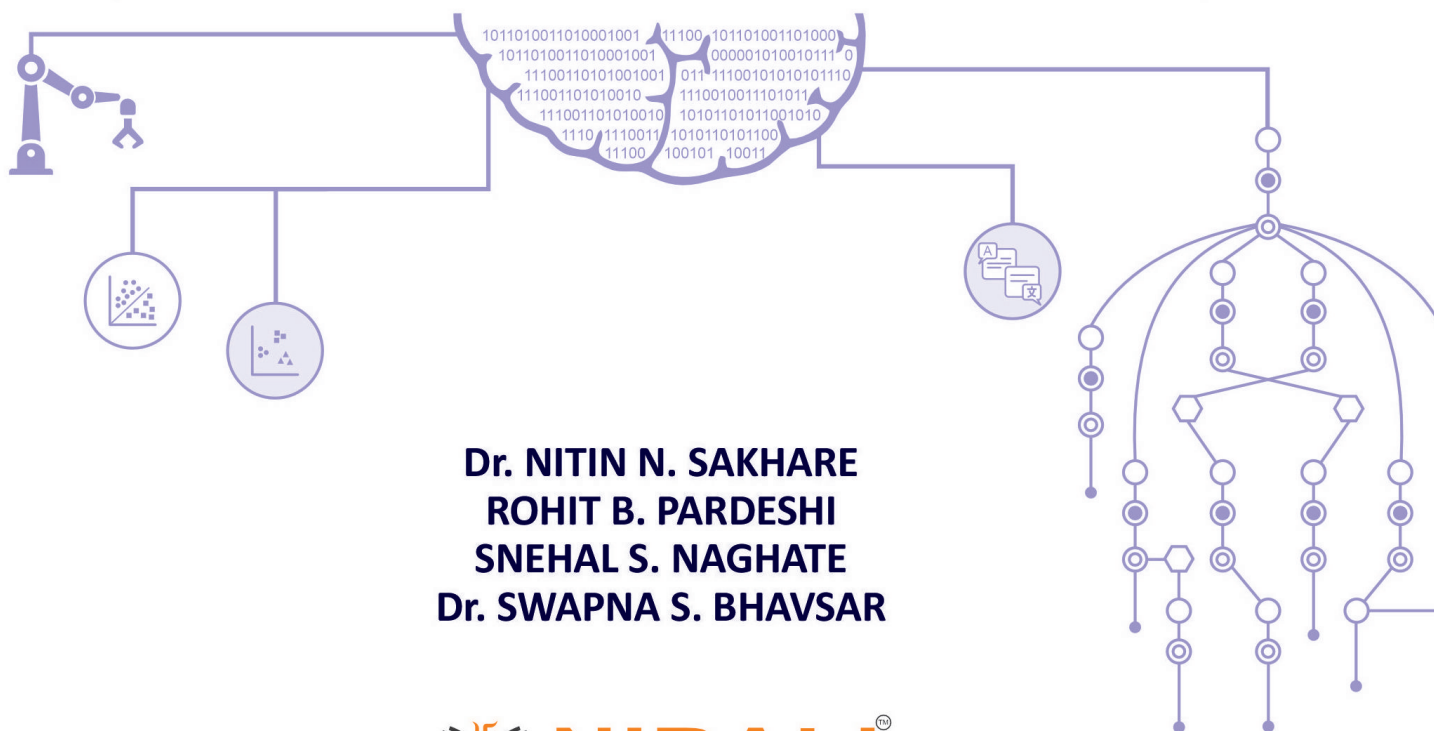


DATA SCIENCE AND MACHINE LEARNING



Dr. NITIN N. SAKHARE
ROHIT B. PARDESHI
SNEHAL S. NAGHATE
Dr. SWAPNA S. BHAVSAR

DATA SCIENCE AND MACHINE LEARNING

Dr. Nitin N. Sakhare

*Ph.D. Computer Science & Engineering
Assistant Professor,
Department of Computer Engineering,
Vishwakarma Institute of Information Technology,
Kondhwa, Pune.*

Rohit B. Pardeshi

*B.E. Computer
Senior Engineer,
eInfochips,
Pune.*

Snehal S. Naghate

*M.E. Computer Science & Information Technology
Assistant Professor,
Department of Computer Engineering,
Ballarpur Institute of Technology,
Chandrapur.*

Dr. Swapna S. Bhavsar

*Ph.D. Computer Science & Engineering
Assistant Professor,
Department of Information Technology,
PES's Modern College of Engineering,
Pune.*

Preface

We are glad to present this book entitled '**Data Science and Machine Learning**'. The world is experiencing a data revolution where vast amounts of information are generated every second. Data Science and Machine Learning have emerged as transformative technologies, enabling businesses, researchers and professionals to extract meaningful insights, make informed decisions and develop intelligent systems.

