LOGISTIC REGRESSION

WWW.TANTROSTAV.CO.IN

- Regression refers to a statistical method used to model the relationship between a **dependent variable** and **one or more independent variables**.

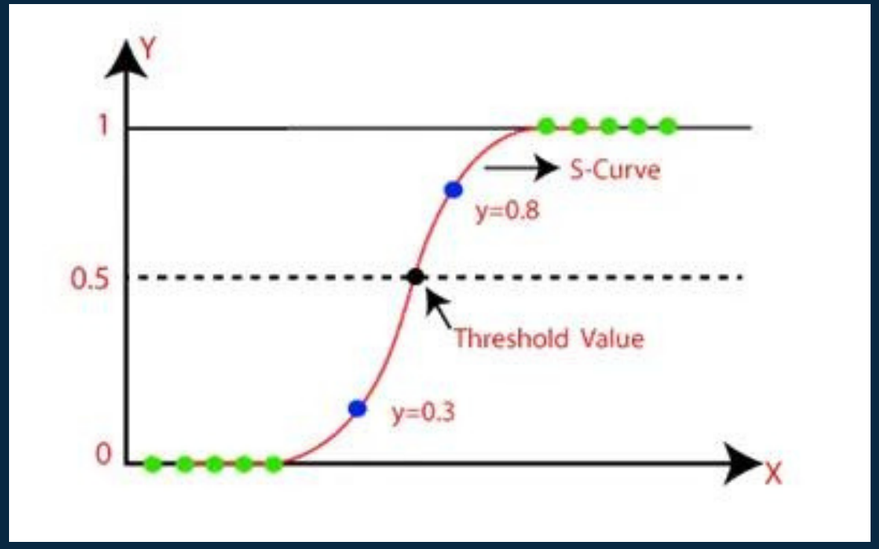
- It involves finding the line or curve that **best fits the data points** to predict the value of the dependent variable based on the values of the independent variable(s).

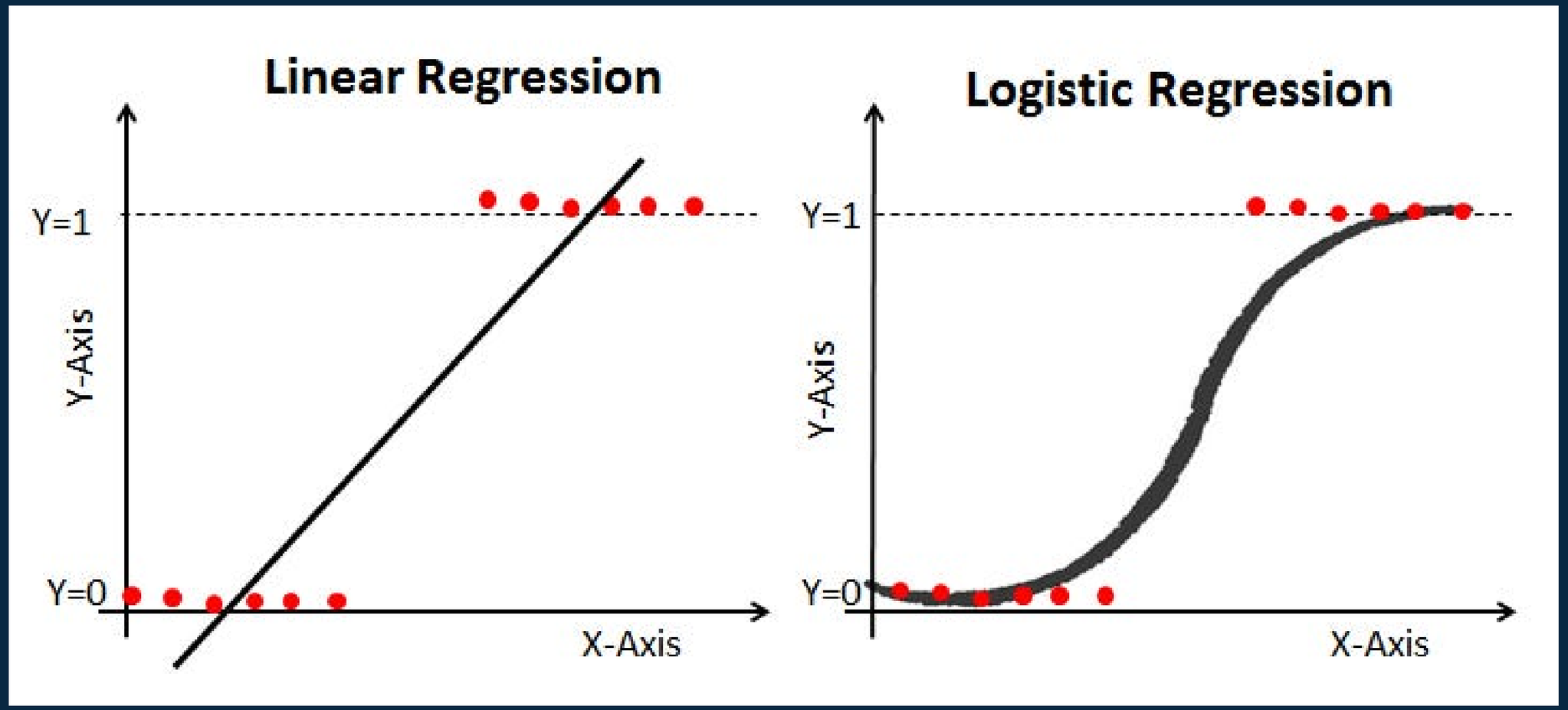
- What if our task is not just finding the best fit? What if we needed to see the possibility of the occurence of a particular thing?

