UNIT-1

Introduction to Data Science-

What is Data Science?

Phases of Data Science: Data Acquisition, Cleansing, Exploratory Data Analysis, Data Preparation, Data Modeling.

Engineering Aspects of Data Science: Business Understanding, Data Understanding, Data Preparation, Model Building, Model Evaluation, Hyper Parameter Optimization and Deployment.

Data Science is an interdisciplinary field that involves using scientific methods, algorithms, and systems to extract knowledge and insights from structured and unstructured data. Here are some real-time examples of how Data Science is used in various industries:

1. E-commerce: Online retailers use Data Science techniques to analyze customer data, such as purchase history, search queries, and social media activity, to personalize product recommendations, improve inventory management, and optimize pricing strategies.
2. Healthcare: Healthcare providers use Data Science techniques to analyze patient data, such as medical records, test results, and lifestyle information, to develop personalized treatment plans, predict disease outcomes, and optimize clinical workflows.
3. Finance: Financial institutions use Data Science techniques to analyze market data, such as stock prices, interest rates, and economic indicators, to develop trading strategies, assess risk, and detect fraud.
4. Marketing: Marketers use Data Science techniques to analyze customer data, such as demographic information, online behavior, and purchase history, to develop targeted advertising campaigns, measure campaign effectiveness, and optimize marketing spend.
5. Transportation: Transportation companies use Data Science techniques to analyze data from sensors, GPS devices, and other sources, to optimize route planning, improve fleet management, and reduce fuel consumption.
6. Sports: Sports teams use Data Science techniques to analyze performance data, such as player statistics, training metrics, and game footage, to optimize team strategies, identify talent, and assess player value.
7. Social Media: Social media platforms use Data Science techniques to analyze user data, such as posts, likes, and comments, to develop personalized recommendations, improve content moderation, and detect fake news and hate speech.

Data Science is being used in various industries and domains to make informed decisions, improve processes, and create value for businesses and customers.

Here is the topic-wise course material for each of the phases and engineering aspects of Data Science:

1. **Data Acquisition:**

- Introduction to data sources and formats
- Techniques for collecting data from APIs, databases, web scraping, etc.
- Data quality and data validation techniques
- Handling missing data and dealing with outliers

2. **Data Cleansing:**

- Data pre-processing techniques, such as normalization, standardization, and scaling
- Data cleaning techniques, such as deduplication, imputation, and outlier detection
- Handling inconsistent and incomplete data
- Data transformation techniques, such as encoding, discretization, and feature extraction

3. **Exploratory Data Analysis:**

- Data visualization techniques, such as histograms, scatterplots, and heatmaps
- Statistical techniques, such as correlation analysis, hypothesis testing, and regression analysis
- Exploring relationships between variables and identifying patterns and trends in the data
- Identifying potential biases and limitations in the data

4. **Data Preparation:**

- Techniques for splitting data into training, validation, and test sets
- Feature selection and feature engineering techniques
- Handling class imbalances and other data-related issues
- Strategies for dealing with large datasets and distributed computing

5. **Data Modeling:**

- Introduction to machine learning algorithms, such as regression, classification, clustering, and deep learning
- Techniques for selecting and optimizing models, such as cross-validation and hyperparameter tuning
- Model evaluation techniques, such as accuracy, precision, recall, and F1 score
- Ensembling techniques, such as bagging, boosting, and stacking

6. **Model Deployment:**

- Techniques for deploying models to production, such as web services, containers, and cloud platforms
- Monitoring and debugging models in production
- Strategies for updating and retraining models as new data becomes available
- Ethical considerations in deploying models, such as fairness, accountability, and transparency

7. **Business Understanding:**

- Understanding the business problem and defining the objectives
- Identifying the stakeholders and their requirements
- Defining the success criteria and metrics for evaluating the model
- Communicating the results and insights to stakeholders

