DATA SCIENCE FOR ENTERPRISES

Deployment and beyond

Get an overview of machine learning fundamentals
Delve into the depths of machine learning platform design
Build and manage data science teams at any scale
Create and nurture large scale data science programs

Mr. Chandim Sett

Ms. Samadrita Ghosh

Dr. Srinivas Telukunta

DATA SCIENCES FOR ENTERPRISES: DEPLOYMENT AND BEYOND

First Edition

Mr. Chandim Sett,
Ms. Samadrita Ghosh,
Dr. Srinivas Telukunta

Data Sciences for enterprises: Deployment and beyond, 1e

- Chandim sett,
- Samadrita Ghosh,
- Dr. Srinivas Telukunta

First published in India 2020, by Amazon Publishing Services

ISBN-16: 87493-346-3983-9
Printed in India

This publication is designed to provide information and guidance for enabling data science teams to deploy production-grade machine learning models within enterprises. It is sold under the express understanding that any decisions or actions you take as a result of reading this book must be based on your independent judgment and will be at your sole risk. The author(s) or publishers will not be held responsible for the consequences of any actions and/or decisions taken as a result of any information given or recommendations made in this publication.

Introduction

Machine learning is increasingly becoming the crux of not only the digital world but of all major industries that generate any sort of data. Data has become the new gold and one cannot deny that data is generated even by the most remote and archaic systems. With suitable data, we can teach our machines to learn historical patterns and come up with smart insights that can foster decision making and innovation. The power of machine learning and data science is so immense that even in its preliminary stages on this day, it has helped businesses rocket sky-high.

It is anticipated that the businesses and industries of today that are in denial regarding the power of data will be incurring heavy losses due to incompetence in future markets. These elements in the market today are being compared to the businesses and industries back in the late 20th century which were in complete denial of the power of the Internet to transform the markets. However, data is restructuring our world heavily and so much so, that we might just be at the tipping point of another great revolution fostered by mankind.

With the aid of machine learning, the base architecture of software design is going through a grand makeover. Rule-based and hard-coded techniques are taking a back seat since the business trends are increasingly witnessing a dynamic landscape and rule-based methods are turning obsolete in no time. Machine learning algorithms, with highly advanced statistical foundations, ensure that the data trends are captured and change with time and suitable action items are assigned to the business elements.

The value of machine learning is in the fact that it allows us to predict the future by learning continuously from incoming data. Industries across the world are trying to optimize their processes and obtain insights into trends and anomalous behavior within their data. However, machine learning techniques cannot persist on their own; a well-equipped team comprising of data scientists, data engineers, business analysts, and business leaders are imperative for a successful end to end implementation of the machine learning solutions. Without such collaboration, the solutions, however efficient they may be, cannot reach the end customer.

About this book

Data Sciences for enterprises: deployment and beyond, 1e, will not only walk you through an end to end machine learning pipeline which involves an overview of data management techniques, machine learning algorithms, and deployment methods but also cover advanced enterprise governance aspects such as project management, program management, machine learning development process management, enterprise compliance adherence, etc. It is thus, a handbook for business leaders and also data scientists who have just arrived into the realm of studying data and business but aspire to go farther in the domain of enterprise data science management. The book is designed to not only improve the reader's skill assets but also to help them enable teams using our suggested end-to-end pipeline methodologies and suggestions.

What this book covers

Data Science for Enterprises: deployment and beyond, will walk you through an end to end machine learning pipeline which involves an overview of data management techniques, machine learning algorithms, and deployment methods.

Chapter 1: Introduction to Data Science

This chapter discusses how raw data can be effectively processed and modeled on, such that learning is maximal and effective. Different aspects of data processing will be discussed. A brief walkthrough of the modeling algorithms such as regression and classification techniques is made along with how to measure the health of these models.

Chapter 2: Machine Learning in Practice

This chapter walks you through the various problems which can be solved with machine learning and delves into the mechanics of implementing machine learning using "pipelines". The pipelines cover various aspects of machine learning such as exploratory data analysis, feature engineering, feature selection, model evaluation, deployment, etc.

Chapter 3: Scale and Speed, Need for Data Sciences as Software Engineering

This chapter describes different aspects of the deployment of the machine learning models and discusses the various formats for storing machine learning models. Then various software development practices are presented which need to be adopted by machine learning teams such as automation, version control, monitoring and logging, package management, documentation, etc. which are essential for enterprise-scale deployment of machine learning models.

Chapter 4: Machine Learning System Design

This chapter discusses how a machine learning system can be developed from scratch in an enterprise setting. Firstly, it addresses the challenges that are faced by enterprises that use legacy systems. A framework in python is built to integrate with Java via a command-line interface (CLI). Then the microservices architecture is discussed along with the use of containerization to build a real-time prediction engine for short-running jobs. The CLI framework and real-time prediction engine, as a singular unified microservice called Machine Learning as a Service, is described. Besides the frameworks, the important design considerations, logging, version control, and dependency management are also discussed.

