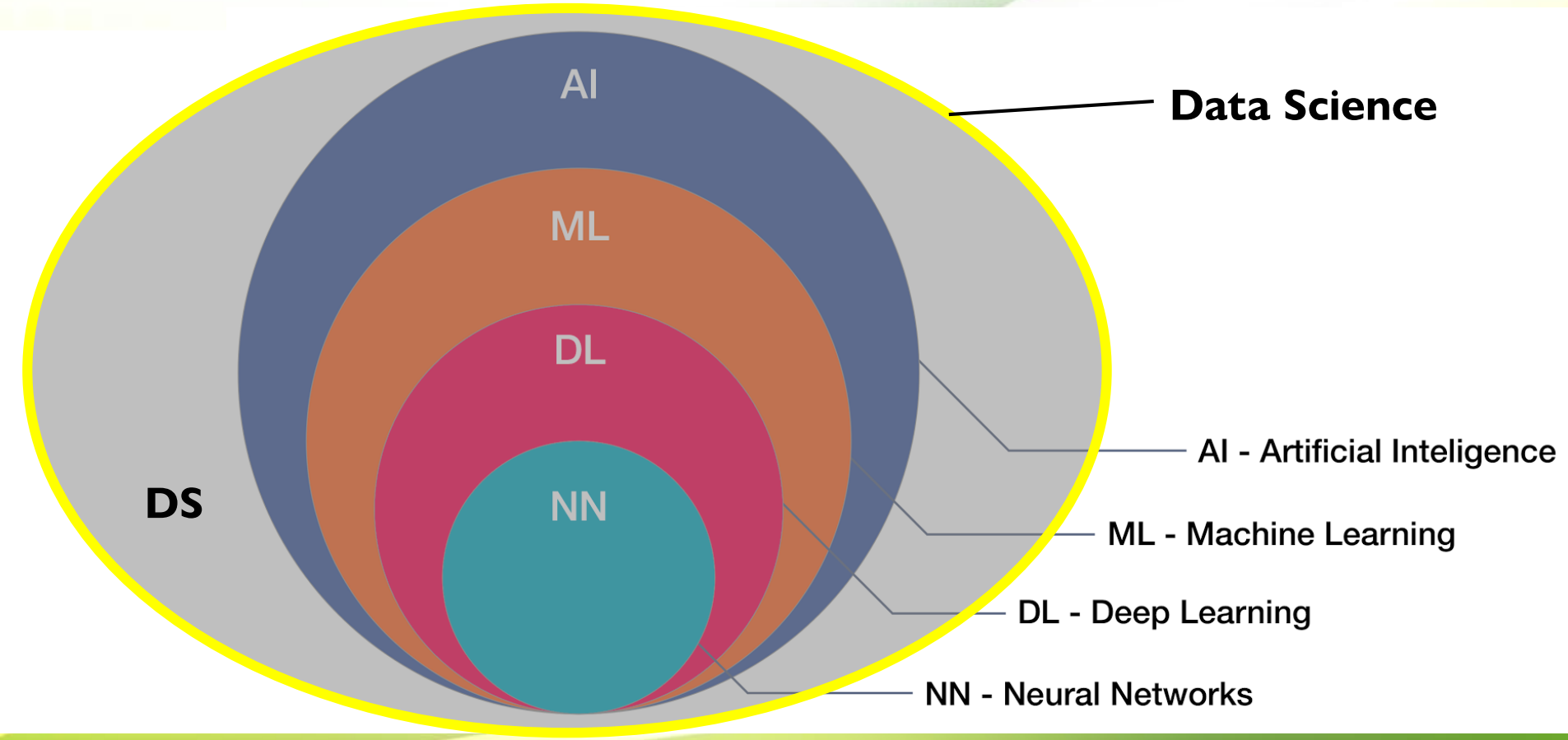
INTRODUCTION TO PREDICTIVE ANALYTICS

2nd Sem, MCA

CONTENT

- Introduction Predictive analytics
  - Predictive modeling
  - Model performance measure
  - Regression Models – Linear, Logistic and others
  - K-nearest neighbor
  - Classification & Regression model

PREDICTIVE ANALYTICS



DATA SCIENCE

- **Data is EVERYTHING;** a new form of revenue.
- Data gives better business insights; helps to uncover *(hidden)* patterns in data.
- Example:
  - One guy order for a computer. He also purchases a mouse and keyboard.
  - Data modeling will build this pattern.
  - Companies use this patter/relation for better business policy. Predictive analytics for future buying.
  - Companies use DS to build recommendation engines. Prescriptive analytics in DS.
- Various algorithms could be applied on data to get more accurate results.
- Running these algorithms on huge datasets needs AI, ML, DL.
- ML is used in DS to make predictions by discovering hidden patterns in data.

ARTIFICIAL INTELLIGENCE

- AI enables Machine to think.
- Adding intelligence to our system in artificial way.
- Ability that enables machines to understand data, learn from data, and make decisions based on patterns hidden in the data, or inferences that could otherwise be very difficult *(almost impossible)* for humans to make manually.
- AI enables machines to adjust/use the gained "knowledge" based on new inputs that were not part of the data used for training these machines.
- Collection of mathematical algorithms that make computers understand relationships between different types & pieces of data to use knowledge to make decisions that could be accurate to a very high degree.

MACHINE LEARNING

- Machine's ability to learn.
- ML is an implementation of AI.
- Establish Relationship between independent & dependent variables present in data.
- ML is used in situations where machine should learn from huge amounts of data given to it ***(training dataset)***, and then apply that knowledge on new pieces of data that streams into the system.
- After learning/training phase is complete (with training data); ML model is tested on data which machine never encountered before ***(testing dataset).***
- Statistical tool to analyze data to get conclusive knowledge.

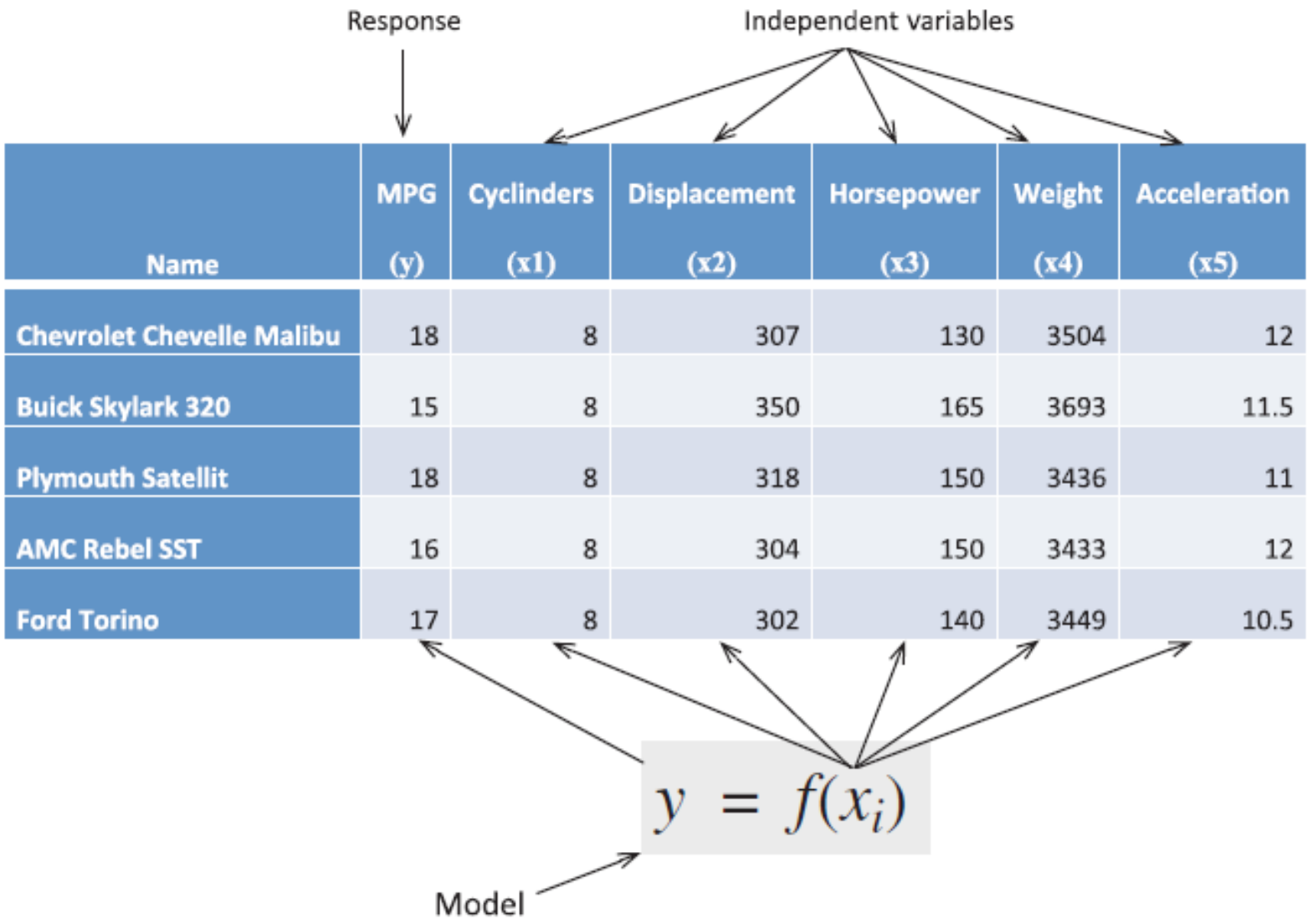
DEEP LEARNING

- Fills the gaps of ML limitations.
- ML involves structured data.
- DL also deals with unstructured data (video, audio, text, social media posts and images — anything and everything that humans communicate with that are not numbers or metric reads).
- Training dataset is relatively small for ML.
- When huge amounts of data to train a model, with data having too many features, and if high level of accuracy is critical, DL is more appropriate.
- DL requires much powerful hardware to run on (GPUs).
- Takes significantly more time to train models; generally more difficult to implement compared to ML.

Predictive Modeling

- **Predictive modeling** is a commonly used statistical technique to predict future behavior.
- Technology that analyzes historical/current data and generate model to help predict future outcomes.
- Data is collected, statistical model is formulated, predictions are made, and model is validated (or revised) as additional data becomes available.
- Various algorithms in ML used for *prediction problems, classification problems, regression problems,* etc.
- Predictive analytics models are:
  - *Classification model*
  - *Clustering model*
  - *Forecast model*
  - *Outliers model*
  - *Time series model*

Predictive Modeling

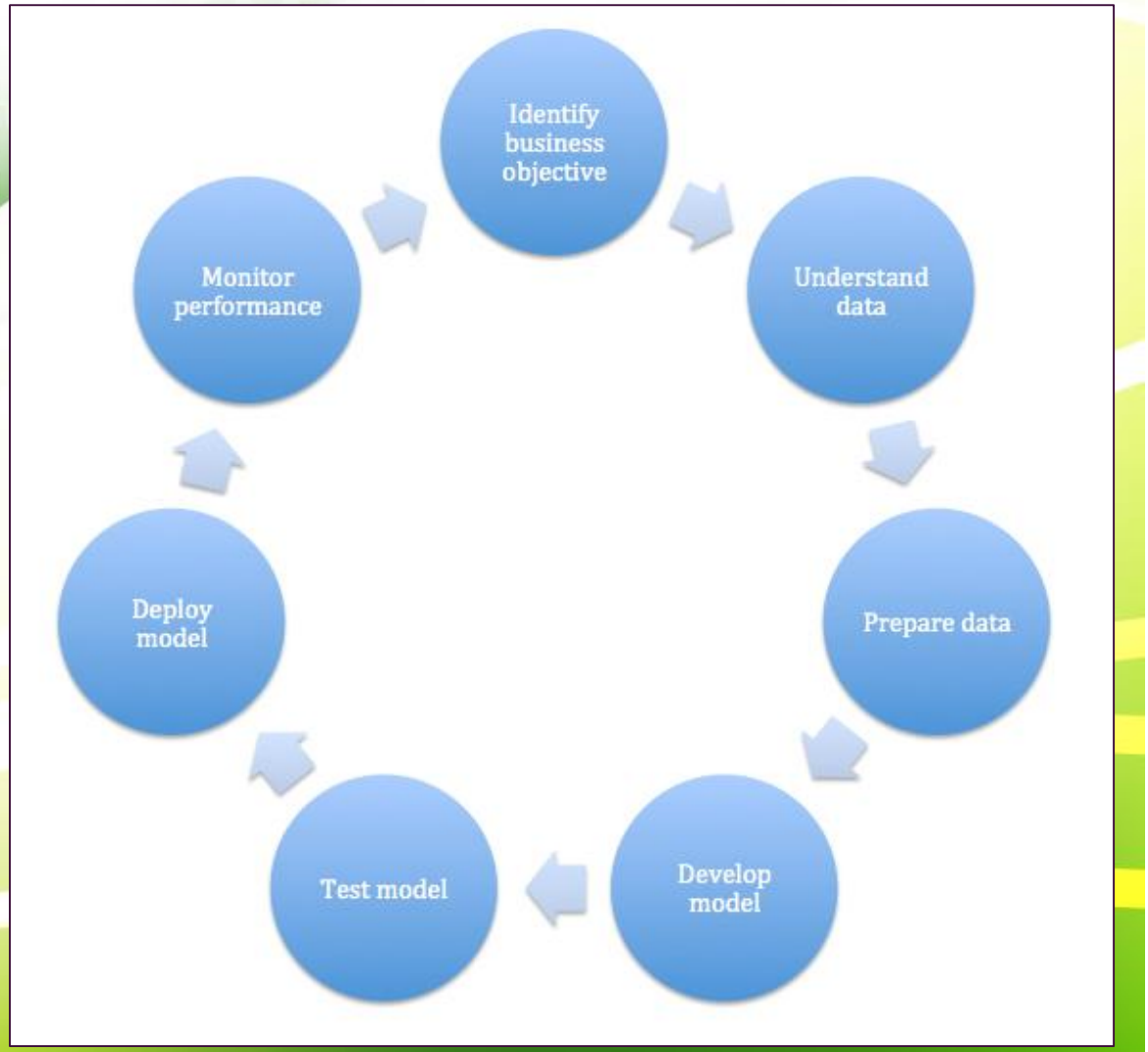


Name	MPG (y)	Cyclinders (x1)	Displacement (x2)	Horsepower (x3)	Weight (x4)	Acceleration (x5)
Chevrolet Chevelle Malibu	18	8	307	130	3504	12
Buick Skylark 320	15	8	350	165	3693	11.5
Plymouth Satellit	18	8	318	150	3436	11
AMC Rebel SST	16	8	304	150	3433	12
Ford Torino	17	8	302	140	3449	10.5