This book is designed as a comprehensive guide and lucid presentation of subject matter for students, academicians and professionals looking to gain a strong foundation in data science and machine learning. It covers essential concepts such as data pre-processing, exploratory data analysis, feature engineering, supervised and unsupervised learning, deep learning, model evaluation and real-world applications. Each chapter balances theoretical understanding with practical implementation, ensuring a well-rounded learning experience.

One of the key highlights of this book is its hands-on approach, with numerous Python-based examples and real-world case studies. Readers will learn how to implement machine learning algorithms using popular libraries like NumPy, Pandas, Matplotlib, Scikit-learn and TensorFlow. By working through the exercises and projects, readers will develop the skills needed to apply machine learning techniques to real-world problems.

This book is suitable for beginners and experienced practitioners alike. Whether you are a student starting your data science journey, a researcher exploring new methods or a professional looking to upskill, this book provides a structured and practical approach to mastering the field.

We also take this opportunity to express our sincere thanks to Shri Dineshbhai Furia, Shri Jignesh Furia, Mrs. Nirali Verma, Mrs. Deepali Lachake (coordinator) and the entire team at Nirali Prakashan for their keen interest and tireless efforts in publishing this book.

The advice and suggestions of our esteemed readers to improve the text are most welcome and will be highly appreciated.

Authors

Contents

1.	INTRODUCTION TO DATA SCIENCE	1
1.1	Basics and Needs of Data Science and Big Data	1
1.2	Applications of Data Science	3
1.3	Data Explosion	6
1.4	5V's of Big Data	7
1.5	Relationship between Data Science and Information Science	8
1.6	Business Intelligence vs Data Science	9
1.7	Data Science Life Cycle	10
1.8	Data: Data Types, Data Collection	12
1.9	What are Association Rules?	16
1.10	Calculating Association Rule Parameters	17
1.11	Need of Data Wrangling	19
1.12	Methods	20
1.13	Recommendation Engines	27
1.14	Recommendation Engines Working	30
1.15	Collaborative Filtering	33
1.16	Content Based Filtering	41
1.17	Case Study: Data Pre-processing using Techniques of Data Cleaning and Data Transformation. (Use suitable dataset)	42
	• Summary	43
	• Exercise	44
2.	DATA ANALYTICS LIFE CYCLE	45
2.1	Introduction to Big Data	45
2.2	Sources of Big Data	48
2.3	Data Analytic Lifecycle: Overview	49
2.4	Data Analytics Lifecycle Example	54
2.5	Need of Life Cycle	54
2.6	Case Study: Global Innovation Network and Analysis (Gina)	54
	• Summary	60
	• Exercise	60
3.	STATISTICAL INFERENCE	61
3.1	Need of Statistics in Data Science and Big Data Analytics	61
3.2	Measures of Central Tendency: Mean, Median, Mode, Mid-Range	62
3.3	Measures of Dispersion: Range, Variance, Mean Deviation, Standard Deviation	72
3.4	Bayes' Theorem	76
3.5	Basics and Need of Hypothesis and Hypothesis Testing	79
3.6	Pearson Correlation	86
3.7	Sample Hypothesis Testing	88
3.8	Chi-Square Tests, T-Test	90
3.9	Case Study: For an Employee Dataset, Create Measure of Central Tendency and Its Measure of Dispersion for Statistical Analysis of Given Data	95
	• Summary	97
	• Exercise	97

