NEW EDITION

DATA SCIENCE

Dr. Rajeswari Purushothaman

This Page Intentionally Left Blank

Data Science

Authors:

Dr. Rajeswari Purushothaman

@ All rights reserved with the Publisher / Authors

First Edition May 2021

ISBN 978-81-947070-1-1

ISBN: 978-81-947070-1-1

Pages: 282

Price: 300/-

Publisher & Imprint:

MAGESTIC TECHNOLOGY SOLUTIONS PVT. LTD.

We Accomplish Your Vision.

544, Anna Street, Kathirvedu

Chennai, Tamil Nadu, India.

Email: info@magesticts.com

This Page Intentionally Left Blank

Copyright Disclaimer

Copyright © 2021 by Magestic Technology Solutions (P) Ltd

All rights reserved. No part of this publication may be reproduced, distributed, or transmitted in any form or by any means, including photocopying, recording, or other electronic or mechanical methods, without the prior written permission of the publisher, except in the case of brief quotations embodied in critical reviews and specific other non-commercial uses permitted by copyright law. For permission requests, write to the publisher, addressed "Attention: Permissions Coordinator," at the address below.

Publisher's Address:

Magestic Technology Solutions (P) Ltd
544, Anna Street, Kathirvedu
Chennai, Tamil Nadu, India.
Email: info@magesticts.com

This Page Intentionally Left Blank

Title Verso

Title: Data Science

Author's Name:

Dr. Rajeswari Purushothaman

Published By:

Magestic Technology Solutions (P) Ltd

Publisher's Address:

544, Anna Street, Kathirvedu
Chennai 600 066. Tamil Nadu, India.

Printer's Details:

Magestic Technology Solutions (P) Ltd

Edition Details: I

ISBN: 978-81-947070-1-1

Copyright @ Magestic Technology Solutions (P) Ltd

This Page Intentionally Left Blank

Preface

In order to discover patterns in large quantities of organised and unstructured big data, data science is used by companies to analyse the data. Increasing efficiency and cost management and the identification of new market possibilities enable businesses to gain an edge in the marketplace.

In order to get a suggestion from a personal assistant like Alexa or Siri, data science is required. Operating a self-driving vehicle, utilising a search engine that returns relevant results, or interacting with a chatbot for customer support all qualifies as technological advancements. All of these are real-world examples of data science in action.

When it comes to data science, the discipline involves massive mining datasets of raw data, both organised and unstructured, to discover patterns and derive meaningful knowledge from them. Because data science is a multidisciplinary subject, its underlying principles include statistics, inference, computer science, predictive analytics, machine learning algorithm development, and new tools for extracting insights from large amounts of data.

Please start with the life cycle of data science to better understand it and enhance data science project management. The first step of the data science pipeline process is data collection, which includes collecting data, extracting data where necessary, and putting it into the system. Data management is the following step, and it involves data warehousing, data cleaning, data processing, data staging, and data architecture, among other things.

Data processing is the next step and is considered to be one of the foundations of data science. Data scientists and data engineers are distinguishable from one another when it comes to data exploration and processing. At this stage, data mining, data classification and clustering, data

modelling, and summarising insights gained from the data are all performed, which are the procedures that result from inadequate data.

Following that comes data analysis, which is an equally important step. During their time here, data scientists carry out exploratory and confirmation work and regression analysis, forecasting analysis, qualitative analysis, and text mining. When data science is done correctly, this step demonstrates why there is no cookie-cutter data science.

The data scientist discusses his or her findings in the final step. Specifically, data visualisation, data reporting, and the use of different business intelligence tools are all part of helping companies, policymakers, and others make more informed decisions.

Data preparation and analysis are the essential data science abilities, yet data preparation alone takes up 60 to 70% of a data scientist's work on average, according to industry estimates. Data is seldom produced in a format that is corrected, organised, and free of noise. In this phase, the data is converted and prepared for future usage by the application.

