Data Analysis(4)

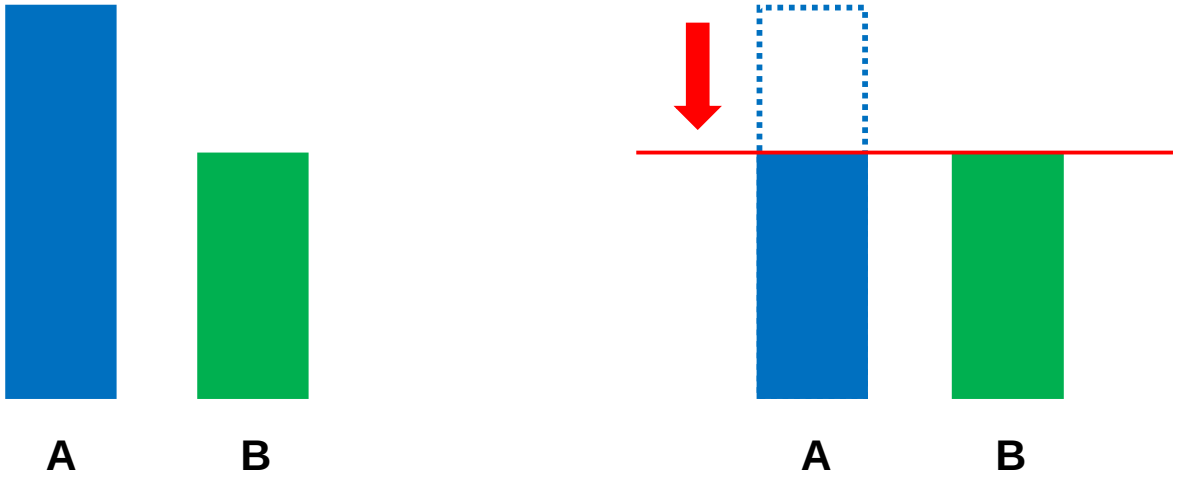
Dept. of Mechanical System Design Engineering,
Seoul National University of Science and Technology

Prof. Ju Yeon Lee
(jylee@seoultech.ac.kr)

