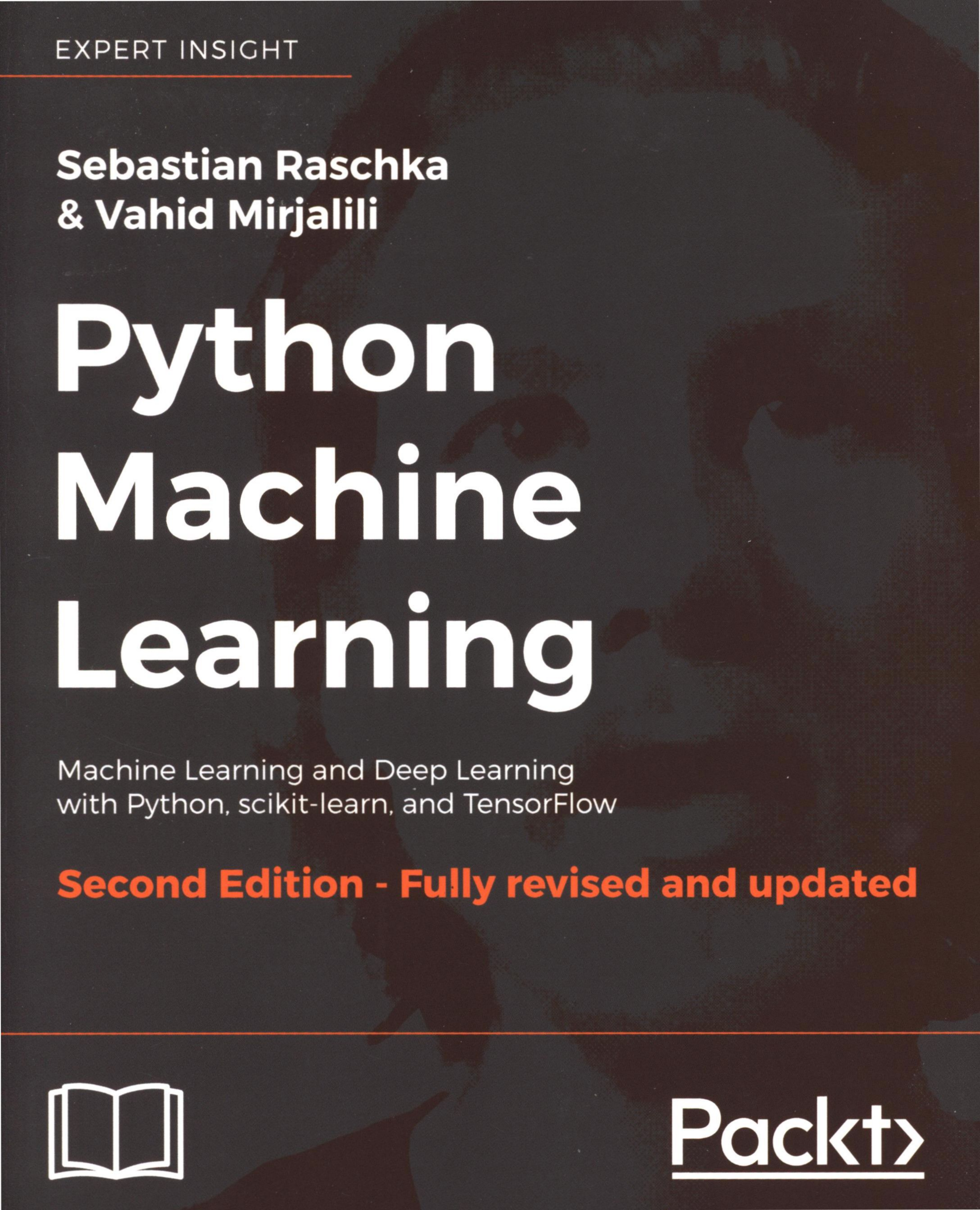
EXPERT INSIGHT

Sebastian Raschka
& Vahid Mirjalili

Python Machine Learning

Machine Learning and Deep Learning
with Python, scikit-learn, and TensorFlow

Second Edition - Fully revised and updated



Packt>

Python Machine Learning

Second Edition

Machine Learning and Deep Learning with Python, scikit-learn, and TensorFlow

Sebastian Raschka

Vahid Mirjalili

Packt>

BIRMINGHAM - MUMBAI

Python Machine Learning

Second Edition

Every effort has been made in the preparation of this book to ensure the accuracy of the information presented. However, the information contained in this book is sold without warranty, either express or implied. Neither the authors, nor Packt Publishing, and its dealers and distributors will be held liable for any damages caused or alleged to be caused directly or indirectly by this book.

