

Ensembles: From Beginner to Expert

Russ Conte

2024-05-23

Contents

1	Welcome!	11
1.1	Ensembles: The New AI, from beginner to expert	12
1.2	What you will be able to do by the end of the book	14
1.3	How this book is organized so you learn the material as easily as possible	15
1.4	How you can learn the skills as fast as possible: How the exercises are organized	16
1.5	Going from student to teacher: You are required to post on social media and help others understand the results	16
1.6	Helping you use the power of pre-trained ensembles and individual models	16
1.7	Helping you master the material: One of your own exercises . . .	17
1.8	Keeping it real: Actual business data and problems as the source of all the data sets	17
1.9	Helping you check your work—and verifying that your results beat previously published results	17
1.10	Helping you work as a team with fully reproducible ensembles and individual models	18
1.11	The Final Comprehensive Project will put everything together for you	18
1.12	Exercises to help improve your skills	19
2	Introduction and your first ensembles	21
2.1	How a Chicago blizzard led to the very unlikely story of the best solutions to supervised data	21
2.2	Saturday, October 15, 2022 at 4:58 pm. The exact birth of the Ensembles system	24
2.3	Here is what an ensemble of models looks like at the most basic level, using the Boston Housing data set as an example:	28
2.4	The steps to build your first ensemble from scratch	30
2.5	Building the first actual ensemble	30
2.6	We’re ready to make our first ensemble!!	32

2.7	Principle: What is one improvement that can be made? Use a diverse set of models and ensembles to get the best possible result	35
2.8	Principle: Randomizing the data before the analysis will make the results more general (and is very easy to do!)	36
2.9	Try it yourself: Repeat the previous analysis, but randomize the rows before the analysis. Otherwise keep the process the same. Share your results on social media.	36
2.10	The more we can randomize the data, the more our results will match nature	38
2.11	Principle: "Is this my very best work?"	40
2.12	"Where do I get help with errors or warnings?"	40
2.13	Is there an easy way to save all trained models?	40
3	Numerical data: How to make 23 individual models, and basic skills with functions	47
4	Set initial values to 0	55
5	Create the function:	57
6	Set up random resampling	59
7	Changes the name of the target column to y	61
8	Moves the target column to the last column on the right	63
9	Breaks the data into train, test and validation sets	65
10	Set initial values to 0	67
11	Set initial values to 0	69
12	Create the function:	71
13	Set up random resampling	73
14	Changes the name of the target column to y	75
15	Moves the target column to the last column on the right	77
16	Breaks the data into train, test and validation sets	79
17	Fit boosted random forest model on the training data, make predictions on holdout data	81
18	Set initial values to 0	83
19	Create the function:	85

<i>CONTENTS</i>	5
20 Set up random resampling	87
21 Changes the name of the target column to y	89
22 Moves the target column to the last column on the right	91
23 Breaks the data into train, test and validation sets	93
24 Fit the model on the training data, make predictions on the holdout data	95
25 Set initial values to 0	97
26 Create the function:	99
27 Set up random resampling	101
28 Changes the name of the target column to y	103
29 Moves the target column to the last column on the right	105
30 Breaks the data into train, test and validation sets	107
31 Set up the elastic model	109
31.1 Elastic using the validation data set	109
32 Set initial values to 0	111
33 Create the function:	113
34 Set up random resampling	115
35 Changes the name of the target column to y	117
36 Moves the target column to the last column on the right	119
37 Breaks the data into train, test and validation sets	121
38 Set up to fit the model on the training data	123
39 Names of columns with ≥ 4 unique vals	125
40 Model data	127
41 Set initial values to 0	129
42 Set up random resampling	131
43 Changes the name of the target column to y	133

44 Moves the target column to the last column on the right	135
45 Breaks the data into train, test and validation sets	137
46 Set initial values to 0	139
47 Set up random resampling	141
48 Changes the name of the target column to y	143
49 Moves the target column to the last column on the right	145
50 Breaks the data into train, test and validation sets	147
51 Set initial values to 0	149
52 Create the function:	151
53 Set initial values to 0	153
54 Set up random resampling	155
55 Changes the name of the target column to y	157
56 Moves the target column to the last column on the right	159
57 Breaks the data into train, test and validation sets	161
58 Set initial values to 0	163
59 Set up random resampling	165
60 Changes the name of the target column to y	167
61 Moves the target column to the last column on the right	169
62 Breaks the data into train, test and validation sets	171
63 Set initial values to 0	173
64 Set up random resampling	175
65 Changes the name of the target column to y	177
66 Moves the target column to the last column on the right	179
67 Breaks the data into train, test and validation sets	181
68 Set initial values to 0	183

<i>CONTENTS</i>	7
69 Set up random resampling	185
70 Changes the name of the target column to y	187
71 Moves the target column to the last column on the right	189
72 Breaks the data into train, test and validation sets	191
73 Set initial values to 0	193
74 Set up random resampling	195
75 Changes the name of the target column to y	197
76 Moves the target column to the last column on the right	199
77 Breaks the data into train, test and validation sets	201
78 Set initial values to 0	203
79 Set up random resampling	205
80 Changes the name of the target column to y	207
81 Moves the target column to the last column on the right	209
82 Breaks the data into train, test and validation sets	211
83 Set initial values to 0	213
84 Create the function:	215
85 Set initial values to 0	217
86 Set up random resampling	219
87 Changes the name of the target column to y	221
88 Moves the target column to the last column on the right	223
89 Breaks the data into train, test and validation sets	225
90 Set initial values to 0	227
91 Set up random resampling	229
92 Changes the name of the target column to y	231
93 Moves the target column to the last column on the right	233

94 Breaks the data into train, test and validation sets	235
95 Set initial values to 0	237
96 Set initial values to 0	239
97 Set up random resampling	241
98 Changes the name of the target column to y	243
99 Moves the target column to the last column on the right	245
100 Breaks the data into train, test and validation sets	247
101 Set initial values to 0	249
102 Set up random resampling	251
103 Changes the name of the target column to y	253
104 Moves the target column to the last column on the right	255
105 Breaks the data into train, test and validation sets	257
106 Set initial values to 0	259
107 Set up random resampling	261
108 Changes the name of the target column to y	263
109 Moves the target column to the last column on the right	265
110 Breaks the data into train, test and validation sets	267
111 define predictor and response variables in test set	269
112 define predictor and response variables in validation set	271
113 define final train, test and validation sets	273
114 define watchlist	275
115 fit XGBoost model and display training and validation data at each round	277
116 Building weighted ensembles to model numerical data	279
116.1 Think before you do something. This will help when we start at the end and work backwards toward the beginning.	282

116.2	One of your own: Add one model to the list of seven individual models, see how it impacts results.	283
116.3	Plan ahead as much as you can, that makes the entire model building process much easier.	283
116.4	One of your own: Add a model to the individual models, and a model to the ensemble of models	293
116.5	Post your results on social media in a way that a non-technical person can understand them. For example:	293
116.6	Exercises to help you improve your skills:	293
116.7	Post the results of your new ensemble on social media in a way that helps others understand the results or methods.	293
1174.	Classification data: How to make 14 individual classification models	295

Chapter 1

Welcome!

Welcome to Ensembles! This book will guide you through the entire process of building your own ensemble models from beginning to end. It will also give you full access to the Ensembles package that automates the entire process.

I've done my very best to make the book very interesting, fun, and practical. There are lots of examples using real world data with all the steps included.

Nature is most accurately
modeled using a diverse set of
individual models and ensembles
of models. This book will give
you tools to make a diverse
range of individual models and
ensembles of models for
numeric, logistic, classification
and time series data.

