Francesca Lazzeri, PhD

Machine Learning for Time Series Forecasting with Python

WILEY

Machine Learning for Time Series Forecasting with Python®

Francesca Lazzeri, PhD

WILEY

Machine Learning for Time Series Forecasting with Python®

ISBN: 978-1-119-68236-3
ISBN: 978-1-119-68237-0 (ebk)
ISBN: 978-1-119-68238-7 (ebk)

Manufactured in the United States of America

Limit of Liability/Disclaimer of Warranty: The publisher and the author make no representations or warranties with respect to the accuracy or completeness of the contents of this work and specifically disclaim all warranties, including without limitation warranties of fitness for a particular purpose. No warranty may be created or extended by sales or promotional materials. The advice and strategies contained herein may not be suitable for every situation. This work is sold with the understanding that the publisher is not engaged in rendering legal, accounting, or other professional services. If professional assistance is required, the services of a competent professional person should be sought. Neither the publisher nor the author shall be liable for damages arising herefrom. The fact that an organization or Web site is referred to in this work as a citation and/or a potential source of further information does not mean that the author or the publisher endorses the information the organization or website may provide or recommendations it may make. Further, readers should be aware that Internet websites listed in this work may have changed or disappeared between when this work was written and when it is read.

For general information on our other products and services please contact our Customer Care Department within the United States at (877) 762-2974, outside the United States at (317) 572-3993 or fax (317) 572-4002.

Wiley publishes in a variety of print and electronic formats and by print-on-demand. Some material included with standard print versions of this book may not be included in e-books or in print-on-demand. If this book refers to media such as a CD or DVD that is not included in the version you purchased, you may download this material at `booksupport.wiley.com`. For more information about Wiley products, visit `www.wiley.com`.

Library of Congress Control Number: 2020947403

About the Author

Francesca Lazzeri, PhD, is an experienced scientist and machine learning practitioner with over a decade of both academic and industry experience. She currently leads an international team of cloud AI advocates and developers at Microsoft, managing a large portfolio of customers and building intelligent automated solutions on the cloud.

Francesca is an expert in big data technology innovations and the applications of machine learning–based solutions to real-world problems. Her work is unified by the twin goals of making better sense of microeconomic data and using those insights to optimize firm decision making. Her research has spanned the areas of machine learning, statistical modeling, and time series econometrics and forecasting as well as a range of industries—energy, oil and gas, retail, aerospace, healthcare, and professional services.

Before joining Microsoft, she was a research fellow at Harvard University in the Technology and Operations Management Unit. Francesca periodically teaches applied analytics and machine learning classes at universities and research institutions around the world. You can find her on Twitter @frlazzeri.

About the Technical Editor

James York-Winegar holds a bachelor's degree in mathematics and physics and a master's degree in information and data science. He has worked in academia, healthcare, and technology consulting. James currently works with companies to enable machine learning workloads by enabling their data infrastructure, security, and metadata management. He also teaches machine learning courses at the University of California, Berkeley, focused on scaling up machine learning technology for big data.

Prior to leaving academia, James originally was focused on the cross section between experimental and theoretical physics and materials science. His research was focused on photo-structural transformations of non-oxide glasses or chalcogenide glasses. This introduced James to processing extremely large amounts of data and high-performance computing, where his work still leads him today.

James has had exposure to many industries through his consulting experience, including education, entertainment, commodities, finance, telecommunications, consumer packaged goods, startups, biotech, and technology. With this experience, he helps companies understand what is possible with their data and how to enable new capabilities or business opportunities. You can find his LinkedIn profile at `linkedin.com/in/winegarj/`.

Acknowledgments

In the past few years, I had the privilege to work with many data scientists, cloud advocates, developers, and professionals from Microsoft and Wiley: all these people inspired me and supported me through the creation and writing of this book. I am particularly grateful to the Cloud Advocacy team at Microsoft, for their trust and encouragement and for making my job so much easier and more enjoyable.

