DEEP LEARNING WITH KERAS WORKSHOP

MUHAMMAD RAJABINASAB @ TARBIAT MODARES UNIVERSITY

Chapter 6 :

Model Evaluation

INTRODUCTION

- In this chapter, we will learn about some different evaluation techniques other than accuracy
- The easiest way to evaluate a model is through its accuracy
- In real-world scenarios, accuracy is not always sufficient
- In this chapter, we will explore many other evaluation techniques

ACCURACY

- The simplest metric for model evaluation is accuracy
- Accuracy is the fraction of predictions that our model gets right

Accuracy =

(Number of correct predictions) / (Total number of predictions)

NULL ACCURACY

- Null accuracy is the accuracy that can be achieved by predicting the most frequent class

Null accuracy =

(Total number of instances of the frequently occurring class) / (Total number of instances)

EXERCISE 6.01

- We have a dataset documenting whether a hurricane has been observed in the Pacific Ocean
- It has two columns, Date and hurricane
- The Date column indicates the date of the observation, while the hurricane column indicates whether there was a hurricane on that date

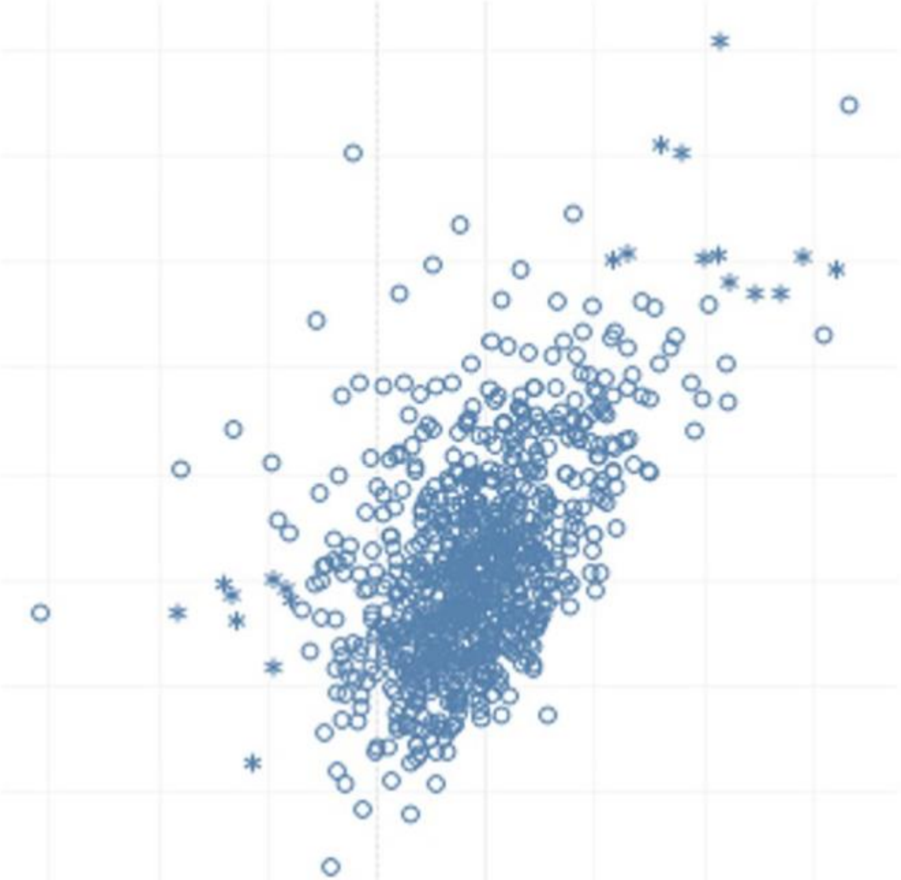
ADVANTAGES AND LIMITATIONS OF ACCURACY

- Easy to use
- Popular compared to other techniques
- Good for general comparison
- No representation of response variable distribution
- Type 1(FP) and type 2 (FN) errors

IMBALANCED DATASETS

- Imbalanced datasets are a distinct case for classification problems where the class distribution varies between the classes
- The null accuracy of an imbalanced dataset is very high
- E.g. credit card fraud detection, HIV test results, etc.

