Industrial Machine Learning

Using Artificial Intelligence as a Transformational Disruptor

—

Andreas François Vermeulen

Apress®

Industrial Machine Learning

Using Artificial Intelligence as a Transformational Disruptor

Andreas François Vermeulen

Apress®

Industrial Machine Learning: Using Artificial Intelligence as a Transformational Disruptor

Andreas François Vermeulen
West Kilbride, UK

ISBN-13 (pbk): 978-1-4842-5315-1 ISBN-13 (electronic): 978-1-4842-5316-8
https://doi.org/10.1007/978-1-4842-5316-8

Managing Director, Apress Media LLC: Welmoed Spahr
Acquisitions Editor: Susan McDermott
Development Editor: Laura Berendson
Coordinating Editor: Rita Fernando

Cover designed by eStudioCalamar

Cover image designed by Freepik (www.freepik.com)

Distributed to the book trade worldwide by Springer Science+Business Media New York, 233 Spring Street, 6th Floor, New York, NY 10013. Phone 1-800-SPRINGER, fax (201) 348-4505, e-mail orders-ny@springer-sbm.com, or visit www.springeronline.com. Apress Media, LLC is a California LLC and the sole member (owner) is Springer Science + Business Media Finance Inc (SSBM Finance Inc). SSBM Finance Inc is a **Delaware** corporation.

For information on translations, please e-mail rights@apress.com, or visit http://www.apress.com/rights-permissions.

Apress titles may be purchased in bulk for academic, corporate, or promotional use. eBook versions and licenses are also available for most titles. For more information, reference our Print and eBook Bulk Sales web page at http://www.apress.com/bulk-sales.

Any source code or other supplementary material referenced by the author in this book is available to readers on GitHub via the book's product page, located at www.apress.com/9781484253151. For more detailed information, please visit http://www.apress.com/source-code.

Printed on acid-free paper

Thank you to Denise and Laurence for their support and love.

"Time is an illusion." —Albert Einstein

Table of Contents

About the Author

Andreas François Vermeulen is Chief Data Scientist and Solutions Delivery Manager at Sopra-Steria, and he serves as part-time doctoral researcher and senior research project advisor at University of St Andrews on future concepts in health-care systems, Internet-of-Things sensors, massive distributed computing, mechatronics, at-scale data lake technology, data science, business intelligence, and deep machine learning in health informatics.

Andreas maintains and incubates the "Rapid Information Factory" data processing framework.

He is active in developing next-generation data processing frameworks and mechatronics engineering with over 36+ years of global experience in complex data processing, software development, and system architecture. Andreas is an expert data scientist, doctoral trainer, corporate consultant, and speaker/author/columnist on data science, business intelligence, machine learning, decision science, data engineering, distributed computing, and at-scale data lakes.

He holds expert industrial experience in various areas (finance, telecommunication, manufacturing, government service, public safety, and health informatics).

Andreas received his bachelor degree at the North West University at Potchefstroom, his Master of Business Administration at University of Manchester, Master of Business Intelligence and Data Science degree at University of Dundee, and Doctor of Philosophy (PhD) at University of St Andrews.

About the Technical Reviewer

Chris Hillman is a Data Science Practice lead with over 25 years of experience working with analytics across many industries including retail, finance, telecoms, and manufacturing. Chris has been involved in the pre-sales and start-up activities of analytics projects helping customers to gain value from and understand advanced analytics and machine learning. He has spoken on data science and AI at Teradata events such as Universe and Partners and also industry events such as Strata, Hadoop World, Flink Forward, and IEEE Big data conferences. Chris gained a Doctor of Philosophy (PhD) researching real-time distributed feature extraction at the University of Dundee.

Acknowledgments

To all my past tutors, thank you for the wisdom you shared with me.
To my numerous associates, thanks for sharing your established wisdom!
Our deliberating and concepts on machine learning produce these ideas.
To the people at Apress, your skills transformed an idea into a book.
Well done!

