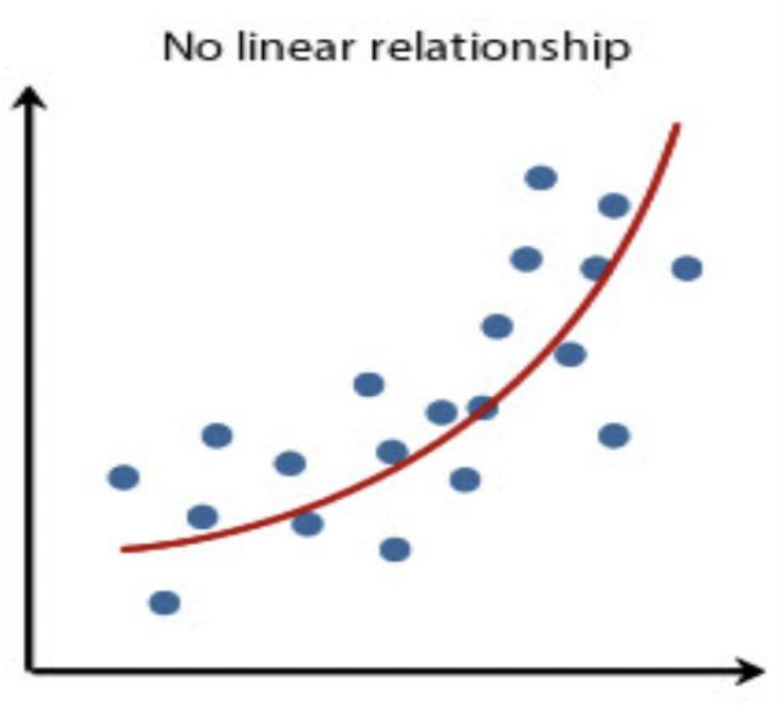
Overfitting, underfitting, and the bias-variance tradeoff are foundational concepts in machine learning. A model is **overfit** if performance on the training data, used to fit the model, is substantially better than performance on a test set, held out from the model training process

At the opposite end of the spectrum, if a model is not fitting the training data very well, this is known as **underfitting**, and the model is said to have **high bias**.


Underfit model
Overfit model
Data

Design


Lineart Regression
Values
Time
Underfitted
Values
Time
Good Fit/Robust
Values
Time
Overfitted
Y
X
Underfitting
Y
X
Just right!
Y
X
overfitting


No linear relationship

Test

Regression Formula

Test Data by Calculating linear regression and Nonlinear Regression

Regression Equation(y) = $a + bx$
Slope(b) = $(N\Sigma XY - (\Sigma X)(\Sigma Y)) / (N\Sigma X^2 - (\Sigma X)^2)$
Intercept(a) = $(\Sigma Y - b(\Sigma X)) / N$

Where:
x and y are the variables.
b = The slope of the regression line
a = The intercept point of the regression line and the y axis.
N = Number of values or elements
X = First Score
Y = Second Score
ΣXY = Sum of the product of first and Second Scores
ΣX = Sum of First Scores
ΣY = Sum of Second Scores
ΣX^2 = Sum of square First Scores

