Community Experience Distilled

Learning Data Mining with Python

Harness the power of Python to analyze data and create insightful predictive models

Robert Layton

[PACKT] open source*
PUBLISHING
community experience distilled

Learning Data Mining with Python

Harness the power of Python to analyze data and create insightful predictive models

Robert Layton

[PACKT] PUBLISHING open source community experience distilled

BIRMINGHAM - MUMBAI

Learning Data Mining with Python

First published: July 2015

Production reference: 1230715

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

ISBN 978-1-78439-605-3

www.packtpub.com

Credits

Author
Robert Layton

Reviewers
Asad Ahamad
P Ashwin
Christophe Van Gysel
Edward C. Delaporte V

Commissioning Editor
Taron Pereira

Acquisition Editor
James Jones

Content Development Editor
Siddhesh Salvi

Technical Editor
Naveenkumar Jain

Copy Editors
Roshni Banerjee
Trishya Hajare

Project Coordinator
Nidhi Joshi

Proofreader
Safis Editing

Indexer
Priya Sane

Graphics
Sheetal Aute

Production Coordinator
Nitesh Thakur

Cover Work
Nitesh Thakur

About the Author

Robert Layton has a PhD in computer science and has been an avid Python programmer for many years. He has worked closely with some of the largest companies in the world on data mining applications for real-world data and has also been published extensively in international journals and conferences. He has extensive experience in cybercrime and text-based data analytics, with a focus on behavioral modeling, authorship analysis, and automated open source intelligence. He has contributed code to a number of open source libraries, including the scikit-learn library used in this book, and was a Google Summer of Code mentor in 2014. Robert runs a data mining consultancy company called dataPipeline, providing data mining and analytics solutions to businesses in a variety of industries.

About the Reviewers

Asad Ahamad is a data enthusiast and loves to work on data to solve challenging problems.

He did his master's degree in industrial mathematics with computer application at Jamia Millia Islamia, New Delhi. He admires mathematics a lot and always tries to use it to gain maximum profit for businesses.

He has good experience working in data mining, machine learning, and data science and has worked for various multinationals in India. He mainly uses R and Python to perform data wrangling and modeling. He is fond of using open source tools for data analysis.

He is an active social media user. Feel free to connect with him on Twitter at `@asadtaj88`.

P Ashwin is a Bangalore-based engineer who wears many different hats depending on the occasion. He graduated from IIIT, Hyderabad at in 2012 with an M Tech in computer science and engineering. He has a total of 5 years of experience in the software industry, where he has worked in different domains such as testing, data warehousing, replication, and automation. He is very well versed in DB concepts, SQL, and scripting with Bash and Python. He has earned professional certifications in products from Oracle, IBM, Informatica, and Teradata. He's also an ISTQB-certified tester.

In his free time, he volunteers in different technical hackathons or social service activities. He was introduced to Raspberry Pi in one of the hackathons and he's been hooked on it ever since. He writes a lot of code in Python, C, C++, and Shell on his Raspberry Pi B+ cluster. He's currently working on creating his own Beowulf cluster of 64 Raspberry Pi 2s.

Christophe Van Gysel is pursuing a doctorate degree in computer science at the University of Amsterdam under the supervision of Maarten de Rijke and Marcel Worring. He has interned at Google, where he worked on large-scale machine learning and automated speech recognition. During his internship in Facebook's security infrastructure team, he worked on information security and implemented measures against compression side-channel attacks. In the past, he was active as a security researcher. He discovered and reported security vulnerabilities in the web services of Google, Facebook, Dropbox, and PayPal, among others.

Edward C. Delaporte V leads a software development group at the University of Illinois, and he has contributed to the documentation of the Kivy framework. He is thankful to all those whose contributions to the open source community made his career possible, and he hopes this book helps continue to attract enthusiasts to software development.

www.PacktPub.com

Support files, eBooks, discount offers, and more

For support files and downloads related to your book, please visit `www.PacktPub.com`.

Did you know that Packt offers eBook versions of every book published, with PDF and ePub files available? You can upgrade to the eBook version at `www.PacktPub.com` and as a print book customer, you are entitled to a discount on the eBook copy. Get in touch with us at `service@packtpub.com` for more details.

At `www.PacktPub.com`, you can also read a collection of free technical articles, sign up for a range of free newsletters and receive exclusive discounts and offers on Packt books and eBooks.

`https://www2.packtpub.com/books/subscription/packtlib`

Do you need instant solutions to your IT questions? PacktLib is Packt's online digital book library. Here, you can search, access, and read Packt's entire library of books.

Why subscribe?

- Fully searchable across every book published by Packt
- Copy and paste, print, and bookmark content
- On demand and accessible via a web browser

Free access for Packt account holders

If you have an account with Packt at `www.PacktPub.com`, you can use this to access PacktLib today and view 9 entirely free books. Simply use your login credentials for immediate access.

Table of Contents