Figure 1.1: How nature is most accurately modeled

You will be able to do wonderful things as you complete the skills in this book. As the book will show, ensembles are much more accurate than any other method to help us understand and model nature. This will be done with a level of accuracy that has not been achieved previously. And you can do all of it.

The phrase “wonderful things” is very intentional. When Howard Carter was doing archaeology, at one point in November, 1922, he was quite sure he found something important. Carter made a small hole to see through. Lord Carnarvon (who was paying for all of this!) asked Howard Carter, “Can you see anything?”. Howard Carter’s famous reply, “Yes, wonderful things!”. When they opened everything up, they found the intact tomb of Tutankhamun. It contained more than 5,000 items, and enriched our knowledge of ancient Africa beyond any other find.

Here is a tiny taste of one of the more than 5,000 the “wonderful things” found by Howard Carter, Lord Carnarvon, and the team of archaeologists.

I will do my very best to share many “wonderful things” through the entire book as you explore the world of ensembles.

The Ensembles package I’ve made does the entire analysis process for you automatically. This will put the power of ensembles in your hands, give you the strongest foundation for your work, with the highest degree of accuracy.

All of the examples in the book will come from real data. For example (and there are many more examples in the book):

- HR Analytics
- Predicting the winning time in the London Marathon
- World’s most accurate score to a very difficult classification problem
- Beat the best score in student Kaggle competitions

We will have many more practical examples from a very wide range of fields for you to enjoy.

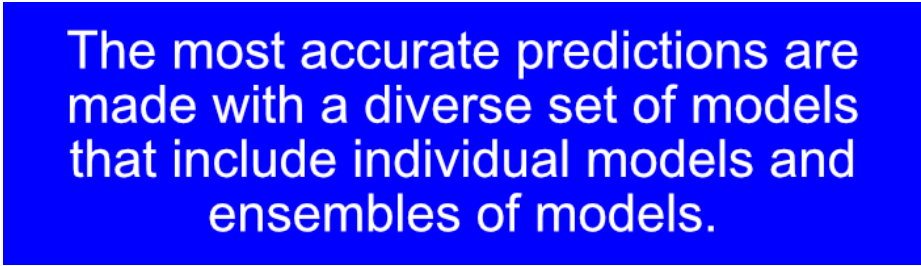
This book will show you how ensembles improve our understanding of nature, and how you can use ensembles in your work. The results using ensembles are much more accurate than has ever been possible before, and that will be demonstrated over and over again in this book. You will be able to use ensembles to understand the world, and build your own models of data, at a level of accuracy that has not been achieved before.

1.1 Ensembles: The New AI, from beginner to expert

As you will see, Ensembles are the new AI. Science has gone from the calculus of Newton and Leibnetz, to differential equations, to the modern world of creating models, and many points in-between. Ensembles are the most powerful way to put models together to achieve the best possible results. This book will guide



Figure 1.2: King Tut Mask



The most accurate predictions are made with a diverse set of models that include individual models and ensembles of models.

Figure 1.3: The most accurate predictions

you through the process, and show you how you can build ensembles that will pass all testing.

This is the new AI. Welcome to the path, it's extremely fun, and I look forward to sharing it with you!

1.2 What you will be able to do by the end of the book

- Make your own customized ensembles of models of numerical, classification, logistic and time series data.
- Use the Ensembles package which does the entire process automatically (but little customization is possible).
- Make your ensemble solutions into packages that can be shared with other users.
- Make your ensemble solutions into totally self-contained solutions that can be shared with anyone.
- Learn how ensembles of models can help to make the wisest possible decision based on the data.
- Learn how to present the results at different levels, from a regular user to a CEO and board of directors.
- How to present results that are social media friendly.
- Find your own data and create the ensemble solution from beginning to end (called One Of Your Own in the each of the chapter exercises)
- Solve real world examples in this book where ensembles achieve such results as:
- Beat the top score in a student data science competition by over 90% (numerical ensembles).

- Correctly predict the winning time for the 2024 Men's London Marathon (time series ensembles).
- Produce a 100% accurate solution to the dry beans classification problem (first in the world with this data set, done using classification ensembles).
- Make recommendations how Lebron James can improve his performance on the basketball court (logistic ensembles).
- Complete a comprehensive Final Project that will put all of your new skills with ensembles together. This result can be shared with employers, advisors, on social media, job interviews, or anywhere else you would like to share your work.

1.3 How this book is organized so you learn the material as easily as possible

The book begins with the foundations of making ensembles of models. We will look at:

- Individual numerical models
- Ensembles of numerical models
- Individual classification models
- Ensembles of classification models
- Individual logistic models
- Ensembles of logistic models
- Individual forecasting models
- Ensembles of forecasting models
- Advanced data visualizations
- Multiple ways to communicate your results. This will range from other people in the field, to customers, to the C-Suite (CEO, CTO, board of directors, etc.)
- We will look at how to treat data science as a business. In particular we will pay close attention to showing return on investment (ROI) in data science, using ensembles of models.
- The book will conclude showing eight examples of a final comprehensive project. There will be two examples each of numerical data, classification data, logistic data and forecasting data. One example of each pair is a regularly formatted paper, the other is professionally formatted. The source files for each of the eight files are available in a github repository.

1.4 How you can learn the skills as fast as possible: How the exercises are organized

As a young child, I learned that I have much better retention with a system I have always called delayed repetition. This means that I learn best and fastest when I see a worked out example, do several practice examples, and then repeat that after a delay in time. The delay can range from an hour to a few days.

For example, the exercises in the Individual Classification Models chapter will ask you to build models using techniques from the classification models and the prior chapters. The exercises for logistic ensembles will ask you to build models from the content in the logistic models chapter, and each of the previous chapters. It has been my experience that repeating this over and over is the fastest way for me to learn new content, and retain it for the longest period of time.

By the time you get to the Final Comprehensive Project, your skills will be sharp for each of the modeling techniques.

1.5 Going from student to teacher: You are required to post on social media and help others understand the results

One of the most important parts of your role in data science is communicating your findings. I will present many examples of summaries and reports for you to adapt and use on your projects. You are also required to post your results on social media. You may use any appropriate choice of social media, but it needs to be publicly available. This has a number of very important benefits to you:

- You will build a body of work that shows your skill level
- The results will demonstrate your ability to communicate in a way that works with a wide variety of people
- You will work to demonstrate very good skills with video and/or audio production
- Use the hashtag #AIEnsembles when you post on social media

1.6 Helping you use the power of pre-trained ensembles and individual models

Another important part of the skills you will learn here includes building pre-trained ensembles and models. The book will walk you through the process of

building the pre-trained models and ensembles for each of the four types of data (numerical, classification, logical, and time series).

1.7 Helping you master the material: One of your own exercises

One of the differences with the exercises in Ensembles is the inclusion of One of Your Own exercises. Each set of exercises will include one which asks you to find your own data (with many hints given to help you find data), define the problem, make the ensemble, and report the results.

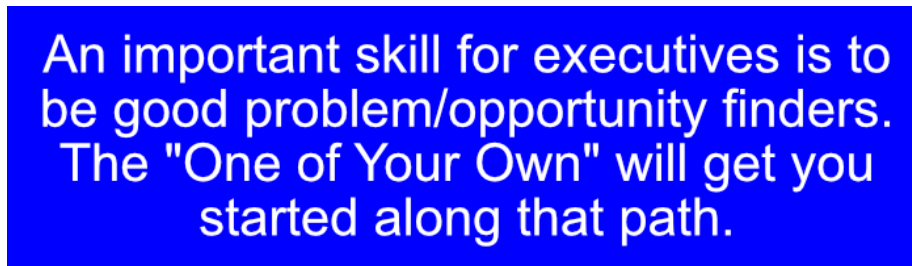


Figure 1.4: Good problem finder

1.8 Keeping it real: Actual business data and problems as the source of all the data sets

All of the data sets in this book use real data. No exceptions, no synthetic data. The sources of the data are all cited, and the real world implications can be found by a simple search. All of the data is absolutely real.

