Optimizing AI and Machine Learning Solutions

Your ultimate guide to building high-impact ML/AI solutions

Mirza Rahim Baig

www.bpbonline.com

First Edition 2024

ISBN: 978-93-55519-818

LIMITS OF LIABILITY AND DISCLAIMER OF WARRANTY

Dedicated to

My lovely wife

About the Author

Since 2009, **Mirza Rahim Baig** has been exploring, practicing, learning, and teaching all things machine learning / artificial intelligence. He is a seasoned data science expert and renowned thought leader. Rahim is adept at solving complex business problems using AI / ML in a career spanning multiple domains and geographies. Numerous job titles aside, his focus has always been using data science to solve business problems and create high impact.

Rahim is also an author, a speaker and educator. In addition to this book, he has authored two well received books *The Deep Learning Workshop* and *Data Science for Marketing Analytics*. In all his titles, Rahim focuses on application and outcomes, and shares best practices derived from his experience. A key feature of his work is the approach of focusing not merely on model building but on the entire end-to-end solution building process. The explanations are easy to follow, even for non-technical learners, thanks to his style developed over years of teaching.

Rahim is a renowned subject matter expert in data science topics, having worked with most of the popular ed-tech platforms in designing courses and delivering content for master's level programs. He is a visiting faculty at NMIMS for MBA programs. In addition to having been published multiple times, Rahim speaks on various platforms about data science.

About the Reviewer

Sujit Pal is a Technology Research Director at Elsevier Health Markets, where he works as an applied researcher at the intersection of search and Machine Learning. His interests are in Information Retrieval, Natural Language Processing, Knowledge Graphs and Machine Learning, and more recently, Large Language Models and Generative AI. He has co-authored two books on Deep Learning and one on Graph-based Natural Language Processing.

Acknowledgement

In writing this book, my aim has been to benefit others with the little knowledge I have. However, it would be insincere to claim that all the insights and knowledge contained in these chapters originated solely from me. A lot comes from my incredible mentors and excellent teachers. What you find here are also their tremendous efforts reaching you through my words.

I would like to acknowledge the achievements of the open-source community. In the past few years, this community has democratized artificial intelligence and machine learning like never before, making state-of-the-art models accessible to anyone with an internet connection. We are truly witnessing a revolution and I am genuinely grateful for it.

I am also grateful to the team at BPB Publications for their guidance and patience throughout the journey of writing this book. Working with some extremely competent editors, reviewers, and technical experts was an enriching experience.

Finally, I would like to thank all the readers who have taken an interest in my book and for their support in making it a reality. Your encouragement has been invaluable.

Preface

Building High-impact Machine Learning / Artificial Intelligence solutions is intended for several types of readers:

- Data science developers (data analysts / scientists) who will create the data science solutions using data,
- Data science professionals who manage data science teams or data science projects,
- Aspiring data scientists who want to learn more than just the techniques,
- Anyone interested in learning the hows of building great solutions and the whys in making the multiple choices in data science solution building.

A key difference between this book and the others in the market is that this book provides a holistic view on data science problem solving. Throughout, the book keeps reminding that business impact and utility are paramount and reminds the learner to be pragmatic in their approach. The book approaches data science solution building using a principled framework. Machine learning model building is one of the many steps in the framework, and this book treats it accordingly. The book will teach what optimization means at each step. Whether it is problem formulation or hyper-parameter tuning for deep learning models, this book provides frameworks and case studies, with extensive hands-on wherever possible.

This book is not focused on the algorithms for building solutions and is not a replacement for algorithm focused books. The book will however provide you enough intuition, functional information, and references to help you make even complex deep learning models. Using this book you will create complex models without implementing the detailed mathematical operations yourself. Instead, you will use abstractions and helpful libraries wherever possible to create clean and simple code. In the later chapters, the book will even teach you how to use transfer learning for using readymade ML and AI models as a part of your solutions. You do not need extensive coding experience for machine learning, the book will guide you step-by-step and provide code and explanations.

With this book, you will be able to approach a ML / AI solution in a systematic way, optimizing each step of the solution building process. You will thus be able to create high impact, high value machine learning / artificial intelligence solutions – which is the ultimate goal of every data science professional.