This stage of the process includes the transformation and sampling of data, the verification of both characteristics and observations, and statistical methods to eliminate noise from the data. This phase also reveals if the various characteristics of the data set are independent of one another and whether there are any missing values in the data set, among other things.

By 2023, there will be about 40 zettabytes of data—or 40 trillion gigabytes—on the planet. The quantity of data that exists is increasing at an exponential rate. According to IBM and SINTEF, about 90 percent of this massive quantity of data is produced in the most recent two years at any given point in time.

It is estimated that every day, internet users produce about 2.5 quintillion bytes of data. Approximately 146,880 gigabytes of data will be generated per day by 2025, and by 2025, this will amount to 165 zettabytes of data per year for every individual on Earth.

This implies a tremendous amount of work to be done in data science—a great deal more to be discovered. Said, simple data analysis may be used to analyse information from a single source or a small quantity of data. In order to comprehend extensive data and data from many sources in a meaningful manner, however, the use of data science techniques is essential. This concept is shown by a look at some of the particular data science applications in business, which serve as an excellent introduction to data science.

This book focuses on the fundamentals of Data Science, which will be beneficial for all aspiring Science and Engineering graduates to understand the basic levels and enrich their knowledge with the readiness to fine-tune themselves to the next level of development.

- **Dr. Rajeswari Purushothaman**

This Page Intentionally Left Blank

Table of Contents

S.No	Title	Page No
1	UNIT 1 – INTRODUCTION TO CORE CONCEPTS AND TECHNOLOGIES	1
	1.0 INTRODUCTION	3
	1.2 DATA SCIENCE TERMINOLOGY	10
	1.3 DATA SCIENCE PROCESS	18
	1.4 DATA SCIENCE TOOLKIT	24
	1.5 TYPES OF DATA IN DATA SCIENCE	30
	1.6 EXAMPLE APPLICATIONS IN DATA SCIENCE	35
2	UNIT 2 – DATA COLLECTION AND MANAGEMENT	44
	2.1 INTRODUCTION	46
	2.2 SOURCES OF DATA	50
	2.3 DATA COLLECTION AND APIs	56
	2.4 EXPLORING AND FIXING DATA	66
	2.5 DATA STORAGE AND MANAGEMENT	76
	2.6 USING MULTIPLE DATA SOURCES	89
3	UNIT 3 – DATA ANALYSIS	94
	3.1 INTRODUCTION	96
	3.2 DATA ANALYSIS: TERMINOLOGY AND CONCEPTS	100
	3.3 DATA ANALYSIS: INTRODUCTION TO STATISTICS	107
	3.4 DATA ANALYSIS: CENTRAL TENDENCIES AND DISTRIBUTIONS	112
	3.5 DATA ANALYSIS: VARIANCE	125
	3.6 DATA ANALYSIS: DISTRIBUTION PROPERTIES AND ARITHMETIC	129
	3.7 DATA ANALYSIS: SAMPLES/CLT	135
	3.8 DATA ANALYSIS: BASIC MACHINE LEARNING ALGORITHMS	143

	3.9 DATA ANALYSIS: LINEAR REGRESSION	154
	3.10 DATA ANALYSIS: SVM	160
	3.11 DATA ANALYSIS: NAÏVE BAYES	167
4	UNIT 4 – DATA VISUALISATION	175
	4.1 INTRODUCTION	177
	4.2 TYPES OF DATA VISUALISATION TOOLS	179
	4.3 DATA FOR VISUALISATION: DATATYPES	185
	4.4 DATA FOR VISUALISATION: DATA ENCODINGS	190
	4.5 DATA FOR VISUALISATION: RETINAL VARIABLES	206
	4.6 DATA FOR VISUALISATION: MAPPING VARIABLES TO ENCODING	209
	4.7 DATA FOR VISUALISATION: VISUAL ENCODINGS	213
5	UNIT 5 – APPLICATIONS OF DATA SCIENCE	223
	5.1 TECHNOLOGIES FOR VISUALISATION	225
	5.2 BOKEH (PYTHON)	230
	5.3 RECENT TRENDS IN VARIOUS DATA COLLECTION AND ANALYSIS TECHNIQUES	236
	5.4 VARIOUS VISUALISATION TECHNIQUES	248
	5.5 APPLICATIONS IN DATA SCIENCE	256
6	Bibliography	267

