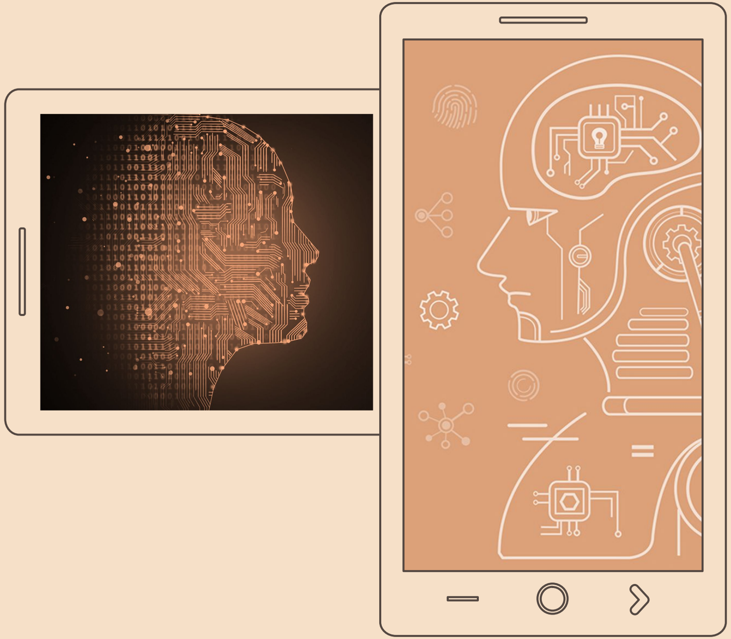
Machine



# Machine Learning

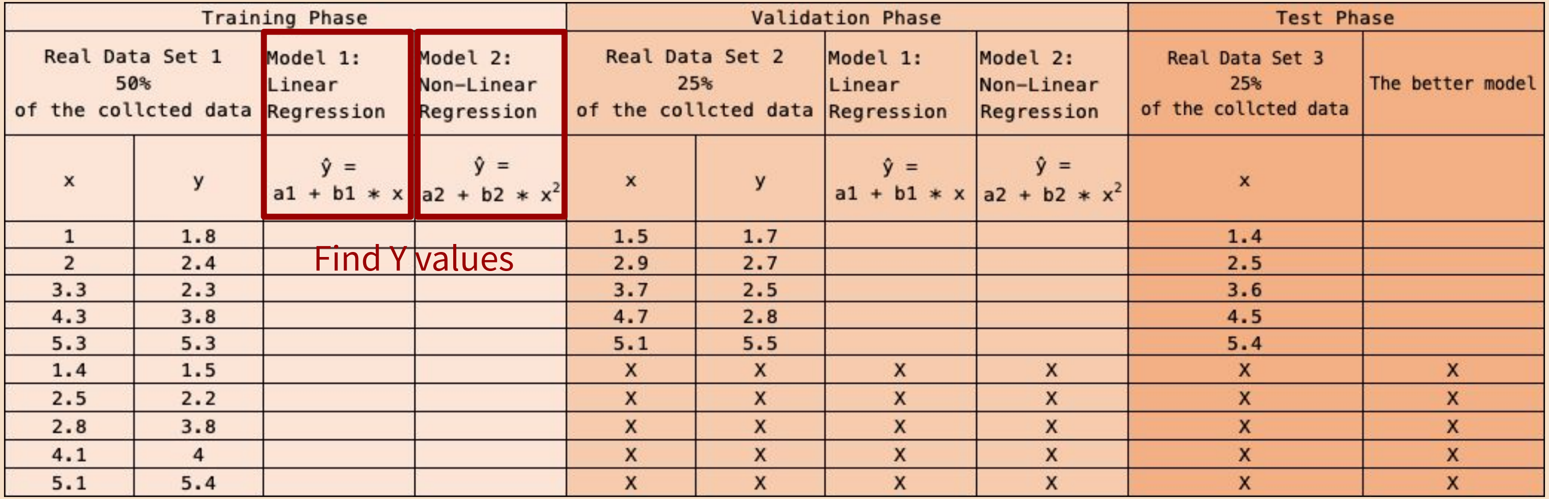
Use Overfitting To Evaluate Different Models