4.	PREDICTIVE DATA ANALYTICS WITH PYTHON	99
4.1	Introduction to Predictive Big Data Analytics	99
4.2	Essential Python Libraries and Basic Examples	102
4.3	Data Preprocessing	106
4.4	Analytics Types	123
4.5	Association Rules	129
4.6	Regression	148
4.7	Classification Requirement	154
4.8	Introduction to Scikit-Learn	171
4.9	Case Study: Use Iris Dataset from Scikit and Apply Data Preprocessing Methods	180
	• Summary	186
	• Exercise	186
5.	INTRODUCTION TO MACHINE LEARNING	187
5.1	Introduction to Machine Learning	187
5.2	Comparison of Machine Learning with Traditional Programming	191
5.3	ML vs. AI vs. Data Science	193
5.4	Types of Learning	197
5.5	Reinforcement Learning Techniques	204
5.6	Models of Machine Learning	213
5.7	Important Elements of Machine Learning	223
	• Case Study	228
	• Summary	229
	• Exercise	230
6.	FEATURE ENGINEERING	231
6.1	Concept of Feature	231
6.2	Pre-processing of Data	234
6.3	Introduction to Dimensionality Reduction	253
6.4	Introduction to Various Feature Selection Techniques	258
6.5	Statistical Feature Engineering	271
6.6	Feature Vector Creation	278
6.7	Multidimensional Scaling	279
6.8	Matrix Factorization Techniques	280
	• Summary	283
	• Exercise	284
7.	SUPERVISED LEARNING	285
7.1	Bias	285
7.2	Variance	286
7.3	Generalization	289
7.4	Underfitting	291
7.5	Overfitting	292
7.6	Linear Regression	293
7.7	Gradient Descent Algorithm	304
7.8	Evaluation Metrics	308
	• Summary	312
	• Exercise	312

8.	CLASSIFICATION	313
8.1	Classification	313
8.2	Ensemble Learning	331
8.3	Binary-vs-Multiclass Classification	348
8.4	Balanced and Imbalanced Multiclass Classification Problem	350
8.5	Variants of Multiclass Classification	359
8.6	Metrics and Score	368
•	Summary	373
•	Exercise	374
9.	DATA ANALYTICS AND MODEL EVALUATION (I)	375
9.1	Introduction to Unsupervised Learning	375
9.2	What is Clustering?	377
9.3	Cluster Analysis	381
9.4	Distance Measure	382
9.5	Using K-means for Flat Clustering	385
9.6	Using K-means from Sklearn	386
9.7	Implementing Fit and Predict Functions	389
9.8	Implementing K-Means Class	391
9.9	Partitioning Method	393
9.10	Hierarchical Methods	399
9.11	Top Down / Divisive Approach	405
9.12	Bottom Up / Agglomerative Approach	407
9.13	Linkage Metrics in Hierarchical Method	409
9.14	Working of Dendrogram in Hierarchical Clustering	414
9.15	PCA – 1	415
9.16	PCA – 2	424
•	Important Formulae	449
•	Summary	449
•	Exercise	450
10.	DATA ANALYTICS AND MODEL EVALUATION (II)	451
10.1	Time Series Analysis	451
10.2	Importance of TSA	467
10.3	Components of TSA	467
10.4	White Noise	468
10.5	AR (Auto-Regressive) Model	473
10.6	MA (Moving Average) Model	473
10.7	ARMA (Auto Regressive Moving Average) Model	474
10.8	ARIMA (Auto-Regressive Integrated Moving Average) Model	475
10.9	Stationarity	476
10.10	ACF and PACF	483
10.11	Introduction to Text Analysis	492
10.12	Need and Introduction to Social Media Analysis	507
•	Important Formulae	510
•	Summary	511
•	Exercise	512

