beingdatum
the data society

Machine Learning

BeingDatum.com
contact@beingdatum.com

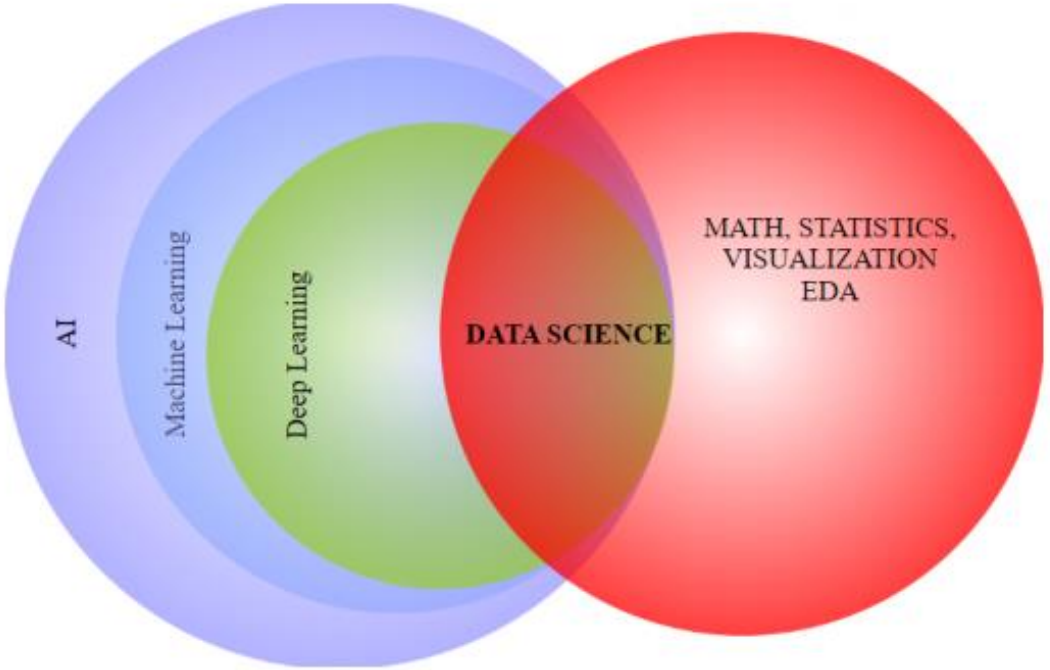
Agenda

- Introduction
- Use Cases
- Essential libraries
- Types of Learning
- Feature Scaling
- Classification Algos
- Regression Algos
- Clustering Algos
- Association Rule Learning
- Ensemble Techniques
- Time Series Analysis
- Dimensionality Reduction

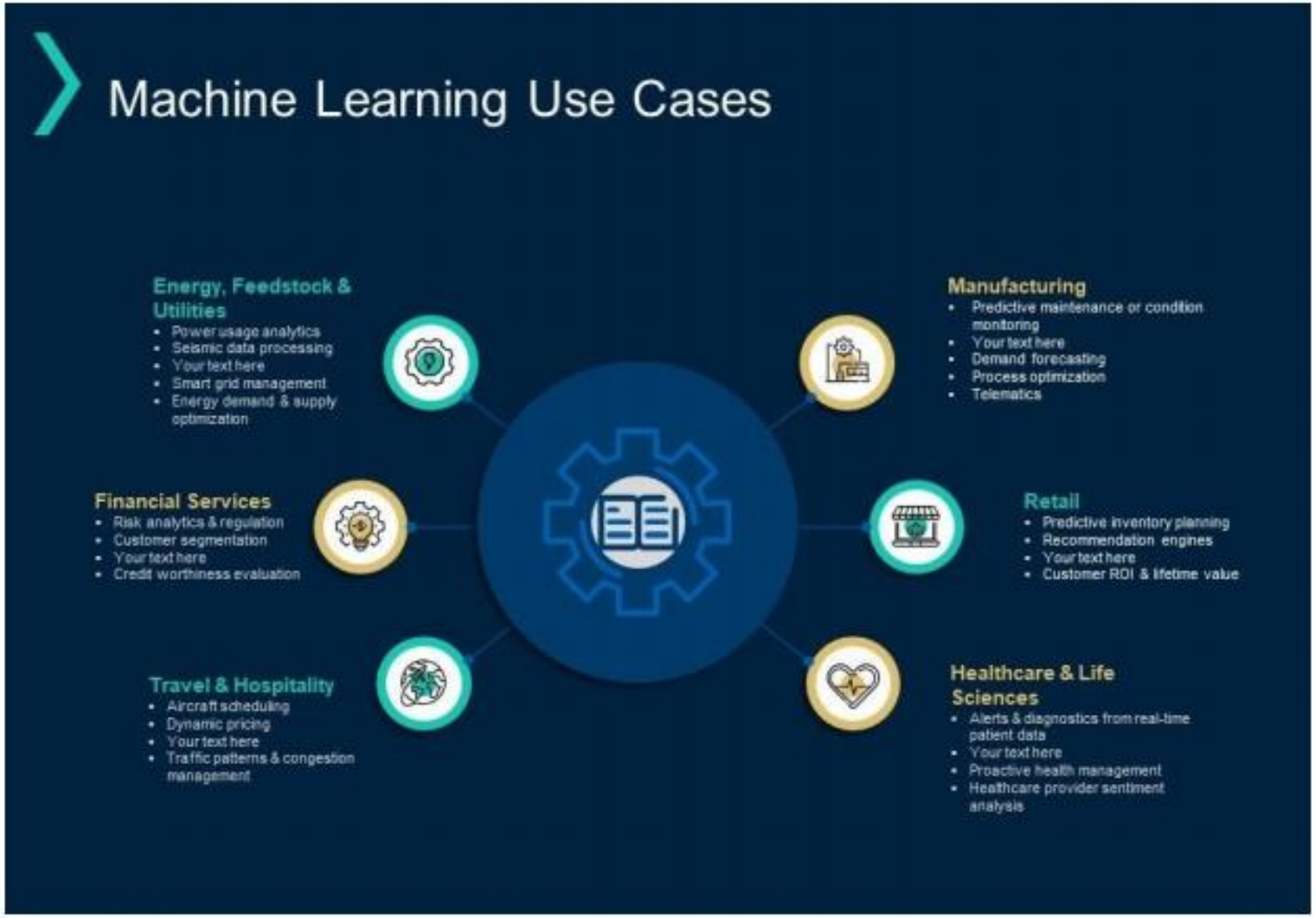
Intro

ML is a subset of AI.

