DATA ANALYTICS

Unit 2:Confusion matrices and Metrics

Mamatha.H.R
Department of Computer Science and Engineering

DATA ANALYTICS

Unit 2:Confusion matrices and Metrics

Mamatha H R
Department of Computer Science and Engineering

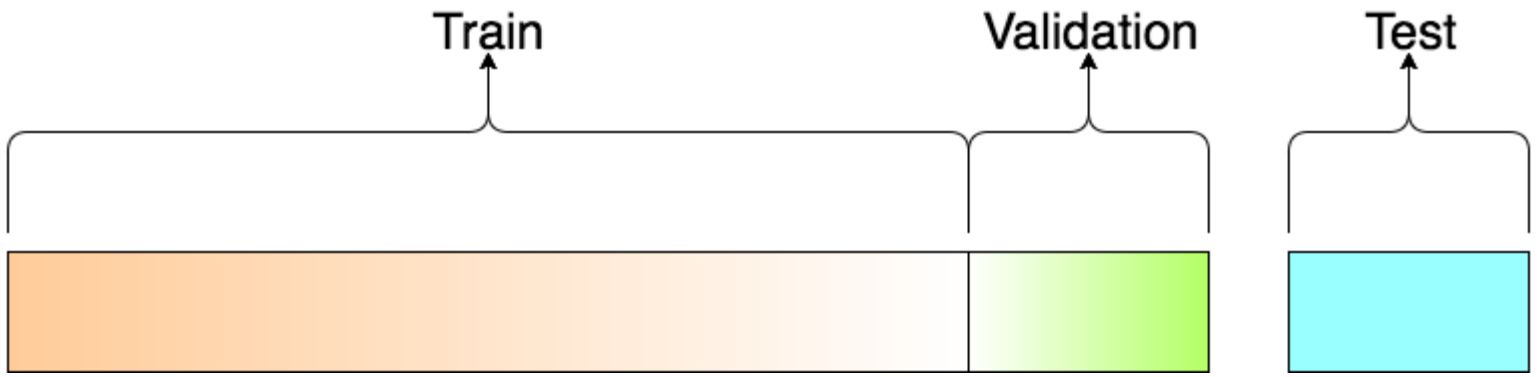
Train, Validation and Test Sets

Training Dataset: The sample of data used to fit the model. The actual dataset that we use to train the model .The model sees and learns from this data.

Validation Dataset: The sample of data used to provide an unbiased evaluation of a model fit on the training dataset while tuning model hyperparameters. The evaluation becomes more biased as skill on the validation dataset is incorporated into the model configuration.

Test Dataset: The sample of data used to provide an unbiased evaluation of a final model fit on the training dataset.

Dataset Split Ratio



A visualization of the splits

Data Split depends on 2 things.

First, the total number of samples in your data

second, on the actual model you are training.

Dataset Split Ratio

Some models need substantial data to train upon, so in this case you would optimize for the larger training sets.

Models with very few hyperparameters will be easy to validate and tune, so you can probably reduce the size of your validation set,

but if your model has many hyperparameters, you would want to have a large validation set as well.

Also, if you happen to have a model with no hyperparameters or ones that cannot be easily tuned, you probably don’t need a validation set too!

Confusion matrix

The confusion matrix is a metric that is often used to measure the performance of a classification algorithm.

In binary classifiers as with the spam filtering example, in which each email can be either spam or not spam.

The confusion matrix will be of the following form:

	Predicted: Real Email	Predicted: Spam Email
Actual: Real Email	True Negatives (TN)	False Positives (FP)
Actual: Spam Email	False Negatives (FN)	True Positives (TP)

Confusion matrix

The predicted classes are represented in the columns of the matrix, whereas the actual classes are in the rows of the matrix.

We then have four cases:

True positives (TP): the cases for which the classifier predicted 'spam' and the emails were actually spam.

True negatives (TN): the cases for which the classifier predicted 'not spam' and the emails were actually real.

False positives (FP): the cases for which the classifier predicted 'spam' but the emails were actually real.

False negatives (FN): the cases for which the classifier predicted 'not spam' but the emails were actually spam.

Classification Performance Metrics

Accuracy: Out of all the classes, how much we predicted correctly

$$\text{Accuracy} = \frac{\text{TP} + \text{TN}}{\text{TP} + \text{TN} + \text{FP} + \text{FN}}$$

The ability of the model to correctly classify positives and negatives are called sensitivity and specificity

$$\text{Sensitivity} = \frac{\text{True Positive (TP)}}{\text{True Positive (TP)} + \text{False Negative (FN)}}$$

where True Positive (TP) is the number of positives correctly classified as positives by the model and False Negative (TN) is positives misclassified as negative by the model. Sensitivity is also called as recall.