Learning

# Project Explanation:

**Determining which model is the better model.**

Suppose we collect a set of sample data and distribute the sample data by
Training phase: 50%, Validation phase: 25%, Test phase: 25%



| Training Phase | | | | Validation Phase | | | | Test Phase | |
|---|---|---|---|---|---|---|---|---|---|
| Real Data Set 1<br>50%<br>of the collcted data | | Model 1:<br>Linear<br>Regression | Model 2:<br>Non-Linear<br>Regression | Real Data Set 2<br>25%<br>of the collcted data | | Model 1:<br>Linear<br>Regression | Model 2:<br>Non-Linear<br>Regression | Real Data Set 3<br>25%<br>of the collcted data | The better model |
| x | y | $\hat{y} = a1 + b1 * x$ | $\hat{y} = a2 + b2 * x^2$ | x | y | $\hat{y} = a1 + b1 * x$ | $\hat{y} = a2 + b2 * x^2$ | x | |
| 1 | 1.8 | | | 1.5 | 1.7 | | | 1.4 | |
| 2 | 2.4 | | | 2.9 | 2.7 | | | 2.5 | |
| 3.3 | 2.3 | | | 3.7 | 2.5 | | | 3.6 | |
| 4.3 | 3.8 | | | 4.7 | 2.8 | | | 4.5 | |
| 5.3 | 5.3 | | | 5.1 | 5.5 | | | 5.4 | |
| 1.4 | 1.5 | | | X | X | X | X | X | X |
| 2.5 | 2.2 | | | X | X | X | X | X | X |
| 2.8 | 3.8 | | | X | X | X | X | X | X |
| 4.1 | 4 | | | X | X | X | X | X | X |
| 5.1 | 5.4 | | | X | X | X | X | X | X |

Find Y values

# Model 1: Linear Regression

Find Y values

## 01

**N = 10**

Count the number of values.

## 02

**Find**
**X * Y and X * X.**

Real Data X * Y and X * X.

## 03

**Find**
**$\Sigma X$, $\Sigma Y$, $\Sigma XY$, $\Sigma P$.**

$\Sigma P = \Sigma X*X = 121.34$,
$\Sigma X = 31.8$, $\Sigma Y = 32.5$,
$\Sigma XY = 120.8$.

## 04

**Slope(b) = 0.863177681**

$(N\Sigma XY - (\Sigma X)(\Sigma Y)) / (N\Sigma X\hat{}2 - (\Sigma X)\hat{}2)$

## 05

**Intercept(a) = 0.505094974**

$(\Sigma Y - b(\Sigma X)) / N$

## 06

**Regression Equation(y)**
**= a + bx**

Use real data X to calculate.

# Model 2: Non-Linear Regression

Find Y values

## 01

**N = 10**

Count the number of values.

## 02

**Find X * Y and X * X.**

Real Data X * Y and X * X.

## 03

**Find $\Sigma X, \Sigma Y, \Sigma XY, \Sigma P, \Sigma PY, \Sigma P^2$**

$\Sigma P = \Sigma X*X = 121.34$, $\Sigma X = 31.8$,
$\Sigma Y = 32.5$, $\Sigma XY = 120.8$,
$\Sigma PY = 509.762$,
$\Sigma PP = 2329.9862$