11.	DATA ANALYTICS AND MODEL EVALUATION (III)	513
11.1	Introduction to Business Analysis	513
11.2	Model Evaluation and Selection	515
11.3	Need of Model Selection	518
11.4	Cross – Validation	519
11.5	Boosting	521
11.6	Boosting Algorithms	525
11.7	Types of Boosting Algorithms	526
11.8	Adaptive Boosting	530
11.9	Holdout Method and Random Subsampling	533
11.10	Parameter Tuning and Optimization	536
11.11	Clustering and Time-Series Analysis Using Scikit-Learn	543
11.12	Sklearn.Metrics	556
11.13	Classification Metrics	563
11.14	AUC-ROC	564
11.15	Log Loss	565
11.16	Elbow Method	567
11.17	Silhouette Algorithm to Choose K the Silhouette Method	570
11.18	Case Study: Use Iris Dataset from Scikit Learn and Apply K-Means Clustering Methods	572
	• Important Formulae	579
	• Summary	580
	• Exercise	581
12.	INTRODUCTION TO NEURAL NETWORKS	583
12.1	Artificial Neural Networks	583
12.2	Back Propagation Learning	597
12.3	Functional Link Artificial Neural Network	604
12.4	Radial Basis Function	604
12.5	Activation Functions	609
12.6	Introduction to Recurrent Neural Networks	620
12.7	Convolutional Neural Networks	628
	• Important Formulae	638
	• Summary	639
	• Exercise	640
13.	NATURAL LANGUAGE PROCESSING (NLP)	641
13.1	What is NLP	641
13.2	What Is Natural Language	641
13.3	Approaches to NLP	646
13.4	Current Applications of Heuristic Methods In Natural Language Processing	648
13.5	Challenges in NLP	650
	• Important Formulae	651
	• Summary	652
	• Exercise	652

Introduction to Data Science



OUTLINE

- 1.1** *Basics and Needs of Data Science and Big Data*
 - 1.2** *Applications of Data Science*
 - 1.3** *Data Explosion*
 - 1.4** *5V'S of Big Data*
 - 1.5** *Relationship Between Data Science and Information Science*
 - 1.6** *Business Intelligence vs Data Science*
 - 1.7** *Data Science Life Cycle*
 - 1.8** *Data: Data Types, Data Collection*
 - 1.9** *What are Association Rules?*
 - 1.10** *Calculating Association Rule Parameters*
 - 1.11** *Need of Data Wrangling*
 - 1.12** *Methods*
 - 1.13** *Recommendation Engines*
 - 1.14** *Recommendation Engines Working*
 - 1.15** *Collaborative Filtering*
 - 1.16** *Content Based Filtering*
 - 1.17** *Case Study*
-

1.1 BASICS AND NEEDS OF DATA SCIENCE AND BIG DATA

Data Science and Big Data are extremely important fields and concepts that are becoming increasingly critical in this modern era. The world has never collected or stored as much data and as fast as it does today. In addition, the variety and volume of data is growing at an alarming rate. Data is analogous to gold in many ways. It is extraordinarily valuable and has many uses, but you often have to pay for it in order to realize its value. There are many debates as to whether Data Science is a new field. Many argue that similar practices have been used and branded as statistics, analytics, business intelligence and so forth. In either case, Data Science is a very popular and prominent term used to describe many different data-related processes and techniques. Big Data on the other hand is relatively new in the sense that the amount of data collected and the associated challenges continues to require new and innovative hardware and techniques for handling it.

1.1.1 Data Science

Data Science is complex and involves many specific domains and skills, but the general definition is that Data Science encompasses all the ways in which information and knowledge are extracted from data. Data is everywhere and is found in huge and exponentially increasing quantities.

Data Science as a whole reflects the ways in which data is discovered, conditioned, extracted, compiled, processed, analyzed, interpreted, modeled, visualized, reported on and presented regardless of the size of the data being processed. Big Data (as defined soon) is a special application of Data Science. Data Science is a very complex field, which is largely due to the diversity and number of academic disciplines and technologies it draws upon. Data Science incorporates mathematics, statistics, computer science and programming, statistical modeling, database technologies, signal processing, data modeling, artificial intelligence and learning, natural language processing, visualization, predictive analytics and so on. Data Science is highly applicable to many fields, including social media, medicine, security, health care, social sciences, biological sciences, engineering, defence, business, economics, finance, marketing, geo-location and many more.