Packt Publishing has endeavored to provide trademark information about all of the companies and products mentioned in this book by the appropriate use of capitals. However, Packt Publishing cannot guarantee the accuracy of this information.

First published: September 2015

Second edition: September 2017

Production reference: 3231017

Published by Packt Publishing Ltd.
Livery Place
35 Livery Street
Birmingham B3 2PB, UK.

ISBN 978-1-78712-593-3

`www.packtpub.com`

Credits

Authors

Sebastian Raschka

Vahid Mirjalili

Reviewers

Jared Huffman

Huai-En, Sun (Ryan Sun)

Acquisition Editor

Frank Pohlmann

Content Development Editor

Chris Nelson

Project Editor

Monika Sangwan

Technical Editors

Bhagyashree Rai

Nidhisha Shetty

Copy Editor

Safis Editing

Project Coordinator

Suzanne Coutinho

Proofreader

Safis Editing

Indexer

Tejal Daruwale Soni

Graphics

Kirk D'Penha

Production Coordinator

Arvindkumar Gupta

About the Authors

Sebastian Raschka, the author of the bestselling book, *Python Machine Learning*, has many years of experience with coding in Python, and he has given several seminars on the practical applications of data science, machine learning, and deep learning including a machine learning tutorial at SciPy—the leading conference for scientific computing in Python.

While Sebastian's academic research projects are mainly centered around problem-solving in computational biology, he loves to write and talk about data science, machine learning, and Python in general, and he is motivated to help people develop data-driven solutions without necessarily requiring a machine learning background.

His work and contributions have recently been recognized by the departmental outstanding graduate student award 2016-2017 as well as the ACM Computing Reviews' Best of 2016 award. In his free time, Sebastian loves to contribute to open source projects, and the methods that he has implemented are now successfully used in machine learning competitions, such as Kaggle.

I would like to take this opportunity to thank the great Python community and developers of open source packages who helped me create the perfect environment for scientific research and data science. Also, I want to thank my parents who always encouraged and supported me in pursuing the path and career that I was so passionate about.

Special thanks to the core developers of scikit-learn. As a contributor to this project, I had the pleasure to work with great people who are not only very knowledgeable when it comes to machine learning but are also excellent programmers. Lastly, I'd like to thank Elie Kawerk, who volunteered to review the book and provided valuable feedback on the new chapters.

Vahid Mirjalili obtained his PhD in mechanical engineering working on novel methods for large-scale, computational simulations of molecular structures. Currently, he is focusing his research efforts on applications of machine learning in various computer vision projects at the department of computer science and engineering at Michigan State University.

Vahid picked Python as his number-one choice of programming language, and throughout his academic and research career he has gained tremendous experience with coding in Python. He taught Python programming to the engineering class at Michigan State University, which gave him a chance to help students understand different data structures and develop efficient code in Python.

While Vahid's broad research interests focus on deep learning and computer vision applications, he is especially interested in leveraging deep learning techniques to extend privacy in biometric data such as face images so that information is not revealed beyond what users intend to reveal. Furthermore, he also collaborates with a team of engineers working on self-driving cars, where he designs neural network models for the fusion of multispectral images for pedestrian detection.

I would like to thank my PhD advisor, Dr. Arun Ross, for giving me the opportunity to work on novel problems in his research lab. I also like to thank Dr. Vishnu Boddeti for inspiring my interests in deep learning and demystifying its core concepts.

About the Reviewers

Jared Huffman is an entrepreneur, gamer, storyteller, machine learning fanatic, and database aficionado. He has dedicated the past 10 years to developing software and analyzing data. His previous work has spanned a variety of topics, including network security, financial systems, and business intelligence, as well as web services, developer tools, and business strategy. Most recently, he was the founder of the data science team at Minecraft, with a focus on big data and machine learning. When not working, you can typically find him gaming or enjoying the beautiful Pacific Northwest with friends and family.

I'd like to thank Packt for giving me the opportunity to work on such a great book, my wife for the constant encouragement, and my daughter for sleeping through most of the late nights while I was reviewing and debugging code.

