Mohamed Imran

-Data Scientist

Ganit Inc.

Data Preprocessing

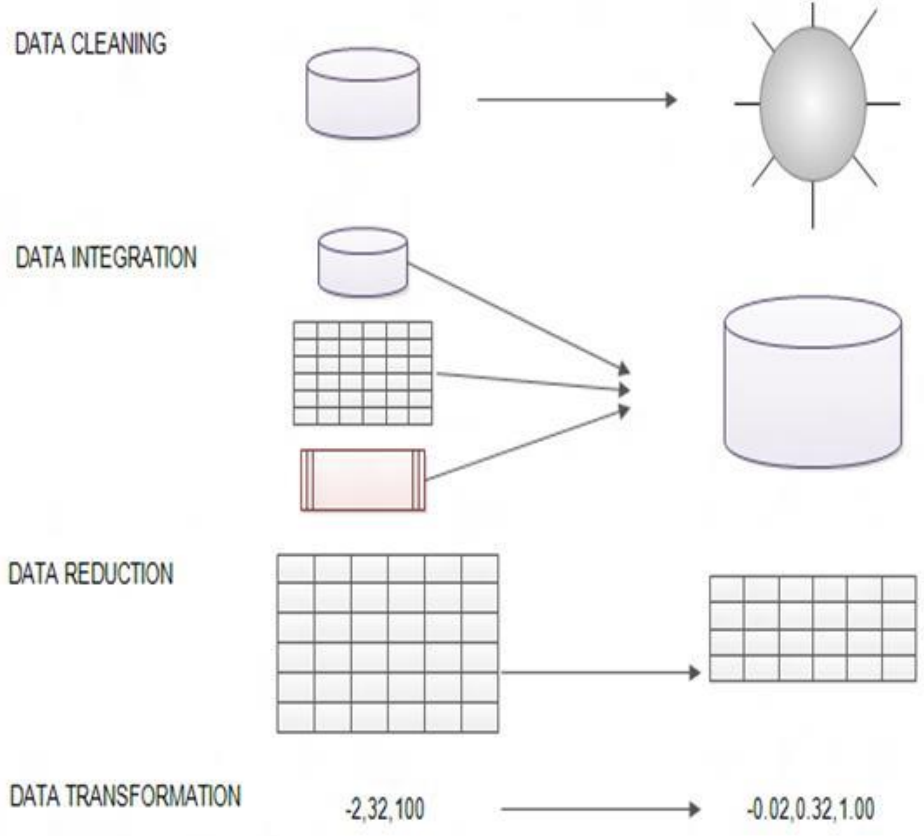
Real World Data

Any Problem?

S.No	Credit_rating	Age	Income	Credit_cards
1	0.00	21	10000	y
2	1.0		2500	n
3	2.0	62	-500	y
4	100.012	42		n
5	yes	200	1	y
6	30	0	Seventy thousand	No

Data Preprocessing

- Data Cleaning
- Data Integration
- Data Reduction
- Data Transformation



Data Cleaning

1. Missing Data
- Central Imputation
- KNN Imputation
- 2. Noisy Data
- Smoothing
- Clustering

1. Outlier Removal
- Using Boxplot

company name	furigana	postal code	address	telephone number
AlphaPurchase Co,. Ltd	Alpha Purchase	107-0061	Aoyama Building 12th floor, 1-2-3, Kita-Aoyama, Minato-ku, Tokyo	03-5772-7801
AAA Foundation	AAA	1500002	Kami-meguro, Meguro-ku X-X-X	0312345678
BBBB, Inc.	BBBB	123	Minami-Azabu, Minato-ku XX-1-1	03(1234)9876

company name	juridical personality	furigana	postal code	all prefectures	address	telephone number
Alpha Purchase	Co,. Ltd	Alpha Purchase	1070061	Tokyo	Aoyama Building 12th floor, 1-2-3, Kita-Aoyama, Minato-ku	0357727801
AAA	Foundation	AAA	1500002	Tokyo	Kami-meguro, Meguro-ku X-X-X	0312345678
BBBB	Inc.	BBBB	1230001	Tokyo	Minami-Azabu, Minato-ku XX-1-1	0312349876

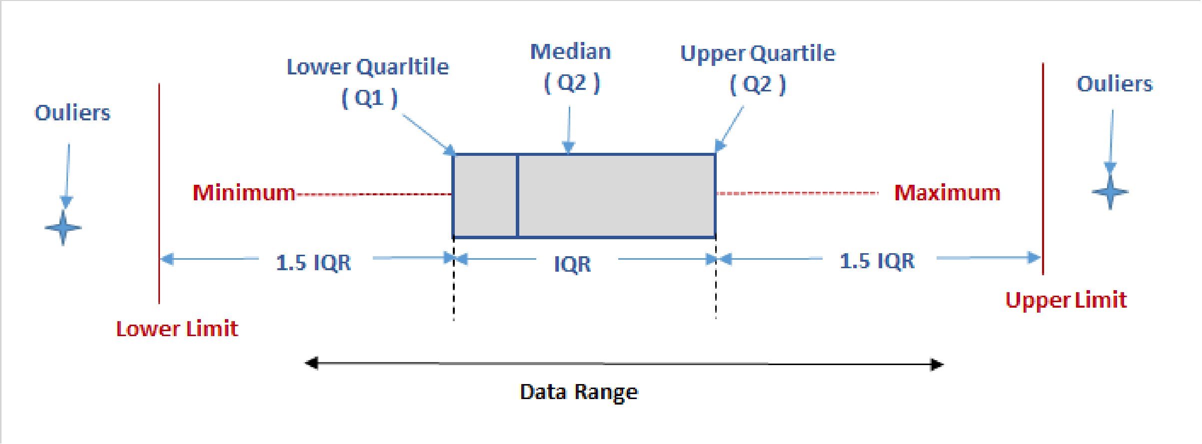
Imputation

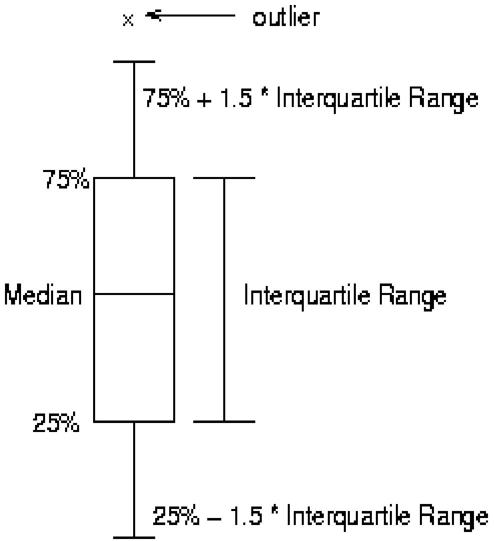
- Replace with mean or a median
- When to use mean?
- Replace with nearest neighbour
- How much nearest to see?