## 04

**Slope(b) = 0.134562411**

$(N\Sigma PY - (\Sigma P)(\Sigma Y)) / (N\Sigma P^2 - (\Sigma P)^2)$

## 05

**Intercept(a) = 1.6172197**

$(\Sigma Y - b(\Sigma P)) / N$

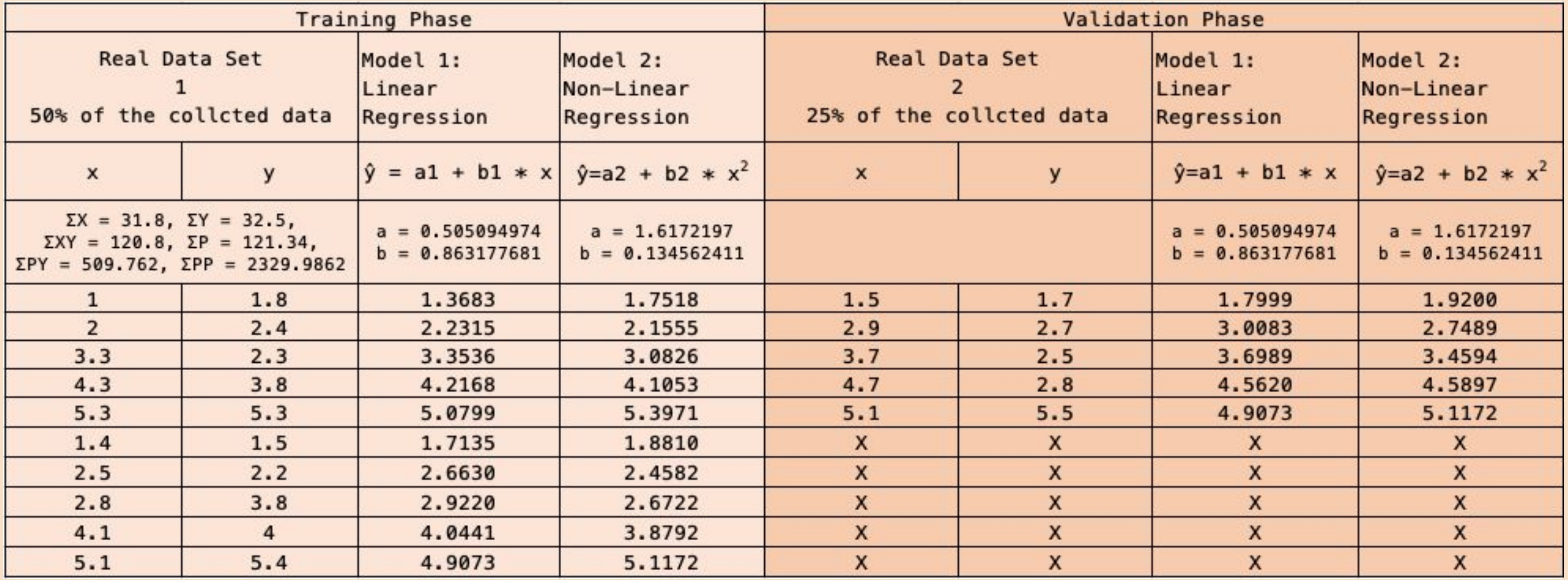
## 06

**Regression Equation(y) = $a + bx^2$**

Use real data X to calculate.

# Data Table:

**After used both linear and non-linear regression to calculate the Y value. Next step is to determining which model is the better model.**