- Identifying if a person is prone to a disease based on age,gender, and other health variables
- Predicting if a loan is going to be approved or not based on previous records
- Classifying if an emain is spam or not.
- Predicting if s student will fail or not.

- Logistic regression uses a logistic function to map the input variables onto the probability of the binary outcome, which is bounded between 0 and 1.

- Logistic regression predicts the output of a categorical dependent variable. Therefore the outcome must be a categorical or discrete value. It can be either Yes or No, 0 or 1, true or False, etc. but instead of giving the exact value as 0 and 1, it gives the probabilistic values which lie between 0 and 1.




Linear Regression
Y=1
Y-Axis
Y=0
X-Axis
Logistic Regression
Y=1
Y-Axis
Y=0
X-Axis

Why is it logistic regression if all it does is classification?

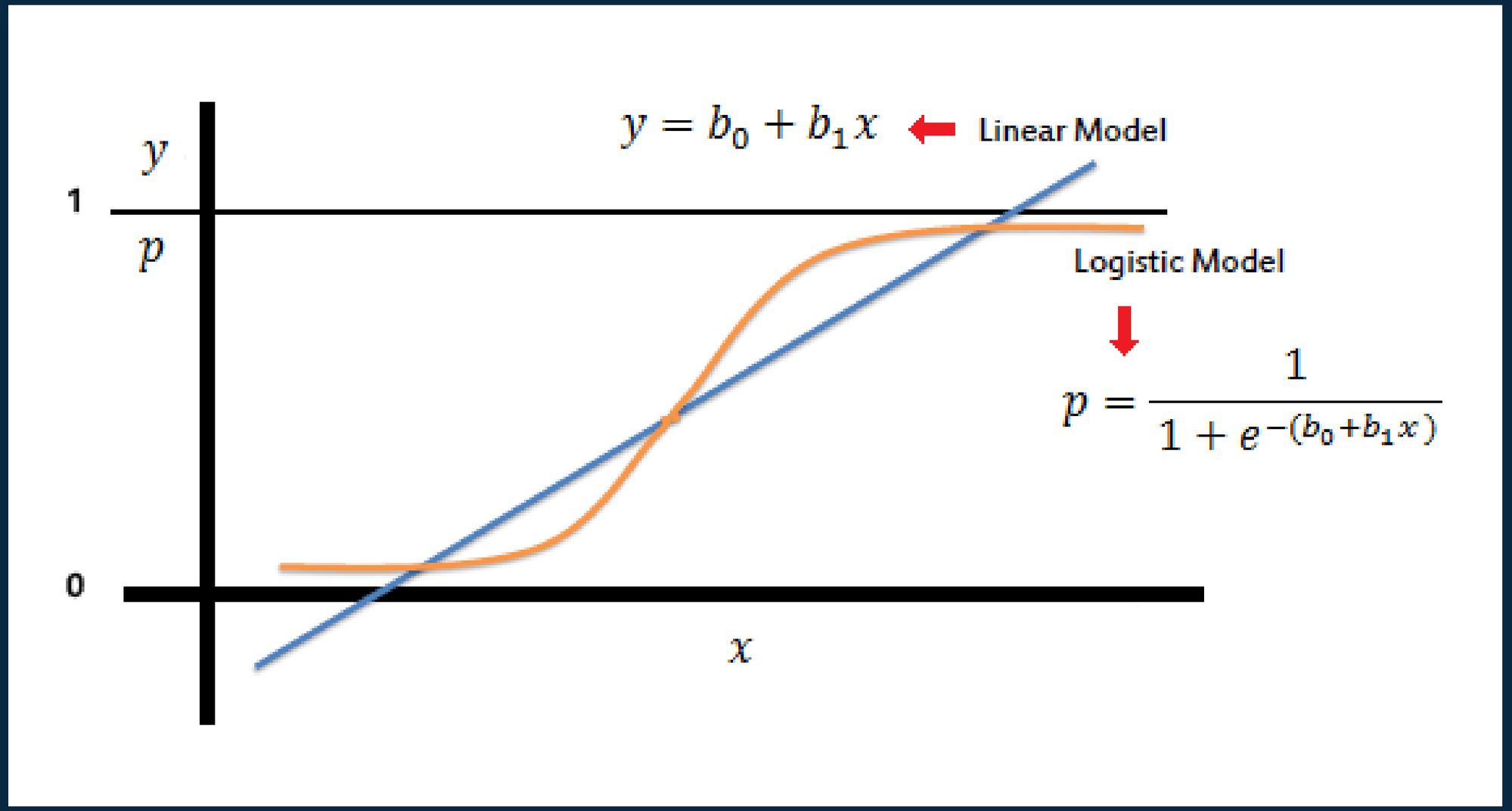
Understanding logistic regression

- The sigmoid function is a mathematical function used to map the predicted values to probabilities. It maps any real value into another value within a range of 0 and 1.

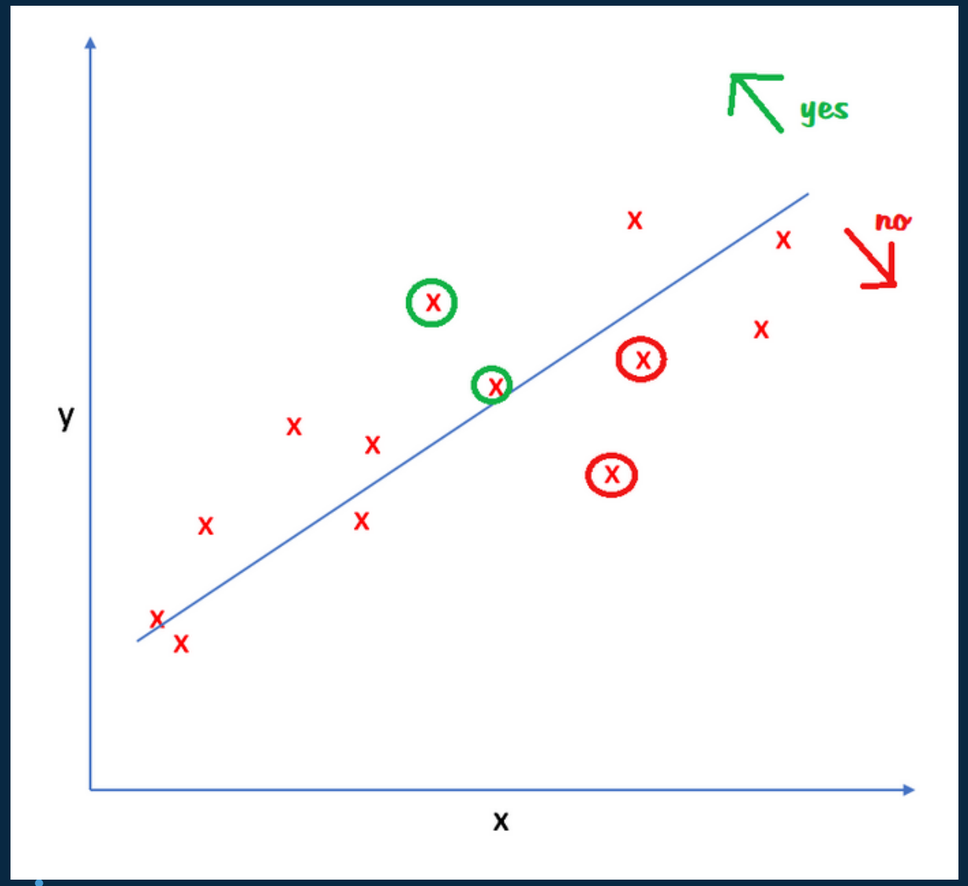
- The value of the logistic regression must be between 0 and 1, which cannot go beyond this limit, so it forms a curve like the "S" form. The S-form curve is called the Sigmoid function or the logistic function.