Predictive Modeling

Phases of Predictive Modeling

- *Understand Business Objective*
- *Define Modelling Goals*
- *Select/Collect Data*
- *Prepare Data*
- *Analyse and Transform Variables,* ***Sampling***
- ***Model*** *Selection, Develop Models (Training)*
- *Validate Models (Testing), Optimize, Profitability*
- *Document Methodology and Models*
- *Implement Models*
- *Monitoring and Performance Tracking*



Predictive Modeling

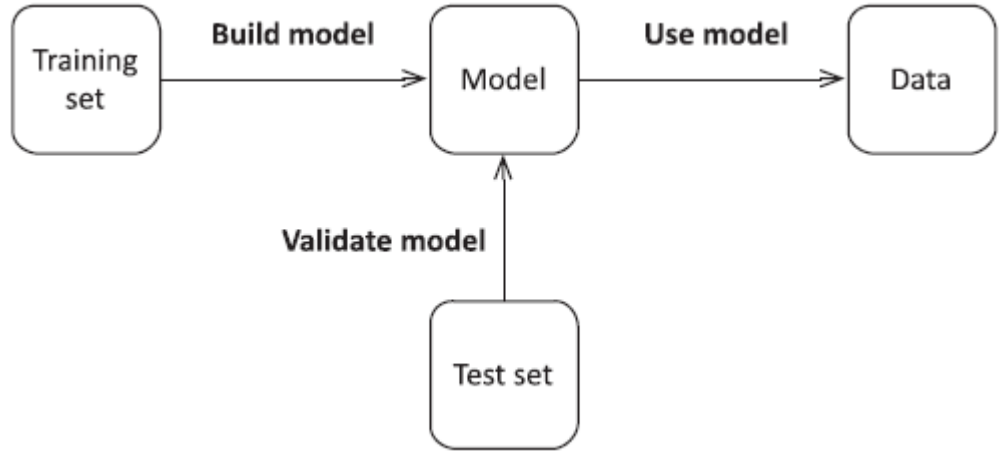
Predictive modeling limitations :

- Errors in data labeling.
- Shortage of massive data sets needed to train machine learning.
- Machine's inability to explain what and why it did what it did.
- Generalizability of learning, or rather lack thereof.
- Bias in data and algorithms.

Predictive Modelling

- ***Training** set:* data set used to build a model.
- ***Test set**:* data set used to test the model
- **Cross-validation:** validating model efficiency by training it on subset of input data and testing on previously unseen subset of input data.
- Three steps involved in cross-validation:
  - Reserve some portion of sample data-set.
  - Using the rest data-set train the model.
  - Test the model using the reserve portion of the data-set.
- Methods used for Cross-Validation:
  - Validation Set Approach (50%)
  - Leave-P-out cross-validation (p, n-p)
  - Leave one out cross-validation (1, n-1)
  - K-fold cross-validation
  - Stratified k-fold cross-validation

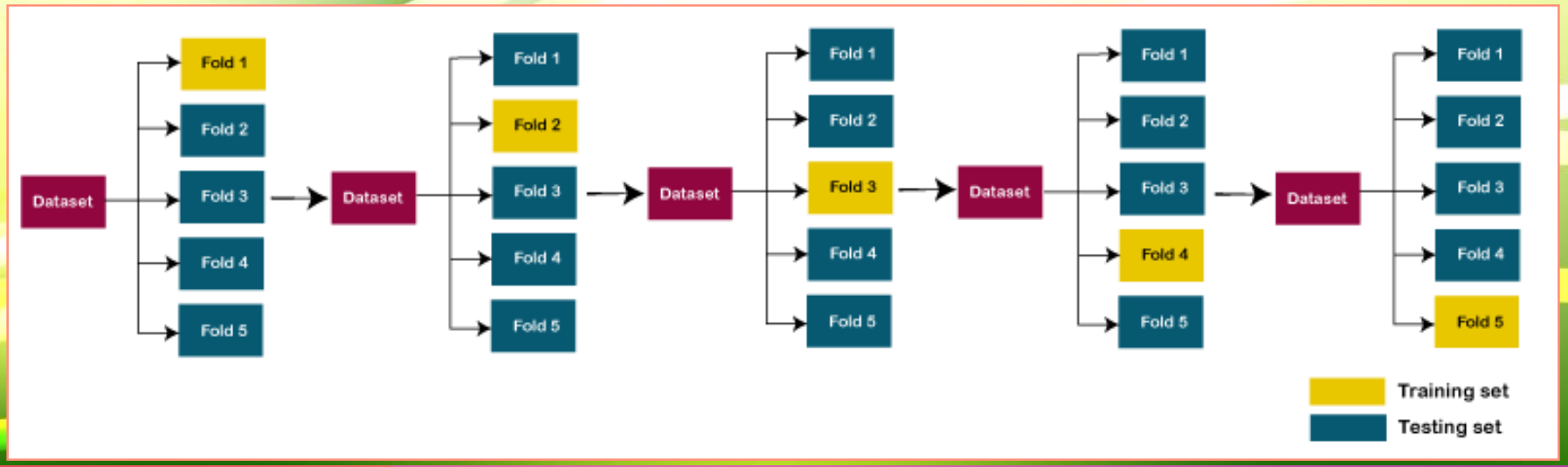
Name	MPG (y)	Cyclinders (x1)	Displacement (x2)	Horsepower (x3)	Weight (x4)	Acceleration (x5)
Chevrolet Chevelle Malibu	18	8	307	130	3504	12
Buick Skylark 320	15	8	350	165	3693	11.5
Plymouth Satellit	18	8	318	150	3436	11
AMC Rebel SST	16	8	304	150	3433	12
Ford Torino	17	8	302	140	3449	10.5



Predictive Modelling

K-fold Cross-validation:

1. Split input dataset into K groups/folds (equal size)
2. Repeat this step by each group on rotation:
   - Take one group as reserve or test dataset.
   - Use remaining groups as training dataset.
   - Fit the model on training set and evaluate performance of the model using the test set.
3. Accuracy of the model is based on average of the k scores.

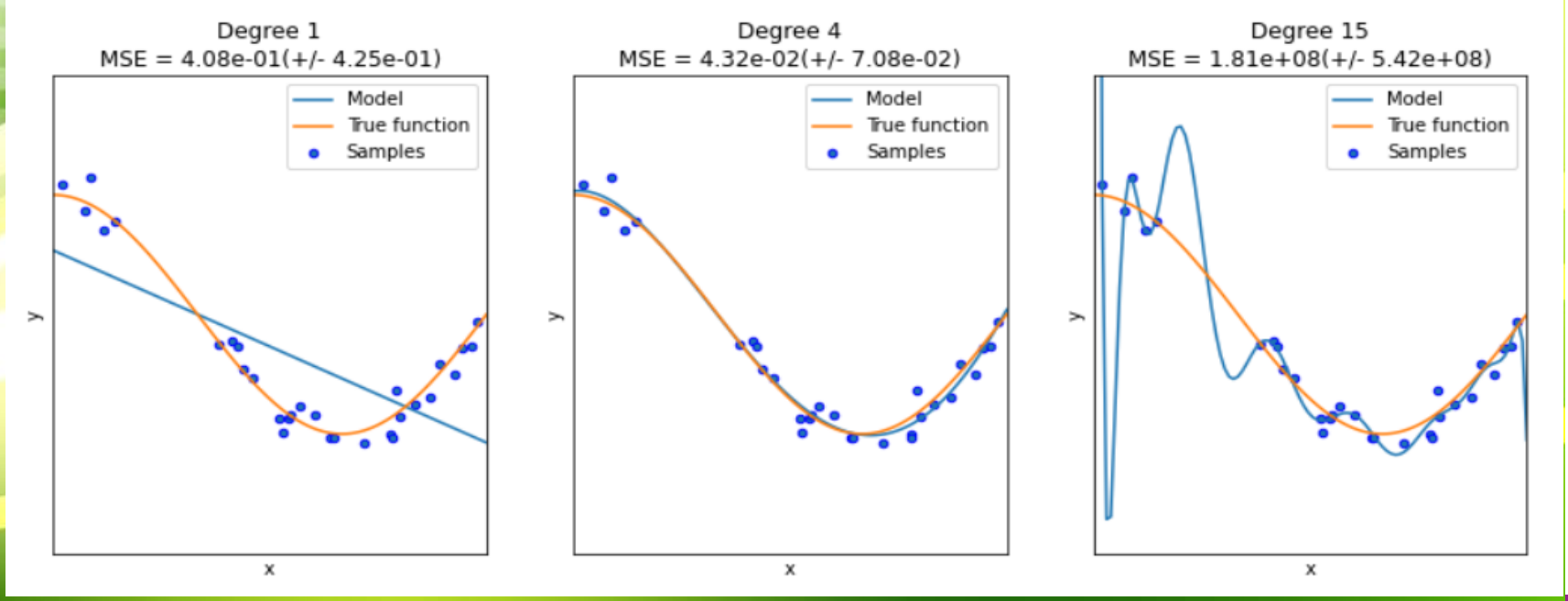


Predictive Modelling

- In statistics a **fit** is, **how close model is to target class/function/value**.
- Key components in ML modelling:
  - **Signal:** true underlying pattern of data that helps ML model to learn from data.
  - **Noise:** unnecessary and irrelevant data that reduces performance of model.
  - **Bias:** measure of model accuracy.
    - difference between predicted values and actual values.
    - prediction error that is introduced in model due to oversimplifying/optimizing ML algorithms.
  - **Variance:** If ML model performs well (low error) with training dataset, but does not perform well (high error) with test dataset.
- Overfitting and underfitting are two biggest causes of poor performance of ML models.

