

Max Kuhn · Kjell Johnson

Applied Predictive Modeling

应用预测建模

Springer

世界图书出版公司
www.wpcbj.com.cn

Max Kuhn • Kjell Johnson

Applied Predictive Modeling

 Springer

Max Kuhn
Division of Nonclinical Statistics
Pfizer Global Research and
Development
Groton, Connecticut, USA

Kjell Johnson
Arbor Analytics
Saline, Michigan, USA

ISBN 978-1-4614-6848-6 ISBN 978-1-4614-6849-3 (eBook)
DOI 10.1007/978-1-4614-6849-3
Springer New York Heidelberg Dordrecht London

Library of Congress Control Number: 2013933452

© Springer Science+Business Media New York 2013

This work is subject to copyright. All rights are reserved by the Publisher, whether the whole or part of the material is concerned, specifically the rights of translation, reprinting, reuse of illustrations, recitation, broadcasting, reproduction on microfilms or in any other physical way, and transmission or information storage and retrieval, electronic adaptation, computer software, or by similar or dissimilar methodology now known or hereafter developed. Exempted from this legal reservation are brief excerpts in connection with reviews or scholarly analysis or material supplied specifically for the purpose of being entered and executed on a computer system, for exclusive use by the purchaser of the work. Duplication of this publication or parts thereof is permitted only under the provisions of the Copyright Law of the Publisher's location, in its current version, and permission for use must always be obtained from Springer. Permissions for use may be obtained through RightsLink at the Copyright Clearance Center. Violations are liable to prosecution under the respective Copyright Law.

The use of general descriptive names, registered names, trademarks, service marks, etc. in this publication does not imply, even in the absence of a specific statement, that such names are exempt from the relevant protective laws and regulations and therefore free for general use.

While the advice and information in this book are believed to be true and accurate at the date of publication, neither the authors nor the editors nor the publisher can accept any legal responsibility for any errors or omissions that may be made. The publisher makes no warranty, express or implied, with respect to the material contained herein.

Reprint from English language edition:

Applied Predictive Modeling

by Max Kuhn, Kjell Johnson

Copyright © Springer Science+Business Media New York 2013

Springer is a part of Springer Science+Business Media

All Rights Reserved

This reprint has been authorized by Springer Science & Business Media for distribution in China Mainland only and not for export therefrom.

图书在版编目 (CIP) 类

应用预测建模 = Applied Predictive Modeling : 英文 / (美) M.
库恩 (M. Kuhn), (美) K. 约翰逊 (K. Johnson) 著 . —
影印本 . —北京 : 世界图书出版公司北京公司 , 2016.11
ISBN 978-7-5192-2089-1

I . ①应… II . ①库… ②约… III . ①系统建模—英文
IV . ① O2

中国版本图书馆 CIP 数据核字 (2016) 第 271804 号

著 者 : Max Kuhn, Kjell Johnson
责任编辑 : 刘 慧 高 蓉
装帧设计 : 任志远

出版发行 : 世界图书出版公司北京公司
地 址 : 北京市东城区朝内大街 137 号
邮 编 : 100010
电 话 : 010-64038355 (发行) 64015580 (客服) 64033507 (总编室)
网 址 : <http://www.wpcbj.com.cn>
邮 箱 : wpcbjst@vip.163.com
销 售 : 新华书店
印 刷 : 北京博图彩色印刷有限公司
开 本 : 711mm × 1245mm 1/24
印 张 : 26
字 数 : 499 千
版 次 : 2017 年 6 月第 1 版 2017 年 6 月第 1 次印刷
版权登记 : 01-2016-6778
定 价 : 159.00 元

版权所有 翻印必究

(如发现印装质量问题 , 请与所购图书销售部门联系调换)

Applied Predictive Modeling

To our families:

Miranda and Stefan

Valerie, Truman, and baby Gideon

Preface

This is a book on *data analysis* with a specific focus on the *practice of predictive modeling*. The term predictive modeling may stir associations such as machine learning, pattern recognition, and data mining. Indeed, these associations are appropriate and the methods implied by these terms are an integral piece of the predictive modeling process. But predictive modeling encompasses much more than the tools and techniques for uncovering patterns within data. The practice of predictive modeling defines the process of developing a model in a way that we can understand and quantify the model's prediction accuracy on future, yet-to-be-seen data. The *entire* process is the focus of this book.

