EDA and Feature Engineering

Core ML Pipeline:

- Data Ingestion
- EDA
- Pre-Processing or Feature Engineering
- Model Building
- Evaluation or Validation of Model

Data Ingestion :

Step 1 : Find the dataset from tools such as Bigdata tool (Hadoop, NoSQL, spark streaming, Kafka)

Step 2 : Find the dataset from tools such as Remote location for example SQL DB or No SQL DB

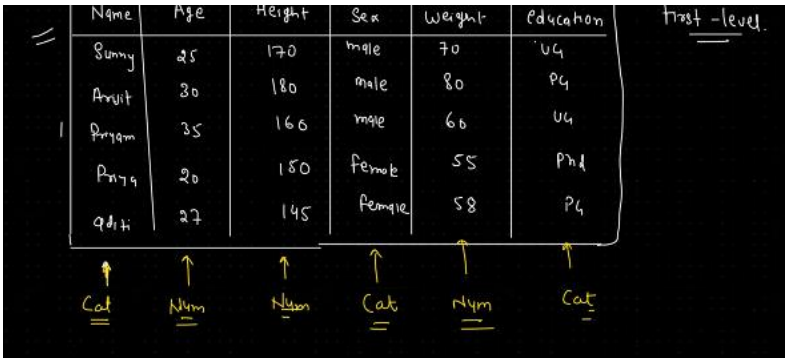
Step 3 : Find the dataset from CSV , TSV, SML , JSON , EXCEL

Step 4 : Fetch the dataset by Scrapping the information or features from website

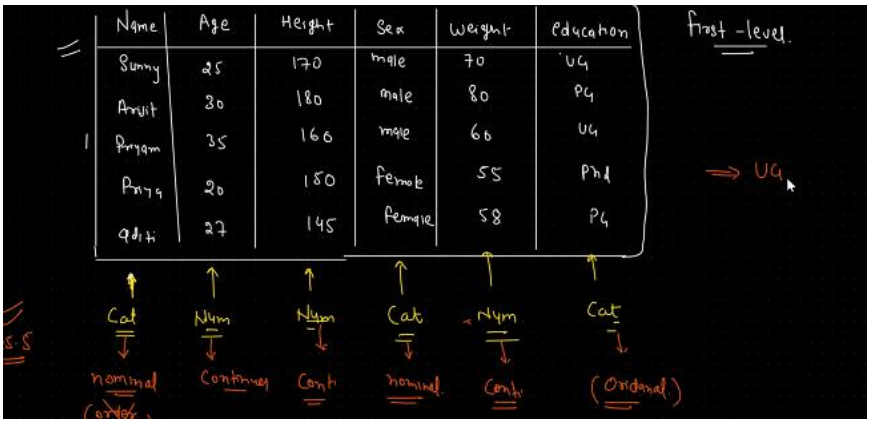
Step 5 : Find the Types of the data whether it is

- Batch data or historical or periodic data,
- Streaming or Continuous or Live data
- Structure Data : Table

Step 6 : Check if the feature is **Categorical or Numerical feature**



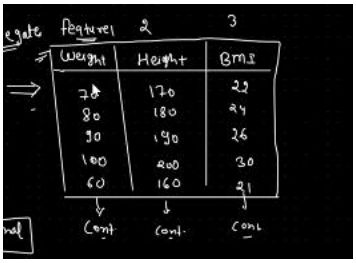
Step 7 : Based on Step 6, we then define whether the **feature is Nominal, Ordinal feature** etc..,



Step 8 : If it is Numeric Data -> then find if it is **Continuous and Discreet feature**

Step 9 : If it is categorical data - > then find if it is **Nominal, Ordinal**

Example of Structured and Continuous data



Example of Discreet Data: Number of students in a class

Example of Categorical : Male or Female, Black or White

Example of Nominal : Order does not matter : Male or female

Example of Ordinal : Order matters -> 10th, 12th, Grad, PG, PHD

Step 10 : Find if it is **Univariate : One/Single column analysis**

Step 11 : Find if it is **Bivariate : More than one column/Two columns analysis**

Step 12 : Find if it is **Multivariate : More than two columns analysis**

Examples of Univariate, Bivariate or Multivariate Analysis :

- If we want to **check only Height = Univariate Analysis**
- If we want to **check height with Age = Bivariate Analysis**
- If we want to **check Height with Age and Name = Multivariate Analysis**

