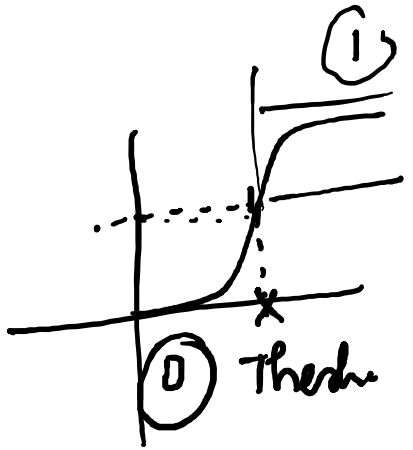
Logistic Reg → classification → O/p → Y < 0
1

Y < A
B
C



DATA ANALYTICS

Unit 2:Confusion matrices and Metrics

$$\hat{p} = \frac{1}{1+e^{-(\beta_0+\beta_1 x)}} \quad \text{or} \quad \frac{e^{(\beta_0+\beta_1 x)}}{1+e^{(\beta_0+\beta_1 x)}}$$

Cut-off probability P_c

Mamatha H R

Department of Computer Science and Engineering

"Cut off probability P_c" & std methods

DATA ANALYTICS

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: error matrix

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

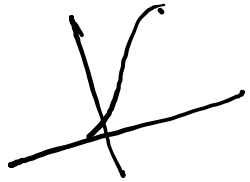
Ground truth

DATA ANALYTICS

Confusion matrix : Exercise :Space Shuttle Challenger Data

Space shuttle orbiter Challenger (Mission STS-51-L) was the 25th shuttle launched by NASA on January 28, 1986 (Smith, 1986; Feynman 1988). The Challenger crashed 73 seconds into its flight due to the erosion of O-rings which were part of the solid rocket boosters of the shuttle. Before the launch, the engineers at NASA were concerned about the outside temperature which was very low (the actual launch occurred at 36°F). Data in Table 11.1 shows the O-ring erosion and the launch temperature of the previous shuttle launches, where 'damage to O-ring = 1' implies there was a damage to O-ring and 'damage to O-ring = 0' implies there was no damage to O-ring during that launch. In this case, the outcome is binary − either there is a damage to O-ring or there is no damage to O-ring. **We can develop a logistic regression model to predict the probability of erosion of O-ring based on the launch temperature.**

Confusion matrix: Exercise :Space Shuttle Challenger Data



$$\hat{p} = \frac{1}{1+e^{-(\beta_0+\beta_1 x)}}$$

Cutoff Prob

Class 1.

Pc. 0.5

< 0.5 → class 0

n = 24

TABLE 11.8 Challenger crash data — predicted probability using logistic regression model

S. No.	Flight Number	Launch Temperature	Damage to O-ring	Predicted Probability
1	STS 1	66.00	0	0.43
2	STS 2	70.00	1	0.23
3	STS 3	69.00	0	0.27
4	STS 4	80.00	0	0.03
5	STS 5	68.00	0	0.32
6	STS 6	67.00	0	0.37
7	STS 7	72.00	0	0.15
8	STS 8	73.00	0	0.13
9	STS 9	70.00	0	0.23
10	STS 41B	57.00	1	0.86
11	STS 41C	63.00	1	0.61
12	STS 41D	70.00	1	0.23
13	STS 41G	78.00	0	0.04
14	STS 51A	67.00	0	0.37
15	STS 51C	53.00	1	0.94
16	STS 51D	67.00	0	0.37
17	STS 51B	75.00	0	0.08
18	STS 51G	70.00	0	0.23
19	STS 51F	81.00	0	0.02
20	STS 51I	76.00	0	0.07
21	STS 51J	79.00	0	0.03
22	STS 61A	75.00	1	0.08
23	STS 61B	76.00	0	0.07
24	STS 61C	58.00	1	0.83

DATA ANALYTICS

Confusion matrix

The Challenger data had 17 no-damage (negative) cases and 7 damage (positive) cases. The probability of damage to O-ring is calculated using the logistic function.

When the probability is less than 0.5, the observation is classified as negative (coded as 0) and when the probability is greater than or equal to 0.5, the observation is classified as positive (coded as 1).

For the classification cut-off probability value of 0.5, the model has classified all 17 negatives (coded as 0) as negatives and 4 positives (coded 1) as positives and remaining 3 positives as negatives.

TABLE 11.6 Classification Table[a]

Observed			Predicted		
			Damage to O-ring		Percentage Correct
			0 (Negative)	1 (Positive)	
Step 1	Damage to O-ring	0 (Negative)	17 (TN)	0 (FP)	100.0
		1 (Positive)	3 (FN)	4 (TP)	57.1
	Overall Percentage				87.5

[a]The cut value is 0.500.

24

DATA ANALYTICS

Confusion matrix

TABLE 11.6 Classification Table[a]

	Observed		Predicted		
			Damage to O-ring		Percentage Correct
			0 (Negative)	1 (Positive)	
Step 1	Damage to O-ring	0 (Negative)	17 (TN)	0 (FP)	100.0
		1 (Positive)	3 (FN)	4 (TP)	57.1
	Overall Percentage				87.5

[a]The cut value is 0.500

$P_c = 0.5$

TABLE 11.7 Classification Table[a]

	Observed		Predicted		
			Damage to O-ring		Percentage Correct
			0	1	
Step 1	Damage to O-ring	0	9	8	52.9
		1	1	6	85.7
	Overall Percentage				62.5

[a]The cut value is 0.200.

0.2

Classification Performance Metrics

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

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

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

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

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

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

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

Sample

Data	Predict	Actual
1	1	1
2	1	1
3	1	0
4	1	0
5	1	0
6	1	0

TPR / Sensitivity / Recall $= \frac{TP}{TP+FN} = \frac{2}{2+0} = \frac{2}{2} = 1$

100%. — We cannot Perfect..

Precision $= \frac{TP}{TP+FP} = \frac{2}{2+4} = \frac{2}{6} = 0.33$

33%

Confusion matrix to compare two classifiers

- Which is the better classification model (wrt Class A)?

Model 1

Actual \ Predicted	A	A'
A	60	34
A'	1	12

Recall(A) = 60/94
Precision(A) = 60/61
F1_score(A) = 2RP/(R+P)
= 0.774
Accuracy(Model1)
= 0.673

Recall(A')=12/13
Precision(A')=12/34
F1_score(A')=0.51

Model 2

Actual \ Predicted	A	A'
A	90	4
A'	8	5

Recall(A) = 90/94
Precision(A) = 90/98
F1_score(A) = 2RP/(R+P)
= 0.937
Accuracy(Model2)
= 0.888

Recall(A')=5/13
Precision(A')=5/9
F1_score(A')=0.45

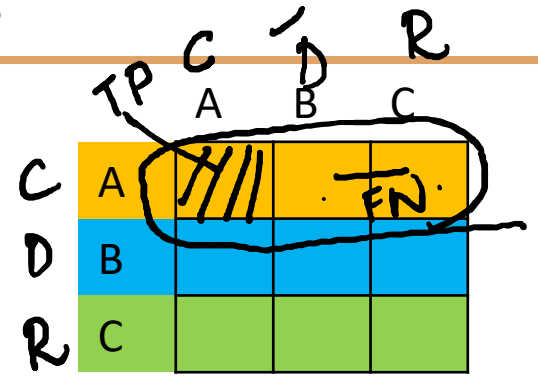
DATA ANALYTICS

Confusion matrices for multiple classes

- What is Recall(A)?
- What is Specificity(B)?
- What is Precision(C)?
- What is the average accuracy of this model?