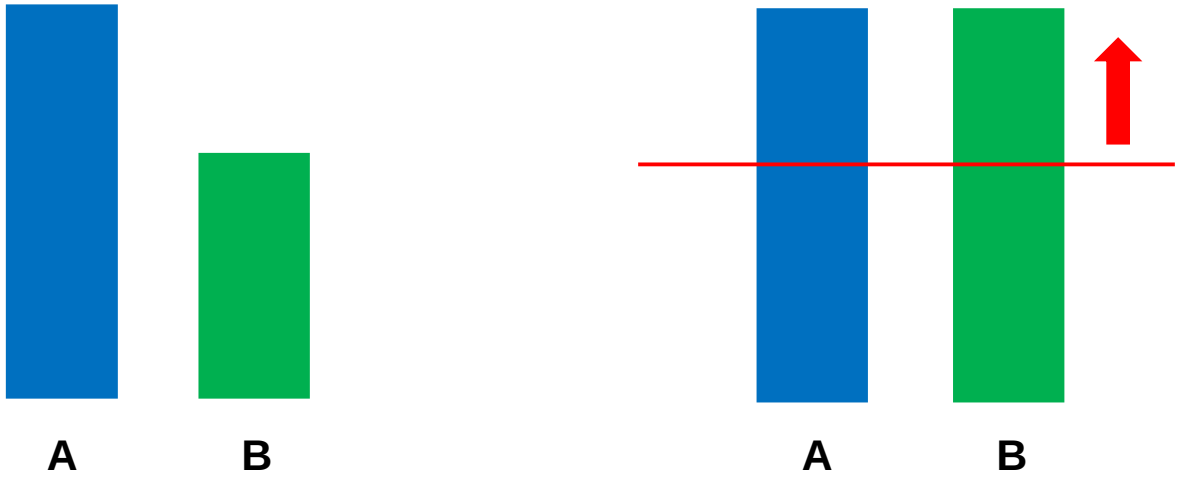
Review

Imbalanced sampling

1. Under/Down Sampling

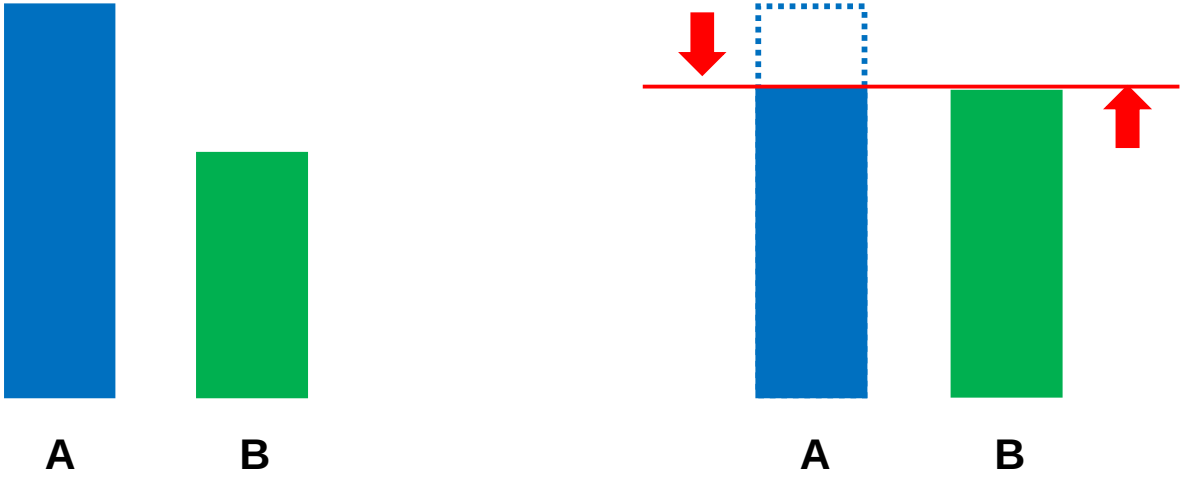


2. Over/Up Sampling



Imbalanced sampling

3. Combination Sampling : Under + Over

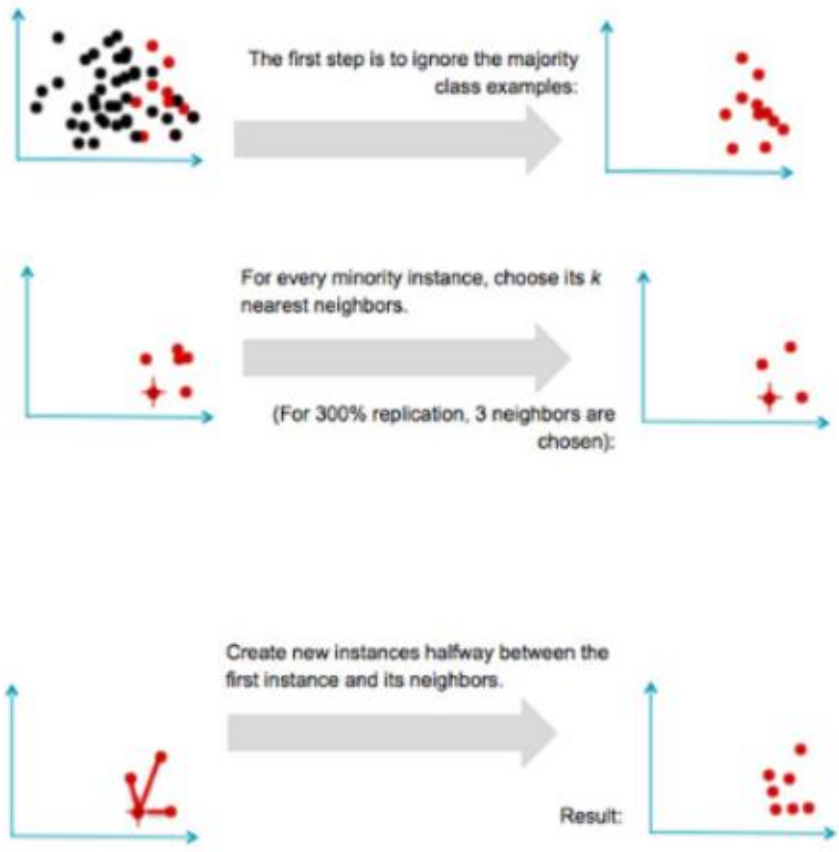


SMOTE