Step 13 : Define or set the "**dependent**" (X) and "**independent" (Y)** features or variable" here **independent variable (Y) is always a "Predictive Variable"** in Machine Learning

Example : If we want to define a weight with Age, Name and Height then the weight is "**Independent Variable**"

Deep Learning data types:

- **Unstructured Data** : Videos, Images, Voice, Sound, text
- **Semi Structured data** : XML, JSON -> This can be converted to structure data and perform EDA

Statistics Uses: To get the insight from the data

- Collect the data
- Organize the data
- Interpretation of the data
- Analysis of the data

- Statistics are used in every domain.
- Every Domain requires EDA and feature Engineering
- EDA : Major role of statistics
- Pre-Processing : Major role of statistics
- Models
- Evaluate and validate

Problem Statement Example : Why the sales are going down?

Sale of a product (Data Analysis) :

Step 1 : Ask as many questions as possible such as below
Why is the sale going down?
What are the reasons sales is down?
Which region the sale is down
Are we paying enough attention to the customers?
Is leadership is good or not
Marketing strategy is not good
We are not looking at the competition

Step 2 : Create a Dataset

Step 3 : Perform the EDA or analysis on why the sale is going down

Step 4 : Conclusion

Scope of Jobs:

- Project Manager
- Business Analytics
- Data Scientist

Core ML Pipeline Stepwise:

- Data Ingestion
- EDA -> Analysis of Data
- Preprocessing(Feature Engineering) : Missing Value/Outlier/Scaling
- Model Building
- Evaluation or Validation of Model

Problem Statement Example of EDA and Preprocessing or Feature-Engineering with the help of "Biryani Preparation" Example:

Step 1: Ensemble the ingredients

Step 2 : EDA -> The quantity of ingredients

Step 3 : Preprocessing data -> Cleaning the ingredients

EDA Stepwise :

Step 1 : How many Rows and columns

Step 2 : Missing values treatment

Step 3 : Categorical or numerical data

Step 4 : Duplicate values treatment

Step 5 : Data types

Step 6 : How much RAM is it consuming by using "pandas profiling"

Step 7 : Statistical based analysis

Step 8 : Box Plot : Outlier and Distribution and statistical profile

Step 9 : Scatter Plot : Outlier

Step 10 : Histogram : Distribution

Step 11 : Heatmap : correlation

Step 12 : Count box : Rows /columns

Step 13 : Based on the above EDA steps, we can perform the preprocessing

Preprocessing or Feature Engineering Stepwise :

Step 1 : Missing value handle

Step 2 : Outlier handle

Step 3 : Scaling of data within certain range

Step 4 : Transformation (Log, Box core, Square, cube)

Step 5 : Encoding to be performed on categorical data into "1" and "0" integers

Step 6 : Imbalance Data can be handled

Step 7 : Important Feature selection can be done

Step 8 : Dimensionality Reduction and combine two or more feature into a single feature (PCA)

EDA Automation Tools :

- Pandas Profiling
- Knime
- Mito
- Please check in google regarding the tool and perform EDA using these automate tools
- These 10 Datasets will be available on resource

FAQ's with Solutions:

Do we need to identify the data types whether it is categorical or numerical based on every feature or whole dataset?
A:Every features

How do we perform statistical data analysis such as T-test, Z-test, Chi square test manually or do we have libraries for all these test available. If libraries are available, what are those libraries
A: Python libraries numpy and pandas' sci pi, Sk learn libraries

Which Statistical analysis is very important and can be used in any datasets
A: Based on the every features, correlation distribution perform statistical analysis

Can you share the PDF for EDA and Preprocessing Interview questions
A: Tomorrow class we will discuss

How do we convert XML and JSON data into a structed datasets?
A: There are Python libraries to convert

How do we know the data is imbalanced and how do we treat imbalanced dataset, Any easy and quick way to handle it?
A:SK learn library , Imb Learn

How do we know which feature needs to be selected and which feature can be dropped. Is there any quick way to choose the features?
A:Statistical method , ML algorithm importance of feature , manually , correlation and sunny will share the code and I can have my own hands on

What do you mean by scaling the data?

A: Data is very scatter and compressed the data or scale a data with different formula in a range

Min max scaler, unit , standard scaler

Which is the best and most famous and wisely used automation tool in 2022 for EDA of any datasets?

A: Discuss on creating own automation tool or pandas profiling also can be used