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

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



Predicted class

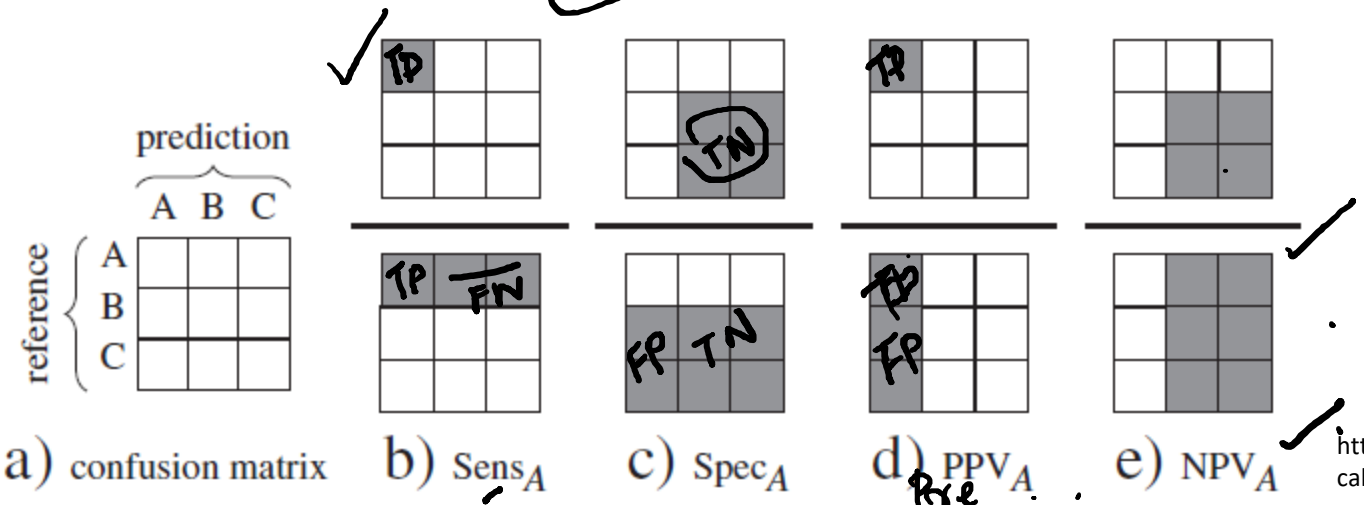
Actual class:	Cat	Dog	Rabbit
Cat	5	3	0
Dog	2	3	1
Rabbit	0	2	11

Predicted class

Actual class:	Cat	NotCat
Cat	5	3
NotCat	2	17

Predicted class

Actual class:	Dog	NotDog
Dog	3	3
NotDog	5	16



PPV → positive predictive value / precision

NPV → Negative predictive value

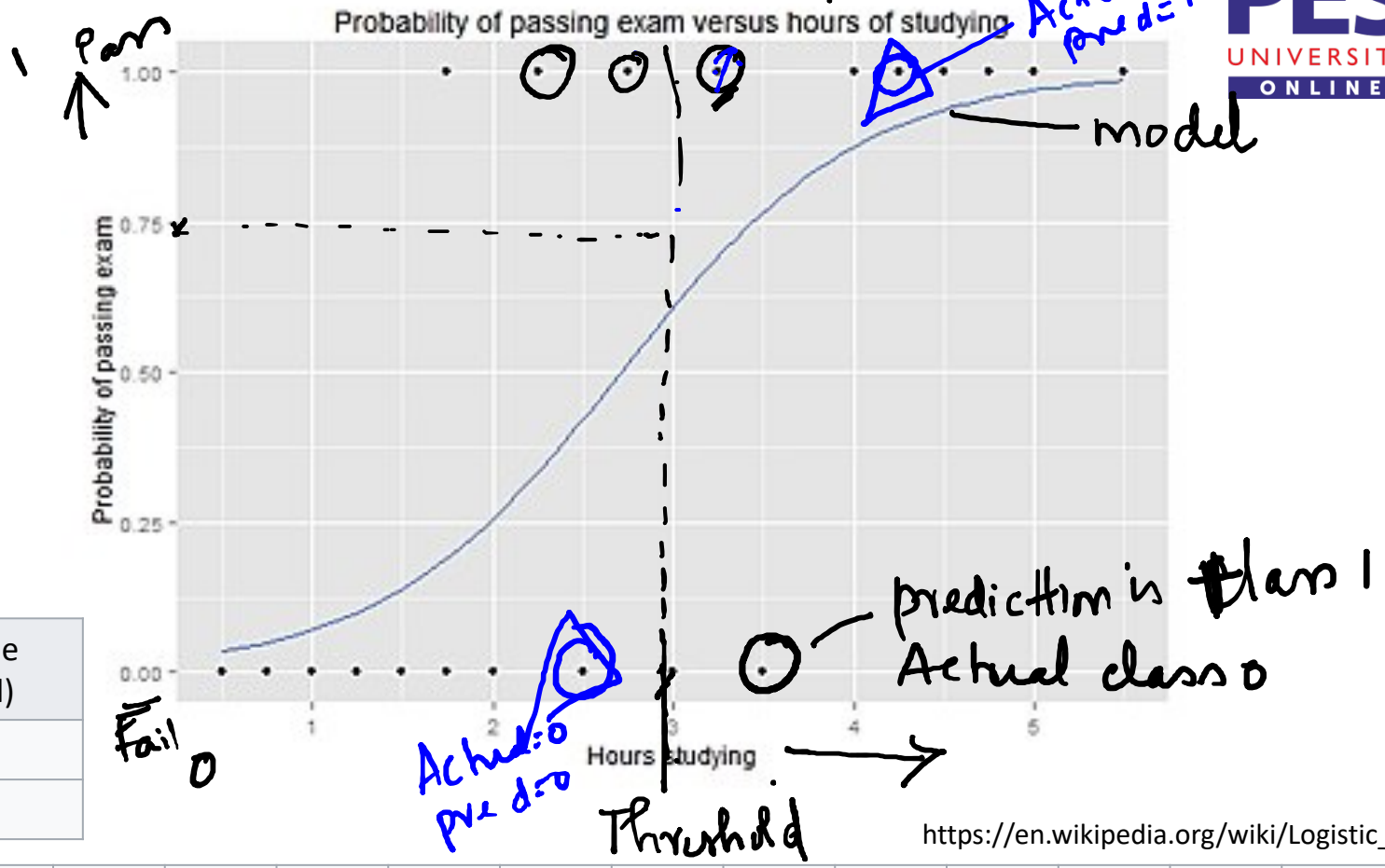
https://stats.stackexchange.com/questions/91044/how-to-calculate-precision-and-recall-in-a-3-x-3-confusion-matrix

Revisiting the example

Hours of study	Passing exam		
	Log-odds	Odds	Probability
1	−2.57	0.076 ≈ 1:13.1	0.07
2	−1.07	0.34 ≈ 1:2.91	0.26
3	0.44	1.55	0.61
4	1.94	6.96	0.87
5	3.45	31.4	0.97

One hr of study increases log odds of passing by 1.5046

	Coefficient	Std.Error	P-value (Wald)
Intercept	−4.0777	1.7610	0.0206
Hours	1.5046	0.6287	0.0167



https://en.wikipedia.org/wiki/Logistic_regression

Hours	0.50	0.75	1.00	1.25	1.50	1.75	1.75	2.00	2.25	2.50	2.75	3.00	3.25	3.50	4.00	4.25	4.50	4.75	5.00	5.50
Pass	0	0	0	0	0	0	1	0	1	0	1	0	1	0	1	1	1	1	1	1

DATA ANALYTICS

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.
- 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.
- 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.
- 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.

