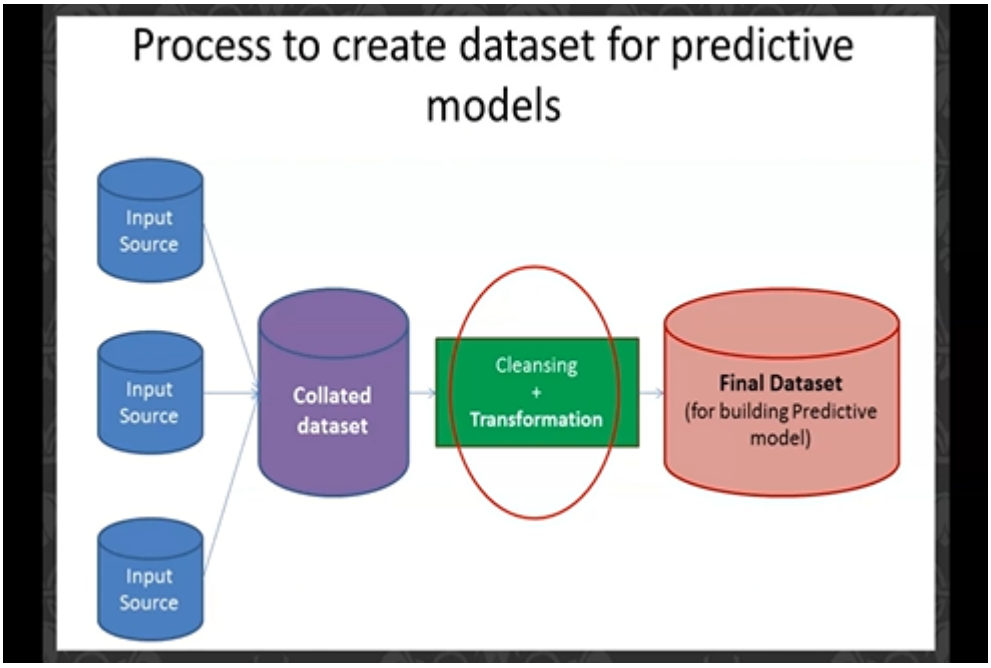
Exploratary Data Analysis (EDA)

Introduction

--> Exploratary data ananlysis or EDA is nothing but a data exploration technique to understand the various aspects of the data.

--> EDA is parent of machine learning.



The above image shows the architecture of the database let's assume the accenture componay has contract to handling the data of the dmart, relaince, spar

. The input source is local data base of the company Example(Dmart, Relaince, spar).

. The collated dataset is the extracted data of the companies put together by accenture.

. The process of extraction is called as "ETL" (Etract Transpose Load"). The process of extraction is automated.

. The collated dataset also called as "MASTER DATABASE" or "DATA WAREHOUSE"

. The data we get from input source will be mostly raw data.

. raw data will be accumilated and data analyst will do cleaning and transfer raw data to clean data by entering phyton code and SQL code to build maching learning model(ML) and Artifical inteliegence model (AI).

. Cleaning and tranfroming of raw data to clean, This technique is called as EDA.

. Once data is tranformed into clean data, then we can biuld "machine learning algoritham" at back end and "artifical intelligence" at front end.

What are the alogorithms we biuld

1. REGRESSION ALGORITHM
2. CLASSIFICATION ALGORITHM
3. CLUSTERING ALGORITHM

Machine Learning (Few topics)

==> Machine learning divided into 3 parts

1. Regression
2. Classification
3. Clustering

Regression alogrithm :

1. If Dependent variable continues then it is called as regression. (Only price realted)

Eample of Continues Variable: (This are front end application " APPLICATION SERVER")

1. Pertrol price
2. Gold price
3. Electricity bill price
4. Hosue price
5. Stock price
6. Crypto price
7. EV manufacture
8. Filght price
9. Train price
10. Hotel booking

re discrete

The follwing are the regerssion algorithms: (This are back end applications "PRODUCTION SERVER")