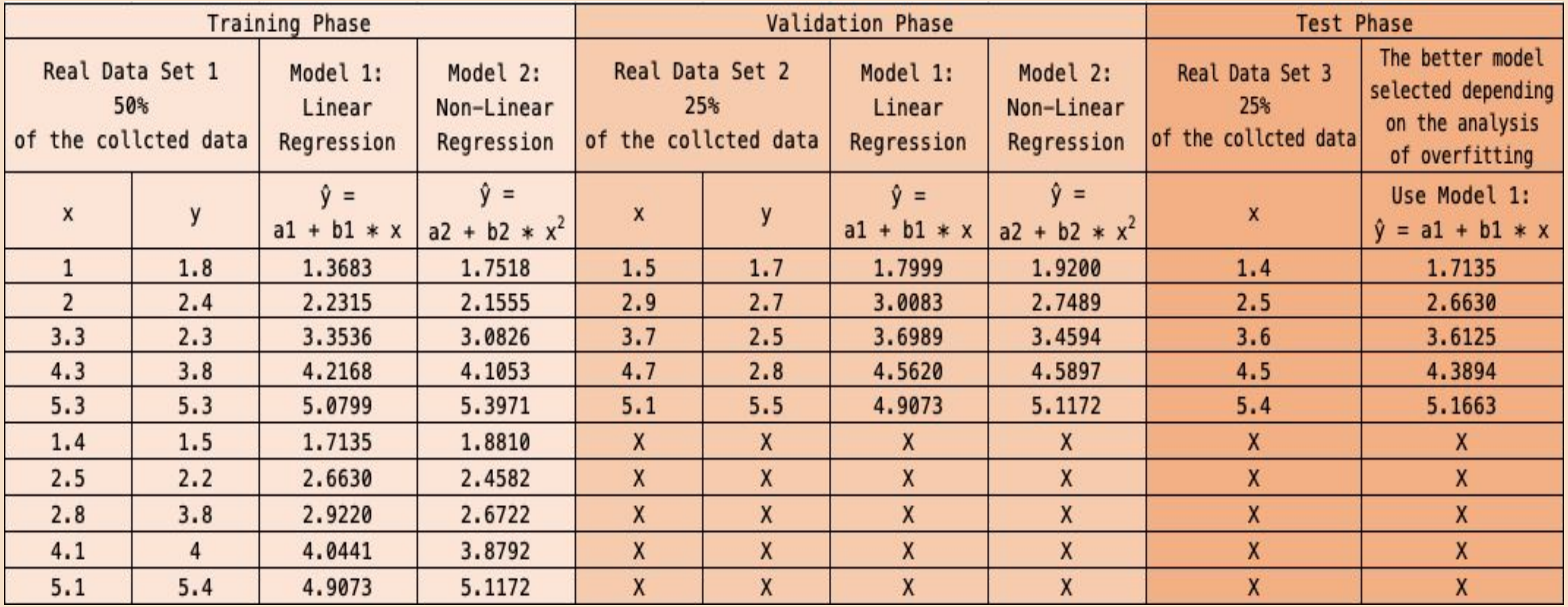
| Training Phase | | | | Validation Phase | | | |
|---|---|---|---|---|---|---|---|
| Real Data Set<br>1<br>50% of the collcted data | | Model 1:<br>Linear<br>Regression | Model 2:<br>Non-Linear<br>Regression | Real Data Set<br>2<br>25% of the collcted data | | Model 1:<br>Linear<br>Regression | Model 2:<br>Non-Linear<br>Regression |
| x | y | $\hat{y} = a1 + b1 * x$ | $\hat{y}=a2 + b2 * x^2$ | x | y | $\hat{y}=a1 + b1 * x$ | $\hat{y}=a2 + b2 * x^2$ |
| ΣX = 31.8, ΣY = 32.5,<br>ΣXY = 120.8, ΣP = 121.34,<br>ΣPY = 509.762, ΣPP = 2329.9862 | | a = 0.505094974<br>b = 0.863177681 | a = 1.6172197<br>b = 0.134562411 | | | a = 0.505094974<br>b = 0.863177681 | a = 1.6172197<br>b = 0.134562411 |
| 1 | 1.8 | 1.3683 | 1.7518 | 1.5 | 1.7 | 1.7999 | 1.9200 |
| 2 | 2.4 | 2.2315 | 2.1555 | 2.9 | 2.7 | 3.0083 | 2.7489 |
| 3.3 | 2.3 | 3.3536 | 3.0826 | 3.7 | 2.5 | 3.6989 | 3.4594 |
| 4.3 | 3.8 | 4.2168 | 4.1053 | 4.7 | 2.8 | 4.5620 | 4.5897 |
| 5.3 | 5.3 | 5.0799 | 5.3971 | 5.1 | 5.5 | 4.9073 | 5.1172 |
| 1.4 | 1.5 | 1.7135 | 1.8810 | X | X | X | X |
| 2.5 | 2.2 | 2.6630 | 2.4582 | X | X | X | X |
| 2.8 | 3.8 | 2.9220 | 2.6722 | X | X | X | X |
| 4.1 | 4 | 4.0441 | 3.8792 | X | X | X | X |
| 5.1 | 5.4 | 4.9073 | 5.1172 | X | X | X | X |