Hours of study	Passing exam	
	Probability	Label
1	0.070	0
2	0.260	0
3	0.610	0
4	0.870	1
5	0.950	1
6	0.970	1
7	0.980	0

(1,5): concordant pair

(4,7): discordant pair

DATA ANALYTICS

Concordant and Discordant Pairs : Exercise :Space Shuttle Challenger Data

TABLE 11.8 Challenger crash data — predicted probability using logistic regression model

S. No.	Flight Number	Launch Temperature	Damage to O-ring	Predicted Probability
1	STS 1	66.00	0	0.43
2	STS 2	70.00	1	0.23
3	STS 3	69.00	0	0.27
4	STS 4	80.00	0	0.03
5	STS 5	68.00	0	0.32
6	STS 6	67.00	0	0.37
7	STS 7	72.00	0	0.15
8	STS 8	73.00	0	0.13
9	STS 9	70.00	0	0.23
10	STS 41B	57.00	1	0.86
11	STS 41C	63.00	1	0.61
12	STS 41D	70.00	1	0.23
13	STS 41G	78.00	0	0.04
14	STS 51A	67.00	0	0.37
15	STS 51C	53.00	1	0.94
16	STS 51D	67.00	0	0.37
17	STS 51B	75.00	0	0.08
18	STS 51G	70.00	0	0.23
19	STS 51F	81.00	0	0.02
20	STS 51I	76.00	0	0.07
21	STS 51J	79.00	0	0.03
22	STS 61A	75.00	1	0.08
23	STS 61B	76.00	0	0.07
24	STS 61C	58.00	1	0.83

STS 1 (for which Y = 0) and STS 2 (for which Y = 1), there is no cut-off probability that can classify both positive (Y = 1) and negative (Y = 0) correctly. Such pairs are called Discordant Pairs. That is, 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

If we use a classification cut-off probability between 0.23 and 0.86, then we will classify STS 9 (Y = 0) and STS 41B (Y = 1) correctly. Such pairs are called Concordant Pairs.
That is, 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.

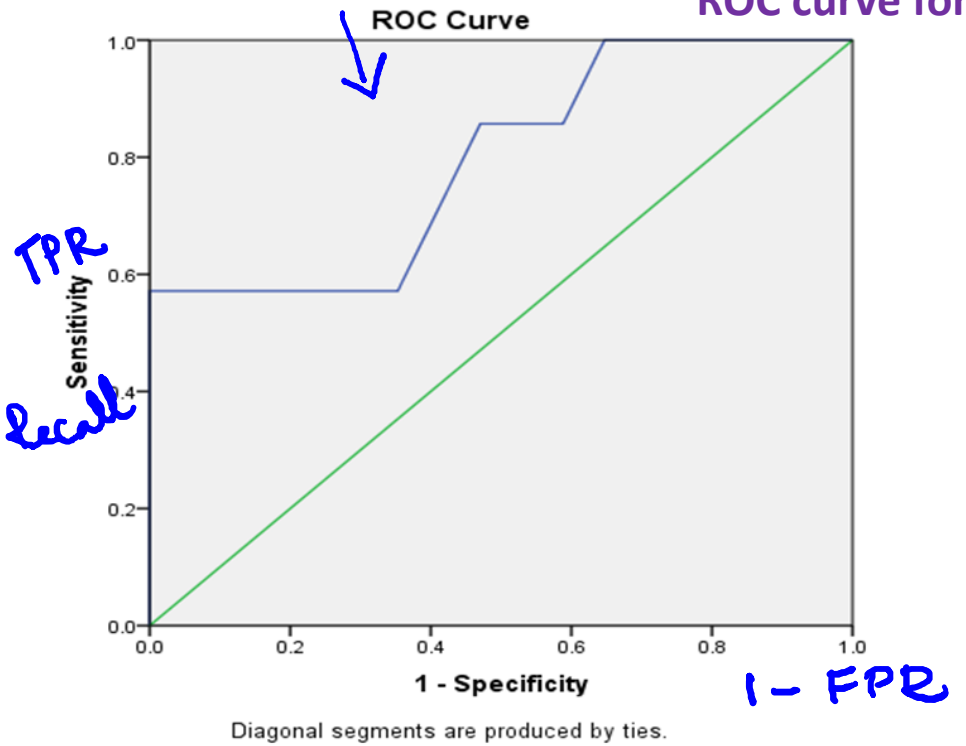
A logistic regression model with high proportion of concordant pairs is preferred.

RoC & AuC

Receiver Operating Characteristics (ROC) Curve

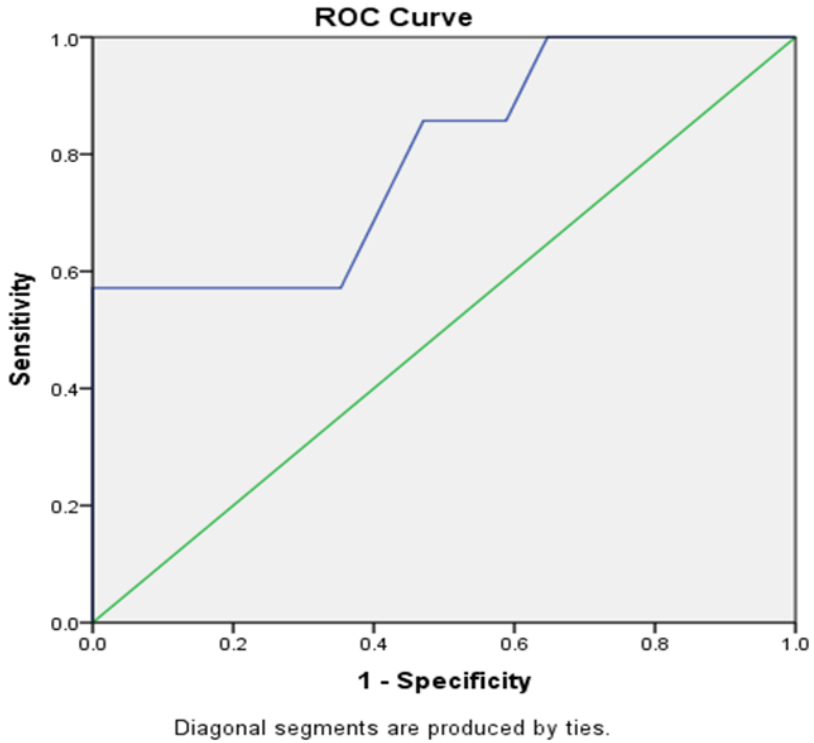
- ROC curve is a plot between sensitivity (true positive rate) in the vertical axis and 1 – specificity (false positive rate) in the horizontal axis.
- The higher the area under the ROC curve, the better the prediction ability.

ROC curve for challenger crash data.



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.



