Fun with Machine Learning

Simplify the Data Science process by automating repetitive and complex tasks using AutoML

Arockia Liborious
Dr. Rik Das

www.bpbonline.com

Every effort has been made in the preparation of this book to ensure the accuracy of the information presented. However, the information contained in this book is sold without warranty, either express or implied. Neither the author, nor BPB Online or its dealers and distributors, will be held liable for any damages caused or alleged to have been caused directly or indirectly by this book.

BPB Online has endeavored to provide trademark information about all of the companies and products mentioned in this book by the appropriate use of capitals. However, BPB Online cannot guarantee the accuracy of this information.

First published: 2023

Published by BPB Online
WeWork
119 Marylebone Road
London NW1 5PU

UK | UAE | INDIA | SINGAPORE

ISBN 978-93-55517-852

www.bpbonline.com

Dedicated to

Arockia Liborious

Dedicated to the light of my life, my Wife: Gargi

My pillars of support: Mom: Mary, and Brother: Camillus

Dr. Rik Das

Dedicated to my parents: Mr. Kamal Kumar Das and Mrs. Malabika Das

My wife: Simi

My kids: Sohan and Dikshan

About the Authors

- **Arockia Liborious** stands out as a celebrated leader in analytics, having amassed over a dozen accolades for his work in machine learning over the past 12 years. He is a seasoned professional who has conceptualized and implemented machine learning solutions for businesses across different domains. His expertise includes computer vision and natural language processing, which enables him to provide insights into emerging technologies and their impact on businesses. Arockia has a unique blend of technical and business expertise, which allows him to evaluate the long-term return on investment (eROI) of machine learning solutions, as well as their strategic positioning within the market. He has successfully led several initiatives to develop innovative service and product strategies that leverage machine learning to solve complex business problems.

  Throughout his career, Arockia has worked with clients in diverse industries, including Manufacturing, Fintech, Banking, Food and Beverage, and large Energy clients. His experience has equipped him with the ability to understand and cater to the unique needs of each industry, enabling him to deliver custom-tailored solutions that drive business results. Apart from his work in machine learning, Arockia is also an expert in data analytics, predictive modeling, feature engineering, and data visualization. He is passionate about using data to drive informed decision-making and believes that businesses that leverage the power of data will be the ones that succeed in today’s competitive market. Arockia holds a master’s degree in data science and is a sought-after speaker at industry events and conferences. He has received numerous awards throughout his career, recognizing his contributions to the field of analytics.

- **Dr. Rik Das** is a Lead Software Engineer (Artificial Intelligence Unit) at Siemens Technology and Services Pvt. Ltd. (Siemens Advanta), Bengaluru. He is a thought leader with over 18 years of experience in Industrial and Academic Research. Rik has been a seasoned academician before joining the corporate and has served as a Professor in multiple prestigious Universities, Institutions and EduTechs of the country.

Dr. Das is conferred with the title of ACM Distinguished Speaker by the Association for Computing Machinery (ACM), New York, USA. He is also a Member of International Advisory Committee of AI-Forum, UK. Dr. Das is a Ph.D. (Tech.) in Information Technology from University of Calcutta. He has also received his M.Tech. (Information Technology) from University of Calcutta after his B.E. (Information Technology) from University of Burdwan. As an innovation leader, Rik has carried out collaborative research in multiple domains. He has filed and published two Indian patents consecutively during the year 2018 and 2019 and has over 80 International publications till date with reputed publishers He has published 10 books on applications of AI/ML and Computer Vision. He has also chaired multiple sessions in International Conferences on Machine Learning and has acted as a resource person / invited speaker in multiple events and refresher courses on Information Technology. Dr. Rik Das is recipient of multiple prestigious awards due to his significant contributions in the domain of Technology Consulting, Research Innovations and Academic Excellence. He is awarded with Warner von Siemens Country Award 2021 followed by Divergent category award 2022 consecutively during his current employment at Siemens Advanta. He has received "Best Innovation Award" in Computer Science category at UILA Awards 2021. He was featured in uLektz Wall of Fame as one of the "Top 50 Tech Savvy Academicians in Higher Education across India" for the year 2019. His keen interest towards application of machine learning and deep learning techniques for real life use case solutions has resulted in joint publications of research articles with Professors and Researchers from various reputed Multinationals brands namely Philips-Canada, Cognizant Technology Solutions, TCS, etc. and International Universities including College of Medicine, University of Saskatchewan, Canada, Faculty of Electrical Engineering and Computer Science, VSB Technical University of Ostrava, Ostrava, Czechia, Cairo University, Giza, Egypt and so on. Dr. Rik Das is always open to discuss new research project ideas for collaborative work and for techno-managerial consultancies.

