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

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）数

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

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 Max Kuhn
Saline, MI, USA Kjell Johnson

Contents

Part I General Strategies

Part II Regression Models

Part III Classification Models

Part IV Other Considerations

Appendix

Indicies

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.*

M. Kuhn and K. Johnson, *Applied Predictive Modeling*,
DOI 10.1007/978-1-4614-6849-3_1,

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,