We intend this work to be a practitioner's guide to the predictive modeling process and a place where one can come to learn about the approach and to gain intuition about the many commonly used and modern, powerful models. A host of statistical and mathematical techniques are discussed, but our motivation in almost every case is to describe the techniques in a way that helps develop intuition for its strengths and weaknesses instead of its mathematical genesis and underpinnings. For the most part we avoid complex equations, although there are a few necessary exceptions. For more theoretical treatments of predictive modeling, we suggest Hastie et al. (2008) and Bishop (2006). For this text, the reader should have some knowledge of basic statistics, including variance, correlation, simple linear regression, and basic hypothesis testing (e.g. p -values and test statistics).

The predictive modeling process is inherently hands-on. But during our research for this work we found that many articles and texts prevent the reader from reproducing the results either because the data were not freely available or because the software was inaccessible or only available for purchase. Buckheit and Donoho (1995) provide a relevant critique of the traditional scholarly veil:

An article about computational science in a scientific publication is not the scholarship itself, it is merely advertising of the scholarship. The actual

scholarship is the complete software development environment and the complete set of instructions which generated the figures.

Therefore, it was our goal to be as hands-on as possible, enabling the readers to reproduce the results within reasonable precision as well as being able to naturally extend the predictive modeling approach to their own data. Furthermore, we use the R language (Ihaka and Gentleman 1996; R Development Core Team 2010), a freely accessible software for statistical and mathematical calculations, for all stages of the predictive modeling process. Almost all of the example data sets are available in R packages. The `AppliedPredictiveModeling` R package contains many of the data sets used here as well as R scripts to reproduce the analyses in each chapter.

We selected R as the computational engine of this text for several reasons. First R is freely available (although commercial versions exist) for multiple operating systems. Second, it is released under the *General Public License* (Free Software Foundation June 2007), which outlines how the program can be redistributed. Under this structure anyone is free to examine and modify the source code. Because of this open-source nature, dozens of predictive models have already been implemented through freely available packages. Moreover R contains extensive, powerful capabilities for the overall predictive modeling process. Readers not familiar with R can find numerous tutorials online. We also provide an introduction and start-up guide for R in the Appendix.

There are a few topics that we didn't have time and/or space to add, most notably: generalized additive models, ensembles of different models, network models, time series models, and a few others.

There is also a web site for the book:

<http://appliedpredictivemodeling.com/>

that will contain relevant information.

This work would not have been possible without the help and mentoring from many individuals, including: Walter H. Carter, Jim Garrett, Chris Gennings, Paul Harms, Chris Keefer, William Klinger, Daijin Ko, Rich Moore, David Neuhouser, David Potter, David Pyne, William Rayens, Arnold Stromberg, and Thomas Vidmar. We would also like to thank Ross Quinlan for his help with Cubist and C5.0 and vetting our descriptions of the two. At Springer, we would like to thank Marc Strauss and Hannah Bracken as well as the reviewers: Vini Bonato, Thomas Miller, Ross Quinlan, Eric Siegel, Stan Young, and an anonymous reviewer. Lastly, we would like to thank our families for their support: Miranda Kuhn, Stefan Kuhn, Bobby Kuhn, Robert Kuhn, Karen Kuhn, and Mary Ann Kuhn; Warren and Kay Johnson; and Valerie and Truman Johnson.

Groton, CT, USA
Saline, MI, USA

Max Kuhn
Kjell Johnson

Contents

1	Introduction	1
1.1	Prediction Versus Interpretation	4
1.2	Key Ingredients of Predictive Models	5
1.3	Terminology	6
1.4	Example Data Sets and Typical Data Scenarios	7
1.5	Overview	14
1.6	Notation	15
 Part I General Strategies		
2	A Short Tour of the Predictive Modeling Process	19
2.1	Case Study: Predicting Fuel Economy	19
2.2	Themes	24
2.3	Summary	26
3	Data Pre-processing	27
3.1	Case Study: Cell Segmentation in High-Content Screening	28
3.2	Data Transformations for Individual Predictors	30
3.3	Data Transformations for Multiple Predictors	33
3.4	Dealing with Missing Values	41
3.5	Removing Predictors	43
3.6	Adding Predictors	47
3.7	Binning Predictors	49
3.8	Computing	51
	Exercises	58
4	Over-Fitting and Model Tuning	61
4.1	The Problem of Over-Fitting	62
4.2	Model Tuning	64
4.3	Data Splitting	67
4.4	Resampling Techniques	69

4.5	Case Study: Credit Scoring	73
4.6	Choosing Final Tuning Parameters	74
4.7	Data Splitting Recommendations	77
4.8	Choosing Between Models	78
4.9	Computing	80
	Exercises	89