Chapter 1: Optimizing a Machine Learning /Artificial Intelligence Solution – introduces some core concepts and sets up the foundation for our journey together in this book. It provides an overview of machine learning, followed by addressing the various practical challenges in machine learning. It introduces some key ideas which will be expanded on in the later chapters. It is crucial to distinguish between simply making a model and carefully designing an end-to-end solution to the business problem. A framework to approach such end-to-end solutions is introduced and the chapter will point to the chapters in the book that help you optimize the solution at each step highlighted in the framework, to ultimately develop a truly optimized machine learning / artificial intelligence solution.

Chapter 2: ML Problem Formulation: Setting the Right Objective – discusses perhaps the most important step in any data science project that involves machine learning problem formulation. A problem can be formulated in various ways employing different solutions: machine learning or not. Further, there are multiple possible ML approaches. Making these decisions is not a trivial matter. The chapter also highlights the important of aligning the model objective with the business objective for achieving maximum impact.

Chapter 3: Data Collection and Pre-processing – addresses a basic reality in the ML / AI space – that the sophistication level of the technique is irrelevant if the input data is poor. The chapter explains the various factors that influence the data collection process and how you should tackle associated problems. A principled approach is introduced that helps you optimize the entire pre-processing process as a pipeline, with the help of a case study.

Chapter 4: Model Evaluation and Debugging – teaches model evaluation in the context of solving business problems. Learn various considerations that make a model better than another. The chapter introduces metrics for model evaluation, classification, and regression problems. Learn a principled way to diagnose and debug models to identify the cause of poor performance, along with approaches to handle underfitting and overfitting.

Chapter 5: Imbalanced Machine Learning – shows how to tackle one of the most common and challenging problems in machine learning / artificial intelligence. The chapter explains how imbalance impacts model evaluation and teaches how to diagnose issues due to imbalance. Apart from introducing better metrics for evaluation, it also introduces multiple approaches to tackle imbalance. Learn using several types of under / over sampling techniques to balance the dataset better for a more optimized machine learning solution.

Chapter 6: Hyper-parameter Tuning – is dedicated to the process of finding the best set of *settings* for the model, i.e., hyper-parameters, to get the best generalized performance. The chapter covers the key considerations in defining *best* and proceeds to demonstrate multiple automated approaches to searching for the best hyper-parameters – from exhaustive methods like grid search to approaches like Randomized search.

Chapter 7: Parameter Optimization Algorithms – provides a good understanding of mathematical optimization, the core of machine learning/artificial intelligence models. The chapter contextualizes optimization for machine learning and discusses aspects that make optimization for machine learning unique. It then teaches a general, principled approach to solving optimization problems for classical machine learning and explains the variations developed for deep neural networks. The chapter helps you make important choices in optimization, like loss functions and optimization algorithms, to maximize the performance of the model.

Chapter 8: Optimizing Deep Learning Models – focuses on optimizing neural networks. Like with classical machine learning, hyper-parameter tuning is a key step for optimizing deep neural networks as well. The chapter focusses on the simple fully connected network architecture for these concepts and the accompanying case studies. In addition to automated hyper-parameter tuning for deep neural networks, the chapter will introduce several techniques to improve model performance, like data addition, data augmentation, ensembling, regularization, and injecting your domain knowledge via manual features.

Chapter 9: Optimizing Image Models – dives deeper by teaching optimization of deep learning models for image tasks. This chapter teaches concepts, architectures and parameters that are unique to Image models and then demonstrates how to optimize these using TensorFlow 2.0. The chapter discusses the design approach used in various popular model approaches and teaches you to implement your own versions of VGG and ResNet models. It teaches how to effective use regularization using multiple approaches - dropout, data augmentation, and the powerful batch normalization technique, ending with some general guidelines for designing deep learning models for image tasks.

Chapter 10: Optimizing Natural Language Processin Models – continues the deep learning deep dive by focusing on NLP models. The chapter introduces the peculiarities of natural language data and teaches handling them by appropriate text pre-processing and data representation (using embeddings). The chapter details the recurrent and architecture and the transformer architecture with hands on case studies. In addition to hyper-parameter tuning for NLP models, the chapter shares tricks like using 1D convolutions and pre-trained embeddings (a prelude to transfer learning).

