Python for Beginners

Archer Infotech , PUNE

Python – Logistic Regression

Logistic Regression

Linear Regression

1. Home prices
2. Weather
3. Stock price

Predicted value is continuous

Classification

1. Email is spam or not
2. Will customer buy life insurance?
3. Which party a person is going to vote for?
   1. Democratic
   2. Republican
   3. Independent

Predicted value is categorical

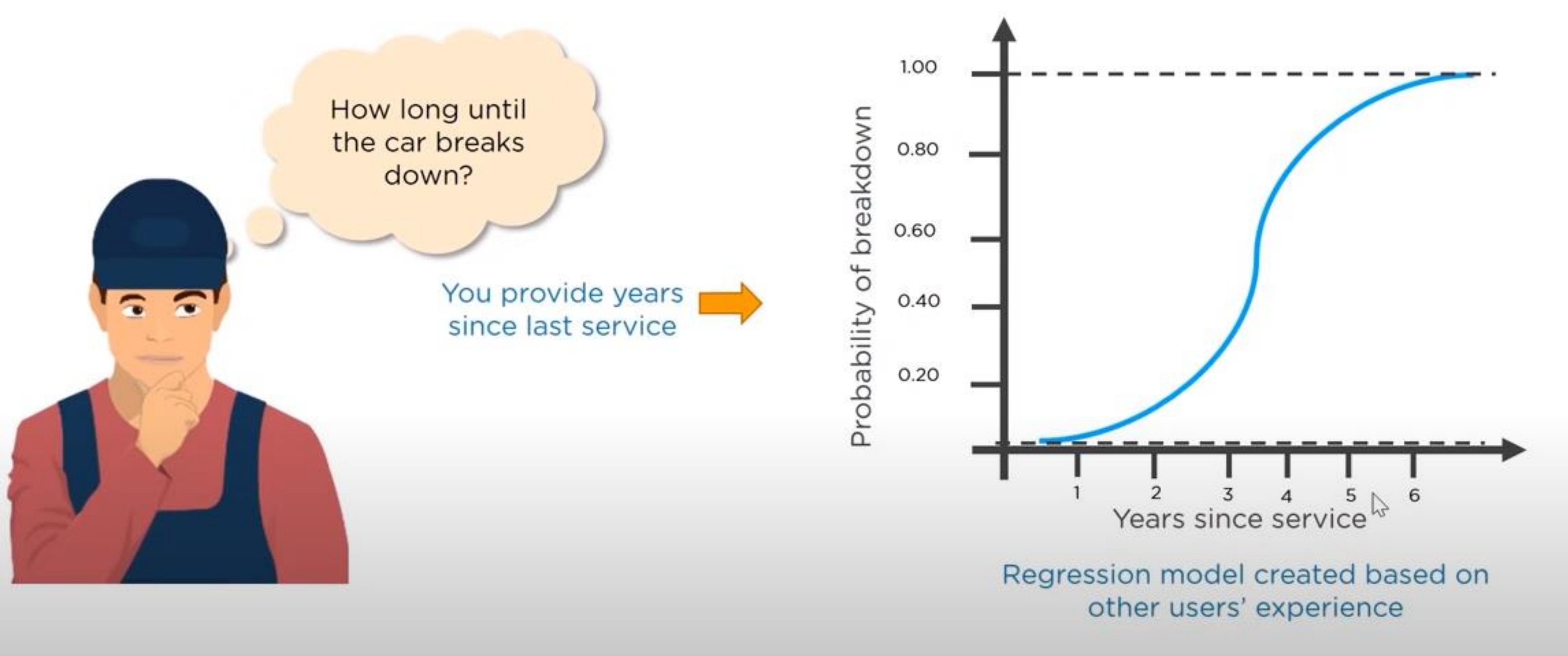
Logistic Regression

- **Logistic regression** is one of the most popular Machine Learning algorithms, which comes under the Supervised Learning technique. It is used for predicting the categorical dependent variable using a given set of independent variables.