About the Reviewers

- **Peter Henstock** is the Machine Learning & AI Lead at Pfizer. His work has focused on the intersection of AI, visualization, statistics, and software engineering. At Pfizer, he has been mostly developing solutions for the drug discovery area but has more recently focused on clinical trials. Prior to Pfizer, he worked at MIT Lincoln Laboratory in computational linguistics and image analysis. Peter holds a Ph.D. in Artificial Intelligence from Purdue University and 7 Master's degrees including an MBA. The Deep Knowledge Analytics group recognized him as among the top 12 leaders in AI and Pharma globally. He also currently teaches two graduate-level courses at Harvard: "Advanced Machine Learning, Data Mining and AI" and the Software Engineering capstone course.
- **Sheena Siddiqui** is a machine learning engineer, working with one of the leading global organizations. She has 5 years of work experience in diverse AI and ML technologies, where she has made her mark in the top 1% of all employees. She has authored several introductory and advanced-level online courses and conducted webinars for learners in her field. She is a postgraduate in Electrical Engineering from Jamia Millia Islamia. In addition to AI and ML, her area of interest includes Quantum Computing and Sustainability Research. She also works as a video editor, graphic designer, and technical writer.

Acknowledgements

We would like to express our deepest gratitude to the many people who have supported us throughout the process of writing this book. First and foremost, we want to thank our family, who have been unwavering in their love and encouragement. Their belief in us and our abilities has been the foundation of our success.

We would also like to thank our friends, who have been a constant source of inspiration and support. Their feedback and encouragement have been invaluable in shaping this book.

We also want to thank BPB Publications for their consideration in publishing this book. Their guidance and suggestions have been instrumental in shaping this work, and we are deeply grateful for their support.

In writing this book, we have been privileged to work with some truly amazing people, and we are grateful to every one of them for their contributions. Thank you all for your support, encouragement, and belief in us.

Preface

Welcome to "Fun with Data Science" - a book that aims to make data-driven decision making accessible and easy for everyone. In today's world, data is everywhere and plays a critical role in every aspect of business. However, not everyone has the skills and expertise to use it effectively. That's where this book comes in.

With the help of auto ML tools, we can make data-driven decision making easier for everyone, even those who are not data scientists. This book is designed to help organizations move from intuition-driven decision making to a more data-driven approach. It provides a step-by-step guide to using auto ML tools to solve business problems and make data-backed decisions.

This book is for anyone who wants to take data-backed decisions but does not know where to start. Whether you are a business leader, a manager, or a professional from any field, this book will help you understand the basics of data science and how you can use it to drive your organization's success.

We hope that you find this book informative, engaging, and most importantly, fun. Let's dive in and explore the world of data science together!

Chapter 1: Significance of Machine Learning in Today's Business- The first chapter of the book emphasizes the growing importance of machine learning in modern business. It explains how machine learning is used to automate business operations, make data-driven decisions, and improve customer experience. The chapter also provides an overview of the different types of machine learning algorithms and their applications, such as predictive analytics, natural language processing, and computer vision. It highlights the potential benefits of machine learning, including increased efficiency, reduced costs, and improved accuracy.

Chapter 2: Know your Data- The second chapter of the book focuses on the importance of data in machine learning. It explains the different types of data, including structured, unstructured, and semi-structured data. The chapter also introduces the concept of "dark data," which refers to the vast amounts of unutilized data that organizations possess. It highlights the importance of collecting and analyzing data to derive valuable insights that can drive business decisions. The chapter emphasizes the need for high-quality, trusted data to ensure accurate results in machine learning applications. It concludes by emphasizing the significance of data governance, which ensures that data is collected, managed, and used in an ethical and compliant manner.

Chapter 3: Up and Running with Analytical Tools- The third chapter of the book focuses on how to get started with analytical tools for addressing business issues quickly. It explains the different data analytics approaches that can be used and emphasizes the importance of predictive modeling. The chapter provides information on how to perform predictive modeling without any prior coding expertise and highlights how data cleansing and visualization can be automated using open-source and commercial solutions. The chapter also discusses various analytical tools, including Excel, Tableau, KNIME, Weka, Rapid Miner, Orange, and many others. It highlights how these tools can be used to analyze data and gain valuable insights, helping businesses to make informed decisions.

Chapter 4: Machine Learning in a nutshell- The fourth chapter of the book focuses on how machine learning can be used to solve business problems and anticipate future problems. The chapter emphasizes the importance of understanding the business problem well enough to select the appropriate data and machine learning algorithm that can help arrive at the right decision-making steps. It also highlights how anticipating problems before they occur can help businesses take corrective steps to eliminate or reduce their impact. The chapter emphasizes the interrelated nature of everything, from the business problem to the solution implementation process.