The top graphic features the words "DATA SCIENCE" in large, bold, black capital letters. The background is a grayscale collage of digital and technical elements, including circuit board traces, binary code, and faint text like "PATH" and "WPINC".

DATA SCIENCE

UNIT - I

INTRODUCTION TO CORE CONCEPTS AND TECHNOLOGIES

This
Page
Intentionally
Left
Blank

UNIT -I

INTRODUCTION TO CORE CONCEPTS AND TECHNOLOGIES

1.0 INTRODUCTION

1.1.1 What is Data Science?

Data science is an area of study that deals with vast volumes of data and uses cutting-edge approaches and techniques to discover secret patterns, extract useful information, and optimize performance. Advanced machine learning algorithms are used in data science to build predictive models.

The data under examination can originate from a diversity of resources and be shown in various formats.

1.1.2 Why Data Science is considered necessary?

- ◆ Data science, also known as data-driven science, allows for more accurate decision-making, statistical analysis, and trend detection.
- ◆ Analyze the findings in an exploratory manner.
- ◆ Various algorithms are used to model the outcomes.
- ◆ To communicate and view the results, charts, analysis tools, or other multimedia resources can be used.

In reality, it is now assisting the airline industry in predicting flight delays, which helps both airlines and travellers avoid discomfort. Airlines may improve their operations in a variety of ways with the aid of data science, including:

- ◆ Plan our routes and determine whether we want to fly straight or link.
- ◆ To forecast flight delays, build predictive analytics models.
- ◆ Provide customized discount packages focused on the booking habits of the clients.

1.1.3 Why We Need Data Science?

We have enormous volumes of data on many facets of society and, at the same time, a lot of cheap processing resources. Most people are aware, including shopping, chatting, reading the papers, listening to songs, looking for facts, and sharing opinions.

People might be unaware that, as the online data processing boom, the "datafication" of digital behaviour has begun as well. Finance, the drug sector, pharmaceuticals, bioinformatics, social care, government, schooling, shopping, and the list goes on are all examples of industries that use data from the Internet. In almost all fields and markets, data is becoming increasingly influential.

Data are the foundations of data items and are primarily done in real-time, including Amazon recommendation services, Facebook friend suggestions, music and movie recommendations, etc. Credit scores, trading algorithms, and simulations are all examples of this in finance.

We see the start of a vast, culturally influenced feedback loop in which our actions influence the product and influence our actions. This is made

possible by technology, including facilities for large-scale data collection, improved memory and bandwidth, and a societal embrace of technology as part of our everyday lives.

Datafication is the process of transforming information into useful information. Datafication is the act of "converting all facets of existence into data." LinkedIn maintains a database of specialist networks."

Datafication is a fascinating phenomenon that has prompted us to understand its significance in people's plans to share their personal information. We are being datafied, or rather, our behaviour is being datafied. We plan to be datafied when we "like" anyone or someone online, but on the Web, we are datafied unwittingly, or at most passively, by cookies that they might or may not be conscious of. Moreover, we are being datafied accidentally as we wander around in a shop or even on the highway, thanks to sensors, cameras, and Google glasses. This intentionality continuum includes everything from us gleefully participating in a social networking experiment

However, it is all about the data. Our motives can be varied, but our outcomes are not. They conclude their description in the paragraph with a line that encapsulates their viewpoint: We can convert the intent of items and convert knowledge into different types of meaning until we have satisfied them. The "value" refers to things like improved productivity by automation.