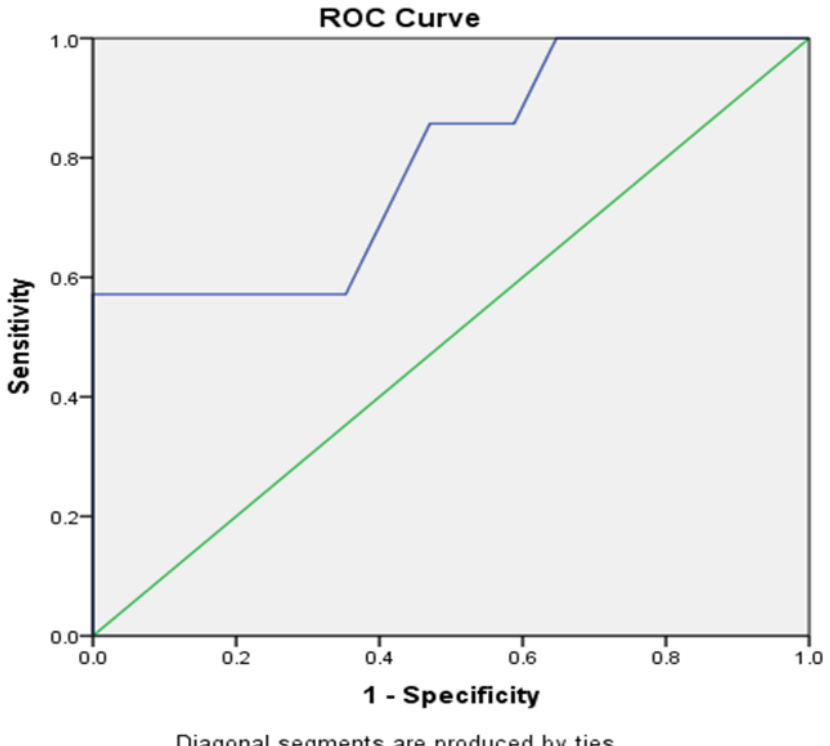
Area Under ROC Curve (AUC)- Space Shuttle Challenger Data

- The higher the area under the ROC curve, the better the prediction ability.

TABLE 11.9 Area under the curve

Area	Std. Error[a]	Asymptotic Sig.[b]	Asymptotic 95% Confidence Interval	
			Lower Bound	Upper Bound
0.794	0.107	0.026	0.585	1.000

The test result variable(s): Predicted probability has at least one tie between the positive actual state group and the negative actual state group. Statistics may be biased. [a]Under the nonparametric assumption [b]Null hypothesis: true area = 0.5.



DATA ANALYTICS

Area Under ROC Curve (AUC)- Space Shuttle Challenger Data

- The higher the area under the ROC curve, the better the prediction ability.

TABLE 11.9 Area under the curve

Area	Std. Error[a]	Asymptotic Sig.[b]	Asymptotic 95% Confidence Interval	
			Lower Bound	Upper Bound
0.794	0.107	0.026	0.585	1.000

The test result variable(s): Predicted probability has at least one tie between the positive actual state group and the negative actual state group. Statistics may be biased. [a]Under the nonparametric assumption [b]Null hypothesis: true area = 0.5.

AUC is the proportion of concordant pairs in the data if the model is used for classification. AUC is one of the criteria used for final model selection; higher AUC is assumed to be a better model.

For challenger crash data, the AUC is 0.794. The area under the ROC curve can be interpreted as follows:

1. If we use the logistic regression model, then there will be 79.4% concordant pairs and 20.6% discordant pairs.
2. For a randomly selected pair of positive and negative observations, probability of correctly classifying them is 0.794.

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.

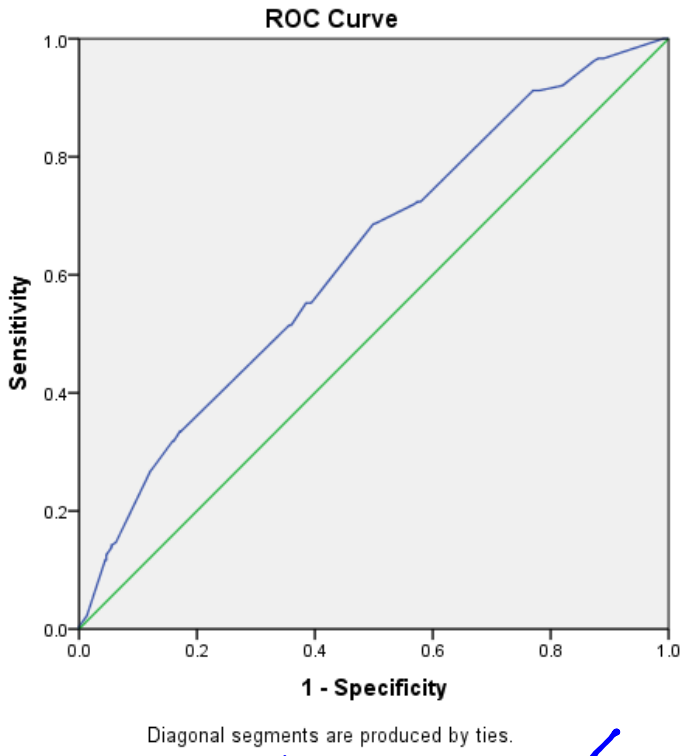
- If n_1 is the number of positives (1s) and n_2 is the number of negatives (0s), then the area under the ROC curve is the proportion of cases in all possible combinations of (n_1, n_2) such that n_1 will have higher probability than n_2.

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.

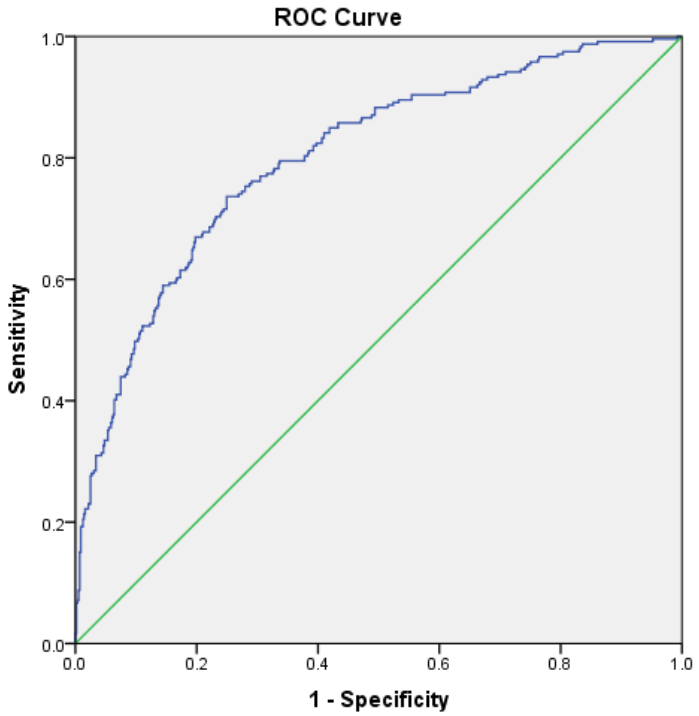
DATA ANALYTICS

Area Under ROC Curve (AUC)



