

Tidy Modeling with R

A Framework for Modeling in the Tidyverse

Max Kuhn and Julia Silge

Beijing • Boston • Farnham • Sebastopol • Tokyo



Table of Contents

Preface.....	xii
--------------	-----

Part I. Introduction

1. Software for Modeling.....	3
Fundamentals for Modeling Software	4
Types of Models	5
Descriptive Models	6
Inferential Models	8
Predictive Models	9
Connections Between Types of Models	10
Some Terminology	11
How Does Modeling Fit into the Data Analysis Process?	12
Chapter Summary	15
2. A Tidyverse Primer.....	17
Tidyverse Principles	17
Design for Humans	18
Reuse Existing Data Structures	19
Design for the Pipe and Functional Programming	20
Examples of Tidyverse Syntax	23
Chapter Summary	25
3. A Review of R Modeling Fundamentals.....	27
An Example	27
What Does the R Formula Do?	33
Why Tidiness Is Important for Modeling	34

Combining Base R Models and the Tidyverse	38
The tidymodels Metapackage	39
Chapter Summary	41

Part II. Modeling Basics

4. The Ames Housing Data	45
Exploring Features of Homes in Ames	47
Chapter Summary	52
5. Spending Our Data	55
Common Methods for Splitting Data	56
What About a Validation Set?	59
Multilevel Data	59
Other Considerations for a Data Budget	60
Chapter Summary	61
6. Fitting Models with <code>parsnip</code>	63
Create a Model	63
Use the Model Results	69
Make Predictions	71
<code>parsnip</code> -Extension Packages	73
Creating Model Specifications	73
Chapter Summary	74
7. A Model Workflow	75
Where Does the Model Begin and End?	75
Workflow Basics	78
Adding Raw Variables to the <code>workflow()</code>	80
How Does a <code>workflow()</code> Use the Formula?	81
Tree-Based Models	82
Special Formulas and Inline Functions	82
Creating Multiple Workflows at Once	84
Evaluating the Test Set	86
Chapter Summary	87
8. Feature Engineering with <code>Recipes</code>	89
A Simple <code>recipe()</code> for the Ames Housing Data	90
Using <code>Recipes</code>	92
How Data Are Used by the <code>recipe()</code>	94
Examples of Steps	94

Encoding Qualitative Data in a Numeric Format	95
Interaction Terms	97
Spline Functions	99
Feature Extraction	101
Row Sampling Steps	102
General Transformations	103
Natural Language Processing	103
Skipping Steps for New Data	103
Tidy a recipe()	104
Column Roles	106
Chapter Summary	107
9. Judging Model Effectiveness.....	109
Performance Metrics and Inference	110
Regression Metrics	112
Binary Classification Metrics	114
Multiclass Classification Metrics	117
Chapter Summary	122
Part III. Tools for Creating Effective Models	
10. Resampling for Evaluating Performance.....	125
The Resubstitution Approach	125
Resampling Methods	128
Cross-Validation	129
Repeated Cross-Validation	132
Leave-One-Out Cross-Validation	133
Monte Carlo Cross-Validation	133
Validation Sets	134
Bootstrapping	135
Rolling Forecasting Origin Resampling	136
Estimating Performance	138
Parallel Processing	143
Saving the Resampled Objects	144
Chapter Summary	147
11. Comparing Models with Resampling.....	149
Creating Multiple Models with Workflow Sets	149
Comparing Resampled Performance Statistics	152
Simple Hypothesis Testing Methods	155
Bayesian Methods	157

A Random Intercept Model	158
The Effect of the Amount of Resampling	163
Chapter Summary	164
12. Model Tuning and the Dangers of Overfitting.	165
Model Parameters	165
Tuning Parameters for Different Types of Models	166
What Do We Optimize?	168
The Consequences of Poor Parameter Estimates	173
Two General Strategies for Optimization	176
Tuning Parameters in tidymodels	177
Chapter Summary	183
13. Grid Search.	185
Regular and Nonregular Grids	185
Regular Grids	186
Nonregular Grids	188
Evaluating the Grid	191
Finalizing the Model	196
Tools for Creating Tuning Specifications	198
Tools for Efficient Grid Search	199
Submodel Optimization	199
Parallel Processing	200
Benchmarking Boosted Trees	204
Access to Global Variables	206
Racing Methods	208
Chapter Summary	210
14. Iterative Search.	213
A Support Vector Machine Model	213
Bayesian Optimization	216
A Gaussian Process Model	216
Acquisition Functions	218
The tune_bayes() Function	222
Simulated Annealing	227
Simulated Annealing Search Process	227
The tune_sim_anneal() Function	230
Chapter Summary	234
15. Screening Many Models.	235
Modeling Concrete Mixture Strength	235
Creating the Workflow Set	238

Tuning and Evaluating the Models	241
Efficiently Screening Models	244
Finalizing a Model	247
Chapter Summary	249

Part IV. Beyond the Basics

16. Dimensionality Reduction.....	253
What Problems Can Dimensionality Reduction Solve?	253
A Picture Is Worth a Thousand...Beans	254
A Starter Recipe	257
Recipes in the Wild	258
Preparing a Recipe	259
Baking the Recipe	260
Feature Extraction Techniques	262
Principal Component Analysis	262
Partial Least Squares	264
Independent Component Analysis	266
Uniform Manifold Approximation and Projection	267
Modeling	268
Chapter Summary	271
17. Encoding Categorical Data.....	273
Is an Encoding Necessary?	274
Encoding Ordinal Predictors	274
Using the Outcome for Encoding Predictors	275
Effect Encodings in tidymodels	276
Effect Encodings with Partial Pooling	278
Feature Hashing	280
More Encoding Options	283
Chapter Summary	284
18. Explaining Models and Predictions.....	285
Software for Model Explanations	286
Local Explanations	287
Global Explanations	291
Building Global Explanations from Local Explanations	293
Back to Beans!	297
Chapter Summary	299

19. When Should You Trust Your Predictions?.....	301
Equivocal Results	301
Determining Model Applicability	307
Chapter Summary	315
20. Ensembles of Models.....	317
Creating the Training Set for Stacking	318
Blend the Predictions	320
Fit the Member Models	324
Test Set Results	325
Chapter Summary	326
21. Inferential Analysis.....	327
Inference for Count Data	328
Comparisons with Two-Sample Tests	329
Log-Linear Models	334
A More Complex Model	337
More Inferential Analysis	341
Chapter Summary	342
Appendix. Recommended Preprocessing.....	343
References.....	345
Index.....	353