

Springer Actuarial Lecture Notes

Michel Denuit
Donatien Hainaut
Julien Trufin

Effective Statistical Learning Methods for Actuaries II

Tree-Based Methods and Extensions

 Springer

Springer Actuarial

Springer Actuarial Lecture Notes

Editors-in-Chief

Hansjoerg Albrecher, University of Lausanne, Lausanne, Switzerland

Michael Sherris, UNSW, Sydney, NSW, Australia

Series Editors

Daniel Bauer, University of Wisconsin-Madison, Madison, WI, USA

Stéphane Loisel, ISFA, Université Lyon 1, Lyon, France

Alexander J. McNeil, University of York, York, UK

Antoon Pelsser, Maastricht University, Maastricht, The Netherlands

Ermanno Pitacco, Università di Trieste, Trieste, Italy

Gordon Willmot, University of Waterloo, Waterloo, ON, Canada

Hailiang Yang, The University of Hong Kong, Hong Kong, Hong Kong

This subseries of Springer Actuarial includes books with the character of lecture notes. Typically these are research monographs on new, cutting-edge developments in actuarial science; sometimes they may be a glimpse of a new field of research activity, or presentations of a new angle in a more classical field.

In the established tradition of Lecture Notes, the timeliness of a manuscript can be more important than its form, which may be informal, preliminary or tentative.

More information about this subseries at <http://www.springer.com/series/15682>

Michel Denuit · Donatien Hainaut ·
Julien Trufin

Effective Statistical Learning Methods for Actuaries II

Tree-Based Methods and Extensions

 Springer

Michel Denuit
Institut de Statistique, Biostatistique et
Sciences Actuarielles (ISBA)
Université Catholique Louvain
Louvain-la-Neuve, Belgium

Donatien Hainaut
Institut de Statistique, Biostatistique et
Sciences Actuarielles (ISBA)
Université Catholique Louvain
Louvain-la-Neuve, Belgium

Julien Trufin
Département de Mathématiques
Université Libre de Bruxelles
Brussels, Belgium

ISSN 2523-3262
Springer Actuarial

ISSN 2523-3270 (electronic)

ISSN 2523-3289
Springer Actuarial Lecture Notes

ISSN 2523-3297 (electronic)

ISBN 978-3-030-57555-7

ISBN 978-3-030-57556-4 (eBook)

<https://doi.org/10.1007/978-3-030-57556-4>

Mathematics Subject Classification: 62P05, 62-XX, 68-XX, 62M45

© Springer Nature Switzerland AG 2020

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.

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.

The publisher, the authors and the editors are safe to assume that the advice and information in this book are believed to be true and accurate at the date of publication. Neither the publisher nor the authors or the editors give a warranty, expressed or implied, with respect to the material contained herein or for any errors or omissions that may have been made. The publisher remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

This Springer imprint is published by the registered company Springer Nature Switzerland AG
The registered company address is: Gewerbestrasse 11, 6330 Cham, Switzerland

Preface

The present material is written for students enrolled in actuarial master programs and practicing actuaries, who would like to gain a better understanding of insurance data analytics. It is built in three volumes, starting from the celebrated Generalized Linear Models, or GLMs and continuing with tree-based methods and neural networks.

This second volume summarizes the state of the art using regression trees and their various combinations such as random forests and boosting trees. This second volume also goes through tools enabling to assess the predictive accuracy of regression models. Throughout this book, we alternate between methodological aspects and numerical illustrations or case studies to demonstrate practical applications of the proposed techniques. The R statistical software has been found convenient to perform the analyses throughout this book. It is a free language and environment for statistical computing and graphics. In addition to our own R code, we have benefited from many R packages contributed by the members of the very active community of R-users. The open-source statistical software R is freely available from <https://www.r-project.org/>.

The technical requirements to understand the material are kept at a reasonable level so that this text is meant for a broad readership. We refrain from proving all results but rather favor an intuitive approach with supportive numerical illustrations, providing the reader with relevant references where all justifications can be found, as well as more advanced material. These references are gathered in a dedicated section at the end of each chapter.