1. SIMPLE LINEAR REGRESSION
2. MULTIPLE LINEAR REGRESSION
3. POLYNOMIAL REGRESSION
4. SUPPORT VECTORE REGRESSOR
5. K NEAREST NEIGHBOUR REGRESSION
6. DECIOON TREE REGRESSOR

7. Gradient descent , Stochastic Gradient descent,
8. L1 LASS REGRESSOR
9. L2 RIDGE RERESSOR
10. TIME SERIES ANLAYSIS
11. XGBOOST REGRESSION
12. ANN REGRESSION

Classification algorithm:

. If the dependent variable is binary then it is called as classification.(two options)

Eample of Continues Variable:

1. win | loose
2. Postive | negative
3. hike | not hike
4. cat | dog
5. profit | loss
6. job | not get job
7. spam | non spam
8. Truet | False
9. pass | fail
10. purchase | not purchase
11. Yes | No

Clustering algorithm:

If there is no dependent varialbe, but the varialbe are discrete

The Follwoing are classification algorithmns

1. logistic regression
2. decission tree classifier
3. knn classifier
4. rf classifier
5. xgboost classifier
6. lgbm classifier
7. ann classifier
8. naive bayes classifier (bayesian thoerm)

The EDA have seven techniques

1. Varialble Identification
2. Univariate Analysis

3. Bi-Variate Analysis
4. Outlier Analysis
5. Missing Value Treatment
6. Variable Tranformation
7. Variable Creation

The above all techniques will be applied in the most of the projects.

1. Variable Identification

. There a two types of varialbes

--> Dependent variable

--> Independent variable

. The dependent variable is the main variable where other variable dependetns on this variable

. The independent variable is the variable which depents on the main variable.

Eaxmple: 1

Father --> Govt employee (school fees, house rent, EMI, investment)

mother --> House Wife

son --> 5th grade

daugther --> 3rd grade

Let's assume the above metioned are variables, here father is earnning and paying the bills where as remaning members are not earning instead their are dependent on father, in this case father is "DEPENDENT VARIABLE" AND other as "INPENDENT VARIABLE"

. Dependent varialbe: it is also known as target variable and predicted varialbe and denoted as ='y'

. Independent Variable: it is also known as non-target variable and non predicted varialbe and denoted as ='x'

Father --> DEPENDENT VARIABLE = y

mother --> INDEPENDENT VARIABLE = x1

son --> INDEPENDENT VARIABLE(X2) = x2

daugther --> INDEPENDENT VARIABLE(X3) = x3

Math Equation: (y = x1 + x2 + x3) ("This is called multiple linear regression algorithm")

Example: 2

Father --> Govt employee (school fees, house rent, EMI, investment) mother --> House Wife

in this case we have only one independent varialbe and one dependent varialbe

Father --> DEPENDENT VARIABLE = y

mother --> INDEPENDENT VARIABLE = x1

Math Equation: (y = mx + c) ("When there is only one independent variable we use this formula")

Note:

. Independent varialbe could be many variables

. But dependent variable is always only "ONE"

Example

X1	X2	X3	X4	X5	X6	X7	X8	X9	X10	X11	Y
NAME	SFT	SCHOOL	SHOPPING	METRO	HOSPITAL	BHK	VASTU	VIEW	NEW	PRICE	PURCH
ALEX	100		Y		Y	1	Y	N	Y	1CR	Y
JAMES	200	Y		N		2	Y	N	N	2CR	N
MARK	300				N	3	Y	Y		3CR	Y

MULTIPLE LINEAR REGRESSION ALGORITHM

1. The above image shows the example of purchasing a house
2. in the example the price attribute is contiues variable (price increasing). we can say it as regression analysis because there is continues price increse
3. The purchase attribute is classification alogorithn, either to buy, or not (binary)

The variable indentification have other concepts

1. Relevant attribute
2. Irrelevant attribute

Relevant variable and Irrelevant variable

x13	X1	X2	X3	X4	X5	X6	X7	X8	X9	X10	X11	Y	x12	x14
skin color	NAME	SFT	SCHOOL	SHOPPING	METRO	HOSPITAL	BHK	VASTU	VIEW	NEW	PRICE	PURCH	gender	weight
wh	ALEX	100		Y		Y	1	Y	N	Y	1CR	Y	m	
bl	JAMES	200	Y		N		2	Y	N	N	2CR	N	f	
wh	MARK	300				N	3	Y	Y		3CR	Y	m	