Name derived from the concept that it deals with “construction & study of systems that can learn from data”



Use Cases



Essential Libraries

1. **numpy: The matrix / numerical analysis layer at the bottom**
2. **scipy: Scientific computing utilities (linalg, FFT, signal/image processing...)**
3. **sklearn: Machine learning (our focus here)**
4. **matplotlib: Plotting and visualization**
5. **opencv: Computer vision**
6. **pandas: Data analysis**
7. **caffe, theano, minerva: Deep neural networks**
8. **spyder: The front end (Scientific Python Development Environment)**

Types of Learning

- **Supervised-**Predict unknown data attributes(outcomes), based on known attributes (predictors), Model built on training data set (has both predictors and outcomes).

Types – Regression (continuous outcome), Classification (classes)

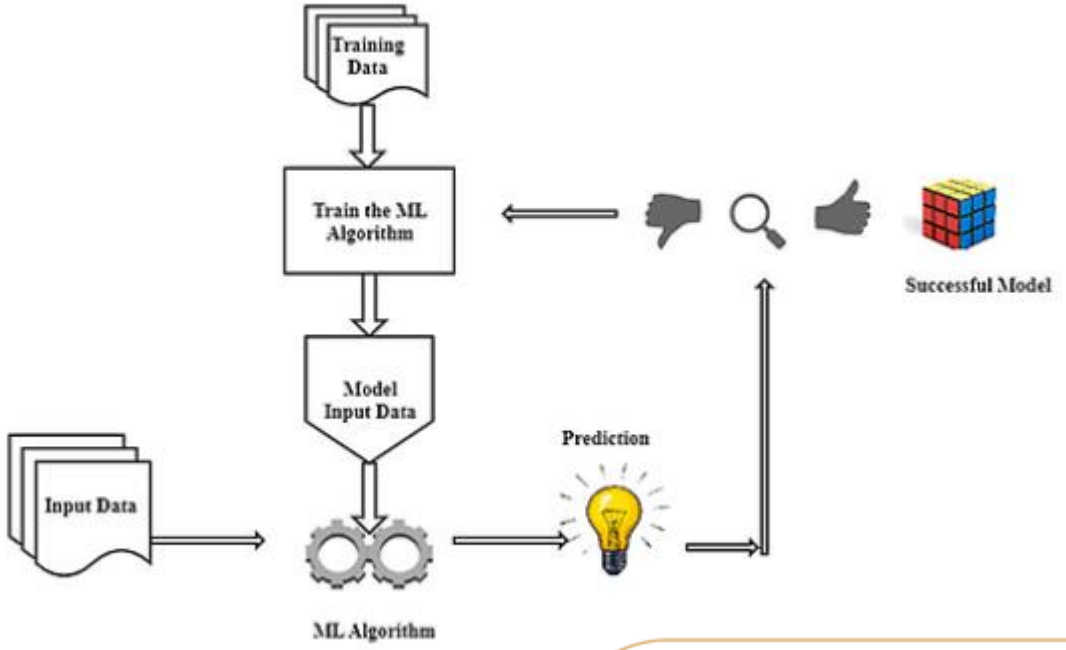
- **Unsupervised** – similarity / grouping entities.

Types- Clustering, Associative Rule Mining, Collaborative filtering

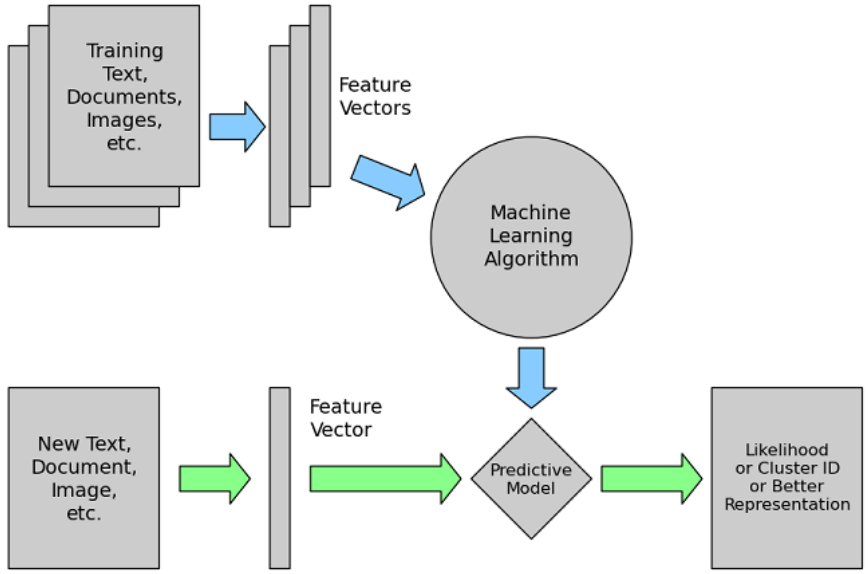
- **Reinforcement-** learn from feedback

Types- Deep approach, Inverse approach, apprentice approach.