SMOTE(Synthetic Minority Over-Sampling Technique) :

generate new data between neighboring minority classes from random minority class data

For numerical features



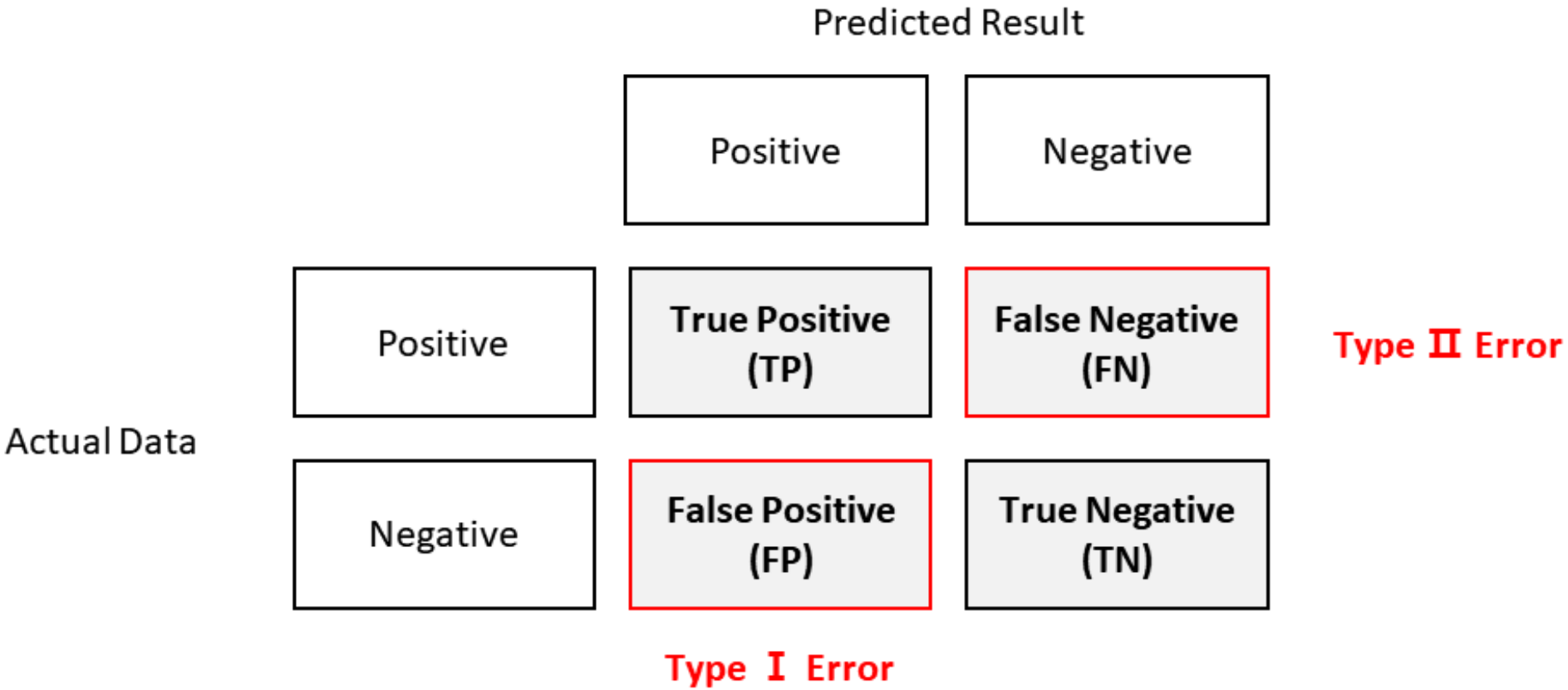
SMOTENC(Synthetic Minority Over-Sampling Technique for Nominal and Continuous) :