S.No	Qualification	Age	Income
1	B.Tech	25	30k
2	M.Tech	30	50k
3	B.Tech	26	32k
4	B.Tech	25	?
5	M.Tech	29	60k
6	B.Tech	?	30k

Outlier

- BoxPlot





Data Transformation

- Normalization

Min-max normalization

1. Min Max Normalization
2. Z - Score Normalization
3. Decimal scaling

Normalization: Example II

- **Min-Max normalization on an employee database**
  - max distance for salary: 100000-19000 = 81000
  - max distance for age: 52-27 = 25
  - New min for age and salary = 0; new max for age and salary = 1

$$x'_i = \frac{x_i - \min x_i}{\max x_i - \min x_i}(new\max - new\min) + new\min$$

ID	Gender	Age	Salary
1	F	27	19,000
2	M	51	64,000
3	M	52	100,000
4	F	33	55,000
5	M	45	45,000

ID	Gender	Age	Salary
1	1	0.00	0.00
2	0	0.96	0.56
3	0	1.00	1.00
4	1	0.24	0.44
5	0	0.72	0.32

Normalization: Example

- **z-score normalization: v' = (v - Mean) / Stdev**
- **Example: normalizing the "Humidity" attribute:**

Humidity
85
90
78
96
80
70
65
95
70
80
70
90
75
80

Mean = 80.3
Stdev = 9.84

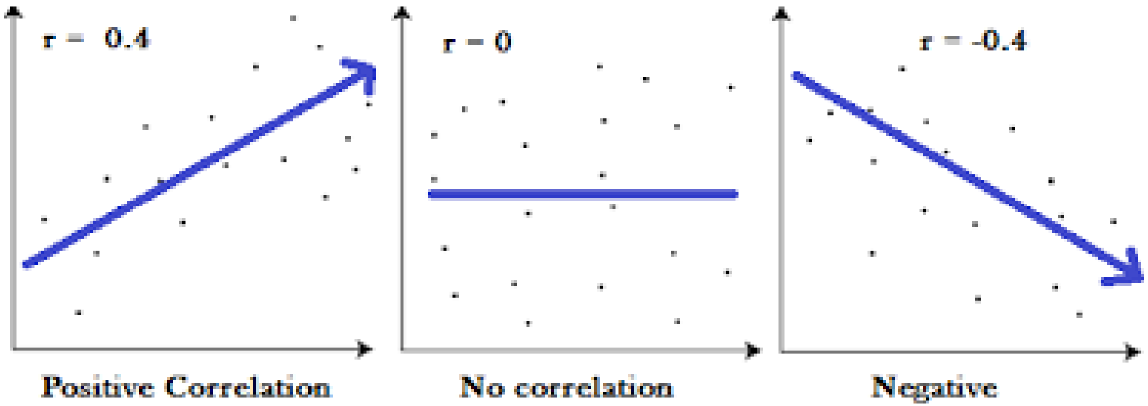
Humidity
0.48
0.99
-0.23
1.60
-0.03
-1.05
-1.55
1.49
-1.05
-0.03
-1.05
0.99
-0.54
-0.03

Decimal scaling

v= v/10^j

Data Integration

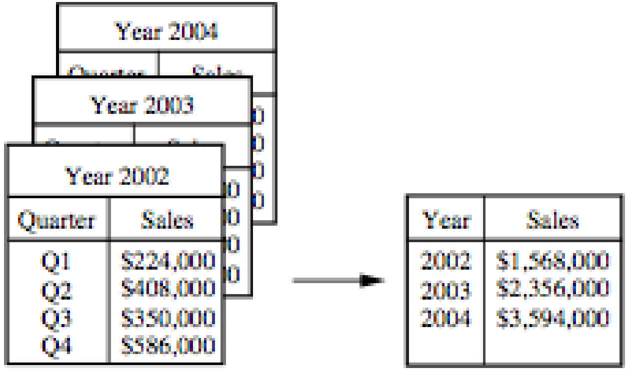
- Check for correlation
- Remove uncorrelated data



$$r = \frac{\sum (x-\bar{x})(y-\bar{y})}{\sqrt{\sum (x-\bar{x})^2 \sum (y-\bar{y})^2}}$$

Data Reduction

- Data Cube Aggregation



Year 2004

Year 2003

Year 2002

Quarter	Sales
Q1	$224,000
Q2	$408,000
Q3	$350,000
Q4	$586,000

Year	Sales
2002	$1,568,000
2003	$2,356,000
2004	$3,594,000

Figure 2.13 Sales data for a given branch of *AllElectronics* for the years 2002 to 2004. On the left, the sales are shown per quarter. On the right, the data are aggregated to provide the annual sales.

Relationship

Y = ??????????

X	Y
2	8
6	20
4	14
3	11
7	23
4	14
2	8
5	17

Relationship

$$Y = 2 + 3(X)$$

X	Y
2	8
6	20
4	14
3	11
7	23
4	14
2	8
5	17

What is 2 here?

Y = 2 + 3(X)

X	Y
2	8
6	20
4	14
3	11
7	23
4	14
2	8
5	17

Find the Y in ?

$$Y = 2 + 3(X)$$

X	Y
2	8
6	20
4	14
3	11
7	23
4	14
2	8
5	17
10	?
1	?

