EDA & FE

26/04/22

Core M.L Pipeline

Data Science Lifecycle :

Statistics → Core ML Pipeline

1. Data Ingestion :- from where are we getting the data
2. EDA (Analysis over data)
3. Pre Processing → Before Model Building clean the data
4. Model Building → Mathematical eqn.
5. Evaluation & validation of Model

Statistics :- Collect, Organise, interpretation, analysis of data.

Insight of Data.

(*) In every field like, Scientific, healthcare, Social problems etc we use Statistics to get the insight about data.

(*) In a company → Sales of product is going down. I.I.T Sales,

Reason:

(i) Product is not good
(ii) Not Paying attention to Customer
(iii) Leadership is not good with the Project
(iv) Marketing Strategy is not good
(v) Not looking to Competitors

Dataset → Analysis → Conclusion

why sale is down?

EDA & Feature Engineering

① Project Manager
② Business Analytics
③ Data scientist

} They will look into the Dataset and give the conclusion why sale is down and will give benifit to company.

Conclude:- Any domain require EDA & Feature Engg.

Data science Lifecycle :→

① Data Ingestion:- Retrieve data from → HDFS, MOSQL, kafka, sparkstreaming big data tools, remote location (SQL, MoSQL) some file format, CSV, TSV, XML, JSON, excel, scrape from a particular website.

Types of Data:→ Batch Data, Streaming Data } Data

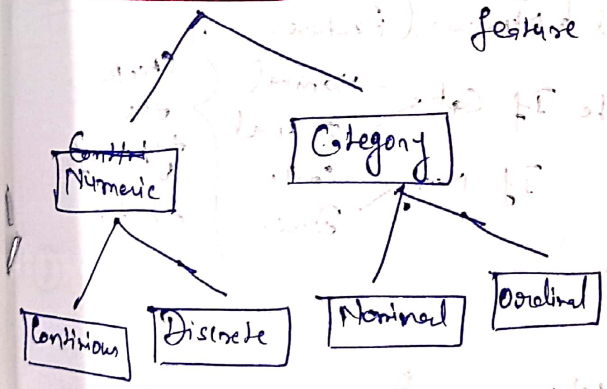
Batch Data → Historical Data or (Periodic Data)
Mini Batch data (little more frequent)
Streaming Data → Continious / Live Data

Data →
① Structured Data :→ Table → M.L (Machine Learning)
② Unstructured Data :→ videos, Images, voice, sound, text → DL (Deep Learning)
③ Semi Structured Data → XML, JSON

② We can Perform EDA + FE in all types of data.

① Structure Data :→

Segregate the feature



Feature1	Feature2	Feature3
Weight	Height	BMI
70	170	22
80	180	24
90	190	26
100	200	30
60	160	21
↓	↓	↓
Continious	Continious	Cont

Continious Data :→ Continious in Nature. Eg:- Height could be 160, 160.5, 160.55 Means b/w 2 values Suppose b/w 160 & 161 there could be infinite no. of values i.e called Continious data.

Discrete data :→ Whole no. Eg:- no. of students in class i.e 10, 20, 50, 100 etc. We can't say 10.5 students in class.

Categorical Data :→ Different-different Category. Suppose,

Gender < Male, Female Colour < Black, White

Nominal :→ Order doesn't matter. eg:- Male, Female
(Categories not in order)
Here order doesn't matter whether we write Male first or female first.

Ordinal :→ Order Matters. Eg:- Degree :- 10th, 12th, Grad, PG, PHD } Order Matters

Categories in order.

Practical Implementation of Dataset

Student Performance Multivariate, Univariate, Bivariate

Feature

Name	Age	Height	Sex	weight	Education
Sunny	25	170	Male	70	UG
Abijit	30	180	Male	80	PG
Paijam	35	160	Male	60	UG
Priya	20	150	Female	55	PHD
Aditi	27	145	Female	58	PG

Class

While selecting give more weightage to Class: UG - 0, PG - 1, PHD - 2

Categorical Nominal | Num Cont. | Num Cont. | Cat Nominal | Num Cont. | Cat Ordinal

EDA

First Level :→ Identify Categorical & Numerical Feature

2nd Level :→ Further Segregate If Cat < Nominal, Ordinal
If Num < Cont, Disc
} Check the type of Data.

