Core ml Pipeline

1. Data colled-
2. EDA (Analysis)
3. Preprocessing or FE
4. model building
5. evalution matrix or validation

EDA

1. Profile
2. Stats based analysis
3. graph based analysis (Python)

Preprocessing

1. missing value
2. Outlier handle
3. Scale
4. transformation
5. encoding
6. handle imbalanced
7. feature Selection
8. dimension reduction (PCA, LDA, tsnE)
9. duplicate value | duplicate col
10. Split | merge | drop | Add → Data → avg

Preprocessing → model

Way of performing feature eng

(1) missing value handle

1. Random
2. forward filling | backward fill
3. Statical approach
   - mean, median, mode
4. end of the distribution
5. drop that row
6. knn-imputer
7. can we take that ml algo which missing value
8. Own ML model you can Predict.

(2) Outlier.

detect
- z-score
- IQR
- box plot
- Scatter Plot.
- Violin Plot

handling
- drop
- median
- replace | trimming

(3) transformation

- box-Cox
- Power transformation
  - log
  - Squre
  - Cube
  - Yeo Johnson.

(4) Scaling

- Standzation
- min-max
- Unit Scaling

(5) encoding

- One hot
- label encoding
- (Binray encoding)
- traget guided encoding
- Hash encoding

(6) imbalanced

1. Collect more.
2. under Sampling
3. over sampling
4. Cluster based over sampling

Dynamic Preprocessing

Data → EDA → Preprocessing → model → 75.1 → 77.1 → 80.1

1. missing →
2. outlier →
3. Scale →
4. encoding →

Structure : Image - text

1. Profile
2. stats
3. graph.

10 dataset-

One dataset

read data

Dataframe

1. EDA

Observation

1. Univariate
2. bivariate
3. multivariate

Interview

= 90 % → EDA | Preprocessing

Single folder

↓

github

↓

2. PreProcessing

- IPYNB 1. missing = 10 - 12
- IPYNB 2. Outlier Step 5
- IPYNB 3. Scaling / transform
4. Encoding
5. feature selection
6. Imbalance data

Automate EDA

- Pandas
- Sweetviz
- autoviz

Pip