Chapter 5: Enterprise Deployment Strategies

This chapter discusses the various strategic operations and considerations that go into enterprise deployments. The ideal characteristics of a good deployment workflow and the various associated responsibilities along with it are described. Various components for an ideal cloud deployment like load balancing, service discovery, auto-scaling and rolling updates are discussed. Various deployment features for two popular orchestration systems such as Docker Swarm and Kubernetes are described.

Chapter 6: Engineering a machine learning workflow management system

This chapter focuses on the empowerment and operationalization of a Data Science team. The aspects such as designing and developing a cloud-native workflow management system using open source software and Docker containers are covered. You will get an opportunity to implement the Housing Price Prediction problem in MLaaS and ML-Cloud tool in an enterprise standard, which we had discussed in Chapter 3.

Chapter 7: Enterprise governance considerations for Data Sciences

This chapter delves into governance aspects of managing such large-scale data science programs including project management, program management, dashboard design, machine learning process management, ethical aspects, maintaining audit standards for enterprise compliance, etc.

Who is this book for

This book is a handbook for business leaders and budding data scientists who have just arrived in the realm of studying data and business. The book is designed to not only improve the reader's skill assets but also to help them enable teams using our suggested end-to-end pipeline methodologies.

Dear Reader,

Thank you for choosing *Data Sciences for enterprises: Deployment and beyond, First Edition*. This book is part of our effort to democratize machine learning and make it accessible to enterprises around the world. As you are aware, machine learning has now become an integral part of most enterprise-level applications but still, there is no dearth of challenges in implementing these programs at a large scale. While there are many textbooks in the market which address the data sciences from a largely theoretical and isolated point of view, not much is covered in the development of platforms for deploying machine learning and this book aims to address this gap.

The authors have many years of experience in enabling machine learning applications for more than 200 of the fortune 500 companies including large scale enterprises such as Walmart, Coke, Bank of America, etc., in the arena of accounts receivables and payables within the constraints of stringent compliance requirements. The authors have tried to leverage their experience working with such large enterprises in producing this publication. With this new title, we are working very hard to set a new standard for data science as a discipline. We hope you see all of this reflected in the following pages.

The authors also want to convey their sincere gratitude to Mr. Ajitesh Shukla, AVP Technology at HighRadius for going over the manuscript and suggesting valuable suggestions which have been incorporated in this work and also for writing the preface to this book. In addition, we also want to convey our deep appreciation to the AI industry leaders Mr. Anjan Purandare, CEO, Ivyclique Technologies; Lt Cdr. (Dr) Ashvini Jakhar, CEO, Prozo Technologies and Mr. Piyush Shah, Co-Founder, InMobi Group & CEO at TruFactor for their review of the work and endorsement.

Reader feedback

Feedback from our readers is always welcome. Let us know what you think about this book - what you liked or disliked. Your feedback is important for us as it will help us improve the upcoming titles. To send us general feedback, simply email your feedback to datasciencebook@gmail.com.

Errata

Although we have taken every care to ensure the accuracy of our content, mistakes do happen. If you find a mistake in either text or code, we would be grateful if you could report this to us by sending us an email to engineer.datascience@gmail.com with errata as to the subject line.

Questions

If you have questions with any aspect of this book, you can contact us at engineer.datascience@gmail.com, and we will do our best to address the problem.

Best Regards,
Chandim Sett,
Samadrita Ghosh,
Dr. Srinivas Telukunta

About the authors

Mr. Chandim sett

Mr. Chandim Sett works as a leading machine learning software architect for HighRadius, a pioneer in the area of SaaS-based products for the Financial Supply Chain Industry. He is a Computer Science graduate from Kalinga Institute of Industrial Technology, Bhubaneswar and has served as the key person in the deployment of machine learning pipelines for large scale enterprise applications. His areas of expertise include enablement of enterprise applications for Data Sciences, DevOps implementation, and Infrastructure automation.

Chandim has empowered the Data sciences team at HighRadius by building a strong foundation in Data Sciences and Software Engineering in the company and by mentoring his peers. He has played a key role in designing and developing machine learning frameworks for deployments and monitoring by adapting the enterprise legacy systems at HighRadius and pushing towards modern cutting-edge technologies. He is an expert in designing and developing enterprise applications and microservices especially for Data Sciences and Machine Learning implementations at scale.

Ms. Samadrita Ghosh

Ms. Samadrita Ghosh works as a data scientist in a key role in solving various machine learning problems in the FinTech sector related to accounts receivables such as collections, deductions, cash application, etc. She is a Computer Science graduate from Kalinga Institute of Industrial Technology, Bhubaneswar and has served as the key person in the development of machine learning models for various collection related problems.

She has served as a technical author in Data Science for various B2B and Educational websites and has also served as the Leading Professor's aid for PGP-AI/ML program in BITS Pilani and has enormous experience in guiding new talent in the AI/ML sector.

