# UNIT 16 DATA SCIENCE FOR SOFTWARE ENGINEERS

## Structure



## 16.0 INTRODUCTION

We are living in data-driven world where making sense out of raw data is the information super power. With each passing day, data is growing exponentially. With every tweet, Facebook post, Instagram picture, YouTube video users are generating massive amounts of data. In addition to the social media popularity, the data generated by Sensors of IoT enabled devices and wearable also generates data at high velocity. Data science is gaining popularity with the emergence of Big Data. Data analysis has been around from last many decades. The data analysis methods have evolved over the period of time. As data scientists can crunch massive amount of data to detect anomalies, patterns, trends, organizations want to leverage data science to reduce cost, explore cross-sell/upsell opportunities, new market opportunities, forecast, recommend for gaining competitive advantage. Data science is an interdisciplinary field consisting of statistics, computer science, Machine learning and others.

Data is considered as strategic asset for businesses. Organizations need to derive insights and gain value from the data they have to solve business problems and succeed in their business.

Data science helps businesses to make data driven decision making. Organizations can apply the data collection, data preparation and data analysis methods to mine massive volume of data to understand customers' behavior and explore the business opportunities to influence their customers.

## 16.1 OBJECTIVES

After going through this unit, you should be able to

- understand key concepts, background and definition of Data Science
- understand the relationship between Data science and Big Data
- understand various lifecycle stages of the Data science solution engineering
- understand the popular methods of data analysis
- deep dive into a case study for predicting the probability of product purchase

## 16.2 APPLICATIONS OF DATA SCIENCE

Given below are the main applications of data science:

- **Smart Recommendations:** Data science applications analyze massive historical data and provide effective recommendations.
- **Big Data Analytics:** Data science applications analyze the massive data such as sensor data, IoT data, social media data, log data and such to gain insights.
- **Web Search:** Analyze the web-scale data to provide accurate search results.
- **Data Analytics:** Through analysis and visualization of data we can gain insights that can help us to make data-driven predictions, forecasting and prescription.
- **Business Intelligence:** Data science applications are mainly used for business intelligence purposes. Organizations can leverage business analytics applications using data science to identify the challenges, identify cross-sell/upsell opportunities, evaluate sales offers, explore marketing opportunities, product pricing, cost optimization avenues, ROI (return on investment) analysis, revenue loss prediction, store planning, customer analytics, seasonality analysis and others.
- **Healthcare:** Data generated by wearable is used for active monitoring and to reduce the health complications.
- **Finance:** Data science methods can be used for many finance applications such as fraud detection, risk detection, data-driven insights, pattern recognition, predictive analytics and such.
- **Automobile:** Training the driverless vehicles using the training data and evaluating the performance of the models.
- **Supply Chain and Inventory Management:** Leverage data science for forecasting, demand planning and revenue forecasting.
- **Telecom:** Use Data science methods to understand the customer churn forecasting.

There are numerous other Data science applications such as election campaigning, chatbots, virtual assistants, automated cars, disaster prediction, product pricing, preventive maintenance, customer retention, pattern recognition, online advertisement, demand forecasting, and trend analysis, identifying campaign effectiveness, recommendations and such.

We have depicted the key business value derived from data science in Figure 16.1.