model: 1 ✓

AUC = 0.629



Model: 2 ✓

AUC = 0.801 ✓

80% Concordant pairs
20% discordant pair

ROC Curve

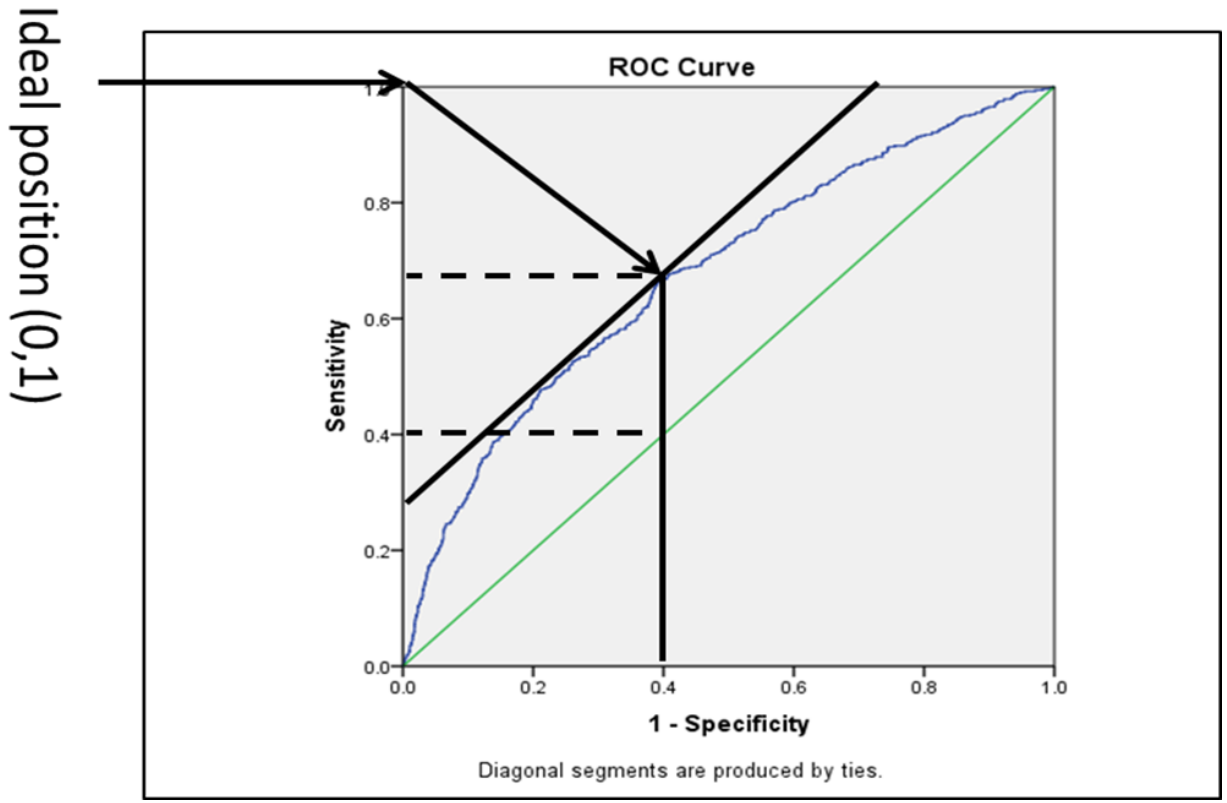
- General rule for acceptance of the model:

- 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

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):

$$\text{Youden's Index} = \text{J Statistic} = \max_{p} [\text{Sensitivity}(p) + \text{Specificity}(p) - 1]$$



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_{00}	C_{01}
1	C_{10}	C_{11}

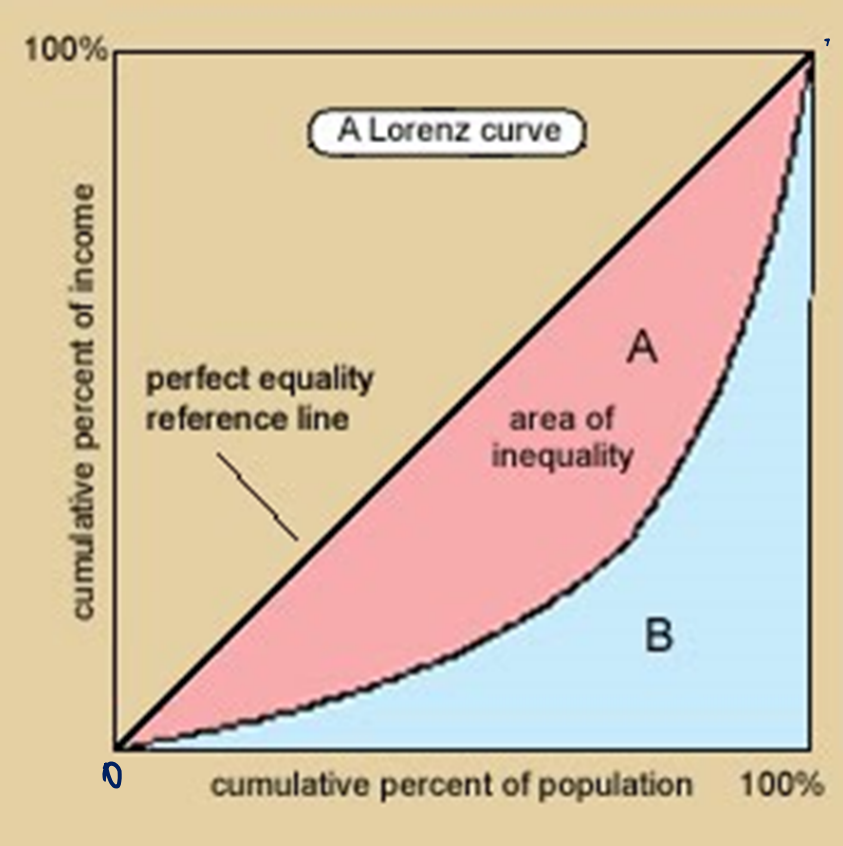
The optimal cut-off probability is the one which minimizes the total penalty cost and is given by

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

DATA ANALYTICS

Lorenz Curve

AUC



Gini Index is a statistical measure of dispersion

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

Questions asked after the class

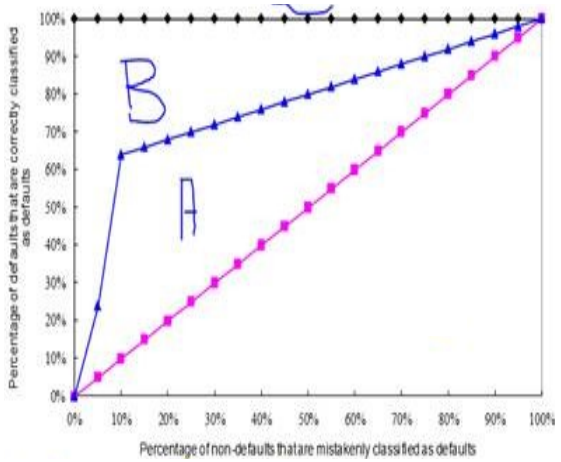
- Can discordant pairs be thought of as outliers?
  - Indeed, discordancy tests are used to detect outliers and one or both the points in the discordant pair could be outliers.
  - However, we must also be aware there are other possibilities:

  (a) The parameters (coefficients) could be better estimated

  (b) The current model (logistic regression) is not suitable for modeling the data on hand (some preprocessing may be required before we can model the data using Logistic Regression or we could explore alternatives)

- Why is Gini coefficient = (2 AUC – 1)?
  - (AUC = Area under the ROC Curve)
  - Gini = A/(A+B)
    - A = area under the curve and the diagonal
    - B = area under the perfect model and diagonal
    - Gini (in RoC) = A/(A+B) = A/0.5 = 2A
    - AUC for this case = A + ½
    - AUC = Gini/2 + ½

    ⇒ Gini = 2AUC-1



https://stats.stackexchange.com/questions/342329/gini-and-auc-relationship

References

Text Book:

"Business Analytics, The Science of Data-Driven Decision Making", U. Dinesh Kumar, Wiley 2017 (Ch. 11.1-11.4, 11.6.5, 11.7.2-11.7.3)

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

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

https://online.stat.psu.edu/stat504/node/216/

https://towardsdatascience.com/logit-of-logistic-regression-understanding-the-fundamentals-f384152a33d1

DATA ANALYTICS

Unit 3: Time Series Analysis

Jyothi R.
Department of Computer Science and Engineering

DATA ANALYTICS

Unit 3: Introduction to Time Series Data

Jyothi R., Gowri Srinivasa

Department of Computer Science and Engineering

DATA ANALYTICS

INTRODUCTION TO FORECASTING

- Forecasting - important and frequently addressed problems in analytics
- Inaccurate forecasting has a significant impact
- For example
  - non-availability of product → customer dissatisfaction
  - too much inventory → erodes the organization's profit
- Necessary to forecast the demand for a product and service as accurately as possible.
- Every organization prepares long-range and short-range planning
  - forecasting demand for product and service is an important input for both long-range and short-range planning
- Budget allocation, manpower, warehouse capacity, machine resource planning, etc., based on forecast of demand for a product