1. The above image shows the attributes which are irrelevant like skin colour,weight, height.
2. We only focus on the relevant attributes

NOTE

Whenever we build any data anayst project, we need to plot the graph only with relevant attributes not to use irrelevant attributes oterwise overfiftting problem or "multicollenarity"

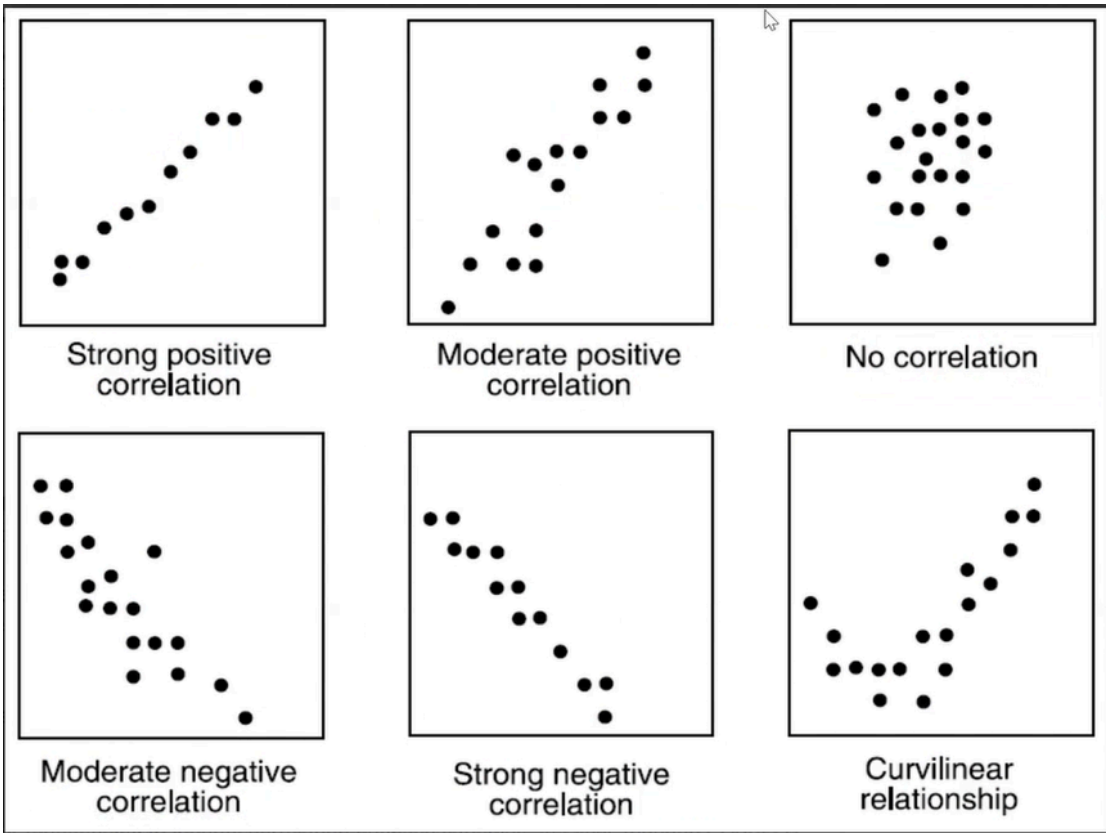
2. Univariate analysis

Plot the graph with one varaible is called Univariate analysis.