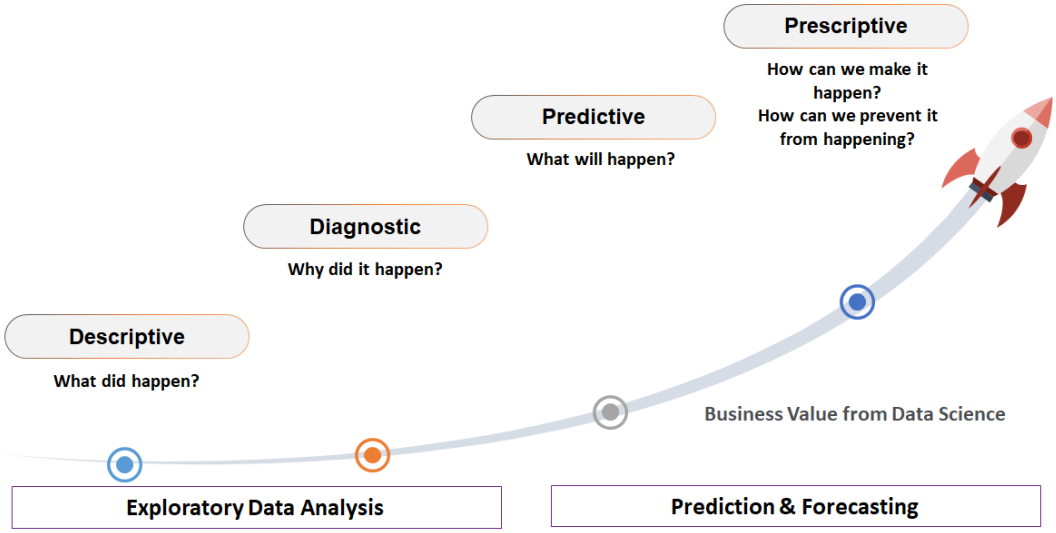


**Figure 16.1 : Business Value from Data Science**

As depicted in Figure 16.1, through exploratory data analysis, we can describe the events, phenomenon and scenarios. We can also look at the data dimensions and do the diagnosis to understand why the event happened. After data preparation, data analysis, model development we can predict what will happen in future and prescribe what can be done to make it happen or what can be done to prevent it from happening.

Given below are the popular applications of data science:

- Computing the most relevant online advertisement based on the user's interests, search history and showing the advertisement in real time.
- Finding the relevant product recommendations based on user's transaction history, browsing history, similar users' interest
- Flagging an email as spam through continuous learning of the features
- Identifying the obstructions by autonomous vehicles in real time.
- Understand the customer behavior in real time.
- Predicting customer churn probability using the signals.

## 16.3 BACKGROUND OF DATA SCIENCE

Data science is closely related to the field of statistics and mainly uses statistical methods for data analysis. As such data analysis and the analysis methods have been since 1960s. The term "Data science" can be traced back to 1974 when it was coined by Peter Naur and was later popularized by C.F.Jeff Wu in 1985.

Let us understand how the data analysis has evolved by looking at the recommendation use case. In the early 2010 we used the formal and rigid business rules for product recommendation; It was followed by linear regression model in 2011. In 2013 we achieved greater accuracy using logistical regression. Decision trees were widely used in 2014. Organizations like Amazon and Netflix leveraged collaborative filtering in 2015 for product recommendations. Bayesian network was used for recommendation in 2017. From 2019 onwards organizations are heavily using the machine learning and deep learning methods for product recommendation to achieve better accuracy.

## 16.4 DEFINITION OF DATA SCIENCE

Data science is an interdisciplinary field that deals with collecting, analyzing, visualizing data using statistical and machine learning methods to convert raw data into data products. Simply put, a data scientist is tasked with making sense and actionable insights out of raw data.
We have depicted various disciplines of data science in Figure 16. 2.

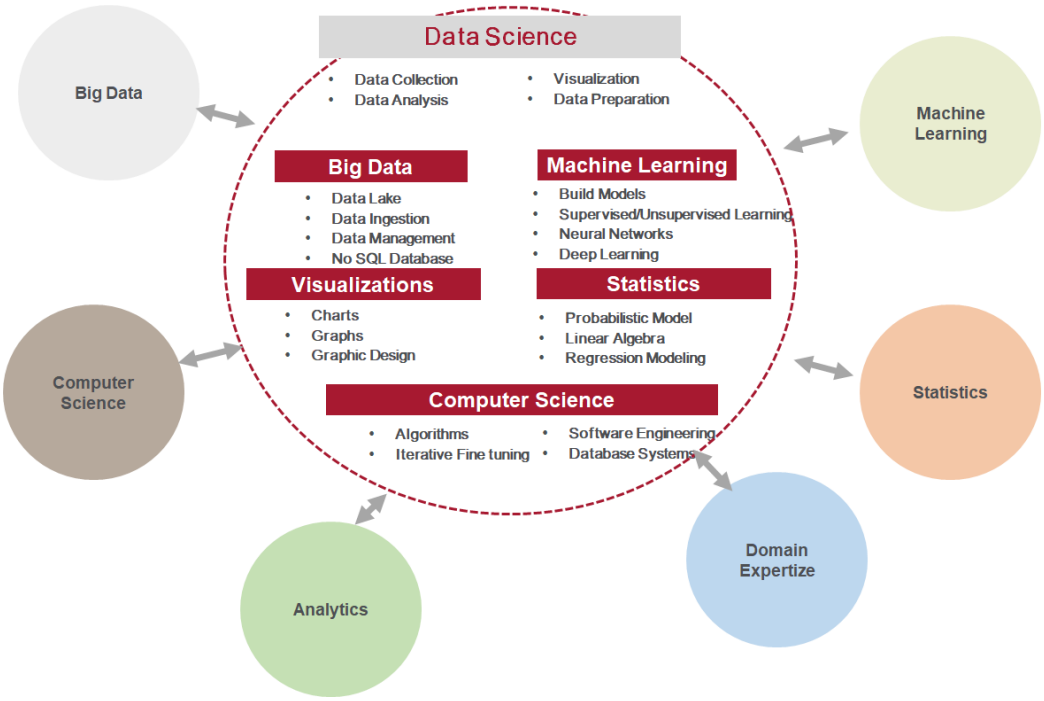


**Figure 16.1 : Data Science Disciplines**

Data science leverages various fields of knowledge such as machine learning, statistics, analytics, computer science and Big Data. Domain expertize is an essential skill needed for data science solutions.

