SCSB4013 INTRODUCTION TO DATA SCIENCE

UNIT 1

Introduction - Need for data science – benefits and uses – facets of data – data science process – setting the research goal – retrieving data – cleansing, integrating, and transforming data – exploratory data analysis – build the models – presenting and building applications.

Introduction To Data Science

Data science is a multidisciplinary field that combines techniques from statistics, mathematics, computer science, and domain knowledge to extract meaningful insights and knowledge from data. It encompasses various stages of data processing, including data collection, cleaning, analysis, visualization, and interpretation.

Key Components of Data Science:

1. **Data Collection:** Gathering relevant data from various sources, which can include databases, APIs, sensors, or other data repositories.
2. **Data Cleaning and Preprocessing:** Ensuring data quality by handling missing values, removing outliers, and transforming data into a usable format.
3. **Exploratory Data Analysis (EDA):** Analyzing and visualizing data to understand its characteristics, uncover patterns, and identify relationships between variables.
4. **Model Building and Machine Learning:** Developing statistical models or machine learning algorithms to make predictions or decisions based on data.
5. **Data Visualization and Communication:** Presenting findings and insights to stakeholders using charts, graphs, and reports that are easy to understand.
6. **Deployment and Maintenance:** Implementing models into production systems and continuously monitoring their performance and updating them as necessary.

Tools and Technologies:

- **Programming Languages:** Python and R are widely used for data science due to their rich libraries and tools for statistical analysis and machine learning.

- **Libraries and Frameworks:** Examples include pandas, NumPy, scikit-learn (Python), and tidyverse R) for data manipulation and analysis, and TensorFlow or PyTorch for deep learning.
- **Big Data Technologies:** Hadoop, Spark, and Kafka are used for processing and analyzing large-scale datasets.
- **Data Visualization Tools:** Such as Matplotlib, Seaborn, ggplot2, and Tableau for creating visual representations of data.

Applications of Data Science:

- **Business Intelligence:** Using data to make informed business decisions and optimize processes.
- **Healthcare:** Analyzing patient records to improve treatment outcomes and predict disease trends.
- **Finance:** Fraud detection, risk assessment, and algorithmic trading based on market data.
- **Internet of Things (IoT):** Analyzing sensor data to optimize performance and predict maintenance needs.

Needs of Data Science:

The field of data science is driven by several key needs and challenges, which reflect its multidisciplinary nature and the demand for specialized skills and technologies. Here are some of the primary needs of data science:

1. **Data Collection and Integration:**
    - **Challenge:** Gathering relevant data from diverse sources, including structured and unstructured data.
    - **Need:** Efficient methods and tools for data extraction, transformation, and integration to ensure data quality and usability.
2. **Data Cleaning and Preprocessing:**
    - **Challenge:** Handling missing values, outliers, and inconsistencies in data.
    - **Need:** Techniques and algorithms for data cleaning, normalization, and feature engineering to prepare data for analysis.
3. **Exploratory Data Analysis (EDA):**
    - **Challenge:** Understanding data distributions, relationships, and patterns.
    - **Need:** Visualization tools, statistical methods, and exploratory techniques to uncover insights and formulate hypotheses.
4. **Machine Learning and Predictive Modeling:**
    - **Challenge:** Developing accurate models that generalize well to new data.
    - **Need:** Algorithms and frameworks for supervised and unsupervised learning, deep learning, and reinforcement learning. Also, techniques for model evaluation, selection, and tuning.
5. **Data Visualization and Communication:**
    - **Challenge:** Effectively communicating insights and findings to stakeholders.
    - **Need:** Tools and techniques for creating clear and informative visualizations, dashboards, and reports that facilitate decision-making.
6. **Scalability and Big Data Management:**
    - **Challenge:** Handling large volumes of data efficiently.
    - **Need:** Big data technologies such as Hadoop, Spark, and distributed computing frameworks for processing and analyzing massive datasets.
7. **Ethical Considerations and Privacy:**
    - **Challenge:** Addressing ethical concerns related to data privacy, bias, and fairness.
    - **Need:** Ethical frameworks, guidelines, and regulations to ensure responsible data use and mitigate potential harms.

8. **Interdisciplinary Knowledge and Collaboration:**
    - **Challenge:** Integrating expertise from statistics, computer science, domain knowledge, and business understanding.
    - **Need:** Collaboration between data scientists, domain experts, and stakeholders to leverage diverse perspectives and domain-specific insights.
9. **Continuous Learning and Adaptation:**
    - **Challenge:** Keeping up with rapid advancements in technology and methodologies.
    - **Need:** Lifelong learning, professional development, and staying updated with emerging tools, techniques, and best practices in data science.
10. **Real-World Application and Impact:**
    - **Challenge:** Translating data-driven insights into actionable strategies and solutions.
    - **Need:** Practical application of data science to solve real-world problems, drive innovation, and create value for businesses, organizations, and society.

Addressing these needs requires a combination of technical skills, domain knowledge, and a strong foundation in data science principles. As the field evolves, data scientists play a crucial role in harnessing the power of data to make informed decisions and drive positive outcomes across various industries and sectors.

