DAY 2 - EDA & FE

25th/9/22

Recall

ML Pipeline

1) Data collection

2) EDA (Analysis)

3) Feature Engineering

4) Model building

5) Model evaluation (validation

EDA

1) profile the data

2) Stat based Analysis

3) Graphical Analysis

Pre processing (or) FE

1) Missing values - handle

2) outlier - "

3) Scaling data "

4) transforming the data

5.) encoding

6.) Imbalance data handling

7.) Feature Selection

8) Dimension Reduction [PCA, LDA, tSNE]

9.) duplicate handling

10.) Split / merge / drop / Add.

pre-processing → Affect → Model building

way of performing FE

1) Missing value - Handle

In ML we have Diff types of technique to handle the missing values

1) to add the Random value

2.) forward filling / backward filling

3) Stat (mean, median, Mode)

4.) end of the Distribution value

5) Drop the Particular row

6) Knn - Imputer

7) ML Algorithm which support missing value handling

⋮

etc.

2) outlier handle

1) Detect the outlier

2) handle the outlier

Detect	handle
1) zscore	1) drop
2.) IQR	2) help of median
3) Box, Scatter, Violin Plot	3) replace any value for outlier
	4.) trimming

3) Transformation

1) box-Cox 2)

2) Power Transformation

3) log

4) Square

5) Cube

6) ...

etc

4) Scaling

1) Standardization

2) Mini-max Scalar

3) Unit Scaling

etc

5) encoding

1) one hot

2) Lable encoding

3) binary encoding

4) Target guider encoding

5) Hash encoding

6) Embalance Data Technique

* 1) Collect more Data

2) Under Sampling

3) over Sampling

4) Cluster based over Sampling

Data → EDA → Preprocessing → model 85%

&

50%

1) missing →

2) outliers →

3) Scale →

4) encoding →

If we want more -%

then we have to

change the preprocess

technique.