## 16.5 DATA SCIENCE TOOLS

As data science is an interdisciplinary field, a data scientist needs to be familiar with multiple tools and technologies as given below:

- AI and Machine Learning: Deep learning methods, regression analysis, ML Models, Support vector machine (SVM), supervised and unsupervised learning, TensorFlow, Pytorch, Supervised learning includes methods such as classification, regression, logistical regression, and recommendation and decision trees. Unsupervised learning includes methods such as clustering, anomaly detection, visualization, association rule engine.
- Big Data: Apache Spark, Apache Hadoop, NoSQL databases (DynamoDB, MongoDB etc.)
- Statistics: Probability, Vector Algebra, Calculus, Linear Regression, Logistical regression,
- Data visualization: Qlik, PowerBI, Tableau
- Domain skills: Thorough understanding of the problem domain (finance, retail, commerce etc.)
- Data science platforms: Matlab, IBM Watson Studio, Anaconda, Numpy, Jupyter,

- Operations Research: Decision Tree, decision modeling (deterministic, probabilistic)
- Programming Languages: R, Python, Scala, Julia, Java, C
- Model Development: Matlab, Octave, MADLib

## 16.5 DATA SCIENCE AND BIG DATA

The proliferation of Big Data has propelled the increase in popularity of Data Science. The sources that generate Big Data increase in volume, velocity (massive rate of increase), veracity and variety (structured, semi-structure, and unstructured) posing challenges in organizing and analyzing the Big Data. Examples of Big Data are sensor data, user click data, tweet data and such. We cannot easily manage Big Data through traditional databases. Hence Big Data systems don't strictly adhere to ACID (Atomicity, Consistency, Integrity and Durability) properties but aim for eventual consistency as per CAP (Consistency, Availability, Partition) theorem. Some of the popular Big Data systems are NoSQL products such as MongoDB, DynamoDB, CouchDB and such.
As part of Data science development, we build, explore and fine tune various models to handle the massive data sets.

Big Data provide the massive data management, data processing technologies such as Apache Hadoop, and NoSQL databases for data science methods.

## 16.6 PHASES INVOLVED IN DATA SCIENCE PROCESS

We have depicted various phases involved in the data science process in the Figure 16.3.

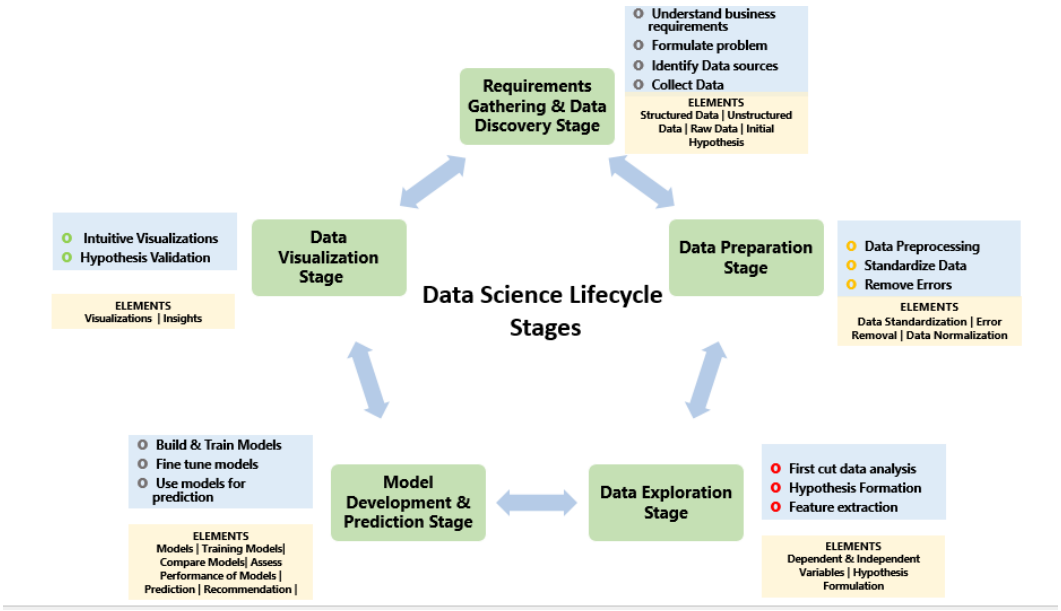


**Figure 16. 2 : Lifecycle Stages of Data Processing**

In data science the data is processed through various stages in the pipeline. We have depicted the stages in Figure 16.3. During each of the stages in the pipeline, the raw data undergoes series of analysis and changes. The lifecycle stages are cyclic as we continuously use the feedback from each stage to its previous stage.