1.9 Helping you check your work—and verifying that your results beat previously published results

Many of the data sets have been solved by previous investigators (such as in competitions), so the results here can be easily compared with published results.

For example, we will look at the Boston Housing data set when we look at numerical data sets. This data set has been used many times in Kaggle competitions, published papers, and Github repositories, among many other sources.

The Ensembles package will automatically solve this data set, and return an RMSE less than 0.20 (there will be slight variation depending on how the parameters are set, as will be explained those chapters). In comparison, the Boston Housing data set was used in this Kaggle student competition: <https://www.kaggle.com/competitions/uou-g03784-2022-spring/leaderboard?tab=public>, and the best score was 2.09684. The Ensembles package will beat the best result in that Kaggle student competition by more than 90%. The Ensembles package only requires one line of code.

1.10 Helping you work as a team with fully reproducible ensembles and individual models

A large part of the skills you will learn include how to make results that are reproducible. This will include:

- Multiple random resamplings of the data
- Learning how to test on totally unseen data for both individual and ensemble models
- How to repeat results (for example, 25 times), and report the accuracy of each resampling

For example, you will make ensembles of models, and then use those trained models to make predictions on totally unseen data.

1.11 The Final Comprehensive Project will put everything together for you

As I was studying data science, one of my professors said that the papers I turned in were “good enough to show to the CEO or Board of Directors” of the Fortune 1000 company he worked for. The chapter on the Final Comprehensive Project will share the highest level of skills in the following:

- Truly understanding the business problem
- Being able to convey the very high value that data science brings to the table
- Being able to back up 100% of your claims with rock solid evidence, facts, and clear reasoning
- How to make a truly professional quality presentation worthy of the C-Suite

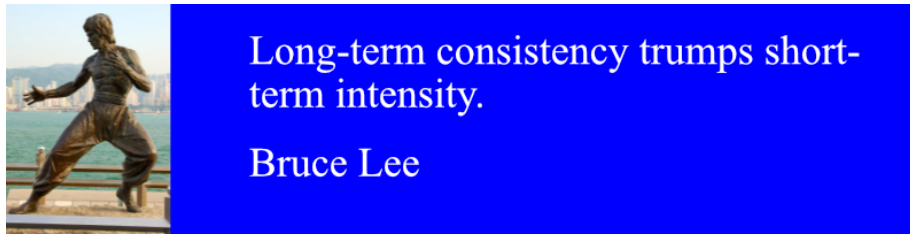


Figure 1.5: Long term consistency wins the race

I've had the incredible pleasure of learning many different skills. A few include being able to play on more than 20 musical instruments, communicate in three languages at a professional level, manage a multi-million dollar division of a Fortune 1000 company, run two non-profit volunteer groups, snowboard a three mile run in Colorado, work as a professional counselor, and much more. The book your reading is only my most recent project. None of these skills were acquired overnight. A huge part of the success is being able to make slow and (usually) steady progress. The next chapter will reveal the big secret to getting results, but for now you are best off if you plan some regular time to work on the contents of the book.

Always remember to test everything, that will save you a ton of problems down the road.

1.12 Exercises to help improve your skills

Exercise 1: Schedule regular time to work on this book

You will gain much more progress if you work at a steady pace. Take everything in small pieces. It's OK to go slow, as long as you keep going. Schedule regular time to work on this book, and you will get the largest possible reward for your efforts.

Exercise 2: Read each chapter at least twice **before** you begin working on the material.

Reading each chapter twice before you begin working on it will actually speed up your progress and results. It will actually take less time for you to complete the chapter. You might not believe it right now, but it's totally true.

Exercise 2a: Read a chapter ahead if you are able to do so.

Exercise 3: Read this chapter again

Chapter 2

Introduction and your first ensembles

2.1 How a Chicago blizzard led to the very unlikely story of the best solutions to supervised data

My journey to the most advanced AI in the world started with an actual blizzard in Chicago. It might seem like Chicago would never get a blizzard, but we did in 2011, and it was incredibly intense, as this video shows:

<https://www.youtube.com/watch?v=cPiFn52ztd8>

What does the Chicago 2011 Snomageddon have to do with the creation of the most advanced AI? Everything. Here's the story.

At the time of the 2011 Blizzard I worked a Recruiter for Kelly Services, where I had worked since 1996. I agreed to work out of the Kelly Services office in Frankfort, Illinois at this time, though I worked out of nearly every Kelly Services office at one time or another. The trip to Frankfort involved a daily commute to the office, but I was able to make the best use of the time on the road.

My manager at the time let me know several days in advance that there was a very large amount of snow forecast, and that I might want to be prepared. The most recent forecasts for large amounts of snow in the Chicago area all amounted to nothing. They were perfectly normal days in the Chicago area, so I predicted this storm would also be nothing, based on the most recent results. This was a great example of a prior prediction not transferring well to a current situation.

That morning I went to work as normal, and did not even look at the weather forecast. Around 2:45 pm my manager came out of her office and said “Russ, you need to come here and look at the weather radar!”. I walked into her office, and saw a map of a winter storm that was incredibly huge. She had the image zoomed out, so it was possible to see several states. From what I could tell, the massive snow storm was barreling down on Chicago, and was about 15 minutes away from our location.

I told the candidate I was interviewing that I was leaving immediately, and that he is not allowed to stay. He has to get home as fast as possible for his own safety.

The storm started dropping snow on my trip north back home. The commute took around 50% longer than normal due to the rapidly falling snow.

As I later learned, the storm was forecast to start in the Chicago area around 3:00 pm, finish up between 11:00 am - 1:00 pm two days later, and leave 17 - 19 inches of snow.

How bad was it? Even City of Chicago snow plows were stopped by the snow:



Figure 2.1: Chicago snow plow stuck on Lake Shore Drive in the 2011 snow storm

To see what the forecasts looked like, check out this news report from the day:

2.1. HOW A CHICAGO BLIZZARD LED TO THE VERY UNLIKELY STORY OF THE BEST SOLUTIONS TO SU

<https://www.nbcchicago.com/news/local/blizzard-unleashes-winter-fury/2096753/>

It turns out all three predictions of the blizzard were accurate to a level that almost seemed uncanny to me: Start time, accumulation, and end time were all spot on. This is the first time I recall ever seeing a prediction at this level of accuracy. I had no idea this type of predictive accuracy was even possible. This level of accuracy in predicting results totally blew me away. I had never seen anything with this level of accuracy, and now I wanted to know how it was done.

I searched and searched for how the accuracy was so high for this forecast.

The power of the method—whatever it was—was obvious to me. I realized that if it could work for the weather, the solution method could work in an incredibly broad range of situations. A few of many other areas include business forecasts, production work, modeling prices, and much, much more. But at this point I had no idea how the accurate prediction was done.

Some months later a person wrote to Tom Skilling, chief meteorologist for WGN TV in Chicago. Tom posted an answer that opened up the solution for me. Here is the relevant part of Tom Skilling's answer to a 2011 storm how the forecast was so accurate:

The Weather Service has developed an interesting "SNOWFALL ENSEMBLE FORECAST PROBABILITY SYSTEM" which draws upon a wide range of snow accumulation forecasts from a whole set of different computer models. By "blending" these model projections, probability of snowfalls falling within certain ranges becomes possible. Also, this "blending" of multiple forecasts "smooths" the sometimes huge model disparities in the amounts being predicted. The resulting probabilities therefore represent a "best case" forecast.