Benefits and Uses of Data Science:

Data science offers a wide range of benefits and applications across various industries and domains, driven by its ability to extract valuable insights from data. Here are some of the key benefits and uses of data science:

Benefits of Data Science:

1. **Informed Decision-Making:**
    - Data science enables organizations to make data-driven decisions by analyzing historical trends, patterns, and relationships within data. This reduces reliance on intuition and gut feeling, leading to more accurate and informed decision-making processes.
2. **Improved Efficiency and Productivity:**
    - By automating repetitive tasks and optimizing processes, data science can significantly improve operational efficiency and productivity. For example, predictive maintenance in manufacturing uses data to anticipate equipment failures, reducing downtime and maintenance costs.
3. **Enhanced Customer Insights and Personalization:**
    - Data science helps businesses understand customer behavior and preferences through techniques like segmentation and recommendation systems. This enables personalized marketing strategies, product recommendations, and customer service enhancements.

4. **Cost Savings and Risk Management:**
    - Through predictive analytics and risk modeling, data science can identify potential risks and opportunities, helping businesses optimize resource allocation and mitigate financial losses. This is particularly valuable in finance, insurance, and supply chain management.
5. **Innovation and Competitive Advantage:**
    - Data science fosters innovation by uncovering new insights, identifying emerging trends, and supporting product development. Companies that effectively leverage data science gain a competitive edge by adapting quickly to market changes and customer demands.
6. **Real-Time Insights and Decision Support:**
    - With advancements in data processing and analytics, data science enables real-time monitoring and decision support systems. This is critical in dynamic environments such as healthcare (patient monitoring) and e-commerce (dynamic pricing).

Uses of Data Science:

1. **Healthcare and Life Sciences:**
    - Predictive analytics for patient outcomes, drug discovery and development, personalized medicine, and healthcare fraud detection.
2. **Finance and Banking:**
    - Fraud detection, risk assessment, algorithmic trading, credit scoring, and customer segmentation for targeted marketing.
3. **Retail and E-commerce:**
    - Market basket analysis, customer churn prediction, personalized recommendations, pricing optimization, and supply chain management.
4. **Telecommunications:**
    - Network optimization, customer churn analysis, predictive maintenance for infrastructure, and service quality improvement.
5. **Manufacturing and Industry 4.0:**
    - Predictive maintenance, quality control, supply chain optimization, and process optimization using IoT and sensor data.
6. **Government and Public Sector:**
    - Policy analysis, predictive policing, smart city initiatives, and fraud detection in public services.
7. **Education:**
    - Personalized learning paths, student performance prediction, and adaptive learning platforms.
8. **Media and Entertainment:**
    - Content recommendation systems, audience segmentation, sentiment analysis, and targeted advertising.
9. **Energy and Utilities:**
    - Demand forecasting, grid optimization, predictive maintenance for infrastructure, and energy consumption analysis.

10. **Transportation and Logistics:**
    - Route optimization, fleet management, predictive maintenance for vehicles, and customer demand forecasting.

Key Components of Data Science:

Very large amount of data will generate in big data and data science. These data is various types and main categories of data are as follows:

a) Structured

b) Natural language

c) Graph-based

d) Streaming

e) Unstructured

f) Machine-generated

g) Audio, video and images

Structured Data

• Structured data is arranged in rows and column format. It helps for application to retrieve and process data easily. Database management system is used for storing structured data.

• The term structured data refers to data that is identifiable because it is organized in a structure. The most common form of structured data or records is a database where specific information is stored based on a methodology of columns and rows.

Structured data is also searchable by data type within content. Structured data is understood by computers and is also efficiently organized for human readers.

• An Excel table is an example of structured data.

Unstructured Data

• Unstructured data is data that does not follow a specified format. Row and columns are not used for unstructured data. Therefore it is difficult to retrieve required information. Unstructured data has no identifiable structure.

• The unstructured data can be in the form of Text: (Documents, email messages, customer feedbacks), audio, video, images. Email is an example of unstructured data.

Even today in most of the organizations more than 80 % of the data are in unstructured form. This carries lots of information. But extracting information from these various sources is a very big challenge.

• Characteristics of unstructured data:

1. There is no structural restriction or binding for the data.

2. Data can be of any type.

3. Unstructured data does not follow any structural rules.

4. There are no predefined formats, restriction or sequence for unstructured data.

5. Since there is no structural binding for unstructured data, it is unpredictable in nature.

Natural Language

• Natural language is a special type of unstructured data.

• Natural language processing enables machines to recognize characters, words and sentences, then apply meaning and understanding to that information. This helps machines to understand language as humans do.

Natural language processing is the driving force behind machine intelligence in many modern real-world applications. The natural language processing community has had success in entity recognition, topic recognition, summarization, text completion and sentiment analysis.

•For natural language processing to help machines understand human language, it must go through speech recognition, natural language understanding and machine translation. It is an iterative process comprised of several layers of text analysis.

Machine - Generated Data

• Machine-generated data is an information that is created without human interaction as a result of a computer process or application activity. This means that data entered manually by an end-user is not recognized to be machine-generated.

• Machine data contains a definitive record of all activity and behavior of our customers, users, transactions, applications, servers, networks, factory machinery and so on.

• It's configuration data, data from APIs and message queues, change events, the output of diagnostic commands and call detail records, sensor data from remote equipment and more.

• Examples of machine data are web server logs, call detail records, network event logs and telemetry.

• Both Machine-to-Machine (M2M) and Human-to-Machine (H2M) interactions generate machine data. Machine data is generated continuously by every processor-based system, as well as many consumer-oriented systems.

It can be either structured or unstructured. In recent years, the increase of machine data has surged. The expansion of mobile devices, virtual servers and desktops, as well as cloud- based services and RFID technologies, is making IT infrastructures more complex.

Graph-based or Network Data

•Graphs are data structures to describe relationships and interactions between entities in complex systems. In general, a graph contains a collection of entities called nodes and another collection of interactions between a pair of nodes called edges.

