Data Science and Deep Reinforcement Learning for Enormous Data

Mrs. Lakhbir Kaur,
Department of Computer Science
Gargi Girl School, Meerut (U.P),India

Dr. Gurusharan kaur
Department of Mathematics,
Career College, Bhopal(M.P),462023, India

Abstract:

 "Data Science and Machine Learning" has been created with the goal to provide beginners seeking to learn about data science, data enthusiasts, and experienced data professionals with a deep understanding of data science application development using open-source programming from start to finish, data science, software development, and open-source based embedded hardware, algorithms, decision engines for data science applications.

Data science and machine learning are the key technologies when it comes to the processes and products with automatic learning and optimization to be used in the automotive industry of the future, it defines the term "optimizing analytics" and illustrates the role of automatic optimization as a key technology in combination with data analytics. It is very important to analyze this data in order to extract some useful information and to develop an algorithm based on this analysis. This can be achieved through data mining and machine learning. Machine learning is an integral part of artificial intelligence, which is used to design algorithms based on the data trends and historical relationships between data. Machine learning is used in various fields such as bioinformatics, intrusion detection, Information retrieval, game playing, marketing, malware detection, image deconvolution.

The automotive industry on the basis of the major sub processes in the automotive value chain (development, procurement; logistics, production, marketing, sales and after-sales, connected customer). Since the industry is just starting to explore the broad range of potential uses for these technologies, visionary application

Keywords:
Data Science, Big Data, Machine Learning, Automatic Optimization, Optimizing analytics, Automotive Industry.

Research and Methodology For present study Data Manipulation.

Data Analysis with Statisticsand Machine Learning. Data Communication with Information Visualization. Data at Scale Working with Big Data. It a leap forward from computer science, statistics, and other emerging applications in the industry. Machine learning can produce accurate results and analysis by developing efficient and fast algorithms and data-driven models for real-time processing of this data. Data Science Methodology indicates the routine for finding solutions to a specific problem. This is a cyclic process that undergoes a critic behaviour guiding business analysts and data scientists to act accordingly

The most used methods are Regression, Clustering, Visualization, Decision Trees/Rules, and Random Forests.

Finding- Machine Learning. At its core, data science is a field of study that aims to use a scientific approach to extract meaning and insights from data. Machine learning, on the other hand, refers to a group of techniques used by data scientists that allow computers to learn from data.

The practice of data science requires a unique combination of skills and experience. A skilled data scientist is fluent in **programming languages** like **R and Python**, has knowledge of statistical methods, an understanding of database architecture and the experience to apply these skills to real-world problems. A **masters in data science** may build upon existing knowledge to ensure that you are best prepared for a long career in this ever-growing field.
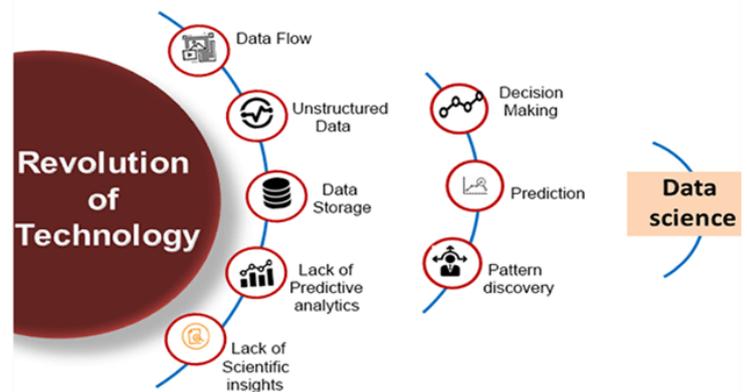
Introduction: -Data Science is about data gathering, analysis and decision-making.

Data Science is about finding patterns in data, through analysis, and make future predictions. It is used in many industries in the world today, e.g. banking, consultancy, healthcare, and manufacturing.

By using Data Science, companies are able to make:

- Better decisions
- Predictive analysis
- Pattern discoveries

Data Science is the area of study which involves extracting insights from vast amounts of data by the use of various scientific methods, algorithms, and processes. It helps you to discover hidden patterns from the raw data. The term Data Science has emerged because of the evolution of mathematical statistics, data analysis, and big data.

Data Science is an interdisciplinary field that allows you to extract knowledge from structured or unstructured data. Data science enables you to translate a business problem into a research project and then translate it back into a practical solution.

Some years ago, data was less and mostly available in a structured form, which could be easily stored in excel sheets, and processed using BI tools.

But in today's world, data is becoming so vast, i.e., approximately **2.5 quintals bytes** of data is generating on every day, which led to data explosion. It is estimated as per researches, that by 2020, 1.7 MB of data will be created at every single second, by a single person on earth. Every Company requires data to work, grow, and improve their businesses.