3. Bi-Variate analysis

Plot the graph with two varailbes are called Bi-Variate analysis (x-axis & y-axis)

1. In bivariate analysis we have concept known as "corelation concept"
2. corealtion means a relation among two varialbe is a corealtion.



corelation has divided into 3 parts

1. +ve corelation (positive) (range 0 to 1)
2. -ve corelaton (negative) (range -1 to 0)
3. 0 corelation (no corelation) (no range)

corelaton range:

corelation range is from -1 to 1

corealtion function in python is:

.corr()

4 Outlier analysis

1. Outlier is the data point which is very far from other observations
2. Oulier also called as "Anomaly detection"

5. Missing value treatment

Being a data analyst how we treat missing value

MissingNumerical Data

Example 1: missing one value

NUMBER
10
17
76
35

1. The above image shows the numeric data with missing value.
2. To fill the missing value we take mean(average) of the total which is "34.5" (see below image)

NUMBER
10
17
34.5
76
35

The above output shows the missing value.

Example 2 : misiing two values

NUMBER
10
17
76
35
40

1. The above image shows the numeric data with missing two values
2. We take total mean and fill the first missing value
3. and again after filling the first filling value we take total mean and fill the second missing value

NUMBER
10
17
35.6
76
35
35.6
40

The above output shows the missing value

NOTE:

in data set if any numerical value are missing we use the below strategy

1. MEAN STRATEGY
2. MEDIAN STRATEGY
3. MODE STRATEGY

Missing catagerical data

Example 1: missing one catagerical data.

1. The above images shows the missing of one catagerical data
2. To the fill the missing data we use the "mode strategy" (frequency) which is repating again "WINTER"

The above out shows the missing catagerical data

Example 2: missing catagerical data with same data repeting twice.

1. The above images shows the missing of one catagerical data, where other data are repeating twice.
2. in this case we check with the neighbour attribute.

SEASON1	TEMP
summer	56
summer	50
WINTER	4
winter	7
winter	8
rainy	12
rainy	15

1. In the above image we can check the missing value neighbour attribute(temp)
2. it is showing as 4, it means "winter"
3. This is called "K NEAREST NEIGHBOUR REGRESSION" (KNN)

NOTE:

1. K-1 is considered as one neighbour attribute to fill the data.
2. k-2 is considered as two neighbour attributes to fill the data.

6. Variable transformation

Eaxmple:

SEASON1
summer
summer
WINTER
winter
winter
rainy
rainy

1. The above image shows one attribute with different data, from this data we do varaible transformation.

SEASON1	SUMMER	WINTER	RAINY
summer	1	0	0
summer	1	0	0
WINTER	0	1	0
winter	0	1	0
winter	0	1	0
rainy	0	0	1
rainy	0	0	1

1. In the above image we have created 3 other attibutes.
2. Where 1 means the data respresent 0 means not represents.
3. The process of converting catagerical data to numerical data called as" TRANSFORMER" OR "IMPUTAION"

Types of Transformer

1. One hot encoder
2. Dummy Variable
3. Label encoder

One hot encoder

one hot encoder		
SEASON1	SUMMER	WINTER
summer	1	0
summer	1	0
WINTER	0	1
winter	0	1
winter	0	1
rainy	0	0
rainy	0	0

1. Removing the one attritube as conidering the values.
2. Any attribute shows as 1 reamning attributes shows as zero
3. any attribute show two zeros it considered as 1

Dummy Variable

dummy variable			
SEASON1	SUMMER	WINTER	RAINY
summer	1	0	0
summer	1	0	0
WINTER	0	1	0
winter	0	1	0
winter	0	1	0
rainy	0	0	1
rainy	0	0	1

1. The above image shows the example of dummy varialbe
2. While converting catagerical data numerical data the attribute what we create are knows as "Dummy Varialbe"
3. Season 1 attribute we call as "classes"
4. summer, winter,rainy we call as "classifirers"

Label Encounter

Lable encoder	
SEASON1	impu
summer	0
summer	0
WINTER	1
winter	1
winter	1
rainy	2
rainy	2

1. When we impute consider the summer as 0, winter as 1, and rainy as 2 is know as label encoder

7. Variable creation

1. Create mutlple variable from one varaibel is called as "Variable Creation"

SEASON1	SEASON1	SUMMER	WINTER	RAINY
summer	summer	1	0	0
summer	summer	1	0	0
WINTER	WINTER	0	1	0
winter	winter	0	1	0
winter	winter	0	1	0
rainy	rainy	0	0	1
rainy	rainy	0	0	1