EDA & FE Part-2

26/02/23

Core ML Pipeline

1. Data Collection
2. EDA (Analysis)
3. Preprocessing or FE
4. Model Building
5. Evaluation Matrix or Validation

EDA

1. Profile Building
2. Stats Based Analysis
3. Graph Based Analysis (using Python)

Preprocessing

1. Missing Values
2. Handle outliers
3. Scale
4. Transformation
5. Encoding
6. Handle Imbalanced data
7. Feature selection
8. Dimension Reduction (PCA, LDA, tSNE)
9. Duplicate value / Duplicate column.
10. Split / merge / Drop / Add

Preprocessing → Directly Affects → Model Building

Different Ways of performing FE

1) Handle Missing Values :→

Different Techniques :→

i) Fill it with any Random Value

ii) Forward Filling / Backward Filling

iii) Statical Approach :→ Can fill with mean, median or mode.

iv) "End of distribution" :→ Can fill the data with the help of end of distribution.

v) Drop that row

vi) KNN - Imputer

vii) Take M.L Algo which support missing values.

viii) Create your own model and can predict missing value.

2) Outlier :→

Detection

i) Z-Score
ii) IQR
iii) Boxplot
iv) Scatter Plot
v) Violine Plot etc

Handling

i) Drop
ii) Fill with median
iii) Replace with any value / Trimming that Part

→ Since outliers affects 'mean' hence we can't choose 'mean' to replace the outliers so we have choosen 'median'.

3) Transformation of Data

i) Box-Cox transformation
ii) power transformation
iii) Log transformation
iv) Square root transformation
v) Cube root transformation.
vi) Yeo Johnson transformation.

4) Scaling

i) Standardization
ii) min-max scaler
iii) Unit scaling

5) Encoding :→

i) One hot encoding
ii) label encoding
iii) Binary encoding
iv) Target guided encoding.
v) Hash encoding.

(6) Imbalanced Data: → Inside the particular column if class ratio is mismatching i.e called Imbalanced data.

Different technique: →

(i) Collect more data
(ii) Under Sampling
(iii) Over Sampling.
(iv) Cluster based over Sampling.

(*) Dataset → Perform → EDA

Preprocessing → Dynamic

(i) missing nos. → mean/mode/median
(ii) Outliers → Boxplot
(iii) Scaling → min-max
(iv) Encoding → one hot encoding

↓

Model → 75%.

(*) Suppose after applying some technique we have done preprocessing and we have done model building & we get → 75% accuracy.

If Using some other technique if we do preprocessing and then after model building we get → 80% Accuracy

~~Hence~~ and if keeps increasing or decreasing.

(*) Hence we can say that preprocessing or FE is completely Dynamic. It is a research area where after every few days new techniques keeps coming

(*) Whether it is structured data or text data or image data we have to perform preprocessing / FE over the data.

Q. Then how can we build the best model?

By doing permutation and combination and it comes with practice.