Now, handling of such huge amount of data is a challenging task for every organization. So to handle, process, and analysis of this, we required some complex, powerful, and efficient algorithms and technology, and that

technology came into existence as data Science. Following are some main reasons for using data science technology:

- o With the help of data science technology, we can convert the massive amount of raw and unstructured data into meaningful insights.

- o Data science technology is opting by various companies, whether it is a big brand or a start-up. Google, Amazon, Netflix, etc, which handle the huge amount of data, are using data science algorithms for better customer experience.

- o Data science is working for automating transportation such as creating a self-driving car, which is the future of transportation.

- o Data science can help in different predictions such as various survey, elections, flight ticket confirmation, etc.

Big data is a blanket term for any collection of data sets so large or complex that it becomes difficult to process them using traditional data management techniques such as, for example, the RDBMS (relational database management systems). The widely adopted RDBMS has long been regarded as a one-size-fits-all solution, but the demands of handling big data have shown otherwise. Data science involves using methods to analyze massive amounts of data and extract the knowledge it contains. You can think of the relationship between big data and data science as being like the relationship between crude oil and an oil refinery. Data science and big data evolved from statistics and traditional data management but are now considered to be distinct disciplines.

To process the massive amounts of data we need more effective algorithms. This is made possible by the Application of Data Analytics. Data Analytics is the application of structured statistical and mathematical techniques on collected data in order to detect underlying patterns as well as make predictions.

Benefits of Data Analytics

The domain of Data Analytics has been embraced by many industries for the outstanding benefits it offers. Data Analytics is a boon to modern-day businesses. Data Analytics helps businesses in making smarter decisions. Data Analytics improves efficiency and controls risks. Data Analytics also results in cost cuttings.

Pay scale of Data Analytics

The domain of Data Analytics is observed to offer one of the highest-paid job roles in the country. Recent industry studies have revealed that median CTC for Data Analytics professionals in India stands at Rs. 12 lakh/annum.

Job Positions of Data Analytics

There is a wide range of job roles offered in this domain. Professionals are seeking data analytics training to fit into the below set of job roles.

The following are the major responsibilities of a data analyst.

- Data Architect
- Applications Architect
- Infrastructure Architect
- Enterprise Architect
- Data Scientist
- Data Analyst
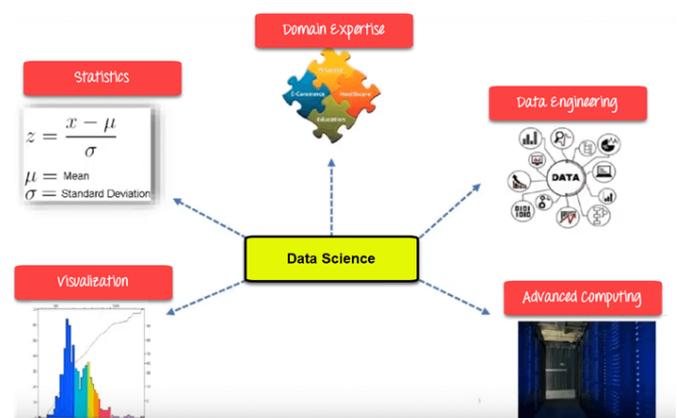
2. Data Science Components

Statistics:

Statistics is the most critical unit of Data Science basics. It is the method or science of collecting and analysing numerical data in large quantities to get useful insights.

Visualization:

Visualization technique helps you to access huge amounts

of data in easy to understand and digestible visuals.

Machine Learning:

Machine Learning explores the building and study of algorithms which learn to make predictions about unforeseen/future data.

Deep Learning:

Deep Learning method is new machine learning research where the algorithm selects the analysis model
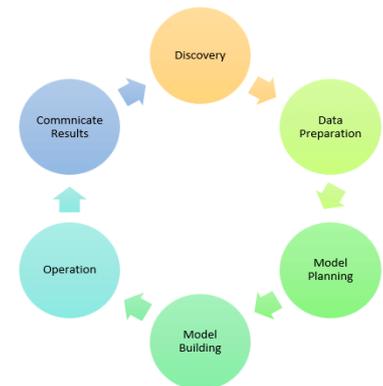
3.Data Science Process

1. Discovery:

Discovery step involves acquiring data from all the identified internal & external sources which helps you to answer the business question.

The data can be:

- Logs from webservers
- Data gathered from social media
- Census datasets
- Data streamed from online sources using APIs

2. Preparation:

Data can have lots of inconsistencies like missing value, blank columns, incorrect data format which needs to be cleaned. You need to process, explore, and condition data before modelling. The cleaner your data, the better are your predictions.

3. Model Planning:

In this stage, you need to determine the method and technique to draw the relation between input variables. Planning for a model is performed by using different statistical formulas and visualization tools. SQL analysis services, R, and SAS/access are some of the tools used for this purpose.