Chapter 11: Transfer Learning – teaches you how to stand on the shoulders of giants and benefit from the excellent work done by the ML/AI community. The chapter explains the different types of transfer learning and how they are helpful in modern machine learning/artificial intelligence, highlighting both the benefits and limitations of it. The chapter demonstrates the usage of a SOTA model out of the box and eventually how to fine tune it for better performance. Transfer learning is slightly different for image processing and NLP tasks. The chapter will employ multiple SOTA models for both image and NLP tasks, employing popular libraries.

Code Bundle and Coloured Images

Please follow the link to download the
Code Bundle and the ***Coloured Images*** of the book:

https://rebrand.ly/zce3wew

The code bundle for the book is also hosted on GitHub at

https://github.com/bpbpublications/Optimizing-AI-and-Machine-Learning-Solutions.

In case there's an update to the code, it will be updated on the existing GitHub repository.

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
Optimizing a Machine Learning /Artificial Intelligence Solution

Introduction

This chapter will provide an overview of **Machine Learning** (**ML**), followed by addressing the various practical challenges in machine learning. This chapter will introduce some key ideas which will be expanded on in the later chapters. We will make the crucial distinction between simply making a model and carefully designing an end-to-end solution to the business problem. We will learn about a framework to approach such end-to-end solutions learn what it means to optimize at each step, and ultimately develop a truly optimized machine learning/artificial intelligence solution.

Structure

In this chapter, we will cover the following topics:

- Case study
- Understanding machine learning
- Machine learning styles
- Challenges in ML/AI
    - o Poor formulation
    - o Invalid assumptions

- o Data availability and hygiene
- o Representative data (lack of)
- o Model scalability
- o Infeasible consumption
- o Misalignment with business objectives
- ML/AI models vs. end-to-end solutions
- CRISP-DM framework for solution development
- Optimization at each step of solution development
  - o Business understanding
  - o Data understanding
  - o Data preparation
  - o Model building
  - o Evaluation
  - o Deployment
- Conclusion

Objectives

In this chapter, we will take a good, holistic look at the field of machine learning. This chapter will introduce some key ideas which will be expanded on in the later chapters. We will make the crucial distinction between simply making a model and carefully designing an end-to-end solution to the business problem. We will learn about a framework to approach such end-to-end solutions and learn what it means to optimize at each step, and ultimately develop a truly optimized machine learning/artificial intelligence solution. The various examples and case studies in this chapter will make the ideas concrete.

Consider this chapter as the gateway – where you get an overview of the steps in creating high-impact, optimized machine learning/artificial intelligence solutions. Each of the steps/ideas we discuss in this chapter will be dealt with in detail in the chapters that follow.

Case study: Text deduplication for online fashion

Consider a data scientist working at the online fashion giant Azra Inc., to make the product detail page most helpful to the shopper. The product page contains detailed information about the product, including ratings, reviews, and questions that users ask about the product. The user questions section is of particular concern. There is a severe duplication of questions. Users are asking the same question with minor variations in language. For example, *Is the material durable*? can be considered a duplicate of *Is the shirt durable*? Due to this, a few common questions are suppressing the visibility of other useful questions and their answers, withholding useful information from users, and affecting product sales. The task for the data scientist is to use their ML / AI expertise to identify duplicate questions as shown in *Figure 1.1*:

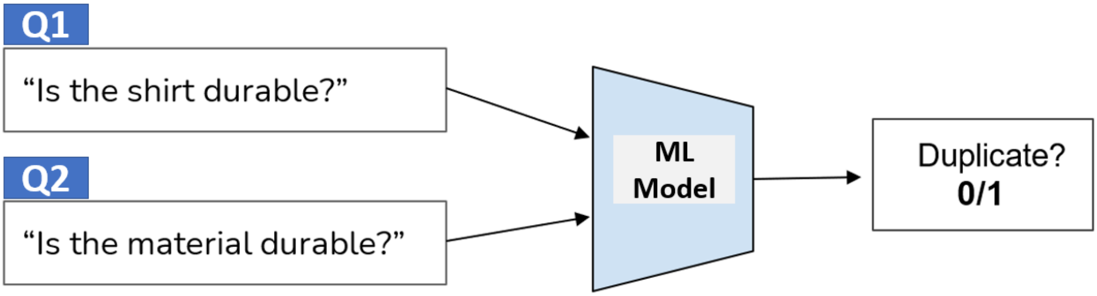

Figure 1.1*: Deduplication using supervised classification*

The data scientist formulates this as a supervised classification problem, as illustrated in *Figure 1.1*, using a deep learning model for text classification. This makes intuitive sense as we expect deep learning methods to shine in such situations. Using the latest transformer architecture should solve this, right? Unfortunately, in this case, the project was stopped after about 3 months of effort. The reason was a lack of sufficient labeled data.

