MON TUE WED THR FRI SAT SUN

Date: 24/09/2024

FEATURE ENGINEERING AND EDA (EXPLORATORY DATA ANALYSIS)

Data Science Life Cycle : (ML Life Cycle)

1) Data Ingestion (collection of Data)
2) EDA (Analysis)
3) Processing (Pre-Processing)
4) Model
5) Evaluate & Validate Model

→ Core ML Pipeline

STATISTICS : Statistics is the Science of Collecting, Organising and analyzing the data.

Collect, Organise, Interpretation, analysis
↓
Insight

Examples : Scientific Problem, Healthcare, Social Problem.

Problem Statements :

Sales of Product → Sales is going down.

Analysis → Product, Paying to Customer, leadership, Marketing, Competitor.

Dataset → Analysis → Conclusion (Domain Expert)

1. Project Manager
2. Business Analyst
3. Data Scientists

→ Used to get conclusion from the dataset.

Only Domain require :- EDA & FE

Scanned with CamScanner

MON TUE WED THR FRI SAT SUN

Page No.
Date / /

Data Ingestion : Collecting of Data.

Examples:- Big Data Tools, remote Location (SQL, NOSQL), Website, webscraping
Some file format: (CSV, XML, JSON, Excel)

TYPES OF DATA :

Data
- BATCH DATA : Historical Data, Mini batch Data (Periodic)
- Streaming Data: Continous Data (Live Data)

① Structure Data : Table (Row x Column) → ML

② UnStructure Data : Video, Images, Voice, Sound, Text → DL

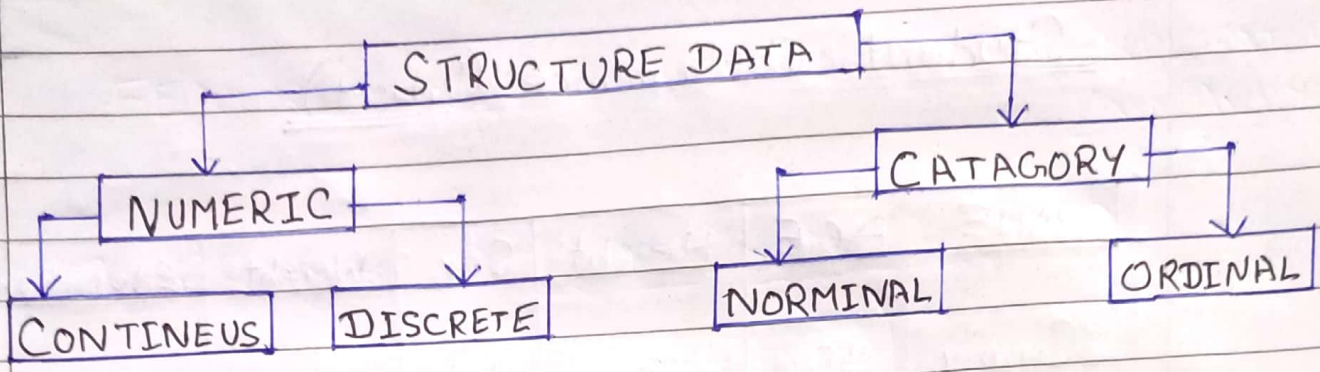
③ Semi Structure Data : JSON, XML

★ Example Of STRUTURE DATA :

(Feature 1) Weight	(Feature 2) Height	(Feature 3) BMI
70	170	22
80	180	24
90	190	26
100	200	20
60	160	21

∴ In Table, Column is also called Feature in Feature Engineering

Scanned with CamScanner



- Numeric Continues : Continuous in Nature (Eg :- Height, Weight)
- Numeric Discrete : Only Whole Number.
- Category :-

  Male, Female > Category

  Black, White > Category.

- Nominal Category : Order Doesn't Matter (Eg :- Male, Female)
- Ordinal Category : Order Matter here.

  Eg :-

  10th, 12th, Grad., Post-Grad., Phd.

Scanned with CamScanner

MON TUE WED THR FRI SAT SUN

Page No.
Date / /

Student Performance (Dataset)

NAME	AGE	Height	Sex	Weight	Education
Sunny	25	170	Male	70	UG
Arjit	30	180	Male	80	PG
Priyam	35	160	Male	60	UG
Priya	20	150	Female	55	Phd
Adita	27	145	Female	50	PG
Ashu	35	175	Male	85	PG
Vijay	21	140	Male	65	PG
↑	↑	↑	↑	↑	↑
Category	Numeric	Numeric	Category	Numeric	Category
↓	↓	↓	↓	↓	↓
Nominal	Continue	Continue	Nominal	Continue	Ordinal

USED To Find Types of Data :

UNIVARIATE : Single Column

Bivariate : Two Column

MULTIVARIATE : More than Two Column

INDEPENDENT AND DEPENDENT VARIABLE :

[Age , (Height , Sex)] , Weight

↓ Independent

↓ Dependent

Scanned with CamScanner

Q) First EDA is required or FE or (PreProcessing).

→ EDA is Required

EDA → Preprocessing ⇒ Model

↓

Imp (Impact)

Example : BRIYANI (MODEL)

Raw Data:
1. Chiken
2. Rice
3. Onion
4. Oil
5. Spices

EDA →

- 1 Kg
- 1 Kg
- 1 Kg
- 1 Ltr
- 500 gm

PreProcessing → Briyani MODEL → Validation → Taste

A) Above is Real Life Example of ML Model :-

1. EDA
2. PreProcessing

EDA : Analysis of the data based on the given data.