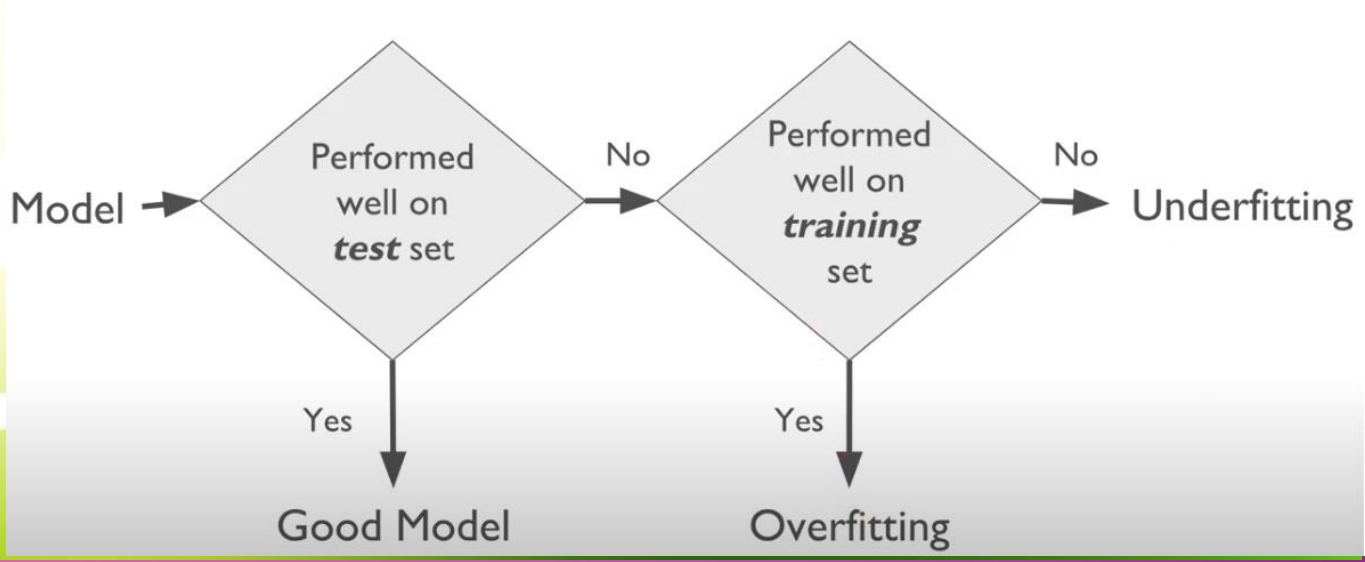
Predictive Modelling

- **Overfitting** model tries to cover all/more than required data points present in given dataset.
  - Model starts catching noise/inaccurate values present in dataset.
  - model can't generalize or fit well on unseen dataset, and reduce model accuracy.
  - Error on testing or validation dataset is much greater than error on training dataset.
  - Overfitted model has low bias and high variance.
- Avoiding Overfitting:
  - Cross-Validation
  - Training with more clean data
  - Removing features
  - Early stopping the training
  - Regularization
  - Ensembling



Predictive Modelling

- Opposite of overfitting is underfitting.
- **Underfitting** when model is not able to capture underlying trend of data.
  - To avoid overfitting in model, volume/time of training data is compromised.
  - Hence, model is not able to learn enough from training data, and hence it reduces accuracy and produces unreliable predictions.
  - Underfitted model has high bias and low variance.
- Avoid underfitting:
  - increasing training time/volume of model.
  - increasing number of features.



Predictive Modelling - Confusion Matrix

ACTUAL	Course-1	Course-2	Course-3
Pass	90	70	80
Fail	30	50	40

PREDICTED	Course-1	Course-2	Course-3
Pass	80	50	70
Fail	40	70	50

Course-1

	Predicted Pass	Predicted Fail	Total
Actual Pass	70	20	90
Actual Fail	10	20	30
Total	80	40	

Predictive Modelling - Confusion Matrix

Contingency Table for Predictive modelling Performance metrics.

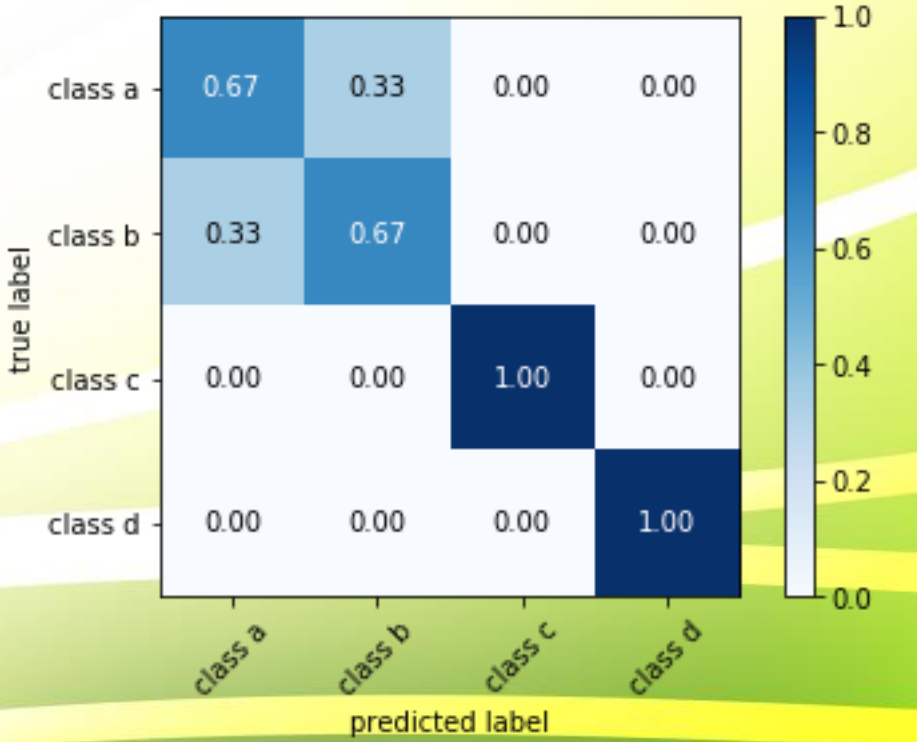
Confusion matrix: N X N matrix, where N is number of classes being predicted.

- **Error / residual:** difference between predicted value and actual value.
- **Accuracy / *Concordance***: proportion of total number of predictions that were correct.
- **Positive Predictive Value/Precision**: proportion of positive cases that were correctly identified.
- **Negative Predictive Value** : proportion of negative cases that were correctly identified.
- **Sensitivity or Recall** : proportion of actual positive cases which are correctly identified.
- **Specificity** : proportion of actual negative cases which are correctly identified.

Confusion Matrix		Target			
		Positive	Negative		
Model	Positive	a	b	*Positive Predictive Value*	a/(a+b)
	Negative	c	d	*Negative Predictive Value*	d/(c+d)
		Sensitivity	*Specificity*	**Accuracy** = (a+d)/(a+b+c+d)	
		a/(a+c)	d/(b+d)		

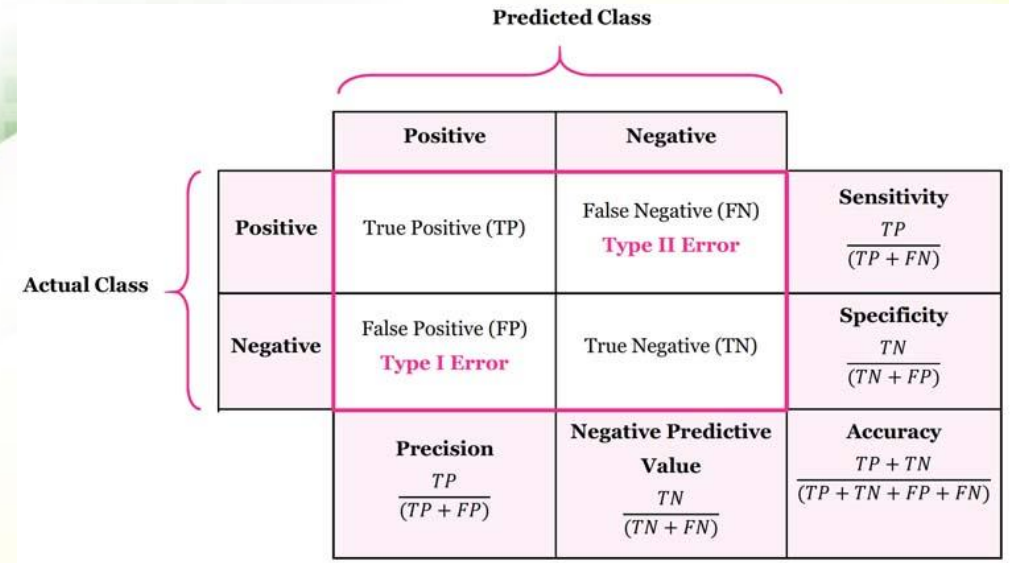
Predictive Modelling - Confusion Matrix

		Predicted Class: Positive	Predicted Class: Negative	
Actual Class	Positive	True Positive (TP)	False Negative (FN) **Type II Error**	**Sensitivity** $\frac{TP}{(TP+FN)}$
	Negative	False Positive (FP) **Type I Error**	True Negative (TN)	**Specificity** $\frac{TN}{(TN+FP)}$
		Precision $\frac{TP}{(TP+FP)}$	**Negative Predictive Value** $\frac{TN}{(TN+FN)}$	**Accuracy** $\frac{TP+TN}{(TP+TN+FP+FN)}$



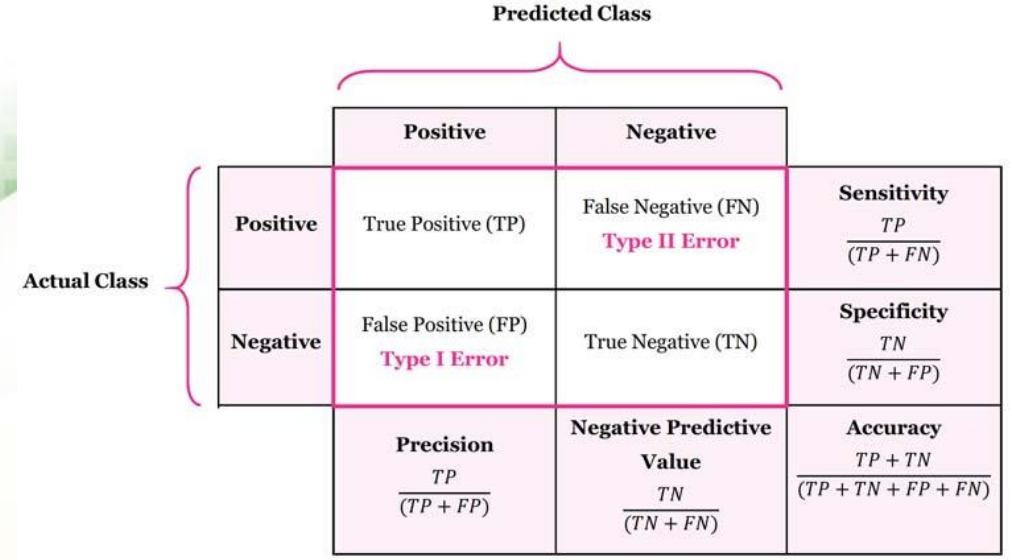
Predictive Modelling - Confusion Matrix

- **Classification accuracy** → total number of correct predictions divided by total number of predictions made for a dataset.
  - Accuracy is not always appropriate for all problems.



- Alternative to using classification accuracy → precision and recall metrics.
  - **Precision** → number of positive class predictions that actually belong to the positive class.
  - **Recall (sensitivity)** → number of positive class predictions made out of all positive examples in the dataset.
  - **F-Measure** → A single score that balances both the concerns of precision and recall.

Predictive Modelling - Confusion Matrix



- **Precision** → number of correct positive predictions made.

  Precision = TruePositives / (TruePositives + FalsePositives)

- 0.0 for no precision and 1.0 for full or perfect precision.

- A model makes predictions and predicts 120 examples as belonging to the positive class, 90 of which are correct, and 30 of which are incorrect.

  Precision = TruePositives / (TruePositives + FalsePositives) = 90 / (90 + 30) = 90 / 120 = 0.75

- Same dataset, another model predicts 50 examples belonging to the positive class, 45 of which are true positives and five of which are false positives.

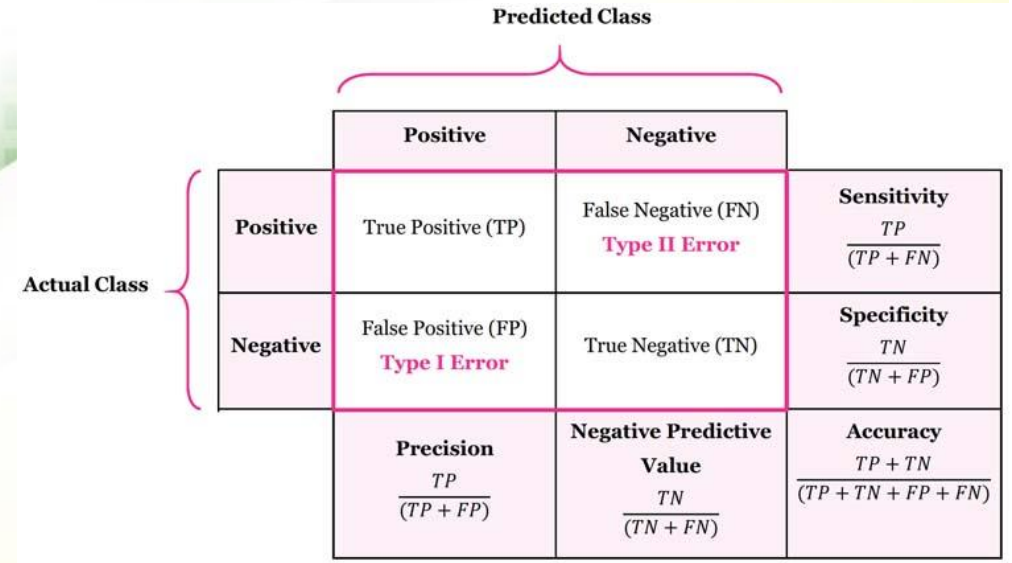
  Precision = TruePositives / (TruePositives + FalsePositives) = 45 / (45 + 5) = 45 / 50 = 0.90

Predictive Modelling - Confusion Matrix

- **Recall** → number of correct positive predictions made out of all positive predictions that could have been made.