• Nodes represent entities, which can be of any object type that is relevant to our problem domain. By connecting nodes with edges, we will end up with a graph (network) of nodes.

• A graph database stores nodes and relationships instead of tables or documents. Data is stored just like we might sketch ideas on a whiteboard. Our data is stored without restricting it to a predefined model, allowing a very flexible way of thinking about and using it.

• Graph databases are used to store graph-based data and are queried with specialized query languages such as SPARQL.

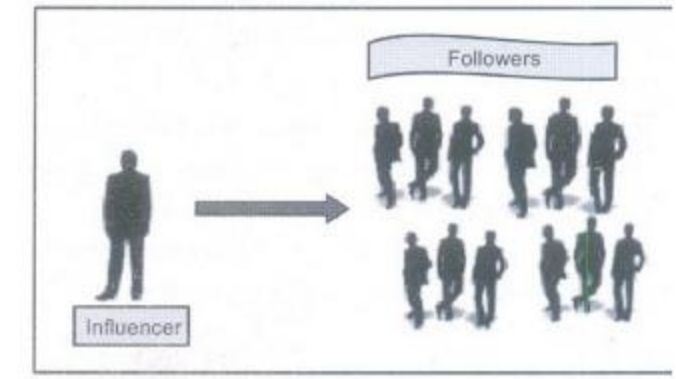
• Graph databases are capable of sophisticated **fraud prevention.** With graph databases, we can use relationships to process financial and purchase transactions in near-real time. With fast graph queries, we are able to detect that, for example, a potential purchaser is using the same email address and credit card as included in a known fraud case.

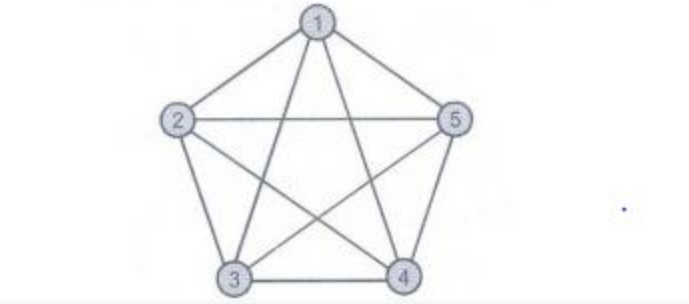
Graph databases can also help user easily detect relationship patterns such as multiple people associated with a personal email address or multiple people sharing the same IP address but residing in different physical addresses.

• Graph databases are a good choice for recommendation applications. With graph databases, we can store in a graph relationships between information categories such as customer interests, friends and purchase history. We can use a highly available graph database to make product recommendations to a user based on which products are purchased by others who follow the same sport and have similar purchase history.

• Graph theory is probably the main method in social network analysis in the early history of the social network concept. The approach is applied to social network analysis in order to determine important features of the network such as the nodes and links (for example influencers and the followers).

• Influencers on social network have been identified as users that have impact on the activities or opinion of other users by way of followership or influence on decision made by other users on the network as shown in below figure





Graph theory has proved to be very effective on large-scale datasets such as social network data. This is because it is capable of by-passing the building of an actual visual representation of the data to run directly on data matrices.

Audio, Image and Video

• Audio, image and video are data types that pose specific challenges to a data scientist. Tasks that are trivial for humans, such as recognizing objects in pictures, turn out to be challenging for computers.

•The terms audio and video commonly refers to the time-based media storage format for sound/music and moving pictures information. Audio and video digital recording, also referred as audio and video codecs, can be uncompressed, lossless compressed or lossy compressed depending on the desired quality and use cases.

• It is important to remark that multimedia data is one of the most important sources of information and knowledge; the integration, transformation and indexing of multimedia data bring significant challenges in data management and analysis. Many challenges have to be addressed including big data, multidisciplinary nature of Data Science and heterogeneity.

• Data Science is playing an important role to address these challenges in multimedia data. Multimedia data usually contains various forms of media, such as text, image, video, geographic coordinates and even pulse waveforms, which come from multiple sources. Data Science can be a key instrument covering big data, machine learning and data mining solutions to store, handle and analyze such heterogeneous data.

Streaming Data

- Streaming data is data that is generated continuously by thousands of data sources, which typically send in the data records simultaneously and in small sizes (order of Kilobytes).

• Streaming data includes a wide variety of data such as log files generated by customers using your mobile or web applications, ecommerce purchases, in-game player activity, information from social networks, financial trading floors or geospatial services and telemetry from connected devices or instrumentation in data centers.

Difference between Structured and Unstructured Data:

S.No	Parameters	Structured Data	Unstructured Data
1	Representation	It is in Discrete form i.e. stored in row and column form	It doesn't follow any specific format
2	Meta Data	Syntax	Semantics
3.	Storage	DBMS	Unmanaged File Structure
4.	Standard	SQL,ADO .net,ODBC	Open XML,SMTO,
5.	Integration Tool	ETL	Batch Processing or manual data entry
6.	Characteristics	With a structure document, certain information always appears in the same location on the page.	In an Unstructured document information can appear in unexpected places on the document
7.	Used by Organization	Low volume operations	High Volume operations

Data Science Process:

It Consists of Six Stages are:

1. Discovery or Setting the research goal

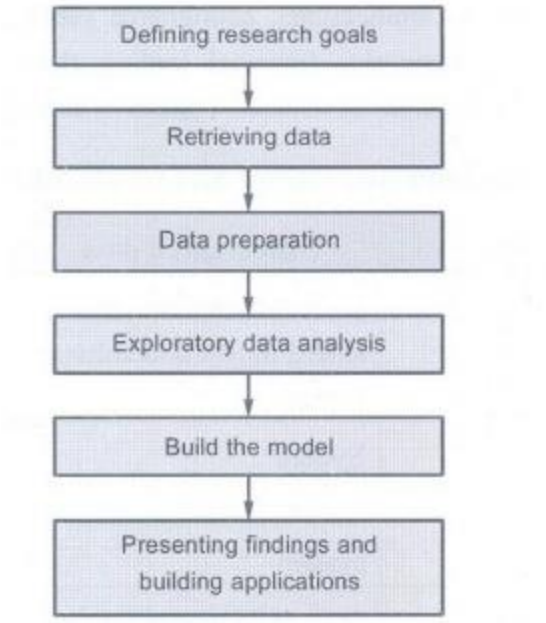
2. Retrieving data

3. Data preparation

4. Data exploration

5. Data modeling

6. Presentation and automation



Data Science Design Process

Step 1: Discovery or Defining research goal

This step involves acquiring data from all the identified internal and external sources, which helps to answer the business question.

• Step 2: Retrieving data

It collection of data which required for project. This is the process of gaining a business understanding of the data user have and deciphering what each piece of data means. This could entail determining exactly what data is required and the best methods for obtaining it. This also entails determining what each of the data points means in terms of the company. If we have given a data set from a client, for example, we shall need to know what each column and row represents.

• Step 3: Data preparation

Data can have many inconsistencies like missing values, blank columns, an incorrect data format, which needs to be cleaned. We need to process, explore and condition data before modeling. The cleandata, gives the better predictions.

• Step 4: Data exploration

Data exploration is related to deeper understanding of data. Try to understand how variables interact with each other, the distribution of the data and whether there are outliers. To achieve this use descriptive statistics, visual techniques and simple modeling. This steps is also called as Exploratory Data Analysis.

• Step 5: Data modeling

In this step, the actual model building process starts. Here, Data scientist distributes datasets for training and testing. Techniques like association, classification and clustering are applied to the training data set. The model, once prepared, is tested against the "testing" dataset.

• Step 6: Presentation and automation

Deliver the final baselined model with reports, code and technical documents in this stage. Model is deployed into a real-time production environment after thorough testing. In this stage, the key findings are communicated to all stakeholders. This helps to decide if the project results are a success or a failure based on the inputs from the model.

Setting Research Goals:

• To understand the project, three concept must understand: what, why and how.

a) What is expectation of company or organization?

b) Why does a company's higher authority define such research value?

c) How is it part of a bigger strategic picture?

• Goal of first phase will be the answer of these three questions.

• In this phase, the data science team must learn and investigate the problem, develop context and understanding and learn about the data sources needed and available for the project.

1. Learning the business domain :

• Understanding the domain area of the problem is essential. In many cases, data scientists will have deep computational and quantitative knowledge that can be broadly applied across many disciplines.

• Data scientists have deep knowledge of the methods, techniques and ways for applying heuristics to a variety of business and conceptual problems.

2. Resources :

• As part of the discovery phase, the team needs to assess the resources available to support the project. In this context, resources include technology, tools, systems, data and people.

3. Frame the problem :

• Framing is the process of stating the analytics problem to be solved. At this point, it is a best practice to write down the problem statement and share it with the key stakeholders.

• Each team member may hear slightly different things related to the needs and the problem and have somewhat different ideas of possible solutions.

4. Identifying key stakeholders:

• The team can identify the success criteria, key risks and stakeholders, which should include anyone who will benefit from the project or will be significantly impacted by the project.

• When interviewing stakeholders, learn about the domain area and any relevant history from similar analytics projects.

5. Interviewing the analytics sponsor:

• The team should plan to collaborate with the stakeholders to clarify and frame the analytics problem.

• At the outset, project sponsors may have a predetermined solution that may not necessarily realize the desired outcome.

In these cases, the team must use its knowledge and expertise to identify the true underlying problem and appropriate solution.

• When interviewing the main stakeholders, the team needs to take time to thoroughly interview the project sponsor, who tends to be the one funding the project or providing the high-level requirements.

• This person understands the problem and usually has an idea of a potential working solution.

6. Developing initial hypotheses:

• This step involves forming ideas that the team can test with data. Generally, it is best to come up with a few primary hypotheses to test and then be creative about developing several more.

• These Initial Hypotheses form the basis of the analytical tests the team will use in later phases and serve as the foundation for the findings in phase.

7. Identifying potential data sources:

• Consider the volume, type and time span of the data needed to test the hypotheses. Ensure that the team can access more than simply aggregated data. In most cases, the team will need the raw data to avoid introducing bias for the downstream analysis.

Retrieving Data

• Retrieving required data is second phase of data science project. Sometimes Data scientists need to go into the field and design a data collection process. Many companies will have already collected and stored the data and what they don't have can often be bought from third parties.

• Most of the high quality data is freely available for public and commercial use. Data can be stored in various format. It is in text file format and tables in database. Data may be internal or external.

1. Start working on internal data, i.e. data stored within the company

• First step of data scientists is to verify the internal data. Assess the relevance and quality of the data that's readily in company. Most companies have a program for maintaining key data, so much of the cleaning work may already be done. This data can be stored in official data repositories such as databases, data marts, data warehouses and data lakes maintained by a team of IT professionals.

• Data repository is also known as a data library or data archive. This is a general term to refer to a data set isolated to be mined for data reporting and analysis. The data repository is a large database infrastructure, several databases that collect, manage and store data sets for data analysis, sharing and reporting.

