Sri.

27/09/22

EDA & FE

Exploratory Data Analysis, Feature Engineering

Data science Life cycle :- 1. Data ingestion [Project] → big data tools

2. EDA [Analysis]

3. Processing [Pre Processing] ← (math)

4. Model

5. Evaluate (Validate)

remote location, some file format csv, tsv, xml, excel website → HDFS, NoSql, Kafka, Spark streaming.

Statistics :- Collect, organize, interpretation analysis

insight

scientific, ascientific, social problem, healtcare

Sales or Product :- Sales is going down

Read

healthcare

y

sales

product, paying to customer, leadership, marketing competetor.

dataset → analysis → conclusion

Project manager
Business analy
date scientist } domain expert

Types of data :- Batch data, Streaming data } data

Batch data → historic data (Periodic)

↳ min batch data (little more frea)

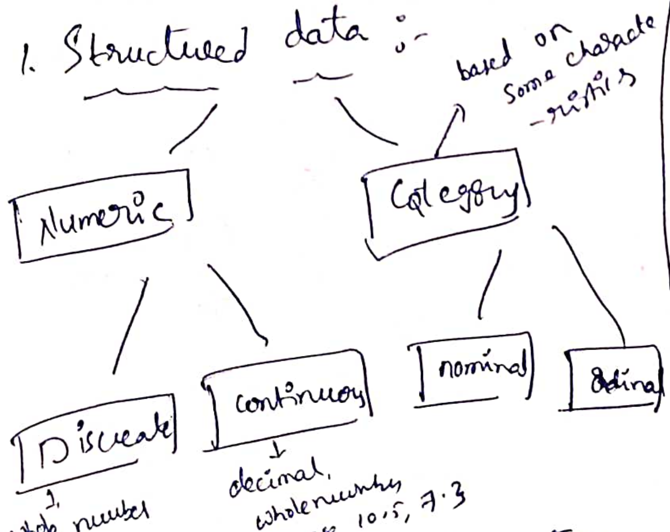
Streaming data → continuous data → live

1. Structured data → tables (RxC) ⇒ ML

2. Unstructured data → Video's, images, voice, sound, text, ⇒ DL

3. Semi structure data → xml, Json

1. Structured data :-



Numeric
Category
based on some characte-ristics
Discreate
continuous
nominal
Ordinal
whole number
decimal, wholenumbers eg 10.5, 7.3

feature 1	2	3
weight	Height	BMI
70	170	22
80	180	24
90	190	26
100	200	30
60	160	21
↓ continuous	↓ continuous	↓ continuous

Segragate the feature

Continuous :- Continuous in nature

eg: Height 160 160.5, 160.55

[160 161]

Nominal :- Oder doesn't matter

eg male, female

Ordinal :- Ex: Qualification education in order
↓
Order matters here

10th
12th
degree
PG
PhD

Practical implementation :- Student Performance

Feature

Name	Age	Height	Sex	weight	education
Sunny	25	170	Male	70	UGD
Arijit	30	180	Male	80	PG
Prijam	35	160	Male	60	UG
Priya	20	150	Female	55	Ph.d
Aditi	27	145	Female	58	PG

Name ↓ Categorical ↓ nomial (no order matter)

Age ↓ Numerical ↓ continuous

Height ↓ numeric ↓ continuous

Sex ↓ Categorical ↓ Nominal

weight ↓ Numerical ↓ continuous

education ↓ Categorical ↓ Ordinal → [here Order matter]
↓
Ex
UG - 0
PG - 1
Phd - 2

EDA → types of data ?

