Thursday, January 11, 2024 8:04 AM

Lifecycle of Data Science Project (3months - 1year) ← vague

Problem Statement ⇒ Business Understanding ⇒ Data Collection ⇒ Basic data Preparation ⇓ EDA

Business Understanding: Who, What, Where, When, Why, How

Data Collection: Ask Right Questions

EDA ⇓ Data preprocessing ⇐ Modelling ⇐ Evaluations ⇐ Deployment

Data preprocessing: FE, FT

Deployment:
- Azure
- AWS → 750 hours free
- GCS
- Heroku → students

Starting Project

→ Import libraries & load the data

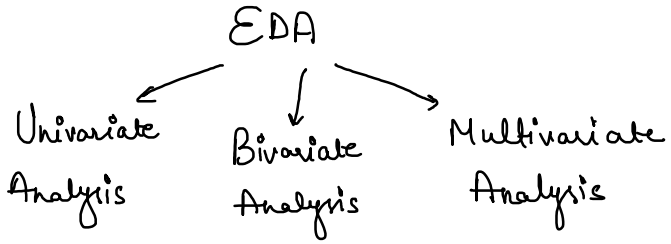
→ Basic Info about data ⇒ df.info() → nulls & non-nulls, metadata, Rows, columns, datatypes

→ Basic description of the data ⇒ df.describe()

→ Basic data study ⇒ unique, nunique & value_counts()

→ Basic Data Preparation ⇒ data quality check / data assessment

⇒ EDA



Univariate ⇒ histogram, countplot, boxplot, kdeplot

Bivariate ⇒ line, lmplot, scatter, bar, pie

Multivariate ⇒ "hue", heatmap, pairplot

UNIVARIATE → BIVARIATE → MULTIVARIATE (Project EDA flow)

Outlier → Boxplot

Z-score

$$Z = \frac{X - \mu}{\sigma}$$

$$-3 > Z \text{ & } Z > 3$$

IQR

$$UL = Q3 + 1.5\,IQR$$

$$LL = Q1 - 1.5\,IQR$$

→ Missing Values → categorical → mode

→ numerical → Skewness ⇒ median

→ no skewness ⇒ mean

Feature Engineering

- Feature Selection
- Feature Extractions
- Standardization ($\mu=0, \sigma=1$)
- Normalization (0 − 1)
- Create New columns
- Drop columns

Encoding ⇒ OHE (Nominal) (One hot encoding) pd.get_dummies , Ordinal Encoding, Label Encoding (I/P variables) → (O/P variable) → ordinal category

Grade	Encoding
A	1
B	2
C	3

One Hot Encoding

	A	B	C	D	E
A	0	0	0	0	0
B	0	1	0	0	0
C	0	0	1	0	0
D	0	0	0	1	0
E	0	0	0	0	1

drop_first = True

Modelling

- Regression → O/p is Continous variable
- Classification → O/p is classes (labels)
- Clustering → No O/p → Create groups

⇒ Evaluations

Regression

Fitness based
→ R^2
→ Adj R^2

Error based
→ MAE, MSE, MAPE, RMSE

Classification

Confusion MATRIX, Accuracy, Precision, Recall, F1score, ROC-AUC

Clustering

Silhouette's score = $\frac{b-a}{\max(b,a)}$

[-1, 1]

INDO-PAK Relations

Tanya

→ India → Hindu Majority
Pak → Muslim Majority

→ Bigger land area of India.

Bharath

→ Health care system is better in India

→ Better Education in India

Monisha

→ Better foreign relations of India

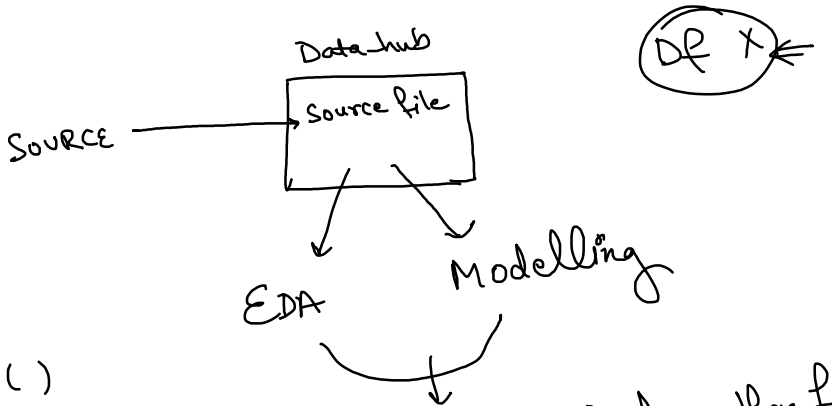
→ GDP is better (India)

PROJECT

AIR TICKET PRICE PREDICTION

Datahub

DP X



def preprocess ()

{ to be filled by you }

return data_eda, data_model

using a single python function