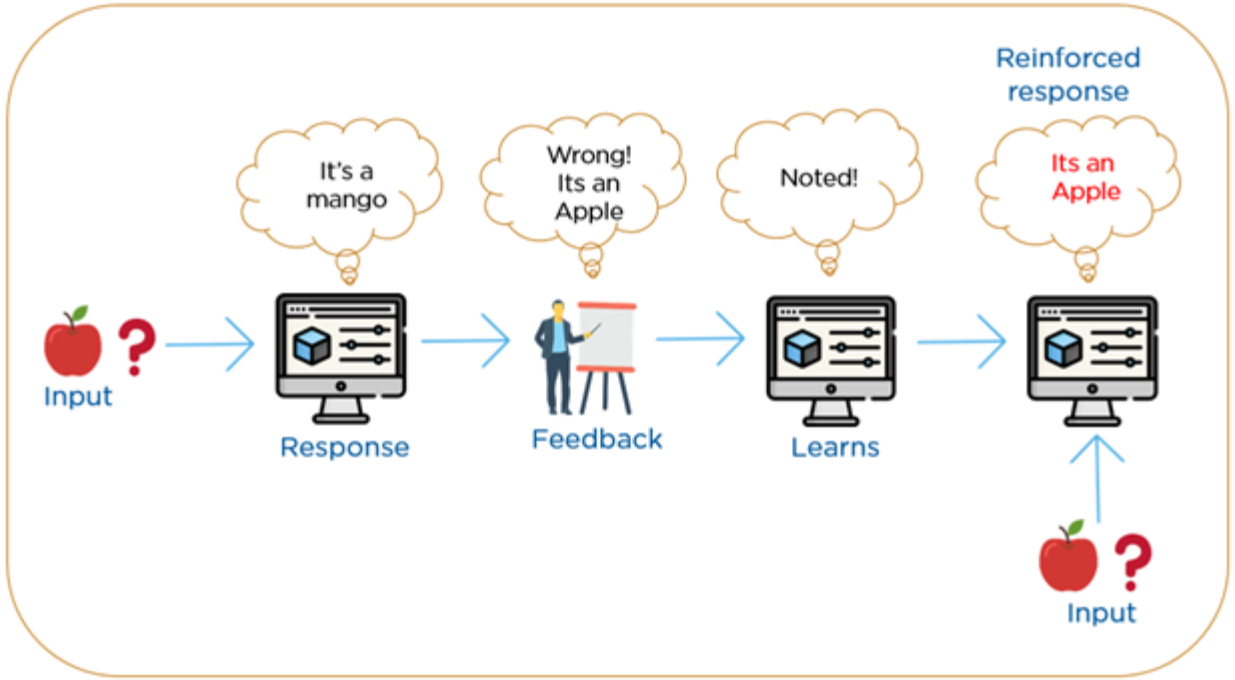
SUPERVISED



UN-SUPERVISED



RE-INFORCEMENT



Additional Concepts

- **Data preprocessing-**Machine only understand numbers, e.g. text data is converted to numerical factors i.e. Document term Matrix
- **Metrics**- Confusion matrix(based on prediction types) , ROC AUC etc.
- **Errors** – In sample, Out of sample - Over fitting, MAE,MASE etc
- **Algorithm Tuning-** Accuracy, Sensitivity, Specificity, Precision etc, Hyperparameter Tuning, Cross validation

Data Preprocessing

Outlier identification

The difference between a good and an average machine learning model is often its ability to clean data. One of the biggest challenges in data cleaning is the identification and treatment of outliers.

Players	Scores
Player1	500
Player2	350
Player3	10
Player4	300
Player5	450

In simple terms, outliers are observations that are significantly different from other data points. Even the best machine learning algorithms will underperform if outliers are not cleaned from the data because outliers can adversely affect the training process of a machine learning algorithm, resulting in a loss of accuracy.

Hands On- Boston House Pricing Dataset which is included in the sklearn dataset API

Discretization/Binning

Converting numeric to categorical

- **Manual**

If age is less than 40 then 1,

If between 40 to 60 then 2,

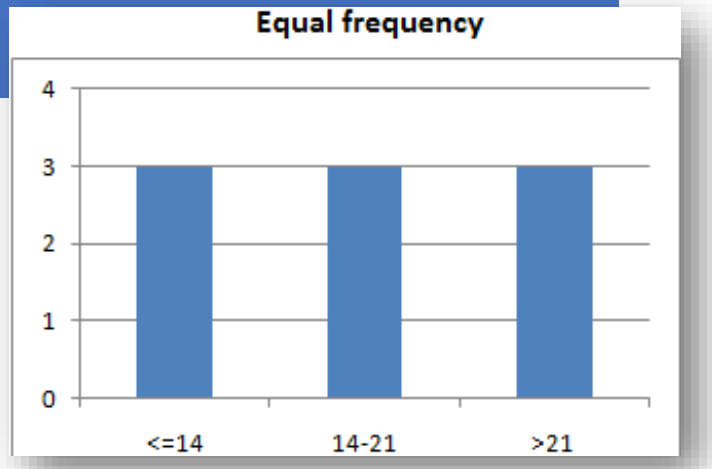
If more than 60 then 3

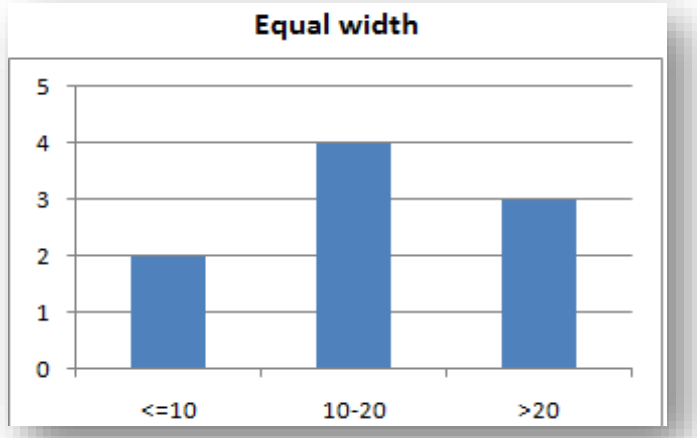
- Equal width

Interval is same, data is divided into *k* groups where each group has approximately the same value .

- Equal frequency

Number of sample in each bin is same, data is divided into *k* intervals of equal size.





Handling missing values

Data is not always fully available. e.g., many tuples have no recorded value for several attributes, such as customer income in sales data

- Missing data may be due to equipment malfunction
  - E.g., system failure
- data not entered due to misunderstanding
  - E.g., cities/zipcodes
- certain data may not be considered important at the time of entry
  - E.g., age of customers
- no registered history or changes of the data
  - E.g., employee promotion or change in designation

Hence, missing data may need to be inferred.

Data Cleansing: action

- Fill in missing values:
  - Ignore the tuple
  - Fill in the missing values manually: tedious + infeasible?
  - Use a global constant to fill in the missing value: e.g., “unknown”, a new class?!
  - Central imputation
    - Use the attribute mean (or majority nominal value) to fill in the missing value.
  - kNN imputation
    - Imputation using k-nearest neighbors. For each record, identify missing features. For each missing feature find the k nearest neighbors which have that feature. Impute the missing value using the imputation function on the k-length vector of values found from the neighbors.

