Data Science : The emerging domain in IT

S. Sharma

Deshbandhu College, University of Delhi, Delhi

ABSTRACT

With the increase in advance data collection tools, the scope of inferential statistics and predictive analytics has increased to a great extent and the dependency on numbers for policy formation, business solutions for complex issues, the data-driven decision taken in cases of various emergencies has led to the development of the field data science.

Keywords: *Data collection, Data cleansing, machine language, predictive analysis*.

1. INTRODUCTION

Here we will present when and how this data is collected, then we will move on to, how we process this data to make it meaningful and ready to be analyzed, in between we will clarify the difference between analysis and analytics which is often misinterpreted as same terms, then we will take you through the sectors where data science is most often used and different roles involved in the entire spectrum of data science. We will also brief you about the tools used in these fields both for data collection and data analytics. We will give a short summary of the transition of data science and lastly we will move towards the machine learning part or modern data science techniques [1,2].

2. DATA COLLECTION

Before we move further we will differentiate between the term analysis and analytics. An analysis is when we use inferential statistics to find answers to questions of the past, whereas analytics is when we use the data collected and the techniques are known to predict future outcomes. So analysis is something which is done prior to analytics as you must know the past to predict the future. With this, we come to data collection, which has revolutionized the data science field as more data means more accurate predictions. The traditional approach to data collection was the method of surveys wherein a selected people from a group of people were selected and questioned and the acquired data was sent for processing. Then came the online surveys which got so hated people would fill then only if they are forced too. With the invention of social media and people providing their information to these platforms along with their likes and dislikes the scope of data collection increased but the technique to process such vast data was missing. Finally, Big data collection tools were brought into use which used high tech computers and even supercomputers to collect raw data from these platforms and process them into categories, etc to be used for analysis [3-5].

3. DATA CLEANSING

Data cleansing is a proper representation of data for analysts to work by fixing missing values issues, repeated entries, spelling errors, etc. Traditional data required proper data cleansing and then categorization but with big data techniques now we can directly cleanse it and categorize huge data, data in zeta bytes!

3.1. Tools required for data collection

A number of tools can be used for data collection and cleansing including programming languages which are quite efficient, amongst which python is the most used, user-friendly ease to understand the programming language. Codes used in the language are very general in nature and hence it’s easy to learn. Other software used in this field are SQL, Matlab, IBM spss. For big data mining, programming languages like JAVA, R, Scala, along with python are also used and a sophisticated software Hadoop makes data collection done easily [5-7].

3.2. Roles related to data collection in the data science field

- Data architect - A data architect is required to create, deploy, and maintain data for an organization.

- Data engineer- These are people who design software and programs to manage big data, collect and store them in required formats.

- Data administrator - He/ she manages the stored data ready to be used for analysis and keeps a check on the flow of data.

4. WHAT IS BUSINESS INTELLIGENCE?

After the data has been collected and cleaned analysts use various statistical techniques to form relations between various variables and analyze the trends in business activities and ultimately find solutions to the ongoing problems. These numbers and data have huge information hidden in them for an analyst to mine and hence they use different software with the pre-installed function for easy and efficient analysis and when the know about about the various trends in business or the way the areas where there is an increase in users it can invest strategically in those areas. This is the way business intelligence helps a business. A more classic example for ease of understanding is, Consider a group of people buying from 7 different stores of a company. Now the company wants to increase its sales to generate more profit. In order to do so, it needs to collect some data so it puts a digital rating where a user can rate his experience with the product. Now with just one extra data, a Business Intelligence analyst can provide a number of inferences of these customers. In recent times, there has been emerging role of artificial intelligence which has proved its roles in different sectors [8-9].

4.1 Customers are divided into groups of four-

- Fans - These are people who buy from only this store and are satisfied with the services. So fans are people with high loyalty as well as high satisfaction.

- Supporters - These are people who are loyal to the store which means they continue to buy products from the store nut are not satisfied with the services offered.

- Roamers - These are people who are satisfied with the services but shop in multiple stores. They don’t shop from a particular company.