For the text deduplication task, using a transformer architecture would require at least a few thousand pairs of questions labeled. The problem is labeling tens of thousands of question pairs. Manual labeling would take time and solid guidelines for the labelers so that their labels are in agreement. This is an expensive and time-consuming approach. This logical approach failed because of a presumption of data availability. The solution that eventually worked used an unsupervised clustering approach. The lesson is that improper problem formulation and presumptions can spell disaster for a ML / AI solution.

For success in ML / AI solutions, there are various considerations and decisions at various steps that need to be optimized; this is why data science projects fail. Before we discuss those, let us take a step back. Let us establish the understanding of machine learning that we will employ throughout this book. It is imperative that we take a holistic look at what ML is and more importantly, what it is good for, and how to make it work.

Understanding machine learning

Our modern, data-driven world seeks to make decisions based on data insights to increasingly employ machines to perform repetitive tasks, drive cars, diagnose patients, allocate ads, recommend connections and songs, summarize news, etc. If data is oil in this new world, machine learning is the closest thing we have to the engine of this machinery.

Machine learning is the process that makes it possible to learn patterns from the provided data. The patterns learned by the machines can then be used to make some estimations/ predictions.

The outcome of the pattern learning process is often a mathematical model, capturing how the output relates to the input. The process of learning is also often referred to as model building or data mining. *Figure 1.2* illustrates this process. The historical data is input into the data mining/ model building process. The model-building process learns the patterns, which are expressed as a machine learning model. This model captures the relation between the inputs and the output and can therefore be employed to make estimations/ predictions. Depending on the technique employed, the model could be simpler and easily interpretable (e.g., a simple decision tree, or a linear regression equation), or a complex, hard-to-interpret from a deep neural network (complex series of matrix multiplications) that requires additional effort in post hoc explanations.

Note: Data mining is a broader term that encapsulates the entire process of building a machine learning solution from data, the end-to-end process. Model building is merely one part of this process. We will discuss this in detail later in the chapter in the section titled CRISP-DM Framework.

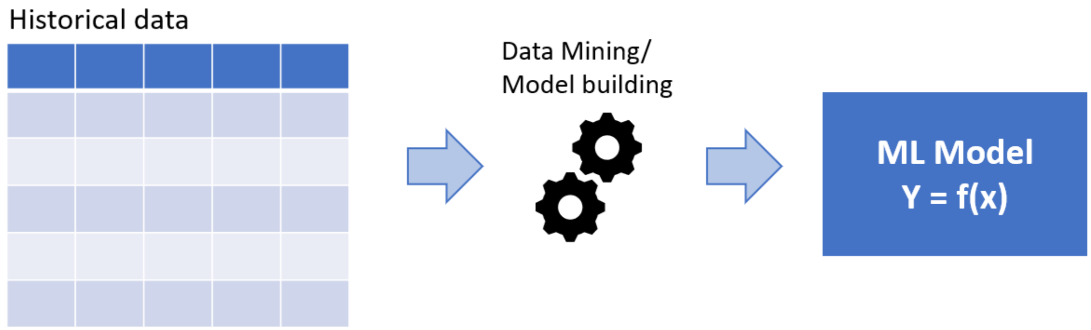

Figure 1.2*: Machines learning patterns from data*

Machine learning styles

Let us now learn how to make machines learn the relevant patterns. We will have to make several decisions. For instance, we need to decide if we want to provide feedback and if yes, how to do that. We must define the kind of estimations that the machine needs to make. We need to define whether the model would be used to make predictions for the future or to uncover some patterns to aid human decision-making. Also important is to define the kind of data we input into the model. The specific solution depends on these considerations, but over the decades we have arrived at broadly three different machine learning styles, as illustrated in *Figure 1.3*:

