Machine Learning

Muhammad Bilal Shahnawaz

bilal.shahnawaz@numl.edu.pk

Outline

- Last Class
  - Machine Learning and Pattern Recognition
  - Supervised Learning, Unsupervised Learning, Reinforcement Learning
  - Decision process
  - Feature selection
  - Pattern recognition process
- Today
  - ML Life cycle
  - Performance Measure Parameters
    - *Classification Accuracy (mostly used)*
    - *Confusion Matrix*
    - *Area under Curve*
    - *Mean Absolute Error*
    - *Mean Squared Error*
  - Common ML Problems

Machine learning Life cycle

- A cyclic process to build an efficient machine learning project.
- The main purpose of the life cycle is to find a solution to the problem or project.
- Machine learning life cycle involves seven major steps, which are given below:
  - Gathering Data
  - Data preparation
  - Data Wrangling (Data cleansing)
  - Analyze Data
  - Train the model
  - Test the model
  - Deployment

1. Gathering Data

- The first and most important step of the ML life cycle.
- The goal is to identify and obtain all data-related problems.
- The quantity and quality of the collected data will determine the efficiency of the output.
  - The more will be the data, the more accurate will be the prediction.
- This step includes the below tasks:
  - **Identify various data sources** such as **files, database, internet** etc.
  - **Collect data**
  - **Integrate the data obtained from different sources**
- By the end of this step, a **dataset** is gathered to be used in further steps.

2. Data preparation

- Data preparation is a step where we put our data into a suitable place and prepare it to use in our machine learning training.
    - Put all data together, and then randomize the ordering of data.
- This step can be further divided into two processes:
    - **Data exploration:**
        - Understand the nature of data
        - Understand the characteristics, format, and quality of data.
    - **Data pre-processing:**
        - Manipulating the data so that it can run through the appropriate machine learning technique with no problems.
        - converting categorical columns into numerical columns

3. Data Wrangling

- The process of cleaning and converting raw data into a useable format, to make it more suitable for analysis in the next step
- May involve dealing with following issues:
  - **Missing Values**
  - **Duplicate data**
  - **Invalid data**
  - **Noise**

- So, Data can be cleaned by either removing rows with missing/ duplicate values or imputing the missing values
- Typically done by a data scientist or business analyst to change views on a dataset and for features engineering.

4. Data Analysis

- The cleaned and prepared data is passed on to the analysis step that involves:
  - Determination of the type of the problem
  - Selection of analytical techniques such as Classification, Regression, Cluster analysis, Association
  - Building models, e.g. setting the parameters

Model Training, Testing and Deployment

- Train Model
  - Feed dataset to ML algorithms so that it can understand the various patterns, rules, and, features.

- Test Model
  - Test the trained model using test unseen dataset
  - Goal is to check the accuracy of the model

- Deployment
  - Deploy the model in the real-world system, if it is producing an accurate result as per our requirement with acceptable speed
  - similar to making the final report for a project.

Evaluating matrices of model

- The performance of a ML model is evaluated using some or all of these evaluation metrics
  - *Classification Accuracy (mostly used)*
  - *Confusion Matrix*
  - *Area under Curve*
  - *F1 Score*
  - *Mean Absolute Error*
  - *Mean Squared Error*

Contd..

- **Classification Accuracy**
  - The ratio of number of correct predictions to the total number of input samples.

$$Accuracy = \frac{Number\ of\ correct\ Predictions}{Total\ number\ of\ Predictions\ made}$$

  - works well if there are equal number of samples belonging to each class.

- **Confusion matrix**
- Confusion Matrix gives us a matrix as output and describes the complete performance of the model.

Confusion Matrix

- Lets assume we have a binary classification problem. We have some samples belonging to two classes : **YES or NO**. Also, we have our own classifier which predicts a class for a given input sample.
- On testing our model on 165 samples ,we get the following result:

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

Yes-> Patient
No-> No Disease

Total Predictions= 165
    Yes=110 times
    No= 55 times
Actual:
    Yes =105 times
    No = 60 times

Contd..