So that was the first step. Ensembles were the way they achieved such extraordinary prediction accuracy.

My next goal was to figure out how ensembles were made. As I looked up information, it became obvious that ensembles had been used for a while, such as the winning entry in the Netflix Prize Competition:

The Netflix Prize Competition was sponsored by Netflix to create a method to accurately predict user ratings on films. The minimum winning score needed to beat the Netflix method (named Cinematch) by at least 10%. Several years of work went into solving this problem, and the results even included several published papers. The winning solution was an ensemble of methods that beat the Cinematch results by 10.09%.

So it was now clear to me that ensembles were the path forward. However, I had no idea how to make ensembles.



Figure 2.2: Netflix Prize Competition

I went to graduate school to study data science and predictive analytics. My degree was completed in 2017, from Northwestern University. However, I still was not sure how ensembles of models were built, nor could I find any clear methods to build them (except for pre-made methods, such as random forests). While it is true there were packages that could do some of the work, nothing I found did what I was looking for: How to build ensembles of models in general. Despite playing with the idea and looking online, I was not able to build the ensembles I wanted to build.

2.2 Saturday, October 15, 2022 at 4:58 pm. The exact birth of the Ensembles system

Everything changed on Saturday, October 15, 2022 at 4:58 pm. I was playing with various methods to make an ensemble, and got an ensemble that worked for the very first time. While the results were extremely modest by any standards, it was clear to me that the foundation was there to build a general solution that can work in an extremely wide range of areas. Here is my journal entry:

You might be asking yourself how I know the day and time. That is a very reasonable question. I've been keeping a journal since I was 19 years old, and have thousands of entries. As soon as I realized how to correctly build ensembles,

It just hit me how to create ensemble models - take all of the y-predicted values from each of the models, make those the X values in a data frame, and have the true y values (in the ensembles data frame) as the true y values, then run that data frame through each of the modeling solutions to find the best performer. W0w!

Figure 2.3: Birth of the ensembles method (typo of W0w in the original)

I made this entry, which contains the key elements to make an ensemble, and we will do these steps in just a moment. Notice that the subject line in the journal matches the text above.

One of the ways to improve your skills is to keep a journal, and we'll be looking at that in more depth in this chapter and future chapters. The journal I use is MacJournal, though there are a large number of other options available on the market.

2.2. SATURDAY, OCTOBER 15, 2022 AT 4:58 PM. THE EXACT BIRTH OF THE ENSEMBLES SYSTEM27

Inspector

Document

Journal

Entry

Topic:

It *just* hit me how to create e

Date:

10/15/2022, 4:58 PM

Tags:

Annotation:

Status:

Unknown

Priority:

None

Due:

☐

5/19/2024, 10:08 AM

Rating:

☐

☐

☐

☐

☐

☐

Created:

October 15, 2022, 4:58 PM

Modified:

June 14, 2023, 9:30 PM

Time Edited:

10 minutes 20 seconds

Size:

408 KB

☐ Editable

☐ Flagged

Icon:

Mood:

Word Goal:

Inherit

Label:

Carnation

Background:

Inherit

Blog:

Inherit

Link:

Location:

0.000000

0.000000

Time Zone:

America/Chicago

Related Files:

+

-

Birth of ensembles, Saturday, October 15, 2022 at 4:58 pm



Figure 2.4: Keep a journal

2.3 Here is what an ensemble of models looks like at the most basic level, using the Boston Housing data set as an example:

2.3.1 Head of Boston Housing data set

```
> head(MASS::Boston, n = 10) # look at the first ten (out of 505) rows of the Boston Housing data set
      crim    zn  indus chas   nox    rm    age    dis rad tax ptratio  black lstat medv
1  0.00632 18.0   2.31    0 0.538 6.575 65.2 4.0900 1 296   15.3 396.90  4.98 24.0
2  0.02731  0.0   7.07    0 0.469 6.421 78.9 4.9671 2 242   17.8 396.90  9.14 21.6
3  0.02729  0.0   7.07    0 0.469 7.185 61.1 4.9671 2 242   17.8 392.83  4.03 34.7
4  0.03237  0.0   2.18    0 0.458 6.998 45.8 6.0622 3 222   18.7 394.63  2.94 33.4
5  0.06905  0.0   2.18    0 0.458 7.147 54.2 6.0622 3 222   18.7 396.90  5.33 36.2
6  0.02985  0.0   2.18    0 0.458 6.430 58.7 6.0622 3 222   18.7 394.12  5.21 28.7
7  0.08829 12.5   7.87    0 0.524 6.012 66.6 5.5605 5 311   15.2 395.60 12.43 22.9
8  0.14455 12.5   7.87    0 0.524 6.172 96.1 5.9505 5 311   15.2 396.90 19.15 27.1
9  0.21124 12.5   7.87    0 0.524 5.631 100.0 6.0821 5 311   15.2 386.63 29.93 16.5
10 0.17004 12.5   7.87    0 0.524 6.004 85.9 6.5921 5 311   15.2 386.71 17.10 18.9
```

Figure 2.5: Head of Boston Housing data set

We will start our first ensemble with a data set that only has numerical values. Our first example will use the Boston Housing data set, from the MASS package. While the Boston Housing data set is controversial (and we will discuss some of the controversies in our example making professional quality reports for the C-Suite), for now it works as a very well known data set to begin our journey into ensembles.

Overview of the most basic steps to make an ensemble:

We will be using the Boston Housing data set, so let's have a look at some Boston images:

2.3. HERE IS WHAT AN ENSEMBLE OF MODELS LOOKS LIKE AT THE MOST BASIC LEVEL, USING THE B



Figure 2.6: Boston

2.4 The steps to build your first ensemble from scratch

- Load the packages we will need (MASS, tree)
- Load the Boston Housing data set, and split it into train (60%) and test (40%) sections.
- Create a linear model by fitting the linear model on the training data, and make predictions on the Boston Housing test data. Measure the accuracy of the predictions against the actual values.
- Create a model using trees by fitting the tree model on the training data, and making predictions on the Boston Housing test data. Measure the accuracy of the predictions against the actual values.
- Make a new data frame. This will be our ensemble of model predictions. One column will be the linear predictions, and one will be the tree predictions.
- Make a new column for the true values—these are the true values in the Boston Housing test data set
- Once we have the new ensemble data set, it's simply another data set. No different in many ways from any other data set (except how it was made).
- Break the ensemble data set into train (60%) and test (40%) sections.
- Fit a linear model to the ensemble training data. Make predictions using the testing data, and measure the accuracy of the predictions against the test data.
- Summarize the results.

I suggest reading the over of the most basic steps to make an ensemble a couple of times, to make sure you are very familiar with the steps.