IMBALANCED DATASETS



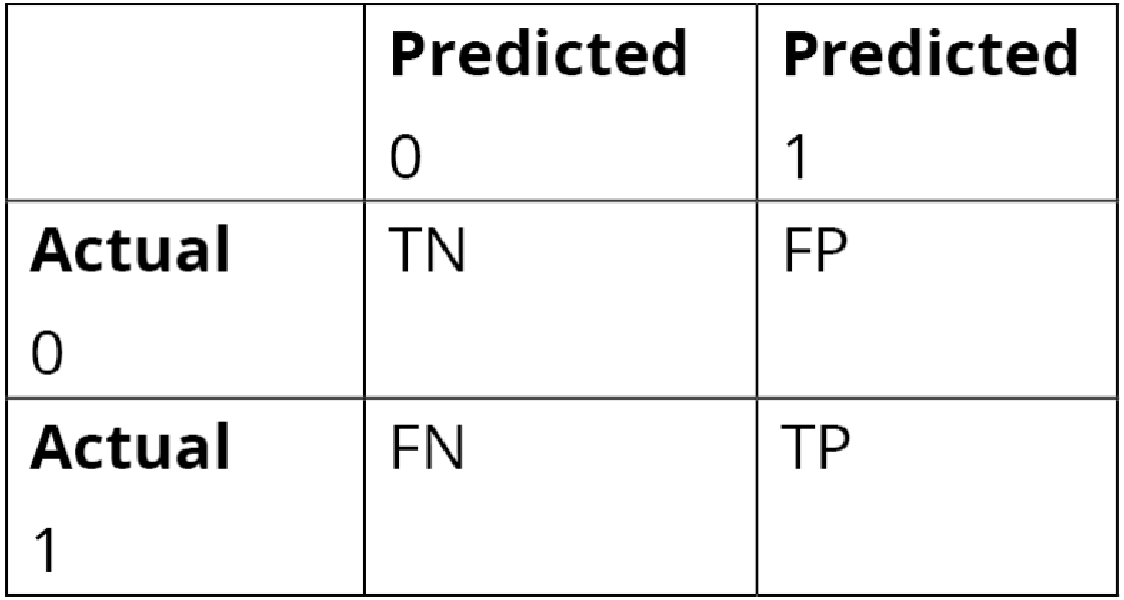
WORKING WITH IMBALANCED DATASETS

- Sampling techniques
- Modifying model evaluation techniques

CONFUSION MATRIX

- A confusion matrix describes the performance of the classification model
- The confusion matrix is the basis of many evaluation techniques
- Although its easier to represent it as a part of a 2-class classification problem for educational purposes, it can be extended to n-class classification problems as well

CONFUSION MATRIX



	Predicted 0	**Predicted** 1
Actual 0	TN	FP
Actual 1	FN	TP

SENSITIVITY (RECALL)

- This is the number of positive predictions divided by the total actual number of positives

$$\textit{Sensitivity} = TP\ /\ (TP+FN)$$

SPECIFICITY

- This is the number of negative predictions divided by the total number of actual negatives

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

PRECISION

- This is the true positive prediction divided by the total number of positive predictions

$$Precision = TP / (TP+FP)$$

FALSE POSITIVE RATE (FPR)