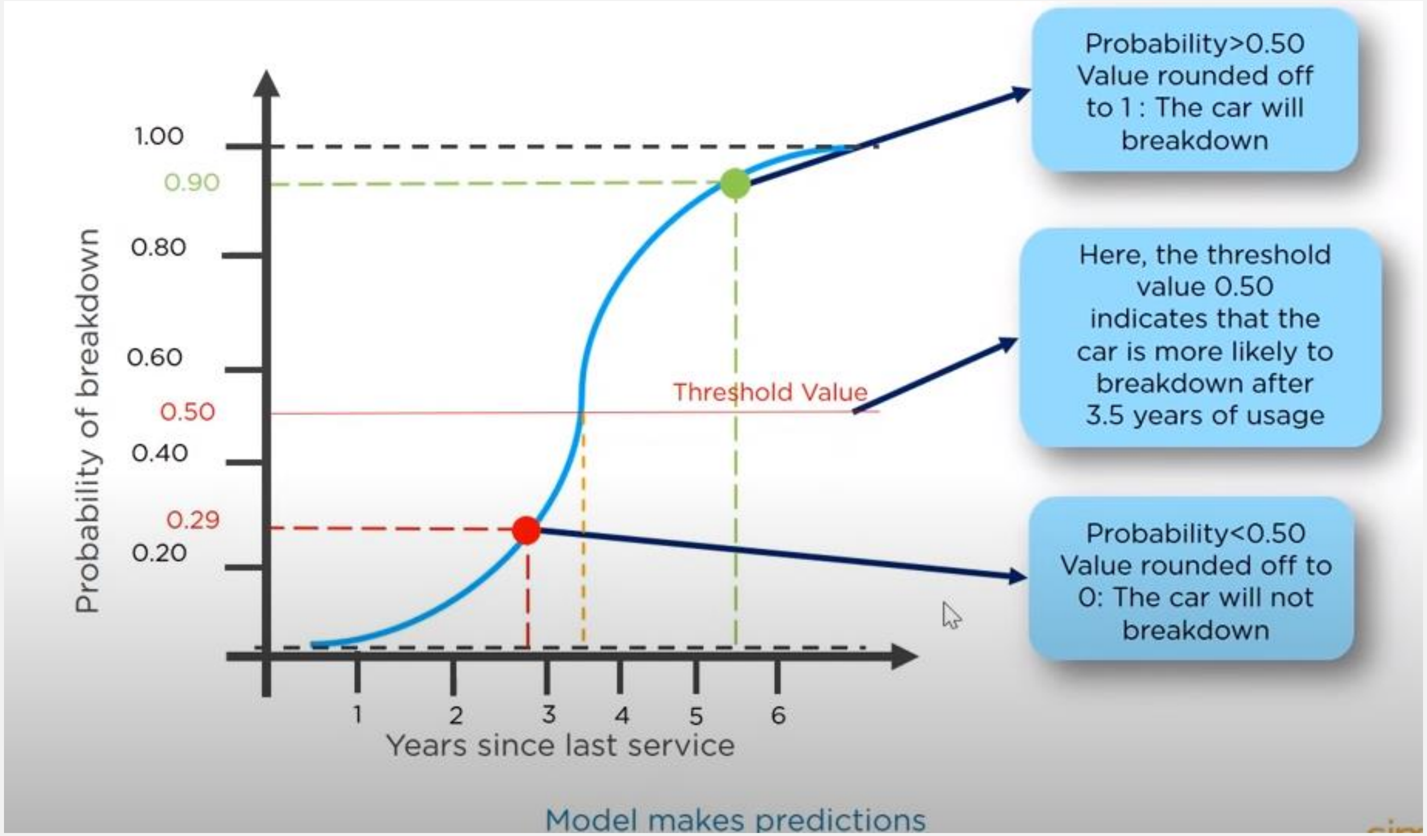
- Logistic regression predicts the output of a categorical dependent variable. Therefore the outcome must be a categorical or discrete value. It can be either Yes or No, 0 or 1, true or False, etc. but instead of giving the exact value as 0 and 1, **it gives the probabilistic values which lie between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

- In Logistic regression, instead of fitting a regression line, we fit an "S" shaped logistic function, which predicts two maximum values (0 or 1).