  Recall = TruePositives / (TruePositives + FalseNegatives)

- 0.0 for no recall and 1.0 for full or perfect recall.



- A model makes predictions and predicts 90 of the positive class predictions correctly and 10 incorrectly.

  Recall = TruePositives / (TruePositives + FalseNegatives)

  Recall = 90 / (90 + 10) = 90 / 100 = 0.9

Predictive Modelling - Confusion Matrix

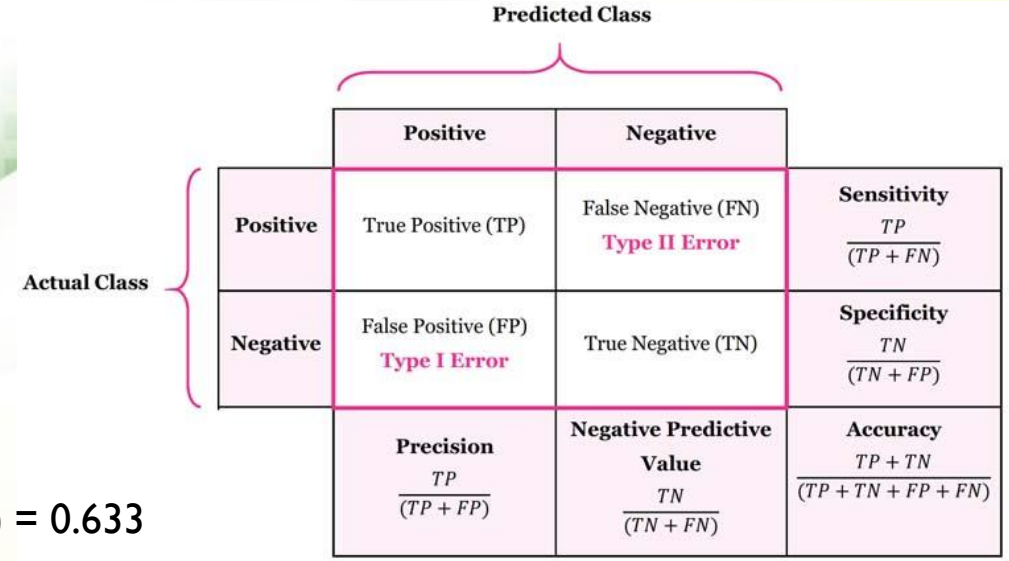
		Predicted Class: Positive	Predicted Class: Negative	
Actual Class	Positive	True Positive (TP)	False Negative (FN) **Type II Error**	**Sensitivity** $\frac{TP}{(TP+FN)}$
	Negative	False Positive (FP) **Type I Error**	True Negative (TN)	**Specificity** $\frac{TN}{(TN+FP)}$
		Precision $\frac{TP}{(TP+FP)}$	**Negative Predictive Value** $\frac{TN}{(TN+FN)}$	**Accuracy** $\frac{TP+TN}{(TP+TN+FP+FN)}$

- Maximizing precision will minimize the number false positives.
- Maximizing the recall will minimize the number of false negatives.
  - For excellent predictions both high precision and high recall needed.
  - Neither precision or recall tells whole story.
    - *Excellent precision with terrible recall, or terrible precision with excellent recall!!!*
  - Increases in recall often come at the expense of decreases in precision, and vice-versa.
- Instead of picking any one measure, a new metric can be used that combines both precision and recall into one score → F/F1-score (harmonic mean of both)

  F-Measure = (2 * Precision * Recall) / (Precision + Recall)
- 0.0 is poor F-Measure score and 1.0 is best or perfect F-Measure score

Predictive Modelling - Confusion Matrix

- Consider a model that predicts 150 examples for the positive class, 95 are correct (true positives), meaning five were missed (false negatives) and 55 are incorrect (false positives).



Precision = TruePositives / (TruePositives + FalsePositives) = 95 / (95 + 55) = 0.633

Recall = TruePositives / (TruePositives + FalseNegatives) = 95 / (95 + 5) = 0.95

Model has poor precision, but excellent recall.

F-Measure = (2 * Precision * Recall) / (Precision + Recall)

= (2 * 0.633 * 0.95) / (0.633 + 0.95) = (2 * 0.601) / 1.583 = 0.759

- Good recall levels-out poor precision, giving an okay or reasonable F-measure score.

Predictive Modelling - Confusion Matrix

Example: Suppose we are trying to create a model that can predict the result for the disease that is either a person has that disease or not. The table is given for the two-class classifier, which has two predictions "Yes" and "NO." **Yes** defines that patient has the disease, and **No** defines that patient does not has that disease.

The classifier has made a total of 100 predictions. Out of 100 predictions, 89 are true predictions, and 11 are incorrect predictions.

The model has given prediction "yes" for 32 times, and "No" for 68 times. Whereas the actual "Yes" was 27, and actual "No" was 73 times.

In 3 instances, an actual patient was wrongly diagnosed.

Prepare the Confusion matrix and calculate all the performance parameters.

Predictive Modelling - Confusion Matrix

Example: Suppose we are trying to create a model that can predict the result for the disease that is either a person has that disease or not. The table is given for the two-class classifier, which has two predictions "Yes" and "NO." **Yes** defines that patient has the disease, and **No** defines that patient does not has that disease.

The classifier has made a total of 100 predictions. Out of 100 predictions, 89 are true predictions, and 11 are incorrect predictions.

The model has given prediction "yes" for 32 times, and "No" for 68 times. Whereas the actual "Yes" was 27, and actual "No" was 73 times. In 3 instances, an actual patient was wrongly diagnosed.

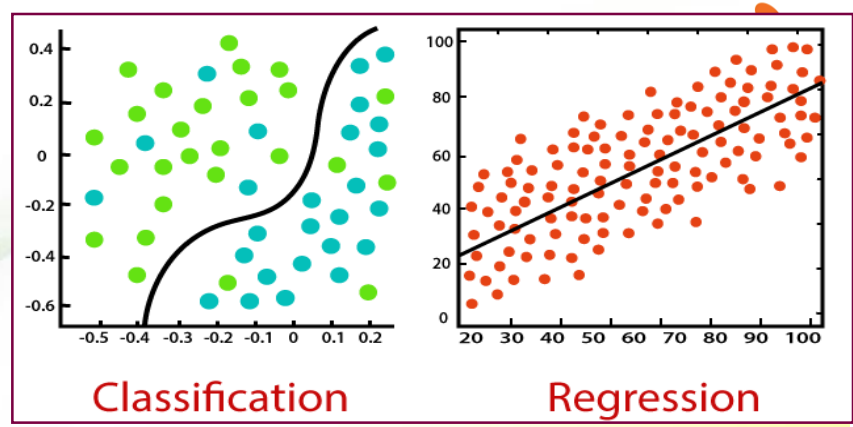
Prepare the Confusion matrix and calculate all the major performance parameters.

n = 100	Actual: No	Actual: Yes	
Predicted: No	TN: 65	FP: 3	68
Predicted: Yes	FN: 8	TP: 24	32
	73	27	

Predictive Modeling

- **Supervised learning**: providing labeled data to machine learning model.
  - labeled dataset is usually data gathered from experience/historical data.
- **Unsupervised learning**: working with unlabeled data.
  - labels in these use cases are often difficult to obtain (not enough knowledge/historical info or labeling is too expensive).
  - lack of labels makes it difficult to set goals for trained model → complicated to measure whether results are accurate.
- **Semi-supervised learning:** Datasets are split into two parts: labeled and unlabeled part.
- **Reinforcement learning:** System learns exclusively from a series of reinforcements.
  - Reinforcements can be positive or negative in relation to a system goal.
  - Positive ones are known as "rewards"; and negative ones as "punishments".

Predictive Modeling



Classification: Process of finding a model/function which helps in separating data into multiple categorical classes.

- Data is categorized under different labels according to some parameters given in input and then labels are predicted for the data.

Regression: Process of finding a model for distinguishing data into continuous real values instead of using classes.

- Helps to identify distribution movement depending on historical data.

- Regression algorithms are used to **predict continuous** values such as price, salary, age, etc.
- Classification algorithms are used to **predict/Classify discrete values** (True or False, Spam or Not Spam, etc.)

Clustering: Process data to find a structure/pattern/cluster in a collection of uncategorized data.

- Hierarchical clustering, K-means clustering

Association: Discovering exciting relationships between variables in large databases.

- Example: Shopping-cart recommendation, video streaming recommendation.

Predictive Modeling

Classification Algorithms:

- Logistic Regression
- K-Nearest Neighbours
- Support Vector Machines
- Naïve Bayes
- Decision Tree Classification
- Random Forest Classification

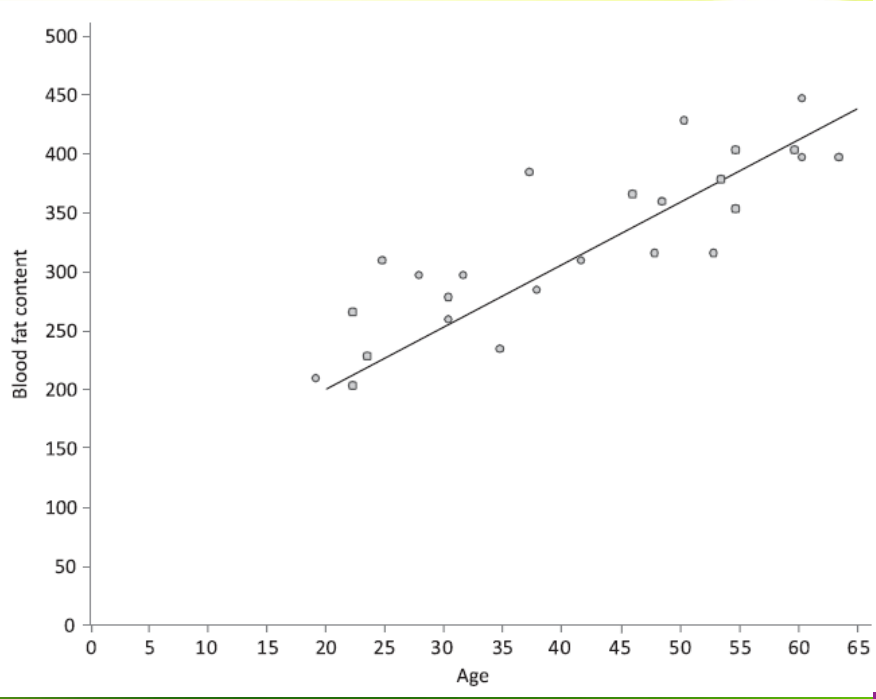
Regression Algorithms:

- Simple Linear Regression
- Multiple Linear Regression
- Polynomial Regression
- Support Vector Regression
- Decision Tree Regression
- Random Forest Regression

Regression Models

Regression: build a function/model that describe relationship between one/more independent variables and a single response variable.

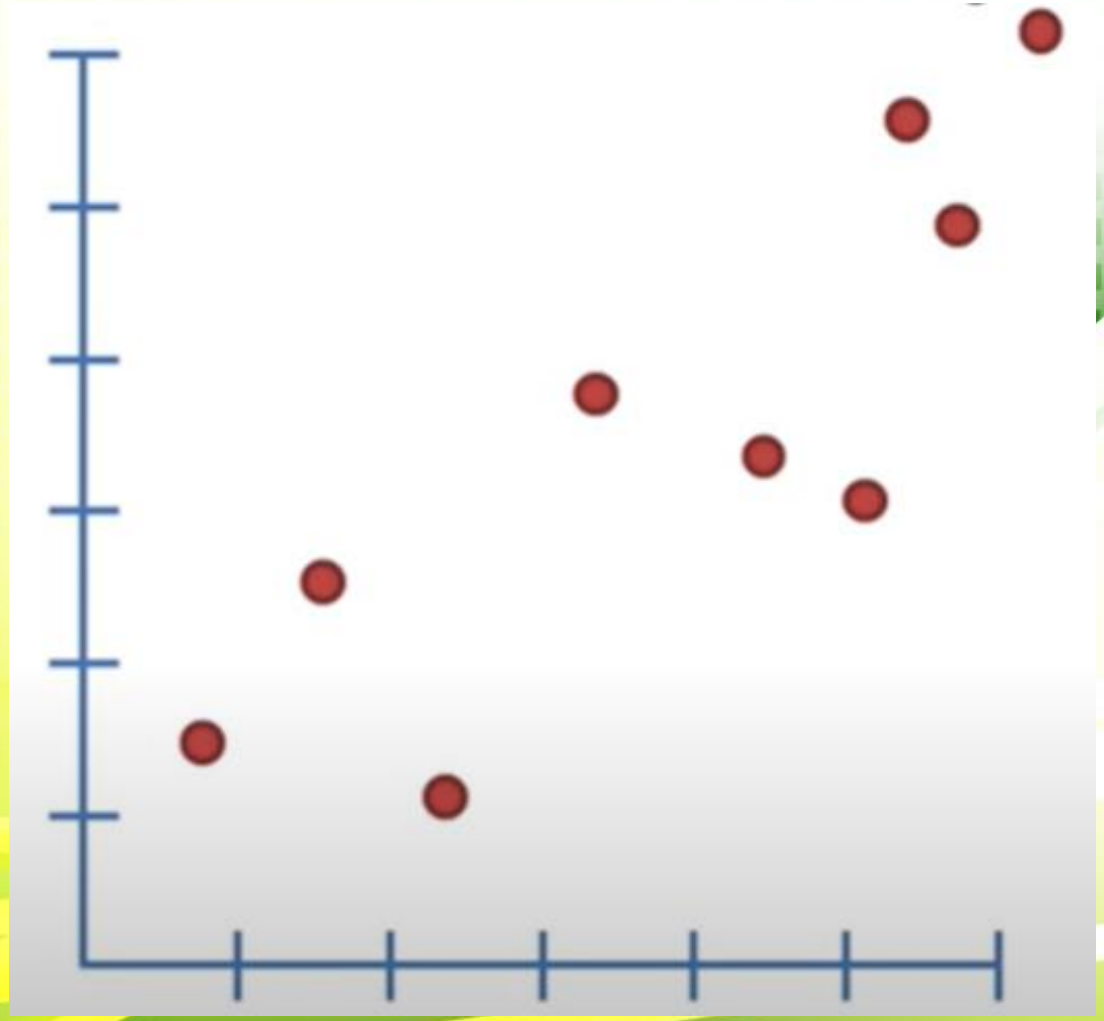
- Plot a graph between the variables which best fit the given data points.
- *Regression shows a line or curve that passes through all data points on a target-predictor graph in such a way that distance between data points and regression line is minimum.*