    - Scenarios of NULL -> 0

    ""

    " "

    NA

    " "

    NULL

Stock	Price
Day1	22.4
Day2	20
Day3	19
Day4	
Day5	22.7
Day6	18.5

Central imputation

- Suppose the price value for day 4 is missing.
  - Compute the mean/median price for the given data
  - Substitute the value

- Average price = (22.4 + 20 + 19 + 22.7 + 18.5)/5
- = 20.52
- Median price = 20

- Hence, the Price for Day 4 = 20.5 or 20

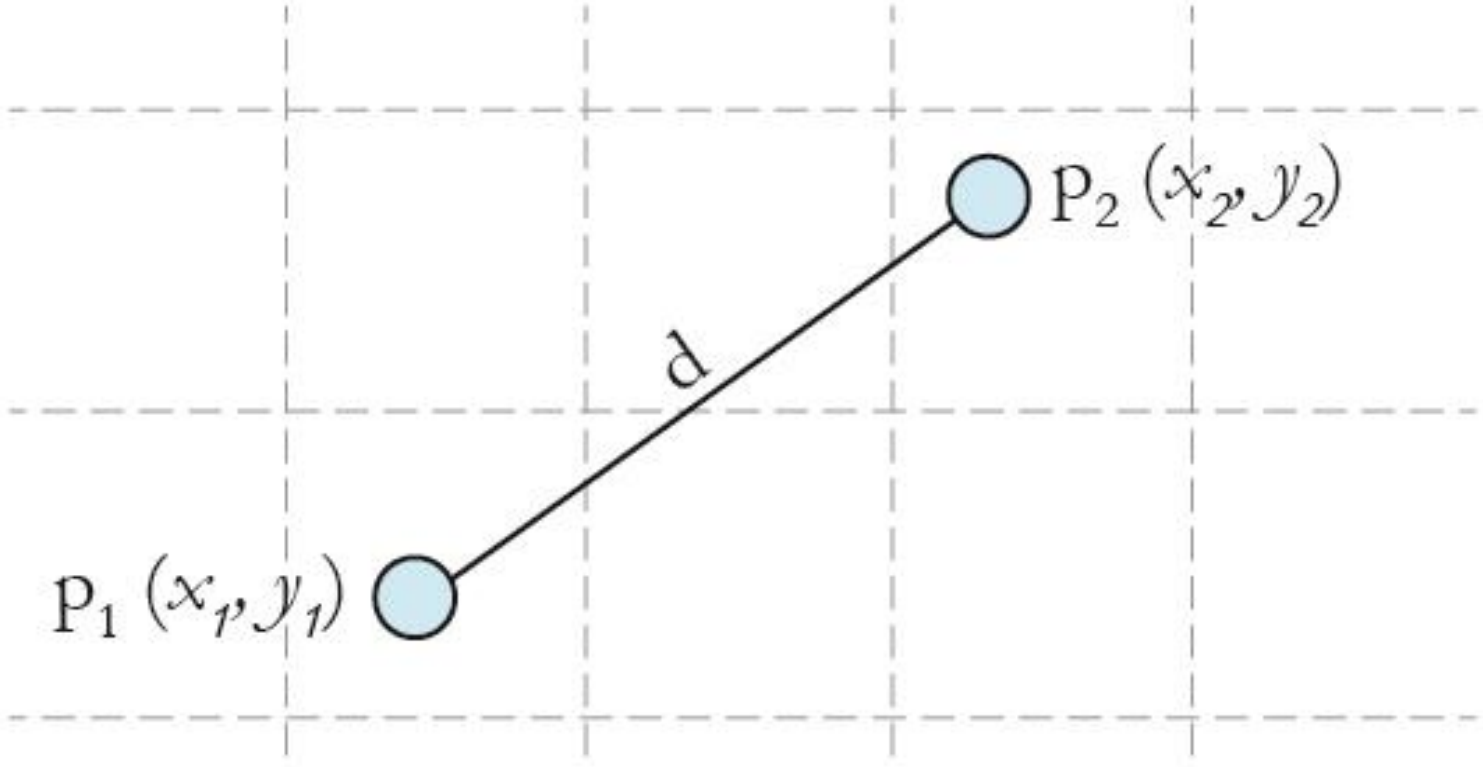
ToyCategory	Price
Infant	46
Infant	20
1 year	19
	45
2-3 years	22.7
4+	18.5

Central imputation

In the case, when the attribute is categorical then substitute with mode

Mode = Infant

Euclidean distance



$$\text{Euclidean distance (d)} = \sqrt{(x_2\text{-}x_1)^2 + (y_2\text{-}y_1)^2}$$

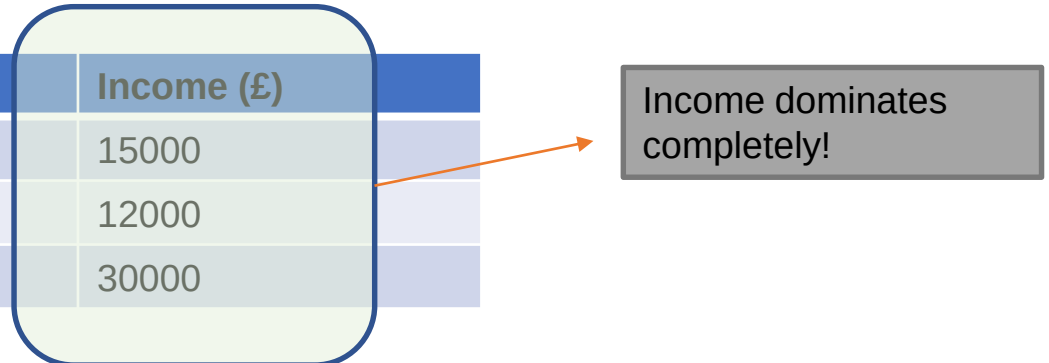
Feature Scaling

Sepal Length, Sepal Width → What is the magnitude & Units.

Most ML's model use Euclidean distance, hence without feature scaling, most of the algo's neglect the Units and focus on Magnitude, in that case the Euclidean distance may vary significantly → Hence, o/p will be impacted.

For e.g., the variable Age that ranges between 0 – 100 outweighs the Income that ranges between 10,000 – 50,000.

Age	Income (£)
24	15000
30	12000
28	30000



Feature Scaling techniques

Standard Scaler

Min-Max Scaling

Normalization

Others..

Scaling techniques- Min-Max

- **Min-Max scaling/0-1**: each variable in the data set is recalculated as

$$(V - \min V)/(\max V - \min V)$$

where V represents the value of the variable in the original data set.

This method allows variables to have differing means and standard deviations but equal ranges.

In this case, there is at least one observed value at the 0 and 1 endpoints.

Example

Age	Income (£)	New value
24	15000	(15000 – 12000)/18000 = 0.16667
30	12000	(12000 – 12000)/18000 =0
28	30000	(30000 – 12000)/18000 =1

Income_Minimum = 12000
Income_Maximum = 30000
(Max – min) = (30000 – 12000) = 18000

Please note, the new values have
Minimum = 0
Maximum = 1

Hence, we have converted the income values between 0 and 1.