INTRODUCTION TO FORECASTING

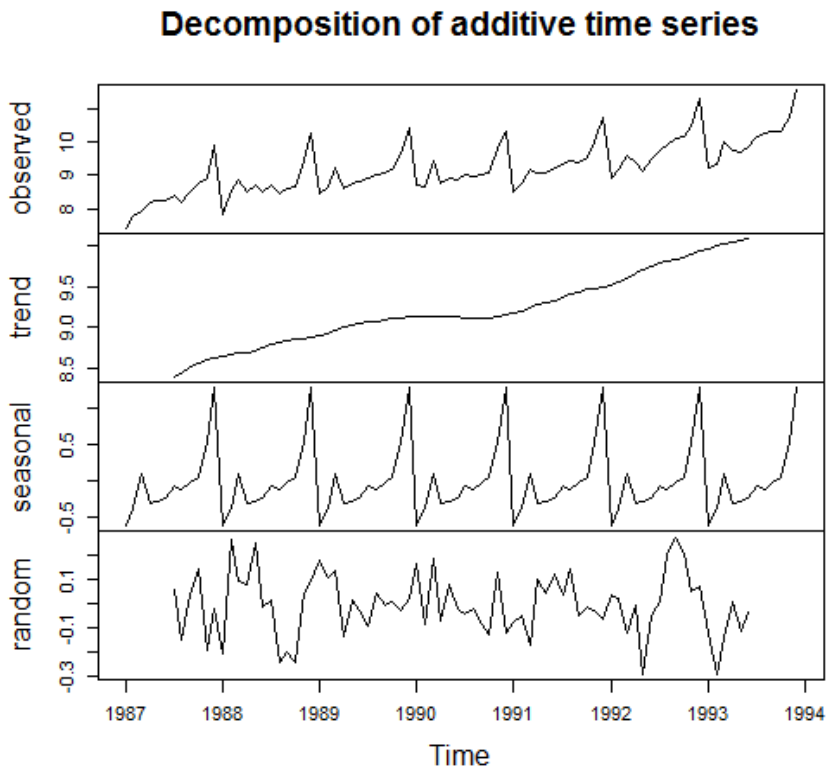
- Forecasting can be very challenging with stock keeping units (SKUs) running into several millions.

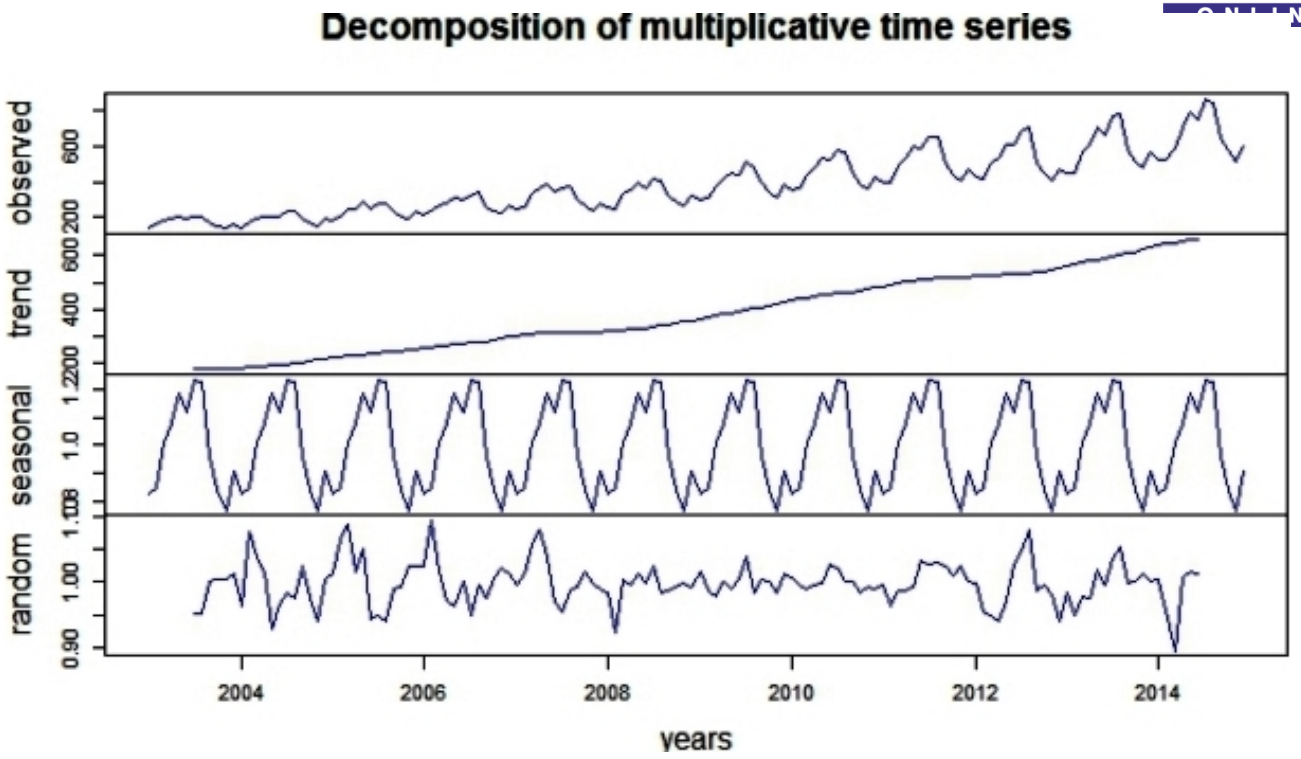
1. Boeing 747-400 has more than 6 million parts and several thousand unique parts (Hill, 2011). Forecasting demand for spare parts is important since non-availability of mission critical parts can result in aircraft on ground (AOG) which can be very expensive for airlines.

2. Amazon.com sells more than 350 million products through its E-commerce portal. Amazon itself sells about 13 million SKUs and has more (about 2 million) retailers selling their products through Amazon (Ali, 2017).

3. Walmart sells more than 142,000 products through their supercenters. Being a brick-and-mortar retail store, Walmart has to maintain stock for each and every product sold and predict demand for the products as accurately as possible.

Additive and Multiplicative Time Series Data





$$Y_t = T_t + S_t + C_t + I_t$$

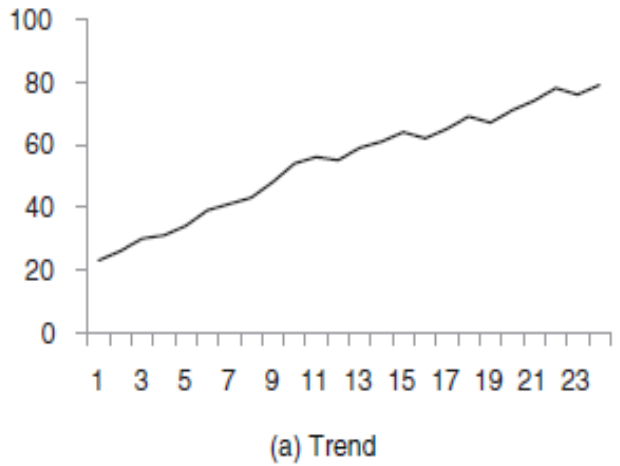
$$Y_t = T_t \times S_t \times C_t \times I_t$$

DATA ANALYTICS

COMPONENTS OF TIME-SERIES DATA

From a forecasting perspective, a time-series data can be broken into the following components

1. **Trend Component (T_t):** Trend is the consistent long-term upward or downward movement of the data over a period of time.



(a) Trend

https://school.stockcharts.com/doku.php?id=chart_analysis:trend_lines

COMPONENTS OF TIME-SERIES DATA Contd.

