**Scalar masterclass: 16/Aug/2024**

**Teacher: Avijit Roy**

Roadmap of today’s class:

1. Roadmap to become a Data scientist.
2. Resources and preparation.
3. Different job roles and salaries.
4. End to end project using LLM.

1. Roadmap for a Data scientist:

Two categories: Data analysis and Data science jobs.

Task of DA: Actionable insights from data.

Task of DS: Task of DA + forecasting/prediction (e.g., predict revenue in future, predict number of cars in next 3 months etc.).

Exploratory Data Analysis: (to become a data analyst)

1. Data should be stored (TB, PB data) → Database (SQL for storing and retrieving data).
2. Visualization or Dashboarding → PowerBI and Tableau to present/visualize data (One single image speaks thousand words).
3. Python → programming language.
4. Python libraries → Numpy, Pandas, Matplotlib (Don't write everything from scratch. There are many useful libraries which help just by writing a few lines of codes.)
5. Probability and Statistics.
6. Hypothesis testing. Helps to understand if multiple datasets are similar/correlated or different (statistically significant difference).

EDA takes around 6-8 months (8 hours/week) to learn.

Even for data scientists, 70% time goes in data analysis.

Model building (ML and AI models): (to become a data scientist)

1. Traditional ML algorithms (probability, statistics, calculus etc).

Math to know for ML → Linear algebra, Calculus and Vector algebra (Need not be experts in all of these, just know a little bit of these things).

Regression algorithms, Classification, Clustering, Outlier detection, Time series analysis (e.g., weather, stock price. Periodic patterns), Recommendation system (e.g., this masterclass is shown to me by google advertisement by automation, Google knows which people are interested in which things. Recommendation systems hugely impact business.).

1. Deep learning.

Transformed the entire domain of AI. No limit.

ML is a kind of ABCD of DL and DL is a kind of writing essays out of many sentences. DL gives better results.

Computer vision → seeing the world from a computer's eye. Deal with data related to images/videos (how many humans in a video, face detection, automated challans, Tesla cars etc.). GenAI (Youtube video: Everybody dance now. Transfer dance steps of an expert to any human image. AI is very powerful and can generate human face like in reality).

Natural Language Processing (NLP).

NLP is text data. ChatGPT, Google Gemini are using NLP. Many jobs in NLP.

DS takes around 6-8 months (8 hours/week) to learn.

Consistent study is very important.

2. Resources and preparation:

SQL (w3school).

Python, Data structure algorithms or DSA (Leetcode, Hackaermk, Python official documentation and tutorials).

Mathematics (MIT opencourseware, Stanford courses, NPTEL courses).

Machine Learning (Coursera course by Andrew Ng).

Practice for different datasets → Kaggale.com.

Scalar approach to Data Science course:

Learning from Industry experts from big companies having industry experience.

Divided into two parts: Data Analysis (6-8 months if you put 2 hours per day) and Data Science.

Do industry perspective projects or business cases just learning after each ML algorithm (theory + application).

For DA, Scalar teaches: SQL, Dashboarding (PowerBI or Tableau), Python, Libraries, Probability and Statistics, Hypothesis testing, Product analytics (how to apply all the concepts for business purpose).

For ML, Scalar teaches: Math for ML, Traditional ML, Time series and Recommendation system, Deep Learning (NLP, Computer vision), ML operations (MLOPS) for cloud: deploy your model (take raw data to proper ML model and then deploy model in reality). END of DS journey.

3. Job roles, Job opportunities and salaries:

Job roles and salaries: 1. Data analyst (entry level job and best for starting. 11-16 LPA), 2. Data scientist (16-23 LPA), 3. ML engineer (DS + needs a little bit of software development. 19-25 LPA), 4. Applied scientist (senior role. DS + taking decisions. 50-70 LPA), 5. Research scientist (need PhD. > 1 Cr).

DA and DS can join as a fresher. DA is more for freshers. No age limit.

Search ML jobs in LinedIN.

**Networking** is very important no matter how much you know about a topic. Start building networking with HR, job givers in LinkedIN. Build a good Linkedin profile. Like posts, whatsapp connection with HR.

4. End to end project using LLM:

Have google chrome, stable internet and laptop.

Today’s project is **Financial summarizer**: Create summary of a financial pdf document. We use HDFC bank's last quarter’s document. Type in google HDFC bank financial results. Click on key parameters July 2024. Download this pdf. Rename to HDFC\_Results.pdf. Go to google colab. Search for Google AI studio (to access Google Gemini). Sign in. Click on API key and create. Click on Copy. Then go to google colab for writing codes. Gemini AI summarizes everything very nicely from the pdf doc.