We shall look at each of the stages in next sections.

### 16.6.1 Requirements Gathering and Data Discovery

We need to understand the business requirements, the goals, focus areas and challenges. We need to understand the key data sources, data inventories, existing data analysis tools and technologies, the subject matter experts (SMEs), project sponsor, main stakeholders, time and budget for the program. It is essential to also understand any existing data analysis initiatives undertaken to learn from the past experience.

Domain experts, business analysts for the business domain help us to provide the heuristics related to the business domain using which we can formulate the initial hypothesis. Once we thoroughly understand the business domain we can then formulate the problem for applying the data science methods.

Once we fully understand the requirements, we need to collect data from various sources such as data lakes, relational databases, ERP systems, internal and external social media systems, collaboration systems and such. For instance for the customer behavioral analysis, we need to gather data from various sources such as CRM system (user case details), user profile details, order history, product review system (product name, review details), email system, social media (user's social media handle, user details) and such.

### 16.6.2 Data Preparation stage

To improve the quality of the extracted data, we need to cleanse the data to standardize the format, structure and remove any inconsistencies (duplicates, inconsistencies, missing values, errors and such). Data preparation is a time consuming stage wherein a data scientist separates signal from noise. The data conditioning can be performed by ETL (Extract, Transform and Load) tools. As part of data conditioning, we do the data normalization, conversion to same case and such.

Given below are some of the main activities as part of the data conditioning stage:

- Removing duplicates
- Bring the data to a common or standard format
- Infering the missing values
- Smoothing the data by removing the noise
- Filtering and sampling the data
- Making the data types consistent across the entire data set
- Replace missing values with their mean or median values

The conditioned data will be used for exploration and model development.

### 16.6.3 Data exploration stage

During this stage, we extract the features from the data and process the data. Based on the extracted features we can perform first cut analysis and form the initial hypothesis. We also classify, cluster the data into the logical categories. We identify the relationship between variables during this stage. We then select the key predictors that are essential for the recommendation and for prediction.

For instance, in order to model the price of a used car, we need to examine massive amount of historical used car sales transactions consisting of various attributes. However key features such as car mileage, car age, kms driven, engine capacity, purchase price have direct correlation with the car's market price. These key features directly influence the final price for the used car. Data scientists look at the data and try to correlate the useful features and form the hypothesis. In the used car price

prediction use case, the hypothesis could be “The car age is inversely correlated to the final price (with the increase in car age, the final price decreases)”.

### 16.6.4 Model development and Prediction stage

We design and develop the machine learning models in this stage. We train and fine tune the model and test their performance on an iterative basis. Data scientists evaluate various models for the selected data and finally the model that provides the highest performance will be selected for prediction.

We use many statistical methods such as regression analysis, qualitative methods, decision tree models, deep learning models for data analysis during this stage. We evaluate the performance of the model using various data points such as the accuracy of the model prediction, model performance on various training and test data sets and such.

We then use the selected models for the prediction.

### 16.6.5 Data Visualization stage

In the final stage, we visually communicate the obtained insights using reports, business intelligence (BI) tools to help the audience make informed decisions. We also document the key findings, observations and the behavior of the model for various test cases. Often we use the visual representations such as charts, infographics to communicate the findings to the stakeholders.

Once the model is accepted by the stakeholders, we operationalize the model and use it for business prediction. We need to continuously fine tune the model based on the feedback and learning from the real world data.

A sample data visualization is shown in Figure 16.4.

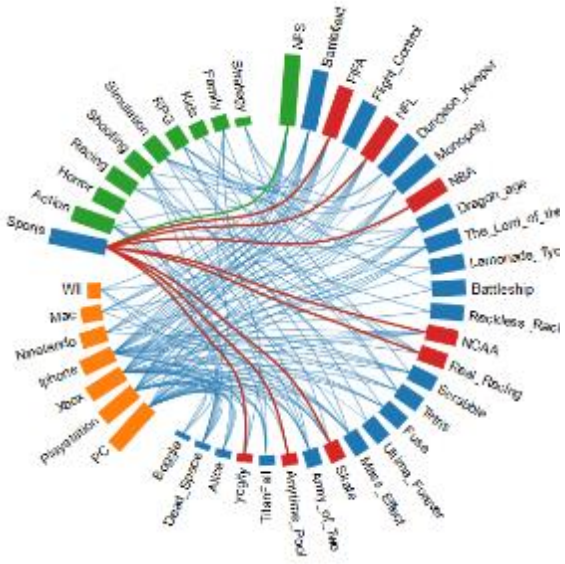


**Figure 16.4 : Sample Data Visualization**

### 16.6.6 A Sample Case Study

In this example we shall look at a sample case study to understand the various stages of data science pipeline events.
The use case we have considered below is for predicting the most popular (and moving) product at a store. Based on the product’s popularity, the organization can chose to replenish the stock of the product faster.

