



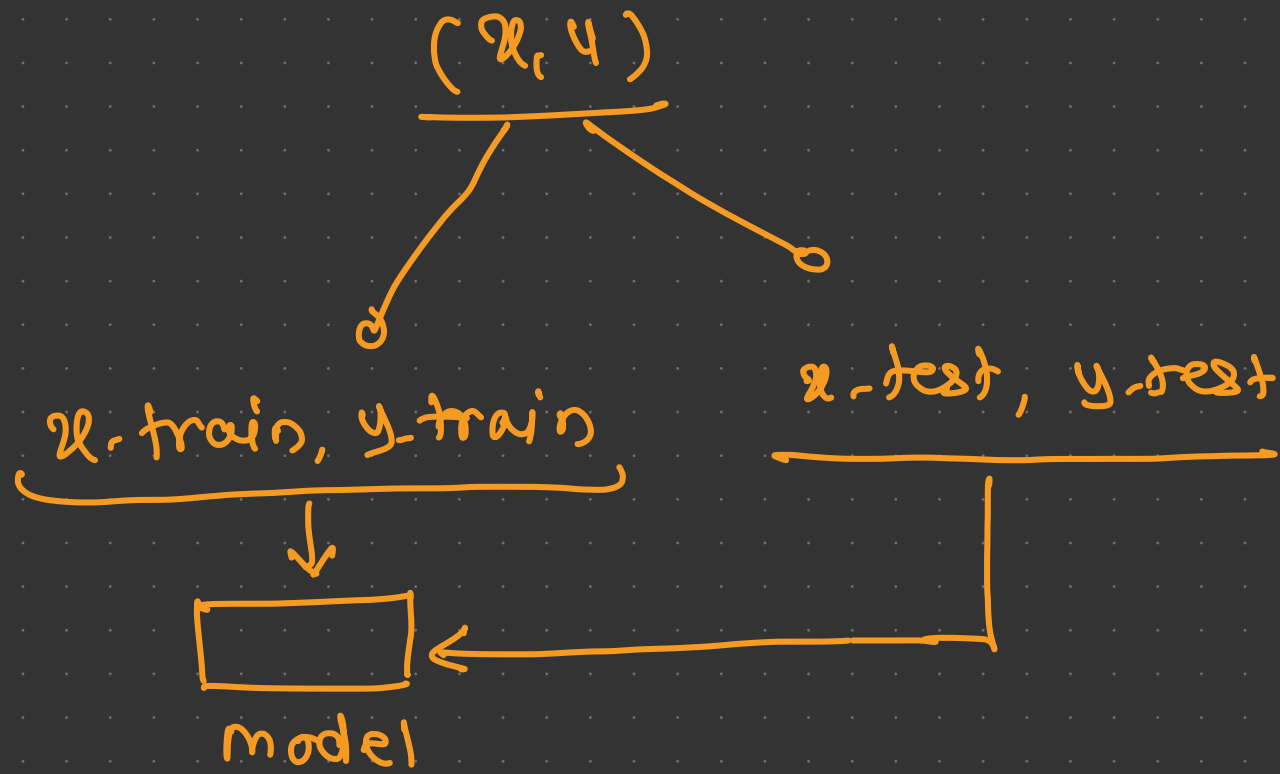
Model Evaluation

Regression

Regression Model Evaluation Metrics



- For evaluation of regression model, following metrics are used
 - MAE
 - MSE
 - RMSE
 - R²
 - Adjusted R²



Mean Absolute Error (MAE)

$$\text{Error} = (y - \hat{y}), \quad \text{MAE} = \frac{\sum |y - \hat{y}|}{n}$$



- The MAE measures the average magnitude of the errors in a set of forecasts, without considering their direction → absolute value
- It measures accuracy for continuous variables
- The MAE is the average over the verification sample of the absolute values of the differences between forecast and the corresponding observation
- The MAE is a linear score which means that all the individual differences are weighted equally in the average

$$\text{Error} = \text{calculated} - \text{observed}$$

Mean Squared Error (MSE)

$$\text{Error} = (y - \hat{y})$$

$$\text{MSE} = \frac{\sum (y - \hat{y})^2}{n}$$



- In statistics, the mean squared error (MSE) or mean squared deviation (MSD) of an estimator (of a procedure for estimating an unobserved quantity) measures the average of the squares of the error
- That is, the average squared difference between the estimated values and the actual value
- MSE is a risk function, corresponding to the expected value of the squared error loss
- The fact that MSE is almost always strictly positive (and not zero) is because of randomness or because the estimator does not account for information that could produce a more accurate estimate
- The MSE is a measure of the quality of an estimator
- As it is derived from the square of Euclidean distance, it is always a positive value with the error decreasing as the error approaches zero