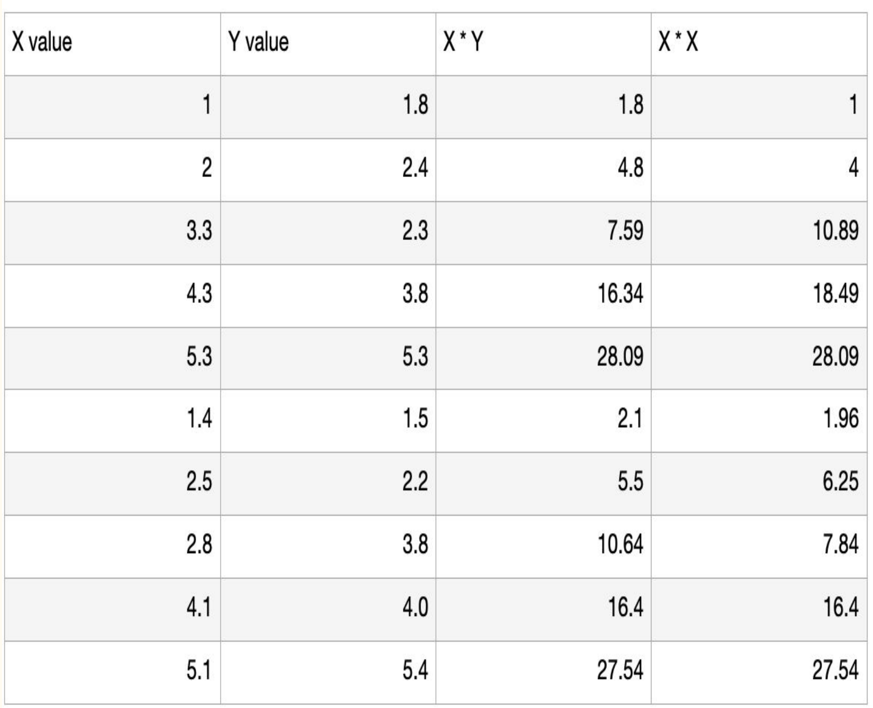
Linear Regression- Training phase model 1:

To find regression equation, we will first find slope, intercept and use it to form regression equation.

Step 1:
Count the number of values. N = 5
Step 2:
Find X * Y, X^2
See the table on right side

Step 3:
Find ΣX, ΣY, ΣXY, ΣX^2.

ΣX = 31.8
ΣY = 32.5
ΣXY = 120.8
ΣX^2 = 121.34



X value	Y value	X * Y	X * X
1	1.8	1.8	1
2	2.4	4.8	4
3.3	2.3	7.59	10.89
4.3	3.8	16.34	18.49
5.3	5.3	28.09	28.09
1.4	1.5	2.1	1.96
2.5	2.2	5.5	6.25
2.8	3.8	10.64	7.84
4.1	4.0	16.4	16.4
5.1	5.4	27.54	27.54

Linear Regression- Training phase model 1:

Step 4:
Substitute in the above slope formula given.

Slope(b) = $(N\Sigma XY - (\Sigma X)(\Sigma Y)) / (N\Sigma X^2 - (\Sigma X)^2)$
= ((10)*(120.8)-(31.8)*(32.5))/((10)*(121.34)-$(31.8)^2$)
= (1208-1033.5)/(1213.4 - 1011.24)
= 174.5/202.16
= 0.86

Step 5:
Now, again substitute in the above intercept formula given.

Intercept(a) = $(\Sigma Y - b(\Sigma X)) / N$
= (32.5 -0.86(31.8))/10
= (32.5 - 27.35)/10
= 5.15/10
= 0.512

Step 6:
Then substitute Intercept(a) and Slope(b) in regression equation formula

Regression Equation(y) = a + bx
= 0.512 + 0.86x.

Training Phase-Non-Linear Regression of Model 2:

Regression Equation(y) = a + b^2

We can still use Linear Regression formula

Slope(b) = (NΣPY - (ΣP)(ΣY)) / (NΣP^2 - $(ΣP)^2$)

Intercept(a) = (ΣY - b(ΣP)) / N

Where P = X * X

Step 1:

Count the number of values. N = 10

Step 2:

Find X * Y, X^2

See the table on right side

Step 3:

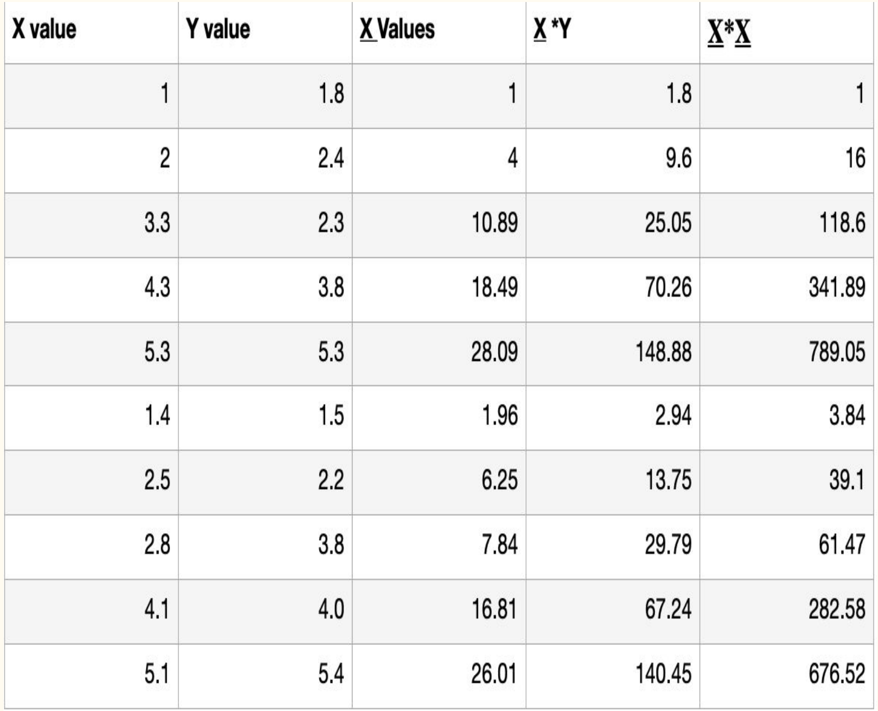
Find ΣX, ΣY, ΣXY, ΣX^2.

ΣX = 121.34

ΣY = 32.5

ΣXY = 509.76

ΣX^2 = 2330.05



X value	Y value	X Values	X *Y	X*X
1	1.8	1	1.8	1
2	2.4	4	9.6	16
3.3	2.3	10.89	25.05	118.6
4.3	3.8	18.49	70.26	341.89
5.3	5.3	28.09	148.88	789.05
1.4	1.5	1.96	2.94	3.84
2.5	2.2	6.25	13.75	39.1
2.8	3.8	7.84	29.79	61.47
4.1	4.0	16.81	67.24	282.58
5.1	5.4	26.01	140.45	676.52

Training Phase-Non-Linear Regression of Model 2:

Step 4:
Substitute in the above slope formula given.

Slope(b) = $(N\Sigma\underline{X}Y - (\Sigma\underline{X})(\Sigma Y)) / (N\Sigma\underline{X}^2 - (\Sigma\underline{X})^2)$
$= ((10)*(509.76)-(121.34)*(32.5))/((10)*(2330.05)-(121.34)^2)$
= (5097.6 -3943.55)/(23300.5 - 14723.4)
=1154.05/8577.1
= 0.134

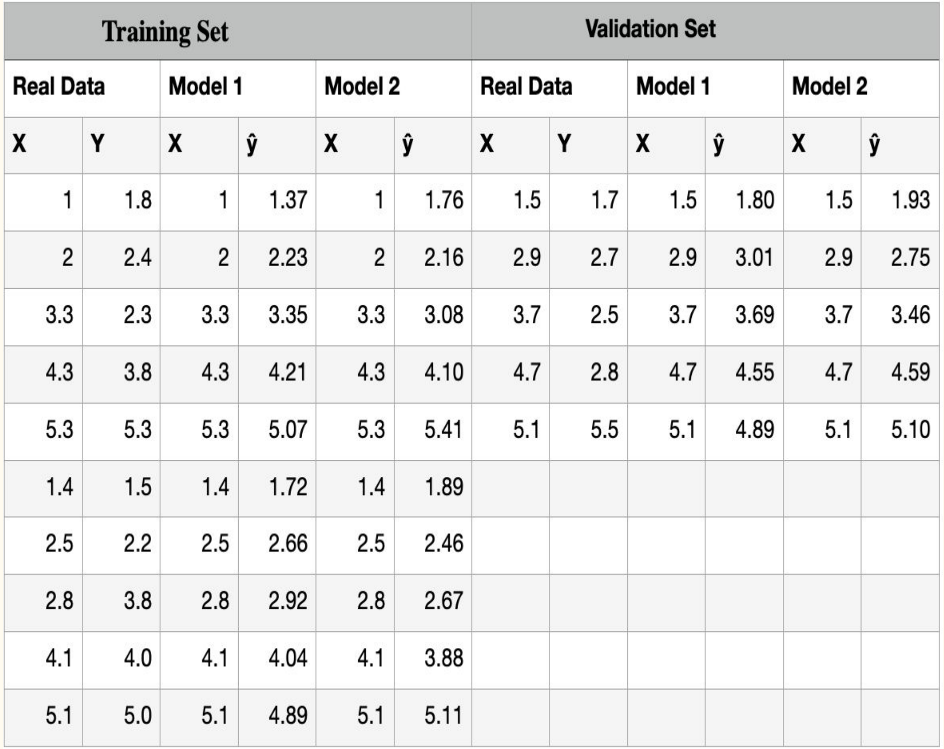
Step 5:
Now, again substitute in the above intercept formula given.
Intercept(a) = $(\Sigma Y - b(\Sigma\underline{X})) / N$
= (32.5 - 0.134(121.34))/10
= (32.5 - 16.26)/10
= 16.24/10
= 1.624