1.1.4 A Brief Data Science History with the Current Landscape

Many people tend to use data science, computer learning, and machine learning interchangeably at gatherings, meetings, and in general communication with others.

Data science is a powerful combination of various disciplines.

Figure 1: Data Science - A combination of various disciplines.

Data science is the art of transforming data into actionable knowledge. At unprecedented new levels, data is being created and, even more critically, digitally recorded. However, an abundance of data is just indicative of future benefit. It must be processed, distilled, and extracted in order to be used. Data science collects knowledge, comprehends, and learns through raw data to inform constructive, systemic, and generalizable decision-making. A critical feature of data science is applying the empirical method to generate and test theories regarding fundamental trends in data.

Data science practice necessitates the application of a diverse set of skills. Data scientists must be trained to scan and manipulate vast swaths of data, which necessitates a good background in computer science. Additionally, experience with math and science enables a firm grasp of the algorithms often applied and tuned.

Data mining is the process of combining a large number of computational resources with advanced algorithms. However, a significant risk associated with this technique is the possibility of mistaking noise for the signal. Domain experience is a valuable asset for verifying causal and conceptual associations in models and assumptions.

In addition, a data scientist should have a greater understanding of analytics than an average programmer, a greater understanding of computing than a typical statistician, and the ability to adapt these abilities to market issues.

Although data science's ultimate purpose may seem simple, execution is a highly dynamic method that requires various phases earlier to detect the utility service. This is how it appears:

- ◆ Business Comprehension
- ◆ Data Interpretation
- ◆ Modelling Data Preparation
- ◆ Evaluation of stakeholders
- ◆ Deploy

A data scientist can use “statistical learning techniques known as machine learning methods or algorithms” during the simulation stage to ease out information in original data. “Machine learning” is a crucial process using predictive computation towards deducing trends, patterns, properties, characteristics, besides laws in original data to make possible results based on new observations. Rather than writing a programme by hand with unique directions and custom guidelines, machine learning is "trained" on broad datasets and "induce" how to

execute a given mission. Scalability and cost are two areas where machine learning approaches outperform human alternatives. A successful machine learning activity occurs often enough that it is not cost-efficient to do it manually.

“Linear Regression, Logistic Regression, Decision Tree, Support Vector Machines (SVM), Naive Bayes, K-Nearest Neighbor, K-Means, Random Forest, Principal Component Analysis (PCA), Gradient Boost & Adaboost, Neural Networks, and Deep Learning Neural Networks Deep Learning” are all standard machine learning algorithms.

A data scientist can employ one or more of these techniques independently in these combinations. “Machine learning” is classified into three comprehensive styles based upon the accessibility of data, challenge essence with desired outcomes.

Supervised Learning – Testing set using a list of inputs and identifiers. This technique involves building a model from scratch. For instance, if we create a model that could forecast home selling prices, we might begin a training collection for previously vended houses, associated characteristics (place, chambers, etc.) besides concluding selling values.

Unsupervised Learning – Develops a construct from a sequence of unlabeled input data sets. Unsupervised learning is used to discover the fundamental framework, distribution, or relationships within records.

Reinforcement Learning – Aims to create agents who improve by contact with their surroundings, improving with each interaction through maximizing a reward signal. For example, reinforcement learning is used to train autonomous driving vehicles.

Artificial intelligence (A.I.) - It is a concept that has a long and illustrious tradition and has amassed many connections. It is used as the term for all aspects related to computer knowledge and automation. AI remained devised by the mid-twentieth century when scientists were enthusiastic about machine algorithms prospects imitating the reasoning ability of human beings. In addition to few noteworthy achievements, the potential of machine learning has to be comprehended for better understanding.