# Use Overfitting To Evaluate Different Models

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

| | Training Set | Validation Set | |
|---|---|---|---|
| **Model 1** | ((1.3686-1.8)^2 + (2.2315-2.4)^2 + (3.3536-2.3)^2 + (4.2168-3.8)^2 + (5.0799-5.3)^2 + (1.7135-1.5)^2 + (2.6630-2.2)^2 + (2.9220-3.8)^2 + (4.0441-4)^2 + (4.9073-5.4)^2)) = 2.82227077 | ((1.7999-1.7)^2 + (3.0083-2.7)^2 + (3.6989-2.5)^2 + (4.5620-2.8)^2 + (4.9073-5.5)^2 ) = 4.9983274 | **Better**<br>4.9983274<br>---------------------<br>2.82227077<br>= 1.771030425 |
| **Model 2** | ((1.7518-1.8)^2 + (2.1555-2.4)^2 + (3.0826-2.3)^2 + (4.1053-3.8)^2 + (5.3971-5.3)^2 + (1.8810-1.5)^2 + (2.4582-2.2)^2 + (2.6722-3.8)^2 + (3.8792-4)^2 + (5.1172-5.4)^2)) = 2.35553231 | ((1.9200-1.7)^2 + (2.7489-2.7)^2 + (3.4594-2.5)^2 + (4.5897-2.8)^2 + (5.1172-5.5)^2 ) = 4.3208015 | 4.3208015<br>---------------------<br>2.35553231<br>= 1.834320625 |

Machine



| Training Phase | | | | Validation Phase | | | | Test Phase | |
|---|---|---|---|---|---|---|---|---|---|
| Real Data Set 1<br>50%<br>of the collcted data | | Model 1:<br>Linear<br>Regression | Model 2:<br>Non-Linear<br>Regression | Real Data Set 2<br>25%<br>of the collcted data | | Model 1:<br>Linear<br>Regression | Model 2:<br>Non-Linear<br>Regression | Real Data Set 3<br>25%<br>of the collcted data | The better model selected depending on the analysis of overfitting |
| x | y | $\hat{y} =$<br>$a1 + b1 * x$ | $\hat{y} =$<br>$a2 + b2 * x^2$ | x | y | $\hat{y} =$<br>$a1 + b1 * x$ | $\hat{y} =$<br>$a2 + b2 * x^2$ | x | Use Model 1:<br>$\hat{y} = a1 + b1 * x$ |
| 1 | 1.8 | 1.3683 | 1.7518 | 1.5 | 1.7 | 1.7999 | 1.9200 | 1.4 | 1.7135 |
| 2 | 2.4 | 2.2315 | 2.1555 | 2.9 | 2.7 | 3.0083 | 2.7489 | 2.5 | 2.6630 |
| 3.3 | 2.3 | 3.3536 | 3.0826 | 3.7 | 2.5 | 3.6989 | 3.4594 | 3.6 | 3.6125 |
| 4.3 | 3.8 | 4.2168 | 4.1053 | 4.7 | 2.8 | 4.5620 | 4.5897 | 4.5 | 4.3894 |
| 5.3 | 5.3 | 5.0799 | 5.3971 | 5.1 | 5.5 | 4.9073 | 5.1172 | 5.4 | 5.1663 |
| 1.4 | 1.5 | 1.7135 | 1.8810 | X | X | X | X | X | X |
| 2.5 | 2.2 | 2.6630 | 2.4582 | X | X | X | X | X | X |
| 2.8 | 3.8 | 2.9220 | 2.6722 | X | X | X | X | X | X |
| 4.1 | 4 | 4.0441 | 3.8792 | X | X | X | X | X | X |
| 5.1 | 5.4 | 4.9073 | 5.1172 | X | X | X | X | X | X |

Learning

Machine

# THANKS!

Learning