- $\sigma(z) = 1 / (1 + e^{(-z)})$
- where z is the linear equation: $z = \beta0 + \beta1x1 + \beta2x2 + ... + \beta nxn$

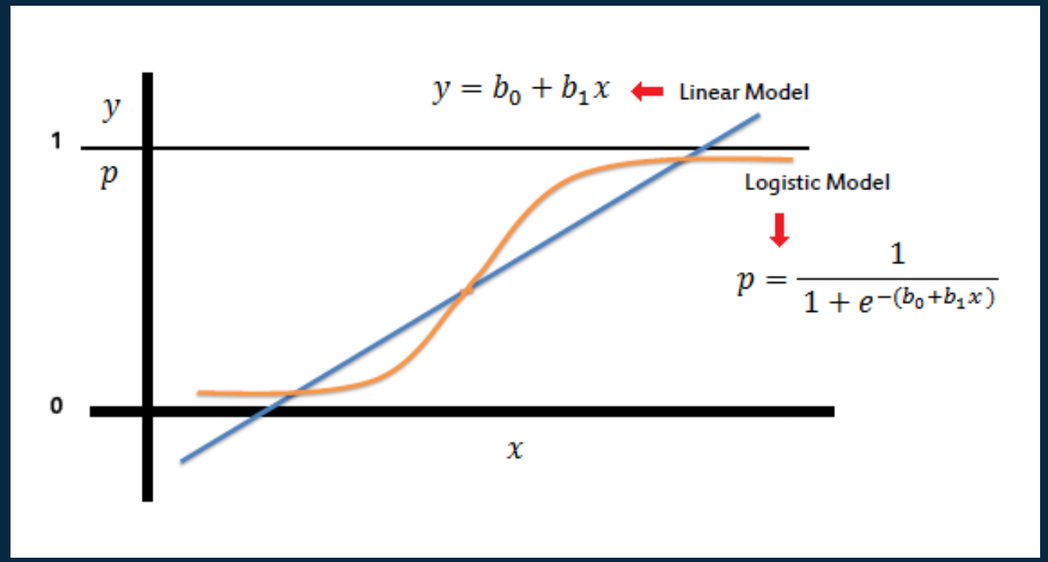
Here, β0, β1, β2, ..., βn are the coefficients of the regression equation and x1, x2, ..., xn are the independent variables.


$y = b_0 + b_1 x$
Linear Model
y
1
p
Logistic Model
$p = \frac{1}{1 + e^{-(b_0 + b_1 x)}}$
0
x

Where did these terms come from?



