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

http://www.springer.com/978-1-4614-6848-6

Applied Predictive Modeling

Kuhn, M.; Johnson, K.

2013, XIII, 600 p. 204 illus., Hardcover
ISBN: 978-1-4614-6848-6