Classification Performance Metrics

Specificity can be calculated using the following equation:

$$\text{Specificity} = \frac{\text{True Negative (TN)}}{\text{True Negative (TN)} + \text{False Positive (FP)}}$$

where True Negative (TN) is number of the negatives correctly classified as negatives by the model and False Positive (FP) is number of negatives misclassified as positives by the model.

Precision measures the accuracy of positives classified by the model.

$$\text{Precision} = \frac{\text{True Positive (TP)}}{\text{True Positive (TP)} + \text{False Positive (FP)}}$$

Classification Performance Metrics

F Score (F Measure) is another measure used in binary logistic regression that combines both precision and recall and is given by:

$$F - Score = \frac{2 \times Precision \times Recall}{Precision + Recall}$$

Concordant and Discordant Pairs

- **Discordant Pairs.** A pair of positive and negative observations for which the model has no cut-off probability to classify both of them correctly are called discordant pairs.

- Divide the dataset into positives (y=1) and negatives (y=0).

- For a randomly chosen positive and negative, if the probability of positive (obtained using logistic regression model) is greater than probability of negative then such pairs are called concordant pairs.

Concordant and Discordant Pairs

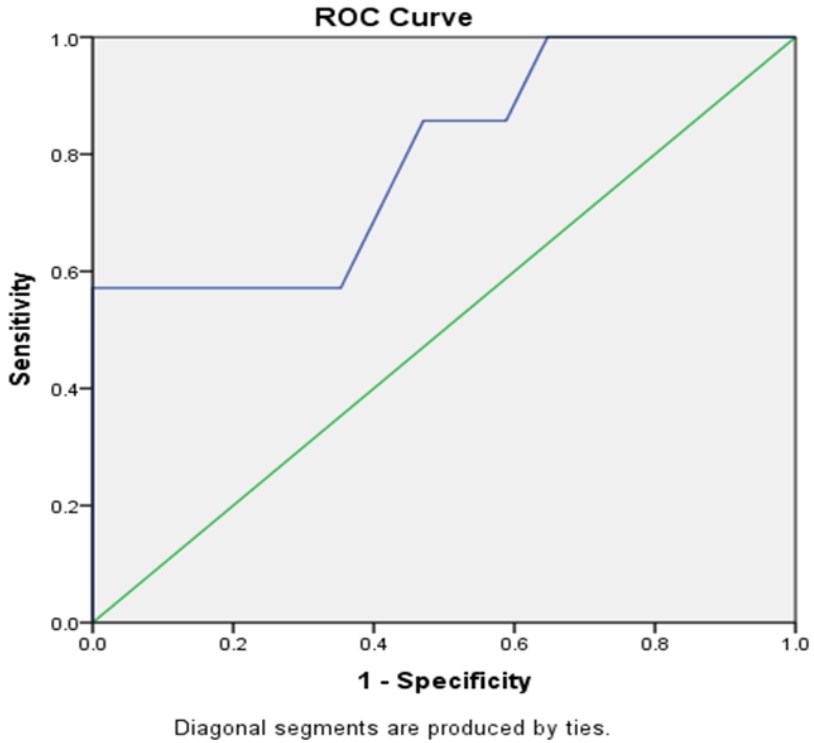
Concordant Pairs. A pair of positive and negative observations for which the model has a cut-off probability to classify both of them correctly are called concordant pairs.

For a randomly chosen positive and negative, if the probability of positive is less than probability of negative then such pairs are called discordant pairs.

Area under the ROC curve is the proportion of concordant pairs in the dataset.

Receiver Operating Characteristics (ROC) Curve

- ROC curve is a plot between sensitivity (true positive rate) in the vertical axis and 1 – specificity (false positive rate) in the horizontal axis.
- The higher the area under the ROC curve, the better the prediction ability.



Area Under ROC Curve (AUC)

- Area under the ROC (AUC) curve is interpreted as the probability that the model will rank a randomly chosen positive higher than randomly chosen negative.

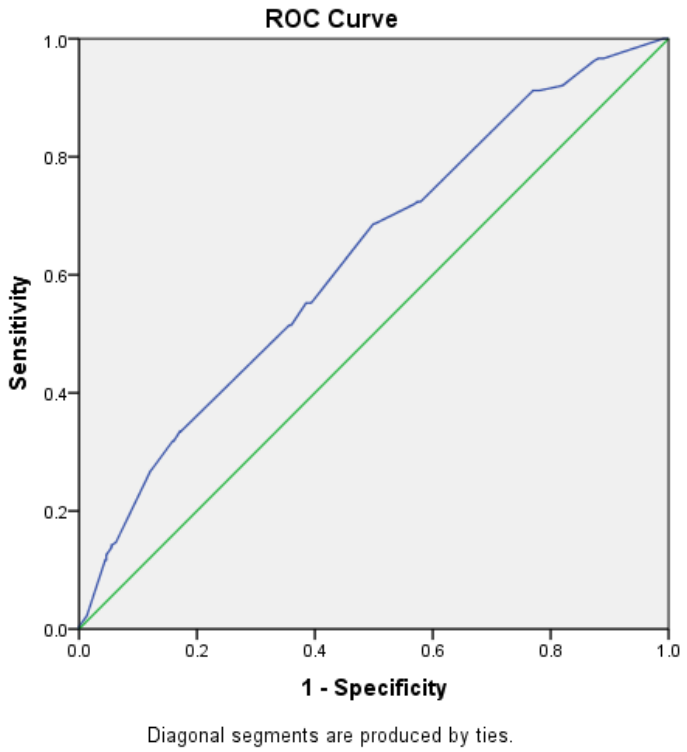
- If n1 is the number of positives (1s) and n2 is the number of negatives (0s), then the area under the ROC curve is the proportion of cases in all possible combinations of (n1, n2) such that n1 will have higher probability than n2.