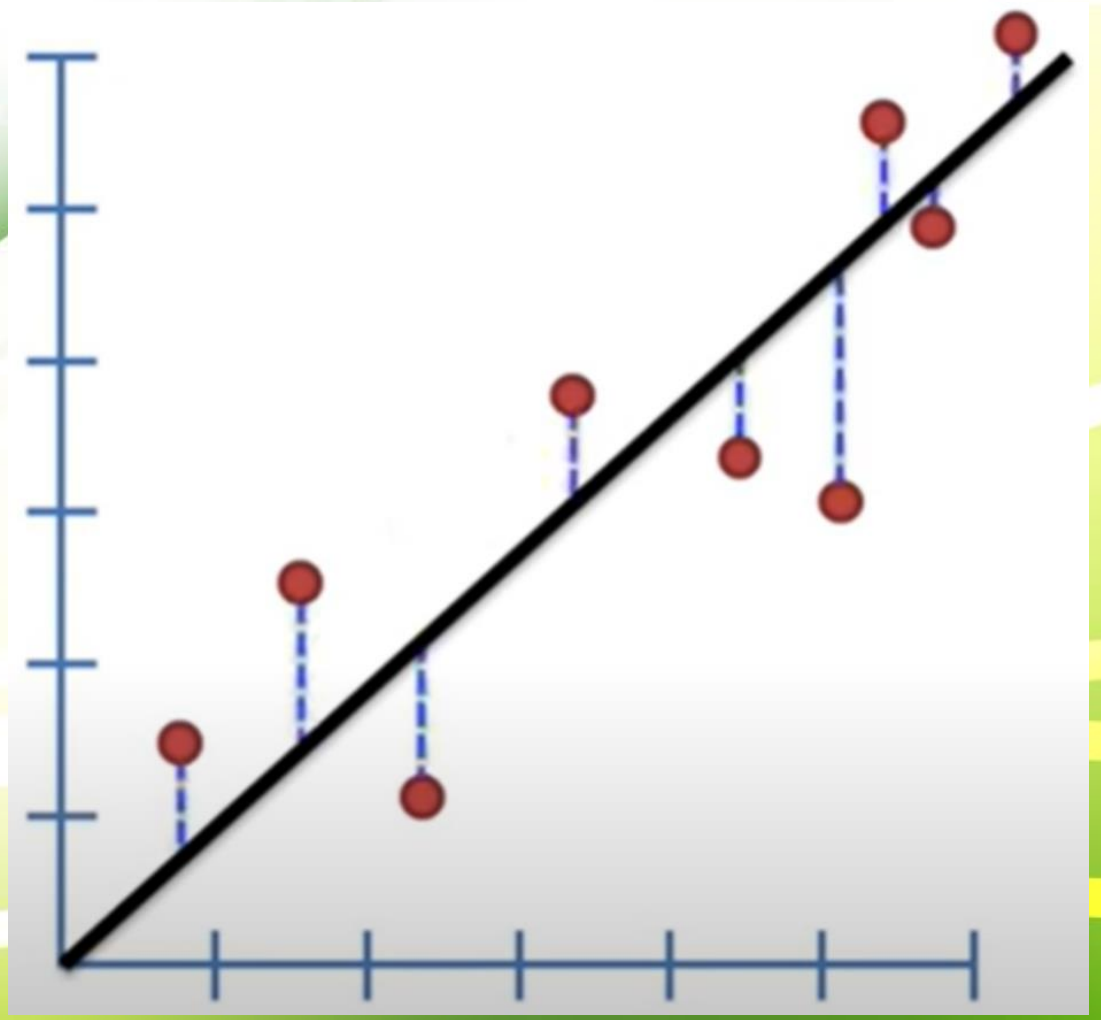


Types of Regression models

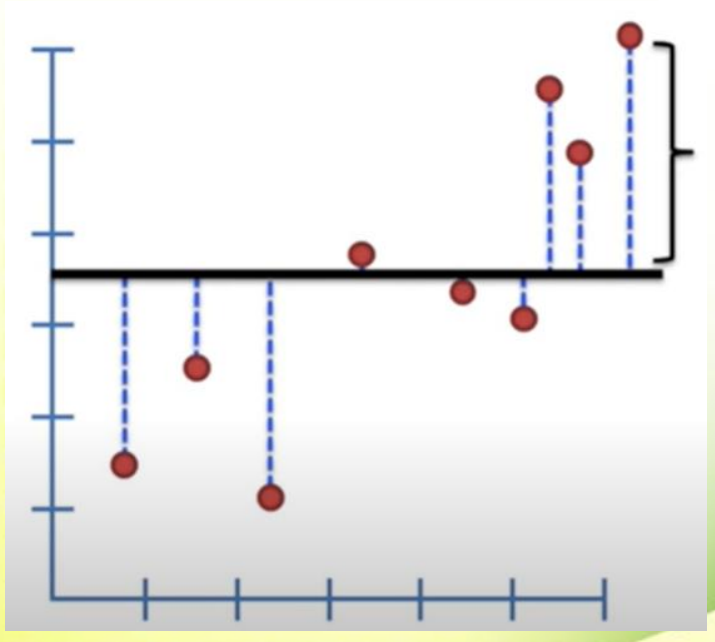
- Linear Regression
- Polynomial Regre6ssion
- Logistic Regression

Linear Regression

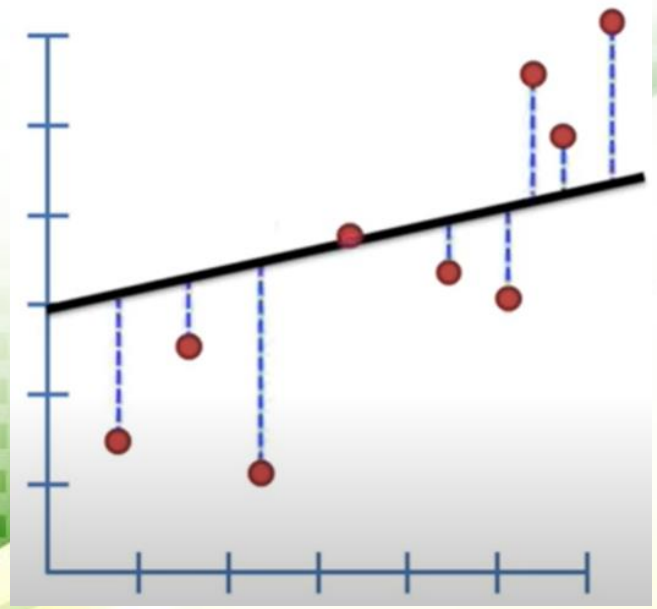




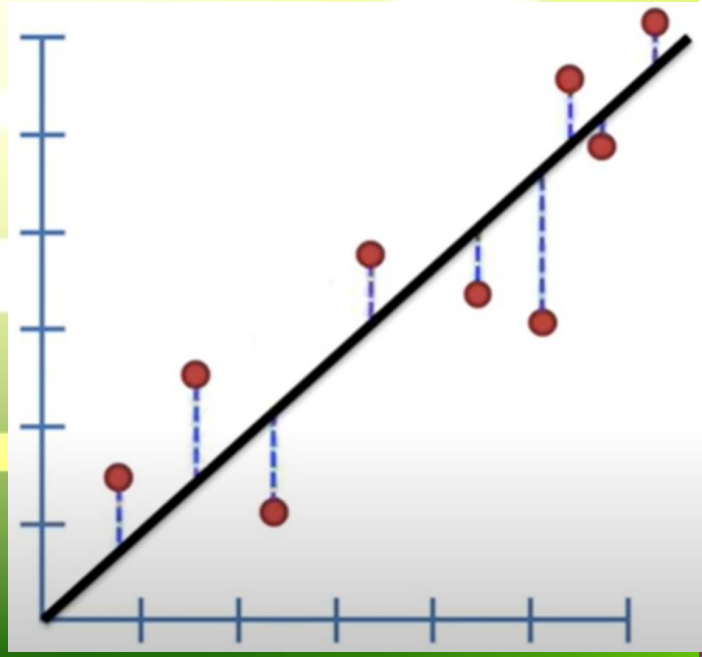
Linear Regression



- Draw a random line through the points.
- Calculate distance between data point & corresponding point on line → Residual.
- Square of least distance.
- Sum of square.



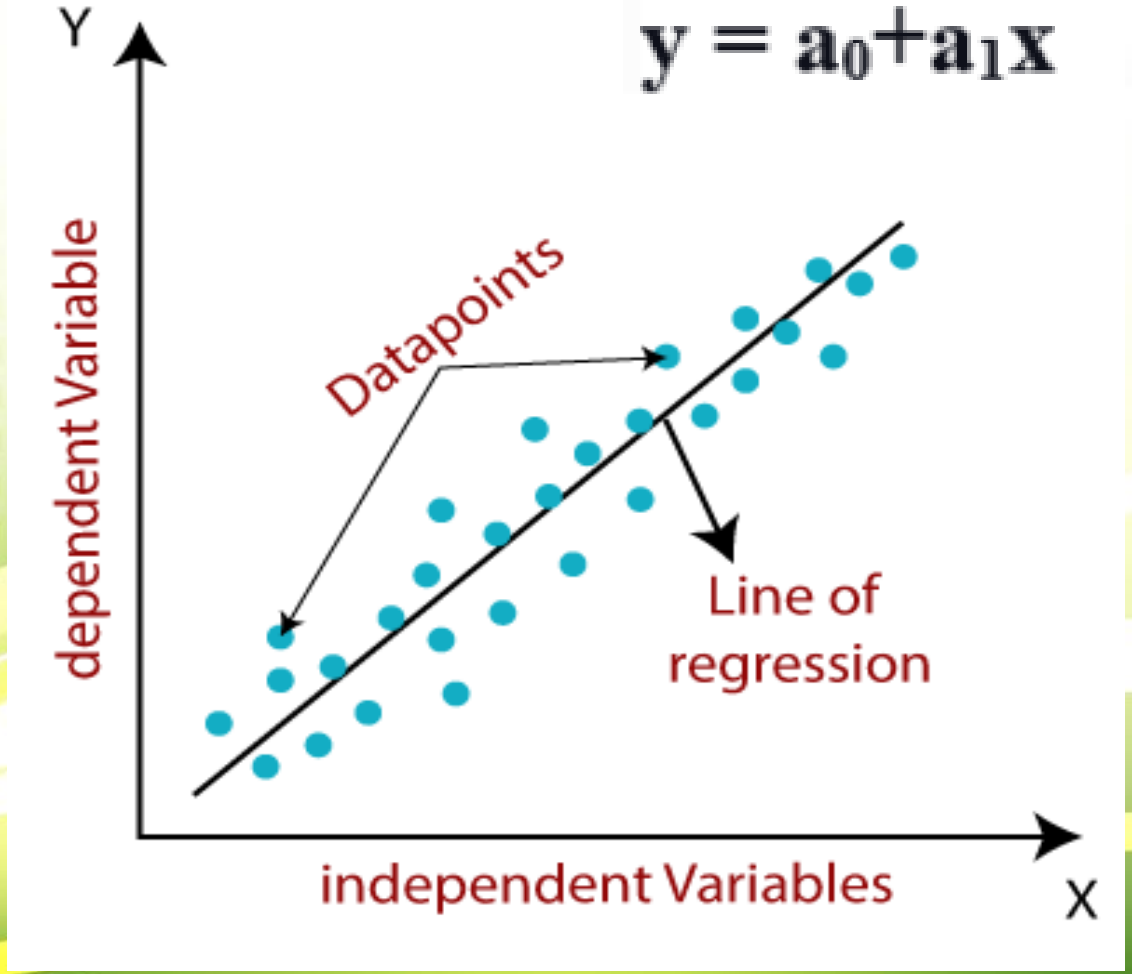
- Rotate the line aiming to reduce error.
- Find each Distance → square → add.



