EDA & Feature Engineering

Data Science life cycle:

1. Data ingestion
2. EDA (Expletory Data analysis)
3. Processing (pre)
4. Model building
5. Evaluate & validation

Statistics: Collect data, organise data, Interpretation data, analysis data.

We can get some meaning full insight from above.

Ex. Problem statement:

Sales of product is going down

What is reason behind sales are going down it can be product, payment to customer, Marketing, Leadership, Competitor etc

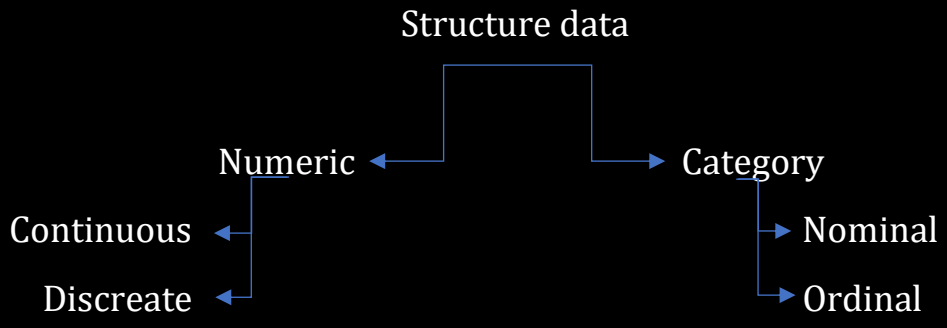
Data set → analysis → Conclusion

Types of data:

1.Structure data (Ex. table format data)

2.Unstructured data (Ex. Videos, image)

3.Semi structure data (Ex. XML, JSON)



Ex.

Feature 1	Feature 2	Feature 3
Weight	Height	BMS
75.6	152	23
85	165	30
55	151	32
Continuous	Continuous	Continuous

Continuous: a value in decimal point is called Continuous.

Ex. Life of bulb so life of bulb it can be any thing 1 year 2 month 3 days 3hr 26 min 56 sec 59 milli second and continue.

Discreate: a whole number is called discreate.

Ex. Number of student present in class

Nominal: Order does not matter.

Ex. Male, Female here order does not matter

Ordinal: Order matters.

Ex. 10^{th}, 12^{th}, UG, PG, PHD here order matter.

Ex.

Name	Age	Height	Sex	Weight	Education
Dipak	25	152	M	70	UG
Priya	36	156	F	65	UG
Raj	30	1556	M	85	PG
Harry	29	169	F	69	PHD
Aditya	26	163	M	86	10^{TH}
karan	39	180	M	90	12^{TH}
Categorical	Numerical	Numerical	Categorical	Numerical	Categorical
Nominal	Continuous	Continuous	Nominal	Continuous	Continuous

Univariate: single column

Bivariate: 2 columns

Multivariate: more than 2 columns

Ex. If a person does consistent study he will get good marks

Independent: study is independent

Dependent: marks are dependent

Profile of data:

1. No of rows
2. No of columns
3. Missing values
4. Categorical variable
5. Numeric variable
6. Duplicate
7. Datatype
8. Memory

Statistics base analysis:

1. Variation
2. Covariance
3. Std
4. Correlation
5. Chi sq. test
6. T test
7. Z test
8. Anova test
9. Mean, median, mode etc.

Graph base analysis

1. Box plot (we can check outlier, distribution of data)

2. Scatter plot (outlier, linear)
3. Pie plot
4. Histogram (distribution of data)
5. KDE plot
6. Count bar (bar chart)
7. Heat map (correlation)

Conclusion: base on EDA we can do processing of data.

Suppose we have found missing value next.

Feature engineering/pre-processing of data.

1. Missing value handle
2. Outlier handle
3. Scaling of data
4. Transformation (log, box-cox, sq. root, cube root)
5. Encoding
6. Imbalance data
7. Feature selection
8. Dimension reduction (PCA, tsne)

If there is null value in my data set we can handle it using missing value.

If there is out liar in my data set we can find and remove.

Categorical (man, woman) → encoding

Skewed range → scale (with a certain range)

Count of feature → handle imbalance data

Suppose we have 1000 set but we use only 200 subsets then we can say we select 200 feature → feature selection

Suppose we have X and Y feature we combine them together in the single feature → Dimension reduction (PCA)

For EDA

1. Profile
2. Stat base analysis
3. Graph base analysis

Pre-processing of data:

1. Missing value
2. Outlier handle
3. Scale
4. Transformation
5. Encoding
6. Handle imbalance data
7. Feature selection
8. Dimension reduction (PCA)
9. Duplicate value/duplicate column
10. Split/merge/drop/add

Way of performing feature engineering

1. Missing value handle

- Random value
- Forward filling/ backward filling
- Stat approach (mean, median, mode)
- End of distribution
- Drop that row
- KNN imputer
- Can we take that ML algorithm that support missing value.
- Own ml model for this

2. Outlier

A) detect outlier

- Z score
- IQR

- Boxplot
- Scatter plot
- Violin plot

B) Handling

- Drop
- Median
- Replace/trim that part

3. Transformation

- Box-cox
- Power
- Log
- square
- cube
- yeo Johnson

4. scaling

- standardization
- min max
- unit scaling

5. encoding

- one hot
- label encoding
- binary encoding
- target guided encoding
- hash encoding

6. imbalance (class ratio mismatch that time we use)

- collect more data
- under sampling
- over sampling

- cluster base over sampling

Thank You.

- krish naik sir
- Sudhanshu kumar sir
- sunny Savita sir
- ineuron.io