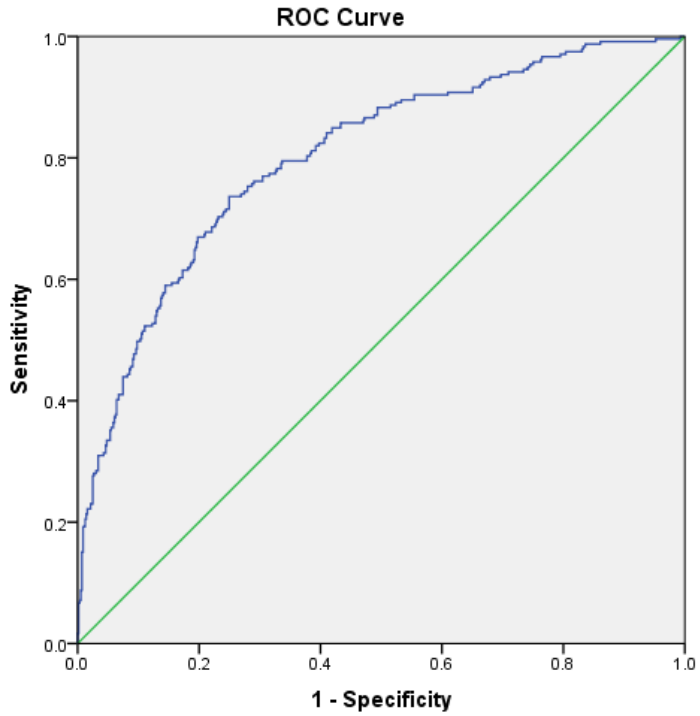
AUC = P (Random Positive Observation) > P(Random Negative Observation)

Area Under the ROC Curve (AUC) is a measure of the ability of the logistic regression model to discriminate positives and negatives correctly.

DATA ANALYTICS

Area Under ROC Curve (AUC)





AUC = 0.629

AUC = 0.801

ROC Curve

- General rule for acceptance of the model:

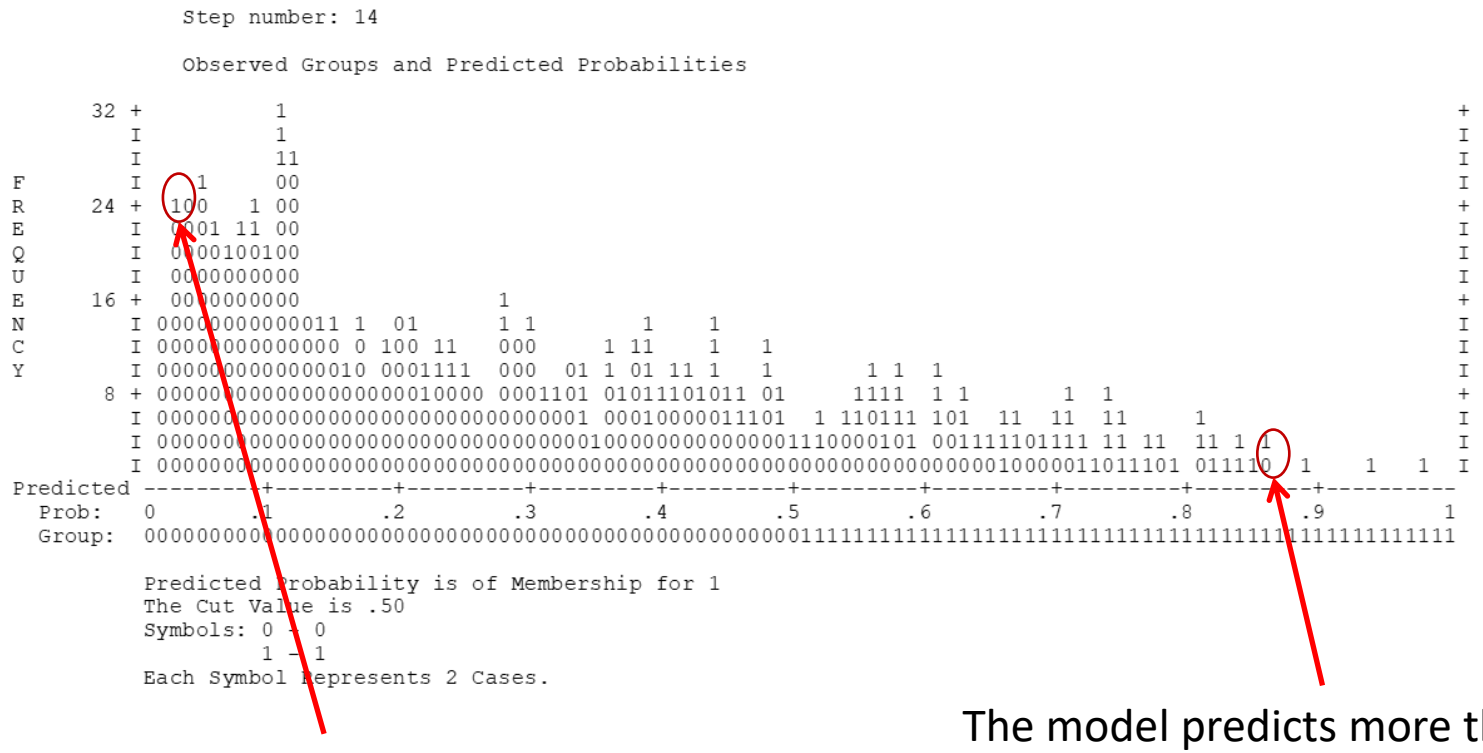
- If the area under ROC is:

  - 0.5 $\Rightarrow$ No discrimination
  - 0.7 $\leq$ ROC area < 0.8 $\Rightarrow$ Acceptable discrimination
  - 0.8 $\leq$ ROC area < 0.9 $\Rightarrow$ Excellent discrimination
  - ROC area $\geq$ 0.9 $\Rightarrow$ Outstanding discrimination

DATA ANALYTICS

Classification Plot for Selection of Cut-off Probability

- Classification plot is a plot between the probability on horizontal axis and predicted probability for each of the observations using LR model.



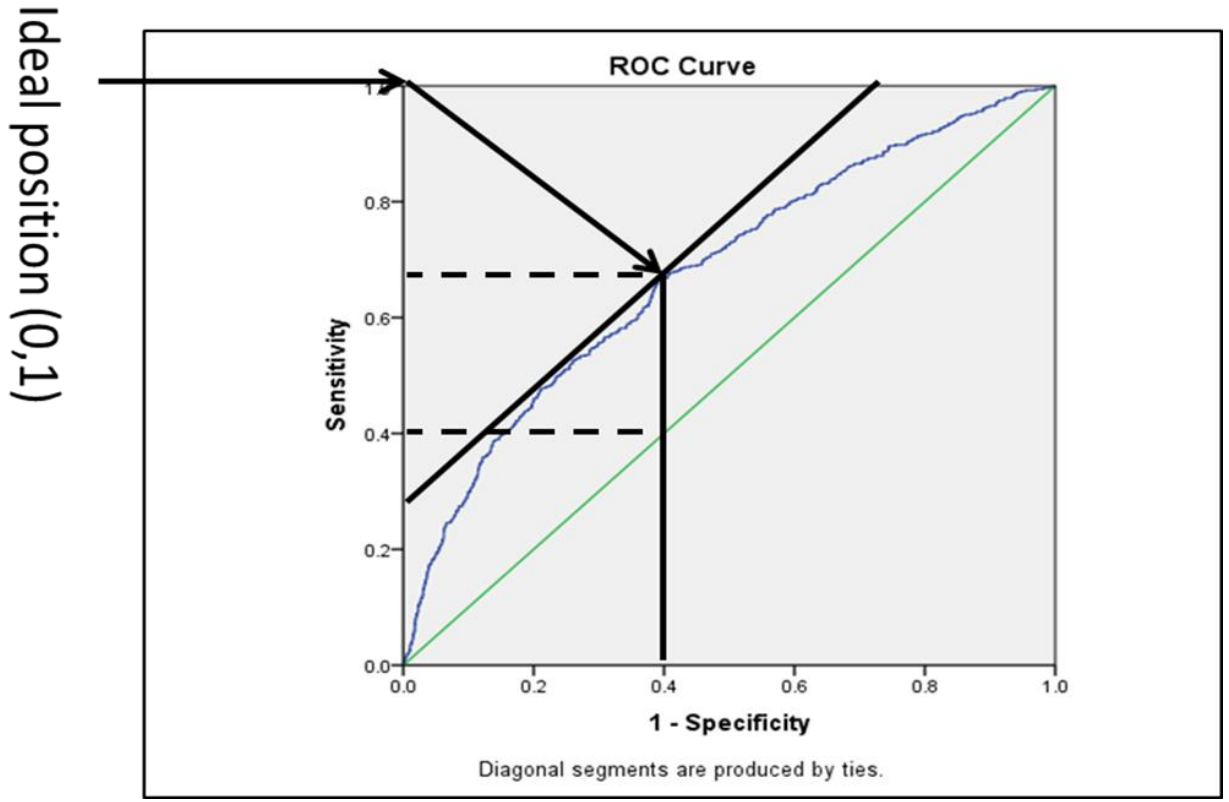
The model predicts less than 5% chance of default