2. Seasonal Component (S_t): Seasonal component is the repetitive upward or downward movement (or fluctuations) from the trend that occurs within a calendar year such as seasons, quarters, months, days of the week, etc.

- The upward or downward fluctuation may be caused due to festivals, customs within a society, school holidays, business practices within the market such as 'end of season sale', and so on.

- For example, in India demand for many products surge during the festival months of October - December.

- Seasonal fluctuation occurs at fixed intervals (such as months, quarters) known as periodicity of seasonal variation and repeats over time.

Seasonal Component (S_t): Contd.

The seasonal component consists of effects that are reasonably stable with respect to timing, direction and magnitude. It arises from systematic, calendar related influences such as:

- **Natural Conditions:** Weather fluctuations that are representative of the season (uncharacteristic weather patterns such as snow in summer would be considered irregular influences)

- **Business and Administrative procedures:**
  Start and end of the school term

- **Social and Cultural behavior:**
  Christmas

Seasonal Component (S_t):

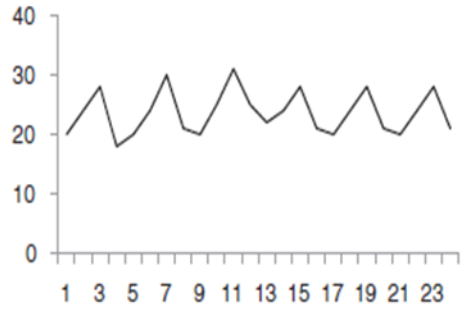
It also includes calendar related systematic effects that are not stable in their annual timing or are caused by variations in the calendar from year to year, such as:

- **Trading Day Effects**

the number of occurrences of each of the day of the week in a given month will differ from year to year
- There were 4 weekends in March in 2000, but 5 weekends in March of 2002

- **Moving Holiday Effects**

holidays which occur each year, but whose exact timing shifts
- Diwali, Easter, Ramadan



(b) Seasonality (fixed periodicity)

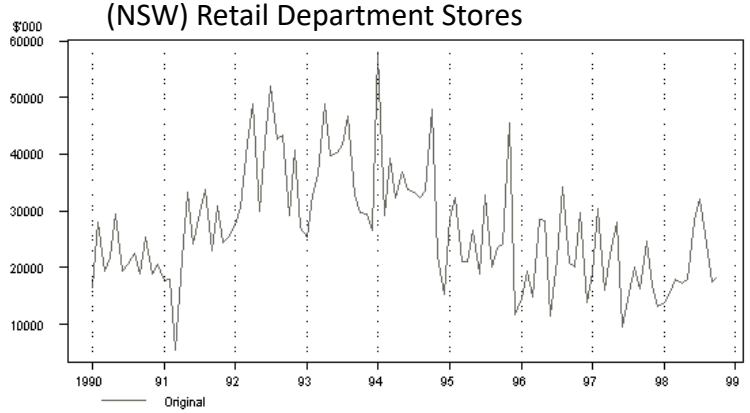
Identifying seasonal components:
Regularly spaced peaks and troughs which have a consistent direction and approximately the same magnitude every year, relative to the trend

COMPONENTS OF TIME-SERIES DATA

Seasonal Component (S_t):

- Seasonality in a time series can be identified by regularly spaced peaks and troughs which have a consistent direction and approximately the same magnitude every year, relative to the trend.
- In this example, the magnitude of the seasonal component increases over time, as does the trend.

Monthly Retail Sales in New South Wales (NSW) Retail Department Stores

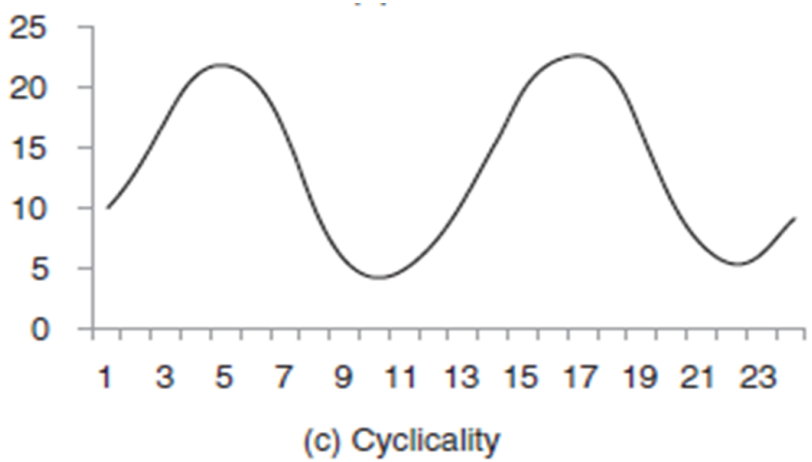


Obvious large seasonal increase in December retail sales in New South Wales due to Christmas shopping

COMPONENTS OF TIME-SERIES DATA

3. Cyclical Component (C_t): Cyclical component is fluctuation around the trend line that happens due to macro-economic changes such as recession, unemployment, etc.

- Cyclical fluctuations have repetitive patterns with a time between repetitions of more than a year

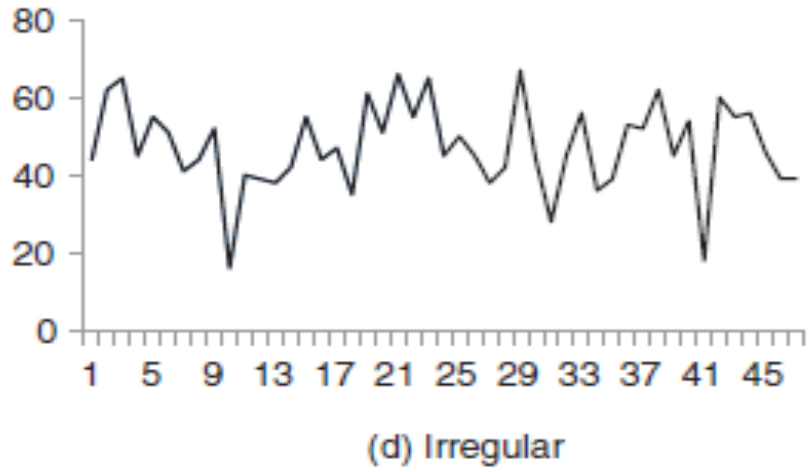


(c) Cyclicality

- A major difference between seasonal fluctuation and cyclical fluctuation is that seasonal fluctuation occurs at fixed period within a calendar year, whereas cyclical fluctuations have random time between fluctuations.
- That is, periodicity of seasonal fluctuations is constant, whereas the periodicity of cyclical fluctuations is not constant.

COMPONENTS OF TIME-SERIES DATA contd.

4. Irregular Component (I_t): Irregular component is the white noise or random uncorrelated changes that follow a normal distribution with mean value of 0 and constant variance.



(d) Irregular

- What remains after the seasonal and trend components of a time series have been estimated and removed.
- It results from short term fluctuations in the series which are neither systematic nor predictable

Additive and Multiplicative Time Series Revisited

- The additive time-series model is given by

$$Y_t = T_t + S_t + C_t + I_t$$

- The additive models assume that the seasonal and cyclical components are independent of the trend component.
- Additive models are not very common since in many cases the seasonal component may not be independent of trend.
- The additive model is appropriate if the seasonal component remains constant about the level (or mean) and does not vary with the level of the series.