For dataset containing numerical and categorical features

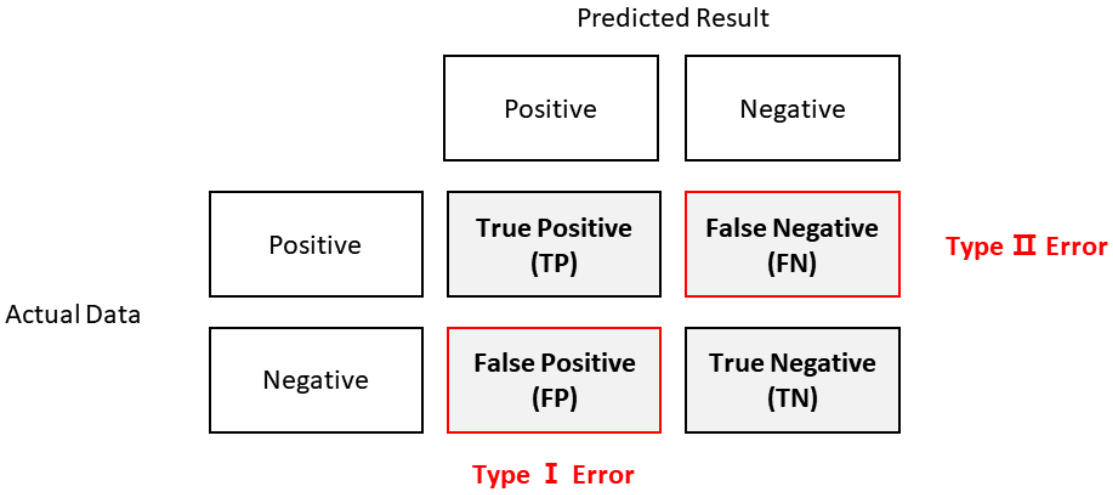
However, it is not designed to work with only categorical features

Confusion Matrix

Confusion Matrix (혼합 행렬)



Accuracy (정확도)