- There are 4 important terms :
  - **True Positives** : The cases in which we predicted YES and the actual output was also YES.
  - **True Negatives** : The cases in which we predicted NO and the actual output was NO.
  - **False Positives** : The cases in which we predicted YES and the actual output was NO.
  - **False Negatives** : The cases in which we predicted NO and the actual output was YES.

n=165	**Predicted: NO**	**Predicted: YES**	
Actual: NO	TN = 50	FP = 10	60
Actual: YES	FN = 5	TP = 100	105
	55	110	

Computations from confusion matrix

- List of rates that are often computed from a confusion matrix for a binary classifier:
- **Accuracy:** Overall, how often is the classifier correct?

$$Accuracy = \frac{TP + TN}{Total\ number\ of\ Predictions\ made}$$

$$Accuracy = \frac{100 + 50}{165} = 0.91$$

- **Misclassification Rate (Error Rate)**: Overall, how often is it wrong?

$$Error\ rate = \frac{FP + FN}{Total\ number\ of\ Predictions\ made}$$

$$Error\ rate = \frac{10 + 5}{} = 0.09 \quad \text{i.e. 1 minus Accuracy}$$

Contd..

- **True Positive Rate:** When it's actually yes, how often does it predict yes?

$$\text{True Positive Rate}(\boldsymbol{\mathit{Sensitivity}}/\text{Recall}) = \frac{TP}{Actual\ yes\ (TP + FN)}$$

$$Sensitivity = \frac{100}{105} = 0.95$$

- **True Negative Rate** : When it's actually no, how often does it predict no?

$$\text{True Negative Rate } (\textbf{Specificity}) = \frac{TN}{Actual\ no\ (TN + FP)}$$

$$\textbf{Specificity} = \frac{50}{60} = 0.83 \quad \text{i.e. 1 minus } False\ Positive\ Rate$$

- **False Positive Rate** : When it's actually no, how often does it predict yes?

$$\text{False Positive Rate} = \frac{FP}{Actual\ no\ (TN + FP)}$$

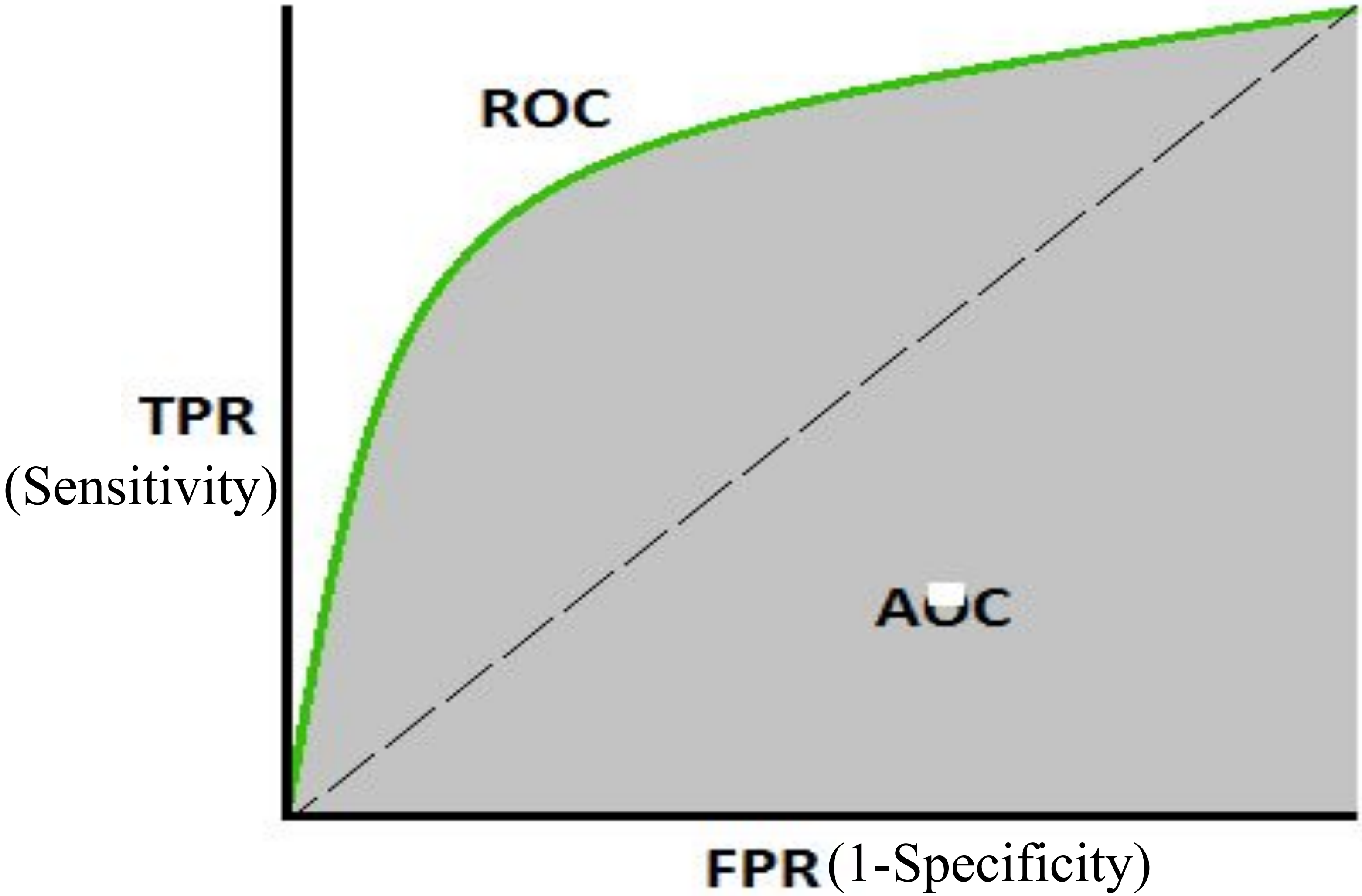
$$\text{False Positive Rate} = \frac{10}{60} = 0.17 \quad \text{i.e. 1 minus Specificity}$$