• Data repository can be used to describe several ways to collect and store data:

) Data warehouse is a large data repository that aggregates data usually from multiple sources or segments of a business, without the data being necessarily related.

b) Data lake is a large data repository that stores unstructured data that is classified and tagged with metadata.

c) Data marts are subsets of the data repository. These data marts are more targeted to what the data user needs and easier to use.

d) Metadata repositories store data about data and databases. The metadata explains where the data source, how it was captured and what it represents.

e) Data cubes are lists of data with three or more dimensions stored as a table.

Advantages of data repositories:

i. Data is preserved and archived.

ii. Data isolation allows for easier and faster data reporting.

iii. Database administrators have easier time tracking problems.

iv. There is value to storing and analyzing data.

Disadvantages of data repositories :

i. Growing data sets could slow down systems.

ii. A system crash could affect all the data.

iii. Unauthorized users can access all sensitive data more easily than if it was distributed across several locations.

2. Do not be afraid to shop around

• If required data is not available within the company, take the help of other company, which provides such types of database. For example, Nielsen and GFK are provides data for retail industry. Data scientists also take help of Twitter, LinkedIn and Facebook.

• Government's organizations share their data for free with the world. This data can be of excellent quality; it depends on the institution that creates and manages it. The information they share covers a broad range of topics such as the number of accidents or amount of drug abuse in a certain region and its demographics.

3. Perform data quality checks to avoid later problem

Allocate or spend some time for data correction and data cleaning. Collecting suitable, error free data is success of the data science project.

• Most of the errors encounter during the data gathering phase are easy to spot, but being too careless will make data scientists spend many hours solving data issues that could have been prevented during data import.

Data Preparation

• Data preparation means data cleansing, Integrating and transforming data.

Data Cleaning

• Data is cleansed through processes such as filling in missing values, smoothing the noisy data or resolving the inconsistencies in the data.

• Data cleaning tasks are as follows:

1. Data acquisition and metadata

2. Fill in missing values

3. Unified date format

4. Converting nominal to numeric

5. Identify outliers and smooth out noisy data

6. Correct inconsistent data

• Data cleaning is a first step in data pre-processing techniques which is used to find the missing value, smooth noise data, recognize outliers and correct inconsistent.

• **Missing value**: These dirty data will affects on miming procedure and led to unreliable and poor output. Therefore it is important for some data cleaning routines. For example, suppose that the average salary of staff is Rs. 65000/-. Use this value to replace the missing value for salary.

• **Data entry errors:** Data collection and data entry are error-prone processes. They often require human intervention and because humans are only human, they make typos or lose their concentration for a second and introduce an error into the chain. But data collected by machines or computers isn't free from errors either. Errors can arise from human sloppiness, whereas others are due to machine or hardware failure. Examples of errors originating from machines are transmission errors or bugs in the extract, transform and load phase (ETL).

• **Whitespace error:** Whitespaces tend to be hard to detect but cause errors like other redundant characters would. To remove the spaces present at start and end of the string, we can use strip() function on the string in Python.

• **Fixing capital letter mismatches:** Capital letter mismatches are common problem. Most programming languages make a distinction between "Chennai" and "chennai".

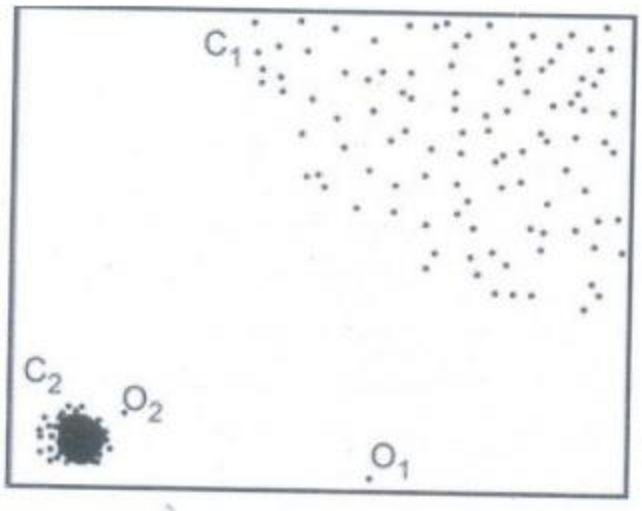
• Python provides string conversion like to convert a string to lowercase, uppercase using lower(), upper().

• The lower() Function in python converts the input string to lowercase. The upper() Function in python converts the input string to uppercase.

Outlier

• Outlier detection is the process of detecting and subsequently excluding outliers from a given set of data. The easiest way to find outliers is to use a plot or a table with the minimum and maximum values.

• Below Figure shows outliers detection. Here O_1 and O_2 seem outliers from the rest.



An outlier may be defined as a piece of data or observation that deviates drastically from the given norm or average of the data set. An outlier may be caused simply by chance, but it may also indicate measurement error or that the given data set has a heavy-tailed distribution.

• Outlier analysis and detection has various applications in numerous fields such as fraud detection, credit card, discovering computer intrusion and criminal behaviours, medical and public health outlier detection, industrial damage detection.

• General idea of application is to find out data which deviates from normal behaviour of data set.

Dealing with Missing Value

• These dirty data will affects on miming procedure and led to unreliable and poor output. Therefore it is important for some data cleaning routines.

How to handle noisy data in data mining?

• Following methods are used for handling noisy data:

1. Ignore the tuple: Usually done when the class label is missing. This method is not good unless the tuple contains several attributes with missing values.