2.5 Building the first actual ensemble

Load the packages we will need (MASS, tree):

```
library(MASS) # for the Boston Housing data set
library(tree) # To make models using trees
library(Metrics) # To calculate error rate (root mean squared error)
library(tidyverse)
#> -- Attaching core tidyverse packages ---- tidyverse 2.0.0 --
#> v dplyr      1.1.4      v readr      2.1.5
#> v forcats    1.0.0      v stringr    1.5.1
#> v ggplot2    3.5.1      v tibble     3.2.1
```

```
#> v lubridate 1.9.3      v tidyr      1.3.1
#> v purrr      1.0.2
#> -- Conflicts ----- tidyverse_conflicts() --
#> x dplyr::filter() masks stats::filter()
#> x dplyr::lag()      masks stats::lag()
#> x dplyr::select() masks MASS::select()
#> i Use the conflicted package (<http://conflicted.r-lib.org/>) to force all conflicts to become
```

Load the Boston Housing data set, and split it into train (60%) and test (40%) sections.

```
df <- MASS::Boston
train <- df[1:400, ]
test <- df[401:505, ]

# Let's have a quick look at the train and test sets
head(train)
#>      crim zn indus chas   nox    rm  age    dis rad tax
#> 1 0.00632 18  2.31    0 0.538 6.575 65.2 4.0900  1 296
#> 2 0.02731  0  7.07    0 0.469 6.421 78.9 4.9671  2 242
#> 3 0.02729  0  7.07    0 0.469 7.185 61.1 4.9671  2 242
#> 4 0.03237  0  2.18    0 0.458 6.998 45.8 6.0622  3 222
#> 5 0.06905  0  2.18    0 0.458 7.147 54.2 6.0622  3 222
#> 6 0.02985  0  2.18    0 0.458 6.430 58.7 6.0622  3 222
#>  ptratio  black lstat medv
#> 1    15.3 396.90  4.98 24.0
#> 2    17.8 396.90  9.14 21.6
#> 3    17.8 392.83  4.03 34.7
#> 4    18.7 394.63  2.94 33.4
#> 5    18.7 396.90  5.33 36.2
#> 6    18.7 394.12  5.21 28.7

head(test)
#>      crim zn indus chas   nox    rm  age    dis rad tax
#> 401 25.04610  0 18.1    0 0.693 5.987 100.0 1.5888 24 666
#> 402 14.23620  0 18.1    0 0.693 6.343 100.0 1.5741 24 666
#> 403  9.59571  0 18.1    0 0.693 6.404 100.0 1.6390 24 666
#> 404 24.80170  0 18.1    0 0.693 5.349  96.0 1.7028 24 666
#> 405 41.52920  0 18.1    0 0.693 5.531  85.4 1.6074 24 666
#> 406 67.92080  0 18.1    0 0.693 5.683 100.0 1.4254 24 666
#>  ptratio  black lstat medv
#> 401    20.2 396.90 26.77  5.6
#> 402    20.2 396.90 20.32  7.2
#> 403    20.2 376.11 20.31 12.1
#> 404    20.2 396.90 19.77  8.3
#> 405    20.2 329.46 27.38  8.5
```

```
#> 406      20.2 384.97 22.98  5.0
```

Create a linear model by fitting the linear model on the training data, and make predictions on the Boston Housing test data. Measure the accuracy of the predictions against the actual values.

```
Boston_lm <- lm(medv ~ ., data = train) # Fit the model to the training data
Boston_lm_predictions <- predict(object = Boston_lm, newdata = test)
```

```
# Let's have a quick look at the model predictions
head(Boston_lm_predictions)
```

```
#>      401      402      403      404      405      406
#> 12.618507 19.785728 20.919370 13.014507  6.946392  5.123039
```

Calculate the error for the model

```
Boston_linear_RMSE <- Metrics::rmse(actual = test$medv, predicted = Boston_lm_predictions)
Boston_linear_RMSE
#> [1] 6.108005
```

The error rate for the linear model is 6.108005. Let's do the same using the tree method.

Create a model using trees by fitting the tree model on the training data, and making predictions on the Boston Housing test data. Measure the accuracy of the predictions against the actual values.

```
Boston_tree <- tree(medv ~ ., data = train) # Fit the model to the training data
Boston_tree_predictions <- predict(object = Boston_tree, newdata = test)
```

```
# Let's have a quick look at the predictions:
```

```
head(Boston_tree_predictions)
#>      401      402      403      404      405      406
#> 13.30769 13.30769 13.30769 13.30769 13.30769 13.30769
```

Calculate the error rate for the tree model:

```
Boston_tree_RMSE <- Metrics::rmse(actual = test$medv, predicted = Boston_tree_predictions)
Boston_tree_RMSE
#> [1] 5.478017
```

The error rate for the tree model is lower (which is better). The error rate for the tree model is 5.478017.

2.6 We're ready to make our first ensemble!!

Make a new data frame. This will be our ensemble of model predictions, and one column for the true values. One column will be the linear predictions, and

one will be the tree predictions. We'll make a third column, the true values.

Make a new column for the true values—these are the true values in the Boston Housing test data set

```
ensemble <- data.frame(
  'linear' = Boston_lm_predictions,
  'tree' = Boston_tree_predictions,
  'y' = test$medv
)

# Let's have a look at the ensemble:
head(ensemble)
#>      linear      tree      y
#> 401 12.618507 13.30769  5.6
#> 402 19.785728 13.30769  7.2
#> 403 20.919370 13.30769 12.1
#> 404 13.014507 13.30769  8.3
#> 405  6.946392 13.30769  8.5
#> 406  5.123039 13.30769  5.0

dim(ensemble)
#> [1] 105  3
```

Once we have the new ensemble data set, it's simply another data set. No different in many ways from any other data set (except how it was made).

Break the ensemble data set into train (60%) and test (40%) sections. There is nothing special about the 60/40 split here, you may use any numbers you wish.

```
ensemble_train <- ensemble[1:60, ]
ensemble_test  <- ensemble[61:105, ]

head(ensemble_train)
#>      linear      tree      y
#> 401 12.618507 13.30769  5.6
#> 402 19.785728 13.30769  7.2
#> 403 20.919370 13.30769 12.1
#> 404 13.014507 13.30769  8.3
#> 405  6.946392 13.30769  8.5
#> 406  5.123039 13.30769  5.0

head(ensemble_test)
#>      linear      tree      y
#> 461 23.88984 13.30769 16.4
#> 462 23.29129 13.30769 17.7
#> 463 22.54055 21.84327 19.5
#> 464 25.50940 21.84327 20.2
```

```
#> 465 22.71231 21.84327 21.4
#> 466 20.83810 21.84327 19.9
```

Fit a linear model to the ensemble training data. Make predictions using the testing data, and measure the accuracy of the predictions against the test data. Notice how similar this is to our linear and tree models.

```
# Fit the model to the training data
ensemble_lm <- lm(y ~ ., data = ensemble_train)

# Make predictions using the model on the test data
ensemble_lm_predictions <- predict(object = ensemble_lm, newdata = ensemble_test)

# Calculate error rate for the ensemble predictions
ensemble_lm_rmse <- Metrics::rmse(actual = ensemble_test$y, predicted = ensemble_lm_predictions)

# Report the error rate for the ensemble
ensemble_lm_rmse
#> [1] 4.826962
```

Summarize the results.

```
results <- data.frame(
  'Model' = c('Linear', 'Tree', 'Ensemble'),
  'Error' = c(Boston_linear_RMSE, Boston_tree_RMSE, ensemble_lm_rmse)
)

results
#>      Model      Error
#> 1  Linear 6.108005
#> 2   Tree 5.478017
#> 3 Ensemble 4.826962
```

Clearly the ensemble had the lowest error rate of the three models. The ensemble is easily the best of the three models because it has the lowest error rate of all the models.