AUC- ROC Curve

- A commonly used graph that summarizes the performance of a classifier over all possible thresholds.
- It is generated by plotting the True Positive Rate (y-axis) against the False Positive Rate (x-axis)
- ROC is a probability curve and AUC represents the degree or measure of separability (simply the area under the curve).
  - It tells how much the model is capable of distinguishing between classes.
  - Higher the AUC, the better the model is at predicting 0 classes as 0 and 1 classes as 1.

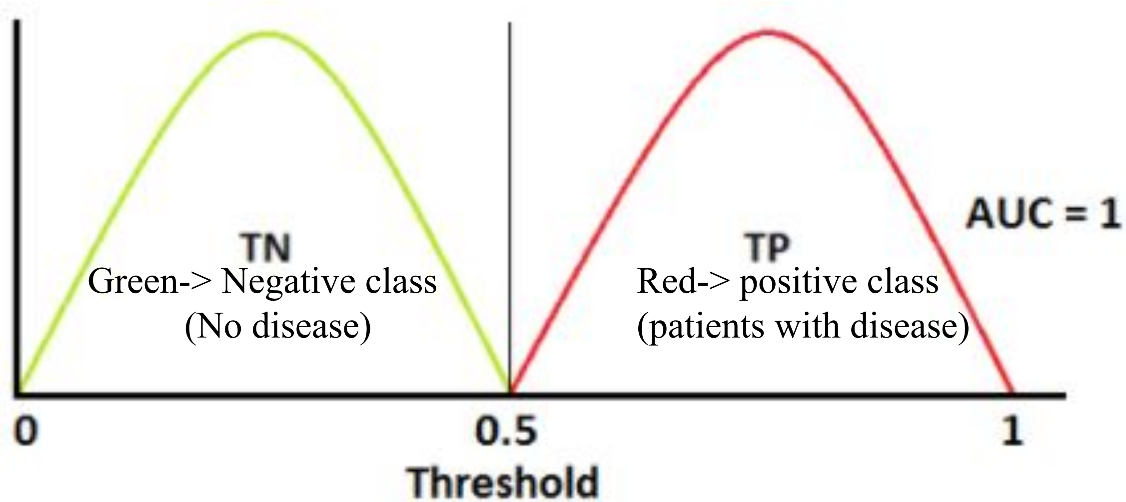
Contd..