> *"A man who dares to waste one hour of time has not discovered the value of life."*
>
> —Charles Darwin

So thank you, as the reader, for investing time into my knowledge distribution.

CHAPTER 1

Introduction

Industrialized Machine Learning (IML) is evolving as a disruptor in the world around us, and people are finally recognizing the true impact it has already had and will continue to have on our future. Throughout this book, I will share my knowledge and insights acquired from more than ten years of consulting, including setting up three data science teams that design and implementation digital transformations for business and research practices across the world.

The uses of IML techniques and algorithms have progressively accumulated in velocity, volume, and impact. The impact over just the last three years has evolved into an immense disruptor of many principal business processes. There is no industry that machine learning is not impacting daily, and predictions of IML growth patterns are 300% plus in volume and 250% in complexity.

In this book, I will provide a sample of wide-ranging insights and advice on the methodologies and techniques that I have found to be most useful in my current consulting ecosystem. These will assist you in capitalizing on the industrialization of your capabilities through the field of machine learning.

To apply machine learning to a selection of data sets, you need the following common process checklist:

- There must be a distinguishable pattern in the data. The determination of a pattern in the data is the primary goal of machine learning. Without the pattern, the process does not work.
- Patterns must not be solvable with a mathematical formula. It is always a good idea to investigate the likelihood that a data pattern is directly correlated to a mathematical calculation. With a proven mathematical formula, you do not need machine learning; you simply perform the mathematical calculation every time you need a result.

A. F. Vermeulen, *Industrial Machine Learning*, https://doi.org/10.1007/978-1-4842-5316-8_1

- You must have access to all aspects of the data sets to learn the insights. The data you need to perform your insights must be available as a "true" single source. Sadly, I have observed numerous projects performing significant machine learning solutions, just to discover the data they used is not for the area of interest and not allowed to be used due to restrictions. Take the time to validate the lineage and provenance of all data sources you include in the source data.

This book covers IML applications for the following industries:

- Health Informatics
- Hospitals and Other Medical Facilities
- Automotive
- Aerospace
- Communications
- Contact Centers
- Datacenters
- Finance and Banking
- Application and Software Development
- Gaming and Virtual Reality (VR)
- Augmented Reality (AR)
- IoT/Embedded Systems
- Mobility and Robotics
- Retail Planning
- E-commerce
- Wireless Carriers
- Cybersecurity

I will supply an introduction to specific machine learning techniques and explain how to industrialize them into real-world applications with comprehensive applied examples.

The general mainstream view of machine learning currently is that humans need protection against their abuse. There are more than a few groups active in opposing camps to actively control the impact of autonomics machines.

Please take note of these forces at play, as it will have a major impact on the level and nature of the IML techniques and algorithms you may use in the future.

One of the most famous concepts is Isaac Asimov's "Four Laws of Robotics" (Isaac Asimov drafted these in a 1942 short story called "Runaround"):

- First Law: A robot may not injure a human being or, through inaction, allow a human being to come to harm.
- Second Law: A robot must obey orders given it by human beings except where such orders would conflict with the First Law.
- Third Law: A robot must protect its own existence if such protection does not conflict with the First or Second Law.
- Zeroth Law: A robot may not harm humanity, or, by inaction, allow humanity to come to harm.

These four basic laws could be the basis of several hours of deep philosophical discussion. I will return to these laws and the meta-ethical issues created by the emerging field of Machine Ethics and Industrialized Machine Learning. My personal advice is this: Just because you can achieve it, doesn't mean you should implement it! Always do no harm! In simple terms: always act with responsibility to an improved overall outcome for everybody involved with your machine learning.

The basic machine learning theory will be covered first before I discuss its implementation and consequences with numerous legal restrictions during Chapter 14.

I will discuss the basic background information of the various methodologies and techniques next. Later on, I will offer examples to show how you can deploy and use IML in your daily life.

The readership of this book is set at an intermediate to advanced level. The information shared makes the following assumptions:

- You have working understanding of the Python programming ecosystem.
- You have knowledge of working with Jupyter Notebooks.
- You have experience using the Jupyter Notebooks as a data science tool.