4. Model Building:

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.

5. Operationalize:

In this stage, you deliver the final baselined model with reports, code, and technical documents. Model is deployed into a real-time production environment after thorough testing.

6. Communicate Results

In this stage, the key findings are communicated to all stakeholders. This helps you to decide if the results of the project are a success or a failure based on the inputs from the model.

4.Data science Jobs:

As per various surveys, data scientist job is becoming the most demanding Job of the 21st century due to increasing demands for data science. Some people also called it "the **hottest job title of the 21st century**". Data scientists are the experts who can use various statistical tools and machine learning algorithms to understand and analyze the data.

The average salary range for data scientist will be approximately **$95,000 to $ 165,000 per annum**, and as per different researches, about **11.5 millions** of job will be created by the year **2026**.

Types of Data Science Job

If you learn data science, then you get the opportunity to find the various exciting job roles in this domain. The main job roles are given below:

1. Data Scientist

2. Data Analyst

3. Machine learning expert

4. Data engineer

5. Data Architect

6. Data Administrator

7. Business Analyst

8. Business Intelligence Manager

5. Tools for DataScience

Following are some tools required for data science:

- ○ **Data Analysis tools:** R, Python, Statistics, SAS, Jupyter, R Studio, MATLAB, Excel, RapidMiner.
- ○ **Data Warehousing:** ETL, SQL, Hadoop, Informatica/Talend, AWS Redshift
- ○ **Data Visualization tools:** R, Jupyter, Tableau, Cognos.
- ○ **Machine learning tools:** Spark, Mahout, Azure ML studio.

6. Applications of Data Science

Now in this Data Science Tutorial, we will learn about Applications of Data Science:

Internet Search:

Google search use Data science technology to search a specific result within a fraction of a second

Recommendation Systems:

To create a recommendation system. Example, "suggested friends" on Facebook or suggested videos" on YouTube, everything is done with the help of Data Science.

Image & Speech Recognition:

Speech recognizes system like Siri, Google assistant, Alexa runs on the technique of Data science. Moreover, Facebook recognizes your friend when you upload a photo with them, with the help of Data Science.

Gaming world:

EA Sports, Sony, Nintendo, are using Data science technology. This enhances your gaming experience. Games are now developed using Machine Learning technique. It can update itself when you move to higher levels.

Online Price Comparison:

Price Runner, Junglee, Shopzilla work on the Data science mechanism. Here, data is fetched from the relevant websites using APIs.

7. Challenges of Data science Technology

- High variety of information & data is required for accurate analysis
- Not adequate data science talent pool available
- Management does not provide financial support for a data science team
- Unavailability of/difficult access to data
- Data Science results not effectively used by business decision makers
- Explaining data science to others is difficult
- Privacy issues
- Lack of significant domain expert
- If an organization is very small, they can't have a Data Science team

8. Machine learning in Data Science

To become a data scientist, one should also be aware of machine learning and its algorithms, as in data science, there are various machine learning algorithms which are broadly being used. Following are the name of some machine learning algorithms used in data science:

- Regression
- Decision tree
- Clustering
- Principal component analysis
- Support vector machines
- Naive Bayes
- Artificial neural network
- Apriori

1. Linear Regression Algorithm: Linear regression is the most popular machine learning algorithm based on supervised learning. This algorithm work on regression, which is a method of modeling target values based on independent variables. It represents the form of the linear equation, which has a relationship between the set of inputs and predictive output. This algorithm is mostly used in forecasting and predictions. Since it shows the linear relationship between input and output variable

2. Decision Tree: Decision Tree algorithm is another machine learning algorithm, which belongs to the supervised learning algorithm. This is one of the most popular machine learning algorithms. It can be used for both classification and regression problems.

In the decision tree algorithm, we can solve the problem, by using tree representation in which, each node represents a feature, each branch represents a decision, and each leaf represents the outcome.

3. K-Means Clustering: K-means clustering is one of the most popular algorithms of machine learning, which belongs to the unsupervised learning algorithm. It solves the clustering problem.

If we are given a data set of items, with certain features and values, and we need to categorize those set of items into groups, so such type of problems can be solved using k-means clustering algorithm.

K-means clustering algorithm aims at minimizing an objective function, which known as squared error function, and it is given as:

$$J(V) = \sum_{i=1}^{c} \sum_{j=1}^{c_i} \left(\left\| x_i - v_j \right\| \right)^2$$

Methodology:-This Data Science Bootcamp program is ideal for all working professionals, covering job-critical topics like R, Python programming, machine learning algorithms, and NLP concepts, data visualization with Tableau in great detail via our interactive learning model with live sessions by global practitioners, practical labs, IBM Hackathons, and industry projects.