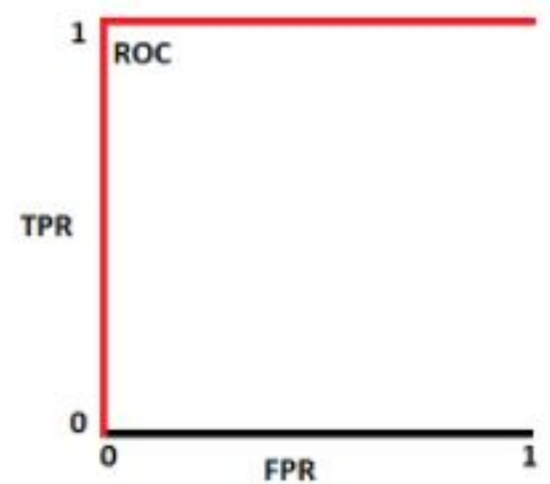
- .



AUC- ROC Curve representation(1)

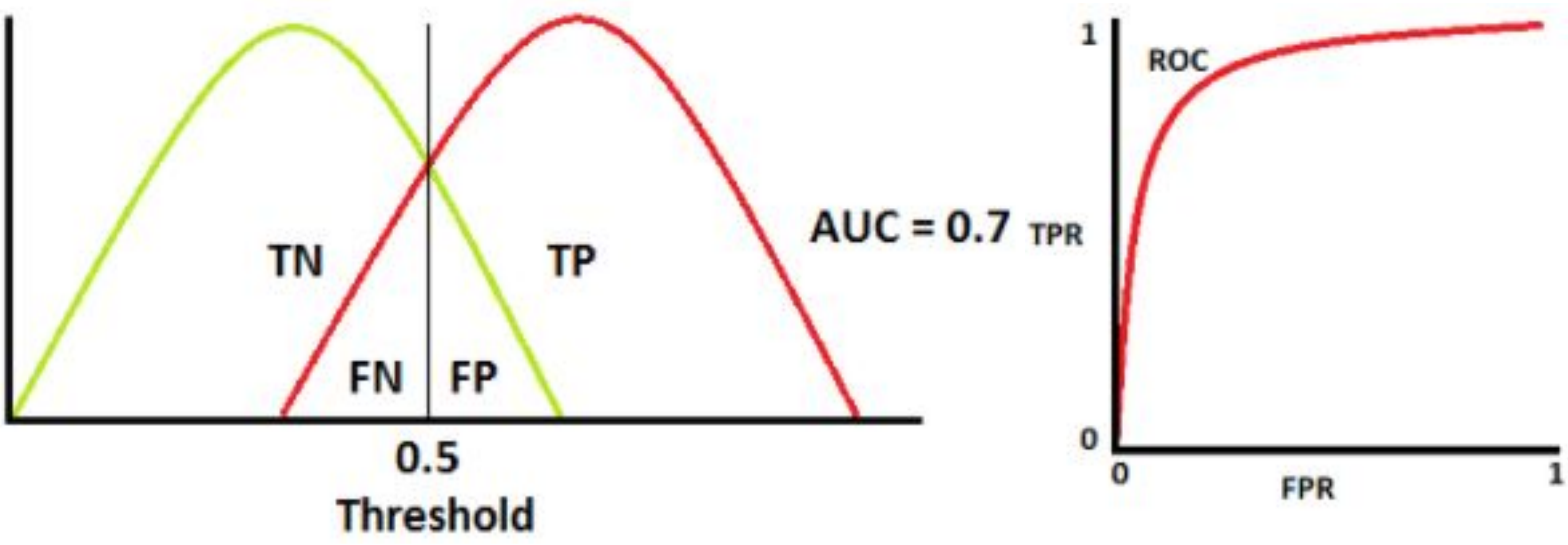
- An ideal situation: Two curves don’t overlap at all
  - Model has an ideal measure of separability (AUC=1).
  - It is perfectly able to distinguish between positive class and negative class.