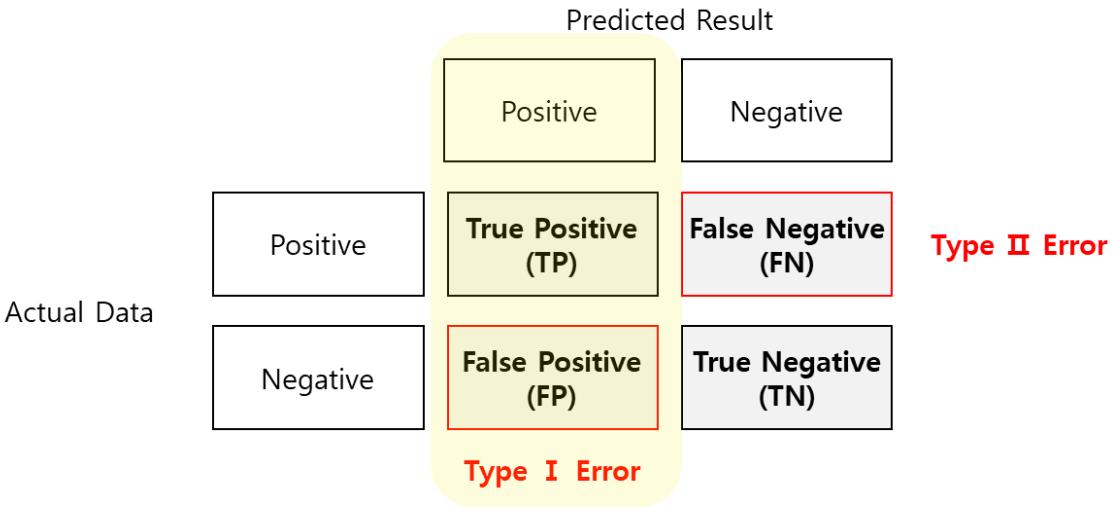


Accuracy =
(TP+TN) / (TP+TN+FP+FN)

Accurate prediction ratio to total

What about unbalanced classes?

Precision (정밀도)

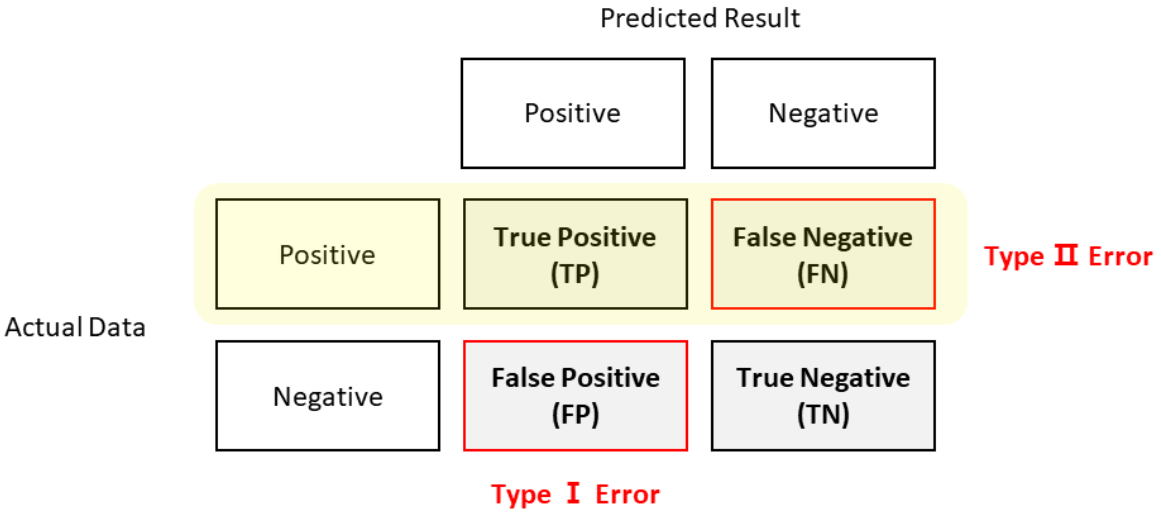


Precision =
TP / (TP + FP)

Actual Positive Ratio in Positive Prediction
= Accuracy of positive prediction model

What if you predict that a defective product is a good product?

