**Machine Learning**

**Days-09(VIMP)**

**Machine Learning Development Life Cycle(MLDLC/MLDC)**

* Software development life cycle(SDLC)
* Machine learning development life cycle

**Steps:-**

1. **Frame the Problem ( Identify the Problem)**
2. **Gathering Data**

Data Sources :-

1. CSV
2. API
3. Web Scrapping(from websites)
4. Data WareHouse
5. BigData(spark cluster)
6. **Data Preprocessing**

Notes:- data to be preprocess that data easily consume machine learning algorithm called as data preprocessing.

1. Remove duplicates
2. Remove missing values
3. Outlier detection
4. Outlier removes
5. Values Scaling
6. **Exploratory Data Analysis (EDA)**
7. Visualization
8. Univariate analysis/Bivariate analysis/Multivariate analysis (mean, std deviation, curve)
9. Outlier detection
10. Handle imbalanced dataset (convert imbalanced dataset into balanced dataset)
11. **Feature Engineering and Selection**
12. **Model Training, Evaluation and Selection**
13. Different different algorithm train by giving data
14. Evaluate to all models, by using r2 method, mean square error…
15. Model selection (one or multiple algorithms select) and set settings parameters(tuning) for improve performance called as hyperparameter tuning.
16. **Model Deployment**

JSON

API (URL)

Model

Predict

Binary file (pickle tool)

Input through user

Serves Python app

USER

**Days-10**

**Various Data Based Jobs Roles**

1. **Data Engineer 2. Data Analyst 3. Data Scientists 4. Machine Learning Engineer**
2. **Data Engineer**

Responsibility: -

* + - Scrape Data from the given sources.
    - Move/Store the data in optimal servers/warehouses.
    - Build data pipelines/APIs for easy access to the data.
    - Handle databases/data warehouses.

Skills Required: -

* + - Strong grasp of algorithms and data structures
    - Programming Languages (Java/R/Python/Scala) and script writing
    - Advanced DBMS's (SQL/NOSQL)
    - BIG DATA Tools (Apache Spark, Hadoop, Apache Kafka, Apache Hive)
    - Cloud Platforms (Amazon Web Services, Google Cloud Platform)
    - Distributed Systems
    - Data Pipelines

1. **Data Analyst**

Responsibility: -

* + - Cleaning and organizing Raw data.
    - Analyzing data to derive insights.
    - Creating data visualizations.
    - Producing and maintaining reports.
    - Collaborating with teams/colleagues based on the insight gained.
    - Optimizing data collection procedures

Skills Required: -

* + - Statistical Programming
    - Programming Languages (R/SAS/Python)
    - Creative and Analytical Thinking
    - Business Acumen Medium to High preferred
    - Strong Communication Skills
    - Data Mining, Cleaning, and Munging
    - Data Visualization
    - Data Story Telling
    - SQL
    - Advanced Microsoft Excel

1. **Data Scientists**

**“A data scientist is someone who is better statistics than any software engineer and better at software engineering than any statistician.”**

1. **Machine Learning Engineer**

**“A data scientist is someone who is better statistics than any software engineer and better at software engineering than any statistician.”**

Responsibility: -

* + - Deploying machine learning models to production ready environment.
    - Scaling and optimizing the model for production.
    - Monitoring and maintenance of deployed models.

Skills Required: -

* + - Mathematics
    - Programming Languages (R/Python/Java/Scala mainly) Creative and Analytical Thinking
    - Distributed Systems
    - Data model and evaluation
    - Data Mining, Cleaning, and Munging
    - Machine Learning models
    - Software Engineering & Systems design
* **Comparision**

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **Role** | **Analytical Skills** | **Business Acumen** | **Data Storytelling** | **Soft Skills** | **Software Skills** |
| Data Analyst | HIGH | MEDIUM TO HIGH | HIGH | MEDIUM TO HIGH | MEDIUM |
| Data Engineer | MEDIUM | LOW | LOW | MEDIUM | HIGH |
| Data Scientist | HIGH | HIGH | HIGH | HIGH | MEDIUM |
| ML Engineer | MEDIUM TO HIGH | MEDIUM | LOW | HIGH | HIGH |

**Days-11(VVIMP)**

**Tensors: -**

Tensor is nothing but a Data Structure, it is a container to store a number.

1. 0D Tensor/Scaler

Ex:- (2), (3), (5) 🡪 no. of dimension is zero.(ndim = 0)

Note:- tensor and ndimension array are the same

1. 1D Tensor/Vector/1D Array

Ex:- [1,2,3,4,5…] 🡪 list of array, 1D tensor, no of dimension is 1.(ndim = 1)

:- [1,2,3,4], It is 1D tensor but ndim vector. From above example it is 1D tensors but in array is 4 no’s, so it is 4D vector

Vector :- it is collection of scaler

1. 2D Tensor/Metrics

Collection of vectors called as metrics

Ex:- [1,2,3], [4,5,6], [7,8,9]

[ [1,2,3]

[4,5,6]

[7,8,9] ]

1. ND Tensor

4x3x3

1. Rank, Axis and Shape

Rank :- no. of dimensions in tensors

No. axis = no. rank = no of dim

Shape = rows and columns

Size = multiply shape(no. of items)

**Days-12**

**Practical Start from here**

**Installing Anaconda**

**Virtual Environment**

**Days-13**

**Steps:-**

1. **Preprocess + EDA + Feature Selection**
2. **Extracts input & output cols**
3. **Scale the values**
4. **Train test split**
5. **Train the Model**
6. **Evaluate the Model/Model selection**
7. **Deploy the Model**

**Days-14**

How to Frame a Machine Learning Problem/Plan a Data Science Project Effectively

1. Business Problem to ML Problem (Conversion`)
2. Type of Problem
3. Current Solution
4. Getting Data
5. Metrics to Measure
6. Online Vs Batch
7. Check Assumption

**Days-15**

**Gathering Data**

1. CSV
2. JSON/SQL Data

Data Gathering

1. Fetch API
2. Web Scrapping