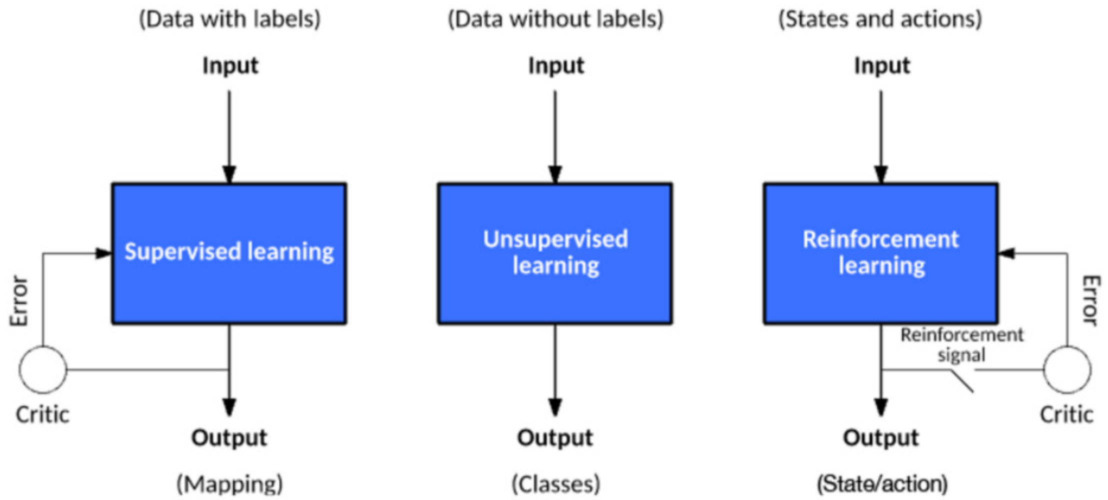

Figure 1.3*: Machine learning styles*

Supervised machine learning

A key feature of supervised machine learning is that the data provided to the model contains the target as well. The input data comprises the features and the target, as shown in *Figure 1.4*. The features contain the information that will be used to predict the target. For using the model in the future, the input features will be available to us and will be used to predict the target. The machine learning process learns to predict the target using the input features, as illustrated in *Figure 1.4*:

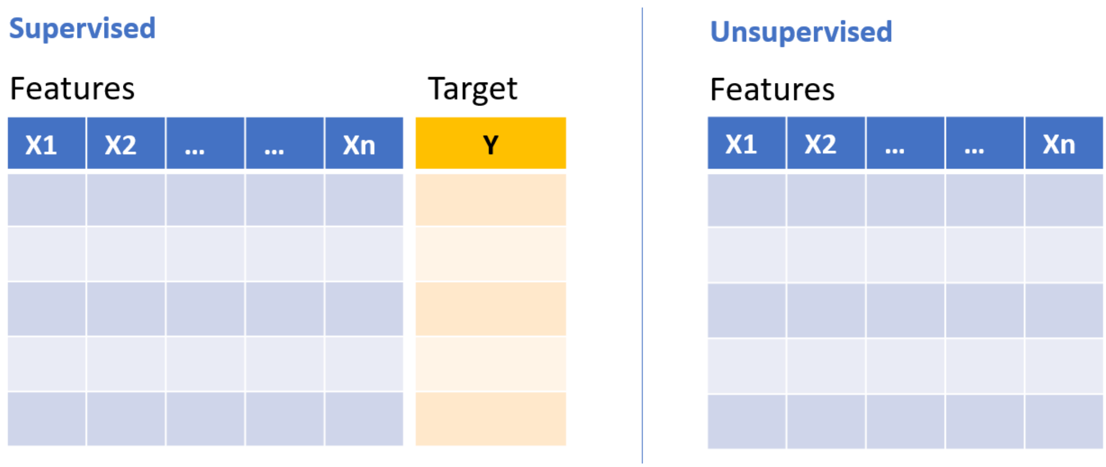

Figure 1.4*: Input data for supervised vs. unsupervised ML*

Figure 1.4 illustrates how the supervision is done. The modeling technique sees the true target values (often called labels) and their predictions, compares them using a notion