1.1.2 Big Data

Big Data is essentially a special application of Data Science, in which the data sets are enormous and require overcoming logistical challenges to deal with them. The primary concern is efficiently capturing, storing, extracting, processing and analyzing information from these enormous data sets. Processing and analysis of these huge data sets is often not feasible or achievable due to physical and/or computational constraints. Special techniques and tools (e.g., software, algorithms, parallel programming, etc.) are therefore required. Big Data is the term that is used to encompass these large data sets, specialized techniques and customized tools. It is often applied to large data sets in order to perform general data analysis and find trends or to create predictive models. Large amounts of data are gathered from mobile devices, remote sensing devices, geo-location, software applications, multimedia devices, radio-frequency identification readers, wireless sensor networks and so on. A primary component of Big Data is the so-called Three Vs (3Vs) model. This model represents the characteristics and challenges of Big Data as dealing with Volume, Variety and Velocity. Companies such as IBM include a fourth “V”, Veracity, while Wikipedia also notes variability. Big Data essentially aims to solve the problem of dealing with enormous amounts of varying-quality data, often of many different types, that is being captured and processed sometimes at tremendous (real-time) speeds. No easy task to say the least. So in summary, Big Data can be thought of being a relative term that applies to huge data sets that require an entity (person, company, etc.) to leverage specialized hardware, software, processing techniques, visualization and database technologies in order to solve the problems associated with the 3Vs and similar characteristic models.

1.1.3 Types of Data and Data Sets

Data is collected in many different ways as mentioned earlier. The life cycle of usable data usually involves capture, pre-processing, storage, retrieval, post-processing, analysis, visualization and so on. Once captured, data is usually referred to as being structured, semi-structured or unstructured. These distinctions are important because they are directly related to the type of database technologies and storage required, the software and methods by which the data is queried and processed and the complexity of dealing with the data. Structured data refers to data that is stored as a model (or is defined by a structure or schema) in a relational database or spreadsheet. Often, it is easily queryable using SQL (Structured Query Language) since the “structure” of the data is known. A sales order record is a good example. Each sales order has a purchase date, items purchased, purchaser, total cost, etc. Unstructured data is data that is not defined by any schema, model or structure and is not organized in a specific way. In other words, it is just stored raw data. Think of a seismometer (earthquakes are a big fear of mine by the way!). You have probably seen the squiggly lines captured by such a device, which essentially represent energy data as recorded at each seismometer location. The recorded signal (that is data) represents a varying amount of energy over time. There is no structure in this case, it is just variations of energy represented by the signal. It follows naturally that Semi-structured data is a combination of the two. It is basically unstructured data that also has structured data (a.k.a. metadata) appended to it. Every time you use your smartphone to take a picture, the shutter captures light reflection information as a bunch of binary data (i.e., ones and zeros). This data has no structure to it, but the camera also appends additional data that includes the date and time the photo was taken, last time it was modified, image size, etc. That is the structured part. Data formats such as XML and JSON are also considered to be semi-structured data.

**TIPS**

1. **Data Science** focuses on extracting knowledge and insights from data using techniques from mathematics, statistics, programming and AI, making it applicable across various fields like healthcare, business and engineering.
2. **Big Data** deals with massive datasets requiring specialized tools and techniques to address challenges related to **Volume, Variety and Velocity**, along with **Veracity** in some models.
3. Data can be categorized into **structured** (organized and schema-defined), **unstructured** (raw and unorganized) and **semi-structured** (a mix of both, like metadata in JSON or XML formats).

1.2 APPLICATIONS OF DATA SCIENCE

The role of Data Science Applications has not evolved overnight. Thanks to faster computing and cheaper storage, we can now predict outcomes in minutes, which could take several human hours to process.

1.2.1 Fraud and Risk Detection

The earliest applications of Data Science were in Finance. Companies were fed up of bad debts and losses every year. However, they had a lot of data which use to get collected during the initial paperwork while sanctioning loans. They decided to bring in data scientists in order to rescue them out of losses. Over the years, banking companies learned to divide and conquer data via customers profiling, past expenditures and other essential variables to analyze the probabilities of risk and default. Moreover, it also helped them to push their banking products based on customer's purchasing power.

1.2.2 Healthcare

The healthcare sector, especially, receives great benefits from Data Science applications.