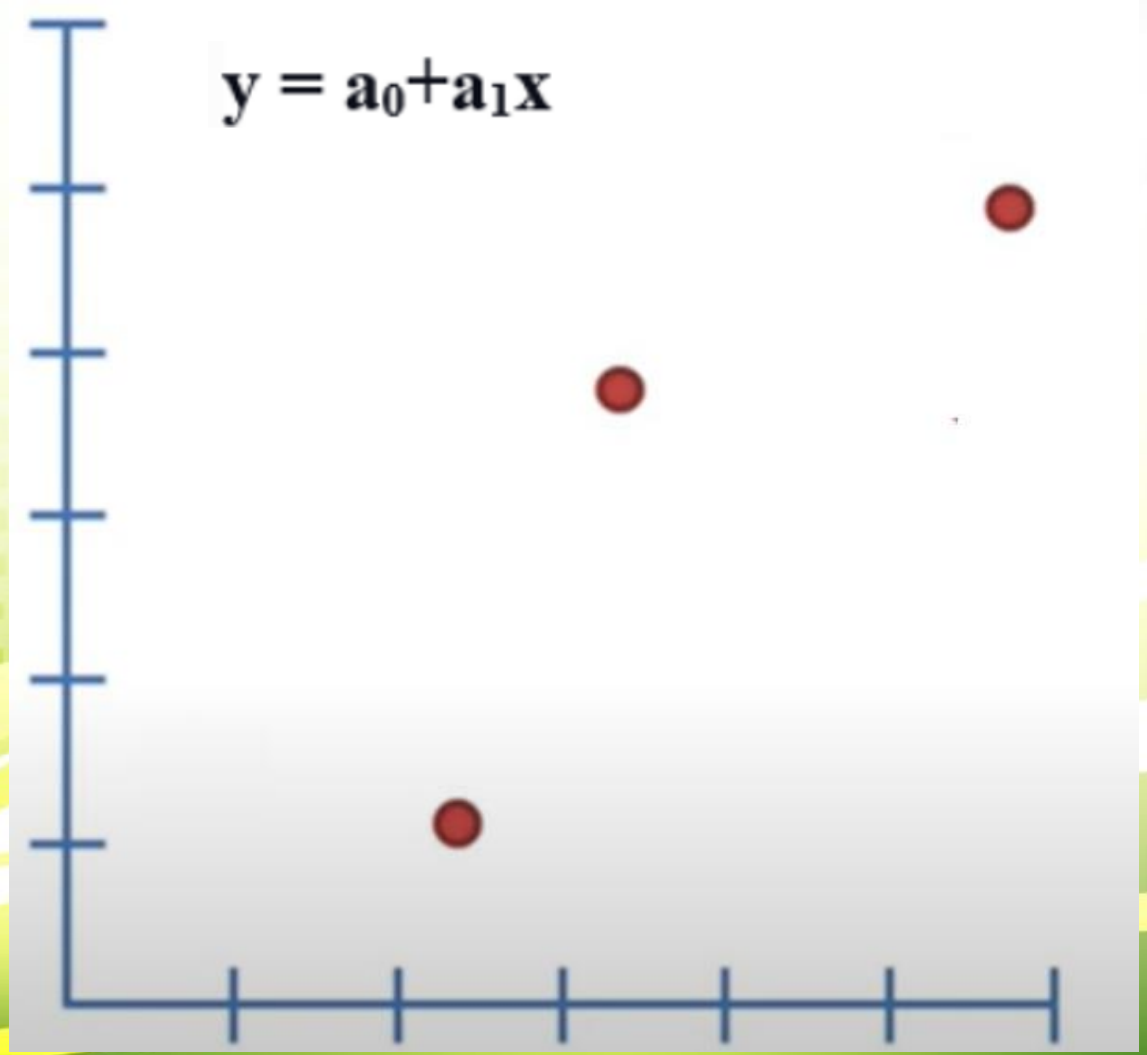
- Regression line
- Lease square distance.

Linear Regression

$$y = a_0 + a_1 x$$

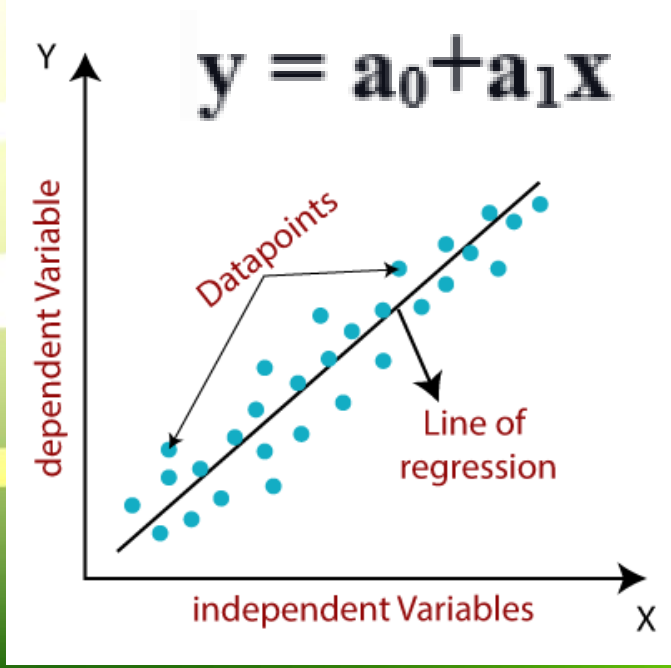


Linear Regression



Linear Regression

- Shows a linear relationship between a dependent (y) and one or more independent (x) variables.
- Finds how value of dependent variable is changing according to value of independent variable.
- Makes predictions for continuous/real or numeric variables; **sales, salary, age, product price,** etc.
- Provides a sloped straight line representing relationship between variables.
- Supervised learning algorithm.
- Goal of linear regression algorithm is to get best values for a_0 and a_1 to find best fit line.
- Best fit line should have least error → error between predicted values and actual values should be minimized.



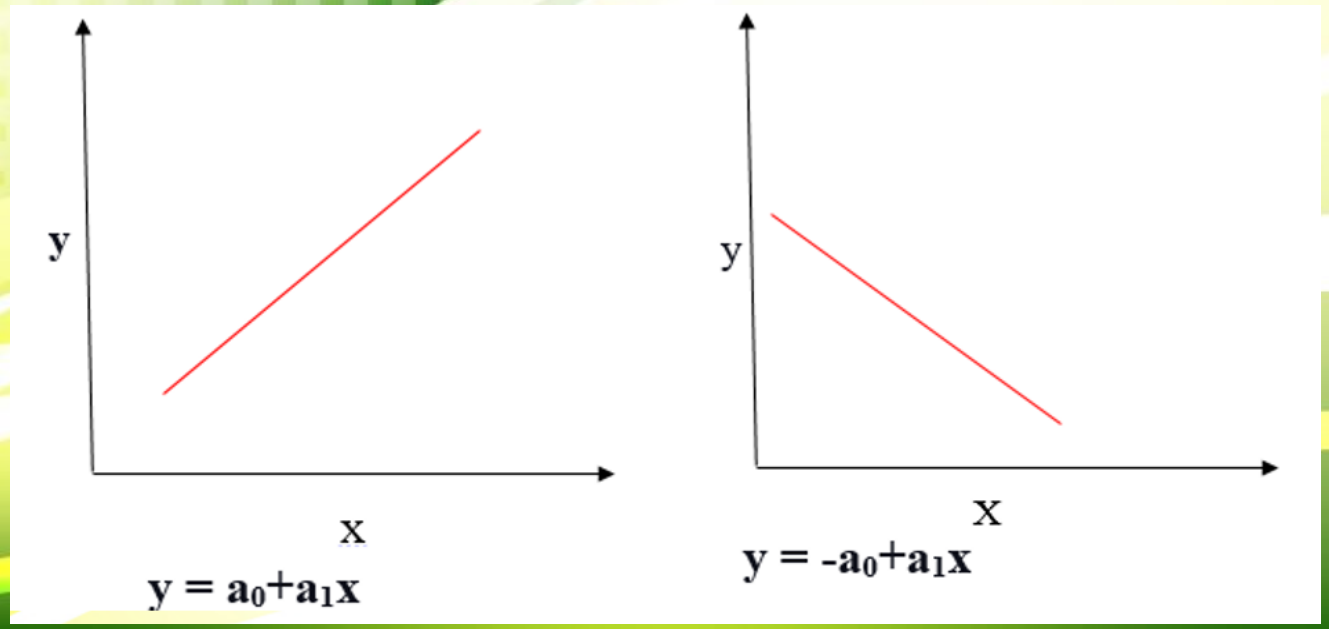
Linear Regression

Positive Linear Relationship

- If dependent variable expands on Y-axis given independent variable progress on X-axis.

Negative Linear Relationship

- If dependent variable decreases on Y-axis given independent variable increases on X-axis.



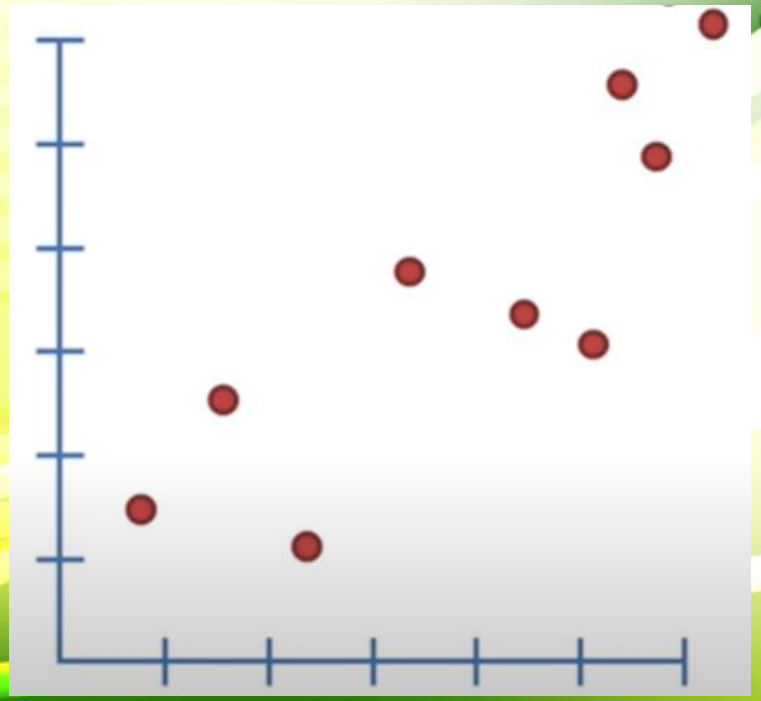
Linear Regression

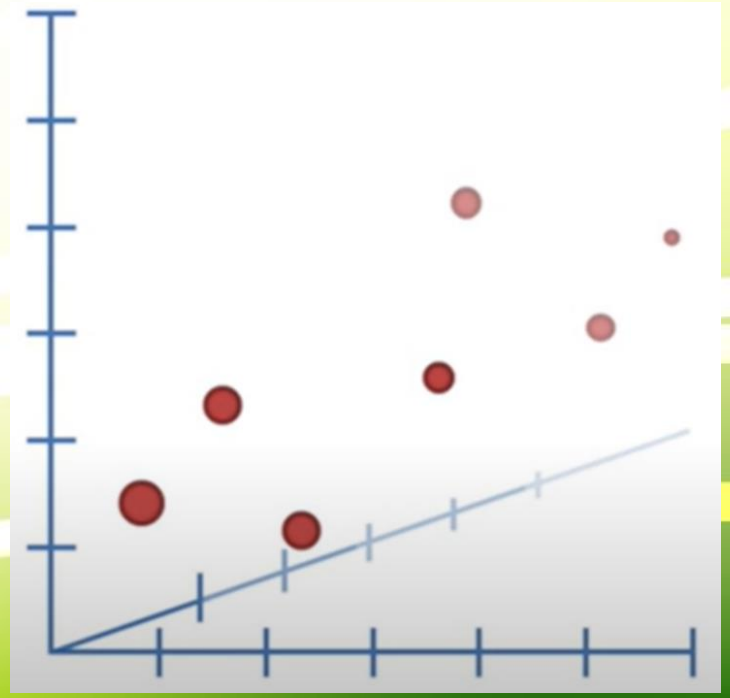
Simple Linear Regression:

- Single independent variable is used to predict value of a numerical dependent variable.

