AI & DS

CHP 1 Data science

- it is an interdisciplinary field that uses scientific methods, algorithms, processes & systems to extract knowledge & insights from structured & unstructured data.
- it uses various techniques from mathematics, statistics, computer science to analyze complex data sets.

Evolution

- Early days (1960 - 1980)
  - data analysis primarily focused on statistical methods & mathematical models eg SAS

- internet BOOM (1990 - 2000)
  - it led to a massive increase in data generation.
  - Technique like web scraping & data mining were used.

- Big data era (2010)
  - Due to social media, IOT devices & other source, data volume exploded
  - Tech like Hadoop were used

- ML (2010 - 2020)
  - ML became central to data science, enabling predictive analytics, recommendat" system & automation

- current trends (2020 - Present)
- DS continues to evolve with advancements in DL, NLP & AI.

Steps in data science project

① problems Definition
- understand the Business problem & define the scope of project.

② Data collection
- Gather relevant data from various sources

③ Data cleaning
- pre process the data to handle missing values, or inconsistencies.

④ Exploratory data analysis - analyze & visualize the data to gain insight & understand patterns.

⑤ feature engineering - Create new features or transform existing ones to improve model performance.

⑥ Model Building - select appropriate algorithms & train predictive models using the data.

⑦ Model evaluation - Evaluate model performance using metrics like accuracy, precision, recall etc.

⑧ Model deployment
- Deploy the model into product environment

⑨ Moitoring & maintenance,
- Monitor model performance over time & update as needed.

Application of DS

- Healthcare
  - Predictive analytics for disease diagnosis, personalized medicine, etc.

- finance
  - fraud detection, algorithmic trading, risk assessment

- Retail
  - Recommendation system, demand forecasting, pricing optimization.

- Manufacturing
  - predictive maintenance, supply chain optimization.

- Transportation

Roles in data science

① Data Scientist
- analyzes complex datasets to extract insights & build predective models using statistical Techniques & ML algo.

② Data engineer
- an Design, develops & maintains the infra for data generatn, storage, & processing.

③ ML engineer
- Focuses on building & deploying ML models into productn system.

④ Business analyst
- Translates business requirement into data-driven solution

⑤ Data analyst
- cleans, processes & analyzes data to provide actionable insights & reports for business stakeholders.

Types of data analytics

① Descriptive analytics - what happened?
- describes what was happened in the past based on historical data
- Often in the form of summary.

② Diagnostic analytics - why did this happen?
- focuses on understanding why certain events occured by identifying patterns & correlations in data.

③ Predictive analytics - what will happen?
- It forecast future events or trends based on historical data & statistical techniques, often using ML algo

④ Prescriptive analytics - How can we make it happen?
- It recommends actions to optimize outcome based on predictive models & business rules.

Data science	Business analytics	Big data
- It is a field which includes working with data & using it for building predictive models	Extracting meaningfull info from data	Enormous & complex collection of data
- Consist of statistics, mathematics, ML, Data mining, BDA	Statistical analysis & predictive modeling	volume, variety, velocity, veracity
- Solving complex problems, creating predictive models	Decision-making, understanding patterns	infra for handling Large data.
eg Building recommendation systems for online platforms	Analyzing customer Purchase behavior to optimize marketing strategies.	stock market data, social media content.

1P2

EDA - exploratory data analysis

- The main purpose of EDA is to help look at data before making any assumptions. It can help identify obvious errors, as well as better understand patterns within the data, find's interesting relations among the variables.

- It is used by data scientist to analyze & investigate data sets & summarize their main characteristics, often employing data visualization methods.

- Classifying EDA into 4 types
  - univariate analysis
    Examines one variable at a time to understand it's distribution, central tea tendency & spread.

  - Bivariate analysis
    Analyzes the relationship between 2 variables to understand how they are related to each other

  - Multivariate analysis
    Studies the relationships between multiple variable simultaneously, exploring complex interaction among them.

. Temporal analysis
focuses on analyzing data over time to identify trends, pattern etc

Quantitative data analysis

① Measure of central tendencies

→ they are statistical metrics that provide insight into the typical or central value of dataset. they help to summarize the distribution of data by identifying a representative value around which the observations cluster.

① Mean
- it is arithmetic average of a set of values
- calculated by summing up all the values & ÷ by total number of observation

② Median
- it is the middle value in a dataset when it is ordered
- If there is an even number of observation, the median is the average of the 2 middle values

③ Mode
- It is a value that occurs most frequently in a dataset

Measure of spread

- Range : Represents the difference between the maximum & minimum values in the data
- variance : Measures the dispersion of data points around the Mean.
- Standard deviation : Indicates the average distance of data points from the mean.