- This line is the line of best fit given out by logistic regression.
- If this line is seen as a boundary the closer a point gets near it the harder it its to classify it.(likelihood)
- The farther it gets it is easier to classify
- This can be related with probability



- Thus as we have the probability as the result we need to convert the the best fine line's range from 0 to 1
- We must also keep in mind of the conclusions that we made earlier.
- Range of the function is (-∞, +∞)
- Thus we can see that the sigmoid function makes the result value from 0 to 1

Evaluation metrics

- Accuracy: Measure of the total number of predictions a model gets right, including both True Positives and True Negatives

$$\text{Accuracy} = \frac{\sum \text{TP} + \text{TN}}{\sum \text{TP} + \text{FP} + \text{FN} + \text{TN}}$$

- When the classes are imbalanced, accuracy can be very high but still be misleading.

- accuracy does not provide information on the specific types of errors the model is making. In some situations, it may be more important to minimize false positives than false negatives or vice versa.

Evaluation metrics

Recall: Indicates the percentage of the response values (that we are interested in) were'actually captured by the model.

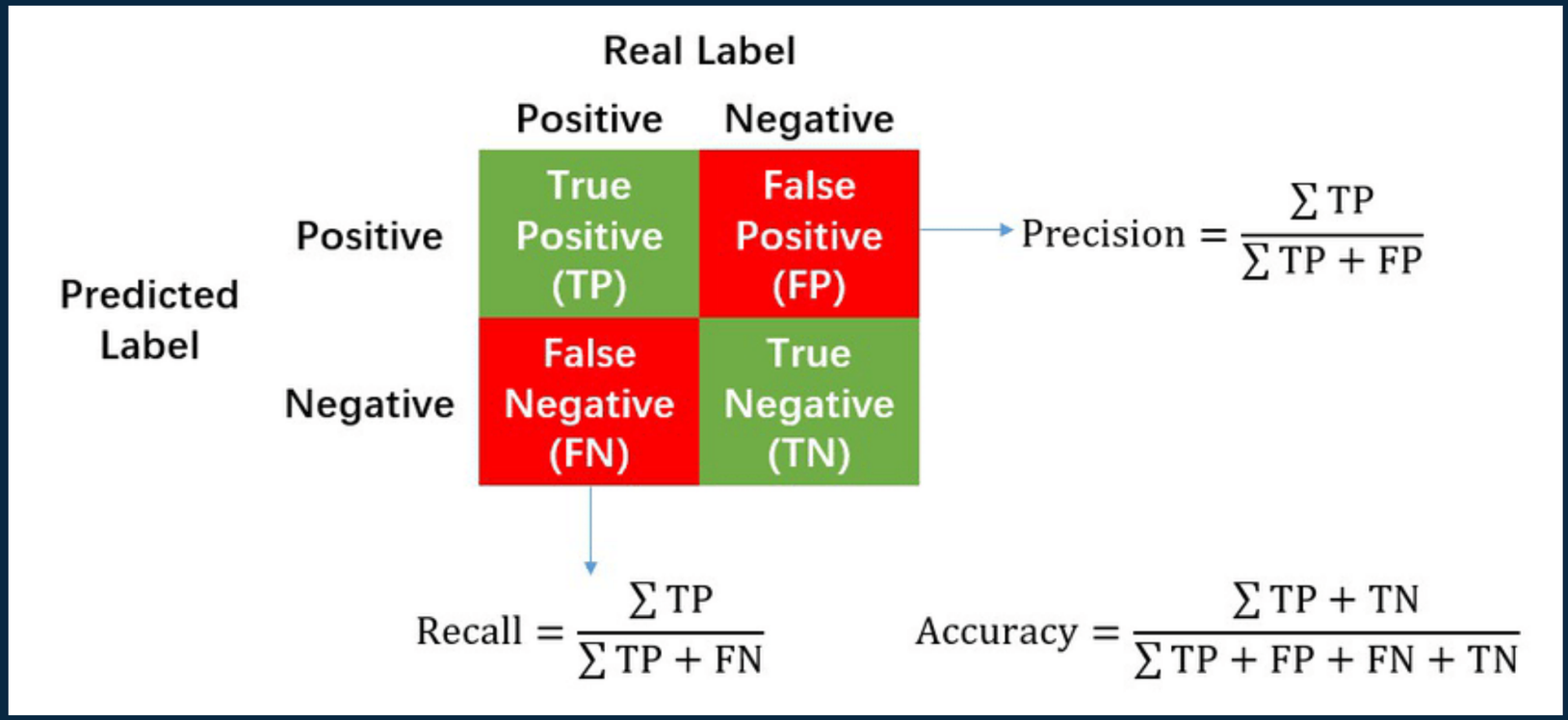
$$\text{Precision} = \frac{\sum \text{TP}}{\sum \text{TP} + \text{FP}}$$