Multiple Linear regression:

- More than one independent variable is used to predict value of a numerical dependent variable.





Linear Regression

- **Simple Linear Regression:**
  - Single independent variable is used to predict value of a numerical dependent variable.
  - $Y = mx + c$
- **Multiple Linear regression:**
  - More than one independent variable is used to predict value of a numerical dependent variable.
  - $Y = a_1x_1 + a_2x_2 + a_3x_3 + a_4x_4 + c$

$$\hat{y} = \hat{b}_0 + \hat{b}_1x_1 + \hat{b}_2x_2 + \cdots + \hat{b}_px_k$$

- a_1, a_2, a_3, a_4 - Regression Coefficient
- capped → estimated coefficient

- $Y = 9 + 3x_1 + 0.9x_2 + 4x_3 + 1.4x_4$
- Regression Coefficient tells which feature(s) have more impact on dependent variable.

Linear Regression

- Size of coefficient for each independent variable tells the impact size of that variable on dependent variable.
- Sign on coefficient (+ or -) tells the direction of effect.

Response

Independent variables

Name	MPG (y)	Cyclinders (x1)	Displacement (x2)	Horsepower (x3)	Weight (x4)	Acceleration (x5)
Chevrolet Chevelle Malibu	18	8	307	130	3504	12
Buick Skylark 320	15	8	350	165	3693	11.5
Plymouth Satellit	18	8	318	150	3436	11
AMC Rebel SST	16	8	304	150	3433	12
Ford Torino	17	8	302	140	3449	10.5

$$\hat{y} = \hat{b}_0 + \hat{b}_1 x_1 + \hat{b}_2 x_2 + \cdots + \hat{b}_p x_k$$

$$y = f(x_i)$$

Model

Linear Regression

Simple Linear Regression:

- b_1 is slope
- b_0 is intercept.

$$y = b_0 + b_1 x$$

$$b_1 = \frac{\sum_{i=1}^{n} (x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{n} (x_i - \bar{x})^2}$$

$$b_0 = \bar{y} - b_1 \bar{x}$$

Linear Regression

Simple Linear Regression:

mean of x = 39.12
mean of y = 310.72

$$y = b_0 + b_1 x$$

$$b_0 = \bar{y} - b_1 \bar{x}$$

$$b_1 = \frac{\sum_{i=1}^{n} (x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{n} (x_i - \bar{x})^2}$$

$Slope\,(b_1) = 19{,}157.84/3{,}600.64$

$Slope\,(b_1) = 5.32$

$Intercept\,(b_0) = 310.72 - (5.32 \times 39.12)$

$Intercept\,(b_0) = 102.6$

Blood fat content* = 102.6 + 5.32 × *Age

X	Y	$(x_i - \bar{x})$	$(y_i - \bar{y})$	$(x_i - \bar{x})(y_i - \bar{y})$	$(x_i - \bar{x})^2$
46	354	6.88	43.28	297.7664	47.3344
20	190	−19.12	−120.72	2,308.1664	365.5744
52	405	12.88	94.28	1,214.3264	165.8944
30	263	−9.12	−47.72	435.2064	83.1744
57	451	17.88	140.28	2,508.2064	319.6944
25	302	−14.12	−8.72	123.1264	199.3744
28	288	−11.12	−22.72	252.6464	123.6544
36	385	−3.12	74.28	−231.7536	9.7344
57	402	17.88	91.28	1,632.0864	319.6944
44	365	4.88	54.28	264.8864	23.8144
24	209	−15.12	−101.72	1,538.0064	228.6144
31	290	−8.12	−20.72	168.2464	65.9344
52	346	12.88	35.28	454.4064	165.8944
23	254	−16.12	−56.72	914.3264	259.8544
60	395	20.88	84.28	1,759.7664	435.9744
48	434	8.88	123.28	1,094.7264	78.8544
34	220	−5.12	−90.72	464.4864	26.2144
51	374	11.88	63.28	751.7664	141.1344
50	308	10.88	−2.72	−29.5936	118.3744
34	220	−5.12	−90.72	464.4864	26.2144
46	311	6.88	0.28	1.9264	47.3344
23	181	−16.12	−129.72	2,091.0864	259.8544
37	274	−2.12	−36.72	77.8464	4.4944
40	303	0.88	−7.72	−6.7936	0.7744
30	244	−9.12	−66.72	608.4864	83.1744
			Sum	19,157.84	3,600.64

Linear Regression

Example:

Build the simple linear regression model/function for the data given below.

Age (x)	Sugar Level (Y)
46	354
20	190
52	405
30	263
57	451

$$y = b_0 + b_1 x$$

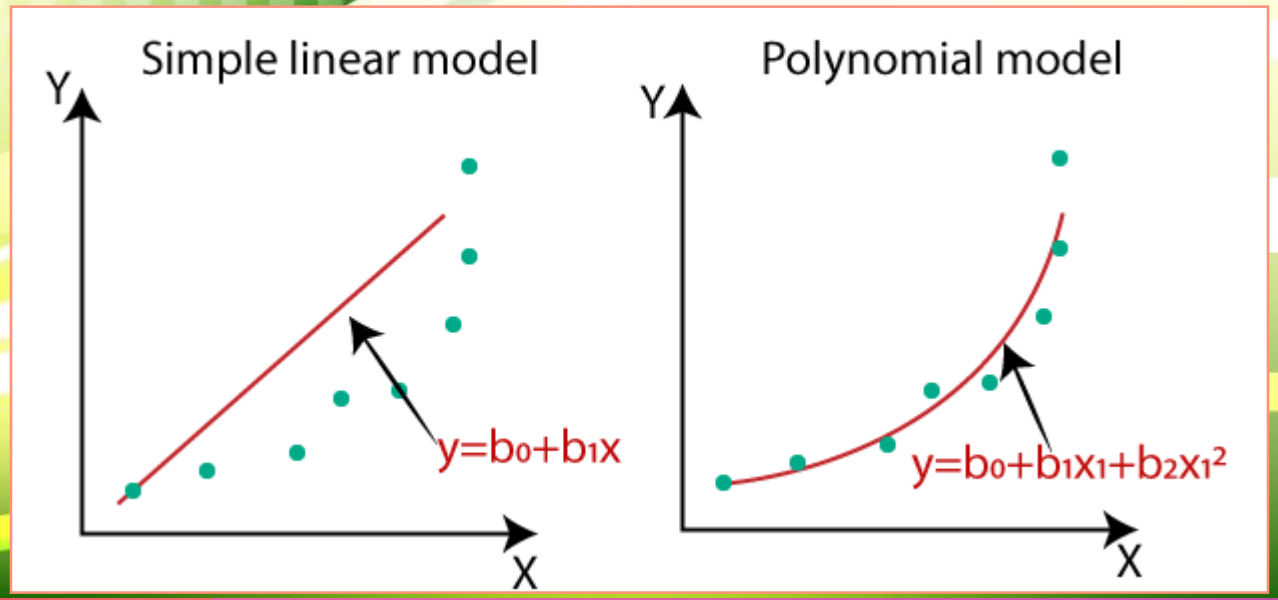
$$b_0 = \bar{y} - b_1 \bar{x}$$

$$b_1 = \frac{\sum_{i=1}^{n} (x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{n} (x_i - \bar{x})^2}$$

Using this model predict the Sugar level for a new patient of age 39.

Polynomial Regression

- If data points clearly will not fit linear regression (straight line through all data points), it might be ideal for polynomial regression.
- (~~Linear~~)Relationship between variables x and y to find best way to draw a line through data points.
- Special case of Multiple Linear Regression, by adding some polynomial terms to it.
- Relationship between a dependent(y) and independent variable(x) as n^{th} degree polynomial.



Model Assessment

- ***Residual***: error term representing difference between observed value (*y*) and predicted value.

$$\hat{e} = y - \hat{y}$$

- Residual analysis helps to better understand how well model is performing.
- ***Sum of squares total (SST)*** : measure of variation of *y*-values about their mean.
- ***Sum of squares due to regression (SSR):*** differences between predicted/regression values and average *y*-value.
- ***Sum of squares of error (SSE):*** differences between actual *y*-values and predicted *y*-values.

$$SST = \sum_{i=1}^{n} (y_i - \bar{y})^2$$

$$SSR = \sum_{i=1}^{n} (\hat{y}_i - \bar{y})^2$$

$$SSE = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2$$

$$SST = SSR + SSE$$

Model Assessment

- ***Coefficient of determination* (R^2)**: proportion of variation.

$$SST = \sum_{i=1}^{n} (y_i - \bar{y})^2 \qquad SSR = \sum_{i=1}^{n} (\hat{y}_i - \bar{y})^2 \qquad R^2 = \frac{SSR}{SST}$$

- R^2 values vary between 0 and 1.
- R^2 closer to 1 → more accurate model predictions (models have a *closer fit).*
- In multiple linear regression, *adjusted* R^2 value (R^2 adj) is usually considered to better account for the multiple independent variables used in analysis as well as sample size.

$$R^2_{\text{adj}} = 1 - \left[\frac{(1 - R^2)(n - 1)}{n - k - 1}\right]$$

n : number of observations
k : number of independent variables.

Model Assessment

- *standard error of the estimate* ($S_{y,\,x}$) : measure of variation of *y*-values about regression line.
  - interpreted in a similar manner to standard deviation.
  - indicates model's accuracy: larger the value for standard error of estimate, lower the precision.

$$s_{y.x} = \sqrt{\frac{\text{SSE}}{n-2}}$$

- t-Test, F-Test performed to assess variable dependencies and model performance.
- **Mean Squared Error (MSE):** most common metric for regression models.
- **Mean Absolute Error (MAE):** simple metric; Not preferred where outliers are prominent.
- **Root Mean Squared Error (RMSE):** square root MSE.
  - RMSE penalizes large errors..

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

$$MAE = \frac{1}{n}\sum_{i=1}^{n}|\hat{y}_i - y_i|$$

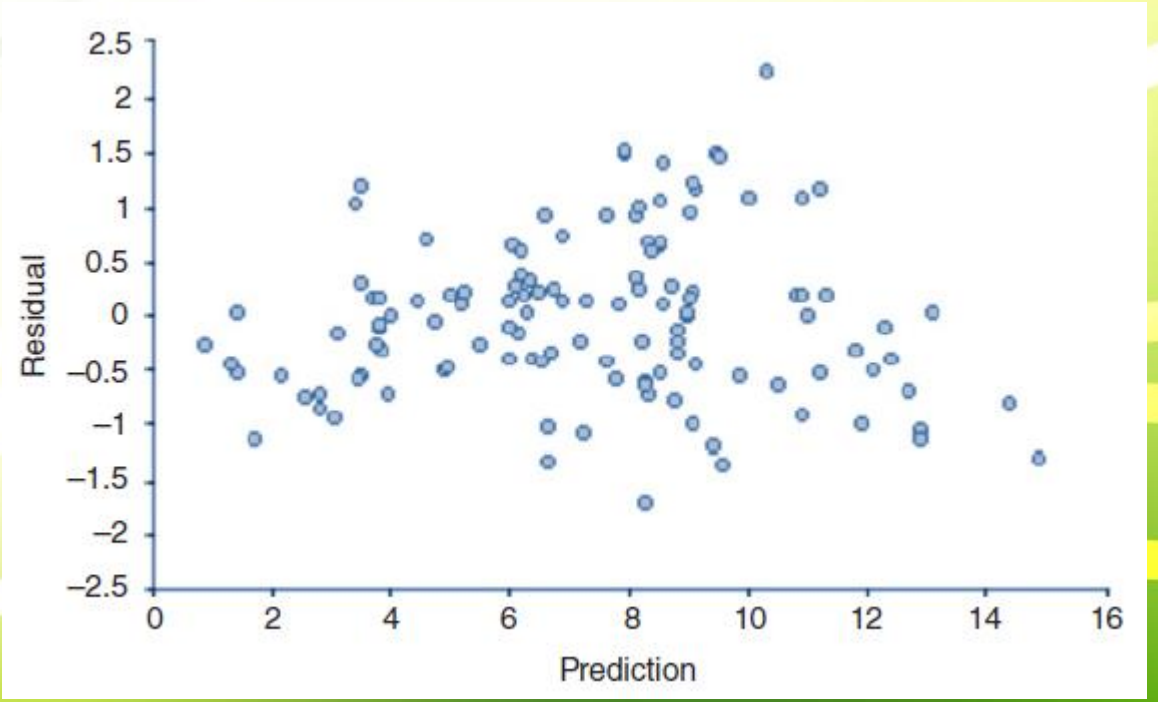
$$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(\hat{y}_i - y_i)^2}$$

Model Assessment

- Linear regression models are based on a series of assumptions.
- If a data set does not conform to these assumptions then either the model needs to be adjusted (by applying mathematical **transformation to data**) or particular linear regression not be suitable for modeling that data set.

