

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

© 2020 Amazon Publishing Services

First published in India 2020, by Amazon Publishing Services

This copy is for sale only in India, Pakistan, Bangladesh, Nepal, and Sri Lanka.

ALL RIGHT RESERVED. No part of this publication may be reproduced or distributed in any form or by any means or stored in a database or retrieval system, without the prior written permission of the author.

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

Introduction to Data Science

- 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

Chapter 1	Introduction to Data Science	2
1.1	Basic Concepts of Data Science.....	7
1.1.1	Data Gathering.....	10
1.1.2	Data Exploration and Processing	12
1.1.3	Feature Engineering.....	19
1.1.4	Feature Selection	29
1.1.5	Machine Learning Modelling	32
1.1.6	Model Tuning	51
1.1.7	Loss Functions	54
1.1.8	Optimization Algorithms.....	56
1.1.9	Performance Measures.....	57
1.1.10	Offline Evaluation Mechanisms	59
1.2	Machine Learning Tools and Platforms.....	62
1.2.1	Platforms.....	62
1.2.2	Libraries.....	65
Chapter 2	Machine Learning in Practice.....	69
2.1	Machine Learning Stages: Business Understanding	72
2.2	Machine Learning Stages: Collect Data and EDA	73
2.3	Machine Learning Stages: Data Preparation.....	78
2.3.1	Removal of duplicate columns.....	78
2.3.2	Removal of columns with constant or quasi-constant values	79
2.3.3	Missing values	80
2.3.4	Analysis of Numerical variables	84
2.3.5	Analysis of outliers	89
2.3.6	Analysis of Categorical variables.....	90
2.3.7	Analysis of relationships between variables and the target	91
2.4	Machine Learning Stages: Feature Engineering.....	91
2.4.1	Missing Value Feature Engineering	93
2.4.2	Nominal Data: Feature Engineering.....	95
2.4.3	Ordinal Data: Feature Engineering	97
2.4.4	Interval Data: Feature Engineering.....	97
2.4.5	Transformation of non-normal data.....	98

2.4.6	Removal of rare labels	98
2.4.7	Variable standardization	99
2.4.8	Feature engineering method-1: Indicator Variables	100
2.4.9	Feature engineering method-2: Interaction Features	101
2.4.10	Feature engineering method-3: Feature Representation	101
2.4.11	Feature engineering method-4: External Data	102
2.4.12	Feature engineering method-5: Error Analysis (Post Modeling)	103
2.5	Feature Engineering for Used car price prediction problem	104
2.5.1	Extraction Features for used car price prediction:	104
2.5.2	Ratio Features for used car price prediction:	106
2.5.3	Grouped Mean value Features for used car price prediction:	106
2.6	Machine Learning Stages: Feature Selection	108
2.6.1	Univariate Selection	109
2.6.2	Correlation Matrix with Heatmap	110
2.6.3	Feature importance (with Lasso regression)	111
2.7	Machine Learning Stages: Modeling	113
2.8	Machine Learning Stages: Evaluation	115
2.9	Machine Learning Stages: Deployment	118
2.10	Machine Learning Model Deployment Overview	120
2.10.1	Procedural Programming Style	121
2.10.2	Object-oriented style	124
Chapter 3	Scale and Speed, need for Data Sciences as Software Engineering	132
3.1	Transferring Machine Learning Models from development to production	133
3.2	Predictive Model Markup Language (PMML)	134
3.3	Need for software development practices in Data Science	136
3.4	Automation: Enable scale and speed in deployment	137
3.5	Version control: To enable seamless team collaboration	138
3.6	Versioning: To track application and model upgrades	141
3.7	Logging: To monitor services and application usage, health	143
3.8	Package Management: To build packages and utilize the private repository	143
3.8.1	Private repository for packages	145
3.9	Project Documentation: To enable knowledge transfer within teams	147