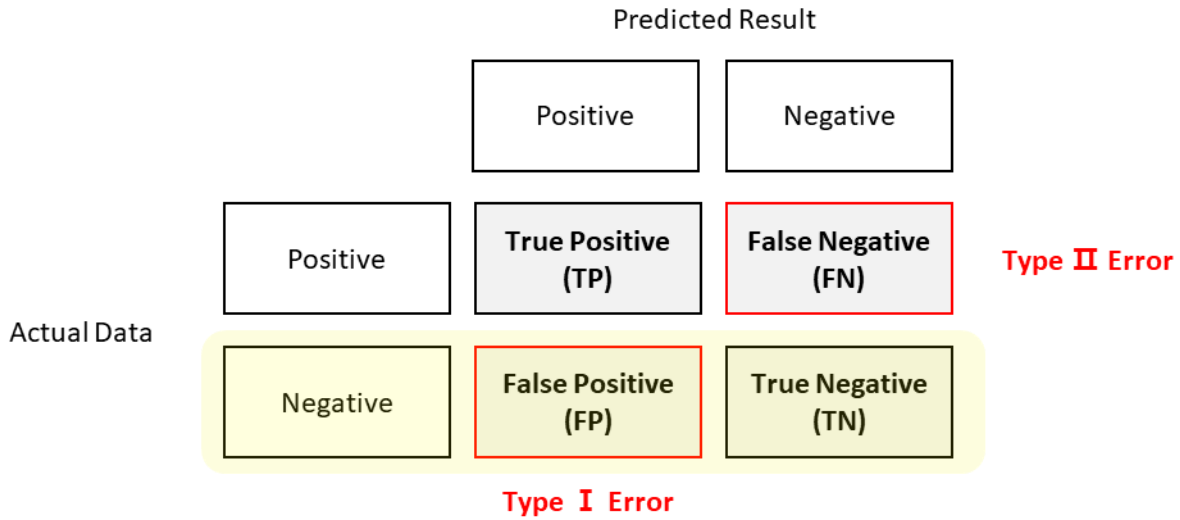
Recall (True Positive Rate, 재현율 / Sensitivity, 민감도)



Recall (TPR) =
TP / (TP + FN)

The percentage of actual positive data predicted to be positive
= Accuracy of positive data prediction

Specificity (특이성, True Negative Rate)



Specificity (TNR) = TN / (TN + FP)

Proportion of predicting the actual negative data as negative
= Accuracy of negative data prediction

FPR = 1−Specificity = FP / (TN + FP)

Percentage of true negatives that are incorrectly predicted as positives

ROC Curve/AUC

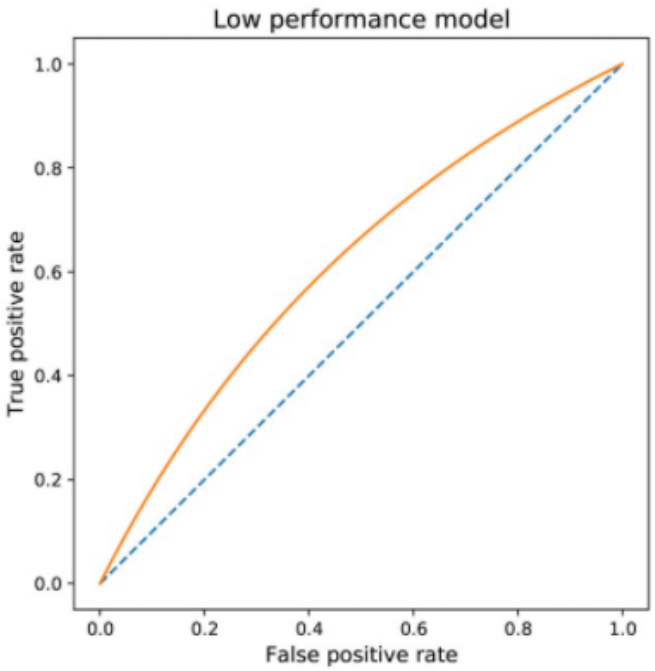
ROC Curve :