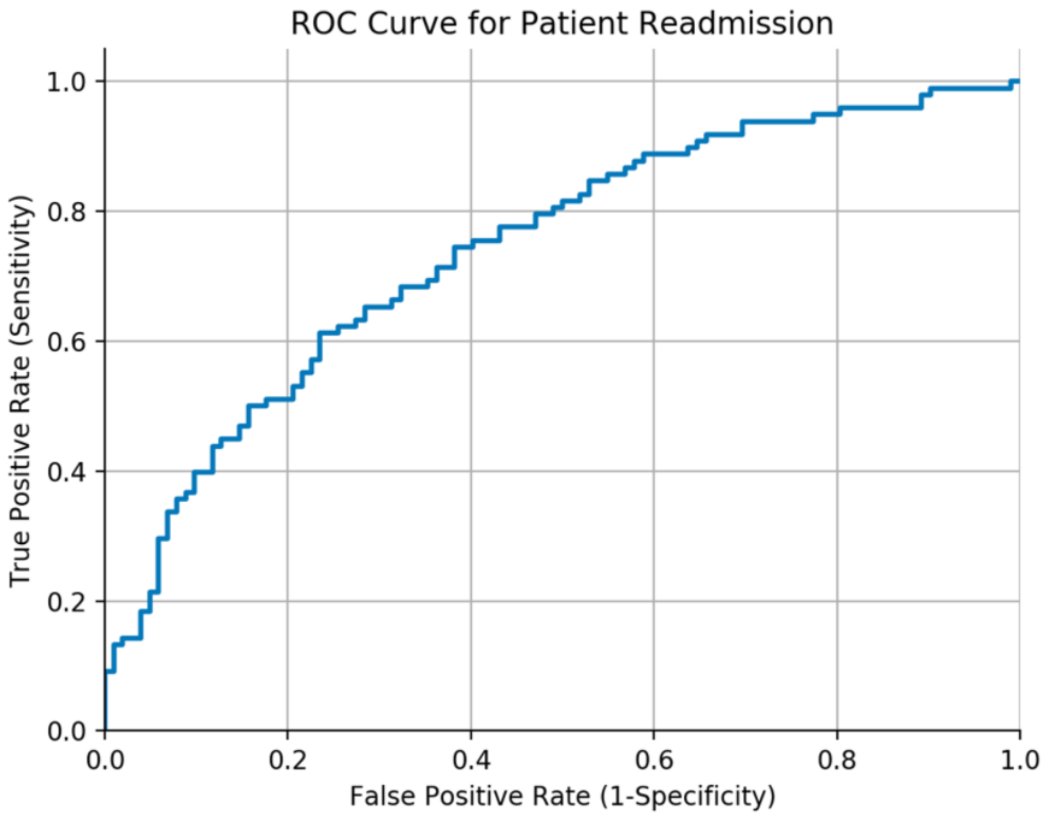
- The FPR is calculated as the ratio between the number of false-positive events and the total number of actual negative events

$$\textit{False positive rate} = 1 - \textit{Specificity} = FP / (FP+TN)$$

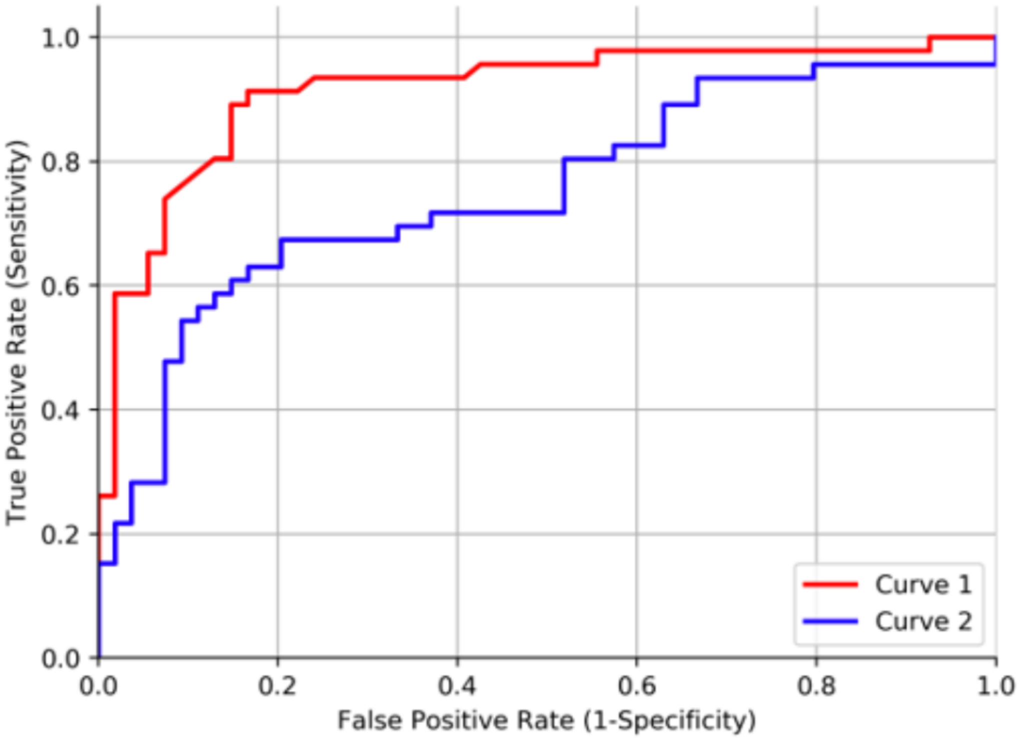
RECEIVER OPERATING CHARACTERISTIC (ROC) CURVE

- Another important way to evaluate a classification model is by using a ROC curve
- A ROC curve is a plot between the true positive rate (sensitivity) and the false positive rate (1 - specificity)
- To decide which ROC curve is the best among multiple curves, we need to look at the empty space on the upper left of the curve—the smaller the space, the better the result