Q) If EDA and FE Both are Same or not ??

→ Both are Same,

Scanned with CamScanner

MON TUE WED THR FRI SAT SUN

Page No :
Date / /

DATASET:

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) :

1. Profile of the Data
2. Statistic Analysis
3. Graph Based Analysis

PROFILE OF THE DATA :

1. Row
2. Column
3. Missing
4. Category
5. Numeric
6. Duplicate
7. DType
8. RAM (Dsize)

$\Rightarrow$ Analysis

Scanned with CamScanner

MON TUE WED THR FRI SAT SUN

Page No :
Date :

STATISTIC Analysis (Interpretation) :

1. Variance
2. Covariance (cov)
3. Std.
4. Correlation
5. Chi Square Test
6. T - Test
7. Z - Test
8. Anova Test
9. Mean, Median, Mode

⇒ Univariate
Bivariate
Mutivariate

GRAPH BASED Analysis (Plotting) :

(EDA)
Observation matter for Conclusion

(also Import Dashboarding)

1. Box Plot ⇒ Outlier, Distribution, Statistical Profile
2. Scatter Plot ⇒ Outlier, Linear or Not
3. Pie Chart
4. Histogram ⇒ Distribution
5. KDE Plot
6. Count Bar ⇒ Used for Counting (Bar Chart)
7. Heat Map ⇒ Correlation

Q) Based on a EDA, can we do a PreProcessing of Data ??

→ True

Scanned with CamScanner

MON TUE WED THR FRI SAT SUN

Page No.
Date

Pre-Processing of Data :

No. of ways to Solves each

1. Missing Value Handle
2. Outlier Handle
3. Scaling of Data
4. Transformation (Log, Box-cox, Square, Cube)
5. Encoding
6. Imbalance Data
7. Feature Selection
8. Dimension Reduction (PCA, tsnE)
9. Duplicate Value / Duplicate Column.
10. Split / Merge / Drop / Add (column)

These are Steps of Feature Engineering :

1. Missing Null Value (EDA) → Missing Value Handle (PP)
2. Outlier → Handle
3. Categorical (Man, Women) → Encoding
4. ~~Sk~~ Skewed Range → Scale (Within a Certain Range)
5. Count of Feature
   - → Handle Imbalanced Data
   - → Feature Selection
   - → Dimension Reduction (PCA, tsnE)

Scanned with CamScanner

MON TUE WED THR FRI SAT SUN

Page No :
Date : / /

Encoding : To Change Categorical Data into Numerical Data is called Encoding.

SOME AUTOMATED TOOLS IN PYTHON FOR EDA:

1. Pandas Profiling
2. Sweet Viz
3. Auto Viz
4. D-Tale
5. Mito
6. Knime
7. Dataprep

Automated EDA Visualization Tools in Python.
&
Many More Tools are there...

Thank You

Scanned with CamScanner

MON TUE WED THR FRI SAT SUN

Page No.
Date 25/09/2022

EDA & FE (Day-2)

Revision of Day-1 :

CORE ML PIPELINE :

1. Data Ingestion (Data Collection)
2. EDA (Analysis)
3. Pre-Processing or Feature Engineering
4. Model Building
5. Evaluation Matrix or Validation

EDA :

1. Profile of the Data
2. Statistical Based Analysis
3. Graph Based Analysis (V.V.I in EDA)

Pre Processing :

1. Missing Value Handle
2. Outlier Handle
3. Scaling of Data
4. Transformation (Log, Box-Cox, Square, Cube)
5. Encoding
6. Handle Imbalanced Data
7. Feature Selection
8. Dimension Reduction (PCA, LDA, tsnE)
9. Duplicate Value / Duplicate Column
10. Split / Merge / Drop / Add (Column)

Scanned with CamScanner

[PreProcessing] → [Model]

WAYS of Performing Feature Engineering :

① Missing Value Handle :

i) Fill with Random Number.
ii) Forward / Backward Filing
iii) Statistical Approach (Mean, Median, Mode)
iv) With the help of end of Distribution, fill the Missing Values
v) Drop the row.
vi) Impute with KNN (KNN-Imputer)
vii) ML Algorithm for Missing Value
viii) Build own ML Model to predict Missing Value.

② Outlier Handling :

Detect the Outlier :

- Z-Score
- IQR
- BOX-Plot
- Scatter-Plot
- Violin-Plot

Handle Outlier :

- Drop
- Fill with Median
- Replace / Trimming

Scanned with CamScanner

③ TRANSFORMATION :

i) Box Cox Transformation
ii) Power Transformation
iii) Log
iv) Square
v) Cube
vi) Yeo Johnson

④ Scaling of Data :

i) Standardization
ii) Min Max Scalar
iii) Unit Scaling

⑤ Encoding Various Method :

i) One Hot Encoding
ii) Label Encoding
iii) Binary Encoding
iv) Target Guided Encoding
v) Hash Encoding

⑥ Imbalanced Dataset Treatment Various Method :

i) Collect More Data
ii) Under Sampling
iii) Over Sampling
iv) Cluster Based over Sampling

Scanned with CamScanner

MON TUE WED THR FRI SAT SUN

Page No :
Date : / /

How To Find the best Model Accuracy Various Method.

① To Increase the Accuracy, We need to change the Pre-Processing technique and use different method or Steps from the above.

② We Needs to use each and every Pro-Processing Steps and Find the best Accuracy.

Q) How do we transform the data ?

⇒ Import numpy as np
np.log (df)
sns.distplot (df)

—— X —— X —— Thank You —— X —— X ——

Scanned with CamScanner