Univariate → single Column

bivariate → two column

Multivariate → more than 2 column

Ex if we want check only Height → Univariate

if " " " Height, with age → bivariate

if " " " Height, weight & age → multivariate

EDA ⇒ Analysis of data

FE / preprocessing ⇒ Cleaning the data

NAME	AGE	Education	Salary	experience
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

EDA → Analysis of data

- Create the Profile of data
- Statical analysis
- Graph based analysis

- Create the Profile of data
  - Rows
  - columns
  - missing
  - Categorical
  - numerical
  - duplicate
  - D type
  - Ram

Statis based (inspolation)

- Variance
- Covariance
- Standard deviation
- correlation
- Chi square test
- t- test
- z- test
- Anova test
- mean / median / mode

} Uni, bi, multivariate

Graph based analysis:-

- Box plot → Outlier, distribution, → (Plotting) → } → dashbording
- Scatter plot → Outliers, linear
- Pie chart
- Histogram → distribution
- Kde plot → Kernal Density Estimater
- Heat map → (Corr)
- count bar B,C

data analysis

Observation / conclusion : Plots → Univariate, bivariance, multivariate

Based on EDA, can we do a processing of the data?

Yes:-

→ Can we handle, missing value handle

Pre Processing of data.

- Outlier handle
- Scaling of data
- Transformation of data (log, Box. cox, Square, cube)
- encoding
- Imbalance data
- feature selection

→ feature Engineering

- feature
- Dimension reduction (PCA, tsne)

Missing null Value → Missing value handle

PD

EDA

Outlier → handle

Cat, (man, woman) → encoding → conversion of categorial into numerical

Skewed range → scale (within a certain range)

Count of feature → handle imbalance

feature selection

Dimention reduction (PCA, tsne)

↓

$x_1, x_2 \rightarrow X$

100

1. 10
2. 40
3. 20
4. 10
5. 20

Ex 1000 → subset

↓

200

EDA

1. Profile
- States
- graph

Preprocessing

- Missing / null value
- Outlier
- Scale value
- Transform
- encode
- Imbalance
- drop / duplication
- Feature selection
- dimentionality reductions.

Automated tool in Python

Pandas Profiling, audoviz, auto viz

mito, sweet viz, Sweet viz

Knime

atleast use 3 automated tools max 5, with respect to 1 dataset.

EDA + FE => 10 in depth analysis => your git hub repo.

Sunny.Savita@ineuron.ai

Krish.naik@ineuron.ai

Automated EDA Libraries by Krish Naik Video :-

- Dtale
- Pandas Profiling
- Sweet viz
- Auto viz
- Lux → Paid one
- Dataprep
- Pandas_visual_analysis

25/09/22 → Core ML Pipelin

- Data collection
- EDA & Analysis
- Preprocessing /FE
- Model building
- Evalution matrix /validation

EDA.

1. Profile
2. Stats based analysis
3. Graph based.

Preprocessing :-

- Missing value
- Outlier handle
- Scale
- Transformation
- encoding
- Handle imbalance
- Feature selection
- Dimention reduction (PCA, LDA, TSNE) → Principle component Analysis.
- duplicate value /duplicate colu
- Split/Merge/drop/add

Preprocessing → Model

ways of Performing FE

1. Missing value handle:-

- Random
- forward filling / backward filling
- Statical approach (mean, median, mode)
- End of the distribution
- drop that row
- KNN - inputes
- Can we take that ML algorithm which missing value
- Create your own ML model, you can predict the missing value

2. Outlies handle:-

- detect the Outlies

- Z-score
- IQR
- Box Plot
- Scatter plot
- Violin Plot

handling

- Drop
- fill with median
- Replace / with any value / trimming

3. Transformation / Scaling of data

- box-Cox transformation
- Power-transformation
- log
- Square
- Cube
- Yeo Johnson

Scaling

- Standardization
- Min-max Scaler
- Unit Scaling

Encoding:

- One hot
- Label encoding
- Binary
- Target guided encoding
- Hash encoding

Imbalanced

- Collect more data
- Under Sampling
- Over sampling
- Custer based over sampling

Dynamic

Data → EDA Preprocessing → Model → 75%, 1, 80%, 77%

- Missing
- Out lies
- Scale
- encoding

- Structured data
- image data / text data / unstructed data