Scaling techniques- Standard

- **Standard/ Z-score scaling**:

variables recalculated as (*V* - mean of *V*)/s,

where "s" is the standard deviation. As a result, all variables in the data set have equal means (0) and standard deviations (1) but different ranges.

Example

Age	Income (£)	New value
24	15000	(15000 - 19000)/9643.65 = -0.4147
30	12000	(12000 - 19000)/9643.65 = -0.7258
28	30000	(30000-19000)/9643.65 = 1.1406

Average = (15000 + 12000 + 30000)/3 = 19000
Standard deviation = 9643.65

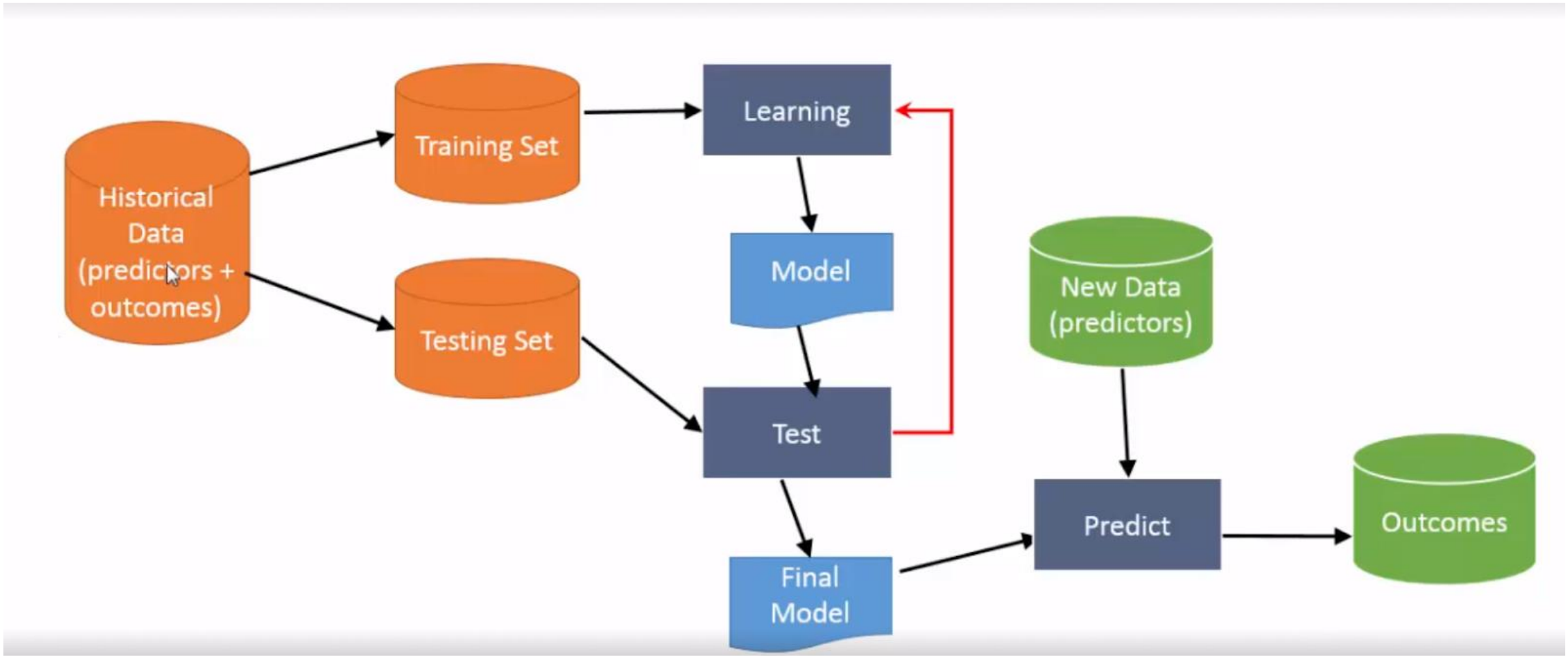
Hence, we have converted the income values to lower values using the z-score method.

x = c(-0.4147, -0.7258, 1.1406)
mean(x) = -0.000003 ~ 0
var(x) = 0.999 ~1

X	Income (£)
10000	15000
3	12000
28	30000

Age	Income
24	-0.4147
30	-0.7258
28	1.1406

Supervised Learning


Historical Data (predictors + outcomes)
Training Set
Testing Set
Learning
Model
Test
Final Model
New Data (predictors)
Predict
Outcomes

Training and Testing Data

- Historical Data contains both predictors and outcomes
- Split as training and testing data
- Training data is used to build the model
- Testing data is used to test the model
- Apply model on testing data
- Predict the outcome
- Compare the outcome with the actual value
- Measure accuracy
- Training and Test fit best practices
- 70-30 split
- Random selection of records. Should maintain data spread in both datasets

Regression

Regression analysis is a form of predictive modelling technique which investigates the relationship between a dependent (target) and independent variable (s) (predictor). This technique is used for forecasting, time series modelling and finding the causal effect relationship between the variables.

Let’s go through various regression algorithms:

1. Linear Regression
2. Multiple Linear Regression
3. Polynomial Regression
4. Stepwise Regression

Linear Regression

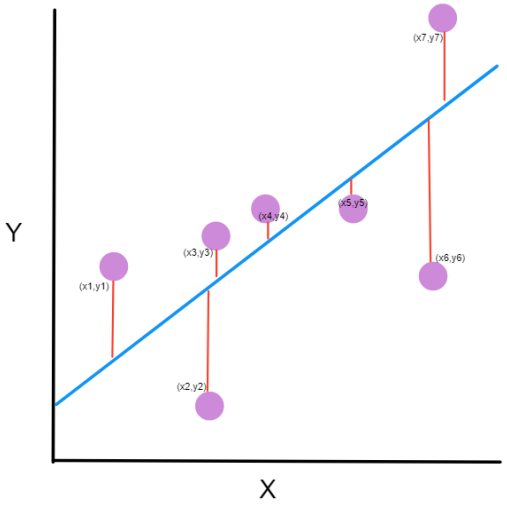
Simple linear regression is a statistical method that enables users to summarise and study the relationships between two continuous (quantitative) variables. Linear regression is a linear model wherein a model that assumes a linear relationship between the input variables (x) and the single output variable (y). Here, y can be calculated from a linear combination of the input variables (x). When there is a single input variable (x), the method is called a simple linear regression. When there are multiple input variables, the procedure is referred to as multiple linear regression.