The model predicts more than 85% chance of default

Youden's Index for Optimal Cut-Off Probability

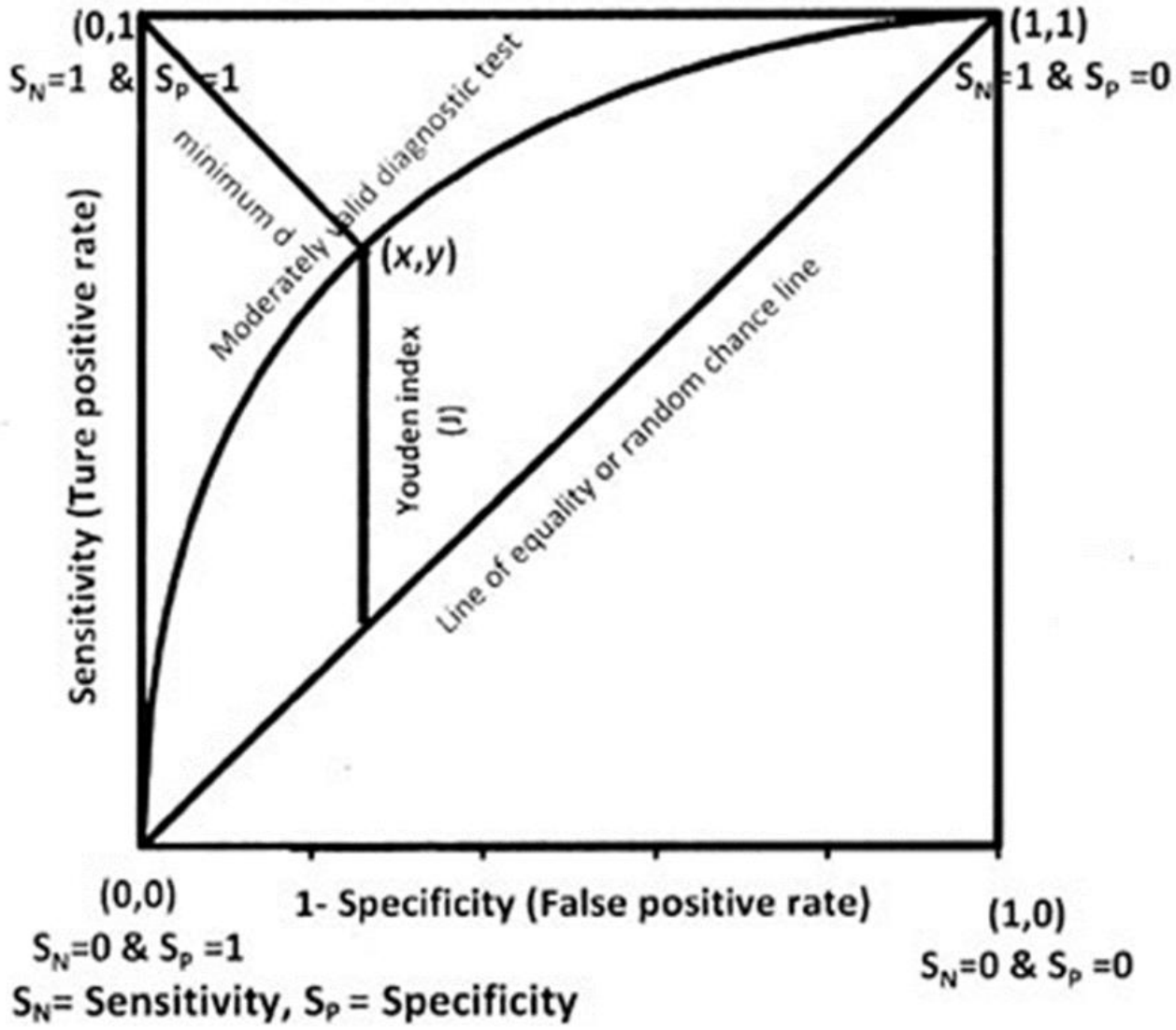
Youden's Index (1950) is a classification cut-off probability, for which the following function is maximized (also known as J statistic):

Youden's Index = J Statistic = $\underset{P}{\text{Max}}$[Sensitivity(p) + Specificity(p) – 1]