Advancements in technology and machine learning techniques such as “deep learning neural networks, technical advances (for example graphics processing units)” and the availability of large training data sets have allowed computer programmes to conduct more human-like tasks such as image processing, speech comprehension, human text recognition, board games, and even driving vehicles! Numerous advancements in artificial intelligence are classified as “Applied or Narrow AI” and often allow the use of “deep learning neural networks” and build deep learning models. The term “Cognitive Computing” is sometimes used interchangeably with narrow artificial intelligence.

In essence, software must also undergo extensive domain-specific testing, configuration, and data collection to succeed at a particular "individual" mission.

1.2 DATA SCIENCE TERMINOLOGY

1.2.1 Basics of Data Science Terminology

There are certain fundamental principles to understand before getting interested in data science.

- 1) **Algorithms:** The algorithm collects instructions that we send to a machine to accept values to transform them into something useful. This is simple to identify and delete any comma in a text or as complicated as formulating a formula that forecasts how often home run a baseball player would reach in 2021.
- 2) **Artificial Intelligence (A.I.):** A field of study that focuses on the research and advancement of self-aware computers. The majority of work in A.I. is focused on using system knowledge to solve challenges or complete tasks. If we are unaware, artificial intelligence is already here: consider self-driving vehicles, automated doctors, and our favourite video game antagonists.
- 3) **Back End:** The back end is comprised of all the programming and hardware that operates behind doors to supply the front end for valuable data. This encompasses files, repositories, and verification procedures, amongst other things.
- 4) **Basics:** There are certain fundamental principles to understand before getting interested in data science.
- 5) **Big Data:** The word "big data" suffers from being much too large to be helpful. Big data is primarily concerned with techniques and tools that assist computers in performing complicated processing on extremely broad (read: 1+ T.B.) dataset sets. Big

data must solve the challenges classified according to the four V's: length, range, factuality, and momentum.

- 6) **Business Intelligence (B.I.):** Analogous to data mining, but with a stronger emphasis on market indicators. The technological aspect of business intelligence entails understanding how to use tools efficiently to produce reports and identify significant patterns. It is analytical rather than predictive.
- 7) **Classification:** Classification is a challenge that requires controlled machine learning. It is concerned with classifying a data point according to its relation to other data points. We collect data in which each object already has a category and examine similar characteristics among the items.
- 8) **Correlation:** Correlation is a statistical indicator of the degree to which one group of values is dependent on another. As values rise in lockstep, they are said to be positively associated. When one set's values rise and the other set's values decline, they are negatively associated. If a move in one set has no impact on a change in the other, there is no connection.
- 9) **Data Analysis:** This is the Wenger sibling in computer science. Data interpretation is mainly concerned with addressing concerns regarding the current and history. It employs four abstract data and aims to recognize trends that may help an organization develop.
- 10) **Data Engineering:** Data engineering is entirely concerned with the back end. These are the individuals that develop technologies that make it simple for data scientists to conduct research. A data scientist can also be a system engineer in smaller teams. Engineers working in broader groups will concentrate

exclusively on speeding up processing and maintaining well-organized, easy-to-access results.

- 11) **Data Journalism:** This practice is mostly about telling compelling and significant stories from a data-driven lens. It occurred spontaneously as more data became accessible. A narrative may be about data or can be told by it.
- 12) **Data Science:** With the field's exponential growth, defining data science may be challenging. Essentially, it is the discipline of making forecasts using evidence and advanced statistics. Additionally, data science is concerned with generating meaning from chaotic and disparate data.
- 13) **Data Visualization:** The graphic art of transmitting substantive results. This is done by the use of infographics, conventional charts, or even complete data dashboards. Nicholas Felton is a leader in this area, and Edward Tufte authored the book in its entirety.
- 14) **Data Warehouse:** "A data warehouse" is a computer infrastructure that enables rapid examination of market patterns by combining data from several sources. They are built to simplify things for people without a PhD in database engineering to address critical statistical questions.
- 15) **Data Wrangling (Munging):** The method of taking raw data and "taming" it to be included in a larger workflow or mission. Taming entails aligning values with a broader data collection, replacing or deleting values that could impact future research or results, and so on. The words wrangling & munging are interchangeable.