2.6.1 Try it yourself: Make an ensemble where the ensemble is made using trees instead of linear models.

```
# Fit the model to the training data
ensemble_tree <- tree(y ~ ., data = ensemble_train)

# Make predictions using the model on the test data
ensemble_tree_predict <- predict(object = ensemble_tree, newdata = ensemble_test)
```

2.7. PRINCIPLE: WHAT IS ONE IMPROVEMENT THAT CAN BE MADE? USE A DIVERSE SET OF MODELS

```
# Let's look at the predictions
head(ensemble_tree_predict)
#>      461      462      463      464      465      466
#> 14.80000 14.80000 18.94286 18.94286 18.94286 18.94286

# Calculate the error rate
ensemble_tree_rmse <- Metrics::rmse(actual = ensemble_test$y, predicted = ensemble_tree_predict)

ensemble_tree_rmse
#> [1] 5.322011
```

How does this compare to our three other results? Let's update the results table

```
results <- data.frame(
  'Model' = c('Linear', 'Tree', 'Ensemble_Linear', 'Ensemble_Tree'),
  'Error' = c(Boston_linear_RMSE, Boston_tree_RMSE, ensemble_lm_rmse, ensemble_tree_rmse)
)

results <- results %>% arrange(Error)

results
#>      Model      Error
#> 1 Ensemble_Linear 4.826962
#> 2 Ensemble_Tree 5.322011
#> 3 Tree 5.478017
#> 4 Linear 6.108005
```

2.6.2 Both of the ensemble models beat both of the individual models in this example

2.7 Principle: What is one improvement that can be made? Use a diverse set of models and ensembles to get the best possible result

As we shall see when we go through and learn how to build ensembles, the numerical method we will use will build 27 individual models and 13 ensembles for a total of 40 results. When the goal is to get the best possible results, a diverse set of models and ensembles, such as the 40 results for numerical data, will produce much better results than a limited number of models and ensembles.

We will do the same principal when we are looking at classification data, logistic, data, and time series forecasting data. We will use a large number of individual models and ensembles with the goal of achieving the best possible result.

2.8 Principle: Randomizing the data before the analysis will make the results more general (and is very easy to do!)

```
df <- df[sample(nrow(df)),] # Randomize the rows before the analysis
```

2.9 Try it yourself: Repeat the previous analysis, but randomize the rows before the analysis. Otherwise keep the process the same. Share your results on social media.

We'll follow the exact same steps, except for randomizing the rows first.

- Randomize the rows
- Break the data into train and test sets
- Fit the model to the training set
- Make predictions and calculate error from the model on the test set

```
df <- df[sample(nrow(df)),] # Randomize the rows before the analysis
```

```
train <- df[1:400, ]
test <- df[401:505, ]
```

```
# Fit the model to the training data
Boston_lm <- lm(medv ~ ., data = train)
```

```
# Make predictions using the model on the test data
Boston_lm_predictions <- predict(object = Boston_lm, newdata = test)
```

```
# Let's have a quick look at the linear model predictions:
```

```
head(Boston_lm_predictions)
#>      168      125      491      316      166      319
#> 23.025233 21.502512 1.981854 20.868281 25.668204 24.502376
```

```
Boston_linear_rmse <- Metrics::rmse(actual = test$medv, predicted = Boston_lm_predictions)
```

```
Boston_tree <- tree(medv ~ ., data = train)
```

```
Boston_tree_predictions <- predict(object = Boston_tree, newdata = test)
```

```
Boston_tree_rmse <- Metrics::rmse(actual = test$medv, predicted = Boston_tree_predictions)
```

2.9. TRY IT YOURSELF: REPEAT THE PREVIOUS ANALYSIS, BUT RANDOMIZE THE ROWS BEFORE THE

```
# Let's have a quick look at the tree model predictions:

head(Boston_tree_predictions)
#>      168      125      491      316      166      319
#> 20.54286 17.47719 17.47719 20.54286 20.54286 20.54286

ensemble <- data.frame( 'linear' = Boston_lm_predictions, 'tree' = Boston_tree_predictions, 'y_ensemble' = y_ensemble )

ensemble <- ensemble[sample(nrow(ensemble)), ] # Randomizes the rows of the ensemble

ensemble_train <- ensemble[1:60, ]
ensemble_test <- ensemble[61:105, ]

ensemble_lm <- lm(y_ensemble ~ ., data = ensemble_train)

# Predictions for the ensemble linear model

ensemble_prediction <- predict(ensemble_lm, newdata = ensemble_test)

# Root mean squared error for the ensemble linear model

ensemble_lm_rmse <- Metrics::rmse(actual = ensemble_test$y_ensemble, predicted = ensemble_prediction)

# Same for tree models

ensemble_tree <- tree(y_ensemble ~ ., data = ensemble_train)
ensemble_tree_predictions <- predict(object = ensemble_tree, newdata = ensemble_test)
ensemble_tree_rmse <- Metrics::rmse(actual = ensemble_test$y_ensemble, predicted = ensemble_tree_predictions)

results <- list( 'Linear' = Boston_linear_rmse, 'Trees' = Boston_tree_rmse, 'Ensembles_Linear' = ensemble_lm_rmse, 'Ensembles_Tree' = ensemble_tree_rmse )

results
#> $Linear
#> [1] 5.130468
#>
#> $Trees
#> [1] 4.961259
#>
#> $Ensembles_Linear
#> [1] 2.903167
#>
#> $Ensemble_Tree
#> [1] 3.665621
```

The fact that our results are a bit different from our first ensemble is useful.

This gives us another solid principle to use in our analysis methods:

2.10 The more we can randomize the data, the more our results will match nature

Just watch: Repeat the results 100 times, return the mean of the results (hint: It's two small changes)

```
for (i in 1:100) {

  # First the linear model with randomized data

  df <- df[sample(nrow(df)),] # Randomize the rows before the analysis

  train <- df[1:400, ]
  test <- df[401:505, ]

  Boston_lm <- lm(medv ~ ., data = train)
  Boston_lm_predictions <- predict(object = Boston_lm, newdata = test)

  # Let's have a quick look at the linear model predictions:

  head(Boston_lm_predictions)

  # Let's calculate the root mean squared error rate of the predictions:

  Boston_linear_rmse[i] <- Metrics::rmse(actual = test$medv, predicted = Boston_lm_predictions)

  Boston_linear_rmse_mean <- mean(Boston_linear_rmse)

  # Let's use tree models

  Boston_tree <- tree(medv ~ ., data = train)

  Boston_tree_predictions <- predict(object = Boston_tree, newdata = test)

  # Let's have a quick look at the tree model predictions:

  head(Boston_tree_predictions)

  # Let's calculate the root mean squared error rate of the predictions:

  Boston_tree_rmse[i] <- Metrics::rmse(actual = test$medv, predicted = Boston_tree_predictions)
  Boston_tree_rmse_mean <- mean(Boston_tree_rmse)
```

2.10. THE MORE WE CAN RANDOMIZE THE DATA, THE MORE OUR RESULTS WILL MATCH NATURE39

```
ensemble <- data.frame('linear' = Boston_lm_predictions, 'tree' = Boston_tree_predictions, 'y_ens

ensemble <- ensemble[sample(nrow(ensemble)), ] # Randomizes the rows of the ensemble

ensemble_train <- ensemble[1:60, ]
ensemble_test <- ensemble[61:105, ]

# Ensemble linear modeling

ensemble_lm <- lm(y_ensemble ~ ., data = ensemble_train)

# Predictions for the ensemble linear model

ensemble_prediction <- predict(ensemble_lm, newdata = ensemble_test)

# Root mean squared error for the ensemble linear model

ensemble_lm_rmse[i] <- Metrics::rmse(actual = ensemble_test$y_ensemble, predicted = ensemble_pred

ensemble_lm_rmse_mean <- mean(ensemble_lm_rmse)

ensemble_tree <- tree(y_ensemble ~ ., data = ensemble_train)

ensemble_tree_predictions <- predict(object = ensemble_tree, newdata = ensemble_test)

ensemble_tree_rmse[i] <- Metrics::rmse(actual = ensemble_test$y_ensemble, predicted =
ensemble_tree_predictions)

ensemble_tree_rmse_mean <- mean(ensemble_tree_rmse)

results <- data.frame(
  'Linear' = Boston_linear_rmse_mean,
  'Trees' = Boston_tree_rmse_mean,
  'Ensembles_Linear' = ensemble_lm_rmse_mean,
  'Ensemble_Tree' = ensemble_tree_rmse_mean )
}

results
#>      Linear      Trees Ensembles_Linear Ensemble_Tree
#> 1 4.865269 4.679484      4.161697      5.138091

warnings() # No warnings!
```