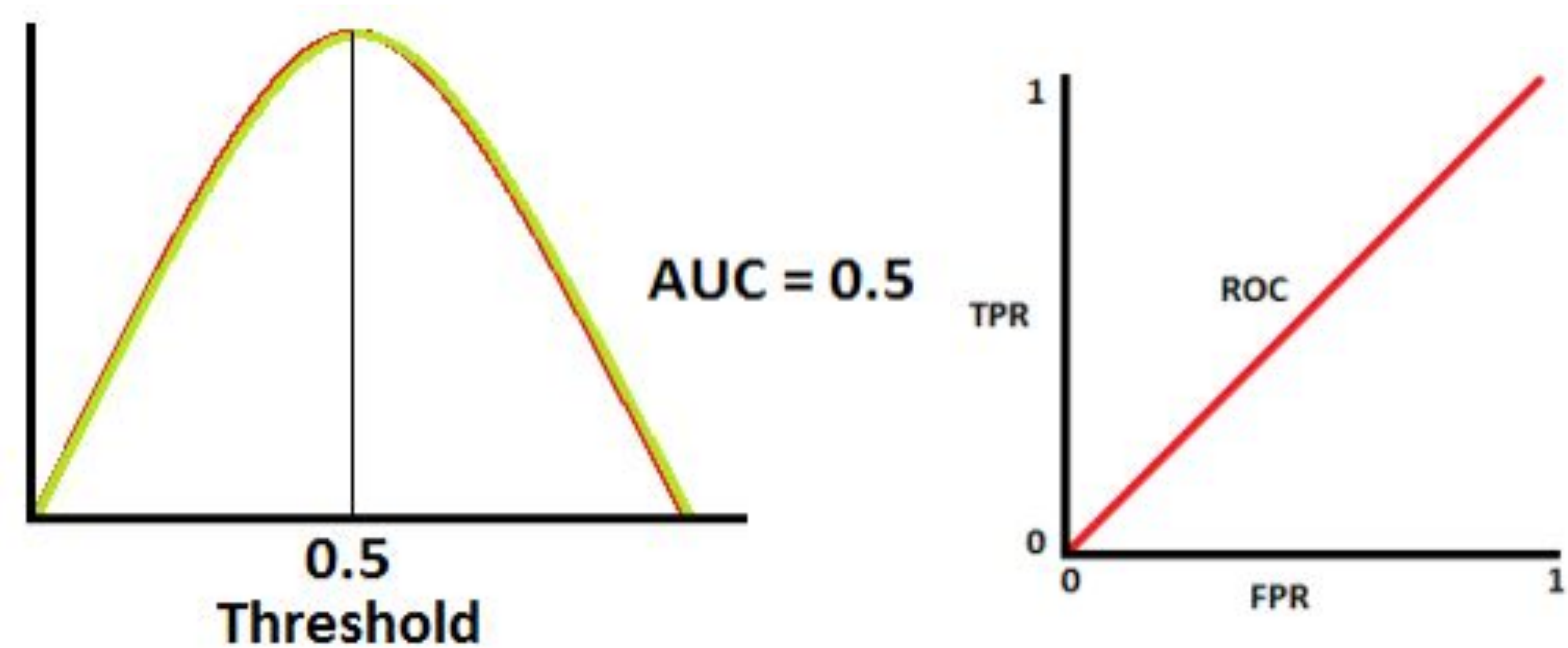
AUC- ROC Curve representation(2)

- Two distributions overlap
- AUC is 0.7, it means there is a 70% chance that the model will be able to distinguish between positive class and negative class.



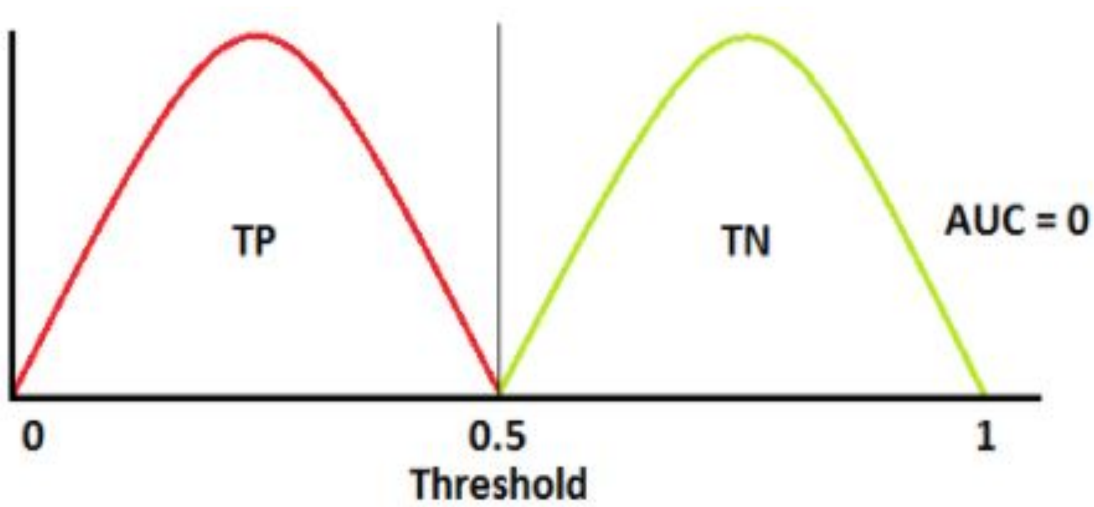
AUC- ROC Curve representation(3)

