EDA PART-2.

FEATURE ENGINEERING

Feature Engineering AND Preprocessing Steps:-

- Handle the missing value.
- Handle the outlier.
- Scaling of data.
- Transformation [log, Box cox, square, Cube]
- Encoding
- We can handle inbalance data.
- Feature selection.
- We can do Dimension Reduction [PCA, tSNE]
- Duplicate value, Duplicate column.
- Split / Merge / drop / Add column or Row.

① **Missing value Handle:-** Method to handle the missing value

①- Fill Random value.
②- Forward Filling / backward Filling
③- Statical Approach.
- Mean - median - Mode
④- End of the distribution.
⑤- Drop that row
⑥- Knn - imputer.
⑦- Can we take that ML algorithm which handle missing value
⑧- Own ML model to handle missing value.

② **Outliers:-**

Some steps to detect the outlier.

→ Z-score.
→ IQR → Scatter plot
→ box plot → Violin plot

Some steps to handle the outlier.

→ drop the outlier.
→ Fill the outlier with median.
→ Replace with any value.
→ ~~Trip~~ Trimming that part

(3) Transformation:-
- Box-Cox
- Power transformation.
- log
- Square root transformation.
- Cube root transformation.
- Yeo Johnson

(4) Scaling:-
- Standardization.
- min-max
- Unit Scaling

(5) Encoding:-
- One hot encoding
- label encoding
- Binary encoding
- Target Guided encoding
- Hash encoding

(6) Handle Imbalanced Data
→ Collect more data.
→ Undersampling
→ oversampling.
→ Cluster based oversampling

(*) **Bernoulli Distribution:-** Outcomes are binary [0 or 1]

$p = P(H) = 0.5$ $\quad\quad$ $q = P(T) = 0.5$

$P(H)$ – Probability of Head.
$P(T)$ – Probability of Tail.

$$p = 1 - q$$

In any Events, when the outcomes are binary, we will considered it as **Bernoulli's Distribution:-**

Most of the classification Problems are Bernoulli Distributed.

(*) **Binomial Distribution:-**

Exp 1	Exp 2	Exp 3	Exp 4	Exp 5
↓	↓	↓	↓	↓
Binary outcome	Binary outcome	Binary outcome	Binary outcome	Binary outcome

Binomial Distribution summarizes the number of trials, or observations when each trial has the same probability of attaining one particular value.

(*) **Power Law Distribution:-** (80-20 Rule)

The distribution that follows this power law distribution is basically called as **Pareto Distribution:-**

⇒ Interview Question:- How can we convert pareto distribution into normal Distribution.

Ans:- By Box cox Transformation.

The Box Cox transformation is defined as.

$$T(Y) = (Y \exp(\lambda) - 1) / \lambda$$

where Y is the response variable and λ is the transformation parameter. λ varies from -5 to 5. In the transformation, all values of λ are considered and the optimal value for a given variable is selected.

df['Age_Boxcox'], parameters = stat.boxcox(df['Age'])