- 16) **Database:** Simply put, this is a data storage facility. We primarily work on databases managed by a Database Management System (DBMS), such as PostgreSQL or MySQL. These are computer programmes that allow us to communicate with a database to capture and interpret the inside data.
- 17) **Feature Engineering:** The method of converting the information we had with humans into a numerical meaning that a machine may comprehend. We may, for example, convert our visual perception of a mug picture into a description of pixel intensities.
- 18) **Field of Focus:** When companies grow increasingly data-driven, new ways for individuals with various ability sets to join the data ecosystem open up. These are individual specializations available within the field of data science.
- 19) **Front End:** Front end refers to anything that a developer or customer sees and interacts with directly. This is accomplished by the use of data dashboards, websites, and forms.
- 20) **Fuzzy Algorithms:** Algorithms that make use of fuzzy logic to reduce a script's runtime. Fuzzy algorithms are usually less accurate than Boolean algorithms. Additionally, they are usually quicker, and computational speed often outweighs the lack of accuracy.
- 21) **Fuzzy Logic:** An abstraction of Boolean logic in which the conventional True and False values are replaced by a set of values between 0 and 1. That is, fuzzy logic allows for expressions such as "somewhat real" or "completely fake."
- 22) **Greedy Algorithms:** A greedy algorithm can decompose a problem into discrete measures. It will then attempt to find the best solution possible at each point to obtain the best overall

solution possible. Dijkstra's algorithm is a clear example since it searches for the shortest path possible in a graph.

- 23) **Machine Learning:** A mechanism in which a machine employs an algorithm to obtain insight into a series of data and then creates predictions based on that insight. There are many machine learning techniques; the majority are categorized as controlled or unsupervised.
- 24) **Mean (Average, Expected Value):** A formula that indicates the "typical" meaning for a set of numbers. The mean is calculated by multiplying a sequence of values by the number of values in the list.
- 25) **Median:** The median value of a series of ordered values is whatever value is in the centre. We often use the median in conjunction with the mean to determine if specific values in the range are exceptionally large or low. This is an early indication that outliers can be investigated.
- 26) **Normalize:** When all of the values in a data collection have been modified to fall into a standard range, the data set is considered normalized. We normalize data sets to facilitate and clarify similarities. For example, aggregating movie reviews from various websites and converting them to a scale of 0 to 100.
- 27) **Outlier:** "An outlier is a data point" that is known to be significantly different from the others. They are often the product of outliers or calculation errors and should be investigated early in a data collection workflow.
- 28) **Overfitting:** Overfitting occurs when a model assumes an excessive amount of data. It is analogous to telling someone to

BIBLIOGRAPHY

1. **Agarwal, R., & Dhar, V. (2014).** Big data, data science, and analytics: The opportunity and challenge for IS research.
2. **Bruce, P., Bruce, A., & Gedeck, P. (2020).** *Practical Statistics for Data Scientists: 50+ Essential Concepts Using R and Python*. O'Reilly Media.
3. **Lahti, L., Marjanen, J., Roivainen, H., & Tolonen, M. (2019).** Bibliographic Data Science and the History of the Book (c. 1500–1800). *Cataloguing & Classification Quarterly*, 57(1), 5-23.
4. **Leskovec, J., Rajaraman, A., & Ullman, J. D. (2020).** *Mining of massive data sets*. Cambridge university press.
5. **O'Neil, C., & Schutt, R. (2013).** *Doing data science: Straight talk from the frontline*. " O'Reilly Media, Inc."
6. **Ruehle, F. (2020).** Data science applications to string theory. *Physics Reports*, 839, 1-117.
7. **Van Der Aalst, W. (2016).** Data science in action. In *Process Mining* (pp. 3-23). Springer, Berlin, Heidelberg.

This
Page
Intentionally
Left
Blank