Logistic Regression ML Model

Definition

- Logistic regression ML model is a classification algorithm. In classification the target variable is categorical. It has different categories also known as classes.

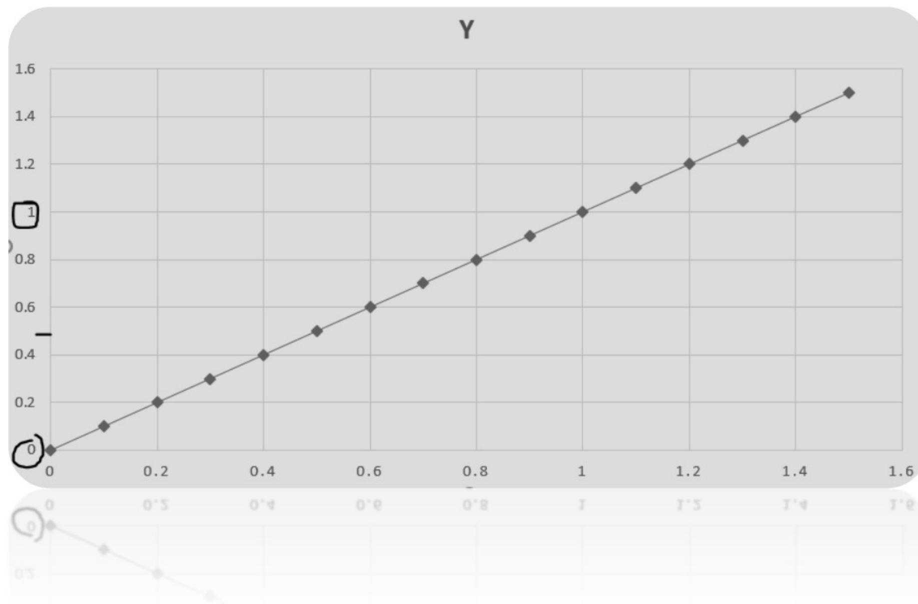
- Example: Yes/No , Pass/Fail , Spam/No Spam , Fraud transaction/Safe transaction, Survived/ Not Survived

Decisions:

- In logistic regression the two classes are defined as success and failure.
- 0 denotes ‘Failure’
- 1 denotes ‘Success’

Why is it called 'regression'?

- It is a classification model , not a regression model.
- But the underlying concept is based on linear regression.
- Here the aim is to create the best fit line, and then limit its values between 0 and 1 only.



- Then the decision boundary is created in the middle at 0.5

- Now if the target value is greater than decision boundary , it is considered as 1 and if it is less than decision boundary then it is considered as 0.
- A decision boundary of 0.5 , ensures that the division happens right from the mid-point leading to unbiasedness.

- So if target value is 0.75 ~~ 1
- And if target value is 0.35 ~~ 0
- Like this TV < 0.5 ~~ 0 & TV > 0.5 ~~ 1

- Hence we can conclude that the regression model classified the line into two categories that is the reason why logistic regression is called a classification model.

- In this model the acceptable value of target variable (Y) is 0 or 1.

- For linear regression the value of Y varies from $-\infty$ to $+\infty$
- Hence we need a function such that:

$$Y(-\infty,\ +\infty) \ \rightarrow\ Y(0,1)$$

- The function that helps us do so is called Sigmoid function

- Where: P is sigmoid function
- e is euler's number
- Y is the response variable

$$P = \frac{1}{1 + e^{-y}}$$

How does sigmoid function help?

- In Sigmoid function plug in
  Y = $-\infty$

$$P = \frac{1}{1 + e^{-(-\infty)}} = \frac{1}{\infty} = 0$$

- In Sigmoid function plug in
  Y = $+\infty$

$$P = \frac{1}{1 + e^{-(+\infty)}} = \frac{1}{1+0} = 1$$

Hence it can be seen that a sigmoid function converts the limits of Y to (0,1)

Transformation of sigmoid function

$$P = \frac{1}{1 + e^{-y}}$$

$$P\left(1 + e^{-y}\right) = 1$$

$$P + Pe^{-y} = 1$$

$$Pe^{-y} = 1 - P$$

$$e^{-y} = \frac{(1 - P)}{P}$$

To remove exponential , take Log on both sides

$$\log_e e^{-y} = \log_e\left(\frac{1 - P}{P}\right)$$

$$-Y = \ln\left(\frac{1 - P}{P}\right)$$

$$Y = \ln\left(\frac{P}{1 - P}\right)$$

- Hence we have converted the sigmoid function such that now it is expressed as providing the value of Y that is the target variable