Root Mean Squared Error (RMSE)

- RMSE is the most popular evaluation metric used in regression problems
- It follows an assumption that error are unbiased and follow a normal distribution
- Here are the key points to consider on RMSE:
 - The power of 'square root' empowers this metric to show large number deviations
 - The 'squared' nature of this metric helps to deliver more robust results which prevents cancelling the positive and negative error values
- It avoids the use of absolute error values which is highly undesirable in mathematical calculations
- When we have more samples, reconstructing the error distribution using RMSE is considered to be more reliable
- RMSE is highly affected by outlier values. Hence, make sure you've removed outliers from your data set prior to using this metric.
- As compared to mean absolute error, RMSE gives higher weightage and punishes large errors



R-Squared (R^2)

- We learned that when the RMSE decreases, the model's performance will improve
- But these values alone are not intuitive
- When we talk about the RMSE metrics, we do not have a benchmark to compare
- This is where we can use R-Squared metric
- In other words how good our regression model as compared to a very simple model that just predicts the mean value of target from the train set as predictions

↪ using $\text{mean}(\hat{y})$

$$[\text{0}, +?]$$



Adjusted R-Squared

- A model performing equal to baseline would give R-Squared as 0
- Better the model, higher the r^2 value
- The best model with all correct predictions would give R-Squared as 1
- However, on adding **new features** to the model, the R-Squared value either increases or remains the same
↳ independent
- R-Squared does not penalize for adding features that add no value to the model
- So an improved version over the R-Squared is the adjusted R-Squared

$$\bar{R}^2 = 1 - (1 - R^2) \left[\frac{n-1}{n-(k+1)} \right]$$

- k : number of features
- n : number of samples



Model Evaluation Classification



Classification Model Evaluation Metrics

- For evaluation of classification model, following metrics are used
 - Confusion Matrix
 - F1 Score
 - AuC-Roc

Confusion Matrix



- A confusion matrix is an N X N matrix, where N is the number of classes being predicted
- The confusion matrix provides more insight into not only the performance of a predictive model, but also which classes are being predicted correctly, which incorrectly, and what type of errors are being made

		Predicted Class		
		Positive	Negative	
Actual Class	Positive	True Positive (TP)	False Negative (FN) Type II Error	Sensitivity $\frac{TP}{(TP + FN)}$
	Negative	False Positive (FP) Type I Error	True Negative (TN)	Specificity $\frac{TN}{(TN + FP)}$
		Precision $\frac{TP}{(TP + FP)}$	Negative Predictive Value $\frac{TN}{(TN + FN)}$	Accuracy $\frac{TP + TN}{(TP + TN + FP + FN)}$

TP vs FP vs TN vs FN



Cat



Cat



Cat



Cat



No Cat



No Cat



No Cat



Cat



Cat

Accuracy



Cat



Cat



Cat



Cat



No Cat



No Cat



No Cat



Cat



Cat

How many we got right ?

Precision



- Precision talks about how precise/ accurate your model is out of those predicted positive, how many of them are actual positive
- Precision is a good measure to determine, when the costs of False Positive is high
- For instance, in email spam detection, a false positive means that an email that is non-spam (actual negative) has been identified as spam (predicted spam). The email user might lose important emails if the precision is not high for the spam detection model.

$$\begin{aligned}\text{Precision} &= \frac{\text{True Positive}}{\text{True Positive} + \text{False Positive}} \\ &= \frac{\text{True Positive}}{\text{Total Predicted Positive}}\end{aligned}$$

Precision



Cat



Cat



Cat



Cat



No Cat



No Cat



No Cat



Cat



Cat

Out of all Cat predictions how many we got right ?