1. Medical Image Analysis

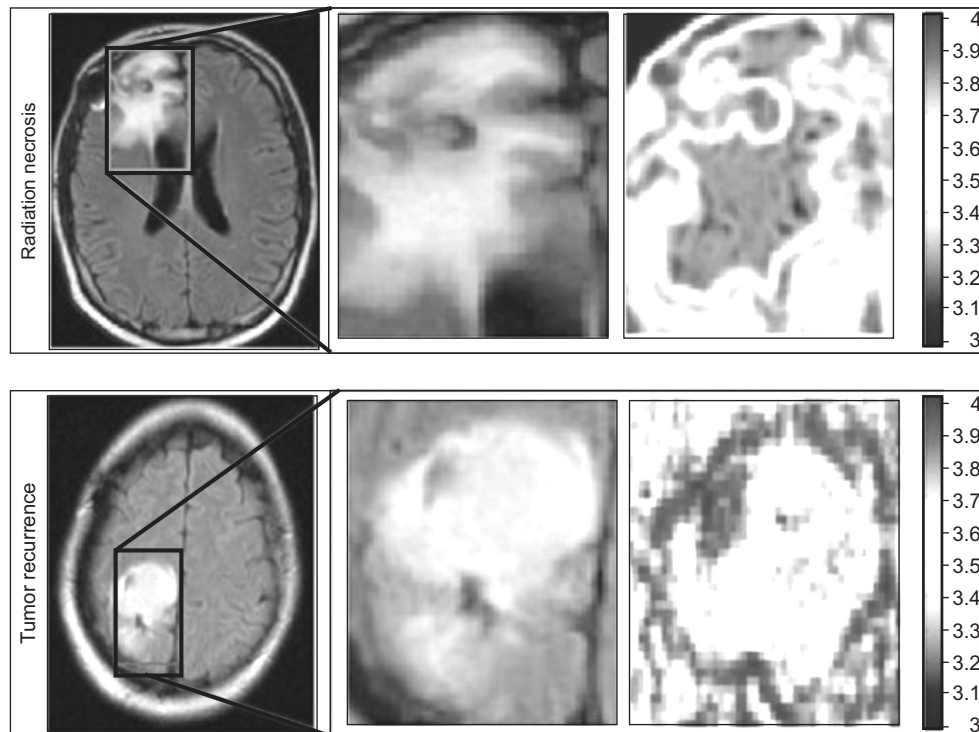


Fig. 1

Procedures such as detecting tumours, artery stenosis and organ delineation employ various different methods and frameworks like MapReduce to find optimal parameters for tasks like lung texture classification. It applies Machine Learning Methods, Support Vector Machines (SVM), content-based medical image indexing and wavelet analysis for solid texture classification.

2. Genetics and Genomics

Data Science applications also enable an advanced level of treatment personalization through research in genetics and genomics. The goal is to understand the impact of the DNA on our health and find individual biological connections between genetics, diseases and drug response. Data Science techniques allow integration of different kinds of data with genomic data in the disease research, which provides a deeper understanding of genetic issues in reactions to particular drugs and diseases. As soon as we acquire reliable personal genome data, we will achieve a deeper understanding of the human DNA. The advanced genetic risk prediction will be a major step towards more individual care.

3. Drug Development

The drug discovery process is highly complicated and involves many disciplines. The greatest ideas are often bounded by billions of testing, huge financial and time expenditure. On average, it takes twelve years to make an official submission. Data Science applications and machine learning algorithms simplify and shorten this process, adding a perspective to each step from the initial screening of drug compounds to the prediction of the success rate based on the biological factors. Such algorithms can forecast how the compound will act in the body using advanced mathematical modeling and simulations instead of the “lab experiments”. The idea behind the computational drug discovery is to create computer model simulations as a biologically relevant network simplifying the prediction of future outcomes with high accuracy.

4. Virtual Assistance for Patients and Customer Support

Optimization of the clinical process builds upon the concept that for many cases it is not actually necessary for patients to visit doctors in person. A mobile application can give a more effective solution by bringing the doctor to the patient instead. The AI-powered mobile apps can provide basic healthcare support, usually as chatbots. You simply describe your symptoms or ask questions and then receive key information about your medical condition derived from a wide network linking symptoms to causes. Apps can remind you to take your medicine on time and if necessary, assign an appointment with a doctor. This approach promotes a healthy lifestyle by encouraging patients to make healthy decisions, saves their time waiting in line for an appointment and allows doctors to focus on more critical cases.