DATA ANALYTICS

Youden's Index for Optimal Cut-Off Probability



Cost-Based Cut-Off Probability

In cost-based approach, we assign penalty cost for misclassification of positives and negatives. Assume that cost of misclassifying negative (0) as positive (1) is C_{01} and cost of misclassifying positive (1) as negative (0) is C_{10} as shown in Table

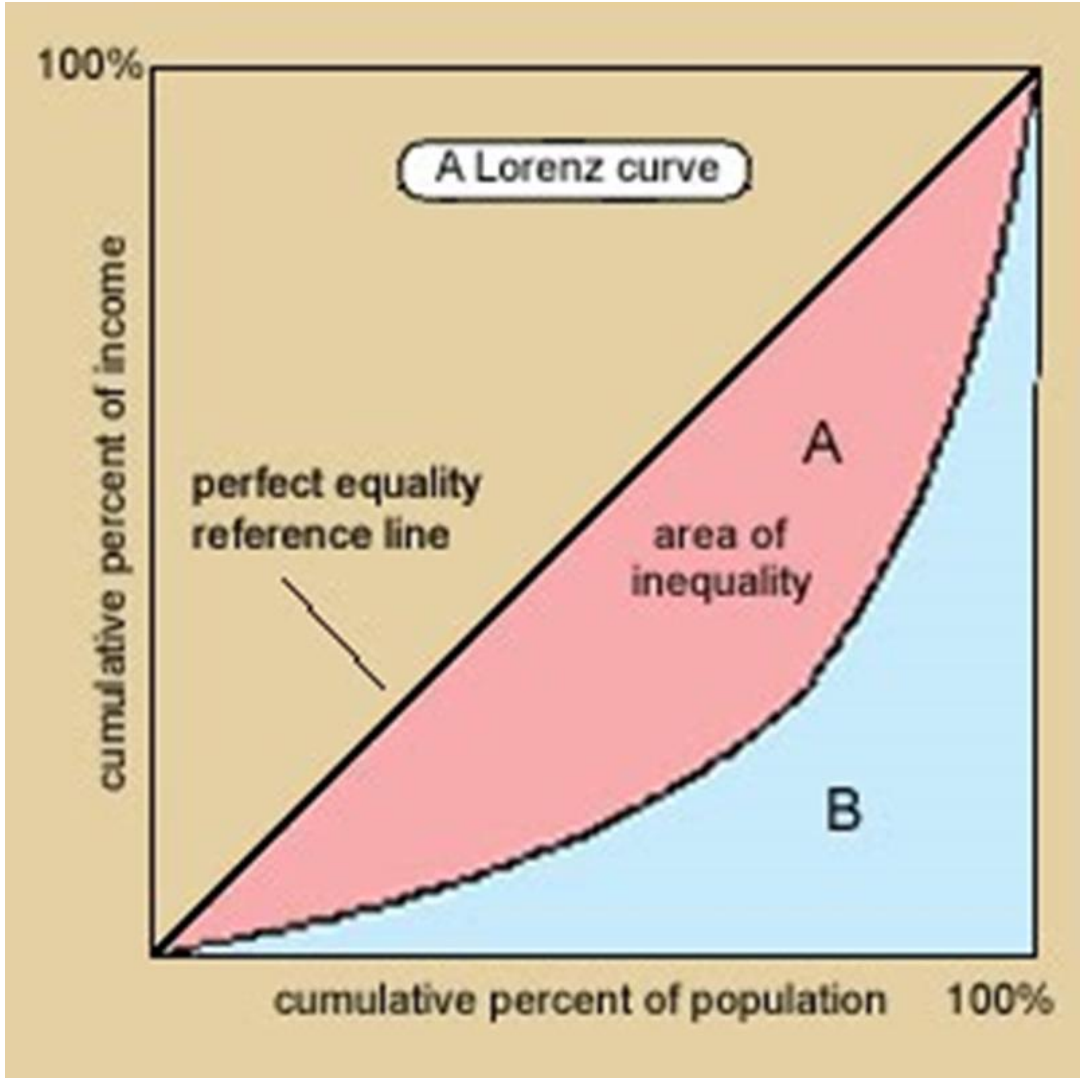
Observed	Classified	
	0	1
0	---	C_{01}
1	C_{10}	---

The optimal cut-off probability is the one which minimizes the total penalty cost and is given by

$$\min_{p} \left[C_{01}P_{01} + C_{10}P_{10} \right]$$

DATA ANALYTICS

Lorenz Curve



Gini Coefficient = A / (A+B)

Gini Coefficient = 2 AUC - 1

AUC = Area Under the ROC Curve

Gini Coefficient

- Gini coefficient measures individual impact of the an explanatory variable.

- Gini coefficient = 2 AUC – 1

- AUC = Area under the ROC Curve

Forward LR (Likelihood Ratio)

In Forward LR (Lawless and Singhal 1987), at each step one variable is added to the model.

Step 1: Start with no variables in the model. Set $i = 0$.