Thanks to Jim Minatel, associate publisher at John Wiley & Sons, who worked with me since the beginning of the publishing process and acted as a bridge between my work and the editorial staff. It was a pleasure to work with Pete Gaughan, content enablement manager; with David Clark, project editor, who managed the process that got us from outline to a finished manuscript; and with Saravanan Dakshinamurthy, content refinement specialist, who managed the last stages of this book development and made them go as smoothly as they did. I appreciate the support and guidance provided by the technical reviewer, James Winegar, and I hope we will work on other projects together in the future.

Last but not least, I will be forever grateful to my daughter, Nicole, who reminds me of the goodness in this world and inspires me to be the greatest version of myself; to my husband, Laurent, for his unconditional and endless support, inspiration, and encouragement; to my parents, Andrea and Anna Maria, and my brother, Marco, for always being there for me and believing in me through all walks of my life.

—Francesca Lazzeri

Contents at a Glance

Contents

Introduction

Time series data is an important source of information used for future decision making, strategy, and planning operations in different industries: from marketing and finance to education, healthcare, and robotics. In the past few decades, machine learning model-based forecasting has also become a very popular tool in the private and public sectors.

Currently, most of the resources and tutorials for machine learning model-based time series forecasting generally fall into two categories: code demonstration repo for certain specific forecasting scenarios, without conceptual details, and academic-style explanations of the theory behind forecasting and mathematical formula. Both of these approaches are very helpful for learning purposes, and I highly recommend using those resources if you are interested in understanding the math behind theoretical hypotheses.

This book fills that gap: in order to solve real business problems, it is essential to have a systematic and well-structured forecasting framework that data scientists can use as a guideline and apply to real-world data science scenarios. The purpose of this hands-on book is to walk you through the core steps of a practical model development framework for building, training, evaluating, and deploying your time series forecasting models.

The first part of the book (Chapters 1 and 2) is dedicated to the conceptual introduction of time series, where you can learn the essential aspects of time series representations, modeling, and forecasting.

In the second part (Chapters 3 through 6), we dive into autoregressive and automated methods for forecasting time series data, such as moving average, autoregressive integrated moving average, and automated machine learning for time series data. I then introduce neural networks for time series forecasting, focusing on concepts such as recurrent neural networks (RNNs) and the comparison of different RNN units. Finally, I guide you through the most important steps of model deployment and operationalization on Azure.

Along the way, I show at practice how these models can be applied to real-world data science scenarios by providing examples and using a variety of open-source Python packages and Azure. With these guidelines in mind, you should be ready to deal with time series data in your everyday work and select the right tools to analyze it.

What Does This Book Cover?

This book offers a comprehensive introduction to the core concepts, terminology, approaches, and applications of machine learning and deep learning for time series forecasting: understanding these principles leads to more flexible and successful time series applications.

In particular, the following chapters are included:

Chapter 1: Overview of Time Series Forecasting This first chapter of the book is dedicated to the conceptual introduction of time series, where you can learn the essential aspects of time series representations, modeling, and forecasting, such as time series analysis and supervised learning for time series forecasting.

We will also look at different Python libraries for time series data and how libraries such as pandas, statsmodels, and scikit-learn can help you with data handling, time series modeling, and machine learning, respectively.

Finally, I will provide you with general advice for setting up your Python environment for time series forecasting.

Chapter 2: How to Design an End-to-End Time Series Forecasting Solution on the Cloud The purpose of this second chapter is to provide an end-to-end systematic guide for time series forecasting from a practical and business perspective by introducing a time series forecasting template and a real-world data science scenario that we use throughout this book to showcase some of the time series concepts, steps, and techniques discussed.

Chapter 3: Time Series Data Preparation In this chapter, I walk you through the most important steps to prepare your time series data for forecasting models. Good time series data preparation produces clean and well-curated data, which leads to more practical, accurate predictions.

Python is a very powerful programming language to handle data, offering an assorted suite of libraries for time series data and excellent support for time series analysis, such as SciPy, NumPy, Matplotlib, pandas, statsmodels, and scikit-learn.