Univariate → Single column
Bivariate → Two column
Multivariate → More than 2 column

Height → Univariate Analysis
Height w.r.t Age → Bivariate Analysis
Height, Age & Sex → Multivariate Analysis

① Independent / Dependent Variable :→

Eg:- Age, Weight, Sex → Weight

Independent | Dependent

(not dependent on any other feature) | Dependent

Weight is dependent in all three 3 feature. So weight is a Dependent Variable

Core ML Pipelines:
1. Data Indigestion
2. ~~Data~~ EDA → Analysis of data. (Based on feature)
3. Preprocessing → (Feature Engg.) → cleaning or wrangling or treating of data.
4. Model building
5. Evaluation or validation of Model

Data → Analysis
Feature / Column
1. Missing value
2. Outlier
3. Scaling

Changes in data ↓ Engg. on features ↓ Feature Engg.

→ First EDA → Preprocessing or FE → impacts → Model Building

Real life Example :→ Cook Biriyani

Things needed | Quantity | Cleaning

Raw Data →
1. Chicken → 1kg → Cleaning
2. rice → 1kg → Cleaning
3. Onion → 1kg → Cutting
4. oil → 1 liter → Pouring
5. spices → 500g → Cooking

EDA

Pre Process → Model: Biriyani → Testing

Test / validation

& are FE and EDA same ? → YES

EDA (Analysis) :→ Dataset :→

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

Steps
1. Profile of Data
2. Statistical Analysis
3. Graph based Analysis

① Profile of Data :→ Person Profile { Name, Age, Place, Designation etc }

i) no. of rows
ii) no. of Column
iii) Missing Values
iv) Categorical Variable
v) Numerical Values
vi) Duplicates Values
vii) Datatype.
viii) Data size (RAM)

② Statical Analysis :→ Interpretation of Data (Statistical Conclusion)

Univariate or Bivariate or multivariate.
i) Variation of Col.
ii) Covariance of Col.
iii) Standard deviation
iv) Correlation of Data (col)
v) Chi square Test
vi) t-test
vii) z-test
viii) Anova test
ix) Mean/median/mode

③ Graph Based Analysis

Observation
i) Box Plot → outlier or distr of data
ii) Scatter Plot → outlier, linear
iii) KDE
iv) Histogram → dist of data
v) Pie chart
vi) Count Bar (Barchart) → no. of no. of Rows, Col.
vii) Heatmap → Corr. b/w Variable

Conclusion:→ Based on EDA we can do processing of the Data.

Steps for Preprocessing or Feature Engg:→

Preprocessing of Data or Feature Engg:

(i) Handle Missing Values
(ii) Handle outlier
(iii) Scaling of data
(iv) transformation of Data (log, Boxcox, Square, cube)
(v) Encoding of Categorical data
(vi) Imbalance Data
(vii) Feature Selection
(viii) Dimension Reduction (PCA, tSNE)
(ix) Drop Duplicates

EDA (only Analysis) FE (Change the value of features)

Finding Missing Values → Handling Missing Values
(EDA) (FE)

Finding outlier → Handling outlier
(EDA) (FE)

Finding Categorical variable (Male, Female) → Encoding (FE)
(EDA)

Finding Skewed Range → Scale within a certain Range
(EDA) (FE)

Count of Feature → Handle Imbalanced data
(EDA) (FE)

→ Feature Selection (Drop the least Correlated feature)
(FE)

→ dimension reduction (PCA, tSNE, LDA)

X_1 X_2 → Combine → X'
features feature

Assignment 1

Perform EDA using Some Python Automated tools.

Eg:- Pandas Profiling
- lenime
- mito
- Sweetviz

} at least 3 ~~auto~~ or 5 Automated tool. w.r.t only one Data.

Ass-2 EDA + FE → 10 Dataset → upload on Github
Analysis