Performance Metrics to Evaluate Classification and Regression Algorithms

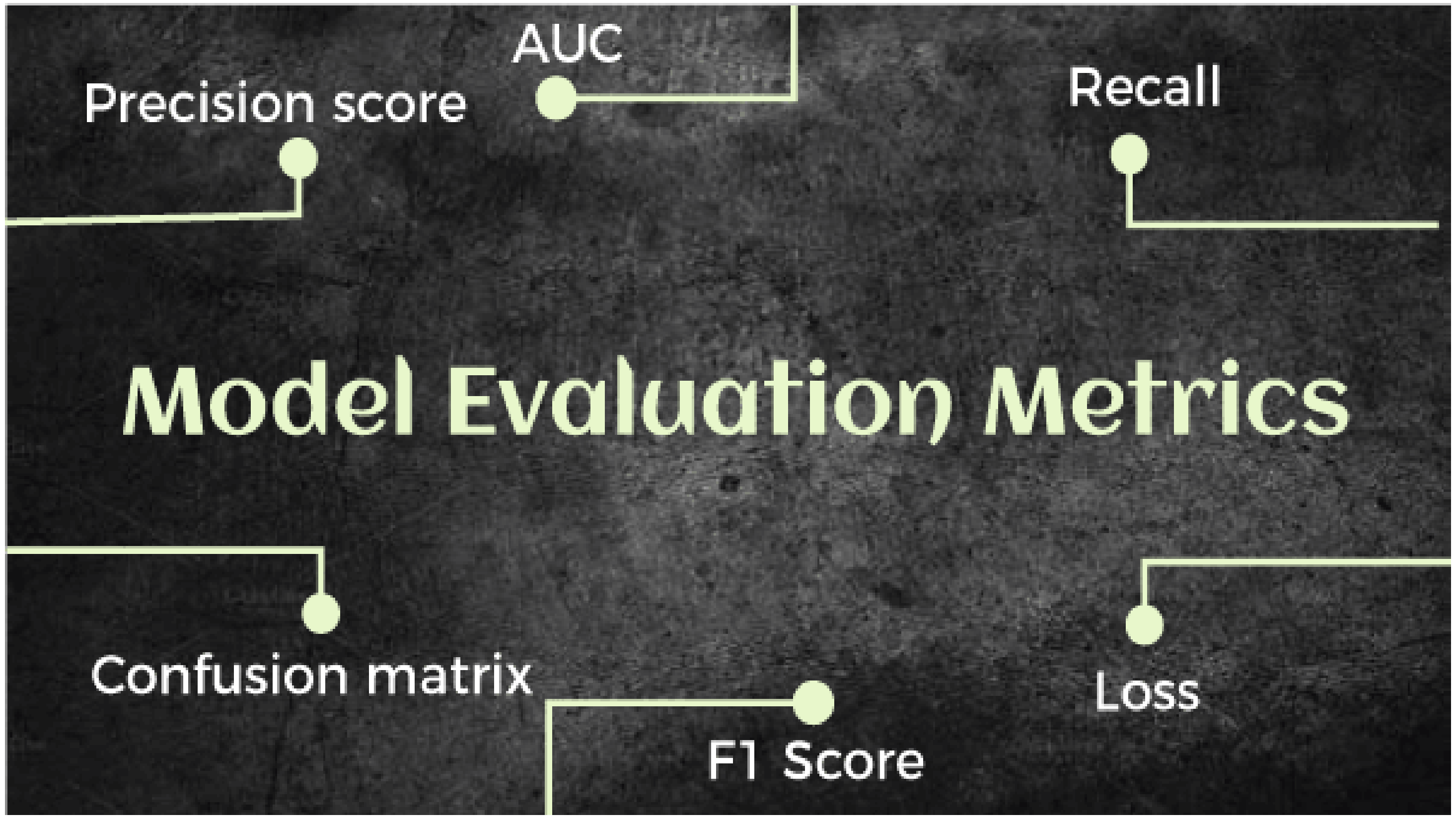
School of Computer Engineering

Evaluating Supervised ML Models

- After implementing a supervised ML algorithm the next step is to find out how effective is the model based on metric and data sets.
- Different performance metrics are used to evaluate the performance or quality of the models and these metrics are known as performance metrics or evaluation metrics.
- These performance metrics help us understand how well our model has performed for the given data.
- In this way, we can improve the model's performance by tuning the hyper-parameters.