**Requirement gathering and Data collection stage**

In this phase, we collect all the data needed to predict the popular product at a given store. We collect the below given data at a product level for each store.

**Table 16.1 : Product sales data in raw format**

| Top 5 in last month (Yes/No) | Top 5 in last year (Yes/No) | Brand campaign (Yes or No) | Buyer Age Group 10-18 yrs. (Yes/No) | Buyer Age Group 19-34 yrs. (Yes/No) | Buyer Age Group 30-50 yrs. (Yes/No) | Buyer Income (10K – 30K) Yes/ No | Product | Store | Sales |
|---|---|---|---|---|---|---|---|---|---|
| 1 | 0 | 1 | 0 | 1 | Nil | 1 | ABC | Store1 | 500 |
| 1 | 1 | Yes | 0 | 0 | 1 | 1 | XYZ | Store1 | 1500 |
| | | | 0 | | | 1 | ABC | Store1 | 500 |

In table 16.1, "1" denotes "yes" and "0" denotes "No"

Likewise we collect all the possible data from all stores for all combinations of identified features

**Data Cleansing Stage**

During data cleansing stage, we identify the following:

- Errors: If there are any errors in the data entry we will fix it
- Missing data: If the data for any features are missing, we can either take the median value to fill the data or drop that feature if it is missing in both training and test data set.

In the above example one of the values is "Yes" that needs to be coded to "1". Another value is "Nil" that will be converted to "0". Third row of data is empty for many features hence we will discard that row for the training purpose. The conditioned data is depicted in table 16.2.

**Table 16.1 : Conditioned Product sales data**

| Top 5 in last month (Yes/No) | Top 5 in last year (Yes/No) | Brand campaign (Yes or No) | Buyer Age Group 10-18 yrs. (Yes/No) | Buyer Age Group 19-34 yrs. (Yes/No) | Buyer Age Group 30-50 yrs. (Yes/No) | Buyer Income (10K – 30K) Yes/ No | Product | Store | Sales |
|---|---|---|---|---|---|---|---|---|---|
| 1 | 0 | 1 | 0 | 1 | 0 | 1 | ABC | Store1 | 500 |
| 1 | 1 | 1 | 0 | 0 | 1 | 1 | XYZ | Store1 | 1500 |

**Data Exploration Stage**

During this stage, we identify the prominent features that influence the overall product's popularity. In the above example, we identify that demographic data has greatest influence on the overall product sales.

We understand the features that have positive/linear correlation with the product sales, non-linear correlation with the product sales.

**Prediction Stage**

In this stage, we predict the product sales for each store. For this, we develop the models that use the dependent features to predict the overall product sales for each store. We can explore various models given below against the test data and compare the accuracy and performance.

- Gaussian Naive Bayes
- Logistic Regression
- Support Vector Machines
- Perceptron
- Decision Tree Classifier
- Random Forest Classifier
- KNN or k-Nearest Neighbors
- Stochastic Gradient Descent
- Gradient Boosting Classifier

Once we finalize the model we can train it further and fine tune to achieve better accuracy. We can then use the model to predict the top selling products.

**Data visualization stage**

In the data visualization stage, we provide the top predicted products for each store so that the store owner can replenish the inventory for the popular products as depicted in the Figure 16.5.