2. Fill in the missing value manually : It is time-consuming and not suitable for a large data set with many missing values.

3. Use a global constant to fill in the missing value: Replace all missing attribute values by the same constant.

4. Use the attribute mean to fill in the missing value: For example, suppose that the average salary of staff is Rs 65000/-. Use this value to replace the missing value for salary.

5. Use the attribute mean for all samples belonging to the same class as the given tuple.

6. Use the most probable value to fill in the missing value.

Correct Errors as Early as Possible

• If error is not corrected in early stage of project, then it create problem in latter stages. Most of the time, we spend on finding and correcting error. Retrieving data is a difficult task and organizations spend millions of dollars on it in the hope of making better decisions. The data collection process is errorprone and in a big organization it involves many steps and teams.

• Data should be cleansed when acquired for many reasons:

a) Not everyone spots the data anomalies. Decision-makers may make costly mistakes on information based on incorrect data from applications that fail to correct for the faulty data.

b) If errors are not corrected early on in the process, the cleansing will have to be done for every project that uses that data.

c) Data errors may point to a business process that isn't working as designed.

d) Data errors may point to defective equipment, such as broken transmission lines and defective sensors.

e) Data errors can point to bugs in software or in the integration of software that may be critical to the company

Combining Data from Different Data Sources

1. Joining table

• Joining tables allows user to combine the information of one observation found in one table with the information that we find in another table. The focus is on enriching a single observation.

• A primary key is a value that cannot be duplicated within a table. This means that one value can only be seen once within the primary key column. That same key can exist as a foreign key in another table which creates the relationship. A foreign key can have duplicate instances within a table.

• Below Figure shows Joining two tables on the CountryID and CountryName keys.

-

Date	CountryID	Units
10/10/2021	1001	100
21/10/2021	3001	50
31/10/2021	4001	75
01/10/2021	3001	90

CountryID	CountryName
1001	India
3001	UK
4001	USA
3001	Spain

Date	CountryID	Units	CountryName
10/10/2021	1001	100	India
21/10/2021	3001	50	USA
31/10/2021	4001	75	Spain
01/10/2021	3001	90	USA

Appending tables

• Appending table is called stacking table. It effectively adding observations from one table to another table. Below Figure shows Appending table.

Table 1

x1	x2	x3
1	a	3
2	b	3
3	c	3
4	d	3
5	e	3

Table 2

x1	x2	x3
11	k	33
12	l	33
13	m	33
14	n	33
15	o	33

Table 3

x1	x2	x3
1	a	3
2	b	3
3	c	3
4	d	3
5	e	3
11	k	33
12	l	33
13	m	33
14	n	33
15	o	33

Fig. 1.6.3 : Appending table

• Table 1 contains x3 value as 3 and Table 2 contains x3 value as 33.The result of appending these tables is a larger one with the observations from Table 1 as well as Table 2. The equivalent operation in set theory would be the union and this is also the command in SQL, the common language of relational databases. Other set operators are also used in data science, such as set difference and intersection.

3. Using views to simulate data joins and appends

• Duplication of data is avoided by using view and append. The append table requires more space for storage. If table size is in terabytes of data, then it becomes problematic to duplicate the data. For this reason, the concept of a view was invented.

• Below Figure shows how the sales data from the different months is combined virtually into a yearly sales table instead of duplicating the data.

Fig. 1.6.4 : View

Transforming Data

• In data transformation, the data are transformed or consolidated into forms appropriate for mining. Relationships between an input variable and an output variable aren't always linear.

• Reducing the number of variables: Having too many variables in the model makes the model difficult to handle and certain techniques don't perform well when user overload them with too many input variables.

• All the techniques based on a Euclidean distance perform well only up to 10 variables. Data scientists use special methods to reduce the number of variables but retain the maximum amount of data.

Euclidean distance :

• Euclidean distance is used to measure the similarity between observations. It is calculated as the square root of the sum of differences between each point.

Euclidean distance = $\sqrt{(X_1-X_2)^2 + (Y_1-Y_2)^2}$

Turning variable into dummies :

• Variables can be turned into dummy variables. Dummy variables canonly take two values: true (1) or false√ (0). They're used to indicate the absence of acategorical effect that may explain the observation.

Customer	Sales	Date	Gender
1	100	Jan-21	M
3	20	Dec-20	F
2	400	May-22	F
1	500	Jan-22	M
10	45	Aug-21	M
7	300	Dec-21	F
9	250	July-22	F

Customer	Sales	Date	Male	Female
1	100	Jan-21	1	0
1	500	Jan-22	1	0
2	400	May-22	0	1
3	20	Dec-20	0	1
7	300	Dec-21	0	1
9	250	July-22	0	1
10	45	Aug-21	1	0

Exploratory Data Analysis

• Exploratory Data Analysis (EDA) is a general approach to exploring datasets by means of simple summary statistics and graphic visualizations in order to gain a deeper understanding of data.

• EDA is used by data scientists to analyze and investigate data sets and summarize their main characteristics, often employing data visualization methods. It helps determine how best to manipulate data sources to get the answers user need, making it easier for data scientists to discover patterns, spot anomalies, test a hypothesis or check assumptions.

EDA is an approach/philosophy for data analysis that employs a variety of techniques to:

1. Maximize insight into a data set;
2. Uncover underlying structure;
3. Extract important variables;
4. Detect outliers and anomalies;
5. Test underlying assumptions;
6. Develop parsimonious models; and
7. Determine optimal factor settings.