Value for Y with given X

$$Y = 2 + 3(X)$$

X	Y
2	8
6	20
4	14
3	11
7	23
4	14
2	8
5	17
10	32
1	5

Terminology

$$Y = 2 + 3(X)$$

Y = Model

X	Y
2	8
6	20
4	14
3	11
7	23
4	14
2	8
5	17
10	32
1	5

Terminology

$$Y = 2 + 3(X)$$

Y = Model

2 = Intercept

X	Y
2	8
6	20
4	14
3	11
7	23
4	14
2	8
5	17
10	32
1	5

Terminology

$$Y = 2 + 3(X)$$

Y = Model

2 = Intercept

3 = Slope

X	Y
2	8
6	20
4	14
3	11
7	23
4	14
2	8
5	17
10	32
1	5

Terminology

$Y = 2 + 3(X)$

Y = Model

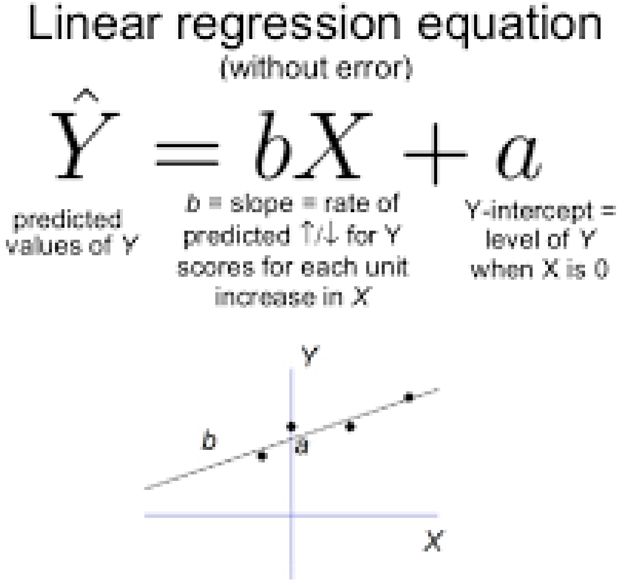
2 = Intercept

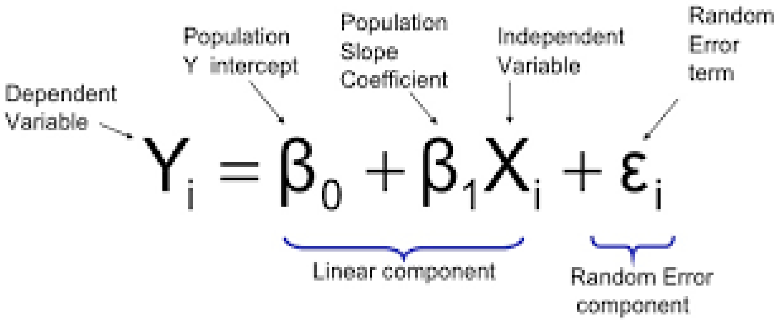
3 = Slope

X = input

X	Y
2	8
6	20
4	14
3	11
7	23
4	14
2	8
5	17
10	32
1	5

Formula for a line





Linear Regression

Welcome to the world of data science

What is linear?

What is linear?

A Straight line

What is Regression?

What is Regression?

Relationship between two points

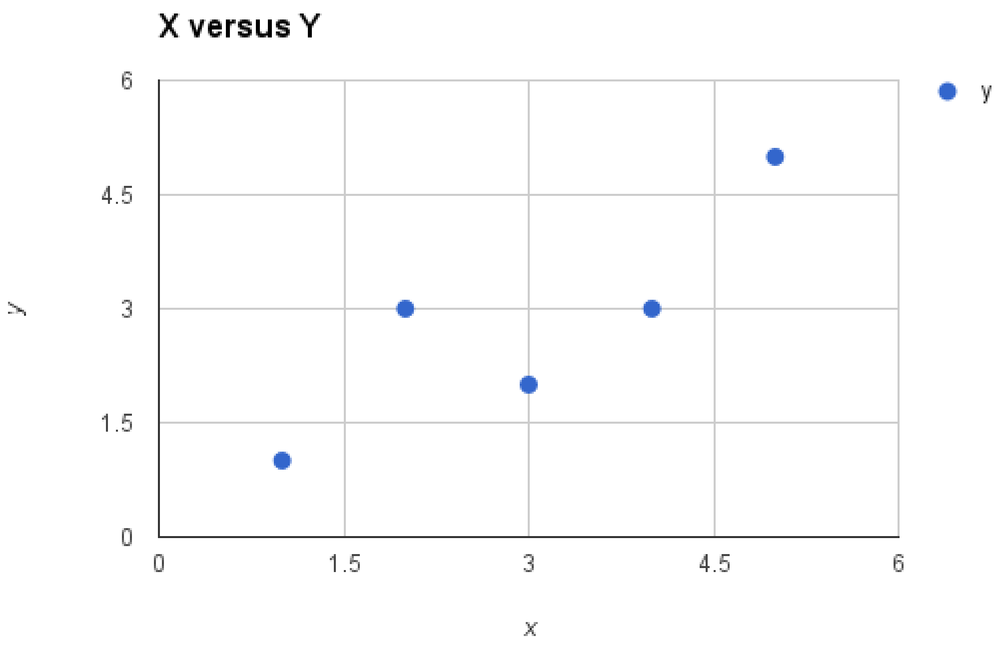
What is Linear Regression?

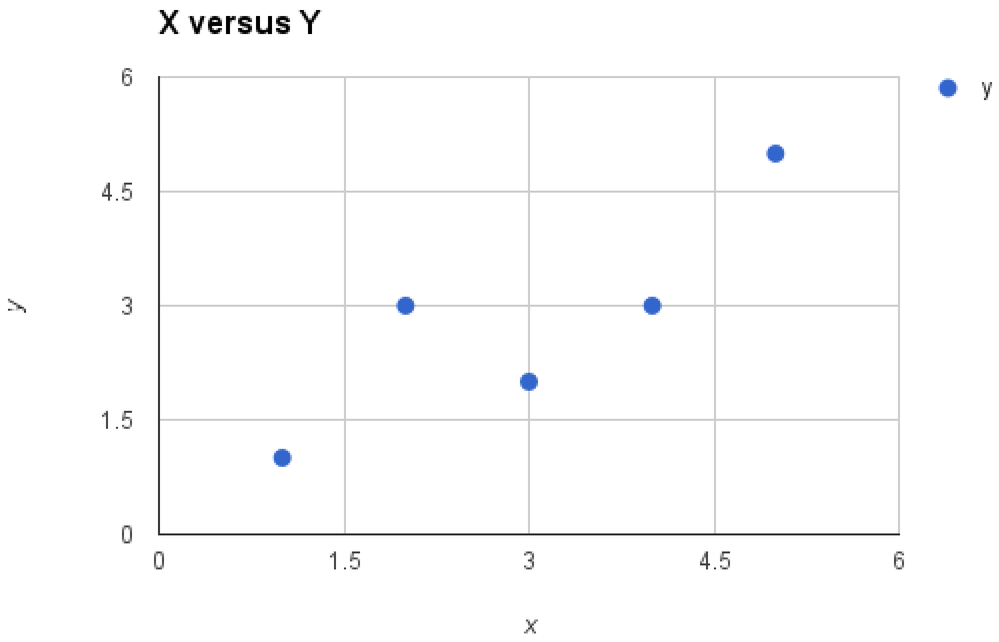
What is Linear Regression?