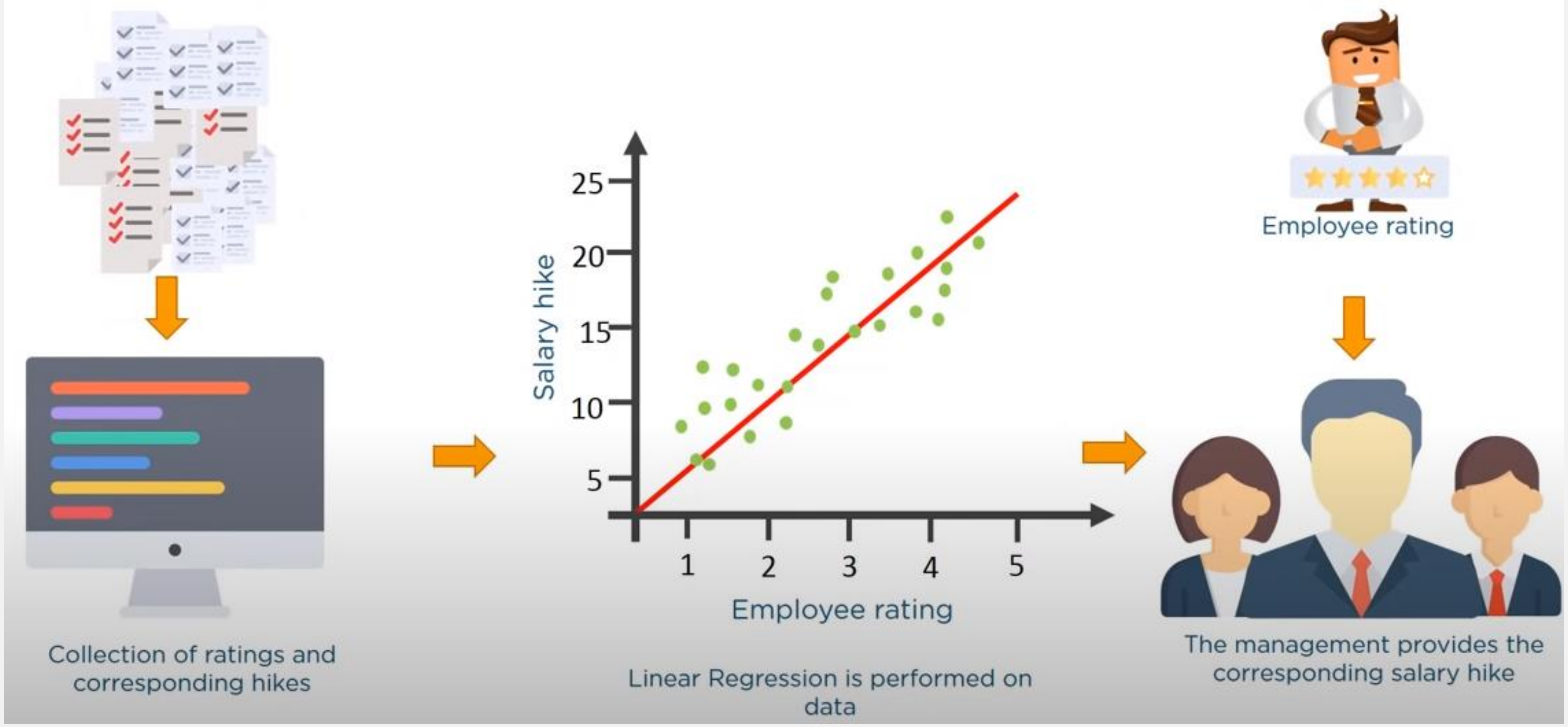
What is Logistic Regression ?