The three authors are professors of actuarial mathematics at the universities of Brussels and Louvain-la-Neuve, Belgium. Together, they accumulate decades of teaching experience related to the topics treated in the three books, in Belgium and throughout Europe and Canada. They are also scientific directors at Detralytics, a consulting office based in Brussels.

Within Detralytics as well as on behalf of actuarial associations, the authors have had the opportunity to teach the material contained in the three volumes of “Effective Statistical Learning Methods for Actuaries” to various audiences of practitioners. The feedback received from the participants to these short courses

greatly helped to improve the exposition of the topic. Throughout their contacts with the industry, the authors also implemented these techniques in a variety of consulting and R&D projects. This makes the three volumes of “Effective Statistical Learning Methods for Actuaries” the ideal support for teaching students and CPD events for professionals.

Louvain-la-Neuve, Belgium
Louvain-la-Neuve, Belgium
Brussels, Belgium
September 2020

Michel Denuit
Donatien Hainaut
Julien Trufin

Contents

| | | |
|----------|---|----------|
| 1 | Introduction | 1 |
| 1.1 | The Risk Classification Problem | 1 |
| 1.1.1 | Insurance Risk Diversification | 1 |
| 1.1.2 | Why Classifying Risks? | 1 |
| 1.1.3 | The Need for Regression Models | 2 |
| 1.1.4 | Observable Versus Hidden Risk Factors | 2 |
| 1.1.5 | Insurance Ratemaking Versus Loss Prediction | 3 |
| 1.2 | Insurance Data | 3 |
| 1.2.1 | Claim Data | 3 |
| 1.2.2 | Frequency-Severity Decomposition | 4 |
| 1.2.3 | Observational Data | 5 |
| 1.2.4 | Format of the Data | 6 |
| 1.2.5 | Data Quality Issues | 7 |
| 1.3 | Exponential Dispersion (ED) Distributions | 8 |
| 1.3.1 | Frequency and Severity Distributions | 8 |
| 1.3.2 | From Normal to ED Distributions | 9 |
| 1.3.3 | Some ED Distributions | 11 |
| 1.3.4 | Mean and Variance | 17 |
| 1.3.5 | Weights | 20 |
| 1.3.6 | Exposure-to-Risk | 21 |
| 1.4 | Maximum Likelihood Estimation | 22 |
| 1.4.1 | Likelihood-Based Statistical Inference | 22 |
| 1.4.2 | Maximum-Likelihood Estimator | 22 |
| 1.4.3 | Derivation of the Maximum-Likelihood Estimate | 23 |
| 1.4.4 | Properties of the Maximum-Likelihood Estimators | 24 |
| 1.4.5 | Examples | 26 |
| 1.5 | Deviance | 28 |