Given our simple linear equation

$$y=mx+b$$

we can calculate MSE as:

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



Note

- N is the total number of observations (data points)
- $\frac{1}{N}\sum_{i=1}^{n}$ is the mean
- y_i is the actual value of an observation and $mx_i + b$ is our prediction

Multiple Linear Regression

The difference between simple linear regression and multiple linear regression, multiple linear regression has (>1) independent variables, whereas simple linear regression has only 1 independent variable.

Simple Linear Regression → y = bo + b1*x1

Multiple Linear Regression → y = bo + b1*x1 + b2*x2 + …. + bn*xn

Where, y → Dependent variable

x1, x2, ….xn → Independent variables

dataset - DataFrame

Index	R&D Spend	Administration	Marketing Spend	State	Profit
0	165349	136898	471784	New York	192262
1	162598	151378	443899	California	191792
2	153442	101146	407935	Florida	191050
3	144372	118672	383200	New York	182902
4	142107	91391.8	366168	Florida	166188
5	131877	99814.7	362861	New York	156991
6	134615	147199	127717	California	156123

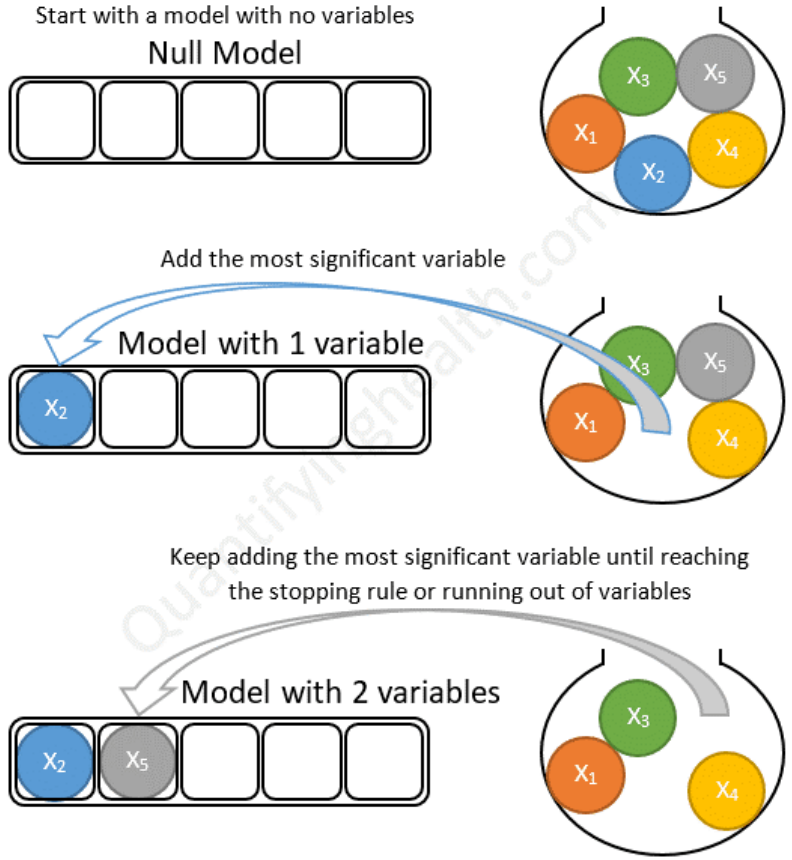
Multiple Linear Regression

5 methods of building models:

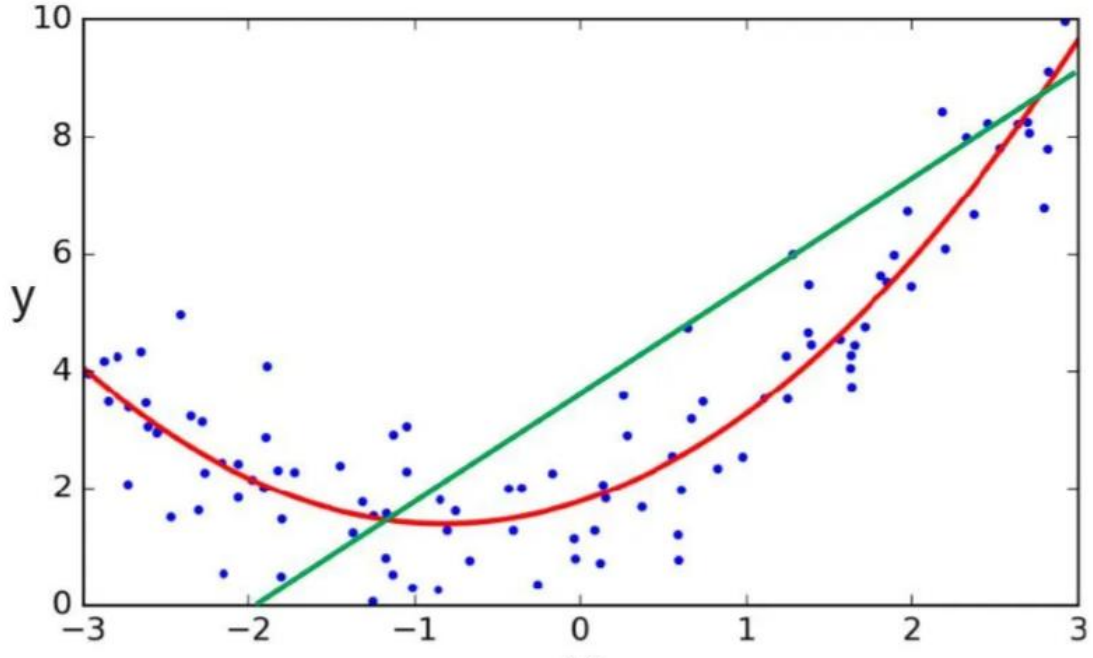
1. All-in
2. Backward Elimination → Stepwise
3. Forward Selection → Stepwise
4. Bidirectional Elimination → Stepwise
5. Score Comparison

All in cases: 2^n-1 = 2^10-1 = 1023 models..

Forward stepwise selection example with 5 variables:



Polynomial Regression



In this type of data, linear equation might not be a good fit as seen in the green line, hence we use Polynomial equation i.e. red line, which completely fits in with the data points and comes with an equation of

$$y = b_0 + b_1 x_1 + b_2 x_1^2$$

The equation of Polynomial Regression is:

Simple Linear Regression	$y = b_0 + b_1 x_1$
Multiple Linear Regression	$y = b_0 + b_1 x_1 + b_2 x_2 + \ldots + b_n x_n$
Polynomial Linear Regression	$y = b_0 + b_1 x_1 + b_2 x_1^2 + \ldots + b_n x_1^n$

Covariance

The **covariance** of two variables *x* and *y* in a data set measures how the two are linearly related. A positive covariance would indicate a positive linear relationship between the variables, and a negative covariance would indicate the opposite.