What is Logistic Regression ?

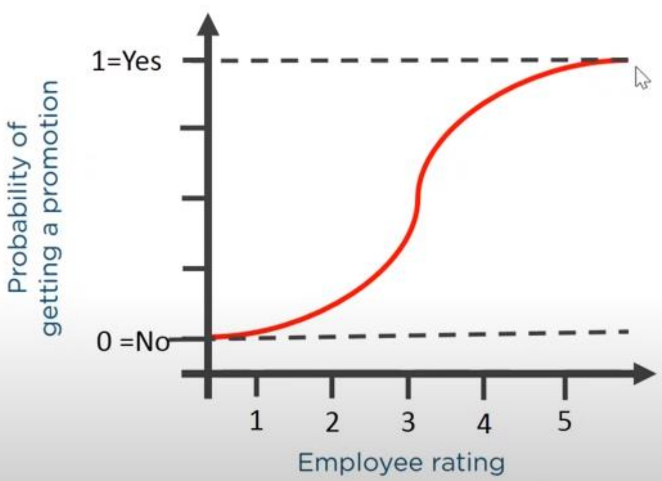


What is Logistic Regression ?



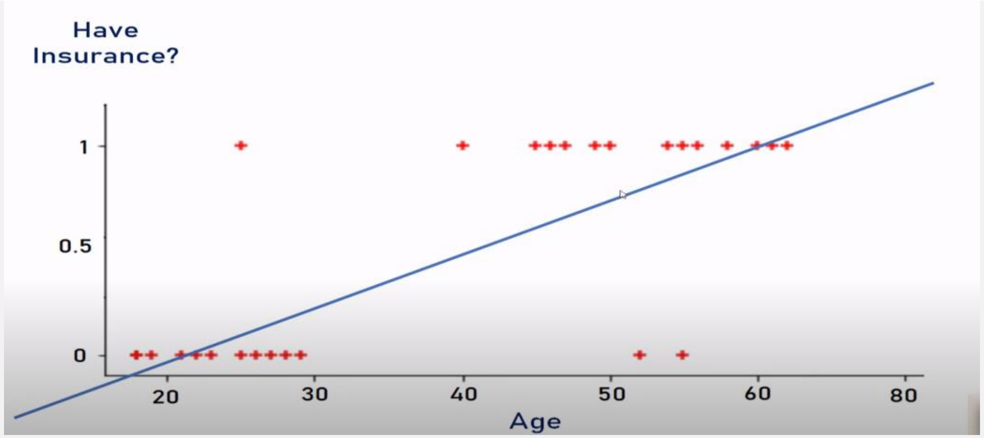
What is Logistic Regression ?

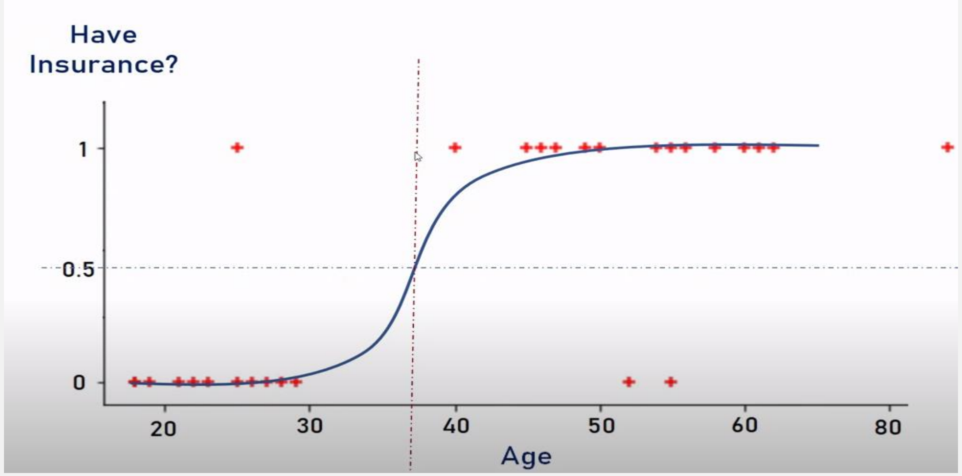
What if you wanted to know whether the employee would get a promotion or not based on their rating