- Each ML model aims to generalize well on unseen/new data, and performance metrics help determine how well the model generalizes on the new dataset.
- In Supervised learning, each task or problem is divided into Classification and Regression.
- Different evaluation metrics are used for both Regression and Classification tasks.
- Each Supervised ML model aims to generalize well on unseen/new data, and performance metrics help determine how well the model generalizes on the new data.
- We will discuss metrics used for classification and regression tasks.

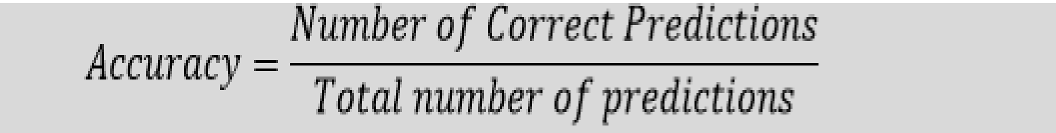
Performance Metrics for Classification



- In a classification problem, the categories or classes of example/observation is identified based on training dataset.
- The model learns from the given training set and then classifies the test set examples into classes or categories based on the learning during training phase.
- To evaluate the performance of a classification model, different metrics are used, and some of them are as follows:
- Accuracy, Confusion matrix (not a metric but fundamental to others), Precision, Recall, F1-score, AUC (Area Under the Curve)-ROC (Receiver Operating Characteristic) curve.

Classification Accuracy

- The accuracy metric is one of the simplest Classification metrics to implement, and it can be determined as the number of correct predictions to the total number of predictions.
- It can be expressed as:



$$Accuracy = \frac{Number\ of\ Correct\ Predictions}{Total\ number\ of\ predictions}$$

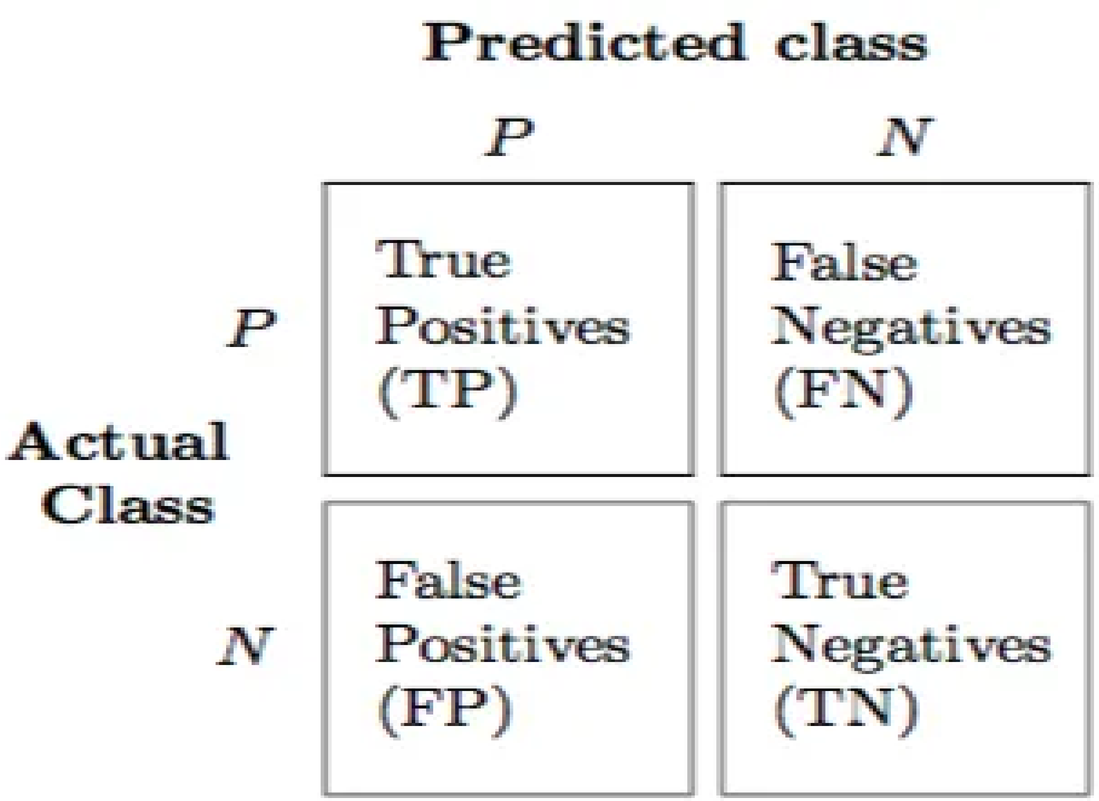
- To implement an accuracy metric, we can compare ground truth and predicted values in a loop.
- Although it is simple to use and implement, it is suitable only for cases where an equal number of samples belong to each class.