3.10	Adoption of standards: To allow interoperability and cross-system communications	148
Chapter 4	Machine Learning System Design	151
4.1	Challenges while Designing a Machine Learning Framework.....	151
4.1.1	Computation and Memory Intensive operations	151
4.1.2	Portable Machine Learning.....	151
4.1.3	Proper Integration with Mainstream/legacy Software	152
4.1.4	Separation of concerns and data privacy management.....	152
4.2	Machine Learning System Design	153
4.2.1	Aligning Data Science and Machine Learning Process to Software Design.....	153
4.2.2	Engineering an Ad-hoc CLI ML-tool	154
4.2.3	Why Ad-hoc?.....	155
4.2.4	What is CLI integration pattern and why to use it now?	155
4.2.5	Overview of cli_ml_automation framework:	168
4.2.6	CLI commands and JSON Input/Output format for the ML modules	171
4.2.7	Need for Micro-services.....	177
4.2.8	Adapting to better Integration patterns.....	181
4.3	Common Project essentials.....	185
4.3.1	Decomposition and Dependency Resolution	186
4.3.2	Project Structure	186
4.3.3	Overview of common_utility	188
4.3.4	Automated Testing	191
4.3.5	Documentation	193
4.3.6	Dependency Resolution using Gemfury	197
4.4	Developing a Real-time Prediction Engine.....	200
4.4.1	Docker basics	201
4.4.2	Key Terminologies in Docker	203
4.4.3	Basic Docker Commands.....	204
4.4.4	Project Structure for real_time_prediction_engine	206
4.4.5	Running our real_time_prediction_engine application	207
4.4.6	Containerization.....	210
4.4.7	I/O formats and Pipelining.....	212
	Framework Design	223

4.5	Machine Learning as a Service	226
4.5.1	Software components for MLaaS	228
4.5.2	Solution architecture	231
4.5.3	Testing our MLaaS.....	236
Chapter 5	Enterprise Deployment Strategies.....	244
5.1	Characteristics of Deployments	244
5.2	Deployment Responsibilities.....	247
5.2.1	Monolithic Architecture.....	248
5.2.2	Microservices Architecture	248
5.3	Various Components of Cloud Deployment	251
5.4	Docker Orchestration Management	254
5.4.1	Docker Swarm	255
5.4.2	Kubernetes.....	259
5.5	Enterprise Release Management.....	265
Chapter 6	Engineering a machine learning workflow management system	268
6.1	Empowerment of Data-Science team	268
6.2	Designing cloud-native workflow management and monitoring platform for Machine Learning.....	270
6.3	Components of ML Cloud.....	272
6.4	Deploying our ML Cloud stack.....	274
6.5	Workflow Management using airflow	280
6.6	Monitoring using Dash	284
6.7	Delivering a machine learning pipeline using MLaaS and ML-Cloud tool.....	288
6.7.1	ML_automation code Repository	288
6.7.2	MLaaS.....	293
6.7.3	ML-Cloud (Airflow).....	295
6.7.4	ML-Cloud (Dash- Analytics):.....	298
6.8	Summary	301
Chapter 7	Enterprise governance considerations for Data Sciences	304
7.1	Project Management for Data Science Teams.....	305
7.2	Program Management and Executive Dashboards for Data Sciences.....	311
7.3	Management of Machine Learning modeling process	317
7.3.1	ModelDB: Purpose and Architecture	318

7.3.2	MLflow and its usage	319
7.4	Ethical AI and Model Fairness for Regulatory compliance	323
7.4.1	Bias in Machine Learning and AI	324
7.4.2	Accountability and Remedial measures	324
7.4.3	Transparency, Interpretability, and Explainability.....	325
7.4.4	Impact of AI and ML.....	325
7.4.5	Tools to Audit Machine Bias in Learning Models	325
7.5	Model Explainability and Interpretability	328
7.6	Designing performance appraisals for data science teams	334
	Traditional appraisal systems:	337
	Quality management-based appraisal systems:.....	339