A Straight line that attempts to predict
the relationship between two points

Help me in finding the relationship?

x	y
1	1
2	3
4	3
3	2
5	5


X versus Y
y
6
4.5
3
1.5
0
0
1.5
3
4.5
6
x
y



$$y = B0 + B1 * x$$

$$B1 = sum((xi-mean(x)) * (yi-mean(y))) / sum((xi – mean(x))^2)$$

$$B0 = mean(y) – B1 * mean(x)$$

x	mean(x)	x - mean(x)
1	3	-2
2	3	-1
4	3	1
3	3	0
5	3	2

y	mean(y)	y - mean(y)
1	2.8	-1.8
3	2.8	0.2
3	2.8	0.2
2	2.8	-0.8
5	2.8	2.2

B1 = 8 / 10

B1 = 0.8

B0 = mean(y) – B1 * mean(x)

or

B0 = 2.8 – 0.8 * 3

or

B0 = 0.4

y = B0 + B1 * x

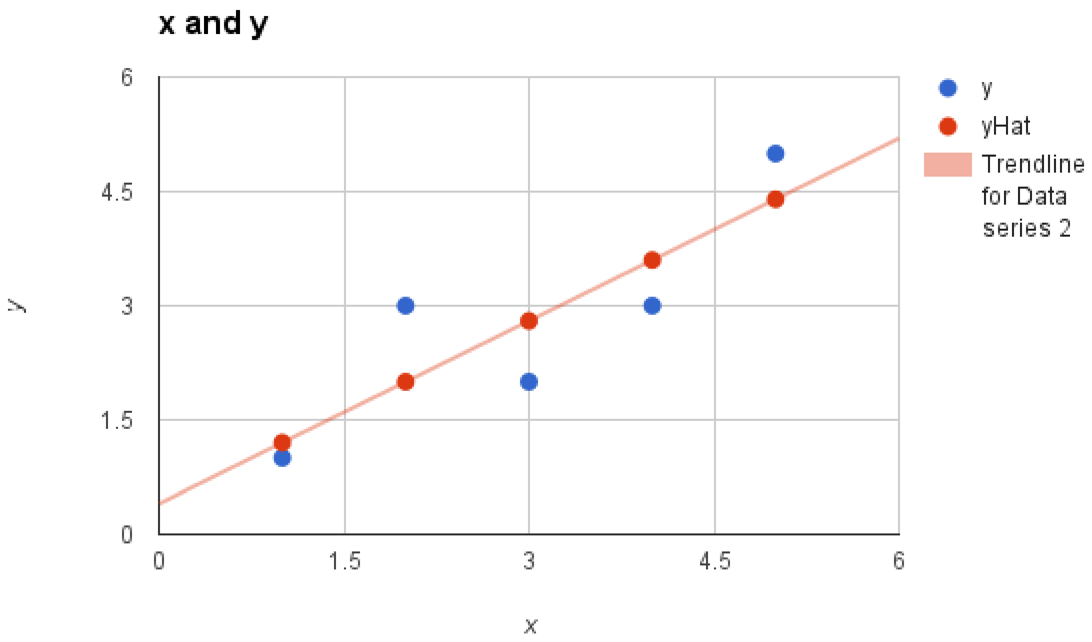
or

y = 0.4 + 0.8 * x

x - mean(x)	y - mean(y)	Multiplication
-2	-1.8	3.6
-1	0.2	-0.2
1	0.2	0.2
0	-0.8	0
2	2.2	4.4
		8

x - mean(x)	squared
-2	4
-1	1
1	1
0	0
2	4
	10

x	y	predicted y
1	1	1.2
2	3	2
4	3	3.6
3	2	2.8
5	5	4.4



RMSE = 0.692

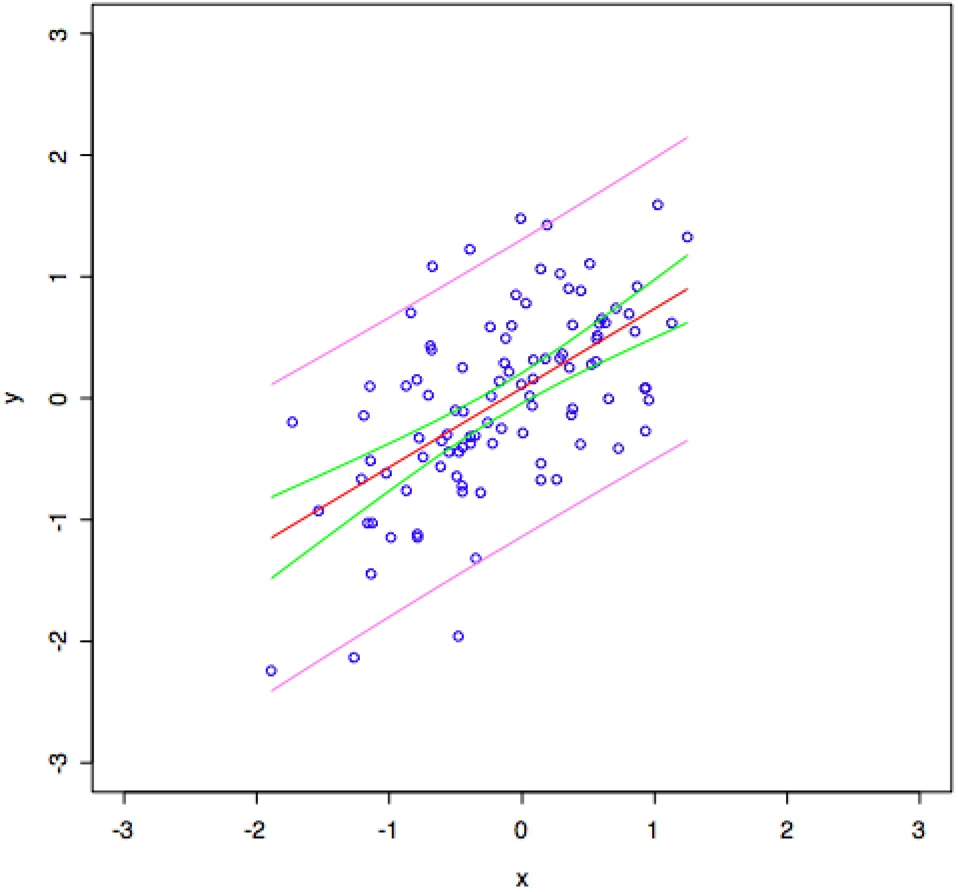
Gradient Descent

Finding the optimum relationship where the error is minimal.