Step 2: For each independent variable, calculate the difference between $-2LL_i$ and $-2LL_{i+1}$ value. When $i = 0$, we will calculate the difference between $-2LL_0$ (model without a variable) and $-2LL_1$ (model with one variable). Add the variable with highest difference in $-2LL_i$ and $-2LL_{i+1}$ if the p-value after adding the variable is statistically significant under likelihood ratio test (omnibus test) at a significance level of α.

Step 3: Repeat step 2, till all the variables are exhausted or the change in $-2LL$ is not significant, that is the p-value after adding a new variable is greater than α.

Forward Selection Wald

In Forward Selection Wald, the variables are entered based on the Wald's test.

Step1: Assume that the data has "*n*" explanatory variables. Develop a univariate LR model and calculate the *p*-value of all the variables and add the variable with the smallest *p*-value if it is less than α (significance level).

Step 2: Add a new variable with smallest *p*-value or largest Wald's statistic value if *p*-value (based on the Wald's test) is less than the significance α.

Step 3: Repeat the procedure till the *p*-value becomes greater than α.

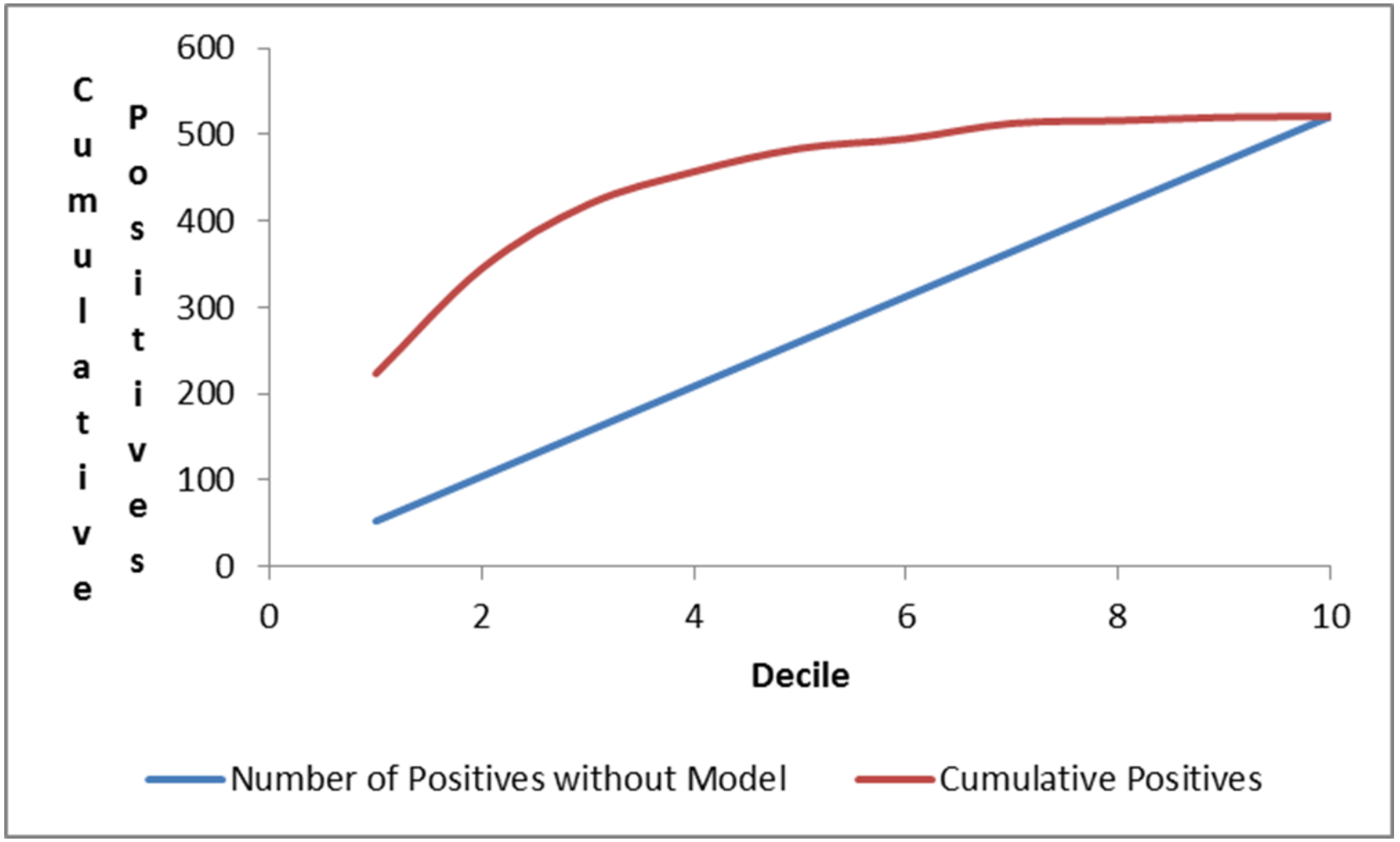
Gain and Lift

$$\text{Gain} = \frac{\text{Cumulative number of positive observations up to decile i}}{\text{Total number of positive observations in the data}}$$

$$\text{Lift} = \frac{\text{Cumulative Gain using LR model}}{\text{Cumulative Gain using Random model}}$$