1.2.3 Internet Search

Now, this is probably the first thing that strikes your mind when you think Data Science Applications. When we speak of search, we think ‘Google’. Right? But there are many other search engines like Yahoo, Bing, Ask, AOL and so on. All these search engines (including Google) make use of Data Science algorithms to deliver the best result for our searched query in a fraction of seconds. Considering the fact that, Google processes more than 20 petabytes of data every day. Had there been no Data Science, Google would not have been the ‘Google’ we know today.

1.2.4 Targeted Advertising

If you thought search would have been the biggest of all Data Science applications, here is a challenger – the entire digital marketing spectrum. Starting from the display banners on various websites to the digital billboards at the airports – almost all of them are decided by using Data Science algorithms. This is the reason why digital ads have been able to get a lot higher CTR (Call-Through Rate) than traditional advertisements. They can be targeted based on a user’s past behavior.

1.2.5 Website Recommendations

Are not we all used to the suggestions about similar products on Amazon? They not only help you find relevant products from billions of products available with them but also adds a lot to the user experience. A lot of companies have fervidly used this engine to promote their products in accordance with user's interest and relevance of information. Internet giants like Amazon, Twitter, Google Play, Netflix, LinkedIn, IMDb and many more use this system to improve the user experience. The recommendations are made based on previous search results for a user.


Compare with similar items				
	This item Bose SoundLink Wireless Around-Ear Headphones with Mic (Black)	Sennheiser HD 4.40-BT Bluetooth Headphones (Black)	Bose 741158-0020 Soundlink wireless Around-Ear Headphones with Mic (White)	Bose 789564-0030 Quiet comfort 35 Wireless Headphone (Blue)-Special Edition
	Add to Cart	Add to Cart	Add to Cart	Add to Cart
Customer rating	★★★★☆(68)	★★★★☆(349)	★★★★☆(22)	★★★★☆(200)
Price	₹ 19,000.00	₹ 7,490.00	₹ 19,000.00	₹ 29,363.00
Shipping	FREE Shipping	FREE Shipping	FREE Shipping	FREE Shipping
Sold By	Appario Retail Private Ltd	Appario Retail Private Ltd	Appario Retail Private Ltd	Appario Retail Private Ltd
Colour	Black	Black	White	Blue
Connectivity technology	Bluetooth wireless	Bluetooth wireless	Bluetooth wireless	Bluetooth wireless

Fig. 2

1.2.6 Advanced Image Recognition

You upload your image with friends on Facebook and you start getting suggestions to tag your friends. This automatic tag suggestion feature uses face recognition algorithm. In their latest update, Facebook has outlined the additional progress they have made in this area, making specific note of their advances in image recognition accuracy and capacity. We have witnessed massive advances in image classification (what is in the image?) as well as object detection (where are the objects?), but this is just the beginning of understanding the most relevant visual content of any image or video. Recently we have been designing techniques that identify and segment each and every object in an image, a key capability that will enable entirely new applications. In addition, Google provides you with the option to search for images by uploading them. It uses image recognition and provides related search results.

1.2.7 Speech Recognition

Some of the best examples of speech recognition products are Google Voice, Siri, Cortana, etc. Using speech-recognition feature, even if you are not in a position to type a message, your life would not stop. Simply speak out the message and it will be converted to text. However, at times, you would realize, speech recognition does not perform accurately.

1.2.8 Airline Route Planning

Airline Industry across the world is known to bear heavy losses. Except for a few airline service providers, companies are struggling to maintain their occupancy ratio and operating profits. With high rise in air-fuel prices and need to offer heavy discounts to customers has further made the situation worse. It was not for long when airlines companies started using Data Science to identify the strategic areas of improvements. Now using Data Science, the airline companies can:

- Predict flight delay.
- Decide which class of airplanes to buy.

- (c) Whether to directly land at the destination or take a halt in between (For example, A flight can have a direct route from New Delhi to New York. Alternatively, it can also choose to halt in any country).
- (d) Effectively drive customer loyalty programs.

Southwest Airlines, Alaska Airlines are among the top companies who have embraced Data Science to bring changes in their way of working.

1.2.9 Gaming

Games are now designed using machine learning algorithms which improve/upgrade themselves as the player moves up to a higher level. In motion gaming also, your opponent (computer) analyzes your previous moves and accordingly shapes up its game. EA Sports, Zynga, Sony, Nintendo, Activision-Blizzard have led gaming experience to the next level using Data Science.