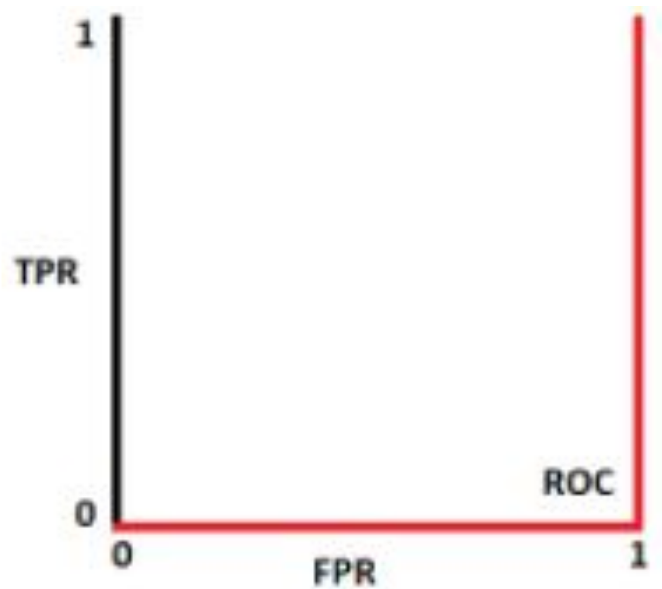
- AUC is approximately 0.5, the model has no discrimination capacity to distinguish between positive class and negative class.



AUC- ROC Curve representation(4)

- AUC is approximately 0, the model is actually reciprocating the classes.
  - It means the model is predicting a negative class as a positive class and vice versa.





Contd..

- **Note**: The ROC curve is normally used to assess the performance of binary classifiers.
    - ROC curve can be extended to problems with three or more classes using the One vs ALL methodology:
    - Suppose you have three classes named X, Y, and Z, you will have three ROCs.
    - one ROC for X classified against Y and Z **(X vs Y&Z)**
    - second ROC for Y classified against X and Z, **(Y vs X&Z)**
    - third one of Z classified against Y and X. **(Z vs X,Y)**
- With imbalanced datasets, AUC score is calculated from ROC and is a very useful metric in imbalanced datasets.
- The AUC score can be calculated with Simpson's Rule using the graph values

Mean Absolute Error (MAE)

- The average of the difference between the Original Values and the Predicted Values
  - gives us the measure of how far the predictions were from the actual output.
- However, don't gives us any idea of the direction of the error i.e. whether we are under predicting the data or over predicting the data.

- Mathematically, it is represented as :

$$\text{Mean Absolute Error} = \frac{1}{N}\sum_{j=1}^{N} |y_j - \hat{y}_j|$$

Mean Squared Error (MSE)

- quite similar to Mean Absolute Error
- Difference is that MSE takes the average of the **square** of the difference between the original values and the predicted values:

- Mathematically, it is represented as :

$$\text{Mean Squared Error} = \frac{1}{N}\sum_{j=1}^{N}(y_j - \hat{y}_j)^2$$

ML/PR Challenges

- There are a lot of challenges that machine learning professionals face to instil ML skills and create an application from scratch.
- Few of them are:
    - Poor Quality of Data
    - Feature Extraction
    - Overfitting and Underfitting of Training Data
    - Lack of Training Data
    - Model Selection
    - Prior Knowledge
    - Missing Features
    - Imperfections in the Algorithm When Data Grows
    - Inadequate Infrastructure
    - Lack of Skilled Resources

END