You will also learn how to perform feature engineering on time series data, with two goals in mind: preparing the proper input data set that is compatible with the machine learning algorithm requirements and improving the performance of machine learning models.

Chapter 4: Introduction to Autoregressive and Automated Methods for Time Series Forecasting In this chapter, you discover a suite of autoregressive methods for time series forecasting that you can test on your forecasting problems. The different sections in this chapter are structured to give you just enough information on each method to get started with a working code example and to show you where to look to get more information on the method.

We also look at automated machine learning for time series forecasting and how this method can help you with model selection and hyperparameter tuning tasks.

Chapter 5: Introduction to Neural Networks for Time Series Forecasting In this chapter, I discuss some of the practical reasons data scientists may still want to think about deep learning when they build time series forecasting solutions. I then introduce recurrent neural networks and show how you can implement a few types of recurrent neural networks on your time series forecasting problems.

Chapter 6: Model Deployment for Time Series Forecasting In this final chapter, I introduce Azure Machine Learning SDK for Python to build and run machine learning workflows. You will get an overview of some of the most important classes in the SDK and how you can use them to build, train, and deploy a machine learning model on Azure.

Through machine learning model deployment, companies can begin to take full advantage of the predictive and intelligent models they build and, therefore, transform themselves into actual AI-driven businesses.

Finally, I show how to build an end-to-end data pipeline architecture on Azure and provide deployment code that can be generalized for different time series forecasting solutions.

Reader Support for This Book

This book also features extensive sample code and tutorials using Python, along with its technical libraries, that readers can leverage to learn how to solve real-world time series problems.

Readers can access the sample code and notebooks at the following link: `aka.ms/ML4TSFwithPython`

Companion Download Files

As you work through the examples in this book, the project files you need are all available for download from `aka.ms/ML4TSFwithPython`.

Each file contains sample notebooks and data that you can use to validate your knowledge, practice your technical skills, and build your own time series forecasting solutions.

How to Contact the Publisher

If you believe you've found a mistake in this book, please bring it to our attention. At John Wiley & Sons, we understand how important it is to provide our customers with accurate content, but even with our best efforts an error may occur.

In order to submit your possible errata, please email it to our customer service team at `wileysupport@wiley.com` with the subject line "Possible Book Errata Submission."

How to Contact the Author

We appreciate your input and questions about this book! You can find me on Twitter at @frlazzeri.

CHAPTER

1

Overview of Time Series Forecasting

Time series is a type of data that measures how things change over time. In a time series data set, the *time* column does not represent a variable per se: it is actually a primary structure that you can use to order your data set. This primary temporal structure makes time series problems more challenging as data scientists need to apply specific data preprocessing and feature engineering techniques to handle time series data.

However, it also represents a source of additional knowledge that data scientists can use to their advantage: you will learn how to leverage this temporal information to extrapolate insights from your time series data, like trends and seasonality information, to make your time series easier to model and to use it for future strategy and planning operations in several industries. From finance to manufacturing and health care, time series forecasting has always played a major role in unlocking business insights with respect to time.

Following are some examples of problems that time series forecasting can help you solve:

- What are the expected sales volumes of thousands of food groups in different grocery stores next quarter?
- What are the resale values of vehicles after leasing them out for three years?

- What are passenger numbers for each major international airline route and for each class of passenger?
- What is the future electricity load in an energy supply chain infrastructure, so that suppliers can ensure efficiency and prevent energy waste and theft?

The plot in Figure 1.1 illustrates an example of time series forecasting applied to the energy load use case.