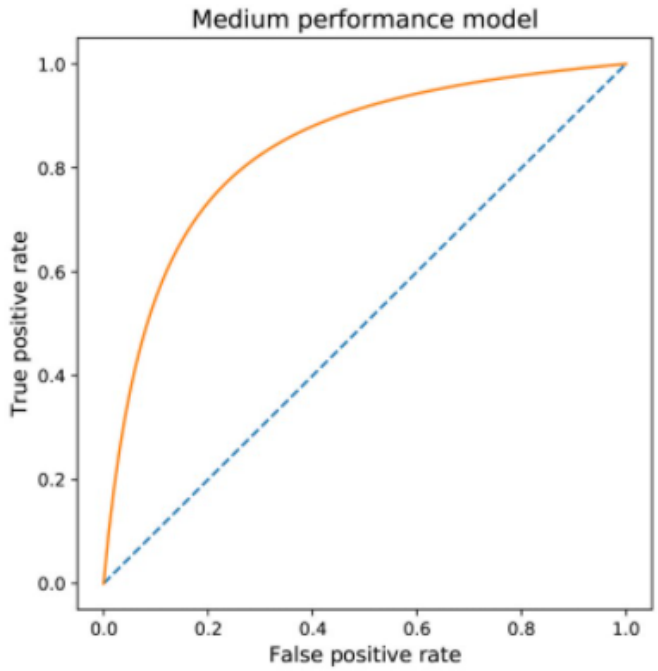
A curve showing **how the True Positive Rate (TPR) changes when the False Positive Rate (FPR) changes**

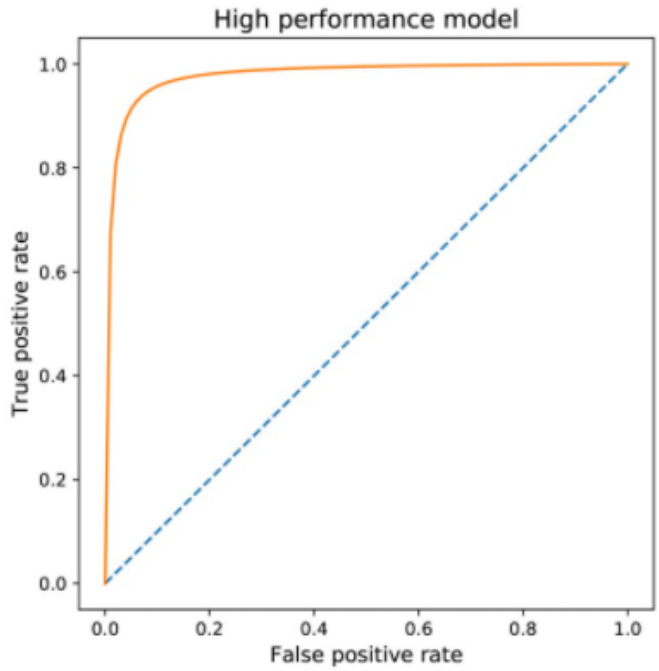
AUC (Area Under Curve) :