Skewness & Kurtosis

- Skewness : Measures the asymmetry of the distribution
- Kurtosis : Measures the peakedness or flatness of the distribution

Histogram

- it is a graphical representation of the frequency distribution of a continous variable.
- it consist of a series of bars, where each bar represents a range (bin) of values and the height of the bar corresponds to the frequency (or count of observations falling within the range).
- Histogram provide a visual depection of the distribution of data, allowing for insights into it's central tendency, dispersion & shape

* - yes, we can perform univariate graphical analysis using a histogram. In fact, histogram are one the most commonly used tools for visualizing the distribution of a single variable

① Displaying distribution
- histogram visualize variable spread accross bins revealing patterns like symmetry / skewness

② Identifying central tendency
- The center of the histogram typically represents the central tendency of variable (mean / median)

③ Assessing Dispersion
- histogram width & spread across bins indicate variability (wider spread means higher variability & narrower spread means Lower variability)

④ Detecting outliers: outliers apper as significantly taller / shorter bars, detectable in the tails of the histogram.

Key Techniques of EDA

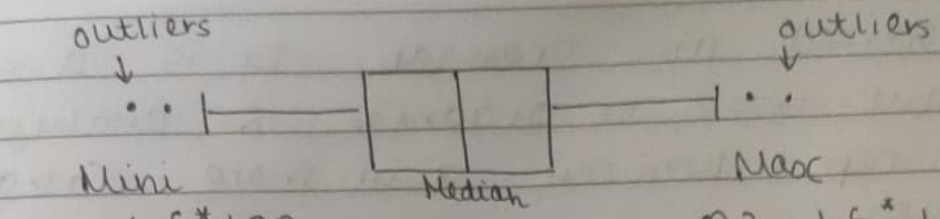
① Descriptive Statisco:
Calculate measure such as mean, median, mode, variance & standard deviation to summarize the central tendency & spread of data

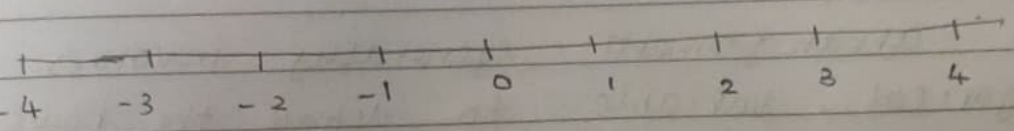
② Data visualization
Create graphical representation like histogram, box plot, scatter plots & heatmaps to visually explore pattern & trends.

③ correlation analysis
Examine the relationships between variable to identify potential dependencies

④ outlier detection
Identify & handle outliers that may significantly impact the analysis

Parts of box plots





Minimum = The min value in the given dataset

first quartile (Q1) : It is the median of the Lower half of the dataset

Median = middle value of the dataset, which ÷ the dataset into 2 equal part. (second quartile)

Third quartile: It is the median of the upper half of the data

Max : Max value in dataset

Interquartile Range (IQR) : Difference between the 1st quartile & 3rd quartile i.e (IQR = Q3 - Q1)

Cross tabulation

- Also referred as crosstab. It is a statistical technique used to organize & analyze the relationship between 2 or more categorical variables.

- The arrangement generally involves one categorical variable to define the rows of the table & another categorical variable to define the columns where the intersect of the rows & columns contain the freq of observatn corresponding to the combinat of variable

- used to Identify patter, trends & dependencies

Machine learning

- It is a subset of Ai that focuses on the development of algo & models that allow computers to learn & improve from experience without being explicitly programmed.

- ML enables computers to recognize patterns, make predictions, & derive insights from data without human intervention.

Features of ML

- Learning from data : ML learn from historical data to identify patterns & relationships

- Adaptability : ML model can adapt & evolve over time as new data becomes available

- Automation : ML automates the process of building a predictive models & making decisions based on data.

- Iterative improvement : ML is an iterative process where models are continuously trained, evaluate & refined based on feedback

- prediction

Types of machine learning:

- supervised learning (classificatn)
  - This involves training a model on a labeled data set, where the correct output is provided for each input
  - The algorithm uses this information to learn the relationship between inputs & outputs & can then make predictions on new, unseen data.

- unsupervised learning
  - This involves training a model on an unlabeled data set where the correct output is not provided
  - The algorithm must find the structure in the data on it's own. (used for clustering).

- semi-supervised learning
  - Combines element of both supervised & unsupervised learning
  - Here a small amount of labelled data along with unlabelled data is provided to improve model performance
  - This approach is useful when obtaining labelled data is expensive or time-consuming.

- Reinforcement learning
  - This involves training an agent to make decisions in an environment where it receives feedback through rewards or punishments.
  - used in decision making task.

Application of machine learning

① image recognition - one of the most common applicatn of ML
- used to identify images

② speech recognition - while using google, we get an option of "search by voice"
- process of converting voice (instructn) into text

③ Traffic prediction - google maps

④ Product recommendation - amazon & netflix

⑤ Online fraud detection

⑥ Stock Market trading.