| | | |
|----------|---|-----------|
| 1.6 | Actuarial Pricing and Tree-Based Methods | 29 |
| 1.7 | Bibliographic Notes and Further Reading | 33 |
| | References | 33 |
| 2 | Performance Evaluation | 35 |
| 2.1 | Introduction | 35 |
| 2.2 | Generalization Error | 35 |
| 2.2.1 | Definition | 35 |
| 2.2.2 | Loss Function | 36 |
| 2.2.3 | Estimates | 37 |
| 2.2.4 | Decomposition | 40 |
| 2.3 | Expected Generalization Error | 44 |
| 2.3.1 | Squared Error Loss | 44 |
| 2.3.2 | Poisson Deviance Loss | 46 |
| 2.3.3 | Gamma Deviance Loss | 46 |
| 2.3.4 | Bias and Variance | 47 |
| 2.4 | (Expected) Generalization Error for Randomized Training Procedures | 47 |
| 2.5 | Bibliographic Notes and Further Reading | 49 |
| | References | 49 |
| 3 | Regression Trees | 51 |
| 3.1 | Introduction | 51 |
| 3.2 | Binary Regression Trees | 51 |
| 3.2.1 | Selection of the Splits | 53 |
| 3.2.2 | The Prediction in Each Terminal Node | 55 |
| 3.2.3 | The Rule to Determine When a Node Is Terminal | 57 |
| 3.2.4 | Examples | 59 |
| 3.3 | Right Sized Trees | 70 |
| 3.3.1 | Minimal Cost-Complexity Pruning | 72 |
| 3.3.2 | Choice of the Best Pruned Tree | 80 |
| 3.4 | Measure of Performance | 89 |
| 3.5 | Relative Importance of Features | 90 |
| 3.5.1 | Example 1 | 91 |
| 3.5.2 | Example 2 | 92 |
| 3.5.3 | Effect of Correlated Features | 93 |
| 3.6 | Interactions | 95 |
| 3.7 | Limitations of Trees | 96 |
| 3.7.1 | Model Instability | 96 |
| 3.7.2 | Lack of Smoothness | 101 |
| 3.8 | Bibliographic Notes and Further Reading | 103 |
| | References | 105 |

- 4 Bagging Trees and Random Forests** 107
 - 4.1 Introduction 107
 - 4.2 Bootstrap 108
 - 4.3 Bagging Trees 109
 - 4.3.1 Bias 111
 - 4.3.2 Variance 111
 - 4.3.3 Expected Generalization Error 114
 - 4.4 Random Forests 119
 - 4.5 Out-of-Bag Estimate 121
 - 4.6 Interpretability 121
 - 4.6.1 Relative Importances 122
 - 4.6.2 Partial Dependence Plots 123
 - 4.7 Example 125
 - 4.8 Bibliographic Notes and Further Reading 128
 - References 130

- 5 Boosting Trees** 131
 - 5.1 Introduction 131
 - 5.2 Forward Stagewise Additive Modeling 131
 - 5.3 Boosting Trees 133
 - 5.3.1 Algorithm 134
 - 5.3.2 Particular Cases 135
 - 5.3.3 Size of the Trees 143
 - 5.4 Gradient Boosting Trees 148
 - 5.4.1 Numerical Optimization 148
 - 5.4.2 Steepest Descent 149
 - 5.4.3 Algorithm 150
 - 5.4.4 Particular Cases 153
 - 5.5 Boosting Versus Gradient Boosting 157
 - 5.6 Regularization and Randomness 160
 - 5.6.1 Shrinkage 160
 - 5.6.2 Randomness 161
 - 5.7 Interpretability 162
 - 5.7.1 Relative Importances 162
 - 5.7.2 Partial Dependence Plots 162
 - 5.7.3 Friedman’s H-Statistics 162
 - 5.8 Example 165
 - 5.9 Bibliographic Notes and Further Reading 171
 - References 172

| | | |
|----------|--|-----|
| 6 | Other Measures for Model Comparison | 175 |
| 6.1 | Introduction | 175 |
| 6.2 | Measures of Association | 176 |
| 6.2.1 | Context | 176 |
| 6.2.2 | Probability of Concordance | 177 |
| 6.2.3 | Kendall's Tau | 181 |
| 6.2.4 | Spearman's Rho | 183 |
| 6.2.5 | Numerical Example | 186 |
| 6.3 | Measuring Lift | 192 |
| 6.3.1 | Motivation | 192 |
| 6.3.2 | Predictors Characteristics | 193 |
| 6.3.3 | Convex Order | 194 |
| 6.3.4 | Concentration Curve | 196 |
| 6.3.5 | Assessing the Performances of a Given Predictor | 203 |
| 6.3.6 | Comparison of the Performances of Two Predictors | 209 |
| 6.3.7 | Ordered Lorenz Curve | 215 |
| 6.3.8 | Numerical Illustration | 217 |
| 6.3.9 | Case Study | 223 |
| 6.4 | Bibliographic Notes and Further Reading | 227 |
| | References | 228 |