- You can install missing Python libraries into the notebook without a great deal of assistance from this book.
- You have working knowledge of mathematical, scientific, and statistical calculations used by machine learning.
- You have accomplished basic machine learning concepts and understand the basic techniques.
- You will need to be able to get example code from a GitHub site to use the example code.

With the required user experience, I will discuss the basic knowledge of the techniques and algorithms at a non-beginner's pace. I will suggest you look at the additional material located in Appendix A as background if you are not sure about the more advanced concepts covered in each chapter.

Get Ready!!!

As we get ready to proceed, you should have a Jupyter Notebook ecosystem ready to run the examples. You can find the book's examples located here: `www.apress.com/9781484253151`.

If you need a Python environment, I personally use the Anaconda ecosystem found here: `https://www.anaconda.com/` When you are ready, I will start with the basic concepts.

Are You Ready?

I assume at this point that you have already been running a completely operational Anaconda Python 3 Jupyter Notebook environment on your computer or have access to the online version of Jupyter Notebook.

You should already have experience in the usage of the Jupyter Notebook ecosystem on your own machine or on a cloud ecosystem. At this point you can progress by two routes: personal computer or Cloud Base. This is exclusively your choice.

Personal Computer

If you are not ready with a Jupyter Notebook ecosystem, I suggest you install Anaconda's ecosystem from:

https://anaconda.org/

On installation completion, you simply run the Jupyter Notebook and you should get a page:

http://localhost:8888/tree

This would provide you the basic ecosystem that you need for this book.

Cloud Base

Want to go to cloud? You have many choices! I use three different cloud providers during my daily work and I have listed them here.

Microsoft Azure Notebooks

Develop and run code from anywhere with Jupyter notebooks on Azure. You can use Microsoft Azure if you have a Microsoft account:

https://notebooks.azure.com/

Google Cloud Platform

Collaboratory is a free Jupyter Notebook environment that requires no setup and runs entirely in the cloud. You can use Google Collaboratory, if you have a Google account:

https://colab.research.google.com

Amazon Web Services

You can use AWS Sagemaker if you have an AWS account:

https://console.aws.amazon.com/sagemaker/

Tip There are many other cloud providers you can use, and if they support the Python 3 and Jupyter Notebook ecosystem, they can be another option.

Note I used my own Jupyter Hub installation for the examples in this book. (See http://jupyter.org/hub)

Let's Get Started

At this point I suggest you run the following to get ready to perform the examples in this book. Open a web browser and link to: www.apress.com/9781484253151.

Download the examples from the Apress Source Code site.

Note For your ease of use, I have bundled each chapter's examples into a single ZIP file. This will help you to get the precise examples for each chapter.

Warning Without the code, the book will be not as effective in supporting your development of new abilities and honing of existing talents.

I will quickly help you check if you are ready to begin. Please open `Chapter_001_Test_System.ipynb` from the example directory for Chapter 1.

Run the complete Jupyter Notebook, which will assist you with getting all the Python libraries you will need throughout the examples in this book. When you have all the steps complete, you can close it and progress with your learning process.

What's Next?

The rest of this book will guide you through theoretical knowledge, supported with examples, to empower you with the knowledge to understand the background skills you need to perform IML. The next chapter will provide background knowledge before you start with the IML theory.

CHAPTER 2

Background Knowledge

The next digital evolution of the world around us is here. Companies are now using industrialized machine learning daily, driven by ever-evolving artificial intelligence capabilities as a Transformational Disruptor of traditional business models. The ability of machine learning to improve the models and methodologies that drive our world around us is enabling machine learning to adapt and evolve to these needs of the next generation of business requirements.

People are storing large amounts of their personal and company assets in massive data lakes. The customers I am consulting with are now openly admitting that without an advanced and well-designed machine learning strategy to effectively and efficiently handle these ever-expanding lakes full of critical business information, they will not survive the fourth industrial revolution.