• With EDA, following functions are performed:

1. Describe of user data
2. Closely explore data distributions
3. Understand the relations between variables
4. Notice unusual or unexpected situations
5. Place the data into groups
6. Notice unexpected patterns within groups

7. Take note of group differences

• Box plots are an excellent tool for conveying location and variation information in data sets, particularly for detecting and illustrating location and variation changes between different groups of data.

• Exploratory data analysis is majorly performed using the following methods:

1. Univariate analysis: Provides summary statistics for each field in the raw data set (or) summary only on one variable. Ex : CDF,PDF,Box plot

2. Bivariate analysis is performed to find the relationship between each variable in the dataset and the target variable of interest (or) using two variables and finding relationship between them. Ex: Boxplot, Violin plot.

3. Multivariate analysis is performed to understand interactions between different fields in the dataset (or) finding interactions between variables more than 2.

• A box plot is a type of chart often used in explanatory data analysis to visually show the distribution of numerical data and skewness through displaying the data quartiles or percentile and averages.

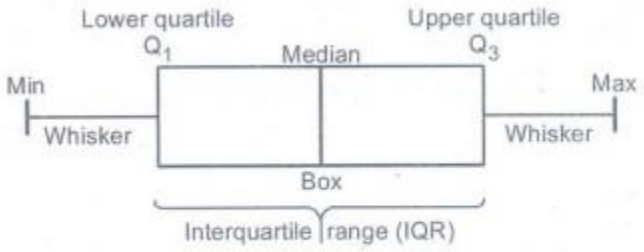


Fig. 1.7.1

Minimum score: The lowest score, exlcuding outliers.

2. Lower quartile : 25% of scores fall below the lower quartile value.

3. Median: The median marks the mid-point of the data and is shown by the line that divides the box into two parts.

4. Upper quartile : 75 % of the scores fall below the upper quartiel value.

5. Maximum score: The highest score, excluding outliers.

6. Whiskers: The upper and lower whiskers represent scores outside the middle 50%.

7. The interquartile range: This is the box plot showing the middle 50% of scores.

• Boxplots are also extremely usefule for visually checking group differences. Suppose we have four groups of scores and we want to compare them by teaching method. Teaching method is our categorical grouping variable and score is the continuous outcomes variable that the researchers measured.

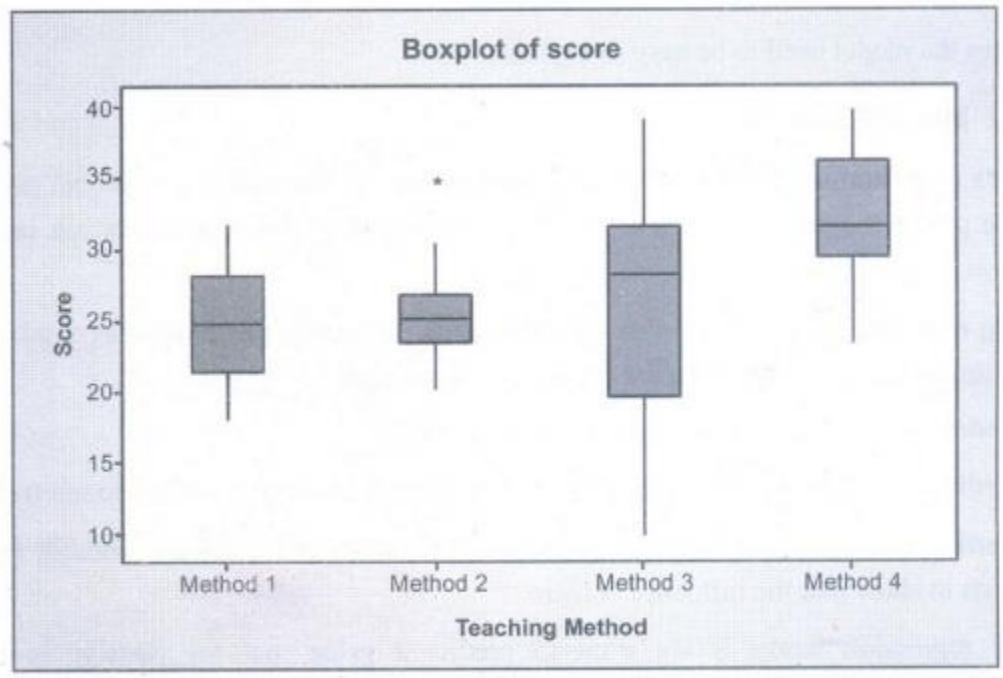


Fig. 1.7.2

Build the Models

• To build the model, data should be clean and understand the content properly. The components of model building are as follows:

a) Selection of model and variable

b) Execution of model

c) Model diagnostic and model comparison

• Building a model is an iterative process. Most models consist of the following main steps:

1. Selection of a modeling technique and variables to enter in the model

2. Execution of the model

3. Diagnosis and model comparison

Model and Variable Selection

• For this phase, consider model performance and whether project meets all the requirements to use model, as well as other factors:

1. Must the model be moved to a production environment and, if so, would it be easy to implement?

2. How difficult is the maintenance on the model: how long will it remain relevantif left untouched?

3. Does the model need to be easy to explain?

Model Execution

• Various programming language is used for implementing the model. For model execution, Python provides libraries like StatsModels or Scikit-learn. These packages use several of the most popular techniques.

• Coding a model is a nontrivial task in most cases, so having these libraries available can speed up the process. Following are the remarks on output:

) Model fit: R-squared or adjusted R-squared is used.

b) Predictor variables have a coefficient: For a linear model this is easy to interpret.

c) Predictor significance: Coefficients are great, but sometimes not enough evidence exists to show that the influence is there.