Chapter 5: Regression Analysis- The fifth chapter of the book focuses on different types of machine learning algorithms, with a specific emphasis on regression analysis. The chapter provides an overview of the various types of regression analysis, such as linear regression, logistic regression, and polynomial regression. It explains the nuances of regression analysis and provides insights into when to use regression analysis and what kind of business problems can be solved using it. The chapter also provides practical examples and use cases to illustrate how regression analysis can be used to solve real-world business problems.

Chapter 6: Classification- The sixth chapter of the book focuses on classification models in machine learning. The chapter provides insights into how to specify the input and output of a classification model and how to solve both binary and multiclass classification problems. It explains how a logistic regression model and a non-linear decision tree model can be implemented to solve classification problems. The chapter also covers several assessment criteria that can be used to evaluate the performance of classification models.

Chapter 7: Clustering and Association- Clustering and Association: The seventh chapter of the book focuses on clustering and association, which are widely used techniques for discovering unknown relationships in data. The chapter provides insights into how clustering and association can serve as a starting point for individuals with minimal or no understanding of the data. Clustering can help identify similar patterns and correlations among data points, like product recommendations based on purchase history on e-commerce websites. The chapter covers different clustering algorithms and provides examples of how they can be used to identify patterns and correlations in data.

Chapter 8:Time series Forecasting- The eighth chapter of the book focuses on time series forecasting, which is a commonly used technique for making scientifically backed predictions on a time stamp basis. The chapter explains that a time series is simply a list of events in chronological order, collected over a fixed interval of time. The dimension of time adds structure and constraint to the data, making it easier to analyze and predict future trends. The chapter covers different time series forecasting techniques, including moving average, exponential smoothing, and ARIMA models, and provides examples of how these techniques can be used to make accurate predictions.

Chapter 9: Image Analysis- In the ninth chapter of the book, the focus is on image analysis and how it enables us to extract useful data from photos through image processing and computer vision. The chapter explains how recent developments in machine learning and deep learning have made it possible to provide imaging data in near-real-time. The chapter also highlights the potential benefits of information extraction, which are far greater than most people realize. For example, image analysis has applications in healthcare, manufacturing, safety, and more, in addition to improving video surveillance.

Chapter 10: Tips and Tricks- The tenth chapter emphasizes the importance of understanding the details of data and data storytelling. Analytics can be challenging for some people, and presenting data in a narrative form can help make it more accessible to everyone. By using storytelling techniques to explain the key aspects of your analytics, you can engage your audience and help them better understand your findings. This can be especially important when presenting data to others who may not be as familiar with analytics.

Coloured Images

Please follow the link to download the
Coloured Images of the book:

https://rebrand.ly/sunkzyn

We have code bundles from our rich catalogue of books and videos available at **https://github.com/bpbpublications**. Check them out!

Errata

We take immense pride in our work at BPB Publications and follow best practices to ensure the accuracy of our content to provide with an indulging reading experience to our subscribers. Our readers are our mirrors, and we use their inputs to reflect and improve upon human errors, if any, that may have occurred during the publishing processes involved. To let us maintain the quality and help us reach out to any readers who might be having difficulties due to any unforeseen errors, please write to us at :

errata@bpbonline.com

Your support, suggestions and feedbacks are highly appreciated by the BPB Publications' Family.

Piracy

If you come across any illegal copies of our works in any form on the internet, we would be grateful if you would provide us with the location address or website name. Please contact us at **business@bpbonline.com** with a link to the material.

If you are interested in becoming an author

If there is a topic that you have expertise in, and you are interested in either writing or contributing to a book, please visit **www.bpbonline.com**. We have worked with thousands of developers and tech professionals, just like you, to help them share their insights with the global tech community. You can make a general application, apply for a specific hot topic that we are recruiting an author for, or submit your own idea.

Reviews

Please leave a review. Once you have read and used this book, why not leave a review on the site that you purchased it from? Potential readers can then see and use your unbiased opinion to make purchase decisions. We at BPB can understand what you think about our products, and our authors can see your feedback on their book. Thank you!

For more information about BPB, please visit **www.bpbonline.com**.

Join our book's Discord space

Join the book's Discord Workspace for Latest updates, Offers, Tech happenings around the world, New Release and Sessions with the Authors:

https://discord.bpbonline.com

Table of Contents

CHAPTER 1

Significance of Machine Learning in Today's Business

Let us be purpose-driven and empowered with data and ethics.

