Tidy Modeling with R

MAX KUHN AND JULIA SILGE

Version 0.0.1.9010 (2021-10-28)

Hello World

This is the website for *Tidy Modeling with R*. This book is a guide to using a new collection of software in the R programming language for model building, and it has two main goals:

- First and foremost, this book provides an introduction to **how to use** our software to create models. We focus on a dialect of R called *the tidyverse* that is designed to be a better interface for common tasks using R. If you've never heard of or used the tidyverse, Chapter 2 provides an introduction. In this book, we demonstrate how the tidyverse can be used to produce high quality models. The tools used to do this are referred to as the *tidymodels packages*.
- Second, we use the tidymodels packages to **encourage good methodology and statistical practice**. Many models, especially complex predictive or machine learning models, can work very well on the data at hand but may fail when exposed to new data. Often, this issue is due to poor choices made during the development and/or selection of the models. Whenever possible, our software, documentation, and other materials attempt to prevent these and other pitfalls.

This book is not intended to be a comprehensive reference on modeling techniques; we suggest other resources to learn such nuances. For general background on the most common type of model, the linear model, we suggest Fox (2008). For predictive models, Kuhn and Johnson (2013) is a good resource. Also, Kuhn and Johnson (2020) is referenced heavily here, mostly because it is freely

available online. For machine learning methods, Goodfellow, Bengio, and Courville (2016) is an excellent (but formal) source of information. In some cases, we describe models that are used in this text but in a way that is less mathematical, and hopefully more intuitive.

> Investigating and analyzing data are an important part of the model process, and an excellent resource on this topic is Wickham and Grolemund (2016).

We do not assume that readers have extensive experience in model building and statistics. Some statistical knowledge is required, such as random sampling, variance, correlation, basic linear regression, and other topics that are usually found in a basic undergraduate statistics or data analysis course.

> *Tidy Modeling with R* is currently a work in progress. As we create it, this website is updated. Be aware that, until it is finalized, the content and/or structure of the book may change.

This openness also allows users to contribute if they wish. Most often, this comes in the form of correcting typos, grammar, and other aspects of our work that could use improvement. Instructions for making contributions can be found in the `contributing.md` file. Also, be aware that this effort has a code of conduct, which can be found at `code_of_conduct.md` .

The tidymodels packages are fairly young in the software lifecycle. We will do our best to maintain backwards compatibility and, at the completion of this work, will archive and tag the specific versions of software that were used to produce it.

This book was written in RStudio using bookdown. The `tmwr.org` website is hosted via Netlify, and automatically built after every push by GitHub Actions. The complete source is available on GitHub. We generated all plots in this book using ggplot2 and its black and white theme (`theme_bw()`). This version of the book was built with R version 4.1.1 (2021-08-10), pandoc version 2.7.3, and the following packages:

package	version	source
applicable	0.0.1.2	standard (@0.0.1.2)
av	0.6.0	standard (@0.6.0)
baguette	0.1.1	standard (@0.1.1)
beans	0.1.0	standard (@0.1.0)
bestNormalize	1.8.2	standard (@1.8.2)
bookdown	0.24	standard (@0.24)
broom	0.7.9	standard (@0.7.9)
corrplot	0.90	standard (@0.90)
corrr	0.4.3	standard (@0.4.3)
Cubist	0.3.0	standard (@0.3.0)
DALEXtra	2.1.1	standard (@2.1.1)
dials	0.0.10	standard (@0.0.10)
digest	0.6.28	standard (@0.6.28)
dimRed	0.2.3	standard (@0.2.3)
discrim	0.1.3	standard (@0.1.3)
doMC	1.3.7	standard (@1.3.7)
dplyr	1.0.7	standard (@1.0.7)
earth	5.3.1	standard (@5.3.1)
embed	0.1.4.9000	Github (tidymodels/embed@6bc4c8f)
fastICA	1.2-3	standard (@1.2-3)
finetune	0.1.0.9000	Github (tidymodels/finetune@9806094)
forcats	0.5.1	standard (@0.5.1)
ggforce	0.3.3	standard (@0.3.3)

package	version	source
ggplot2	3.3.5	standard (@3.3.5)
glmnet	4.1-2	standard (@4.1-2)
gridExtra	2.3	standard (@2.3)
infer	1.0.0	standard (@1.0.0)
kableExtra	1.3.4	standard (@1.3.4)
kernlab	0.9-29	standard (@0.9-29)
kknn	1.3.1	standard (@1.3.1)
klaR	0.6-15	standard (@0.6-15)
knitr	1.36	standard (@1.36)
learntidymodels	0.0.0.9001	Github (tidymodels/learntidymodels@4b9dcb0)
lime	0.5.2	standard (@0.5.2)
lme4	1.1-27.1	standard (@1.1-27.)
lubridate	1.8.0	standard (@1.8.0)
mda	0.5-2	standard (@0.5-2)
mixOmics	6.17.26	standard (@6.17.26)
modeldata	0.1.1	standard (@0.1.1)
nlme	3.1-153	standard (@3.1-153)
nnet	7.3-16	CRAN (R 4.1.1)
parsnip	0.1.7	standard (@0.1.7)
patchwork	1.1.1	standard (@1.1.1)
poissonreg	0.1.1	standard (@0.1.1)
prettyunits	1.1.1	standard (@1.1.1)
probably	0.0.6	standard (@0.0.6)

package	version	source
pscl	1.5.5	standard (@1.5.5)
purrr	0.3.4	standard (@0.3.4)
ranger	0.13.1	standard (@0.13.1)
recipes	0.1.17	standard (@0.1.17)
rlang	0.4.12	standard (@0.4.12)
rmarkdown	2.11	standard (@2.11)
rpart	4.1-15	CRAN (R 4.1.1)
rsample	0.1.0	standard (@0.1.0)
rstanarm	2.21.1	standard (@2.21.1)
rules	0.1.2	standard (@0.1.2)
sessioninfo	1.1.1	standard (@1.1.1)
stacks	0.2.1	standard (@0.2.1)
stringr	1.4.0	standard (@1.4.0)
svglite	2.0.0	standard (@2.0.0)
themis	0.1.4	standard (@0.1.4)
tibble	3.1.5	standard (@3.1.5)
tidymodels	0.1.4	standard (@0.1.4)
tidyposterior	0.1.0	standard (@0.1.0)
tidyverse	1.3.1	standard (@1.3.1)
tune	0.1.6	standard (@0.1.6)
uwot	0.1.10	standard (@0.1.10)
workflows	0.2.4	standard (@0.2.4)
workflowsets	0.1.0	standard (@0.1.0)

package	version	source
xgboost	1.4.1.1	standard (@1.4.1.1)
yardstick	0.0.8	standard (@0.0.8)

REFERENCES

Fox, J. 2008. *Applied Regression Analysis and Generalized Linear Models*. Second. Thousand Oaks, CA: Sage.

Goodfellow, I, Y Bengio, and A Courville. 2016. *Deep Learning*. MIT Press.

Kuhn, M, and K Johnson. 2013. *Applied Predictive Modeling*. Springer.

Kuhn, M, and K Johnson. 2020. *Feature Engineering and Selection: A Practical Approach for Predictive Models*. CRC Press.

Wickham, H, and G Grolemund. 2016. *R for Data Science: Import, Tidy, Transform, Visualize, and Model Data*. O'Reilly Media, Inc.