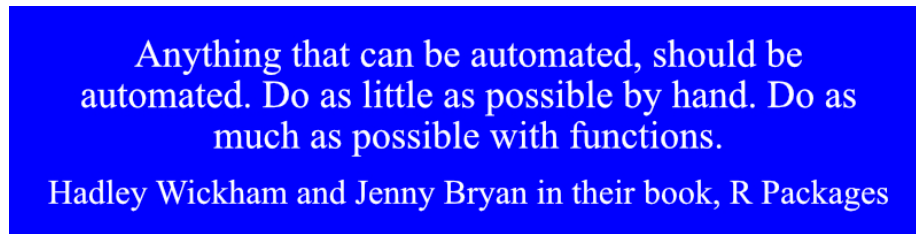


Figure 2.7: Automate as much as possible

2.11 Principle: “Is this my very best work?”

This is your best work to build ensembles at this stage of your skills. We are going to make a number of improvements to the solutions we see here, so our final result will be much stronger than what we have here so far. Always strive to do your very best work, without any excuses.

2.12 “Where do I get help with errors or warnings?”

It is extremely useful to check if your code returns any errors or warnings, and fix those as fast as possible. There are numerous sites to help address errors in your code:

<https://stackoverflow.com>

<https://forum.posit.co>

<https://www.r-project.org/help.html>

2.13 Is there an easy way to save all trained models?

Absolutely! We will simply add the code at the end of this section that saves the four trained models (linear, tree, ensemble_linear and ensemble_tree), as follows:

```
library(MASS)
library(Metrics)
library(tree)

ensemble_lm_rmse <- 0
```



```

ensemble_tree_rmse <- 0

for (i in 1:100) {

  # Fit the linear model with randomized data

  df <- df[sample(nrow(df)),] # Randomize the rows before the analysis

  train <- df[1:400, ]
  test <- df[401:505, ]

  Boston_lm <- lm(medv ~ ., data = train)

  Boston_lm_predictions <- predict(object = Boston_lm, newdata = test)

  # Let's have a quick look at the linear model predictions:

  head(Boston_lm_predictions)

  # Let's calculate the root mean squared error rate of the predictions:

  Boston_linear_rmse[i] <- Metrics::rmse(actual = test$medv, predicted = Boston_lm_predictions)
  Boston_linear_rmse_mean <- mean(Boston_linear_rmse)

  # Let's use tree models

  Boston_tree <- tree(medv ~ ., data = train)

  Boston_tree_predictions <- predict(object = Boston_tree, newdata = test)

  # Let's have a quick look at the tree model predictions:

  head(Boston_tree_predictions)

  # Let's calculate the root mean squared error rate of the predictions:

  Boston_tree_rmse[i] <- Metrics::rmse(actual = test$medv, predicted = Boston_tree_predictions)
  Boston_tree_rmse_mean <- mean(Boston_tree_rmse)

  ensemble <- data.frame( 'linear' = Boston_lm_predictions, 'tree' = Boston_tree_predictions, 'y_en

  ensemble <- ensemble[sample(nrow(ensemble)), ] # Randomizes the rows of the ensemble

  ensemble_train <- ensemble[1:60, ]

```

```

ensemble_test <- ensemble[61:105, ]

# Ensemble linear modeling

ensemble_lm <- lm(y_ensemble ~ ., data = ensemble_train)

# Predictions for the ensemble linear model

ensemble_prediction <- predict(ensemble_lm, newdata = ensemble_test)

# Root mean squared error for the ensemble linear model

ensemble_lm_rmse[i] <- Metrics::rmse(actual = ensemble_test$y_ensemble, predicted = ensemble_prediction)

ensemble_lm_rmse_mean <- mean(ensemble_lm_rmse)

ensemble_tree <- tree(y_ensemble ~ ., data = ensemble_train)

ensemble_tree_predictions <- predict(object = ensemble_tree, newdata = ensemble_test)

ensemble_tree_rmse[i] <- Metrics::rmse(actual = ensemble_test$y_ensemble, predicted = ensemble_tree_predictions)

ensemble_tree_rmse_mean <- mean(ensemble_tree_rmse)

results <- list( 'Linear' = Boston_linear_rmse_mean, 'Trees' = Boston_tree_rmse_mean,
)

results
#> $Linear
#> [1] 4.881756
#>
#> $Trees
#> [1] 4.759513
#>
#> $Ensembles_Linear
#> [1] 4.200395
#>
#> $Ensemble_Tree
#> [1] 5.140953

warnings()

Boston_lm <- Boston_lm
Boston_tree <- Boston_tree

```

```
ensemble_lm <- ensemble_lm  
ensemble_tree <- ensemble_tree
```

2.13.1 What about classification, logistic and time series data?

In subsequent chapters we will do similar processes with classification, logistic and time series data. It's possible to build ensembles with all these types of data. The results are extremely similar to the results we've seen here with numerical data: While the ensembles won't always have the best results, it is best to have a diverse set of models and ensembles to get the best possible results.

2.13.2 Principle: Ensembles can work with many types of data, and we will do that in this book

2.13.3 Can it make predictions on totally new data from the trained models—including the ensembles?

The solutions in this book are independent of the use of the data. We will look at everything from housing prices to business analysis to HR analytics to research in medicine. One of our later examples will do exactly what this question is asking—build individual and ensemble models from data, then use those pre-trained models to make predictions on totally unseen data. You will develop this set of skills later in the book, but it's a minor extension of what you're already seen and completed.

2.13.4 The way I was taught how to write code was totally wrong for me: The best way for me is to start at the end and work backward from there. Do not start coding looking for a solution, instead, start with the ending and work backwards from there.

Start at the end and work backwards from there

The biggest lesson for me in all of this work is how to make ensembles. You've already seen some of the steps, and there are more results to come. The second biggest lesson is that everything I was taught about how to do data science and AI was backwards to what actually works for me in real life. I've learned how I learn, and applied that skill (learning how I learn) to a wide range of skills, including:

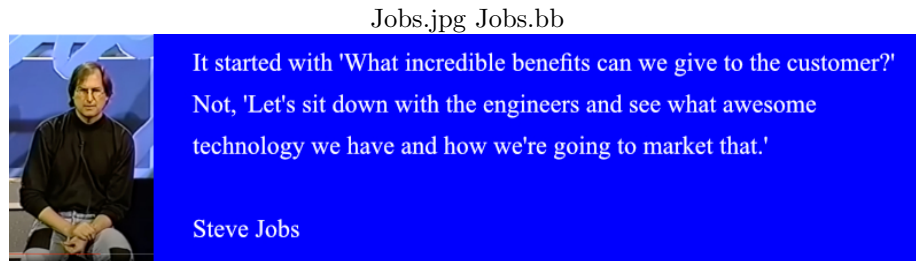


Figure 2.8: Start at the end and work backwards

- Running a multi-million dollar division of a Fortune 1000 company, including full profit and loss responsibility
- Performing at a professional level on many musical instruments
- Able to communicate in English, Spanish and sign language in a professional setting
- Earning the #1 place on the annual undergraduate university mathematics competition—twice
- Completing a Master’s degree in Guidance and Counseling, allowing me to help many people in their path toward a healthier life
- Leader of the Oak Park, Illinois chapter of Amnesty International for ten years, helping to release several Prisoners of Conscience
- President of the Chicago Apple User Group for ten years, helping many people do extremely good work with their hardware and software
- Leg press 1,000 pounds ten times in a row
- Climbed a mountain in Colorado
- Completed multiple skydives (and looking forward to doing more)

