

DAY 2 - EDA & FE

25th 4/22

Recall

ML Pipeline

- 1) Data Collection
- 2) EDA (Analysis) (1)
- 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 - 1/1
- 3) Scaling data *
- 4) transforming the data

5) encoding

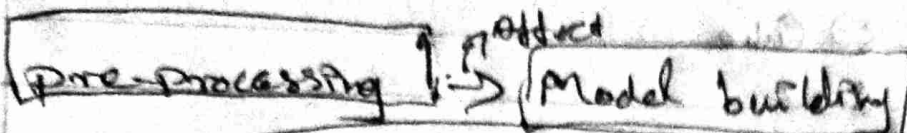
6) Imbalance data handling

7) Feature Selection

8) Dimension Reduction [PCA, LDA, tSNE]

9) duplicate handling

10) Split / merge / drop / add



1) Missing value handling

As 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

2) outlier handling

etc.

1) Detect the outlier

2) handle the outlier

Detect

handle

1) Z score

1) drop

2) IQR

2) help of median

3) Box, Scatter

3) replace with value for outlier

4) Outlier Plot

4) trimming

3) Transformation

- 1) box-Cox
- 2) Power Transformation
- 3) log
- 4) square
- 5) Cube
- 6) etc

4) Scaling

- 1) Standardization
- 2) min-max scalar
- 3) unit scalar
- etc

5) Encoding

- 1) one hot
- 2) label encoding
- 3) binary encoding
- 4) target guided encoding
- 5) Hash encoding

6) Imbalance Data Technique

- 1) Collect more data
- 2) Under sampling
- 3) Over sampling
- 4) Cluster based over sampling

Data \rightarrow EDA \rightarrow Preprocessing \rightarrow model fit

1) missing \rightarrow

2) outliers \rightarrow

3) Scale \rightarrow

4) encoding \rightarrow

If we want ^{80%} \rightarrow

then we have to

choose the Preprocessing
Technique.