

Alexis Perrier

Effective Amazon Machine Learning

Machine learning in the cloud



Packt>

Effective Amazon Machine Learning

Predictive analytics is a complex domain requiring coding skills, an understanding of the mathematical concepts underpinning machine learning algorithms, and the ability to create compelling data visualizations. Following AWS simplifying Machine learning, this book will help you bring predictive analytics projects to fruition in three easy steps: data preparation, model tuning, and model selection.

This book will introduce you to the Amazon Machine Learning platform and will implement core data science concepts such as classification, regression, regularization, overfitting, model selection, and evaluation. Furthermore, you will learn to leverage the Amazon Web Service (AWS) ecosystem for extended access to data sources, implement real-time predictions, and run Amazon Machine Learning projects via the command line and the Python SDK.

Towards the end of the book, you will also learn how to apply these services to other problems, such as text mining, and to more complex datasets.

Things you will learn:

- Learn how to use the Amazon Machine Learning service from scratch for predictive analytics
- Gain hands-on experience of key Data Science concepts
- Solve classic regression and classification problems
- Run projects programmatically via the command line and the Python SDK
- Leverage the Amazon Web Service ecosystem to access extended data sources
- Implement streaming and advanced projects

Packt
www.packtpub.com

\$ 49.99 US
£ 41.99 UK

Prices do not include local sales
Tax or VAT where applicable



Effective Amazon Machine Learning

Alexis Perrier



Effective Amazon Machine Learning

Machine learning in the cloud

Alexis Perrier

Packt>

BIRMINGHAM - MUMBAI

Effective Amazon Machine Learning

Copyright © 2017 Packt Publishing

All rights reserved. No part of this book may be reproduced, stored in a retrieval system, or transmitted in any form or by any means, without the prior written permission of the publisher, except in the case of brief quotations embedded in critical articles or reviews.

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 author, 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: April 2017

Production reference: 1210417

Published by Packt Publishing Ltd.

Livery Place

35 Livery Street

Birmingham

B3 2PB, UK.

ISBN 978-1-78588-323-1

www.packtpub.com

Credits

Author

Alexis Perrier

Reviewer

Doug Ortiz

Commissioning Editor

Veena Pagare

Acquisition Editor

Vinay Argekar

Content Development Editor

Cheryl D'sa

Production Coordinator

Arvinkumar Gupta

Copy Editor

Manisha Sinha

Project Coordinator

Nidhi Joshi

Proofreader

Safis Editing

Indexer

Mariammal Chettiyar

Technical Editor

Karan Thakkar

About the Author

Alexis Perrier is a data scientist at Docent Health, a Boston-based startup. He works with Machine Learning and Natural Language Processing to improve patient experience in healthcare. Fascinated by the power of stochastic algorithms, he is actively involved in the data science community as an instructor, blogger, and presenter. He holds a Ph.D. in Signal Processing from Telecom ParisTech and resides in Boston, MA.

You can get in touch with him on twitter @alexip and by email at alexis.perrier@gmail.com.

About the Reviewer

Doug Ortiz is an independent consultant who has been architecting, developing, and integrating enterprise solutions throughout his whole career. Organizations that leverage his skillset have been able to rediscover and reuse their underutilized data via existing and emerging technologies, such as Microsoft BI Stack, Hadoop, NoSQL databases, SharePoint, .Net, and related toolsets and technologies.

He is the founder of Illustris, LLC, and can be reached at dougortiz@illustris.org.

Interesting aspects of his profession are listed here:

- Has experience integrating multiple platforms and products
- Helps organizations gain a deeper understanding of the value of their current investments in data and existing resources, turning them into useful sources of information
- Has improved, salvaged, and architected projects by utilizing unique and innovative techniques

His hobbies include yoga and scuba diving.

www.PacktPub.com

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://www.packtpub.com/mapt>

Get the most in-demand software skills with Mapt. Mapt gives you full access to all Packt books and video courses, as well as industry-leading tools to help you plan your personal development and advance your career.

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

Customer Feedback

Thanks for purchasing this Packt book. At Packt, quality is at the heart of our editorial process. To help us improve, please leave us an honest review on this book's Amazon page at <https://www.amazon.com/dp/1785883232>.

If you'd like to join our team of regular reviewers, you can e-mail us at customerreviews@packtpub.com. We award our regular reviewers with free eBooks and videos in exchange for their valuable feedback. Help us be relentless in improving our products!

This book is dedicated to the love of my life Janne and to our beautiful children Astrid, Léonard and Lucas.