Step 6:
Then substitute these values in regression equation formula
Regression Equation(y) = $\underline{a} + \underline{b}x^2$
= 1.624+ 0.134 x^2

Validation Phase

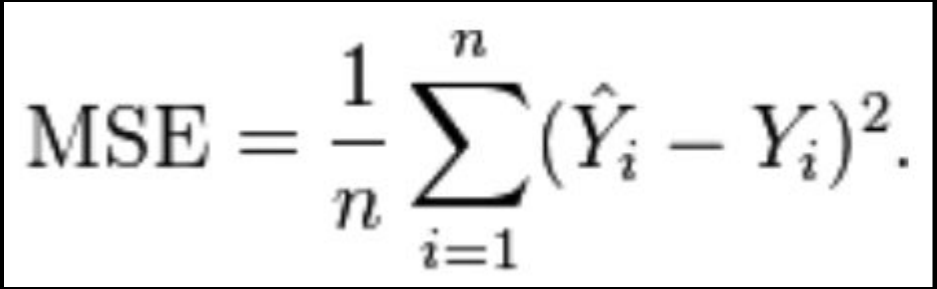
Put the value in these formulas calculated in Training phase to calculate validation phase ŷ :

Linear Regression Model 1 = 0.512 + 0.86x.
Non-Linear Regression Model 2= 1.624+ 0.134 x^2



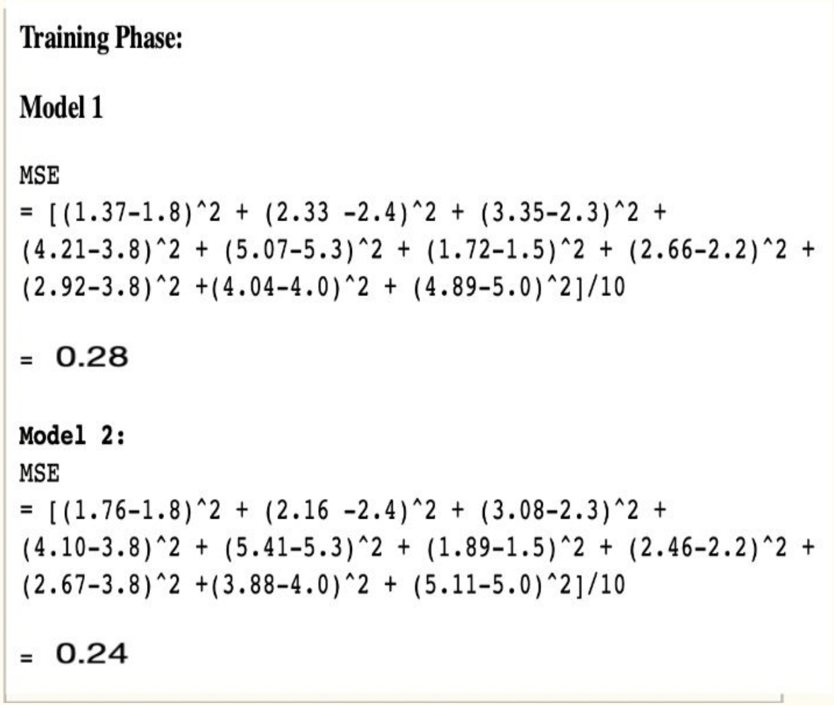
Training Set						Validation Set					
Real Data		Model 1		Model 2		Real Data		Model 1		Model 2	
X	Y	X	ŷ	X	ŷ	X	Y	X	ŷ	X	ŷ
1	1.8	1	1.37	1	1.76	1.5	1.7	1.5	1.80	1.5	1.93
2	2.4	2	2.23	2	2.16	2.9	2.7	2.9	3.01	2.9	2.75
3.3	2.3	3.3	3.35	3.3	3.08	3.7	2.5	3.7	3.69	3.7	3.46
4.3	3.8	4.3	4.21	4.3	4.10	4.7	2.8	4.7	4.55	4.7	4.59
5.3	5.3	5.3	5.07	5.3	5.41	5.1	5.5	5.1	4.89	5.1	5.10
1.4	1.5	1.4	1.72	1.4	1.89						
2.5	2.2	2.5	2.66	2.5	2.46						
2.8	3.8	2.8	2.92	2.8	2.67						
4.1	4.0	4.1	4.04	4.1	3.88						
5.1	5.0	5.1	4.89	5.1	5.11						