Huai-En, Sun (Ryan Sun) holds a master's degree in statistics from the National Chiao Tung University. He is currently working as a data scientist for analyzing the production line at PEGATRON. Machine learning and deep learning are his main areas of research.

Customer Feedback

Preface

Through exposure to the news and social media, you are probably aware of the fact that machine learning has become one of the most exciting technologies of our time and age. Large companies, such as Google, Facebook, Apple, Amazon, and IBM, heavily invest in machine learning research and applications for good reasons. While it may seem that machine learning has become the buzzword of our time and age, it is certainly not a fad. This exciting field opens the way to new possibilities and has become indispensable to our daily lives. This is evident in talking to the voice assistant on our smartphones, recommending the right product for our customers, preventing credit card fraud, filtering out spam from our email inboxes, detecting and diagnosing medical diseases, the list goes on and on.

If you want to become a machine learning practitioner, a better problem solver, or maybe even consider a career in machine learning research, then this book is for you. However, for a novice, the theoretical concepts behind machine learning can be quite overwhelming. Many practical books have been published in recent years that will help you get started in machine learning by implementing powerful learning algorithms.

Getting exposed to practical code examples and working through example applications of machine learning are a great way to dive into this field. Concrete examples help illustrate the broader concepts by putting the learned material directly into action. However, remember that with great power comes great responsibility! In addition to offering a hands-on experience with machine learning using the Python programming languages and Python-based machine learning libraries, this book introduces the mathematical concepts behind machine learning algorithms, which is essential for using machine learning successfully. Thus, this book is different from a purely practical book; it is a book that discusses the necessary details regarding machine learning concepts and offers intuitive yet informative explanations of how machine learning algorithms work, how to use them, and most importantly, how to avoid the most common pitfalls.

Currently, if you type "machine learning" as a search term in Google Scholar, it returns an overwhelmingly large number of publications—1,800,000. Of course, we cannot discuss the nitty-gritty of all the different algorithms and applications that have emerged in the last 60 years. However, in this book, we will embark on an exciting journey that covers all the essential topics and concepts to give you a head start in this field. If you find that your thirst for knowledge is not satisfied, this book references many useful resources that can be used to follow up on the essential breakthroughs in this field.

If you have already studied machine learning theory in detail, this book will show you how to put your knowledge into practice. If you have used machine learning techniques before and want to gain more insight into how machine learning actually works, this book is for you. Don't worry if you are completely new to the machine learning field; you have even more reason to be excited. Here is a promise that machine learning will change the way you think about the problems you want to solve and will show you how to tackle them by unlocking the power of data.

Before we dive deeper into the machine learning field, let's answer your most important question, "Why Python?" The answer is simple: it is powerful yet very accessible. Python has become the most popular programming language for data science because it allows us to forget about the tedious parts of programming and offers us an environment where we can quickly jot down our ideas and put concepts directly into action.

We, the authors, can truly say that the study of machine learning has made us better scientists, thinkers, and problem solvers. In this book, we want to share this knowledge with you. Knowledge is gained by learning. The key is our enthusiasm, and the real mastery of skills can only be achieved by practice. The road ahead may be bumpy on occasions and some topics may be more challenging than others, but we hope that you will embrace this opportunity and focus on the reward. Remember that we are on this journey together, and throughout this book, we will add many powerful techniques to your arsenal that will help us solve even the toughest problems the data-driven way.

What this book covers

Chapter 1, Giving Computers the Ability to Learn from Data, introduces you to the main subareas of machine learning in order to tackle various problem tasks. In addition, it discusses the essential steps for creating a typical machine learning model by building a pipeline that will guide us through the following chapters.

Chapter 2, Training Simple Machine Learning Algorithms for Classification, goes back to the origins of machine learning and introduces binary perceptron classifiers and adaptive linear neurons. This chapter is a gentle introduction to the fundamentals of pattern classification and focuses on the interplay of optimization algorithms and machine learning.

Chapter 3, A Tour of Machine Learning Classifiers Using scikit-learn, describes the essential machine learning algorithms for classification and provides practical examples using one of the most popular and comprehensive open source machine learning libraries: scikit-learn.

Chapter 4, Building Good Training Sets – Data Preprocessing, discusses how to deal with the most common problems in unprocessed datasets, such as missing data. It also discusses several approaches to identify the most informative features in datasets and teaches you how to prepare variables of different types as proper input for machine learning algorithms.