- Alienated - The last among the group are people who are alienated, i.e., these people neither buy much nor receive much satisfaction from the product.

Now, when a Business intelligence analyst will provide you the numbers of these four groups you can easily invest in the right way and increase sales. If the number of supporters is high, then the company needs to work on the services offered, store management, product price, etc.

If the number if roamers are too high the company would issue loyalty cards encouraging people to buy more and these are the ways data science help business to grow and play a key role in the corporate world now.

4.2 WHY USE BUSINESS INTELLIGENCE TOOLS?

For one thing, information revelation, which used to be restricted to the ability of cutting edge investigation pros, is currently something everybody can do utilizing these instruments. What's more, not just that, these devices give you the bits of knowledge you have to accomplish things like development, resolve gives that are critical, gather every one of your information in one spot, conjecture future results thus substantially more [7,9,10].

The top Business Intelligence apparatuses that can assist you with settling on the correct choice.

- **SAP Business Intelligence**

Business Intelligence offers a few progressed investigation arrangements including ongoing BI prescient examination, AI, and arranging and investigation. The Business Intelligence stage specifically, offers detailing and examination, information representation and investigation applications, office incorporation and portable investigation. SAP is a vigorous programming proposed for all jobs (IT, end uses and the executives) and offers huge amounts of functionalities in a single stage.

- **MicroStrategy**

MicroStrategy is a business knowledge instrument that offers incredible (and fast) dashboarding and information examination which help screen pattern, perceive new chances, improve profitability and the sky is the limit from there. It tends to be gotten to from your work area or by means of portable.

- **Datapine**

Datapine is an across the board BI stage that encourages the perplexing procedure of information examination in any event, for non-specialized clients. Because of a far reaching self-administration examination approach, datapine's answer empowers information experts and business clients the same to effectively coordinate various information sources, perform propelled information investigation, fabricate intelligent business dashboards and produce significant business bits of knowledge.

- **Yellowfin BI**

Yellowfin BI is a business insight instrument and 'start to finish' examination stage that joins representation, AI, and coordinated effort. You can likewise effectively channel through huge amounts of information with natural too open up dashboards pretty much anyplace.

- **QlikSense**

QlikSense is a result of Qlik, an organization additionally known for another business knowledge instrument called QlikView. The UI of QlikSense is advanced for touchscreen, which makes it a well known bi apparatus. A major distinction with QlikView is the element Storytelling. Clients add their experience to the information and by utilizing depictions and features settling on the correct investigation and choices has become significantly simpler and better.

- **Microsoft Power BI**

Microsoft Power BI is an electronic business examination apparatus suite which exceeds expectations in information representation. It permits clients to distinguish patterns progressively and has fresh out of the box new connectors that permit you to up your game in crusades. This product additionally permits clients to coordinate their applications and convey reports and constant dashboards.

5. MACHINE LEARNING (ML)

Machine learning is a part of science that makes the computer act without being programmed. Machine learning is a part of data science that deals with artificial intelligence. It in a way helps to make the computers learn from the data. Machine learning is not the same, it differs from the past. Both artificial intelligence (AI) and machine learning are often used interchangeably. Artificial intelligence is also a part of science that makes the computer act according to the human tasks. Most of the machines learning algorithms are used in the advancement of artificial intelligence. These algorithms are used in various applications like email, filtering etc. we go through many activities that is powered by machine learning in our day to day life like the recommendation on Netflix, YouTube, etc and the search engines like Google. Even voice assistants like Google assistant and Siri [10-13].

What is machine learning? ML is a part of AI that makes the computers to learn by itself without being explicitly programmed. An ML algorithm uses the statistical data to compute the output. The knowledge needed to create an efficient machine learning system are

- Whole model

- Scalability
- Advanced algorithms along with the basics
- Capability of preparing data

6. METHODS OF MACHINE LEARNING

These methods drive different ways to train the ML algorithms. We need to look at the kind of data it accepts to know the advantages and disadvantages of each method. There are two types of data .i.e. labeled and unlabeled data.