Everyone in today's fast-moving digital world wishes to be data-driven and wishes to create value. Do we really need to be data-driven, or can data even drive things? A million-dollar question, right? Yes indeed. Humans are emotional beings and always need a "just cause" or purpose, as often cited by leadership and management guru *Simon Sinek*. Data can unravel the mystery of whether we are progressing toward our cause but not where to go. As leaders, students, managers, and decision makers, it is imperative to use data to power our purpose. The issue of poor data analysis has plagued humanity for some time, but it has become increasingly apparent in the current era of widespread digital transformation and interconnectedness.. Everyone must know their way around data and be comfortable talking about it. This chapter will explain that the need for insights from data is stronger than ever before.

Structure

In this chapter, we will cover the following topics:

- Hype behind machine learning and data science
- Benefits of machine learning in business introducing data
- Types of data in business context
- Challenges with data
- Citizen data science
- Data science for leaders

Objectives

After studying this chapter, you should be able to relate how data plays a critical role in the business decisions we make. The chapter will also help you understand how machine learning is helping improve the decision-making process and how to utilize data to our advantage and make decisions based on data insights.

Hype behind machine learning and data science

Imagine the time you first learnt how to add two numbers in your Mathematics class in school. For some years, you would manually add up numbers till you had to use a calculator to perform the same task. Did it mean you could not do it manually any longer? No, it was because the process of calculation had to be quickened so that you could focus on the other critical steps of the problem.

In business too, we can take decisions based on experience and suggestions from experts we trust. However, is this the right step that would help us take faster and accurate decisions in the new normal world that is heavily data-driven? So, the question is, 'Do you want to spend your valuable time in tasks that can be automated, or do you want to spend time utilizing the data insights to make critical decisions? If you want to do the latter, then you made the right choice by getting this book.

We will help you make the best use of your time and effort to utilize data, wrangle it fast and then derive relevant insights from the data. Now let us walk through the history of machine learning and look at how it has evolved over the years. Its origin can be traced back to the 17th century, when people were trying to make sense of data and process to make quick decisions. A simple evolution chart depicting the machine learning journey is shown in *Figure 1.1*:

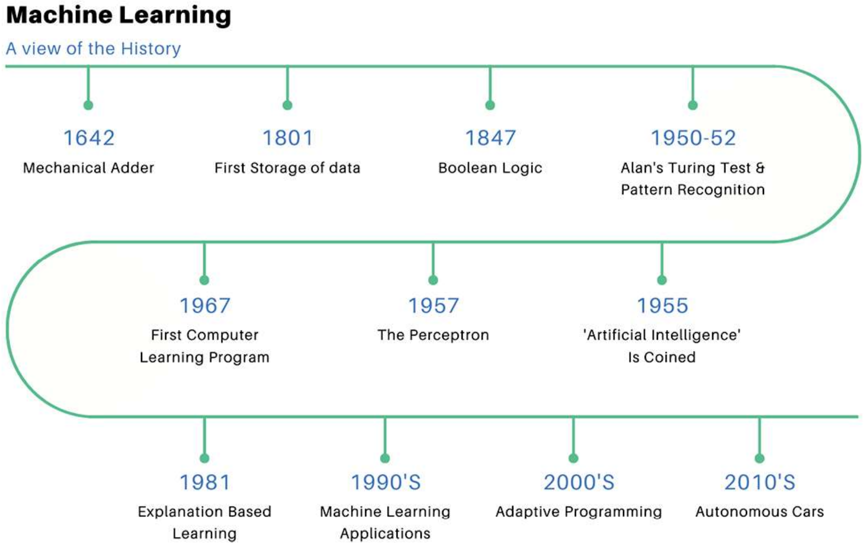


***Figure 1.1**: History of Machine Learning*

Blaise Pascal created the very first mechanical adding machine in 1642. Next, the data storage challenge was overcome using a weaving loom to store data by Joseph Marie Jacquard. Over time, we developed concepts like Boolean logic, statistical calculations, and the Turing test to evaluate whether a computer had intelligence, and eventually, the phrase **artificial intelligence** (**AI**). After a series of other inventions, recent years have seen advancements in machine learning algorithms. Today, three major innovations, as listed here, have fuelled the recent buzz and helped companies and individuals use and experiment machine learning technologies. In a nutshell, they have democratized machine learning to all:

- **Higher volume of data and cheap storage**: Business-critical applications are producing and storing more data than ever before, thanks to cloud-based tools and the decreasing cost of storing data through services like Google Cloud Storage, Amazon Redshift, Microsoft Azure Services, and others. Most of these tools are highly intuitive and user friendly, with easy-to-use click and move features that simplify your work process incredibly.