List of Figures

Figure 1-1: Data Science Venn diagram [Adopted from Steven Geringer].....	2
Figure 1-2: Evolution of Data Science	3
Figure 1-3: Analytics and Data Science Job Growth	7
Figure 1-4: Normal Distribution curve with Value (x-axis) against Frequency (y-axis)	14
Figure 1-5: Correlation on Linear Data	15
Figure 1-6: Data demonstrating sine (non- linear) relationship	16
Figure 1-7: Correlation Scores	16
Figure 1-8: Mutual Information Scores.....	17
Figure 1-9: Gradient Descent Convergence.....	23
Figure 1-10: Distributions of a discrete (left) and continuous (right) variable	24
Figure: 1-11: Bernoulli Distribution, $n = 1$, since only one trial is allowed	24
Figure 1-12: Binomial Distribution; p = probability of one outcome, n = number of trials.....	25
Figure 1-13: Effect of Log Transformation: Skewed Data Distribution to Gaussian Distribution	26
Figure 1-14: Uniform distribution: Probability (y-axis) vs Events (x-axis)	27
Figure 1-15: Poisson distribution	27
Figure 1-16: Exponential Distribution	28
Figure 1-17: Summary of the types of learners along with apt examples; Courtesy: Oracle Blogs	34
Figure 1-18: Random Weight Initialization	36
Figure 1-19: Measurement of distance from the best line of fit	36
Figure 1-20: Gradient Descent to find the optimal point	36
Figure 1-21: Line of best fit	37
Figure 1-22: Sigmoid Function Equation.....	38
Figure 1-23: Polynomial regression	39
Figure 1-24: Different Probabilities used in Naive Bayes Model.....	40
Figure 1-25: Prior Probability.....	40
Figure 1-26: Marginal Likelihood	41
Figure 1-27: Likelihood.....	41
Figure 1-28: KNN Graphical Demonstration	42
Figure 1-29: Representation of SVM on a 2D space using 1D Hyperplane	44
Figure 1-30: Decision Tree Splits (Graphical).....	45
Figure 1-31: Decision Tree	45
Figure 1-32: K Mean Clustering Demonstration	49
Figure 1-33: Elbow Method to find the optimal number of Clusters	49
Figure 1-34: Hierarchical Clustering Demonstration	50
Figure 1-35: Method to find the optimal number of clusters	51
Figure 1-36: Hold Out Validation	60
Figure 1-37: Cross-Validation.....	61
Figure 1-38: Bootstrap	61
Figure 1-39: Leave one out Validation	62
Figure 1-40: Anaconda Navigator	63
Figure 1-41: Spyder IDE.....	63