Confusion Matrix

- A confusion matrix is a tabular representation of prediction outcomes of any binary classifier, which is used to describe the performance of the classification model on a set of test data when true values are known.
- Confusion Matrix as the name suggests gives us a matrix as output as N X N matrix , where N is the number of classes being predicted.
- Confusion matrix, also known as an error matrix.
- This Metric used for finding the correctness and accuracy of the model and even works better for imbalanced data set.

Confusion matrix for Binary Class:

- Confusion matrix for binary classifier

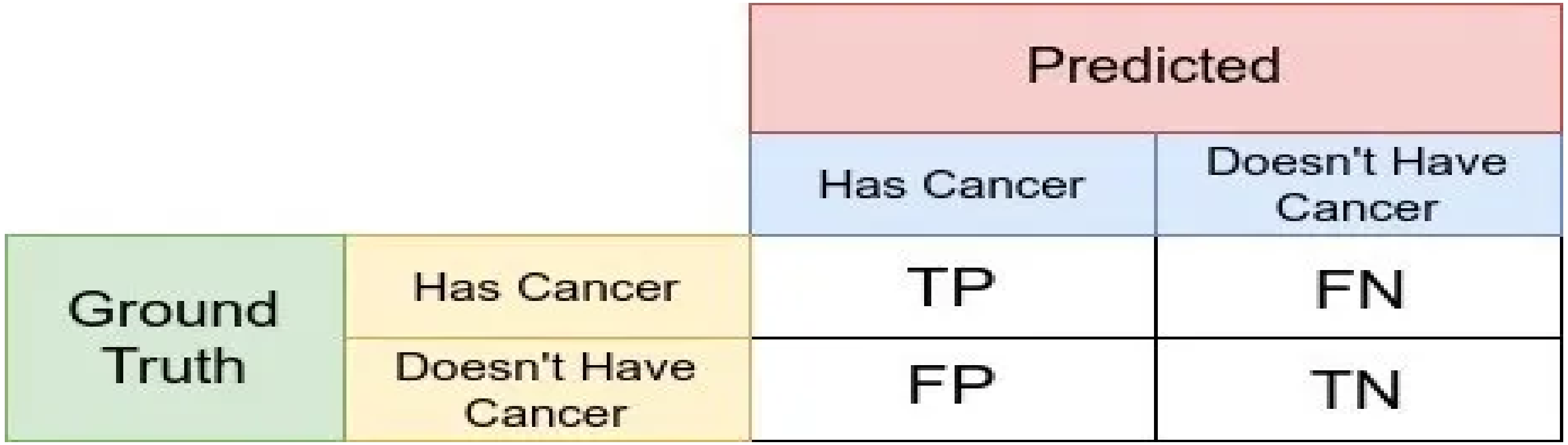


		Predicted class	
		P	*N*
Actual Class	*P*	True Positives (TP)	False Negatives (FN)
	N	False Positives (FP)	True Negatives (TN)

Confusion Matrix

- Confusion Matrix is a tabular visualization of the ground-truth labels versus model predictions.
- Each row of the confusion matrix represents the instances in a predicted class and each column represents the instances in an actual class.
- Confusion Matrix is not exactly a performance metric but sort of a basis on which other metrics evaluate the results.
- In order to understand the confusion matrix, we need to set some value for the null hypothesis as an assumption.

- For example, from our Breast Cancer data, let’s assume our Null Hypothesis (Ho) be “The individual has cancer”.