Assumption - 1:

- ***linearity***: relationship between independent and response variable should be linear.
- A scatterplot of actual response values plotted against predicted values should evenly distribute on both sides of regression line.



Model Assessment

Assumption - 2:

- ***normality of error distribution***: error about the line of regression should be approximately normally distributed for each value of *x*.
  - frequency histogram, statistical measures of skewness/kurtosis, or a normal probability plot.
  - If not a normal distribution, then variable transformations expected.

Assumption - 3:

- ***homoscedasticity of errors***. variation of error/residual across each independent variables should remain constant (as a function of predicted value).

Assumption - 4:

- ***independence of errors***. no trend in residuals based on order in which observations were collected.

Model Assessment

- important to generate simplest possible model that contains only necessary independent variables.
- Ideally number of independent variables should be small and include at least **10 observations** in training set *for every independent variable* included in model. Example:
  - Dataset has 25 independent variables with 300 records.
  - Hypothesis Tests shows 13 important variables.
  - Training set should have minimum 130 records.
- Important to *perform exploratory data analysis to* inspect relationships between variables.
- Perform *transformations* on potential independent variables.
- *Dummy, derived, or composite variables can be generated.*
- Continuous variables may need to be *transformed into a categorical variable.*
- If relationship between a potential independent and response variable needs to be converted from nonlinear to linear, suitable transformations can be used for same.
- Multiple combinations of different independent variables can be used to build set of models from which best performing, most plausible, and simplest model is selected.
- *Standard error*, *t-stat*, and *p-value are* calculated, which can be used to help in selection of independent variables.

Logistic Regression

- Supervised classification algorithm.
- In linear regression problem, target variable(output) y can take only continuous values for a given set of features(inputs) X.
- Logistic regression is popular approach to building models where response variable is categorical.
- Just like Linear regression, it assumes that the data follows a linear function.
- Logistic Regression in its base form is a *Binary Classifier*.
  - Target vector may only take the form of one of two values.
  - Model builds a regression model to predict probability that a given data entry belongs to the category numbered as '1'.
  - A Linear Model, $\beta_0 + \beta_1 x$, is integrated into a Logistic Function (Sigmoid Function).

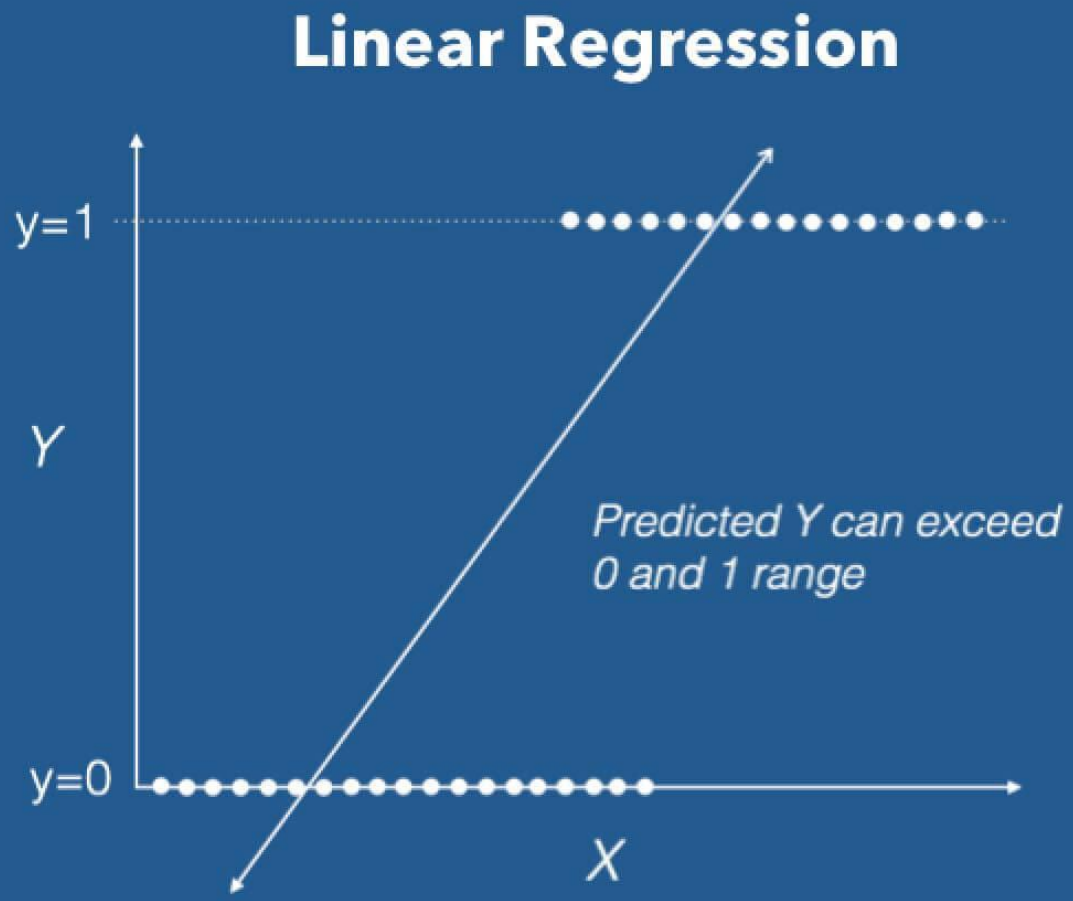
Sugar Level (X)	Diabetes (Y)
354	1
190	0
405	1
263	0
451	1

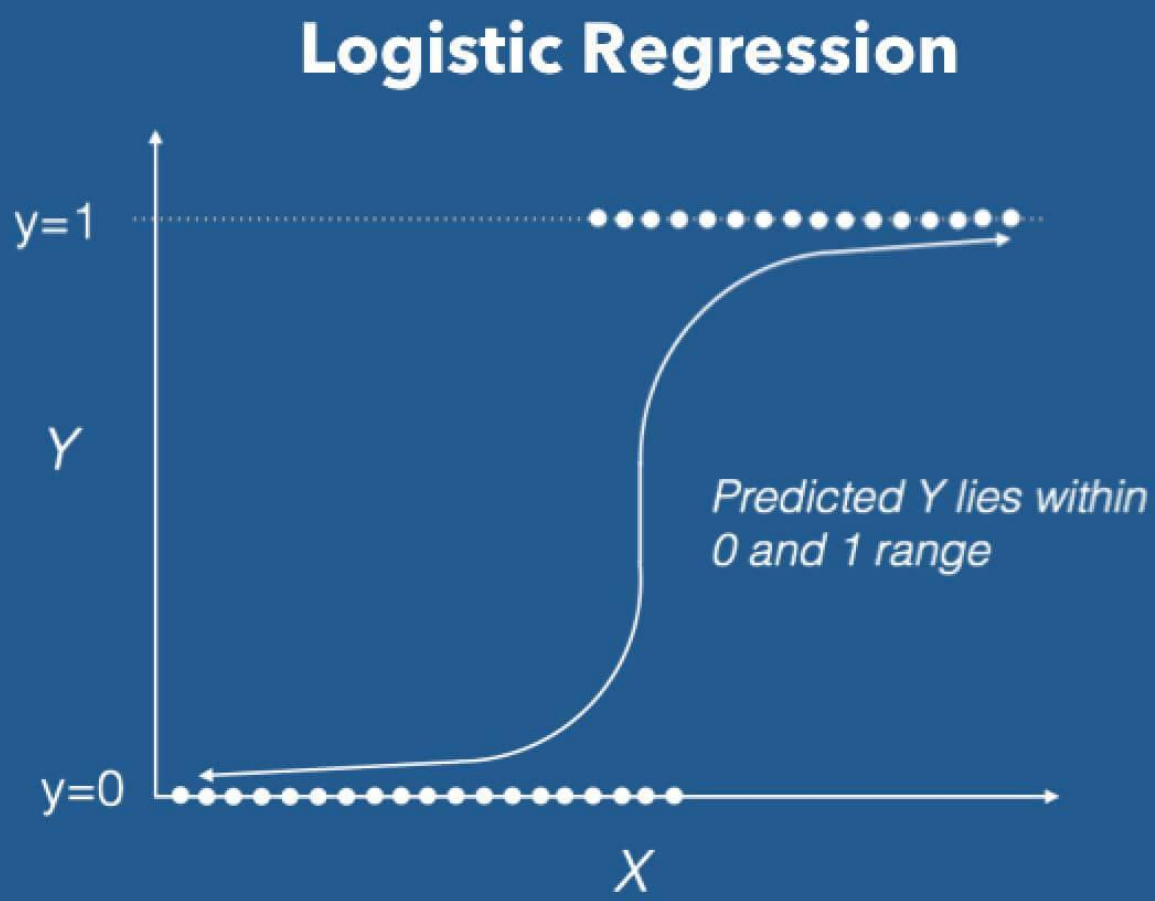
Logistic Regression

Based on the number of categories, Logistic regression can be classified as:

- **binomial:** target variable can have only 2 possible types: “0” or “1”.
  - “win” vs “loss”, “pass” vs “fail”, “dead” vs “alive”, etc.
- **multinomial:** target variable can have 3 or more possible types which are not ordered.
  - “disease A” vs “disease B” vs “disease C”.
- **ordinal:** it deals with target variables with ordered categories.
  - A test score can be categorized: “very poor”, “poor”, “good”, “very good”.

Logistic Regression

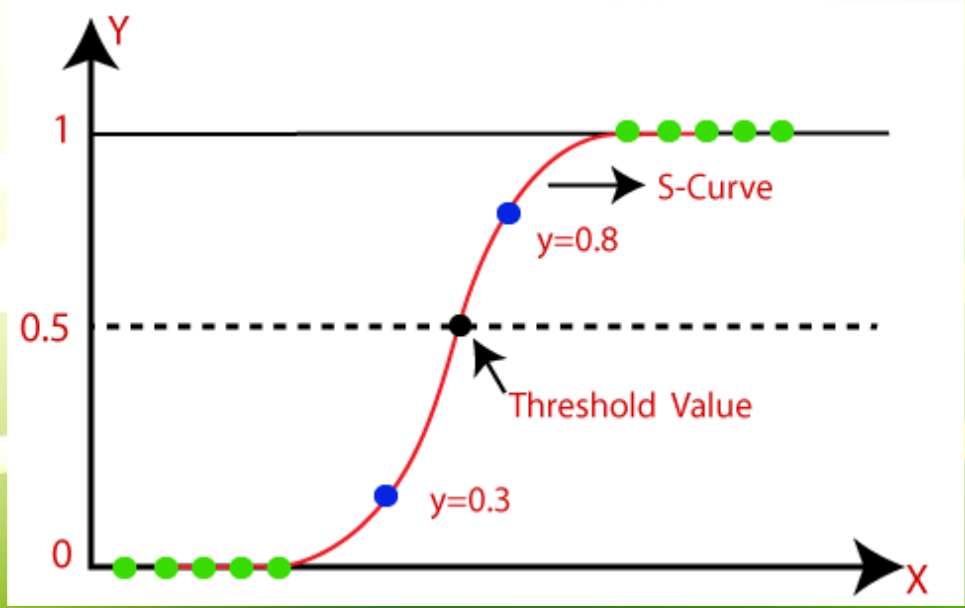




Logistic Regression

Logistic Function (Sigmoid Function):

- A mathematical function used to map the predicted values to probabilities.
- Maps any real value into another value within a range of 0 and 1.
- Value of logistic regression must be between 0 and 1, which cannot go beyond this limit, so it forms a curve like the "S" form.
- S-form curve is called Sigmoid function or logistic function.
- A threshold value defines probability of either 0 or 1.
    - values above threshold value tends to 1,
    - value below the threshold values tends to 0.



Logistic Regression

- standard linear regression formula would compute values outside 0-1 range (not useful)
- Logistic function ensures prediction in 0-1 range

Linear regression/line function $\hat{y} = \hat{b}_0 + \hat{b}_1 x_1 + \hat{b}_2 x_2 + \cdots + \hat{b}_p x_k$

logistic function for response = 1 $P(y = 1) = \frac{1}{1 + e^{(\beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_k x_k)}}$

b_0 is a constant, and b_1 b_k are coefficients to k independent variables (x_1 x_k).

Logistic Regression

User ID	Gender	Age	EstimatedSalary	Purchased
15624510	Male	19	19000	0
15810944	Male	35	20000	0
15668575	Female	26	43000	0
15603246	Female	27	57000	0
15804002	Male	19	76000	0
15728773	Male	27	58000	0
15598044	Female	27	84000	0
15694829	Female	32	150000	1
15600575	Male	25	33000	0
15727311	Female	35	65000	0
15570769	Female	26	80000	0
15606274	Female	26	52000	0
15746139	Male	20	86000	0
15704987	Male	32	18000	0
15628972	Male	18	82000	0
15697686	Male	29	80000	0
15733883	Male	47	25000	1
15617482	Male	45	26000	1
15704583	Male	46	28000	1
15621083	Female	48	29000	1
15649487	Male	45	22000	1
15736760	Female	47	49000	1