RECEIVER OPERATING CHARACTERISTIC (ROC) CURVE



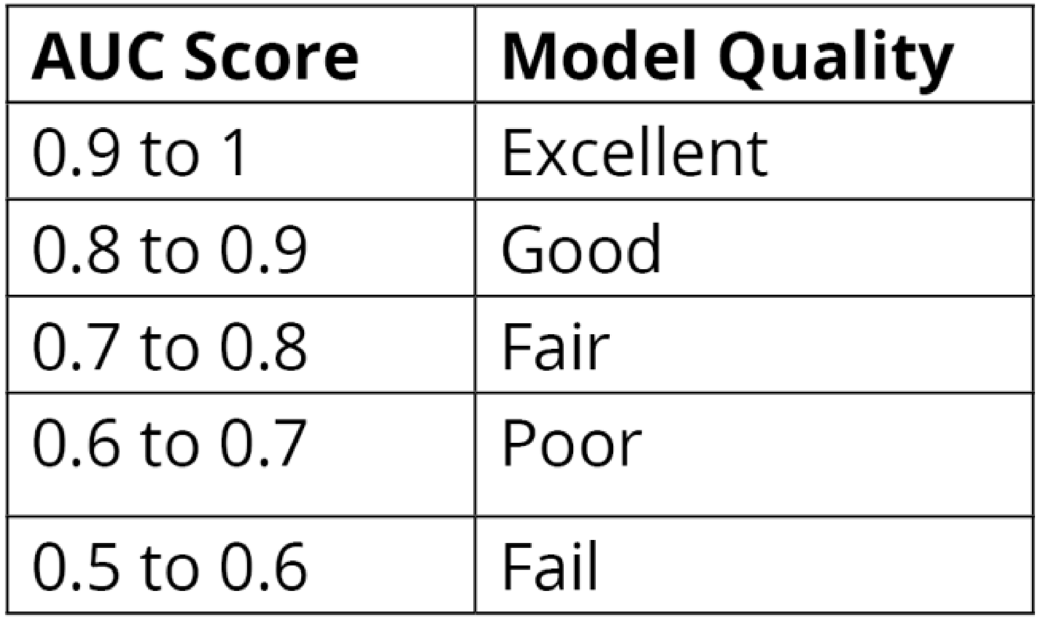
RECEIVER OPERATING CHARACTERISTIC (ROC) CURVE



AREA UNDER CURVE (AUC)

- This is the area under the ROC curve
- Sometimes, AUC is also written as AUROC, meaning the area under the ROC curve
- The larger the area under the ROC, the better, and the bigger the AUC score, the better

AREA UNDER CURVE (AUC)



AUC Score	Model Quality
0.9 to 1	Excellent
0.8 to 0.9	Good
0.7 to 0.8	Fair
0.6 to 0.7	Poor
0.5 to 0.6	Fail

EXERCISE 6.02

- The dataset that we will be using in this exercise consists of data that's been collected from heavy Scania trucks in everyday usage that have failed in some way
- The system in focus is the Air pressure system (APS), which generates pressurized air that is utilized in various functions in a truck, such as braking and gear changes
- The positive class in the dataset represents component failures for a specific component in the APS, while the negative class represents failures for components not related to the APS

ACTIVITY 6.01: COMPUTING THE ACCURACY AND NULL ACCURACY OF A NEURAL NETWORK WHEN WE CHANGE THE TRAIN/TEST SPLIT

- In this activity, we will see that our null accuracy and accuracy will be affected by changing the train/test split
- To implement this, the part of the code where the train/test split was defined has to be changed
- We will use the same dataset that we used in Exercise 6.02

EXERCISE 6.03

- In this exercise we will work on the dataset from exercise 6.02
- We will derive the sensitivity, specificity, precision, and false positive rate of the neural network model to evaluate the model's performance

ACTIVITY 6.02: CALCULATING THE ROC CURVE AND AUC SCORE

- In this activity, we will plot the ROC curve and calculate the AUC score of a mode
- We will work on the dataset from exercise 6.02

SUMMARY

- In this chapter, we learned why accuracy is not sufficient enough to evaluate a model's performance
- We learned about evaluation metrics other than accuracy
- We learned to implement these metrics to evaluate our keras model

THANK YOU FOR YOUR ATTENTION!