Dr. Srinivas Telukunta

Dr. Srinivas Telukunta has rich experience in Product Portfolio Project management (PPM), Product Development, Analytics & Machine Learning with Big Data suited for enterprise-scale deployment. He is proficient in Technology consulting with multi-disciplinary experience across several industries and domains including Semi-conductor, Pharmaceutical, Finance and IT by virtue of corporate work experience in these domains.

He has a distinguished educational background with an MS, Ph.D. (Cornell University), MBA (Indian School of Business), B-Tech (IIT Madras), LLB (Equal to J.D), PG in Public Accounting (Bharathiar University), Specialization in Data Science from University of Michigan and more than 50+ certifications in diverse domains (Project Management, Information Technology, Cloud Automation). He has wide consulting experience with 25+ global companies including Almarai (Saudi Arabia), ADP, CYIENT, Kantar Operations, Qualcomm, Dr. Reddy's Laboratories.

Preface

This book covers topics such as data management techniques, building models using different machine learning (ML) algorithms and deployment methods, in relation to building end-to-end machine learning pipeline. Training ML models in a standalone environment for research or academic purpose is very different than training models and moving them into production such that products across enterprise could invoke them for predictions. In the enterprise setting, there is a need for data science, product engineering and IT teams to collaborate with each other for understanding the problem, gather data from different product data sources, train and test different models, move the best model in production and finally monitor the models at regular intervals. All this requires one to have a good understanding of different aspects of the ML model development lifecycle. This is where this book will come handy for different stakeholders including data scientists, product managers, IT staff and business leaders.

The following are some of the key challenges that will be addressed in this book:

- What are the different kinds of problems which could be solved using ML algorithms?
- How to go about building ML models?
- How to design an ML system that scales?
- What are the software engineering practices to be adopted for ensuring quality models?
- How to operationalize ML models across different environments including production?
- What are some of the best practices and standard guidelines for monitoring ML models?

Here are some of the key topics covered in this book:

- Different aspects of building ML models including exploratory data analysis (EDA), feature engineering, model training and selection
- Different aspects of software engineering including planning and tracking ML model development, version control, build management, etc.
- Designing ML systems which help in deploying ML models
- Different ML models deployment strategies

- ML models governance best practices and standard guidelines
- ML models as AutoML services

Mr. Ajitesh Kumar Shukla,
Assistant Vice President, Technology
HighRadius Inc

Table of Contents

List of Figures

MACHINE LEARNING IS THE FUTURE OF EVERY ENTERPRISE

Machine learning is now becoming pervasive in every enterprise that generates any sort of data. Data has become the new gold and one cannot deny that data is generated even by the most remote and archaic systems. With suitable data, we can teach our machines to learn historical patterns and come up with smart insights which can foster better decision making and innovation. The power of machine learning and data science is so immense that even in its preliminary stages on this day, it has helped businesses rocket sky-high. In light of this, it is no wonder that machine learning is getting embedded in every application within enterprises these days so as to boost collections, create smart checklists, forecast demand, improve supply chain efficiency, treasury cash forecasting etc. This book not only covers the fundamentals of machine learning, but also presents design patterns for building robust machine learning platforms and nurturing machine learning programs and teams with rigorous enterprise governance practices.

THIS BOOK WILL TEACH YOU

- Fundamentals of machine learning
- The business imperative of machine learning
- Deployment strategies and machine learning platform design
- Enterprise governance for machine learning programs

Pothi.com

WHAT AI LEADERS ARE SAYING ABOUT THIS BOOK

Anjan Purandare,
CEO, Ivyclique Technologies

Data Science for Enterprises: Deployment, and beyond is the definitive new age bible for all stakeholders in the enterprise ecosystem. The book is lucid yet detailed and systematically elucidates a clear roadmap to developing and effectively deploying machine learning based AI programs within the enterprise milieu.

Among various standout features, the modelling phase of machine learning cited in the book, leading up to the evaluation stage, and ultimately through to the detailed description of the deployment of the model via API's, is a treasure trove, bound to benefit all decision makers in the enterprise AI space. All in all, an absolute must read – highly recommended!"

Lt Cdr. (Dr) Ashvini Jakhar,
CEO, Prozo Technologies

"Data Science for Enterprises: Deployment and beyond" is a must-read primer for anyone considering Data Sciences as a career. Read this book - and learn from some of the best.

The book not only takes the reader through the journey of developing machine learning models at a swift pace but also takes the reader into a much bigger world of managing the modelling process and looking into practical aspects of deploying them at a large scale all while addressing the enterprise restraints expected out of a large organization.

Shiva Dhawan
Founder and CEO, AttentiveAI

This is an essential book for machine learning-based startups, managers and senior data science engineers at all levels to gain a deep understanding of data science program in relation to the enterprise setting. This book provides a holistic view of the data science initiatives in a larger enterprise setting covering various aspects of critical importance. This book not only teaches the core machine learning principles such as feature engineering, feature selection but also helps discover various aspects of managing large scale data science initiatives such as stakeholder management and novel techniques of performance assessment for data science teams.