I consult with organizations on a regular basis on how to develop their data lake, data science strategy, and machine learning to serve their evolving and ever-changing business strategies. These disruptors require agile and cost-effective, machine-driven information management to handle the priority list of senior managers worldwide.

It is a fact that many unknown insights are captured and stored in a massive pool of unprocessed data in the enterprise. These data lakes have major implications for the future of the business world. It is projected that combined data scientists worldwide will have to handle 40 zettabytes of data by 2020, an increase of 300+ times since 2005.

There are numerous data sources that still need to be converted into actionable business knowledge. The achievement will safeguard the future of the business that can achieve it.

The world's data producers are generating 2.5 quintillion bytes of new data every day. The Internet of Things will cause this volume to be substantially higher. Data scientists and engineers are falling behind on an immense responsibility. The only viable solution is an active drive to enable machine learning to adapt and evolve to these new data needs while data scientists become the trainers of the next generation of artificial intelligence capabilities.

A. F. Vermeulen, *Industrial Machine Learning*, https://doi.org/10.1007/978-1-4842-5316-8_2

The purpose of this book is to prepare you to understand how to use these incredible and powerful processing engines to act as a disruptor of your current business environments. By reading the introduction plus this background, you are already proving to be an innovative person who wants to understand, and perhaps tame, this advanced artificial intelligence. To tame your data lake with artificial intelligence, you will need practical advice on the data science, machine learning, and transformational disruptors.

I propose to teach you how to tame this beast!

I am familiar with the skills it takes to achieve this goal, and I will guide you with the sole purpose of helping you to learn and expand while understanding the practical guidance in this book.

I will get you started using machine learning theory and then advance to deployment against the data lakes from several business applications.

You will then understand the following:

- What machine learning models tame your business' data lake?
- How do you apply data science and machine learning to succeed in this undertaking?

Think of the process as comparable to a natural lake. It is vital to accomplish a sequence of proficient techniques with the lake water to obtain pure water in your glass.

By the end of this book, you will have shared in over 30+ years of working experience with data and extracting actionable business knowledge. I will share the experience I gained in working with data on an international scale with you. You will understand the processing framework that I use on a regular basis to tame data lakes and the collection of monsters that live in and around the lake.

I have included examples at the end of each chapter, along with code, which more serious data scientists can use as you progress throughout the book. But please note that it is not required for you to complete the examples in order to understand the concepts in each chapter.

So welcome to a walk-through of a characteristic machine learning of a data lake project using practical data science techniques and machine learning insights. The objective of the rest of this background chapter is to explain the fundamentals of data science and machine learning.

Data Science

In 1960, Peter Naur started using the term "data science" as a substitute for computer science. He stated to work with data, you need more than just computer science. I agree with his declaration.

Data science is an interdisciplinary science incorporating practices and methods to action knowledge and insights from data in heterogeneous schemas (structured, semi-structured, or unstructured). It amalgamates the science fields of data exploration from thought-provoking research fields like data engineering, information science, computer science, statistics, artificial intelligence, machine learning, data mining, and predictive analytics.

As I enthusiastically researched into the future usage of data science by translating multiple data lakes, I discovered several valuable insights. I will explain with end-to-end examples and share my insights on data lakes. This book explains vital elements from these sciences that you can use to process your data lake into actionable knowledge. I will guide you through a series of recognized science procedures for data lakes. These core skills are a key set of assets to perfect as you start into your encounters using data science.

Data Analytics

Data analytics is the science of fact-finding analysis of raw data with the goal of drawing conclusions from the data lake. It is driven by certified algorithms to statistically define associations between data that produce insights.

The perception of certified algorithms is exceptionally significant when you want to sway other business people about the importance of the data insights you have uncovered.

You should not be surprised if you are asked regularly to substantiate it and explain how you know it is correct!

The best answer is to have the competency to point at a certified and recognized algorithm you used. Associate the algorithm to your business terminology to accomplish success with your projects.