Chapter 5, Compressing Data via Dimensionality Reduction, describes the essential techniques to reduce the number of features in a dataset to smaller sets while retaining most of their useful and discriminatory information. It discusses the standard approach to dimensionality reduction via principal component analysis and compares it to supervised and nonlinear transformation techniques.

Chapter 6, Learning Best Practices for Model Evaluation and Hyperparameter Tuning, discusses the dos and don'ts for estimating the performances of predictive models. Moreover, it discusses different metrics for measuring the performance of our models and techniques to fine-tune machine learning algorithms.

Chapter 7, Combining Different Models for Ensemble Learning, introduces you to the different concepts of combining multiple learning algorithms effectively. It teaches you how to build ensembles of experts to overcome the weaknesses of individual learners, resulting in more accurate and reliable predictions.

Chapter 8, Applying Machine Learning to Sentiment Analysis, discusses the essential steps to transform textual data into meaningful representations for machine learning algorithms to predict the opinions of people based on their writing.

Chapter 9, Embedding a Machine Learning Model into a Web Application, continues with the predictive model from the previous chapter and walks you through the essential steps of developing web applications with embedded machine learning models.

Chapter 10, Predicting Continuous Target Variables with Regression Analysis, discusses the essential techniques for modeling linear relationships between target and response variables to make predictions on a continuous scale. After introducing different linear models, it also talks about polynomial regression and tree-based approaches.

Chapter 11, Working with Unlabeled Data – Clustering Analysis, shifts the focus to a different subarea of machine learning, unsupervised learning. We apply algorithms from three fundamental families of clustering algorithms to find groups of objects that share a certain degree of similarity.

Chapter 12, Implementing a Multilayer Artificial Neural Network from Scratch, extends the concept of gradient-based optimization, which we first introduced in *Chapter 2, Training Simple Machine Learning Algorithms for Classification*, to build powerful, multilayer neural networks based on the popular backpropagation algorithm in Python.

Chapter 13, Parallelizing Neural Network Training with TensorFlow, builds upon the knowledge from the previous chapter to provide you with a practical guide for training neural networks more efficiently. The focus of this chapter is on TensorFlow, an open source Python library that allows us to utilize multiple cores of modern GPUs.

Chapter 14, Going Deeper – The Mechanics of TensorFlow, covers TensorFlow in greater detail explaining its core concepts of computational graphs and sessions. In addition, this chapter covers topics such as saving and visualizing neural network graphs, which will come in very handy during the remaining chapters of this book.

Chapter 15, Classifying Images with Deep Convolutional Neural Networks, discusses deep neural network architectures that have become the new standard in computer vision and image recognition fields—convolutional neural networks. This chapter will discuss the main concepts between convolutional layers as a feature extractor and apply convolutional neural network architectures to an image classification task to achieve almost perfect classification accuracy.

Chapter 16, Modeling Sequential Data Using Recurrent Neural Networks, introduces another popular neural network architecture for deep learning that is especially well suited for working with sequential data and time series data. In this chapter, we will apply different recurrent neural network architectures to text data. We will start with a sentiment analysis task as a warm-up exercise and will learn how to generate entirely new text.

What you need for this book

The execution of the code examples provided in this book requires an installation of Python 3.6.0 or newer on macOS, Linux, or Microsoft Windows. We will make frequent use of Python's essential libraries for scientific computing throughout this book, including SciPy, NumPy, scikit-learn, Matplotlib, and pandas.

The first chapter will provide you with instructions and useful tips to set up your Python environment and these core libraries. We will add additional libraries to our repertoire; moreover, installation instructions are provided in the respective chapters: the NLTK library for natural language processing (*Chapter 8, Applying Machine Learning to Sentiment Analysis*), the Flask web framework (*Chapter 9, Embedding a Machine Learning Model into a Web Application*), the Seaborn library for statistical data visualization (*Chapter 10, Predicting Continuous Target Variables with Regression Analysis*), and TensorFlow for efficient neural network training on graphical processing units (*Chapters 13 to 16*).

Who this book is for

If you want to find out how to use Python to start answering critical questions of your data, pick up *Python Machine Learning, Second Edition*—whether you want to start from scratch or extend your data science knowledge, this is an essential and unmissable resource.

Conventions

In this book, you will find a number of text styles that distinguish between different kinds of information. Here are some examples of these styles and an explanation of their meaning.