Similarly, the **covariance** is defined in terms of the mean as:

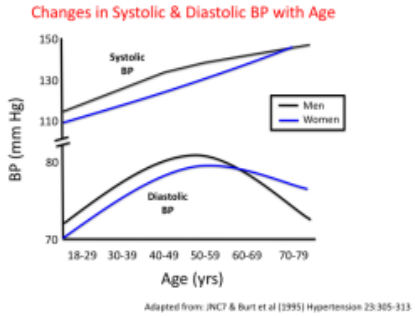
$$s_{xy} = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})$$

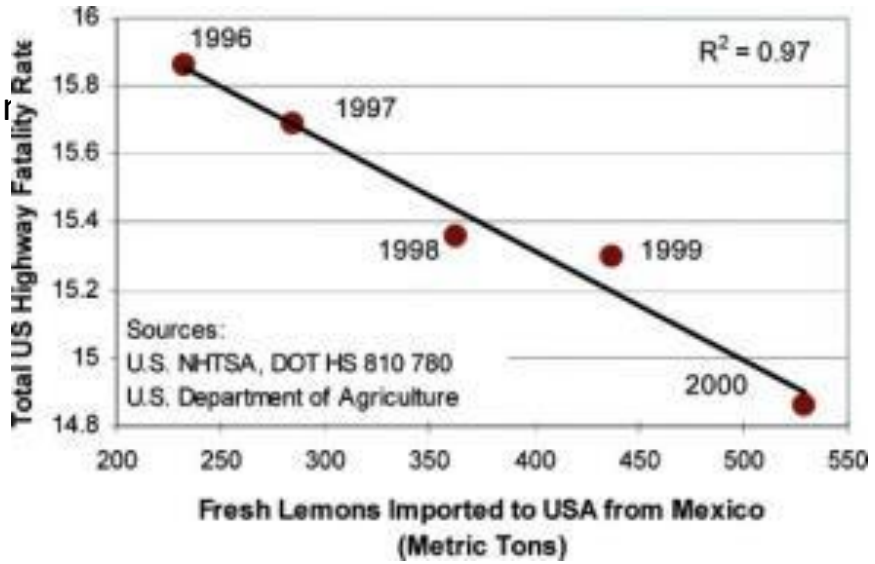
However, covariance can be a very large number. It is best to express it as a normalized number between -1 and 1 to understand the relation between both quantities. This is achieved by normalizing covariance with standard deviations of both variables.

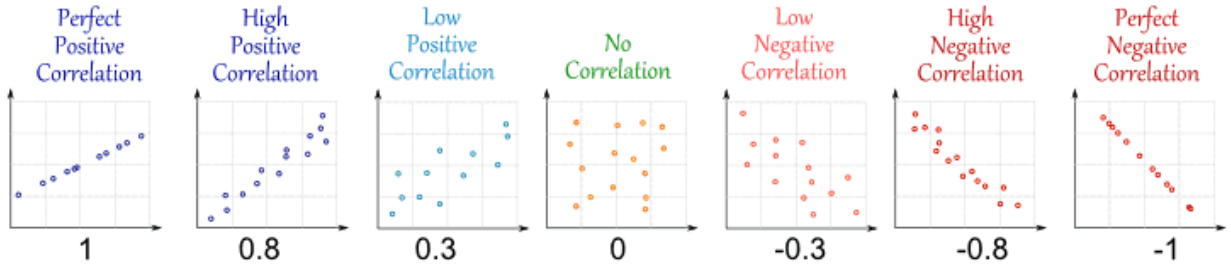
Correlation

- **Correlation** : a mutual relationship or connection between two or more things
  - Interdependence & Correlation between 2 sets of data ar
- **Example** : Age and Blood Pressure
- Pearson's Correlation co-efficient -1 to +1
- **Causation** : The reason for a change in value
- Correlation does not imply causation
- Correlation might be **due** to –

Causation, Common cause, Incidental







$$Cor(x,y) = \rho_{xy}/\sigma_x\sigma_y$$

Correlation Coefficient

Correlation Coefficient, r, gives the strength and direction of the relationship between two variables.

$r = \frac{bs_x}{s_y} = \frac{s_{xy}}{s_x s_y}$ where b is the slope of the line of best fit, s_x is the standard deviation of the x values in the sample, s_y is the standard deviation of the y values in the sample and s_{xy} is the covariance between x and y.

Coefficient of Determination

The coefficient of determination is given by r^2 or R^2. It is the percentage of variation in the y variable that is explainable by the x variable.

If $r^2 = 0$, it means you can't predict the y value from the x value.

If $r^2 = 1$, it means you can predict the y value from the x value without any errors.

Usually, r^2 is between these two extremes.

Covariance, Correlation and R^2

Day	Interest Rate	Futures Index	$x - \bar{x}$	$y - \bar{y}$	$(x - \bar{x}) * (y - \bar{y})$
1	7.43	221	-0.31	-6.08	1.91
2	7.48	222	-0.26	-5.08	1.34
3	8.00	226	0.26	-1.08	-0.28
4	7.75	225	0.01	-2.08	-0.01
5	7.60	224	-0.14	-3.08	0.44
6	7.63	223	-0.11	-4.08	0.47
7	7.68	223	-0.06	-4.08	0.26
8	7.67	226	-0.07	-1.08	0.08
9	7.59	226	-0.15	-1.08	0.17
10	8.07	235	0.33	7.92	2.58
11	8.03	233	0.29	5.92	1.69
12	8.00	241	0.26	13.92	3.56
Mean	**7.74**	**227.08**			**Sum = 12.216**
StDev	**0.22**	**6.07**			