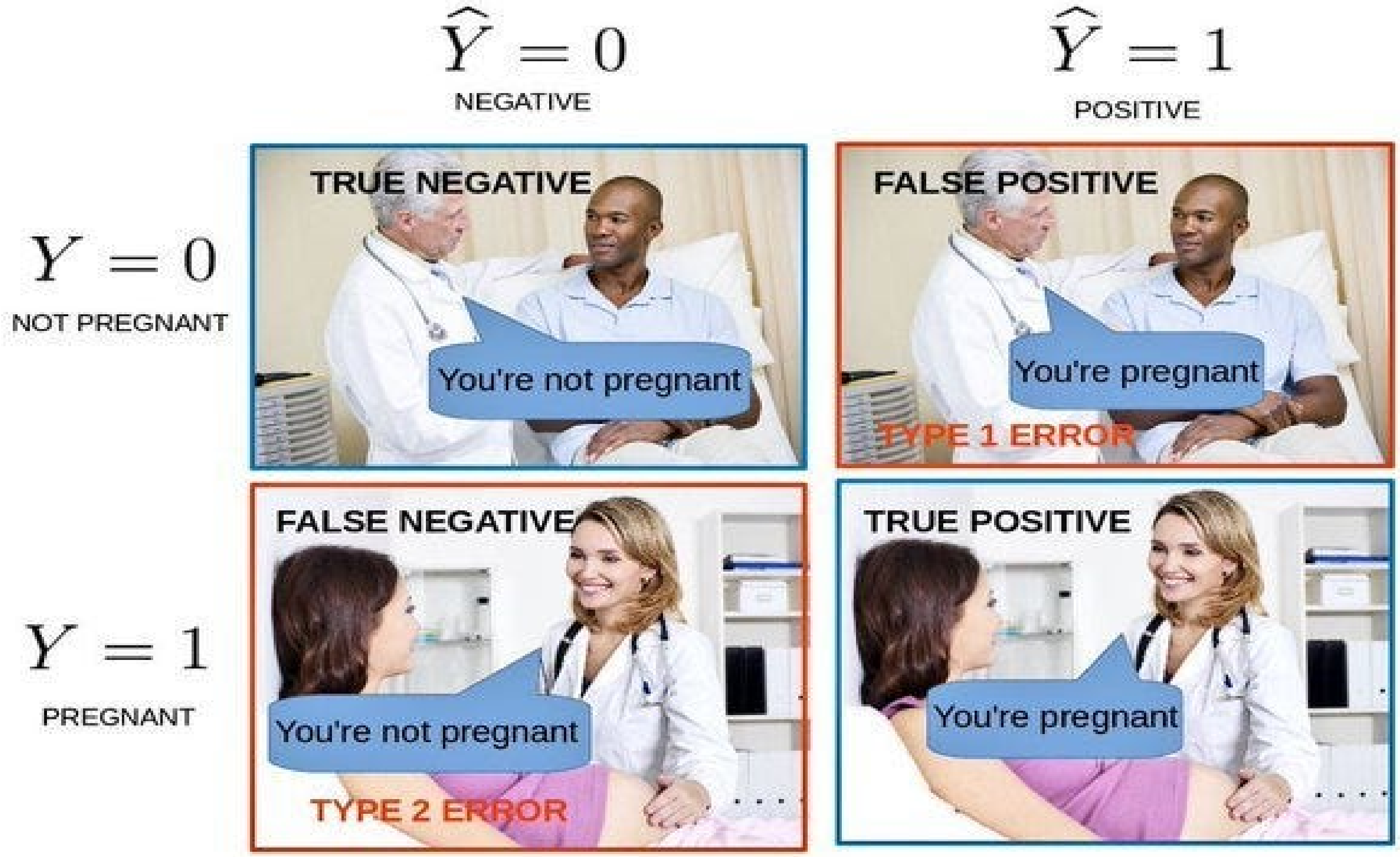


		Predicted	
		Has Cancer	Doesn't Have Cancer
Ground Truth	Has Cancer	TP	FN
	Doesn't Have Cancer	FP	TN

- Let’s understand these factors one by one:

- True Positive(TP) signifies how many positive class samples your model predicted correctly.
- True Negative(TN) signifies how many negative class samples your model predicted correctly.
- False Positive(FP) signifies how many negative class samples your model predicted incorrectly. This factor represents Type-I error in statistical nomenclature. This error positioning in the confusion matrix depends on the choice of the null hypothesis.
- False Negative(FN) signifies how many positive class samples your model predicted incorrectly. This factor represents Type-II error in statistical nomenclature. This error positioning in the confusion matrix also depends on the choice of the null hypothesis.