8. **Data Understanding:**

- Understanding the data sources, formats, and quality
- Exploring the data to identify potential issues and limitations
- Assessing the data biases and ethical considerations
- Creating a data dictionary and data lineage

9. **Data Preparation (in the context of Engineering Aspects):**

- Identifying the data sources and requirements
- Extracting, transforming, and loading (ETL) the data into a suitable format
- Defining the data schema and data quality standards
- Managing the data storage and access

10. **Model Building (in the context of Engineering Aspects):**

- Selecting the appropriate machine learning algorithm and model architecture
- Defining the model inputs and outputs

- Implementing the model using suitable programming languages and frameworks
- Testing and debugging the model

11. Model Evaluation (in the context of Engineering Aspects):

- Evaluating the model performance against the success criteria and metrics
- Assessing the model robustness, scalability, and maintainability
- Conducting A/B testing and user acceptance testing
- Documenting the model design and implementation

12. Hyper Parameter Optimization (in the context of Engineering Aspects):

- Selecting the hyperparameters and their ranges
- Defining the search space and optimization strategy
- Implementing the optimization algorithm and evaluating the results
- Tuning the hyperparameters to improve the model performance

Data Science is a complex and iterative process that involves several phases or steps. Here is a detailed report on the phases of Data Science:

1. Data Acquisition: In this phase, the raw data is collected from various sources such as databases, websites, social media, and IoT devices. The data can be structured or unstructured, and it can be in different formats such as CSV, Excel, JSON, or XML.
2. Data Cleansing: Once the data is acquired, it needs to be cleaned and pre-processed to remove any errors, inconsistencies, or irrelevant data. This phase involves tasks such as data cleaning, data integration, data transformation, and data normalization.
3. Exploratory Data Analysis (EDA): In this phase, the data is explored and analyzed to gain insights into its properties, distributions, and relationships. EDA involves visualizing the data using graphs, charts, and statistical measures, and identifying patterns and anomalies.
4. Data Preparation: After the data has been cleaned and analyzed, it needs to be prepared for modeling. This phase involves feature selection, feature engineering, data sampling, and data splitting. The data is divided into training, validation, and testing sets, and the features are selected and transformed to create a suitable input for the modeling phase.
5. Data Modeling: In this phase, statistical and machine learning models are trained on the prepared data to make predictions or classifications. The choice of model depends on the type of problem and the available data. Popular

models include linear regression, decision trees, random forests, and deep neural networks.

6. Model Evaluation: Once the models are trained, they need to be evaluated on a separate data set to assess their performance and accuracy. This phase involves cross-validation, ROC curves, confusion matrices, and model selection.
7. Hyperparameter Optimization: In this phase, the model's hyperparameters are tuned to improve its performance on the validation set. Hyperparameters are the model settings that are not learned from the data, such as the learning rate, regularization strength, and the number of layers.
8. Model Deployment: After the model has been trained and optimized, it can be deployed in a production environment to make predictions on new data. This phase involves tasks such as model serving, API design, and scalability testing.

In conclusion, the phases of Data Science are iterative and interdependent, and each phase contributes to the project's overall success. A thorough understanding of these phases is essential for anyone working in the field of Data Science.

A real-time example of the Engineering Aspects of Data Science:

Let's say a company wants to develop a machine-learning model to predict customer churn. The engineering aspects of data science involved in this project would include:

1. Business Understanding: The Company needs to define the problem at hand, i.e., customer churn, and identify the stakeholders, such as the sales and marketing teams. The success criteria for the project could be a reduction in customer churn by a certain percentage.
2. Data Understanding: The Company needs to collect and analyze customer data to understand the factors contributing to churn, such as customer demographics, usage patterns, and customer support interactions.
3. Data Preparation: The Company needs to clean and prepare the data for use in machine learning models. This involves tasks such as data cleaning, feature selection, data sampling, and data splitting.
4. Model Building: The Company needs to develop machine learning models based on the cleaned and prepared data. The model selection, architecture, and training methods are all important considerations in this phase.
5. Model Evaluation: The Company needs to evaluate the models to determine their performance and accuracy. This involves cross-validation, ROC curves, confusion matrices, and model selection.

