DAY-8
EDA & Feature engineering

DAY 2 OF EDA & F.E

Revision of Day-1:-

Core ML Pipeline

=> Data ingestion
=> EDA (Analysis)
=> Feature engineering or Pre-Processing
=> Model Building
=> Evaluation matrix or validation

EDA:

(1) Profile of the DATA
2. Statistical Based Analysis
3. Graph Based Analysis (important)

PreProcessing

1. Missing value handle
2. outlier handle
3. Scaling of Data
4. Transformation (Log, Cox-Box, Square, Cube)
5. Encoding
6. Handle Imbalanced Data
7. Feature Selection
8. Dimension Reduction
9. Duplicate value / Duplicate column
10. Split / Merge / Drop / Add (column)

→ Directly effecting Model Building

Pre Processing → Model

Ways of Performing Feature engineering

(1) Missing value Handle

- → Fill with Random no.
- → Forward / Backward filing.
- → Statistical Approch (Mean, Median, Mode)
- → with the help of end of Distribution, fill the Missing value.
- → Drop the Row
- → Impute with KNN (KNN-IMPUTER)
- → ML Algorithm for Missing value
- → Build own ML Model to Predict Missing value.

(2) Outlier Handling

Detect the outliyer

- Z-Score
- IQR
- Box
- Scatter-Plot
- Violin-Plot

Handle outlier

- Drop
- Fill with Median
- Replace / Trimming

3) Transformation

1. Box Cox Transformation
2. Power Transformation
3. log
4. Square
5. cube

4. Scaling of Data

i) Standardization
ii) Min Max Scaler
iii) unit Scaling

5) Encoding Various Method

1. one-hot encoding
2. Label - Encoding
3. Binary Encoding
4. Target Guided Encoding
5. Hash Encoding

6. Imbalanced Dataset (Treatment various Method):-

1. Collect More Data
2. under Sampling
3. over Sampling
4. Cluster Based over Sampling

Q. How to find the Best Model Accuracy various Method?

⇒ To increase the Accuracy, we need to change the Pre-Processing technique & use different Method or Step from the above.

⇒ we needs to use each & every Pre-Processing Step & find Best Accuracy.

Q. How do we transform the Data?

⇒ import numpy as np
np.log(df)
sns.distplot(df)

Q. Does Outlier effect Mean?

⇒ Yes, outlier effect the mean.