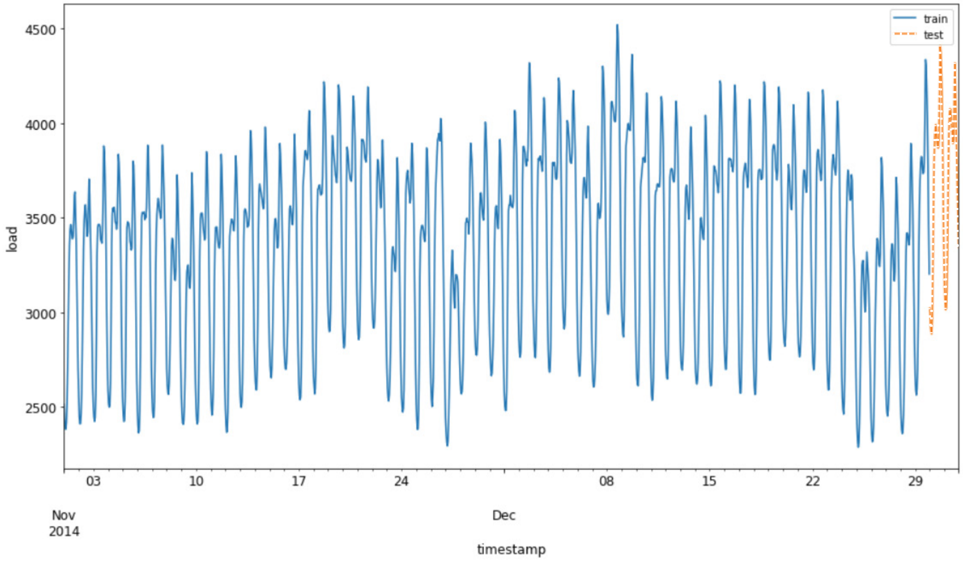

Figure 1.1: Example of time series forecasting applied to the energy load use case

This first chapter of the book is dedicated to the conceptual introduction—with some practical examples—of time series, where you can learn the essential aspects of time series representations, modeling, and forecasting.

Specifically, we will discuss the following:

- *Flavors of Machine Learning for Time Series Forecasting* – In this section, you will learn a few standard definitions of important concepts, such as time series, time series analysis, and time series forecasting, and discover why time series forecasting is a fundamental cross-industry research area.
- *Supervised Learning for Time Series Forecasting* – Why would you want to reframe a time series forecasting problem as a supervised learning problem? In this section you will learn how to reshape your forecasting scenario as a supervised learning problem and, as a consequence, get access to a large portfolio of linear and nonlinear machine learning algorithms.

- *Python for Time Series Forecasting* – In this section we will look at different Python libraries for time series data and how libraries such as pandas, statsmodels, and scikit-learn can help you with data handling, time series modeling, and machine learning, respectively.
- *Experimental Setup for Time Series Forecasting* – This section will provide you general advice for setting up your Python environment for time series forecasting.

Let's get started and learn some important elements that we must consider when describing and modeling a time series.

Flavors of Machine Learning for Time Series Forecasting

In this first section of Chapter 1, we will discover together why time series forecasting is a fundamental cross-industry research area. Moreover, you will learn a few important concepts to deal with time series data, perform time series analysis, and build your time series forecasting solutions.

One example of the use of time series forecasting solutions would be the simple extrapolation of a past trend in predicting next week hourly temperatures. Another example would be the development of a complex linear stochastic model for predicting the movement of short-term interest rates. Time-series models have been also used to forecast the demand for airline capacity, seasonal energy demand, and future online sales.

In time series forecasting, data scientists' assumption is that there is no causality that affects the variable we are trying to forecast. Instead, they analyze the historical values of a time series data set in order to understand and predict their future values. The method used to produce a time series forecasting model may involve the use of a simple deterministic model, such as a linear extrapolation, or the use of more complex deep learning approaches.

Due to their applicability to many real-life problems, such as fraud detection, spam email filtering, finance, and medical diagnosis, and their ability to produce actionable results, machine learning and deep learning algorithms have gained a lot of attention in recent years. Generally, deep learning methods have been developed and applied to univariate time series forecasting scenarios, where the time series consists of single observations recorded sequentially over equal time increments (Lazzeri 2019a).

For this reason, they have often performed worse than naïve and classical forecasting methods, such as exponential smoothing and autoregressive integrated moving average (ARIMA). This has led to a general misconception that deep learning models are inefficient in time series forecasting scenarios, and many data