6. Hyperparameter Optimization: The Company needs to tune the model's hyperparameters to improve its performance on the validation set.
7. Model Deployment: The Company needs to deploy the model in a production environment to make predictions on new customer data. This involves tasks such as model serving, API design, and scalability testing.

By applying these engineering aspects of data science, the company can build a robust and effective machine learning model to predict customer churn and reduce it by a certain percentage, leading to better customer retention and ultimately increased revenue.

Engineering Design:

Engineering Design (ED) can be defined as the process of solving technical problems within requirements and constraints to create new products.

Artificial Intelligence (AI), that "applies advanced analysis and logic-based techniques, including machine learning, to interpret events, support and automate decisions, and take actions"

The CRoss Industry Standard Process for Data Mining (CRISP-DM) is a process model that serves as the base for a data science process. It has six sequential phases:

1. Business understanding – What does the business need?

    - Determine business objectives: You should first "thoroughly understand, from a business perspective, what the customer really wants to accomplish." (CRISP-DM Guide) and then define business success criteria.
    - Assess situation: Determine resources availability, project requirements, assess risks and contingencies, and conduct a cost-benefit analysis.
    - Determine data mining goals: In addition to defining the business objectives, you should also define what success looks like from a technical data mining perspective.
    - Produce project plan: Select technologies and tools and define detailed plans for each project phase.

2. Data understanding – What data do we have / need? Is it clean?

    - Collect initial data: Acquire the necessary data and (if necessary) load it into your analysis tool.
    - Describe data: Examine the data and document its surface properties like data format, number of records, or field identities.
    - Explore data: Dig deeper into the data. Query it, visualize it, and identify relationships among the data.

- Verify data quality: How clean/dirty is the data? Document any quality issues.

3. Data preparation – How do we organize the data for modeling? “data munging”

   - Select data: Determine which data sets will be used and document reasons for inclusion/exclusion.
   - Clean data: Often this is the lengthiest task. Without it, you’ll likely fall victim to garbage-in, garbage-out. A common practice during this task is to correct, impute, or remove erroneous values.
   - Construct data: Derive new attributes that will be helpful. For example, derive someone’s body mass index from height and weight fields.
   - Integrate data: Create new data sets by combining data from multiple sources.
   - Format data: Re-format data as necessary. For example, you might convert string values that store numbers to numeric values so that you can perform mathematical operations.

4. Modeling – What modeling techniques should we apply?

   - Select modeling techniques: Determine which algorithms to try (e.g. regression, neural net).
   - Generate test design: Pending your modeling approach, you might need to split the data into training, test, and validation sets.
   - Build model: As glamorous as this might sound, this might just be executing a few lines of code like “reg = LinearRegression().fit(X, y)”.
   - Assess model: Generally, multiple models are competing against each other, and the data scientist needs to interpret the model results based on domain knowledge, the pre-defined success criteria, and the test design.

5. Evaluation – Which model best meets the business objectives?

   - Evaluate results: Do the models meet the business success criteria? Which one(s) should we approve for the business?
   - Review process: Review the work accomplished. Was anything overlooked? Were all steps properly executed? Summarize findings and correct anything if needed.
   - Determine next steps: Based on the previous three tasks, determine whether to proceed to deployment, iterate further, or initiate new projects.

6. Deployment – How do stakeholders access the results?

   This final phase has four tasks:

   - Plan deployment: Develop and document a plan for deploying the model.

- Plan monitoring and maintenance: Develop a thorough monitoring and maintenance plan to avoid issues during the operational phase (or post-project phase) of a model.
- Produce final report: The project team documents a summary of the project which might include a final presentation of data mining results.
- Review project: Conduct a project retrospective about what went well, what could have been better, and how to improve in the future.