A Data Scientist requires expertise in several backgrounds:

- Machine Learning
- Statistics
- Programming (Python or R)
- Mathematics
- Databases

Finding:-

- Data Science is the area of study which involves extracting insights from vast amounts of data by the use of various scientific methods, algorithms, and processes.
- Statistics, Visualization, Deep Learning, Machine Learning, are important Data Science concepts.
- Data Science Process goes through Discovery, Data Preparation, Model Planning, Model Building, Operationalize and Communicate Results.
- Important Data Scientist job roles are: 1) Data Scientist 2) Data Engineer 3) Data Analyst 4) Statistician 5) Data Architect 6) Data Admin 7) Business Analyst 8) Data/Analytics Manager
- R, SQL, Python, SaS, are essential Data science tools
- The predictions of Business Intelligence is looking backward while for Data Science it is looking forward.

- Important applications of Data science are 1) Internet Search 2) Recommendation Systems 3) Image & Speech Recognition 4) Gaming world 5) Online Price Comparison.
- High variety of information & data is the biggest challenge of Data Science technology.

Role:

A Data Scientist is a professional who manages enormous amounts of data to come up with compelling business visions by using various tools, techniques, methodologies, algorithms, etc.

Languages:

R, SAS, Python, SQL, Hive, Matlab, Pig, Spark

Advantages of Data Science

- Data is the oil for today's world. With the right tools, technologies, algorithms, we can use data and convert it into a distinctive business advantage
- Data Science can help you to detect fraud using advanced machine learning algorithms
- It helps you to prevent any significant monetary losses
- Allows to build intelligence ability in machines
- You can perform sentiment analysis to gauge customer brand loyalty
- It enables you to take better and faster decisions
- Helps you to recommend the right product to the right customer to enhance your business

Reference

1. About Data Science | Data Science Association, www.datascienceassn.org. Retrieved 3 April 2020.
2. ASA Statement on the Role of Statistics in Data Science. AMSTATNEWS. American Statistical Association. 1 October 2015. Archived from the original on 20 June 2019. Retrieved 29 May 2019.
3. Bell, G.; Hey, T.; Szalay, A. (2009). "COMPUTER SCIENCE: Beyond the Data Deluge". Science. 323 (5919): 1297–1298. doi:10.1126/science.1170411. ISSN 0036-8075. PMID 19265007. S2CID 9743327.
4. Basic Example. benfry.com. Retrieved 3 April 2020.
5. DharVasant (1 December 2013). "Data science and prediction". Communications of the ACM. 56 (12): 64–73. doi:10.1145/2500499. S2CID 6107147.
6. Dhar, V. (2013). "Data science and prediction". Communications of the ACM. 56 (12): 64–73. doi:10.1145/2500499. S2CID 6107147. Archived from the original on 9 November 2014. Retrieved 2 September 2015.
7. Hayashi, Chikio (1 January 1998). "What is Data Science? Fundamental Concepts and a Heuristic Example". In Hayashi, Chikio; Yajima, Keiji; Bock, Hans-Hermann; Ohsumi, Noboru; Tanaka, Yutaka; Baba, Yasumasa (eds.). Data Science, Classification, and Related Methods. Studies in Classification, Data Analysis, and Knowledge Organization. Springer Japan. pp. 40–51. doi:10.1007/978-4-431-65950-1_3. ISBN 9784431702085.
8. Introduction: What Is Data Science? - Doing Data Science [Book]". www.oreilly.com. Retrieved 3 April 2020
9. Jeff Leek (12 December 2013). "The key word in "Data Science" is not Data, it is Science". Simply Statistics. Archived from the original on 2 January 2014. **Retrieved** 1 January 2014.
10. Nate Silve: What I need from statisticians - Statistics Views". www.statisticsviews.com. Retrieved 3 April 2020.
11. Posted by Vincent Granville on December 8, 2014 at 5:00pm; Blog, View. "Data science without statistics is possible, even desirable". www.datasciencecentral.com. Retrieved 3 April 2020.
12. Statistics is the least important part of data science « Statistical Modeling, Causal Inference, and Social Science. statmodeling.stat.columbia.edu. Retrieved 3 April 2020.

13. Tony Hey; Stewart Tansley; Kristin Michele Tolle (2009). The Fourth Paradigm: Data-intensive Scientific Discovery. Microsoft Research. ISBN 978-0-9825442-0-4. Archivedfrom the original on 20 March 2017. Retrieved 16 December 2016.

14. The three sexy skills of data geeks". m.e.driscoll: data utopian. Retrieved 3 April 2020.

15. What's the Difference Between Data Science and Statistics?". Priceonomics. Retrieved 3 April 2020.

16. Yau, Nathan (4 June 2009). "Rise of the Data Scientist". FlowingData. Retrieved 3 April2020.

.