What is Logistic Regression ?

age	have_insurance
22	0
25	0
47	1
52	0
46	1
56	1
55	0
60	1
62	1
61	1
18	0
28	0
27	0
29	0
49	1





Logistic Function

$$sigmoid(z) = \frac{1}{1 + e^{-z}}$$

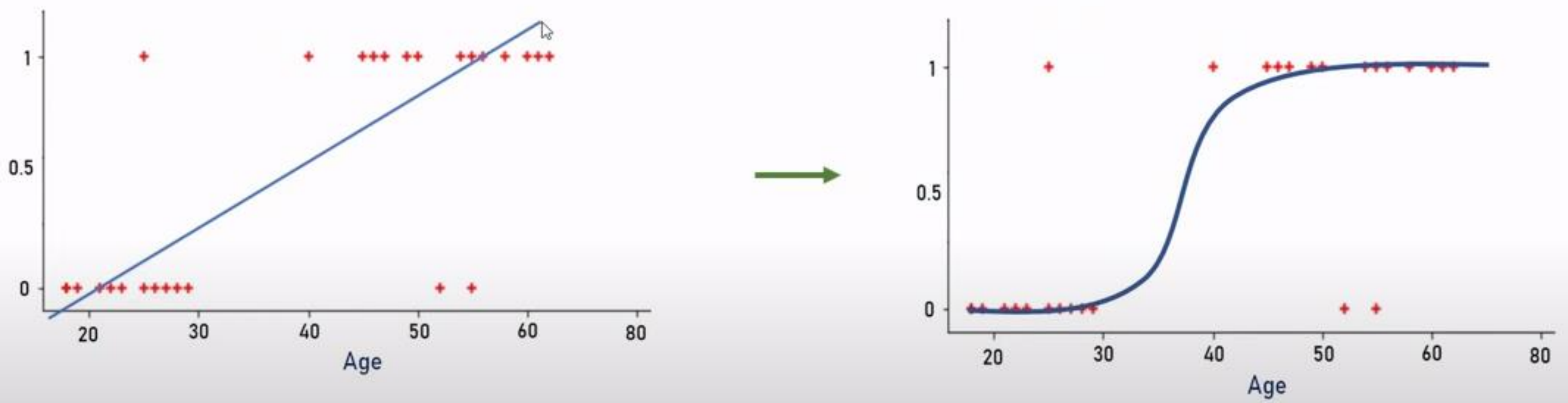
e = Euler's number ~ 2.71828

Sigmoid function converts input into range 0 to 1

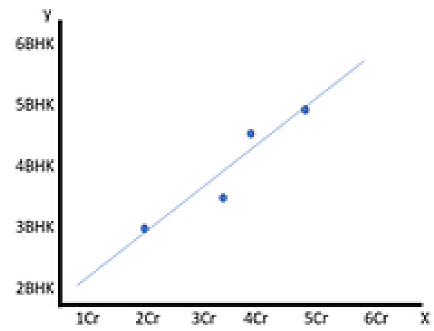
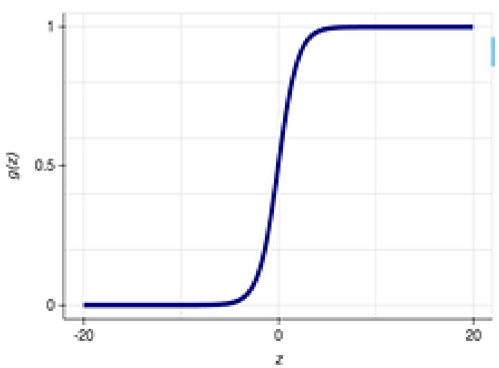
Logistic Function