Calculate Mean Square



$$\mathrm{MSE} = \frac{1}{n}\sum_{i=1}^{n}(\hat{Y}_i - Y_i)^2.$$

Mean Squared Error (MSE)



Training Phase:

Model 1

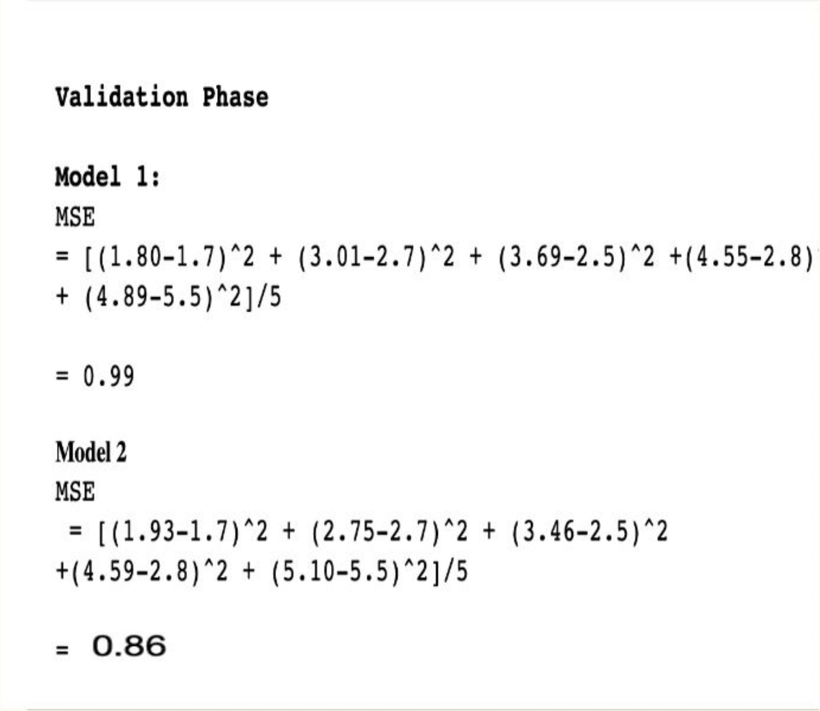
MSE
= [(1.37-1.8)^2 + (2.33 -2.4)^2 + (3.35-2.3)^2 +
(4.21-3.8)^2 + (5.07-5.3)^2 + (1.72-1.5)^2 + (2.66-2.2)^2 +
(2.92-3.8)^2 +(4.04-4.0)^2 + (4.89-5.0)^2]/10

= 0.28

Model 2:

MSE
= [(1.76-1.8)^2 + (2.16 -2.4)^2 + (3.08-2.3)^2 +
(4.10-3.8)^2 + (5.41-5.3)^2 + (1.89-1.5)^2 + (2.46-2.2)^2 +
(2.67-3.8)^2 +(3.88-4.0)^2 + (5.11-5.0)^2]/10

= 0.24



Validation Phase

Model 1:

MSE
= [(1.80-1.7)^2 + (3.01-2.7)^2 + (3.69-2.5)^2 +(4.55-2.8)
+ (4.89-5.5)^2]/5

= 0.99

Model 2

MSE
= [(1.93-1.7)^2 + (2.75-2.7)^2 + (3.46-2.5)^2
+(4.59-2.8)^2 + (5.10-5.5)^2]/5

= 0.86

MSE

max(Training_Set_MSE, Validation_Set_MSE) / min(Training_Set_MSE, Validation_Set_MSE)

Compare Model 1 and Model 2

Mode1

0.99 / 0.28 = 3.53

Model 2

0.86 / 0.24 = 3.58

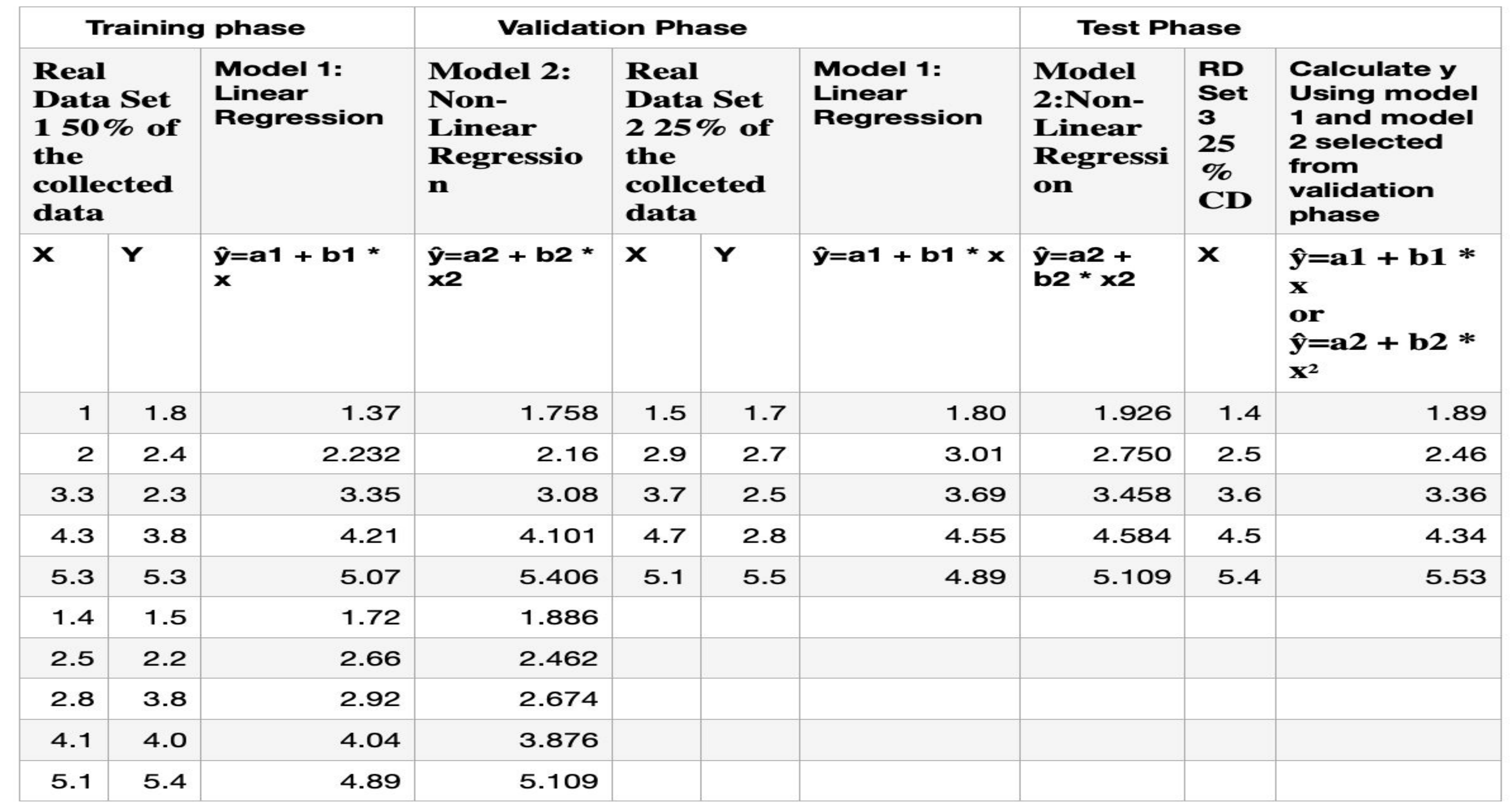
Linear model 1 is better than model 2 (Non -linear)