Let’s try to understand TP, FP, FN, TN with an example of pregnancy analogy



Precision

- Precision is the ratio of true positives and total positives predicted:

$$P = \frac{TP}{TP+FP} = \frac{\text{Cancer patients correctly identified}}{\text{Cancer patients correctly identified+incorrectly labelled cancer patients as non-cancerous}}$$

- $0<P<1$
- The precision metric focuses on Type-I errors(FP). A Type-I error occurs when we reject a true null Hypothesis(Ho). So, in this case, Type-I error is incorrectly labeling cancer patients as non-cancerous.

- A precision score towards 1 will signify that your model didn’t miss any true positives, and is able to classify well between correct and incorrect labeling of cancer patients. What it cannot measure is the existence of Type-II error, which is false negatives – cases when a non-cancerous patient is identified as cancerous.
- A low precision score (<0.5) means your classifier has a high number of false positives which can be an outcome of imbalanced class or untuned model hyperparameters.
- If FP is 0, so the condition is perfect for a 100% precise model on a given hyperparameter setting. In this setting, no type-I error is reported.

Recall/Sensitivity/Hit-Rate

- A Recall is essentially the ratio of true positives to all the positives in ground truth.

$$R = \frac{TP}{TP+FN} = \frac{\text{Cancer patients correctly identified}}{\text{Cancer patients correctly identified+incorrectly labelled non-cancer patients as cancerous}}$$

- 0<R<1
- The recall metric focuses on type-II errors(FN). A type-II error occurs when we accept a false null hypothesis(Ho). So, in this case, type-II error is incorrectly labeling non-cancerous patients as cancerous.

- Recall towards 1 will signify that your model didn’t miss any true positives, and is able to classify well between correctly and incorrectly labeling of cancer patients.
- What it cannot measure is the existence of type-I error which is false positives i.e. the cases when a cancerous patient is identified as non-cancerous.
- A low recall score (<0.5) means your classifier has a high number of false negatives which can be an outcome of imbalanced class or untuned model hyperparameters.
- The major highlight of the above two metrics is that both can only be used in specific scenarios since both of them identify only one set of errors.

When to use Precision and Recall?

- From the above definitions of Precision and Recall, we can say that recall determines the performance of a classifier with respect to a false negative, whereas precision gives information about the performance of a classifier with respect to a false positive.
- So, if we want to minimize the false negative, then, Recall should be as near to 100%, and if we want to minimize the false positive, then precision should be close to 100% as possible.
- In simple words, if we maximize precision, it will minimize the FP errors, and if we maximize recall, it will minimize the FN error.

F1-score

- The F1-score metric uses a combination of precision and recall.
- It is calculated with the help of Precision and Recall.
- It is a type of single score that represents both Precision and Recall.
- So, the F1 Score can be calculated as the harmonic mean of both precision and Recall, assigning equal weight to each of them.

- The formula for calculating the F1 score is given below:

$$F_1 = \frac{2}{\frac{1}{precision} + \frac{1}{recall}}$$

$$F1 - score = 2 * \frac{precision * recall}{precision + recall}$$

- When to use F
- As F-score make use of both precision and recall, so it should be used if both of them are important for evaluation
- It presents a good balance between precision and recall and gives good results on imbalanced classification problems.

AUROC (Area under Receiver operating characteristics curve)

- Sometimes we need to visualize the performance of the classification model on charts; then, we can use the AUC-ROC curve.
- It is one of the popular and important metrics for evaluating the performance of the classification model.
- Firstly, let's understand ROC (Receiver Operating Characteristic curve) curve.
- ROC represents a graph to show the performance of a classification model at different threshold levels.

