AD3491- FUNDAMENTALS OF DATA SCIENCE AND ANALYTICS

Unit I

INTRODUCTIONTO DATA SCIENCE

Topic: Need for data science – benefits and uses – facets of data

By

Dr.M.Gomathy Nayagam

Associate Professor/CSBS

Ramco Institute ofTechnology, Rajapalayam

Course Objectives

- To understand the techniques and processes of data science
- To apply descriptive data analytics
- To visualize data for various applications
- To understand inferential data analytics
- To analysis and build predictive models from data

Pre-Requisites

- GE3151- Problem Solving and Python Programming
- MA3251 - Statistics and Numerical Methods
- Probability

Course Outcome

- Elucidate the pipeline of data analytics process for any data science application.
- Describe, Visualize and examine the data of real world problems using descriptive analytics techniques
- Perform statistical inferences from data
- Analyze the variance in the data for any real world data science problems.
- Build models for predictive analytics

UNIT I INTRODUCTION TO DATA SCIENCE

Need for data science – benefits and uses – facets of data – data science process – setting theresearch goal – retrieving data – cleansing, integrating, and transforming data – exploratory dataanalysis – build the models – presenting and building applications.

UNIT II DESCRIPTIVE ANALYTICS

Frequency distributions – Outliers –interpreting distributions – graphs – averages – describingvariability – interquartile range – variability for qualitative and ranked data - Normal distributions – zscores –correlation – scatter plots – regression – regression line – least squares regression line –standard error of estimate – interpretation of r2 – multiple regression equations – regression towardthe mean.

UNIT III INFERENTIAL STATISTICS

Populations – samples – random sampling – Sampling distribution- standard error of the mean -Hypothesis testing – z-test – z-test procedure –decision rule – calculations – decisions –interpretations - one-tailed and two-tailed tests – Estimation – point estimate – confidence interval –level of confidence – effect of sample size.

UNIT IV ANALYSIS OF VARIANCE

t-test for one sample – sampling distribution of t – t-test procedure – t-test for two independentsamples – p-value – statistical significance – t-test for two related samples. F-test – ANOVA – Twofactor experiments – three f-tests – two-factor ANOVA –Introduction to chi-square tests.

UNIT V PREDICTIVE ANALYTICS

Linear least squares – implementation – goodness of fit – testing a linear model – weightedresampling. Regression using StatsModels – multiple regression – nonlinear relationships – logisticregression – estimating parameters – Time series analysis – moving averages – missing values – serial correlation – autocorrelation. Introduction to survival analysis.

Need for data science – benefits and uses – facets of data

What is Data Science?

- Data Science is a combination of multiple disciplines that uses
  - Statistics
  - Data analysis,
  - Machine Learning
- To analyze data and to extract knowledge and insights from it.
- Data Science is about data gathering, analysis and decision-making
- Data Science is about finding patterns in data, through analysis, and make future predictions
- By using Data Science, companies are able to make:
  - Better decisions (should we choose A or B)
  - Predictive analysis (what will happen next?)
  - Pattern discoveries (find pattern, or maybe hidden information in the data)

- Data science involves using methods to analyze massive amounts of data and extract the knowledge it contains.
- Data science and big data evolved from statistics and traditional data management.
- But are now considered to be distinct disciplines.
- *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
- Example for Traditional Data Management Techniques as RDBMS
- *Data science involves using* methods to analyze massive amounts of data and extract the knowledge it contains.

Definition of Data Science

- Data Science is a field or domain which includes and involves working with a huge amount of data and uses it for building predictive, prescriptive and prescriptive analytical models.
- It’s about digging, capturing, (building the model) analyzing (validating the model) and utilizing the data (deploying the best model).

Definition of Big Data

- It is huge, large or voluminous data, information or the relevant statistics acquired by the large organizations and ventures.
- Many software and data storage created and prepared as it is difficult to compute the big data manually.

Data Science vs Big Data

Data Science	Big Data
Data Science is an area/Domain	Big Data is a technique to collect, maintain and process the huge information.
It is about collection, processing, analyzing and utilizing of data into various operations. It is more conceptual	It is about extracting the vital and valuable information from huge amount of the data.
It is a field of study just like the Computer Science, Applied Statistics or Applied Mathematics, Data Base Management System.	It is a technique of tracking and discovering of trends of complex data sets.
The goal is to build data-dominant products for a venture	The goal is to make data more vital and usable i.e. by extracting only important information from the huge data within existing traditional aspects.
Tools mainly used in Data Science includes SAS, R, Python, etc	Tools mostly used in Big Data includes Hadoop, Spark, Flink, etc.