Part II Regression Models

5	Measuring Performance in Regression Models	95
5.1	Quantitative Measures of Performance	95
5.2	The Variance-Bias Trade-off	97
5.3	Computing	98
6	Linear Regression and Its Cousins	101
6.1	Case Study: Quantitative Structure-Activity Relationship Modeling	102
6.2	Linear Regression	105
6.3	Partial Least Squares	112
6.4	Penalized Models	122
6.5	Computing	128
	Exercises	137
7	Nonlinear Regression Models	141
7.1	Neural Networks	141
7.2	Multivariate Adaptive Regression Splines	145
7.3	Support Vector Machines	151
7.4	K -Nearest Neighbors	159
7.5	Computing	161
	Exercises	168
8	Regression Trees and Rule-Based Models	173
8.1	Basic Regression Trees	175
8.2	Regression Model Trees	184
8.3	Rule-Based Models	190
8.4	Bagged Trees	192
8.5	Random Forests	198
8.6	Boosting	203
8.7	Cubist	208
8.8	Computing	212
	Exercises	218

9	A Summary of Solubility Models	221
10	Case Study: Compressive Strength of Concrete Mixtures	225
10.1	Model Building Strategy	229
10.2	Model Performance	230
10.3	Optimizing Compressive Strength	233
10.4	Computing	236
Part III Classification Models		
11	Measuring Performance in Classification Models	247
11.1	Class Predictions	247
11.2	Evaluating Predicted Classes	254
11.3	Evaluating Class Probabilities	262
11.4	Computing	266
12	Discriminant Analysis and Other Linear Classification Models	275
12.1	Case Study: Predicting Successful Grant Applications	275
12.2	Logistic Regression	282
12.3	Linear Discriminant Analysis	287
12.4	Partial Least Squares Discriminant Analysis	297
12.5	Penalized Models	302
12.6	Nearest Shrunken Centroids	306
12.7	Computing	308
	Exercises	326
13	Nonlinear Classification Models	329
13.1	Nonlinear Discriminant Analysis	329
13.2	Neural Networks	333
13.3	Flexible Discriminant Analysis	338
13.4	Support Vector Machines	343
13.5	<i>K</i> -Nearest Neighbors	350
13.6	Naïve Bayes	353
13.7	Computing	358
	Exercises	366
14	Classification Trees and Rule-Based Models	369
14.1	Basic Classification Trees	370
14.2	Rule-Based Models	383
14.3	Bagged Trees	385
14.4	Random Forests	386
14.5	Boosting	389
14.6	C5.0	392

14.7 Comparing Two Encodings of Categorical Predictors	400
14.8 Computing	400
Exercises	411
15 A Summary of Grant Application Models	415
16 Remedies for Severe Class Imbalance	419
16.1 Case Study: Predicting Caravan Policy Ownership	419
16.2 The Effect of Class Imbalance	420
16.3 Model Tuning	423
16.4 Alternate Cutoffs	423
16.5 Adjusting Prior Probabilities	426
16.6 Unequal Case Weights	426
16.7 Sampling Methods	427
16.8 Cost-Sensitive Training	429
16.9 Computing	435
Exercises	442
17 Case Study: Job Scheduling	445
17.1 Data Splitting and Model Strategy	450
17.2 Results	454
17.3 Computing	457
Part IV Other Considerations	
18 Measuring Predictor Importance	463
18.1 Numeric Outcomes	464
18.2 Categorical Outcomes	468
18.3 Other Approaches	472
18.4 Computing	478
Exercises	484
19 An Introduction to Feature Selection	487
19.1 Consequences of Using Non-informative Predictors	488
19.2 Approaches for Reducing the Number of Predictors	490
19.3 Wrapper Methods	491
19.4 Filter Methods	499
19.5 Selection Bias	500
19.6 Case Study: Predicting Cognitive Impairment	502
19.7 Computing	511
Exercises	518

20	Factors That Can Affect Model Performance	521
20.1	Type III Errors	522
20.2	Measurement Error in the Outcome	524
20.3	Measurement Error in the Predictors	527
20.4	Discretizing Continuous Outcomes	531
20.5	When Should You Trust Your Model's Prediction?	534
20.6	The Impact of a Large Sample	538
20.7	Computing	541
	Exercises	542