Finding the intercept and coefficients value.

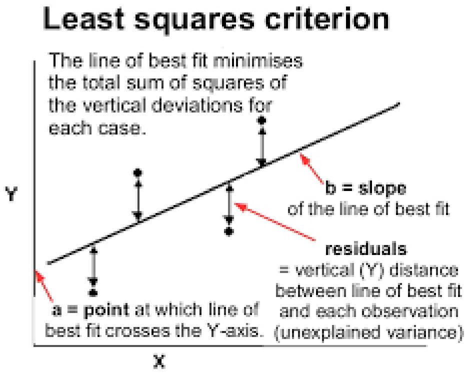
Find the solution?

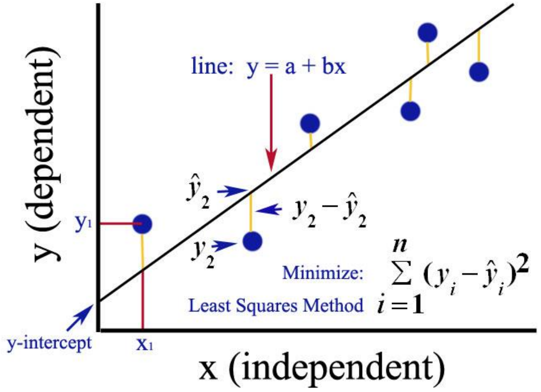
Any Suggestions?

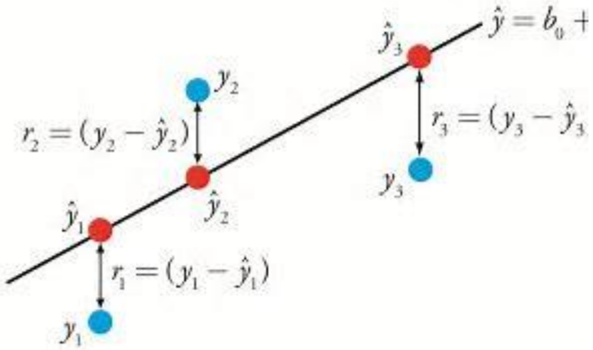


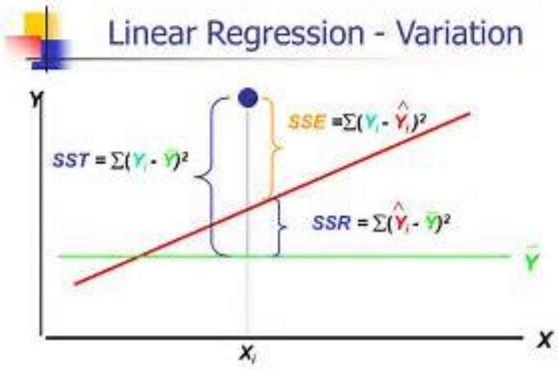
Line of best fit

Ordinary least square line









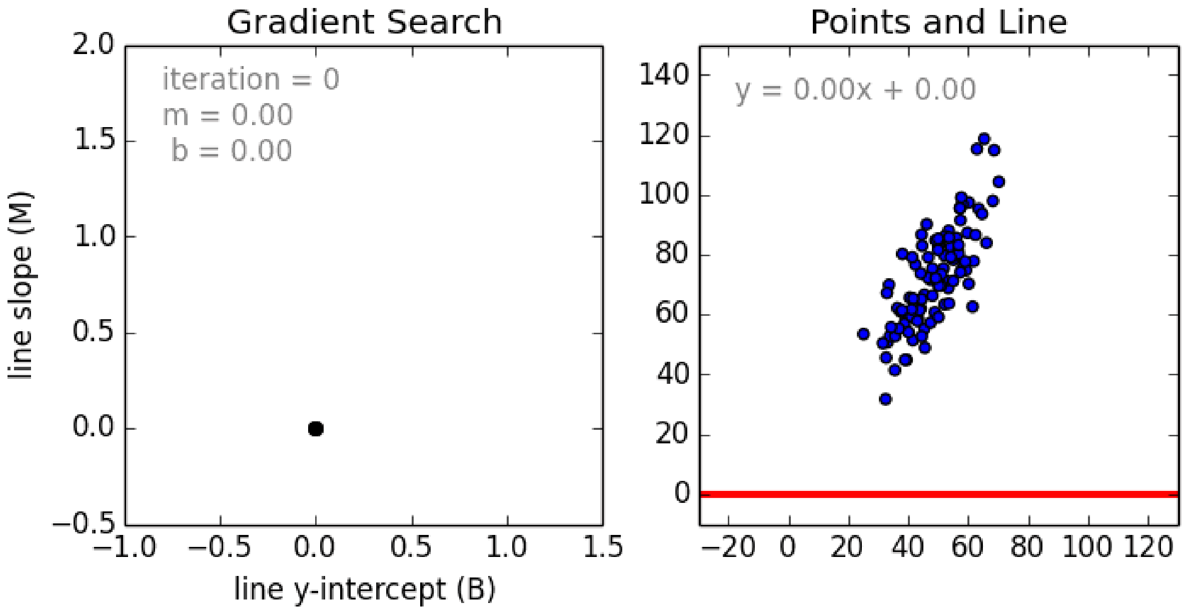
Cost Function

$$\text{Error}_{(\text{m,b})} = \frac{1}{N}\sum_{i=1}^{N}(y_i - (mx_i + b))^2$$

Gradient Descent

Learning Rate

Momentum



Partial Derivative

Finding the direction of coefficient and slope moves in.