1.2.10 Augmented Reality

This is the final of the Data Science application which seems most exciting in the future. Data Science and Virtual Reality do have a relationship, considering a VR headset contains computing knowledge, algorithms and data to provide you with the best viewing experience. A very small step towards this is the high trending game of Pokemon GO. The ability to walk around things and look at Pokemon on walls, streets, things that are not really there. The creators of this game used the data from Ingress, the last app from the same company, to choose the locations of the Pokemon and gyms. However, Data Science makes more sense once VR economy becomes accessible in terms of pricing and consumer use it often like other apps. Though, not much has been revealed about them except the prototypes and neither do we know when they would be available for a common man's disposal. Let's see, what amazing Data Science applications the future holds for us!



TIPS

1. **Fraud and Risk Detection:** Used extensively in finance to analyze customer profiles, spending patterns and risk probabilities to minimize losses and defaults.
2. **Healthcare Applications:** Data Science enhances medical image analysis, genomics, drug development and patient support through predictive models and AI-powered applications.
3. **Internet Search and Recommendations:** Search engines and platforms like Google and Amazon use Data Science algorithms to deliver precise results and personalized recommendations in real-time.

1.3 DATA EXPLOSION

The world is currently used to sparing everything without exception in the electronic space. Processing power, RAM speeds and hard-disk sizes have expanded to level that has changed our viewpoint towards data and its storage. Would you be able to envision having 256 or 512 MB RAM in your PC now? On the off chance that we comprehend idea of byte, we can envision how data growth has expanded over time and how storage systems handle it. We know that 1 byte is equivalent to 8 bits and these 8 bits can represent character or expression. An archive with huge number of bytes will contain huge number of characters, expressions and spaces etc. Similarly, Megabyte (MB) is million bytes of information, Gigabyte (GB) is billion bytes of information and Terabyte (TB) is trillion bytes of information. We use these terms while managing data and storage, on our everyday activities. But it does not end here. Next comes the Petabyte, which is quadrillion bytes or million gigabytes. Ones after that are Exabyte, Zettabyte and Yottabyte. Yottabyte is basically trillion Terabytes of information. There are considerably higher numbers and we will stop here now. The rapid or exponential increase in the amount of data that is generated and stored in the computing systems, that reaches level where data management becomes difficult, is called "Data Explosion".

DATA SCIENCE AND MACHINE LEARNING

Dr. NITIN N. SAKHARE

Ph.D. (Computer Science & Engineering),
Assistant Professor,
Department of Computer Engineering,
Vishwakarma Institute of Information Technology,
Kondhwa, PUNE.

ROHIT B. PARDESHI

B.E. (Computer Engineering),
Senior Engineer,
eInfochips,
PUNE.

SNEHAL S. NAGHATE

M.E. (Computer Science & Information Technology),
Assistant Professor,
Department of Computer Engineering,
Ballarpur Institute of Technology,
CHANDRAPUR.

Dr. SWAPNA S. BHAVSAR

Ph.D. (Computer Science & Engineering),
Assistant Professor,
Department of Information Technology,
PES's Modern College of Engineering,
PUNE.

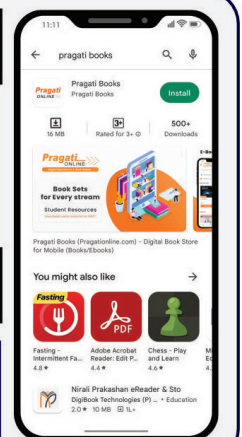
**Your Favourite
Book Brand is now on
PLAY STORE !**

Buy our Engineering e-Books from

**PRAGATI BOOKS APP BOOKSTORE
AND E-READER**



**Scan the QR code
to download our app**



Note : Do not Xerox Or Copy the Book, It is a Punishable Offence.

**BOOKS
AVAILABLE**

**Scan QR code
to access
our Books**



Pragati
ONLINE.COM
Digital Bookstore & E-Book Reader

amazon
Flipkart

Also available at all bookstores in India.



NIRALI PRAKASHAN

www.niralibooks.com
 niralipune@pragationline.com
 (+91-020) 25512336/ 7/ 9

www.facebook.com/niralibooks
 @nirali.prakashan
 www.pragationline.com