Precision: Measures the percentage of the predicted response values (that we are interested in) that were correct.

$$\text{Recall} = \frac{\sum \text{TP}}{\sum \text{TP} + \text{FN}}$$

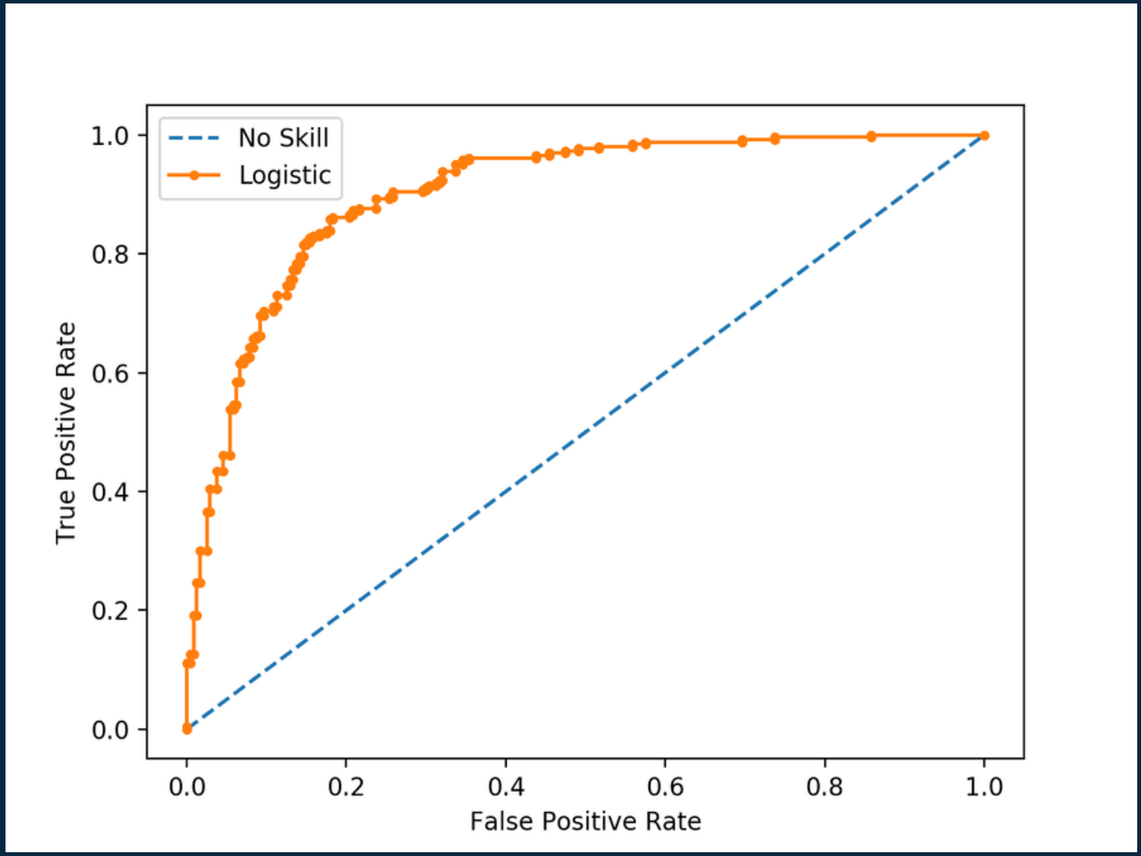
F1 score = 2 * (precision * recall) / (precision + recall), It is the harmonic mean.

Evaluation metrics



Evaluation metrics

ROC AUC



- AUC = ∫(TPR(FPR)) dFPR
- AUC value between 0.7 and 0.9 is considered good, while an AUC value above 0.9 is considered excellent.
- AUC of 0.5 indicates that the classifier is no better than random guessing.