- The multiplicative time-series model is given by

$$Y_t = T_t \times S_t \times C_t \times I_t$$

- Multiplicative models are more common and are a better fit for many data sets.
- In many cases, we will use the form

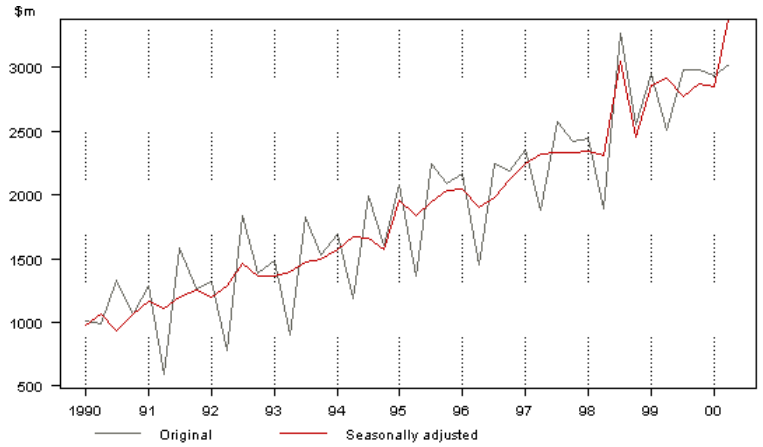
$$Y_t = T_t \times S_t$$

- To estimate the cyclical component we will need a large data set.
- The multiplicative model is more appropriate, if seasonal variation is correlated with level (local mean).

Additive and Multiplicative Time Series Revisited

- The additive time-series model is given by $Y_t = T_t + S_t + C_t + I_t$

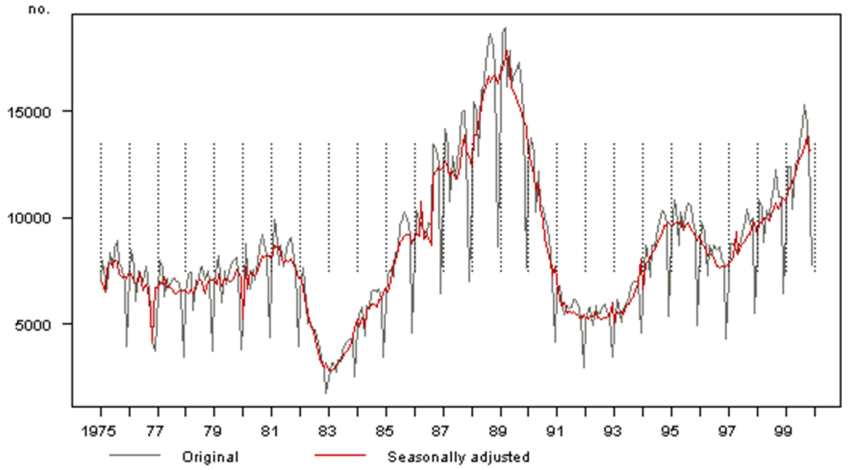
General Government and Other Current Transfers to Other Sectors



The underlying level of the series fluctuates but the magnitude of the seasonal spikes remain approximately stable

- The multiplicative time-series model is given by $Y_t = T_t \times S_t \times C_t \times I_t$

Monthly NSW ANZ Job Advertisements



The trend has the same units as the original series, but the seasonal and irregular components are unitless factors, distributed around 1

Decomposition of Time Series Data - Additive

- Decomposition models are typically additive or multiplicative, but can also take other forms such as pseudo-additive.

Additive Decomposition

In some time series, the amplitude of both the seasonal and irregular variations do not change as the level of the trend rises or falls. In such cases, an additive model is appropriate.

In the additive model, the observed time series (O_t) is considered to be the sum of three independent components: the seasonal S_t, the trend T_t and the irregular I_t.

Observed series = Trend + Seasonal + Irregular

$$O_t = T_t + S_t + I_t$$

Seasonally adjusted series = Observed-Seasonal

$$SA_t = O_t - \hat{S}_t$$

$$= T_t + I_t$$

COMPONENTS OF TIME-SERIES DATA contd.

- **Multiplicative Decomposition**

  In many time series, the amplitude of both the seasonal and irregular variations increase as the level of the trend rises. In this situation, a multiplicative model is usually appropriate.

  In the multiplicative model, the original time series is expressed as the product of trend, seasonal and irregular components.

- Observed series = Trend x Seasonal x Irregular

$$O_t = T_t + S_t + I_t$$

$$\text{Seasonally Adjusted series} = \text{Observed} \div \text{Seasonal}$$
$$= \text{Trend} \times \text{Irregular}$$

$$SA_t = O_t / \hat{S}_t$$
$$= T_t \times I_t$$

COMPONENTS OF TIME-SERIES DATA contd.

Pseudo-Additive Decomposition:

- The multiplicative model cannot be used when the original time series contains very small or zero values
- This is because it is not possible to divide a number by zero
- In these cases, a pseudo additive model combining the elements of both the additive and multiplicative models is used
- This model assumes that seasonal and irregular variations are both dependent on the level of the trend but independent of each other.

The original data can be expressed in the following form:

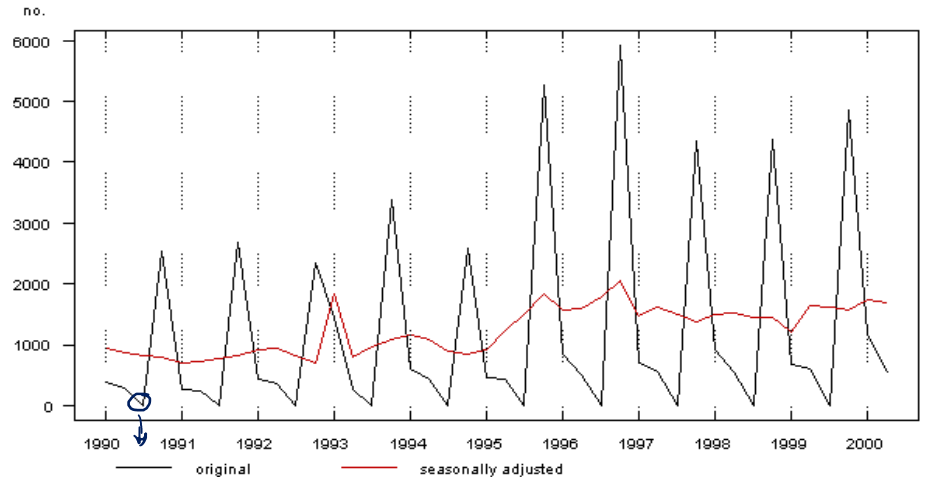
$$O_t = T_t + T_t \times (S_t - 1) + T_t \times (I_t - 1)$$
$$= T_t \times (S_t + I_t - 1)$$

- Both the seasonal factor S_t and the irregular factor I_t centered around one
- We need to subtract one from S_t and I_t to ensure that the terms $T_t \times (S_t - 1)$ and $T_t \times (I_t - 1)$ are centered around zero.
- These terms can be interpreted as the additive seasonal and additive irregular components respectively; the original data O_t will be centered around the trend values T_t .

COMPONENTS OF TIME-SERIES DATA contd.

- An example of series that requires a pseudo-additive decomposition model is shown below.
- This model is used as cereal crops are only produced during certain months, with crop production being virtually zero for one quarter each year.

Quarterly Gross Value for the Production of Cereal Crops



This model is used as cereal crops are only produced during certain months, with crop production being virtually zero for one quarter each year.

References

Text Book:

"Business Analytics, The Science of Data-Driven Making", U. Dinesh Kumar, Wiley 2017 (Chapter 13.1-13.2)

Additional reference and image courtesy:

https://www.abs.gov.au/websitedbs/D3310114.nsf/home/Time+Series+Analysis:+The+Basics

THANK YOU

Dr. Mamatha H R
Professor, Department of Computer Science
mamathahr@pes.edu

+91 80 2672 1983 Extn 834

Ms. Jyothi R.
Assistant Professor, Department of Computer Science
jyothir@pes.edu