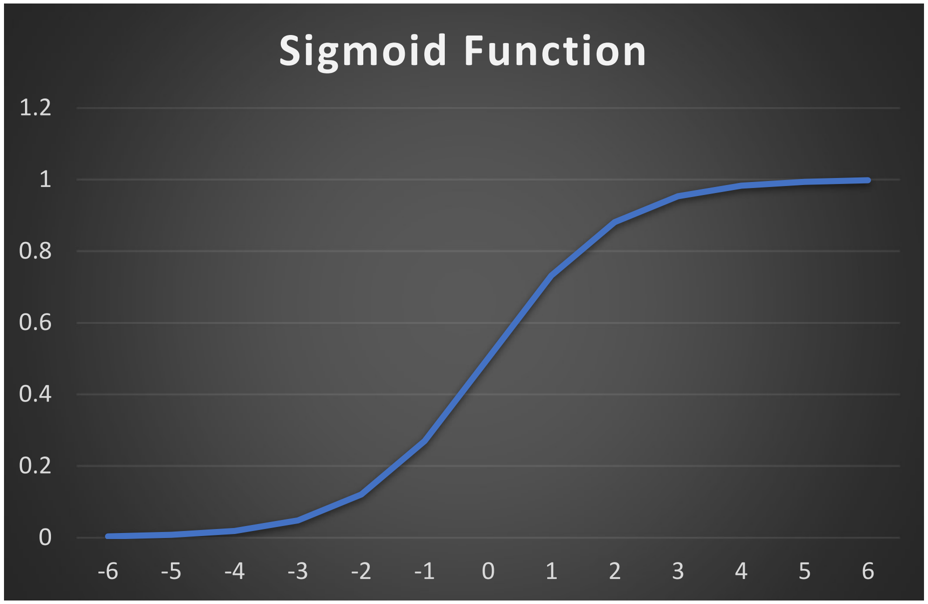
$$Y = \ln\left(\frac{P}{1 - P}\right)$$

- Here (P/(1-P)) is called “odds ratio”.
- Hence Y can be defined as log of odds ratio.
- ***Y is also called as Log Odd Function or Logistic Function or Logit Function***

Graph of sigmoid function

- If the sigmoid function P is plotted on the graph , it can be observed that the curve lies between 0 and 1 on the Y axis.

- Hence it gets a S shaped curve.



Evaluation Matrix for Classification Model

- ✓Confusion matrix
- ✓Accuracy
- ✓Misclassification
- ✓TPR
- ✓FPR
- ✓TNR
- ✓Precision
- ✓Recall
- ✓F1 score
- ✓ROC curve
- ✓AUC

Y (original)	Y(Predicted)
P	P
P	P
N	P
N	N
P	P
P	N
N	N
N	P
P	P
P	N

Total actual P = 6 Correctly predicted = 4

Total actual N = 4 Correctly predicted = 2

Confusion Matrix

- Covid cases: There are two aspects.
- One actual fact: whether a person has covid or not
- Second predicted result: whether the test came positive or not

		Actual		
		+	-	Total
Predicted	+	20	5	
	-	15	60	
	Total			

- The above table so obtained by tabulating the actual counts vs the predicted counts is called confusion matrix

Confusion matrix

Actual

Predicted

	+	-	Total
+	20	5	25
-	15	60	75
Total	35	65	

Actual

Predicted

	+	-	Total
+	True Positive	False Positive	Total Predicted Positve
-	False Negative	True Negative	Total Predicted Negative
Total	Total Actual Positive	Total Actual Negative	

Accuracy of the model

- It is defined as the ratio of **correct** predictions to total predictions.