Table of Contents

Preface	1
Chapter 1: Introduction to Machine Learning and Predictive Analytics	7
Introducing Amazon Machine Learning	8
Machine Learning as a Service	8
Leveraging full AWS integration	9
Comparing performances	10
Engineering data versus model variety	11
Amazon's expertise and the gradient descent algorithm	12
Pricing	12
Understanding predictive analytics	13
Building the simplest predictive analytics algorithm	17
Regression versus classification	20
Expanding regression to classification with logistic regression	20
Extracting features to predict outcomes	22
Diving further into linear modeling for prediction	23
Validating the dataset	26
Missing from Amazon ML	27
The statistical approach versus the machine learning approach	28
Summary	29
Chapter 2: Machine Learning Definitions and Concepts	31
What's an algorithm? What's a model?	32
Dealing with messy data	33
Classic datasets versus real-world datasets	33
Assumptions for multiclass linear models	34
Missing values	34
Normalization	35
Imbalanced datasets	37
Addressing multicollinearity	39
Detecting outliers	40
Accepting non-linear patterns	42
Adding features?	45
Preprocessing recapitulation	45
The predictive analytics workflow	46
Training and evaluation in Amazon ML	47

Identifying and correcting poor performances	49
Underfitting	49
Overfitting	50
Regularization on linear models	52
L2 regularization and Ridge	53
L1 regularization and Lasso	53
Evaluating the performance of your model	54
Summary	57
Chapter 3: Overview of an Amazon Machine Learning Workflow	59
<hr/>	
Opening an Amazon Web Services Account	60
Security	60
Setting up the account	61
Creating a user	63
Defining policies	64
Creating login credentials	65
Choosing a region	68
Overview of a standard Amazon Machine Learning workflow	69
The dataset	70
Loading the data on S3	71
Declaring a datasource	73
Creating the datasource	74
The model	76
The evaluation of the model	79
Comparing with a baseline	79
Making batch predictions	81
Summary	85
Chapter 4: Loading and Preparing the Dataset	87
<hr/>	
Working with datasets	87
Finding open datasets	88
Introducing the Titanic dataset	89
Preparing the data	90
Splitting the data	90
Loading data on S3	91
Creating a bucket	92
Loading the data	93
Granting permissions	93
Formatting the data	95
Creating the datasource	96
Verifying the data schema	98
Reusing the schema	102

Examining data statistics	106
Feature engineering with Athena	109
Introducing Athena	111
A brief tour of AWS Athena	112
Creating a titanic database	113
Using the wizard	114
Creating the database and table directly in SQL	114
Data munging in SQL	116
Missing values	117
Handling outliers in the fare	117
Extracting the title from the name	118
Inferring the deck from the cabin	118
Calculating family size	118
Wrapping up	118
Creating an improved datasource	119
Summary	121
Chapter 5: Model Creation	123
<hr/>	
Transforming data with recipes	124
Managing variables	125
Grouping variables	125
Naming variables with assignments	126
Specifying outputs	127
Data processing through seven transformations	127
Using simple transformations	128
Text mining	128
Coupling variables	130
Binning numeric values	131
Creating a model	132
Editing the suggested recipe	133
Applying recipes to the Titanic dataset	134
Choosing between recipes and data pre-processing.	135
Parametrizing the model	137
Setting model memory	138
Setting the number of data passes	138
Choosing regularization	138
Creating an evaluation	138
Evaluating the model	140
Evaluating binary classification	141
Exploring the model performances	143
Evaluating linear regression	145
Evaluating multiclass classification	147
Analyzing the logs	148
Optimizing the learning rate	149

Visualizing convergence	151
Impact of regularization	153
Comparing different recipes on the Titanic dataset	155
Keeping variables as numeric or applying quantile binning?	155
Parsing the model logs	157
Summary	159
Chapter 6: Predictions and Performances	161
<hr/>	
Making batch predictions	162
Creating the batch prediction job	163
Interpreting prediction outputs	165
Reading the manifest file	166
Reading the results file	166
Assessing our predictions	167
Evaluating the held-out dataset	169
Finding out who will survive	172
Multiplying trials	172
Making real-time predictions	174
Manually exploring variable influence	174
Setting up real-time predictions	175
AWS SDK	176
Setting up AWS credentials	176
AWS access keys	176
Setting up AWS CLI	177
Python SDK	178
Summary	185
Chapter 7: Command Line and SDK	187
<hr/>	
Getting started and setting up	188
Using the CLI versus SDK	188
Installing AWS CLI	189
Picking up CLI syntax	191
Passing parameters using JSON files	192
Introducing the Ames Housing dataset	193
Splitting the dataset with shell commands	194
A simple project using the CLI	195
An overview of Amazon ML CLI commands	195
Creating the datasource	197
Creating the model	201
Evaluating our model with create-evaluation	202
What is cross-validation?	204
Implementing Monte Carlo cross-validation	205
Generating the shuffled datasets	205
Generating the datasources template	207