

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 imbalance 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 low
- 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
- box plot
- Scatter plot
- Violin plot

Some steps to handle the outlier.

- drop the outlier.
- Fill the outlier with median.
- Replace with any value.
- 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]

$$p = P(H) = 0.5$$

$$q = P(T) = 0.5$$

$P(H)$ - Probability of Head.

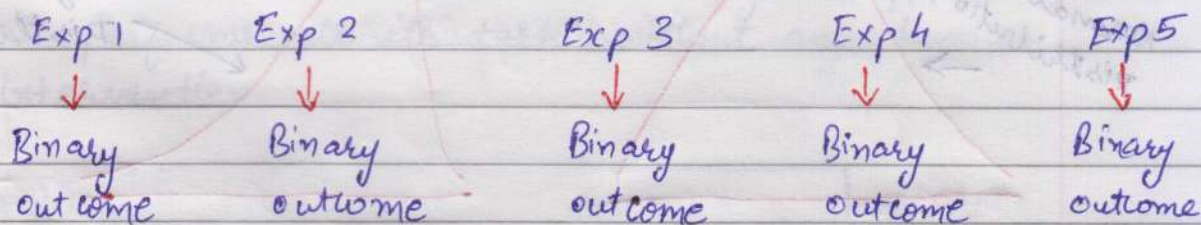
$P(T)$ - Probability of Tail.

$$p = 1 - q$$

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

Most of the Classification Problems are Bernoulli Distributed.

(*) Binomial Distribution:-



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.

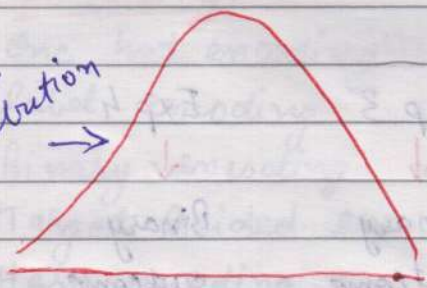
$$T(Y) = (Y \exp(\lambda) - 1) / \lambda$$

where Y is the response variable and λ is the transformation parameter. λ varies from -5 to 5.

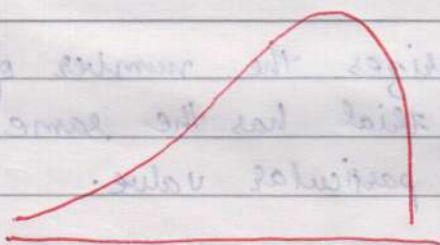
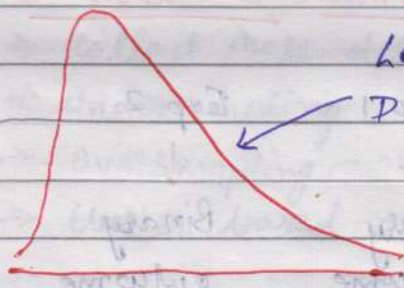
In the transformation, all values of λ are considered and the optimal value for a given variable is selected.

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

Normal
Distribution



Log Normal
Distribution



power Law Distribution.