ROC curve

- ROC represents a graph to show the performance of a classification model at different threshold levels.
- The curve is plotted between two parameters, which are true positive rates(TPR) and false positive rates(FPR).
  - **True Positive Rate (TPR) (Sensitivity)**
  - **False Positive Rate (FPR) (1-Specificity)**
- **TPR is a synonym for Recall, hence can be calculated as:**

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

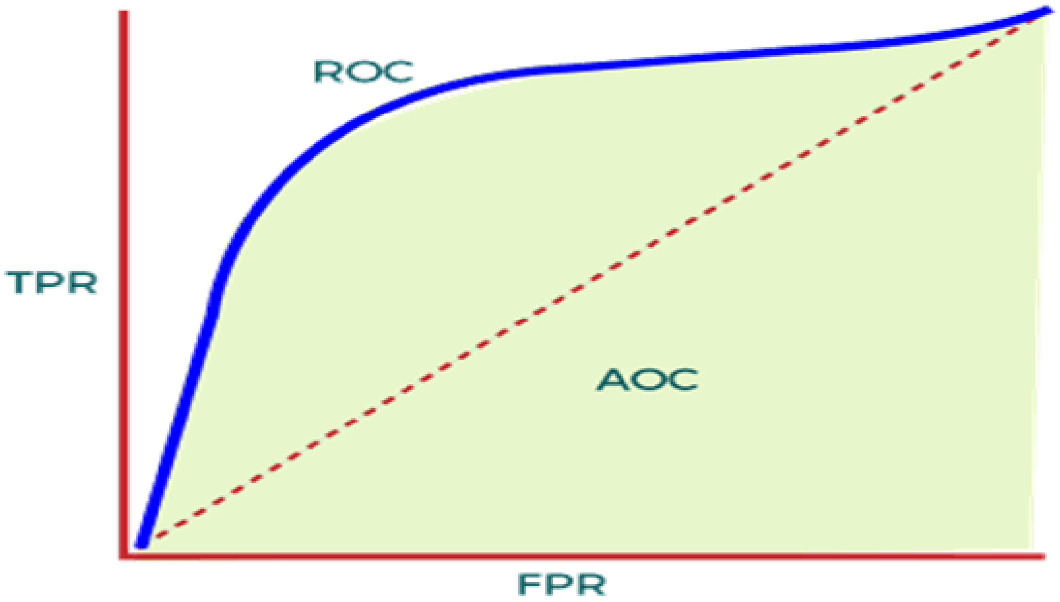
- FPR or False Positive Rate can be calculated as:

$$TPR = \frac{FP}{FP + TN}$$

- Intuitively TPR/recall corresponds to the proportion of positive data points that are correctly considered as positive, with respect to all positive data points. In other words, the higher the TPR, the fewer positive data points we will miss.
- Intuitively FPR/fallout corresponds to the proportion of negative data points that are mistakenly considered as positive, with respect to all negative data points. In other words, the higher the FPR, the more negative data points we will misclassify.

- To combine the FPR and the TPR into a single metric, we first compute the two former metrics with many different thresholds for the classifier, then plot them on a single graph. The resulting curve is called the ROC curve, and the metric we consider is the area under this curve, which we call Area under the ROC curve (AUC).
- As its name suggests, AUC calculates the two-dimensional area under the entire ROC curve, as shown in next slide.
- ROC curves aren’t a good choice when your problem has a huge class imbalance.

- AUC calculates the performance across all the thresholds and provides an aggregate measure. The value of AUC ranges from 0 to 1. It means a model with 100% wrong prediction will have an AUC of 0.0, whereas models with 100% correct predictions will have an AUC of 1.0.



Frequently Asked Questions

- Q1. What are the classification metrics?
- Ans: Classification metrics are evaluation measures used to assess the performance of a classification model. Common metrics include accuracy (proportion of correct predictions), precision (true positives over total predicted positives), recall (true positives over total actual positives), F1 score (harmonic mean of precision and recall), and area under the receiver operating characteristic curve (AUC-ROC).

Conclusion

- Understanding how well a machine learning model will perform on unseen data is the main purpose behind working with these evaluation metrics. Metrics like accuracy, precision, recall are good ways to evaluate classification models for balanced datasets, but if the data is imbalanced then other methods like ROC/AUC perform better in evaluating the model performance.