Figure 1-42: Jupyter IDE	64
Figure 1-43: Pycharm IDE.....	64
Figure 1-44: Pandas Cheatsheet	66
Figure 2-1: Stages of Machine Learning Deployment (Batch Processing)	71
Figure 2-2: Stages of Machine Learning Deployment (Real-time Processing)	71
Figure 2-3: Difference in price for missing (code-0) and non-missing (code-1) observations	84
Figure 2-4: Change in price over the years (Note the variation of mean vs. median house price)	85
Figure 2-5: Plot of two discrete variables (left variable exhibits monotonic relationship)	87
Figure 2-6: histogram plot of continuous variables (LotFrontage and SalePrice)	88
Figure 2-7: histogram plot of log-transformed variables (LotFrontage, SalePrice).....	89
Figure 2-8: Box plots of LotFrontage and SalePrice (Outlier Analysis)	90
Figure 2-9: Relationship between variables and Target (SalePrice)	91
Figure 2-10: Types of Data and meaningful transformations on them	96
Figure 2-11: The initial five rows of the dataset	104
Figure 2-12 : Heatmap to help with feature selection	111
Figure 2-13: “Actual” vs “Predicted” house prices.	115
Figure 2-14: Procedural programming structure overview	122
Figure 3-1: Crisp-DM Model.....	136
Figure 3-2: Versioning Schema	142
Figure 4-1: Phase Overlaps in Machine Learning	154
Figure 4-2: cli_ml_automation project contents	156
Figure 4-3: High-level overview of cli_ml_automation	157
Figure 4-4: Integrating Java Controller Application with Python ML framework	158
Figure 4-5: Custom function input arguments	161
Figure 4-6: Custom Function Output arguments.....	161
Figure 4-7: Flow of the pipeline	165
Figure 4-8: cli_ml_automation project structure	165
Figure 4-9: ml_automation project structure	166
Figure 4-10: cli_ml_automation Framework Design	167
Figure 4-11: Phases of Machine Learning	169
Figure 4-12: JSON Input Structure for merging and preprocessing multiple raw files in DataProcessor	173
Figure 4-13: JSON Output Structure for merging and preprocessing multiple raw files in DataProcessor	173
Figure 4-14: JSON Input Structure for invoking single raw file with two preprocessing functions in the Data Processor module.....	174
Figure 4-15: JSON Output Structure for a single raw file with two preprocessing functions in Data Processor module.....	174
Figure 4-16: JSON Input Structure for invoking ModelGenerator functions.....	175
Figure 4-17: JSON Output Structure for ModelGenerator functions	175
Figure 4-18: JSON Input Structure for invoking ModelPredictor functions	177
Figure 4-19: JSON Output Structure for ModelPredictor functions	177
Figure 4-20: JSON Output Structure for Exception Handling	177

Figure 4-21: Synchronous Communication via a telephone.....	182
Figure 4-22: Asynchronous Communication via Text messages	183
Figure 4-23: Communication pattern in a Real-time Prediction Engine.....	183
Figure 4-24: Communication pattern in MLaaS.....	184
Figure 4-25: Machine Learning Project Decomposition and Dependencies	186
Figure 4-26: General ML Project Structure	187
Figure 4-27: common_utils framework design.....	189
Figure 4-28: common_utility project structure	190
Figure 4-29: Documentation folder structure	193
Figure 4-30: Steps to create documentation	195
Figure 4-31:Project Dependencies.....	197
Figure 4-32: common_utils package contents.....	198
Figure 4-33: Steps for managing package dependencies	199
Figure 4-34: Real-time prediction engine framework	201
Figure 4-35: Docker abstraction design	202
Figure 4-36: Docker in various environments.....	203
Figure 4-37: Docker functionalities.....	203
Figure 4-38: Real-time prediction engine project structure	206
Figure 4-39: Steps to run the application	207
Figure 4-40: configs directory for real_time_prediction_engine	208
Figure 4-41: Request Json Structure for Real-time prediction engine	215
Figure 4-42: Response JSON structure for real_time_prediction_engine.....	215
Figure 4-43: Keras Serialization and Deserialization.....	217
Figure 4-44: Request Json for Keras implementation	219
Figure 4-45: Response Json for Keras implementation	219
Figure 4-46: Request json for Custom algorithm implementation.....	222
Figure 4-47: Response JSON for Custom algorithm implementation.....	223
Figure 4-48:Project Structure for Real-time prediction engine.....	223
Figure 4-49: Module dependency for Real-time prediction engine	224
Figure 4-50: High-Level Sequence Diagram for MLaaS.....	226
Figure 4-51: Distributed nature of MLaaS	228
Figure 4-52: Consumer and Producer model.....	229
Figure 4-53: Celery architecture and the components.....	230
Figure 4-54: MLaaS Project structure	231
Figure 4-55: API server talks with the worker queue via a broker (Redis)	232
Figure 4-56: MLaaS synchronous invocation	232
Figure 4-57: MLaaS asynchronous invocation	233
Figure 4-58: Sequence Diagram for MLaaS components	233
Figure 4-59: Standard Operating Procedure in MLaaS	236
Figure 4-60: Celery Flower Dashboard	239
Figure 4-61: Route to check the task	239
Figure 4-62: Response from the check route	239
Figure 4-63: Celery Flower Table for tracking tasks	242
Figure 5-1: Various Deployment architectures over time	244