Put values in the final table to implement the data:

Implementation

Linear Regression Model 1 = 0.512 + 0.86x.

Non-Linear Regression Model 2= 1.624+ 0.134 x^2



Training phase			Validation Phase				Test Phase		
Real Data Set 1 50% of the collected data		Model 1: Linear Regression	Model 2: Non-Linear Regression	Real Data Set 2 25% of the collceted data		Model 1: Linear Regression	Model 2:Non-Linear Regression	RD Set 3 25 % CD	Calculate y Using model 1 and model 2 selected from validation phase
X	Y	ŷ=a1 + b1 * x	ŷ=a2 + b2 * x2	X	Y	ŷ=a1 + b1 * x	ŷ=a2 + b2 * x2	X	ŷ=a1 + b1 * x or ŷ=a2 + b2 * x^2
1	1.8	1.37	1.758	1.5	1.7	1.80	1.926	1.4	1.89
2	2.4	2.232	2.16	2.9	2.7	3.01	2.750	2.5	2.46
3.3	2.3	3.35	3.08	3.7	2.5	3.69	3.458	3.6	3.36
4.3	3.8	4.21	4.101	4.7	2.8	4.55	4.584	4.5	4.34
5.3	5.3	5.07	5.406	5.1	5.5	4.89	5.109	5.4	5.53
1.4	1.5	1.72	1.886						
2.5	2.2	2.66	2.462						
2.8	3.8	2.92	2.674						
4.1	4.0	4.04	3.876						
5.1	5.4	4.89	5.109						

Enhancement Ideas

- To verify that any increase in accuracy over the training data set actually yields an increase in accuracy over a data set that has not been shown to the model before, or at least the model hasn't trained on it (i.e. validation data set).
    - If the accuracy over the training data set increases, but the accuracy over then validation data set stays the same or decreases, then you're overfitting your model and you should stop training.
- A test set (i.e., validation set) is a set of data that is independent of the training data, but that follows the same probability distribution as the training data.
    - If a model fit to the training set also fits the test set well, minimal overfitting has taken place.
    - If the model fits the training set much better than it fits the test set, overfitting is likely the cause.

Conclusion

- In order to avoid overfitting, when any classification parameter needs to be adjusted, it is necessary to have a validation set in addition to the training and test sets.
    - For example if the most suitable classifier for the problem is sought,
        - Training Data is used to train the candidate algorithms.
        - Validation Data is used to compare their performances and decide which one to take.
        - Test Data is used to obtain the performance characteristics such as accuracy, sensitivity, specificity.

References

- https://en.wikipedia.org/wiki/Overfitting
- https://towardsdatascience.com/overfitting-vs-underfitting-a-complete-example-d05dd7e19765
- https://towardsdatascience.com/overfitting-underfitting-and-the-bias-variance-tradeoff-83b42fb11efb
- https://hc.labnet.sfbu.edu/~henry/sfbu/course/data_science/algorithm/slide/non_linear_regression_example.html#nl
- https://hc.labnet.sfbu.edu/~henry/sfbu/course/data_science/algorithm/slide/linear_regression_example.html#lf