What is Data Science?

INTRODUCTION TO DATA SCIENCE

Husnain Baidar | Tuesday, August 30, 2022

Table of Contents

Agenda of today's video

1. Need for Data Science?

2. What is Data Science?

3. Data Science vs Business intelligence

4. Prerequisites for learning Data Science

5. What does a Data scientist do?

6. Data Science life cycle with use case

7. Demand for Data scientists

Need for Data Science?

One of the examples where data science is used, and the car needs to take a lot of decisions in this whole process whether to speed up whether to apply the brain take a left turn right turn or slow down so all these decisions are basically a part of data science. Data science is also used in airline industry for better weather prediction, better route selection, better prediction of passenger demand. The airline industry can do better route planning so that there are less cancellations occur and people are less frustrated. We can use predictive analytics and predict any delays that are there so that some seats can be rescheduled ahead of time and there are no last-minute changes Data science can also be used to make promotional offers and the finally is what kind of planes should be used. Data science can also be used in the logistics industry so companies like FedEx they use data science models to increase their efficiency drastically to optimize the roads and cut costs and so on so before their delivery truck sets out, they determine which is the best possible route to ship their items. so what is data science used for these are some of the main areas where data science is used of better decision-making there are always tricky decisions to be made so which is the right decision which way to go so that is one area then for predicting for performing predictive analysis like for example can we predict delays like in the case of Airlines can we predict demand for certain products let's say any commerce that is the second area third area is pattern discovery of pattern recognition is there a pattern in which people are buying items.

What Is Data Science?

Data science is the study of using modern tools and techniques to find unseen patterns, derive meaningful information, and make business decisions. Data science is used in many recommendation systems. Like recommendation of movies, tv shows, books, courses etc. Data science is also used in route prediction. Predictive analysis of mechanical parts can also be done via data science predictive modeling. Real time election, cricket, football, F1 race winner prediction is also answered by data science.

Here are the various steps in data science:

- Asking the right question and exploring the data
- Modeling the data using various algorithms
- Communicating and visualizing the results

BI vs DS

Criterion	BI	DS
Data Source	Streamlined structured data.	Uses both structured and unstructured data.
Method	Analytical in nature.	Scientific in nature.
Skills	Mostly used for visualization and statistics purposes.	Same use case as BI but also includes machine learning.
Focus	Past and present data.	Present data and future prediction.

Prerequisites for learning Data Science

For data science there are three essential traits required for to be a data scientist.

1. Curiosity
2. Common sense
3. Communication skills

Skills for Data analysis

1. Machine Learning
2. Modeling
3. Statistics
4. Programing (most common Python and R is also used)

Tools used in Data analysis

1. SAS
2. Jupyter
3. R Studio
4. MATLAB
5. Excel
6. RapidMiner

Skills for Data warehousing

1. ETL
2. SQL
3. Hadoop
4. Apache Spark

Tools used in Data warehousing

1. Informatica / Talend
2. AWS Redshift

Skills for Data Visualization

1. R
2. Python libraries

Tools used in Data Visualization

1. Jupyter
2. Tableau
3. Cognos
4. RAW

Skills for Machine Learning

1. Algebra
2. ML Algos
3. Statistics

Tools used in Machine Learning

1. Spark Mlib
2. Mahot
3. Azure ML studio

What does a Data scientist do?

A data scientist first gets a real word problem for which he/she is required to find an answer so the people can benefit from your findings. After getting a question a data scientist gathers raw data on which he/she performs different processes and analyze the data then get meaningful data through which they provide useful insights.

Basics and important techniques that a data scientist should know

1. Regression
2. Clustering
3. Decision Tree
4. SVM
5. Naïve Baiyes

Regression

In regression you try to let's say come up with a continuous number so the difference between regression and let's say a classification is that in case of classification those are discrete values whereas here we are talking about regression where you I'd say you are trying to predict the temperature which is a continuous value or the share price which is a continuous value so that is regression so you need to know what regression is how to perform regression.

Clustering

clustering is one of the unsupervised learning techniques in this case there is no label data that is available and you get some data and then you want to put this into some shape so that you can analyze it and you try to make sense out of it let's say you have one example as you have a list of cricketers and they have not been marked as bowlers and batsman nor all-rounders or whatever so you just have their names and maybe how many runs they scored how many wickets they have taken and so on but there is no readily available information saying that okay this person is a batsman this person is a bowler. So how do we find out? So, then we put this into a clustering mechanism and then the system will say that okay these are the people who are all who have all scored good amount of so they belong to one cluster these are all the people who have taken good amount of wickets so they belong to one cluster and maybe here are some people who have taken good amount of wickets these belong to the other cluster. So, people who have taken most wickets are labeled as bowler the people who scored most runs are labeled as batters and the hybrids are labeled as al rounders.

Data Science life cycle with use case

Concept study

The first step is the concept study in this step it involves understanding the business problem asking questions get a good understanding of the business model meet up with all the stakeholders understand what kind of data is available.

Here is an example:

Predict the price of a 1.25 carat diamond and there may be relevant information inputs that are available, and we want to predict the price.

Data preparation

The second step is data preparation data gathering and data preparation also known as data munging or sometimes it is also known as data manipulation so what happens here is the raw data that is available may not be usable in its current format for various reasons so that is why in this step a data scientist would explore the data he will take a look at some sample data if we pick there are millions of Records pick a few thousand records and see how the data is looking are there any gaps is the structure appropriate to be fed into the system are there some columns which are probably not adding value may not be required for the analysis very often these are like names of the customers they will probably not add any value or much value for an analysis perspective.

Ways to fill missing values

Firstly, check if the dataset is huge so you can easily remove the missing value rows. But in the case of a small dataset, we can substitute missing value via mean, median and mod using panda's dataframe.

Model planning

EDA is exploring and finding out what are the data types and what is the is the data clean in in each of the columns what is the maximum minimum value.

Visualization techniques

1. Histograms
2. Box plots
3. Scatter plots, etc.

Model building

After deciding model and algorithm you're going to use if you are trying to do machine learning you need to pass your 80% the training data or rather you use that training data to train your model and the training process itself is iterative so the training process you may have to perform multiple times and once the training is done and you feel it is giving good accuracy then you move on to test so you take the remaining twenty percent of the data to test. So, the test data is now used to check the accuracy or how well our model is performing and if there are further issues let's say model is still during testing of the accuracy is not good then you may want to retrain your model or use a different model so this whole thing again can do again. If the model passes the test, then we go ahead with the development of that.

Tools

1. R
2. Python
3. MATLAB
4. SAS

Linear regression

Linear regression is basically a relation between an independent variable and a dependent variable. Here is its mathematical formula:

$Y = mX + c$

Where Y is the dependent variable and X is the independent variable.

Where m is the slope of the line and c is the Y intercept.

Communicate results

After getting a required answer or a required output then we must make a presentation or a dashboard or we visualize the findings via some tools and then communicates/explain these findings to the appropriate stake holders.

Operationalize

After your presentations are accepted then they put it into factors and thereby they will be able to improve or solve the problem that they stated in step one.

Demand for Data scientists

The demand for data scientists is currently huge and the supply is very low so there is a huge gap so what are some of the industries with high demand for data scientists.

Here are some of the industries where data scientists are in demand:

1. Gaming
2. Marketing
3. Healthcare
4. Finance
5. Technology
6. Aviation etc.