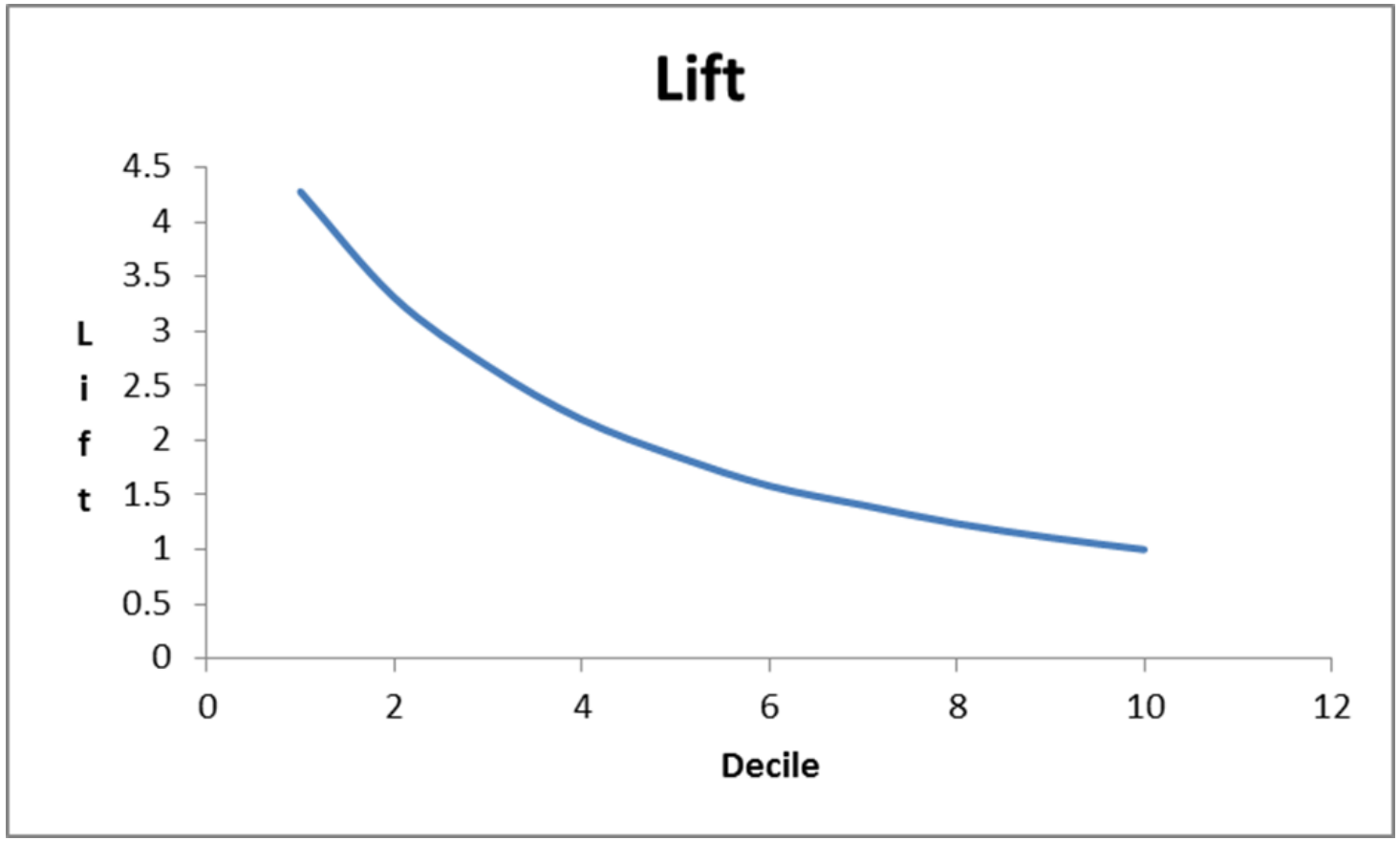
DATA ANALYTICS

GAIN CHART



DATA ANALYTICS

Lift Chart



DATA ANALYTICS

Exercise

- To be done

References

Text Book:

"Business Analytics, The Science of Data-Driven Decision Making", U. Dinesh Kumar, Wiley 2017

https://towardsdatascience.com/train-validation-and-test-sets-72cb40cba9e7

https://medium.com/hugo-ferreiras-blog/confusion-matrix-and-other-metrics-in-machine-learning

THANK YOU

Dr.Mamatha H R

Professor,Department of Computer Science

mamathahr@pes.edu

+91 80 2672 1983 Extn 834