Code words in text, database table names, folder names, filenames, file extensions, pathnames, dummy URLs, user input, and Twitter handles are shown as follows: "Using the `out_file=None` setting, we directly assigned the dot data to a `dot_data` variable, instead of writing an intermediate `tree.dot` file to disk."

A block of code is set as follows:

```
>>> from sklearn.neighbors import KNeighborsClassifier
>>> knn = KNeighborsClassifier(n_neighbors=5, p=2,
...                            metric='minkowski')
>>> knn.fit(X_train_std, y_train)
>>> plot_decision_regions(X_combined_std, y_combined,
...                       classifier=knn, test_idx=range(105,150))
>>> plt.xlabel('petal length [standardized]')
>>> plt.ylabel('petal width [standardized]')
>>> plt.show()
```

Any command-line input or output is written as follows:

```
pip3 install graphviz
```

New terms and **important words** are shown in bold. Words that you see on the screen, for example, in menus or dialog boxes, appear in the text like this: "After we click on the **Dashboard** button in the top-right corner, we have access to the control panel shown at the top of the page."

[Warnings or important notes appear in a box like this.]

[Tips and tricks appear like this.]

Reader feedback

Feedback from our readers is always welcome. Let us know what you think about this book—what you liked or disliked. Reader feedback is important for us as it helps us develop titles that you will really get the most out of.

To send us general feedback, simply email `feedback@packtpub.com`, and mention the book's title in the subject of your message.

If there is a topic that you have expertise in and you are interested in either writing or contributing to a book, see our author guide at `www.packtpub.com/authors`.

Customer support

Now that you are the proud owner of a Packt book, we have a number of things to help you to get the most from your purchase.

Downloading the example code

You can download the example code files for this book from your account at `http://www.packtpub.com`. If you purchased this book elsewhere, you can visit `http://www.packtpub.com/support` and register to have the files emailed directly to you.

You can download the code files by following these steps:

1. Log in or register to our website using your email address and password.
2. Hover the mouse pointer on the **SUPPORT** tab at the top.
3. Click on **Code Downloads & Errata**.
4. Enter the name of the book in the **Search** box.
5. Select the book for which you're looking to download the code files.
6. Choose from the drop-down menu where you purchased this book from.
7. Click on **Code Download**.

You can also download the code files by clicking on the **Code Files** button on the book's web page at the Packt Publishing website. This page can be accessed by entering the book's name in the **Search** box. Please note that you need to be logged in to your Packt account.

Once the file is downloaded, please make sure that you unzip or extract the folder using the latest version of:

- WinRAR / 7-Zip for Windows
- Zipeg / iZip / UnRarX for Mac
- 7-Zip / PeaZip for Linux

The code bundle for the book is also hosted on GitHub at `https://github.com/PacktPublishing/Python-Machine-Learning-Second-Edition`. We also have other code bundles from our rich catalog of books and videos available at `https://github.com/PacktPublishing/`. Check them out!

Downloading the color images of this book

We also provide you with a PDF file that has color images of the screenshots/diagrams used in this book. The color images will help you better understand the changes in the output. You can download this file from http://www.packtpub.com/sites/default/files/downloads/PythonMachineLearningSecondEdition_ColorImages.pdf. In addition, lower resolution color images are embedded in the code notebooks of this book that come bundled with the example code files.

Errata

Although we have taken every care to ensure the accuracy of our content, mistakes do happen. If you find a mistake in one of our books—maybe a mistake in the text or the code—we would be grateful if you could report this to us. By doing so, you can save other readers from frustration and help us improve subsequent versions of this book. If you find any errata, please report them by visiting http://www.packtpub.com/submit-errata, selecting your book, clicking on the **Errata Submission Form** link, and entering the details of your errata. Once your errata are verified, your submission will be accepted and the errata will be uploaded to our website or added to any list of existing errata under the Errata section of that title.

To view the previously submitted errata, go to https://www.packtpub.com/books/content/support and enter the name of the book in the search field. The required information will appear under the **Errata** section.

Piracy

Piracy of copyrighted material on the Internet is an ongoing problem across all media. At Packt, we take the protection of our copyright and licenses very seriously. If you come across any illegal copies of our works in any form on the Internet, please provide us with the location address or website name immediately so that we can pursue a remedy.

Please contact us at copyright@packtpub.com with a link to the suspected pirated material.

We appreciate your help in protecting our authors and our ability to bring you valuable content.

Questions

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