$$\frac{\partial}{\partial \mathrm{m}} = \frac{2}{N}\sum_{i=1}^{N} - x_i(y_i - (mx_i + b))$$

$$\frac{\partial}{\partial \mathrm{b}} = \frac{2}{N}\sum_{i=1}^{N} -(y_i - (mx_i + b))$$

Error Metrics for Regression

$$\mathrm{MAE} = \frac{1}{n}\sum_{j=1}^{n}|y_j - \hat{y}_j|$$

$$\mathrm{MSE} = \frac{1}{n}\sum_{i=1}^{n}(y_i - \tilde{y}_i)^2$$

$$r^2 = 1 - \frac{\text{SS Error}}{\text{SS Total}} = 1 - \frac{\sum(y_i - \hat{y}_i)^2}{\sum(y_i - \bar{y})^2}$$

$$\mathrm{RMSE} = \sqrt{\frac{1}{n}\sum_{j=1}^{n}(y_j - \hat{y}_j)^2}$$

Iteration	Error
1	9.556915033600001
2	9.514033718864932
3	9.471355093177891
4	9.42887819847207
5	9.302648387373978
10	9.302648387373978
20	9.260968926175824
30	8.775918820666949
40	8.392252947074406
50	8.02634104901006
60	7.677361561773854
100	6.160260505649477
200	4.018554474422596
300	2.685046327855845
400	1.854748522005687
800	0.6906129091698867
1000	0.5644839798882763
1600	0.4891352315933852

Step 1

`import` statement:

```
from sklearn import linear_model
```

Step 2

I have the height and weight data of some people. Let's use this data to do linear regression and try to predict the weight of other people.

```
height=[[4.0],[4.5],[5.0],[5.2],[5.4],[5.8],[6.1],[6.2],[6.4],[6.8]]
weight=[  42 ,   44 , 49, 55   , 53   , 58    , 60   , 64   ,   66 ,   69]

print("height weight")
for row in zip(height, weight):
    print(row[0][0],"->",row[1])
```

Step 3

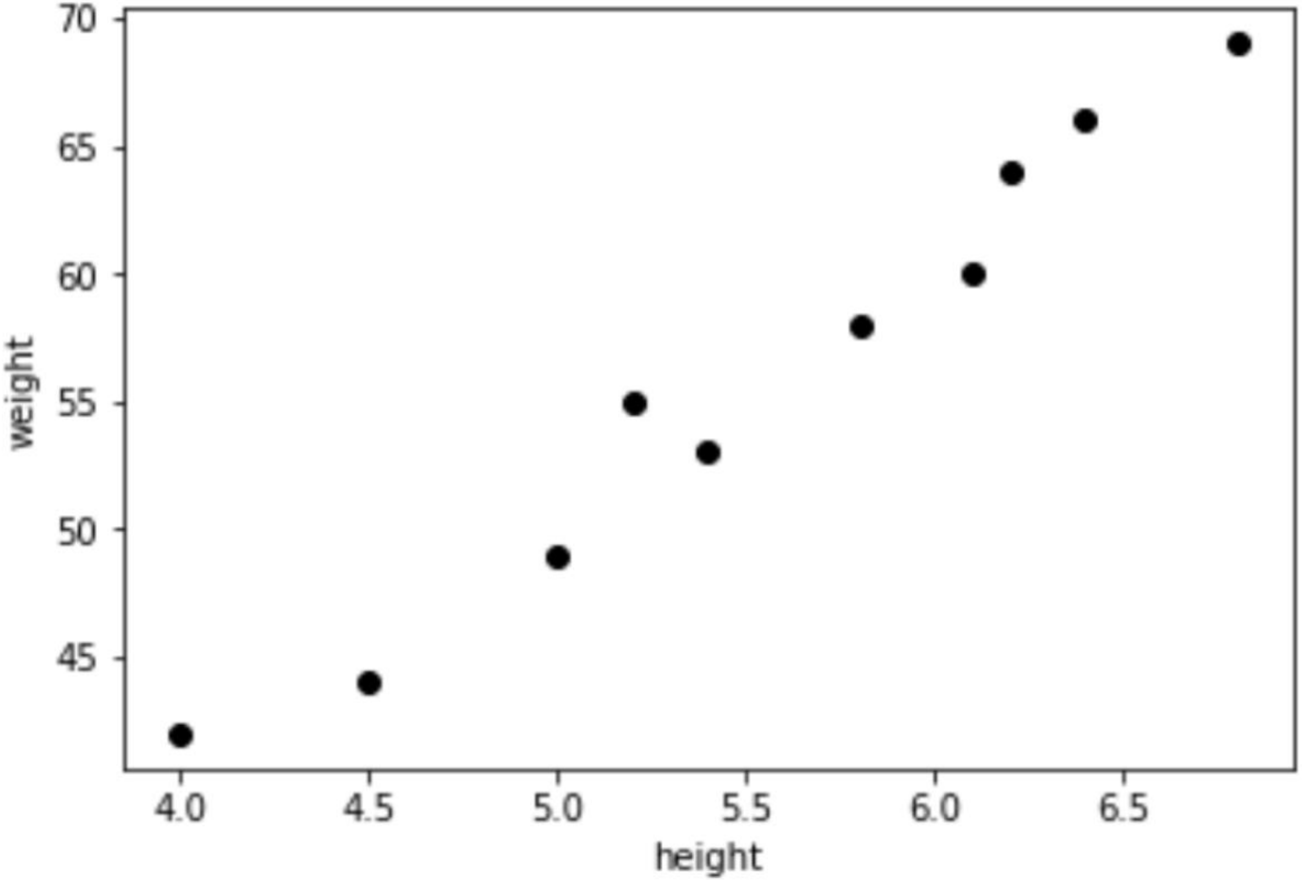
`import` statement to plot graph using `matplotlib`:

```
import matplotlib.pyplot as plt
```

Plotting the height and weight data:

```
plt.scatter(height,weight,color='black')
plt.xlabel("height")
plt.ylabel("weight")
```

Output:



Step 4

Declaring the linear regression function and call the `fit` method to learn from data:

```
reg=linear_model.LinearRegression()
reg.fit(height,weight)
```

Slope and intercept:

```
m=reg.coef_[0]
b=reg.intercept_
print("slope=",m, "intercept=",b)
```