- Total number of correct predictions = TP + TN
- Total Predictions = sum of all 4 cells

		Actual		
		+	-	Total
Predicted	+	20	5	
	-	15	60	
	Total			

$$\text{Accuracy} = \frac{\text{TP} + \text{TN}}{\text{Total number of predictions}}$$

Misclassification

- It is defined as the ratio of incorrect predictions to total predictions.
- Total number of incorrect predictions = FP + FN
- Total Predictions = sum of all 4 cells

		Actual		
		+	-	Total
Predicted	+	20	5	
	-	15	60	
	Total			

$$\text{Misclassification} = \frac{\text{FP} + \text{FN}}{\text{Total number of predictions}}$$

True Positive Rate (Also called 'Recall')

- It is defined as the ratio of TP to total actual positives.

$$\text{TPR} = \frac{\text{TP}}{\text{Total number of Actual Positives}}$$

		Actual		
		+	-	Total
Predicted	+	20	5	
	-	15	60	
	Total			

False Positive Rate

- It is the ratio of FP to Total Actual Negatives

$$FPR = \frac{FP}{\text{Total number of Actual Negatives}}$$

		Actual		
		+	-	Total
Predicted	+	20	5	
	-	15	60	
	Total			

True Negative Rate

- It is defined as the ratio of TN to Total Actual Negatives

$$\text{TNR} = \frac{\text{TN}}{\text{Total number of Actual Negatives}}$$

		Actual		
		+	-	Total
Predicted	+	20	5	
	-	15	60	
	Total			

Precision of the model

- It is defined as the ratio of TP to Total Predicted Positive

$$\text{Precision} = \frac{\text{TP}}{\text{Total number of Predicted Positives}}$$

		Actual		
		+	-	Total
Predicted	+	20	5	
	-	15	60	
	Total			

- It is also known as positive predictive value

Difference between Precision and Recall

$$\text{Precision} = \frac{\text{TP}}{\text{Total number of Predicted Positives}}$$

$$\text{TPR} = \frac{\text{TP}}{\text{Total number of Actual Positives}}$$

Precision relates the true positive to the predicted positives whereas recall or TPR relates the true positive to the actual positives

Actual

Predicted		+	-	Total
	+	True Positive	False Positive	Total Predicted Positive
	-	False Negative	True Negative	Total Predicted Negative
	Total	Total Actual Positive	Total Actual negative	

Use case of precision: Identifying a mail as spam. If a Business Mail is marked as spam , this is FALSE POSITIVE. This harms the business. Here Precision is used to evaluate the model.

Use case of recall: Identifying a transaction as fraud. If a wrong transaction is NOT marked as fraud , this is FALSE NEGATIVE. This harms the business. Here recall is used to evaluate the model.

F1 Score

- This is a measure that shows the combined effect of Precision & Recall
- F1 score is the harmonic mean of Precision and Recall

$$\text{F1 Score} = \frac{2}{\frac{1}{\text{Precision}} + \frac{1}{\text{Recall}}}$$

- When is F1 score used?
- When False Positive and False Negative both are important parameters for the business F1 score helps.

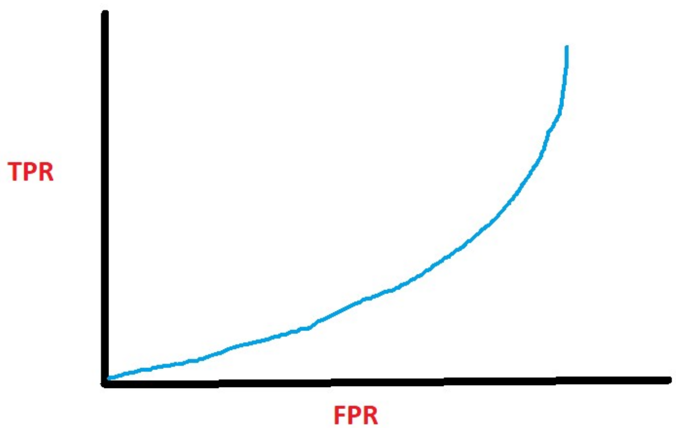
- Drawback of F1 Score:
- Interpretability of F1 score is difficult
- We cannot individually comment about False Negative and False Positive

- It is precisely used to compare two classifiers. If suppose model A has higher Precision and model B has higher Recall. In that scenario the F1 score of model A and B is compared.