Figure 5-2: Relationship between Monolithic and Service-oriented architecture.....	249
Figure 5-3: Management responsibilities of deployments.....	251
Figure 5-4: An illustrative example of deployments using Docker containers.....	253
Figure 5-5: Rolling updates of containers.....	254
Figure 5-6: Docker Swarm architecture.....	255
Figure 5-7: Scaling of celery workers.....	258
Figure 5-8: Kubernetes high-level components.....	260
Figure 5-9: Kubernetes deployment plan.....	262
Figure 5-10: Build and release plan in an enterprise using Docker.....	265
Figure 6-1: An example of ML workflow.....	269
Figure 6-2: Data sciences Implementation in Cloud ML.....	270
Figure 6-3: Project Structuring.....	270
Figure 6-4: Components of Cloud ML.....	272
Figure 6-5: Project Structure of Cloud ML.....	273
Figure 6-6: Docker Containers for Cloud ML.....	276
Figure 6-7: File browser web console.....	277
Figure 6-8: Setting up Codiad Online text editor.....	277
Figure 6-9: Codiad Online text editor.....	278
Figure 6-10: Dash Visualization Index Page.....	279
Figure 6-11: Adminer Signup.....	279
Figure 6-12: Airflow Home Page.....	279
Figure 6-13: DAG Syntax for Airflow.....	280
Figure 6-14: A DAG entry in Airflow.....	281
Figure 6-15: Graph view of a DAG.....	281
Figure 6-16: Tree view of a DAG.....	282
Figure 6-17: Gantt chart of a DAG run.....	283
Figure 6-18: 1.txt file in /storage – file browser.....	284
Figure 6-19: Example of a Dash application Dashboard.....	285
Figure 6-20: Adding a dashboard entry in Dash Visualization Index.....	286
Figure 6-21: Example of a Dash application Graph.....	287
Figure 6-22: cli_ml_automation implementation for Housing Price Prediction.....	291
Figure 6-23: JSON structure for model generation.....	292
Figure 6-24: MLaaS additional content- house_prediction_client codes.....	293
Figure 6-25: Request and Responses in MLaaS for Housing Price Implementation.....	295
Figure 6-26: Airflow Variables for Housing Price Prediction.....	295
Figure 6-27: Airflow Housing Price Prediction DAGs set as active.....	295
Figure 6-28: Testing DAG for Housing Price Prediction.....	296
Figure 6-29: Training DAG for Housing Price Prediction.....	296
Figure 6-30: Tree view of Housing Price Prediction testing DAG.....	297
Figure 6-31: csv files created after Housing Price Prediction testing DAG run.....	298
Figure 6-32: Entry of Housing Price Prediction Dashboard in Datascience visualization index.....	299
Figure 6-33: Dashboard depicting Actual vs Predicted labels for Housing Price Prediction.....	301
Figure 6-34: Housing Price Prediction Implementation plan.....	301
Figure 7-1: Various sections (or Statuses) a Data Science project goes through.....	307

Figure 7-2: Suggested development process for Data Sciences	308
Figure 7-3: Snapshot of the JIRA Project Board for New ML Model Development.....	309
Figure 7-4: View showing products, use cases and the count of supported clients	315
Figure 7-5: Dashboard to illustrate the stages of various ML uses cases and algorithms used.	316
Figure 7-6: ML Model enablement timeline	317
Figure 7-7: Model DB architecture (Image courtesy: https://mitdbg.GitHub.io/modeldb)	319
Figure 7-8: ML Flow Initial Screen.....	322
Figure 7-9: MLFlow screen after single train command	322
Figure 7-10: MLFlow User Interface (After several "train" executions)	323
Figure 7-11: Importances from SHAP model (For Random Forest Regressor)	332
Figure 7-12: LIME explainability for a selected data point.	334

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



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.