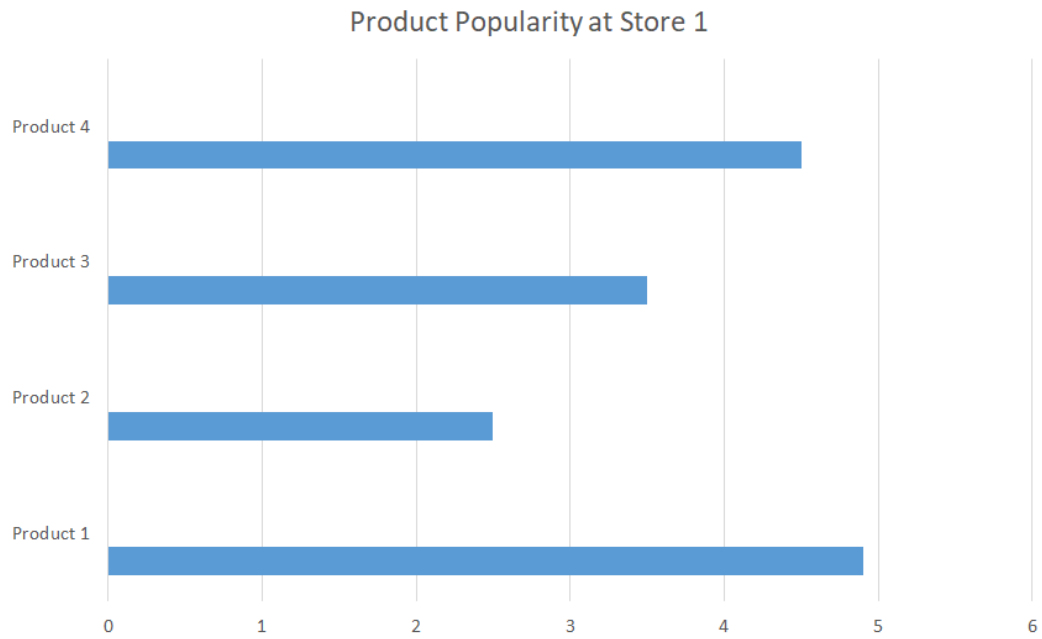


**Figure 16.3 : Top producs at store 1**

## 16.7 DATA SCIENCE METHODS

In this section we shall look at the most popular data analysis methods we use in the data science projects.

### 16.7.1 Clustering Method

As part of clustering method, we group the objects into a logical cluster/group based on the object attribute. As part of unsupervised learning, the model groups the data based on the labels. For instance, as part of demographic data, we can define groups as follows:

- Users less than 20 years
- Users between 20-40 years
- Users more than 40 years

During exploratory analysis, we can form clusters and create rules and predictions based on the clusters. The main clustering methods are hierarchical clustering, fuzzy clustering, K-means clustering, nearest neighboring

Use cases
In Data science we heavily use the clustering methods for group-based personalization, customer segmentation, image processing

### 16.7.2 Collaborative Filtering or Similarity Matching Method

In collaborative filtering the system identifies the users with similar interests (based on likes or purchases of similar products). The information about the set the users with similar interests will then be used for recommending the new products for the users. Collaboration filtering uses set of rules to understand the similarity between users, products and information sources.

For examples if user A has same interest as that of user B while purchasing a book, then we can recommend user A's books to user B which user B has not yet purchased.

A variation of collaboration filtering is to understand the association of products (which products are often purchased together) known as co-occurrence grouping. Based on this insight, we can recommend the closely related products together to the customer.

Use cases
Collaboration filtering is used in recommendation systems for cross-sell and up-sell opportunities such as the following:

"Customers who bought this item also bought these items"

"Based on your purchase history, we recommend these products"

"You may also like these products"

Collaborative filtering is also used for campaigns, product promotions and product offers.

☞ **Check Your Progress 1**

1. In ____ phase we do the data pre-processing.
2. The key features of the data set are extracted in _____ phase
3. we group the objects into a logical cluster/group based on the object attribute in _____ method
4. _____ uses set of rules to understand the similarity between users, products and information
5. Understanding the association of products happens in _____

### 16.7.3 Regression Methods

Regression methods such as linear regression and logistical regression methods identify the relationship (as a function) between dependent variables and independent variable. Based on the learnt relationship the regression methods can predict the independent variable using the key predictor dependent variables.

For instance, we can get the historical prices of house sales from past 5 years. Based on the historical data we can use logistical regression between the independent variable (house price) with the dependent variables (such as number of beds, location, carpet area, amenities, nearby point of interests) to predict the likely price for houses.

Use cases
Regression methods are mainly used for prediction and forecasting scenarios such as demand forecasting, price prediction. Logistical regression is also used for predicting the probability of an event (such as customer churn event or loan default event) given the key dependent variables.

### 16.7.4 Classification Methods

As part of supervised learning, classifiers are given labelled data to understand the data. Once the classifier is fully trained, it can be used to label the new data.

The main classification methods are decision tree method, naïve Bayes classification and Confusion matrix. We will be using decision tree classification for a product purchase use case using the attributes in the next section.

In the decision tree method we structure the decisions in a tree format. The test points (known as nodes) with the Boolean values (yes or no) acting as branches. Once we fully traverse a path in the tree, we can reach the predicted value.

Use cases
Classification methods are mainly used for prediction and recommendation scenarios.

### 16.7.5 Other Methods

We have given other common methods for Data Science based solutions:

- Time series prediction methods to analyze the events occurring at regular time intervals
- Text analysis and sentiment analysis using methods such as bag of words, TFIDF, topic modelling to gain insights from the text.
- Profiling involves characterizing the event or user or an object.
- Reduction involves reducing the large data set into a smaller data set that represents the key samples from the large data set.

## 16.8 DATA SCIENCE PROCESS FOR PREDICTING THE PROBABILITY OF PRODUCT PURCHASE

In this sample case study let us look at various steps involved in the predicting the probability of product purchase and understanding of the key features.