Recall

- Recall actually calculates how many of the Actual Positives our model capture through labelling it as Positive (True Positive)
- Applying the same understanding, we know that Recall shall be the model metric we use to select our best model when there is a high cost associated with False Negative
- For instance, in fraud detection or sick patient detection, if a fraudulent transaction (Actual Positive) is predicted as non-fraudulent (Predicted Negative), the consequence can be very bad for the bank
- Similarly, in sick patient detection, if a sick patient (Actual Positive) goes through the test and predicted as not sick (Predicted Negative), the cost associated with False Negative will be extremely high if the sickness is contagious

$$\begin{aligned}\text{Recall} &= \frac{\text{True Positive}}{\text{True Positive} + \text{False Negative}} \\ &= \frac{\text{True Positive}}{\text{Total Actual Positive}}\end{aligned}$$

Recall



Cat



Cat



Cat



Cat



No Cat



No Cat



No Cat



Cat



Cat

Out of all Cat truth how many we got right ?



F1 Score

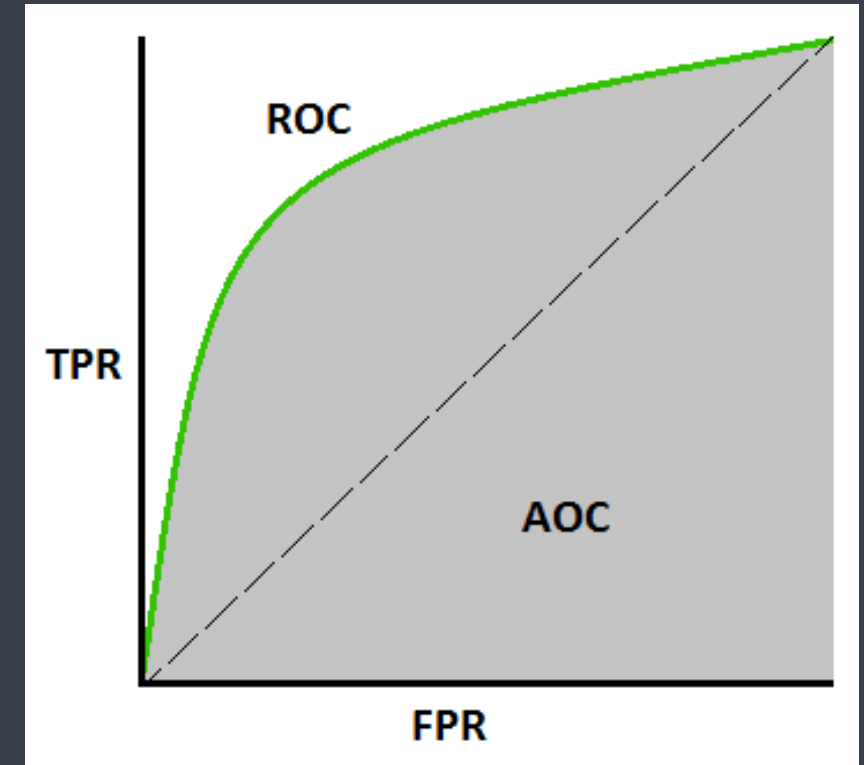
- The F1 score is the harmonic mean of the precision and recall
- The highest possible value of an F-score is 1.0, indicating perfect precision and recall, and the lowest possible value is 0, if either the precision or the recall is zero
- The F1 score is also known as the Sørensen–Dice coefficient or Dice similarity coefficient (DSC)

$$F1 = 2 \times \frac{Precision * Recall}{Precision + Recall}$$



Receiver Operating Characteristic (ROC)

- ROC curve is a metric that assesses the model ability to distinguish between binary classes
- It is created by plotting the true positive rate (TPR) against the false positive rate (FPR) at various threshold settings
- The TPR is also known as sensitivity, recall or probability of detection in machine learning
- The FPR is also known as the probability of false alarm and can be calculated as $1 - \text{specificity}$
- Points above the diagonal line represent good classification (better than random)
- The model performance improves if it becomes skewed towards the upper left corner



Receiver Operating Characteristic (ROC)



TPR (True Positive Rate) / Recall / Sensitivity

$$\text{TPR / Recall / Sensitivity} = \frac{\text{TP}}{\text{TP} + \text{FN}}$$

Image 3

Specificity

$$\text{Specificity} = \frac{\text{TN}}{\text{TN} + \text{FP}}$$

Image 4

FPR

$$\text{FPR} = 1 - \text{Specificity}$$

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