• Linear regression works if we want to predict a value, but for classify something, classification models are used. The k-nearest neighbors method is one of the best method.

• Following commercial tools are used :

1. SAS enterprise miner: This tool allows users to run predictive and descriptive models based on large volumes of data from across the enterprise.

2. SPSS modeler: It offers methods to explore and analyze data through a GUI.

3. Matlab: Provides a high-level language for performing a variety of data analytics, algorithms and data exploration.

4. Alpine miner: This tool provides a GUI front end for users to develop analytic workflows and interact with Big Data tools and platforms on the back end.

• **Open Source tools:**

1. R and PL/R: PL/R is a procedural language for PostgreSQL with R.

2. Octave: A free software programming language for computational modeling, has some of the functionality of Matlab.

3. WEKA: It is a free data mining software package with an analytic workbench. The functions created in WEKA can be executed within Java code.

4. Python is a programming language that provides toolkits for machine learning and analysis.

5. SQL in-database implementations, such as MADlib provide an alterative to in memory desktop analytical tools.

Model Diagnostics and Model Comparison

Try to build multiple model and then select best one based on multiple criteria. Working with a holdout sample helps user pick the best-performing model.

• **In Holdout Method**, the data is split into two different datasets labeled as a training and a testing dataset. This can be a 60/40 or 70/30 or 80/20 split. This technique is called the hold-out validation technique.

Suppose we have a database with house prices as the dependent variable and two independent variables showing the square footage of the house and the number of rooms. Now, imagine this dataset has 30 rows. The whole idea is that you build a model that can predict house prices accurately.

• To 'train' our model or see how well it performs, we randomly subset 20 of those rows and fit the model. The second step is to predict the values of those 10 rows that we excluded and measure how well our predictions were.

• As a rule of thumb, experts suggest to randomly sample 80% of the data into the training set and 20% into the test set.

• The holdout method has two, basic drawbacks :

1. It requires extra dataset.

2. It is a single train-and-test experiment, the holdout estimate of error rate will be misleading if we happen to get an "unfortunate" split.

Presenting Findings and Building Applications

Presenting and building applications in data science involves several key steps and considerations to effectively communicate insights and deploy data-driven solutions. Here's a structured approach to presenting and building applications in data science:

Presenting Insights:

1. **Understand the Audience:**
    - **Identify stakeholders:** Determine who will be consuming the insights—executives, technical teams, or end-users—and tailor the presentation accordingly.
2. **Define Objectives:**
    - **Clarify goals:** Clearly define the purpose of the presentation—whether it's to inform decision-making, showcase findings, or propose recommendations.
3. **Choose Appropriate Visualizations:**
    - **Select relevant charts:** Use visualizations such as bar charts, line graphs, scatter plots, heatmaps, or interactive dashboards based on the nature of the data and the insights to be conveyed.

4. **Tell a Story with Data:**
    - **Narrative structure:** Structure the presentation like a story, with a clear beginning (context and background), middle (analysis and findings), and end (conclusions and actionable insights).
5. **Focus on Key Findings:**
    - **Highlight insights:** Emphasize the most important findings or trends that align with the objectives of the presentation. Use annotations or call-outs to draw attention to critical points.
6. **Provide Context and Interpretation:**
    - **Explain implications:** Help the audience understand the implications of the findings in the context of their business or domain. Interpret the data insights and connect them to actionable outcomes.
7. **Address Questions and Feedback:**
    - **Be prepared:** Anticipate questions and challenges from the audience. Prepare to provide additional context, explanations, or further analysis as needed.
8. **Use Clear and Concise Language:**
    - **Simplify complexity:** Avoid jargon and technical terms that may confuse non-technical stakeholders. Use plain language to ensure clarity and understanding.

Building Applications:

1. **Define Requirements and Scope:**
    - **Gather specifications:** Work closely with stakeholders to define the functional and non-functional requirements of the data science application. Clarify the scope and expected outcomes.
2. **Data Collection and Preparation:**
    - **Acquire and clean data:** Collect relevant datasets and preprocess them to ensure data quality and consistency. This may involve handling missing values, outliers, and formatting data for analysis.
3. **Choose Modeling Techniques:**
    - **Select algorithms:** Depending on the problem and data characteristics, choose appropriate machine learning or statistical modeling techniques. Consider factors like supervised vs. unsupervised learning, regression vs. classification, etc.
4. **Model Training and Evaluation:**
    - **Train and validate models:** Split the data into training and testing sets. Train the model on the training data and evaluate its performance using appropriate metrics (e.g., accuracy, precision, recall, F1-score).
5. **Integration and Deployment:**
    - **Deploy the application:** Integrate the trained model into the application framework or environment. Consider scalability, performance, and compatibility with existing systems.
6. **User Interface Design (UI/UX):**
    - **Create a user-friendly interface:** Design an intuitive user interface (UI) that allows stakeholders or end-users to interact with the application easily. Consider usability, accessibility, and visual design principles.
7. **Testing and Validation:**

- **Ensure functionality:** Conduct thorough testing to validate the application's functionality, reliability, and performance under different scenarios and inputs.

8. **Documentation and Maintenance:**
   - **Document processes:** Provide comprehensive documentation covering data sources, methodology, model details, and application usage instructions. Plan for regular maintenance and updates to ensure the application remains effective and relevant.
9. **Feedback and Iteration:**
   - **Continuous improvement:** Gather feedback from users and stakeholders to identify areas for improvement. Iterate on the application based on feedback and evolving business needs.