**Requirement gathering and Data collection stage**

The sample case study is aimed at understanding the key features that influence the probability of product purchase and develop a prediction model given the key feature values.

We have provided a sample data collected for eleven historical transactions in table 16.3 for reference.

**Table 16.3 : Product purchase data**

| Age_LT30 | Age_GT30 | In_LT50 | In_GT50 | Loyal_CT | onSale | Prod_buy |
|---|---|---|---|---|---|---|
| 0 | 1 | 0 | 1 | 1 | 1 | 1 |
| 0 | 1 | 0 | 1 | 0 | 1 | 1 |
| 1 | 0 | 0 | 1 | 1 | 1 | 1 |
| 0 | 1 | 0 | 1 | 1 | 1 | 1 |
| 1 | 0 | Yes | 0 | 1 | 1 | 0 |
| 0 | 1 | Correct | 0 | 1 | 1 | 1 |
| 1 | 0 | Y | 0 | 0 | 1 | 0 |
| 1 | 0 | Less than 50 | 0 | 0 | 0 | 0 |
| 0 | 1 | Confirmed | 0 | 0 | 1 | 1 |
| 1 | 0 | One | 0 | 1 | 1 | 0 |
| 1 | 0 | NA | 0 | 1 | 1 | 0 |

We have collected the key features that influence a product purchase in an e-commerce portal:

- Age_LT30 stands for age less than 30, indicates if the buyer is aged below 30 years. 1 indicates Yes and 0 indicates no
- Age_GT30 stands for age greater than 30, indicates if the buyer is aged below 30 years. 1 indicates Yes and 0 indicates no
- In_LT50 stands for income less than 50, indicates if the buyer's monthly income is less than 50K. 1 indicates Yes and 0 indicates no
- In_GT50 stands for greater less than 50, indicates if the buyer's monthly income is greater than 50K. 1 indicates Yes and 0 indicates no
- Loyal_CT stands for loyal customer who has enrolled into the loyalty program. 1 indicates Yes and 0 indicates no
- onSale stands for product on sale. 1 indicates Yes and 0 indicates no
- Prod_buy stands for product buy. 1 indicates Yes and 0 indicates no

We have collected eleven historical values for the illustration of the case study.

**Data Cleansing Stage**

In this phase, we fix the data errors and condition the data so that it can be easily modeled using one of the existing models. The highlighted rows in the table need to be properly coded. For accurate model training we need to use the correct code. Table 16.4 provides the correctly coded values.

**Table 16.2 : Product purchase data coded**

| Age_LT30 | Age_GT30 | In_LT50 | In_GT50 | Loyal_CT | onSale | Prod_buy |
|---|---|---|---|---|---|---|
| 0 | 1 | 0 | 1 | 1 | 1 | 1 |
| 0 | 1 | 0 | 1 | 0 | 1 | 1 |
| 1 | 0 | 0 | 1 | 1 | 1 | 1 |
| 0 | 1 | 0 | 1 | 1 | 1 | 1 |
| 1 | 0 | 1 | 0 | 1 | 1 | 0 |
| 0 | 1 | 1 | 0 | 1 | 1 | 1 |
| 1 | 0 | 1 | 0 | 0 | 1 | 0 |
| 1 | 0 | 1 | 0 | 0 | 0 | 0 |

| | | | | | | |
|---|---|---|---|---|---|---|
| 0 | 1 | 1 | 0 | 0 | 1 | 1 |
| 1 | 0 | 1 | 0 | 1 | 1 | 0 |
| 1 | 0 | 1 | 0 | 1 | 1 | 0 |

**Data Exploration Stage**

In this stage, we identify the key features and its relation to the product purchase decision. We apply the statistical methods to understand the data distribution and the feature relationship. We can infer the following based on our analysis

- Feature Age_GT30 is positively and linearly correlated to prod_buy. Conversely Age_LT30 is inversely related to prod_buy.
- Feature IN_GT50 is positively and linearly correlated to prod_buy. Conversely In_LT50 is inversely related to prod_buy.
- Loyal_CT and onSale are positively correlated to prod_buy

We can use these features and insights for building and training models.

**Prediction Stage**

In this stage we build and evaluate various models. We can evaluate various models such as linear regression, logistic regression and others and understand their performance. We have selected the decision tree model as it factors all the key features. We can visualize the combination and impact of various features with the decision tree model.

We validate the decision tree model with the test data to ensure that the decision tree model does the accurate prediction.

**Data visualization stage**

Figure 16.6 provides sample decision tree for our product purchase prediction.