$$y = m * x + b$$

$$y = \frac{1}{1 + e^{-(m*x+b)}}$$

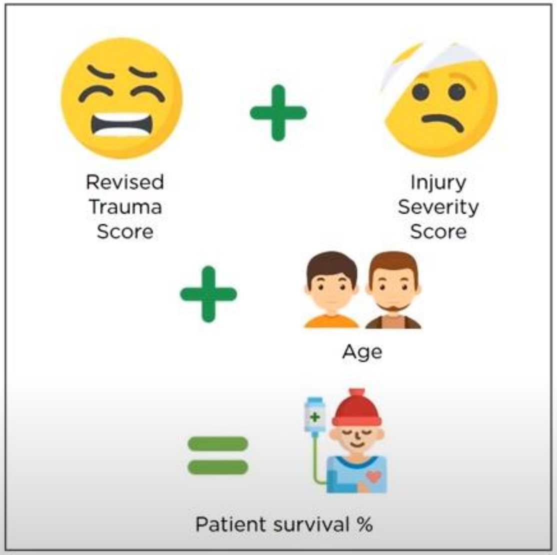


Linear Vs. Logistic Regression

Linear Regression	Logistic Regression
Target is an interval variable	Target is discrete (binary or ordinal) variable
Predicted values are the mean of the target variable at the given values of the input variable	Predicted values are the probability of the particular levels of the given values of the input variable
Solve regression problems	Solve classification problems
Example : What is the Temperature?	Example : Will it rain or not?
Graph is straight line	Graph is S-curve
	

Examples of Logistic Regression

Weather Prediction



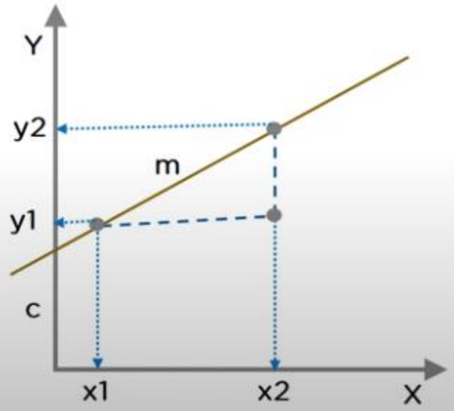
Healthcare (TRISS)

Image Categorization

Regression Equation

The simplest form of a simple linear regression equation with one dependent and one independent variable is represented by:

$$y = m * x + c$$



y ---> Dependent Variable

x ---> Independent Variable

m ---> Slope of the line $\quad m = \dfrac{y2 - y1}{x2 - x1}$

c ---> Coefficient of the line

Model Performance – Confusion Matrix

A confusion matrix is a table that is often used to **describe the performance of a classification model** (or "classifier") on a set of test data for which the true values are known.

n=165	Predicted: NO	Predicted: YES
Actual: NO	50	10
Actual: YES	5	100

- There are two possible predicted classes: "yes" and "no".
- The classifier made a total of 165 predictions (e.g., 165 patients were being tested for the presence of that disease).
- Out of those 165 cases, the classifier predicted "yes" 110 times, and "no" 55 times.
- In reality, 105 patients in the sample have the disease, and 60 patients do not.

Model Performance – Confusion Matrix

True Positive (TP)

- The predicted value matches the actual value
- The actual value was positive and the model predicted a positive value

True Negative (TN)

- The predicted value matches the actual value
- The actual value was negative and the model predicted a negative value

False Positive (FP) – Type 1 error

- The predicted value was falsely predicted
- The actual value was negative but the model predicted a positive value
- Also known as the **Type 1 error**

False Negative (FN) – Type 2 error

- The predicted value was falsely predicted
- The actual value was positive but the model predicted a negative value
- Also known as the **Type 2 error**

Model Performance – Confusion Matrix

		ACTUAL VALUES	
		POSITIVE	NEGATIVE
PREDICTED VALUES	POSITIVE	560	60
	NEGATIVE	50	330

True Positive (TP) = 560; meaning 560 positive class data points were correctly classified by the model

True Negative (TN) = 330; meaning 330 negative class data points were correctly classified by the model

False Positive (FP) = 60; meaning 60 negative class data points were incorrectly classified as belonging to the positive class by the model

False Negative (FN) = 50; meaning 50 positive class data points were incorrectly classified as belonging to the negative class by the model

Why Need Confusion Matrix ?