$$Cov = \frac{12.216}{11} = 1.111$$

$$r = \frac{1.111}{0.22 * 6.07} = 0.815$$

$$R^2 = 0.815^2 = 0.665$$

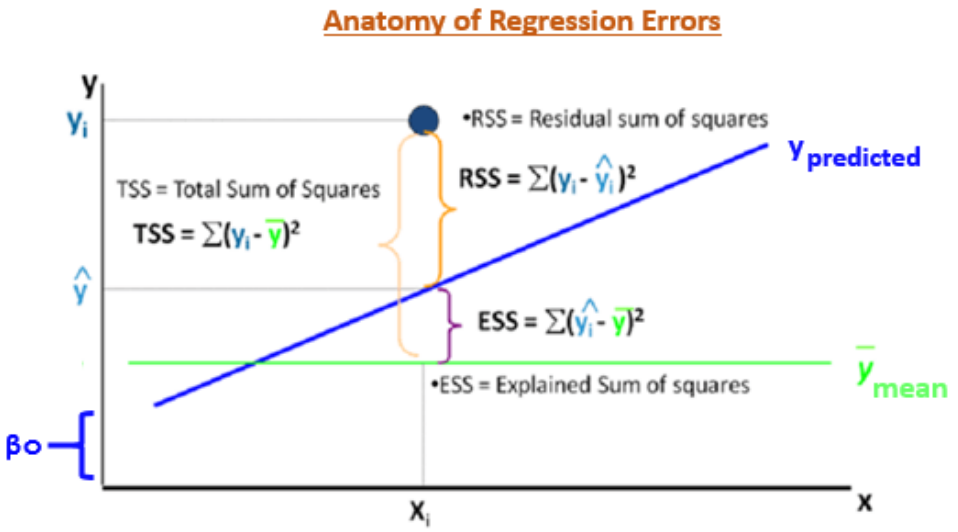
“p-values” and Significance Levels

- p-value suggests the probabiity of our experiment results(stats params with sample population) happening, p-value less than alpha value means there is less chance of experiment results happening
- In Linear Regression, the Null Hypothesis is that the coefficients associated with the variables is equal to zero
- So, p-value of <0.05 means the we can reject null hypothesis that the coefficients associated with the variables is equal to zero. Hence, the variables are contributing to the model.

Regression Error Metrics

MAE, MSE, RMSE, MAPE etc.

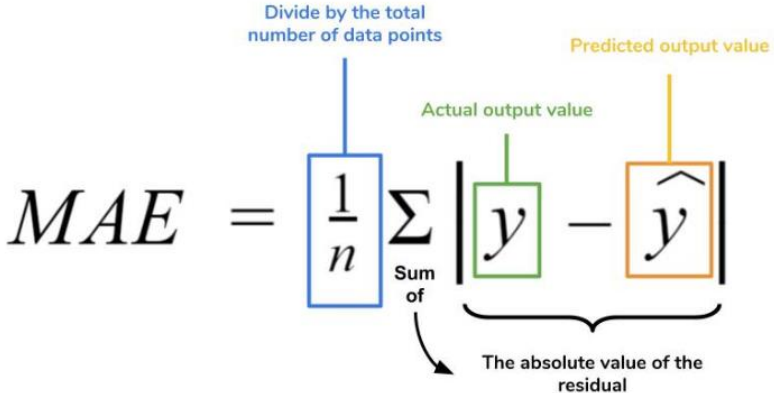
MSE (Mean Squared Rrror)
RMSE (Root Mean Squared Error)
MAE (Mean Absolute Error)
MAPE (Mean Absolute Percentage Error)

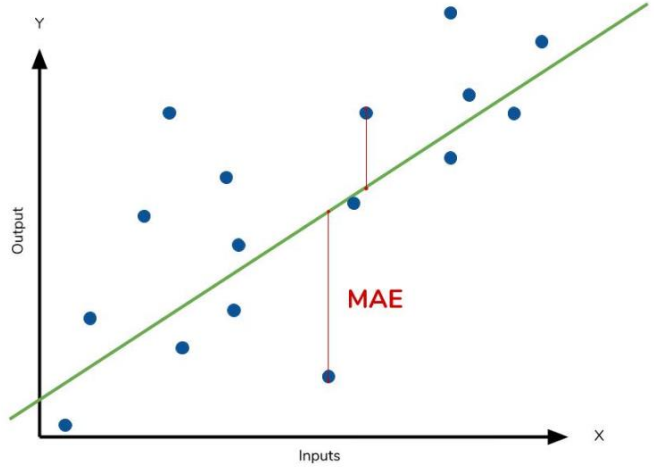


Mean Absolute Error

MAE: We just look at the absolute difference between data and model's predictions

Lower MAE → Better model

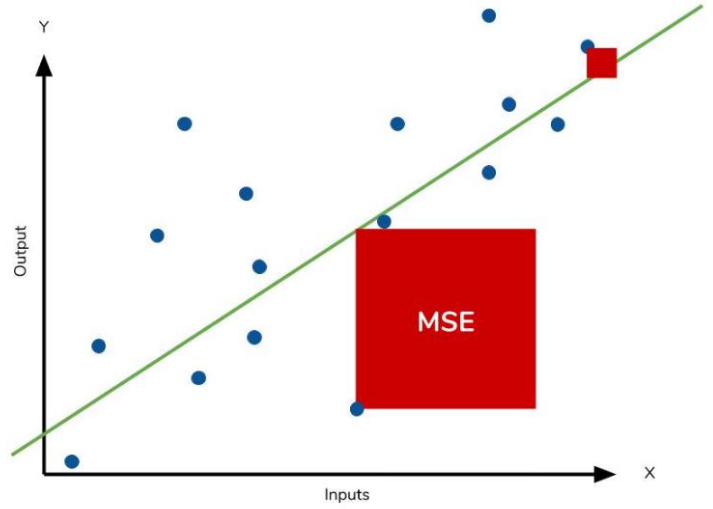




Mean Squared Error

MSE is going to be a huge number because of squaring, hence, we can't compare it with MAE. This ultimately means that outliers in our data will contribute to much higher total error in the MSE than they would the MAE. Similarly, our model will be penalized more for making predictions that differ greatly from the corresponding actual value. This is to say that large differences between actual and predicted are punished more in MSE than in MAE.

$$MSE = \frac{1}{n}\Sigma\underbrace{\left(y - \hat{y}\right)^2}_{\text{The square of the difference}}$$



Root Mean Squared Error

RMSE = Square root(MSE)

As the name suggests, it is the square root of the MSE. Because the MSE is squared, its units do not match that of the original output. Researchers will often use RMSE to convert the error metric back into similar units, making interpretation easier. Since the MSE and RMSE both square the residual, they are similarly affected by outliers.

Improve Accuracy

How can you check and improve the accuracy of a regression model ?

- You can do variable selection based on **p values**. If a variable shows p value > 0.05, we can remove that variable from model since at p> 0.05, we'll always fail to reject null hypothesis. Try for Stepwise Regression.
- Higher R squared
- Error metrics- As low as possible
- Do as much Pre-processing as possible.

QUESTIONS??