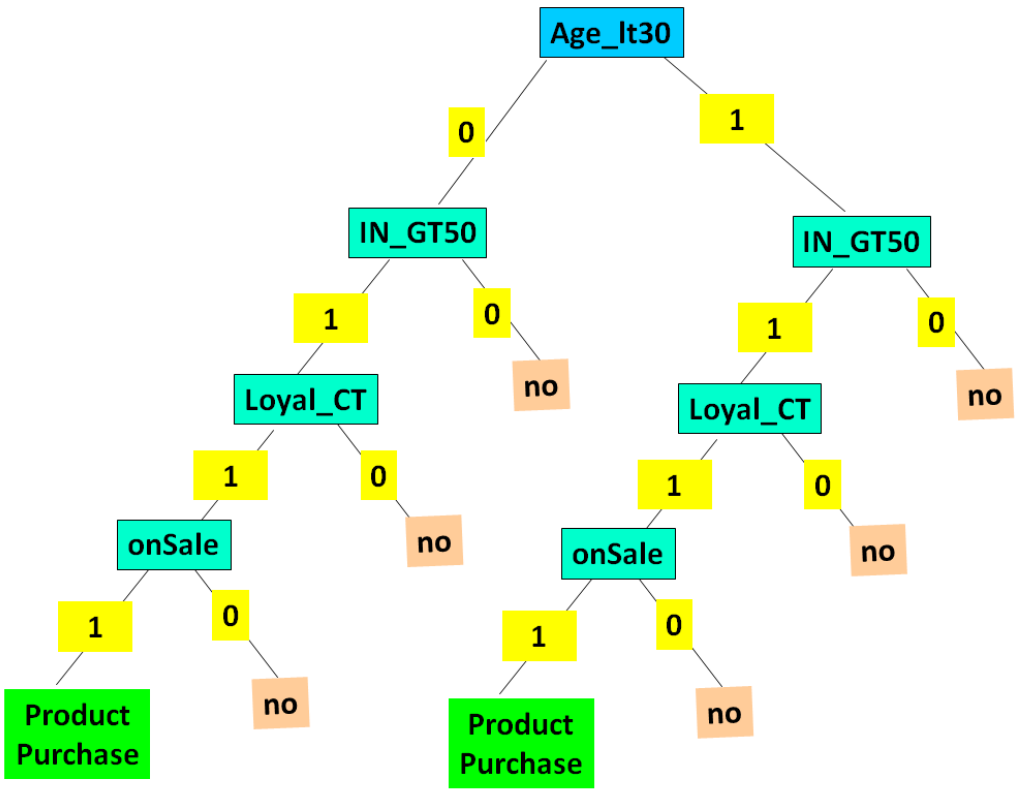


**Figure 16.4 : Decision Tree for Product Purchase**

☞ **Check Your Progress 2**

1. _____methods identify the relationship (as a function) between dependent variables and independent variable.
2. In _____ method, labelled data is used for understanding of the data.
3. _____ involves characterizing the event or user or an object
4. _____ involves reducing the large data set into a smaller data set

## 16.9 SUMMARY

In this unit, we started discussing main applications of the data science. We discussed the interdisciplinary nature of the data science as it involves methods from various fields such as computer science, big data, statistics, analytics, domain knowledge, machine learning and others. We also looked the key tools of data science such as R, Python, Matlab others. We discussed the relationship between data science and Big data. In next section we had detailed discussion on various phases of the data science lifecycle stages. The requirements gathering and data collection stage involves gathering all the data sources, data preparation stage conditions the data and fixes the errors. Data exploration stage involves close examination of the key features and building the hypothesis. In Prediction stage we build models to predict the event. Finally in the data visualization stage we document and showcase our findings. We also had a deep dive discussion on various data science methods related to clustering, collaborative filtering, regression and classification. Finally we look at a detailed case study for using data science method for predicting a product purchase probability.

## 16.10 SOLUTIONS/ANSWERS

**Check Your Progress 1**

1. preparation.
2. data exploration
3. clustering
4. collaborative filtering
5. co-occurrence grouping

**Check Your Progress 2**

1. Regression.
2. Classification
3. Profiling
4. Reduction

## 16.11 FURTHER READINGS

**References**

Provost, F., & Fawcett, T. (2013). *Data Science for Business: What you need to know about data mining and data-analytic thinking.* " O'Reilly Media, Inc.".

Van Der Aalst, W. (2016). Data science in action. In *Process mining* (pp. 3-23). Springer, Berlin, Heidelberg.

Amershi, S., Begel, A., Bird, C., DeLine, R., Gall, H., Kamar, E., ... & Zimmermann, T. (2019, May). Software engineering for machine learning: A case study. In *2019 IEEE/ACM 41st International Conference on Software Engineering: Software Engineering in Practice (ICSE-SEIP)* (pp. 291-300). IEEE.

Agarwal, R., & Dhar, V. (2014). Big data, data science, and analytics: The opportunity and challenge for IS research.