Output:

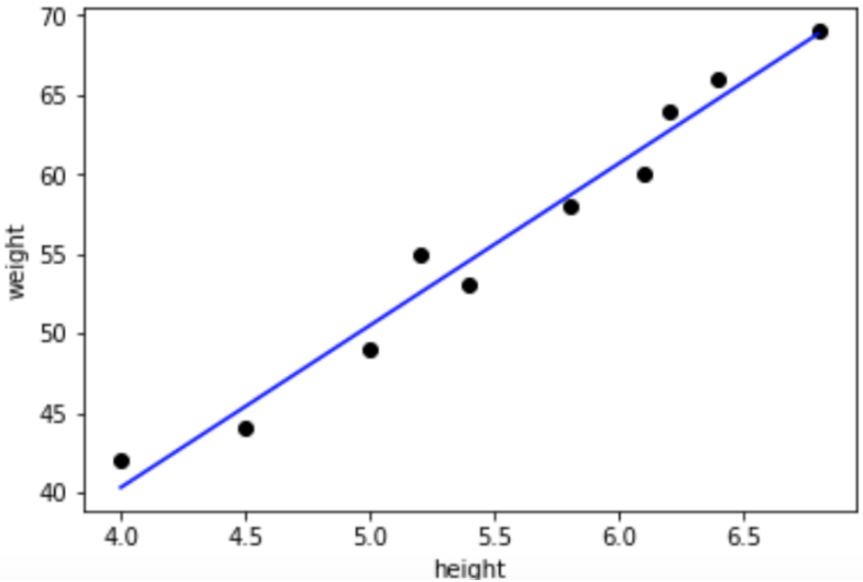
```
slope= 10.1936218679 intercept= -0.4726651480
```

Step 5

Using the values of slope and intercept to construct the line to fit our data points:

```
plt.scatter(height,weight,color='black')
predicted_values = [reg.coef_ * i + reg.intercept_ for i in height]
plt.plot(height, predicted_values, 'b')
plt.xlabel("height")
plt.ylabel("weight")
```

Output:



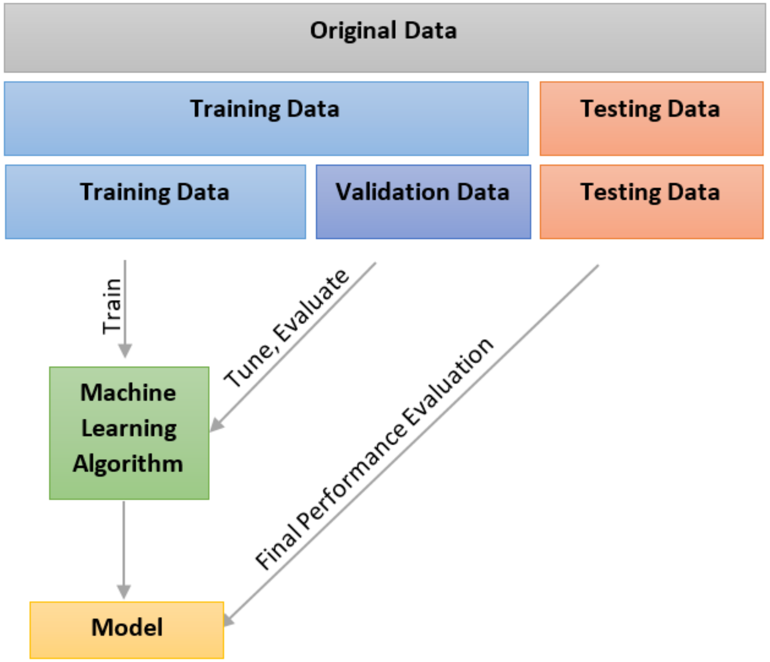
Advantage of Linear Regression

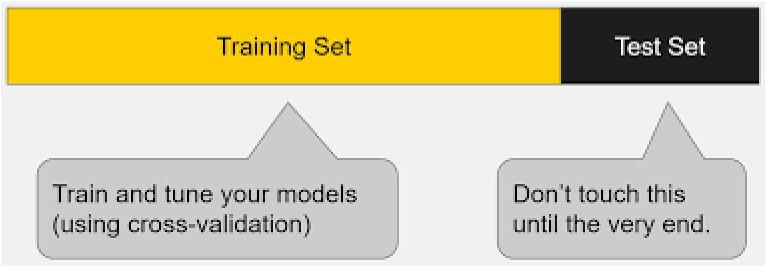
- Linear regression implements a statistical model that, when relationships between the independent variables and the dependent variable are almost linear, shows optimal results.
- Best place to understand the data analysis
- Easily Explicable

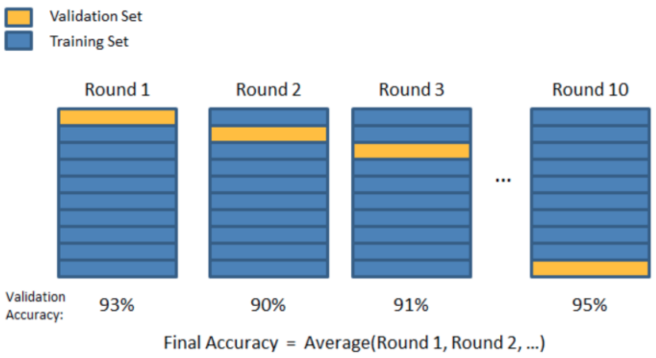
Disadvantages

- Linear regression is often inappropriately used to model non-linear relationships.
- Linear regression is limited to predicting numeric output.
- A lack of explanation about what has been learned can be a problem.
- Prone to bias variance problem

How to evaluate our model?