Retrieved from: https://www.geeksforgeeks.org/difference-between-big-data-and-data-science/

Data Science vs Big Data

Data Science	Big Data
It is a sub set of Data Science as mining activities which is in a pipeline of the Data science.	It is a super set of Big Data as data science consists of Data scrapping, cleaning, visualization, statistics and many more techniques.
It is mainly used for scientific purposes	It is mainly used for business purposes and customer satisfaction
Uses mathematics and statistics extensively along with programming skills to develop a model to test the hypothesis and make decisions in the business	Used by businesses to track their presence in the market which helps them develop agility and gain a competitive advantage over others
Internet search, digital advertisements, textto-speech recognition, risk detection, and other activities	Telecommunication, financial service, health and sports, research and development, and security and law enforcement

Retrieved from: https://www.geeksforgeeks.org/difference-between-big-data-and-data-science/

Characteristics of Big Data

- The characteristics of big data are explained with ‘Five V’ approach.
- If it satisfy the five characteristics then it is known as Big Data.
- **Volume**
  - How much data is there?
  - To determine the value of data, size of data plays a very crucial role.
  - If the volume of data is very large then it is actually considered as a ‘Big Data’.
- **Variety**
  - How diverse are different types of data?
  - It refers to nature of data that is structured, semi-structured and unstructured data.
  - It also refers to heterogeneous sources.

- Velocity
  - At what speed is new data generated?
  - Velocity refers to the high speed of accumulation of data.
  - In Big Data velocity data flows in from sources like machines, networks, social media, mobile phones etc.
- Veracity
  - How accurate is the data?
  - It refers to inconsistencies and uncertainty in data
  - That is data which is available can sometimes get messy and quality and accuracy are difficult to control.

- Value
  - How effectively to transform a tsunami of data into business?
  - Data in itself is of no use or importance
  - But it needs to be converted into something valuable to extract Information.

Challenges of Big Data

- Data Capture
- Curation
- Storage
- Search
- Sharing
- Transfer
- Visualization

- Data Capture
  - Data capture, or electronic data capture, is the process of extracting information from a document and converting it into data readable by a computer
- Curation
  - Data curation includes "all the processes needed for principled and controlled data creation, maintenance, and management, together with the capacity to add value to data".
- Storage
  - Data storage refers to magnetic, optical or mechanical media that records and preserves digital information for ongoing or future operations.

- Search
  - Searching is designed to check for an element/item or retrieve an element from any data storage
- Sharing
  - Data sharing is the practice of making data used for scholarly research available to other investigators
- Transfer
  - Data transfer refers to the secure exchange of large files between systems or organizations
- Visualization
  - Data visualization is the graphical representation of information and data.

Areas of Application of Data Science

- Fraud and Risk Detection
- Healthcare
- Internet Search
- Targeted Advertising
- Website Recommendations
- Advanced Image Recognition
- Speech Recognition
- Airline Route Planning
- Gaming
- Augmented Reality

Where is Data Science Needed?

- Data Science is used in many industries in the world today, e.g. banking, consultancy, healthcare, and manufacturing.
- Examples of where Data Science is needed:
  - For route planning: To discover the best routes to ship
  - To foresee delays for flight/ship/train etc. (through predictive analysis)
  - To create promotional offers
  - To find the best suited time to deliver goods
  - To forecast the next years revenue for a company
  - To analyze health benefit of training
  - To predict who will win elections

- Data Science can be applied in nearly every part of a business where data is available.
- Examples are:
  - Consumer goods
  - Stock markets
  - Industry
  - Politics
  - Logistic companies
  - E-commerce

Benefits and uses/advantage of Data Science

- Commercial Companies in all business wish to
  - analyses and gain insights into their customers, processes, staff, completion, and products.
  - Many companies use data science to offer customers a
    - better user experience,
    - cross-sell, up-sell, and personalize their offerings.
- Human resource professionals use
  - people analytics and text mining to screen candidates
  - monitor the mood of employees
  - study informal networks among coworkers.
- Financial institutions use data science to
  - predict stock markets
  - determine the risk of lending money
  - learn how to attract new clients for their services.

- Many governmental organizations not only rely on internal data scientists to
  - discover valuable information, but also share their data with the public. You can use this data to gain insights or build data-driven applications.
- Nongovernmental organizations (NGOs)
  - can use it as a source for get funding.
  - Many data scientists devote part of their time to helping NGOs, because NGOs often lack the resources to collect data and employ data scientists.
- Universities use data science in their research but also to
  - Enhance the study experience of their students.
  - The rise of massive open online courses (MOOC) produces a lot of data, which allows universities to study how this type of learning can complement traditional classes.

- Data accumulation from multiple sources, including the Internet, social media platforms, online shopping sites, company databases, external third-party sources, etc.
- Real-time forecasting and monitoring of business as well as the market.
- Identify crucial points hidden within large datasets to influence business decisions.
- Promptly mitigate risks by optimizing complex decisions for unforeseen events and potential threats.

References

- David Cielen, Arno D. B. Meysman, and Mohamed Ali, “Introducing Data Science”, Manning Publications, 2016.