ID	Actual Sick?	Predicted Sick?	Outcome
1	1	1	TP
2	0	0	TN
3	0	0	TN
4	1	1	TP
5	0	0	TN
6	0	0	TN
7	1	0	FP
8	0	1	FN
9	0	0	TN
10	1	0	FP
:	:	:	:
1000	0	0	FN

TP = 30, TN = 930, FP = 30, FN = 10

$$Accuracy = \frac{30 + 930}{30 + 30 + 930 + 10} = 0.96$$

96%! Not bad!
But it is giving the wrong idea about the result.

Our model is saying "I can predict sick people 96% of the time". However, it is doing the opposite.

It is predicting the people who will not get sick with 96% accuracy while the sick are spreading the virus!

Precision Vs. Recall

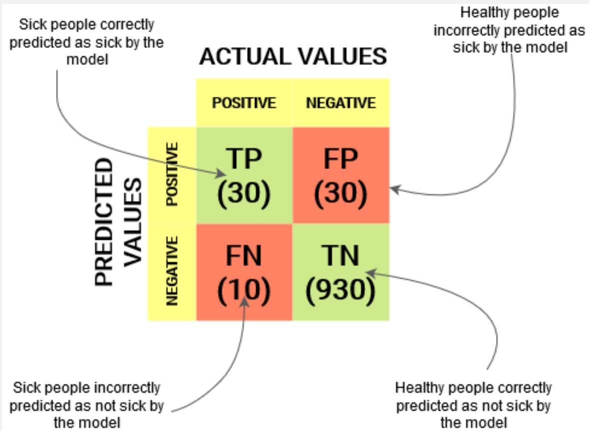
Precision tells us how many of the correctly predicted cases actually turned out to be positive.

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

Recall tells us how many of the actual positive cases we were able to predict correctly with our model.

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

Precision Vs. Recall



$$Precision = \frac{30}{30+30} = 0.5$$

$$Recall = \frac{30}{30+10} = 0.75$$

50% percent of the correctly predicted cases turned out to be positive cases. Whereas 75% of the positives were successfully predicted by our model. Awesome!

Precision is a useful metric in cases where False Positive is a higher concern than False Negativ

Precision is important in music or video recommendation systems, e-commerce websites, etc. Wrong results could lead to customer churn and be harmful to the business.

Recall is a useful metric in cases where False Negative trumps False Positive.

Recall is important in medical cases where it doesn't matter whether we raise a false alarm bu the actual positive cases should not go undetected!

Model Performance – F1 Score

In practice, when we try to increase the precision of our model, the recall goes down, and vice-versa. The F1-score captures both the trends in a single value:

$$F1 - score = \frac{2}{\frac{1}{Recall} + \frac{1}{Precision}}$$

F1-score is a harmonic mean of Precision and Recall, and so it gives a combined idea about these two metrics. It is maximum when Precision is equal to Recall.

Python Matrix

Sklearn has two great functions: **confusion_matrix()** and **classification_report()**.
Sklearn [confusion_matrix()]() returns the values of the Confusion matrix.

It takes the rows as Actual values and the columns as Predicted values. The rest of the concept remains the same.

Sklearn [classification_report()]() outputs precision, recall and f1-score for each target class. In addition to this, it also has some extra values: **micro avg**, **macro avg**, and **weighted avg**

$$\text{Micro avg Precision} = \frac{TP1 + TP2}{TP1 + TP2 + FP1 + FP2}$$

$$\text{Macro avg Precision} = \frac{P1 + P2}{2}$$

THANK YOU !!!

Amol Patil - 9822291613