Appendix

A	A Summary of Various Models	549
B	An Introduction to R	551
B.1	Start-Up and Getting Help	551
B.2	Packages	552
B.3	Creating Objects	553
B.4	Data Types and Basic Structures	554
B.5	Working with Rectangular Data Sets	558
B.6	Objects and Classes	560
B.7	R Functions	561
B.8	The Three Faces of =	562
B.9	The AppliedPredictiveModeling Package	562
B.10	The caret Package	563
B.11	Software Used in this Text	565

C	Interesting Web Sites	567
---	-----------------------------	-----

References	569
------------------	-----

Indices

Computing	591
-----------------	-----

General	595
---------------	-----

Part I
General Strategies

Chapter 1

Introduction

Every day people are faced with questions such as “What route should I take to work today?” “Should I switch to a different cell phone carrier?” “How should I invest my money?” or “Will I get cancer?” These questions indicate our desire to know future events, and we earnestly want to make the best decisions towards that future.

We usually make decisions based on information. In some cases we have tangible, objective data, such as the morning traffic or weather report. Other times we use intuition and experience like “I should avoid the bridge this morning because it usually gets bogged down when it snows” or “I should have a PSA test because my father got prostate cancer.” In either case, we are predicting future events given the information and experience we currently have, and we are making decisions based on those predictions.

As information has become more readily available via the internet and media, our desire to use this information to help us make decisions has intensified. And while the human brain can consciously and subconsciously assemble a vast amount of data, it cannot process the even greater amount of easily obtainable, relevant information for the problem at hand. To aid in our decision-making processes, we now turn to tools like Google to filter billions of web pages to find the most appropriate information for our queries, WebMD to diagnose our illnesses based on our symptoms, and E*TRADE to screen thousands of stocks and identify the best investments for our portfolios.

These sites, as well as many others, use tools that take our current information, sift through data looking for patterns that are relevant to our problem, and return answers. The process of developing these kinds of tools has evolved throughout a number of fields such as chemistry, computer science, physics, and statistics and has been called “machine learning,” “artificial intelligence,” “pattern recognition,” “data mining,” “predictive analytics,” and “knowledge discovery.” While each field approaches the problem using different perspectives and tool sets, the ultimate objective is the same: *to make an accurate prediction*. For this book, we will pool these terms into the commonly used phrase *predictive modeling*.

Geisser (1993) defines predictive modeling as “the process by which a model is created or chosen to try to best predict the probability of an outcome.” We tweak this definition slightly:

Predictive modeling: the process of developing a mathematical tool or model that generates an accurate prediction

Steve Levy of *Wired* magazine recently wrote of the increasing presence of predictive models (Levy 2010), “Examples [of artificial intelligence] can be found everywhere: The Google global machine uses AI to interpret cryptic human queries. Credit card companies use it to track fraud. Netflix uses it to recommend movies to subscribers. And the financial system uses it to handle billions of trades (with only the occasional meltdown).” Examples of the types of questions one would like to predict are:

- How many copies will this book sell?
- Will this customer move their business to a different company?
- How much will my house sell for in the current market?
- Does a patient have a specific disease?
- Based on past choices, which movies will interest this viewer?
- Should I sell this stock?
- Which people should we match in our online dating service?
- Is an e-mail spam?
- Will this patient respond to this therapy?

Insurance companies, as another example, must predict the risks of potential auto, health, and life policy holders. This information is then used to determine if an individual will receive a policy, and if so, at what premium. Like insurance companies, governments also seek to predict risks, but for the purpose of protecting their citizens. Recent examples of governmental predictive models include biometric models for identifying terror suspects, models of fraud detection (Westphal 2008), and models of unrest and turmoil (Shachtman 2011). Even a trip to the grocery store or gas station [everyday places where our purchase information is collected and analyzed in an attempt to understand who we are and what we want (Duhigg 2012)] brings us into the predictive modeling world, and we’re often not even aware that we’ve entered it. Predictive models now *permeate our existence*.

While predictive models guide us towards more satisfying products, better medical treatments, and more profitable investments, they regularly generate inaccurate predictions and provide the wrong answers. For example, most of us have not received an important e-mail due to a predictive model (a.k.a. e-mail filter) that incorrectly identified the message as spam. Similarly, predictive models (a.k.a. medical diagnostic models) misdiagnose diseases, and predictive models (a.k.a. financial algorithms) erroneously buy and sell stocks predicting profits when, in reality, finding losses. This final example of predictive models gone wrong affected many investors in 2010. Those who follow the stock market are likely familiar with the “flash crash” on May 6, 2010,