Overfitting vs Underfitting

Training Data(Less Error)

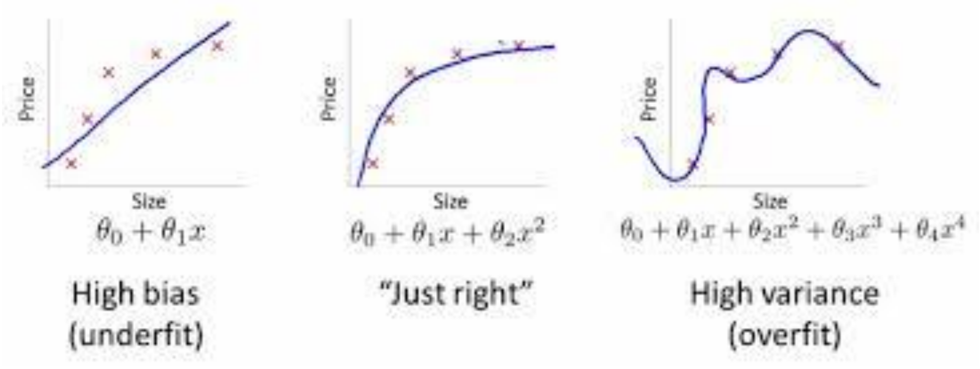
Testing Data (More Error)

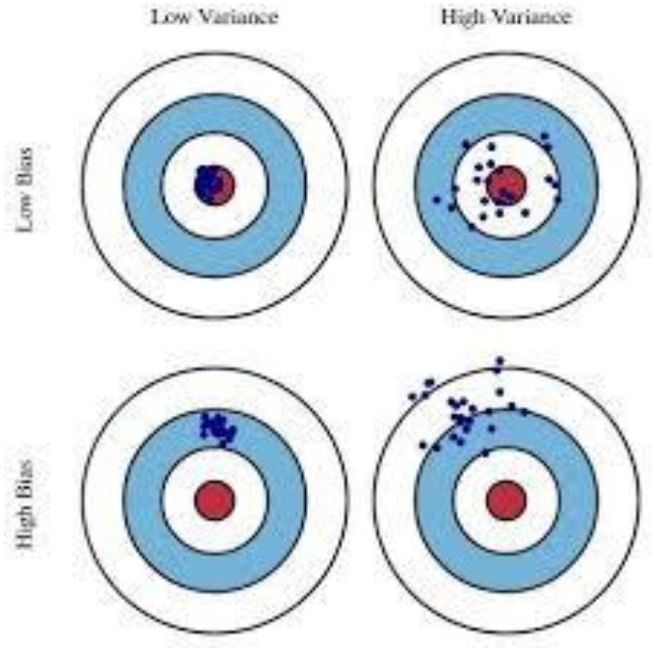
Overfitting vs Underfitting

Training Data (More Error)

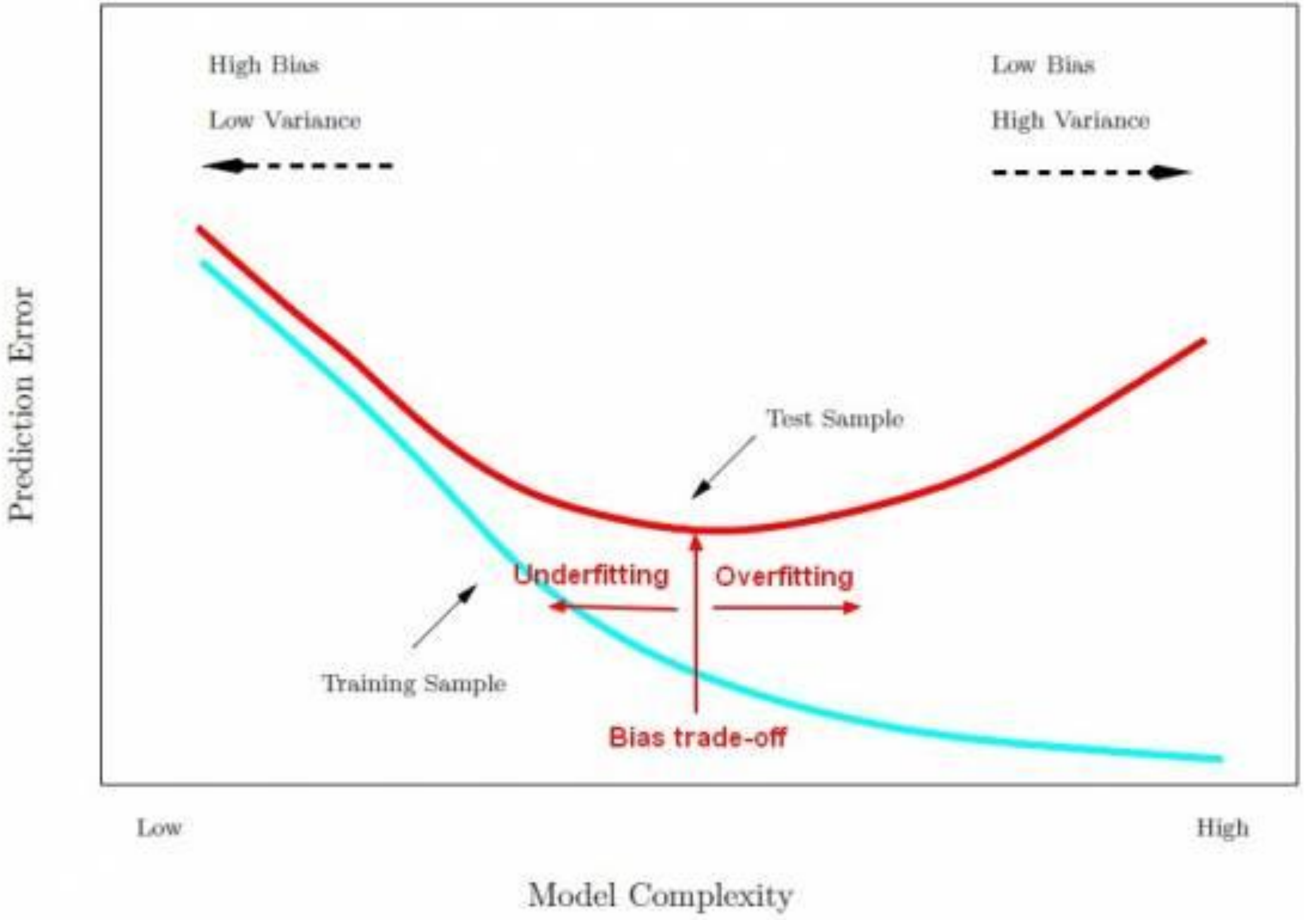
Testing (Still More Error)

Variance and Bias Trade off





Ideal Model should have Low varinance and Low Bias


High Bias
Low Variance
Low Bias
High Variance
Prediction Error
Test Sample
Underfitting
Overfitting
Training Sample
Bias trade-off
Low
High
Model Complexity