- **Open-source:** Open-source machine learning libraries, such as scikit-learn, Google's TensorFlow and Orange, make cutting-edge algorithms more usable and accessible to a larger community of data scientists and engineers.
- **Greater computing power:** With the advent of cloud-based technologies and custom hardware designed for machine learning, these systems can now run faster and at a lower cost, making them more suitable for a wide range of business needs.

Consider machine learning in this light. You, as a person and as a user of technology, carry out such actions, which allow you to make a decisive judgement and classify something. Machine learning has advanced to the point that it can mimic the pattern-matching ability of human brains. Algorithms are now used to teach machines how to recognise features of an object.

To provide just one example, a computer may be shown a cricket ball and instructed to treat it as such. The programme then uses the data to identify the different characteristics of a cricket ball, each time adding new data to the mix. Initially, a machine could identify a cricket ball as round and construct a model that states that everything round is a cricket ball. The programme then discovers that if anything is round and red, it is a cricket ball, when a red colour ball is added later. Then, a reddish-brown colour ball is introduced, and so on.

The machine must update its model as new knowledge becomes available and assign a predictive value to each model, indicating the degree of certainty that an entity is one item over another. Here, predictive value refers to the probability of identifying the ball correctly as cricket or tennis ball. As you can see in *Figure 1.2*, a machine learns the information provided to it by the user, executes certain action and receives feedback for it, and the learning continues as a feedback loop. The learning step is called as "model training", and the feedback loop is called "model retraining".

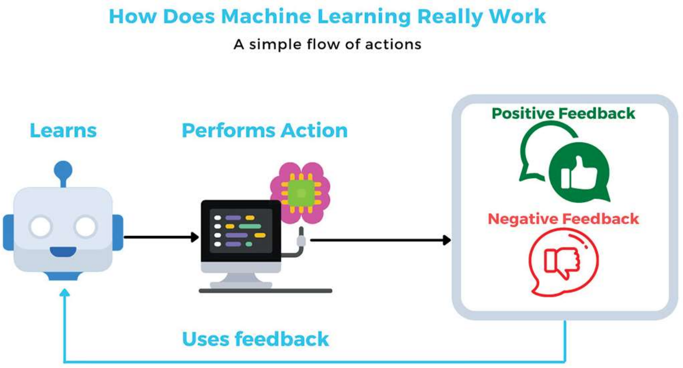
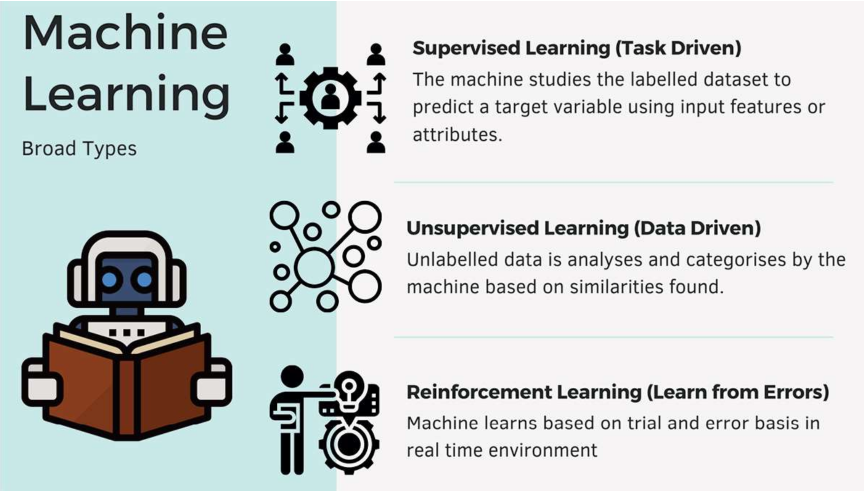


Figure 1.2: *How does machine learning work?*

The following figure gives an overview of the most common types of machine learning techniques available today: supervised learning, unsupervised learning, and reinforcement learning:



Supervised Learning (Task Driven)
The machine studies the labelled dataset to predict a target variable using input features or attributes.

Unsupervised Learning (Data Driven)
Unlabelled data is analyses and categorises by the machine based on similarities found.

Reinforcement Learning (Learn from Errors)
Machine learns based on trial and error basis in real time environment

Figure 1.3: *Types of Machine Learning*

Supervised Learning

When an algorithm is trained to predict an output (also called a label or target) from one or more inputs (also called features or predictors), we say that the algorithm is engaging in supervised learning. As the name implies, "supervised" learning occurs when the algorithm is given labelled training instances that consist of input-output pairs.