- Labeled data processes both input and output data with a readable pattern but it requires lot of human work.
- Unlabeled data does not have any parameters to process. It is processed in a machine readable pattern.

In machine learning, there are three methods

6.1.Supervised learning

Supervised learning is the easiest way of machine learning. It is considered as a paradigm of machine learning. It is basically a task driven process. It firstly needs example with labeled data set to work with. Then feed the algorithm with one example dataset at time and see if the prediction made by the system is correct or not. Follow this process, in the mean time the system starts to find the relationship between the examples and the labels. After the data being processed, the algorithm gains an idea of how the system works and derives the relation between the input and the output. The common applications are

- Popularity of the advertisements – selecting the required ads is based on this algorithm. The ads those are present in Google while browsing is due to this algorithm.

- Face recognition – the Facebook uses this algorithm to recognize your face. If you have a system that takes photos, then works based on the supervised learning.

6.2.Unsupervised learning

Unsupervised learning is the opposite of supervised learning. It uses unlabeled data. It is a data driven process. As it doesn’t have labeled data, the system process hidden structures. These hidden structures make unsupervised learning versatile. It can get adapted to any data by changing its structure. The system where you can see unsupervised learning are

- Recommendation system – the recommendation provided by Netflix, YouTube and many other websites are based on unsupervised learning.

6.3.Reinforcement learning

Reinforcement learning learns from errors the same as where humans learn data from errors that they commonly do. It uses a trial and error method. It doesn’t use both labeled and unlabeled data. It is behavior-driven learning which does lot of mistakes initially and learns from the mistakes. Over a period of time, the system corrects the mistake and makes fewer mistakes than it does before [10-13].

7. REFRENCES:

1. Nadikattu, Rahul Reddy, Research on Data Science, Data Analytics and Big Data (April 17, 2020). INTERNATIONAL JOURNAL OF ENGINEERING, SCIENCE AND - Volume 9, Issue 5, May 2020 Pages: 99-105.. Available at SSRN: https://ssrn.com/abstract=3622844 or http://dx.doi.org/10.2139/ssrn.3622844
2. K. Terao, Machine Learning synthetic data , scanning probe data , and reciprocal space data on quantum materials. (2019).

3. V.Setlur and M. Tory, Exploring Synergies between Visual Analytical Flow and Language Pragmatics. *AAAI Spring Symposia*. (2017).
4. L.A. Enneking, The use of data collection activities in the secondary mathematics classroom. (2008).
5. P.W .Group and G. Garrett, Data Engineering Project (Educating for the Future PhUSE Working Group). (2019).
6. Nadikattu, Rahul Reddy, Data Warehouse Architecture – Leading the Next Generation Data Science (September 11, 2019). Rahul Reddy Nadikattu "Data Warehouse Architecture – Leading the next generation Data Science" International Journal of Computer Trends and Technology 67.9 (2019):78-80.. Available at SSRN: https://ssrn.com/abstract=3622840 or http://dx.doi.org/10.2139/ssrn.3622840
7. J.M .Fernández and A. Valencia, XML Databases, are Ready for Bioinformatics? *Spanish Bioinformatics Conference*. (2004).
8. M.Cox, S.F. Austin and A.B.Gresham, The Role of Customer Service in Small Business Strategic Planning. (1997).
9. P. Khatri, The Emergence of AI through Machine learning and Data Science. *The Journal of Innovations, 14*. (2019).
10. D.C.Desai, C.Dhanasekaran, A.Narayanapur and V.Joshi, (Social Media and Multimedia Data Analytics through Machine Learning.(2017).
11. MODELLING A NEW WORKFLOW BASED ON EMOTIONAL ANALYSIS OF FLOOR-PLANS USING MACHINE LEARNING ALGORITHAMS AND SEMIOTICS. (2020).
12. B.Geluvaraj, P.M .Satwik and T.A. Kumar, The Future of Cybersecurity: Major Role of Artificial Intelligence, Machine Learning, and Deep Learning in Cyberspace. (2019).
13. S.K. Vishwakarma, A Machine Learning Approach for Prediction Analysis in Data Mining. (2020).