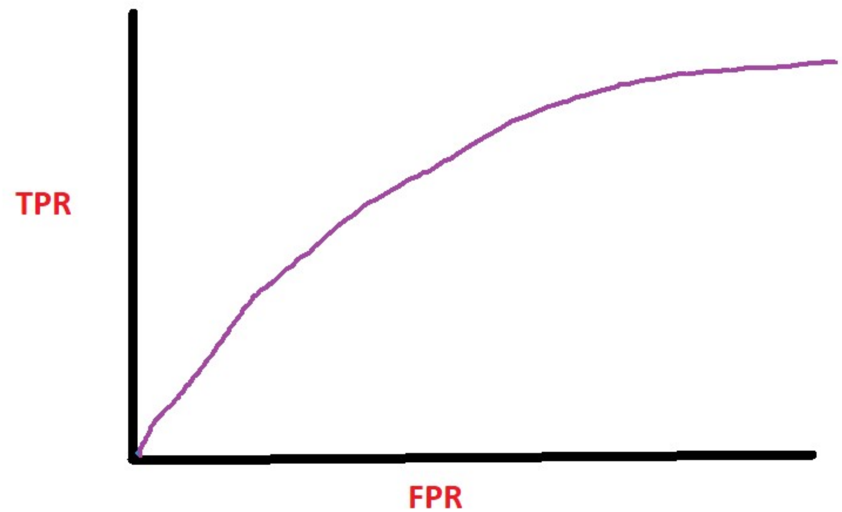
ROC Curve

- ROC means Receiver Operating Characteristics.
- This was initially used by operators of military radar in 1941 , that is why it is named as ROC
- ROC curve is a graph plotted between TPR and FPR
- In Machine Learning Classification models ROC helps to analyze the operating characteristics of the model.

- TPR < FPR



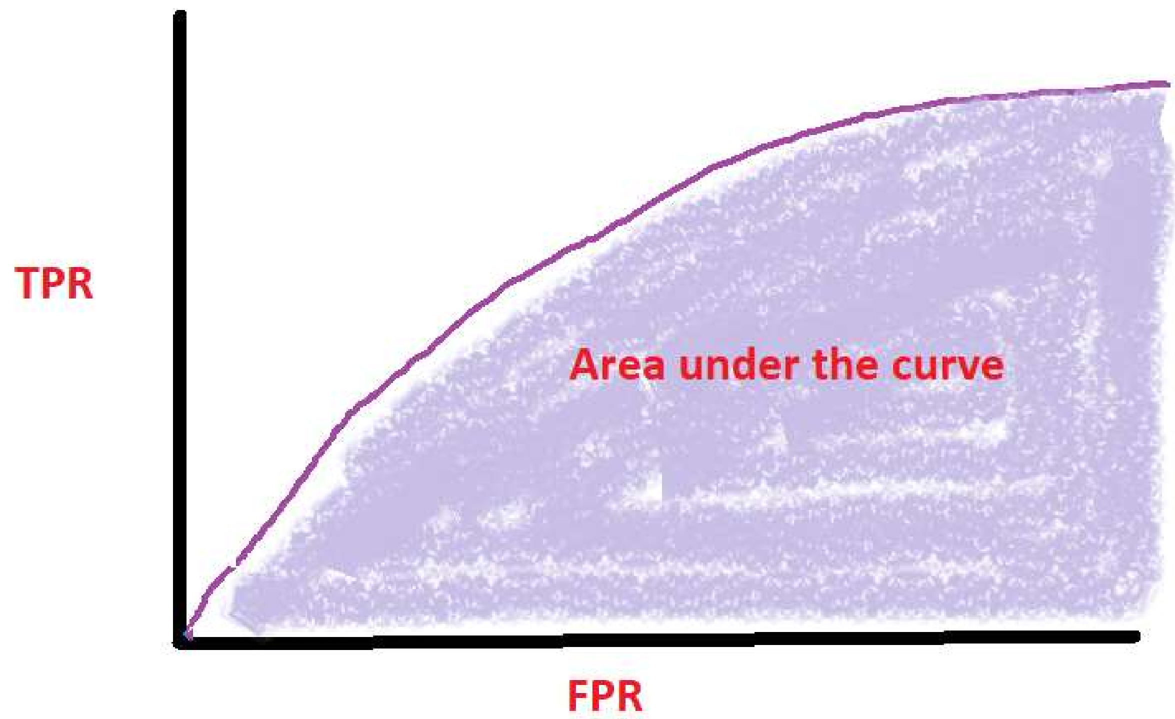
- TPR > FPR



Area under the curve

- Area under the curve is the total area under the ROC curve.

- Higher the Area under the curve, higher is TPR and that indicates that the model is doing a good job!



Confusion Matrix format as dispayed in the output of python

Predicted

		-	+
Actual	-	TN	FP
	+	FN	TP