Area value under the curve,

the closer to 1, the better, 0.5 for a diagonal straight line





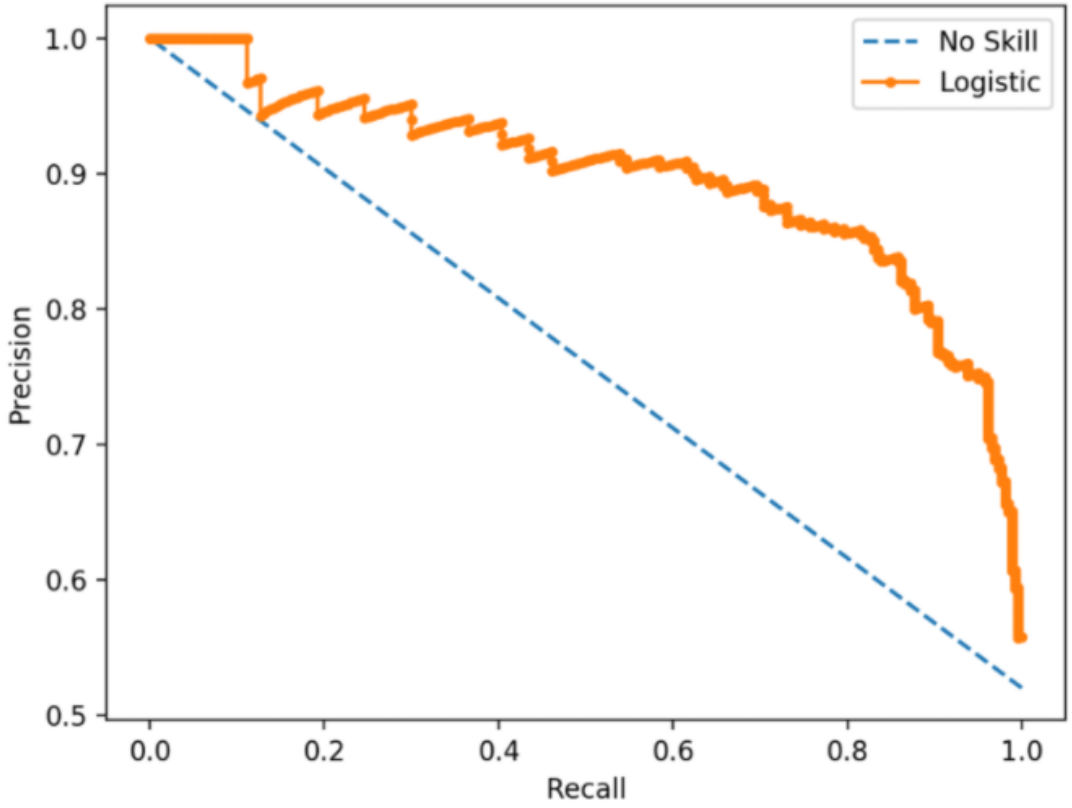


Precision-Recall Curve

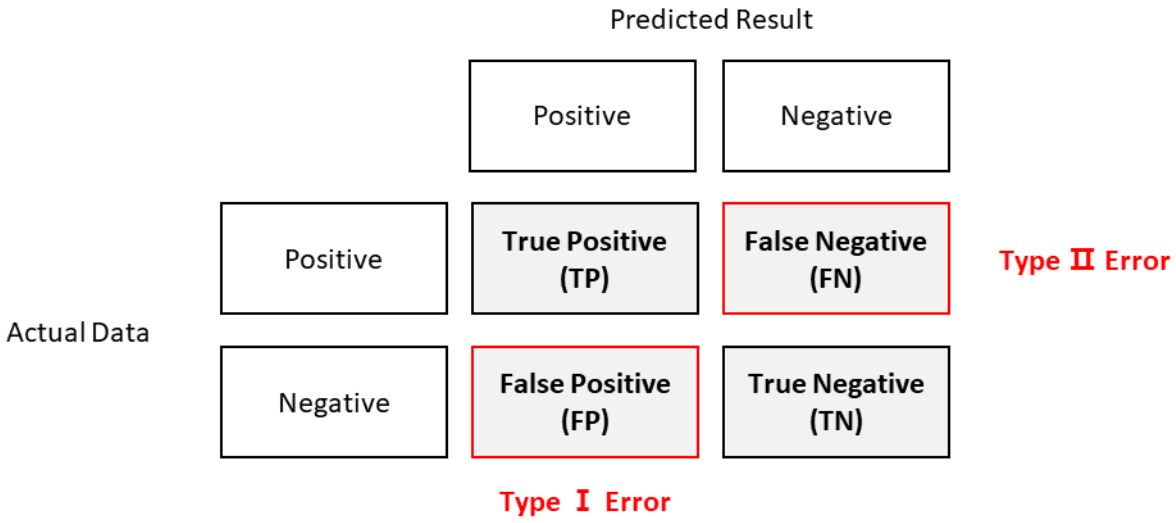
Precision-Recall Curve :

A plot of the precision (y-axis) and the recall (x-axis) for different thresholds

Trade-off between Precision and Recall



F1 Score



F1 Score =
2 * ((Precision * Recall) / (Precision + Recall))

Precision and Recall integrated into a single value through the harmonic average of two performance indicators


Thank you
Q & A