24 Sep22 / Saturday

EDA (Exploratory Data Analysis) & Feature Engineering

EDA - EDA is used by data scientists to analyze and investigate data sets and summazie their main characteristic, often employing data visualization methods.

Data Science Life cycle -

1.) Data ingetion —— Project

(big data Tools), remote location (Sql, nosql), Some file format csv, tsv, Xml, json, website

→ HDFS, NOSQL, Kafka, spark

2.) EDA (analysis)

3.) Processing (Pre)

4.) Model

5.) Evaluate and validate

Statistics -

Collect , Organise, Interpretation, Analysis

↑ Insight

(Scientific, healthcare, Social problem)

Data Types -

Batch data, Streaming data

Batch → Minibatch data (little more frequent)

Batch data → Historic data (Periodic)

Streaming data → Continuous data

1.) Structure data — Table

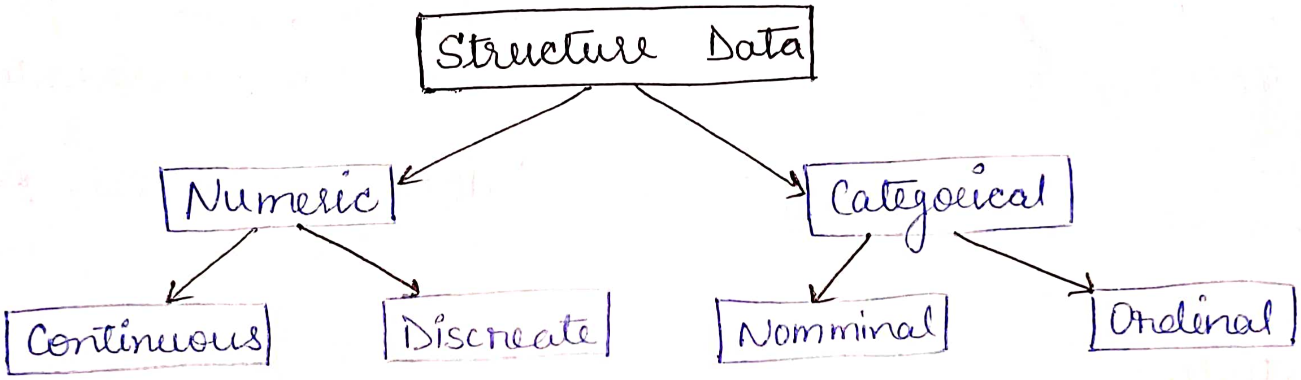
2.) Unstructure data — Video, Images, text

3.) Semi structure data — XML, Json

EDA + FE →

Structure Data -

Weight	Height	BMI
70	170	22
80	180	24
90	190	26
100	200	30
60	160	21


Structure Data

Numeric

Categorical

Continuous

Discreate

Nomminal

Ordinal

STUDENT PERFORMENCE →

Multivariate

Bivariate

Univariate

FEATURE

Name	Age	Height	Sex	Weight	Education
Sunny	25	170	Male	70	UG
Arijit	30	180	Male	80	PG
Priyam	35	160	Male	60	UG
Priya	20	150	Female	55	Phd
Aditi	27	145	Female	58	PG
Categorical ↓ Nominal	Numerical ↓ Continuous	Numerical ↓ Continuous	Categorical ↓ Nominal	Numerical ↓ Continuous	Categorical ↓ Ordinal

EDA - TYPE OF DATA

Univariate - Single coloumn
Bivariale - Two coloumn
Multivariate - More than two coloumn

Independent / Dependent -

Age, height, Sex — Dependent [Weight]
Age, height - Independent

Data - Analysis
1) Missing value
2) Outliers
3) Scaling
} Feature/ Coloumn

Core ML Pipeli-ne
1) Data ingestion
2) EDA
3) Preprocessing → FE
4) Model building
5) Evaluation & validation

First EDA is required on FE ~~or~~ PP? - EDA → PP/FE

Name	Age	Education	Salary	Exp
Sunny	25	UG	25K	2
Deepak	30	PG	30K	3
Rushi	40	UG	40K	5
Aman	50	Phd	50K	10
Shalini	20	UG	35K	1

Steps -
EDA (Analysis)
→ Profile of the data
→ Statistical analysis
→ Graph based analysis

Profile of the data

Row

Coloumn

No of missing value

Categorical

Numerical

duplicate

Data type

RAM

Data Size

Graph based analysis

Box plot

Scatter plot

Pie plot

Histogram

KDE [Kernal density]

Bar Chart

Heatmap

Stats based Interpretation

Varience

Covarience

Standard deviation

Correlation

Chi-squre

T-test

Z-test

anova test

mean/median/mode

★ Based on EDA can we do processing of the data?

Yes

STEPS FOR FEATURE ENGINEERING

1.) Missing value handle

2.) Outliers handle

3.) Scaling of data

4.) Transformation (log, Boxcox, squre, cube)

5.) encoding

6) Imbalance data

7.) Feature selection

8.) Dimention reduction (PCA, ESnE).

9.) Duplicate Value/Coloumn

10.) Split/merge/drop/add

25 Sep 22 / Sunday

WAY OF PERFORMING FEATURE ENGINEERING —

1.) Missing Value handle

1.) Random no filling
2.) Forward filling / backward filling
3.) Statistical approach - mean, median, mode
4.) end of the distribution
5.) drop the row
6.) Knn - imputer
7.) Can we take that ML algorithm which missing value.
8.) Can create own ML model and predict missing value.

2.) Outlier

detect

Z-Score
IQR
box-plot
Scatter plot
Violin plot

handling

drop
median
replace / trimming

3.) Transformation / Sc

box-cox
power transformation
Log
Square
Cube
Yeo Johnson

4.) Scaling

Standrization
Min/Max
Unit scaling

5.) Encoding

One hot
label encoding
binary encoding
target guided encoding
hash encoding

6.) Imbalanced

collect more data
Undersampling
Oversampling
Cluster-based oversampling