The point here is that I have learned how I learn, and I’ve applied that skill to many areas. When I started learning data science/AI/coding, it was all very different from the way I was being creative my whole life. The way that works for me is to start at the end, work backward from there, and never give up. Maybe the best evidence of the success of this method is this fact:

When I started to write the code that led to the Ensembles package, I followed those steps: Start at the end, work backward from there, and never give up. I wound up writing an average of 1,000 lines of clean, error free code per month for 15 months. The Ensembles package is around 15,000 lines of clean, error free code.

I found my attitude was much more important than my skill set, by a long shot.

2.13.5 How I stuck with it all the way to the end: The best career advice I ever received was from a homeless man I never met, and answers the question of what most strongly predicts success.

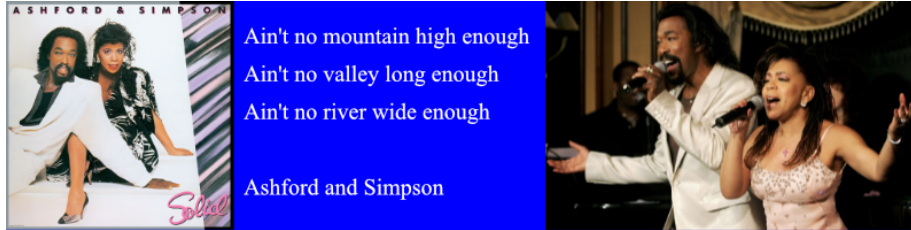


Figure 2.9: Ashford and Simpson

Ashford and Simpson

Learning about building ensembles will help you make more accurate predictions. That's an extremely good skill to have in any setting. But I found the most important thing to predict is success. This has been studied, and there are quite a few good works on the subject, both academic and for the general population.

My favorite career advice—which I listened to nearly every day as I worked on the Ensembles project—is from a man who was homeless at the time he came up with the words.

Nick Ashford was from Willow Run, Michigan. He moved to New York, hoping to get into the entertainment world as a dancer. Unfortunately he ended up homeless on the streets of New York. He slept on park benches, and got food from soup kitchens.

He heard that the people at White Rock Baptist Church would feed him (a homeless man) a normal meal, so Nick went there one Sunday morning. He met the people, especially the choir members, and started working with the piano player in the choir. Her name is Valerie Simpson.

Soon Nick and Valerie were writing songs for the church choir. Nick mentioned that while he was homeless, he realized that New York wasn't going to "do me in". He was determined. The words he put down say:

Ain't no mountain high enough

Ain't no valley low enough

Ain't no river wide enough

Valerie took those words, and set them to music. They sent that song to Motown, who released it with Marvin Gaye and Tammy Terrell covering the vocals. It

was later re-done by Ashford and Simpson and Paul Riser, with Diana Ross singing the lead.

Here is a short video that summarizes that experience, and concludes with the finale of the 1970 version of the song. This attitude that Ashford and Simpson expressed in song is extremely highly predictive of success, no matter what the field of endeavor. I found this extremely motivating, and used it to overcome any obstacles and challenges I had while on the journey.

While I have the skill of knowing how I learn (which I will continue to share with you in this book), this attitude of working no matter how high the mountain or long the valley or wide the river, gives me the how and the why to keep moving toward success, until that success is fully achieved.

Later on we will look at how to make presentations, consider this as an example of the level of quality that can be done:

<https://www.icloud.com/iclouddrive/002bNfVreagRYCYHAZ9GyQ02w#Ain't%5FNo%5FMountain%5F>

2.13.6 Exercises:

1. Find your data science Genesis. The data science idea that totally excites you and gets you out of bed every day. The idea that leads to the creation of many other ideas. The biggest and boldest dreams you can possibly have. The idea that is so strong that you have to do it. Not for yourself, but for the benefit of all who will use it and receive all the good it will create.
2. Keep a journal of your progress. It's much easier to see results over time when there is a record. Set the journal up today (or this week). I did not use Github as a journal. My journal was for crazy ideas, contradictory evidence, writing down my frustrations and successes, inspiration, the one next thing I worked on, and having a rock solid record of the path to success. Seeing the path I traversed was a huge motivation to finishing the project.
3. Do your best to add journal entries to your regular schedule.
4. Make an ensemble using the Boston Housing data set. Model any of the other 13 columns of data, not the median value of the home (14th column) which we have been working on in this chapter.
5. Start planning for your comprehensive project. What types of data are you most interested in? What patterns would you like to discover? Begin looking online now for possible data sets, and so a little basic research. More examples will be provided as we get closer to that section of the book.

Chapter 3

Numerical data: How to make 23 individual models, and basic skills with functions

This is where we will begin building the skills to make ensembles of models of numerical data. However, this is going to be much easier than it might appear at first. Let's see how we can make this as easy as possible.

How to work backwards and make the function we need: Start from the end

We are going to start at the ending, not at the beginning, and work backwards from there. This method is much, much easier than working forward, as you will see throughout this book. While it might be a little uncomfortable at first, this skill will allow you to complete your work at a faster rate than if you work forward.

We'll use the Boston Housing data set, and we'll start with the Bagged Random Forest function. For now we're only going to work with one function, to keep everything simple. In essence, we are going to run this like an assembly line.

We want the ending to be the error rate by model. Virtually any customer you work with is going to want to know, "How accurate is it?" That's our starting point.

How do we determine model accuracy? We already did this in the previous chapter, finding the root mean squared error for the individual models and the ensemble models. We're going to do the same steps here, so the process is familiar to you.

To get the error rate by model on the holdout data sets (test and validation), we're going to need a model (Bagged Random Forest in this first example), fit to the training data, and use that model to make predictions on the test data. We can then measure the error in the predictions, just as we did before. These steps should be familiar to you. If not, please re-read the previous chapter.

But what do we need to complete those steps? We're going to have to go backward (a little) and make a function that will allow us to work with any data set.

What does our function need? Let's make a list:

- The data (such as Boston housing)
- Column number (such as 14, the median value of the property)
- Train amount
- Test amount
- Validation amount
- Number of times to resample

One of the key steps here is to change the name of the target variable to `y`. The initial name could be nearly anything, but this method changes the name of the target variable to `y`. This allows us to make one small change that will allow this to be the easiest possible solution:

3.0.1 All our models will be structured the same way: `y ~ ., data = train`

This means that `y` (our target value) is a function of the other features, and the data set is the training data set. While there will be some variations on this in our 27 models, the basic structure is the same.

3.0.2 Having the same structure for all the models makes it much easier to build, debug, and deploy the completed models.

Then we only need to start with our initial values, and it will run.

One extremely nice part about creating models this way is the enormous efficiency it gives us. Once we have the Bagged Random Forest model working, we will be able to use very similar (and identical in many cases!) processes with other models (such as Support Vector Machines).

The rock solid foundation we lay at the beginning will allow us to have a smooth and easy experience once the foundation is solid and we use it to build more

models. The other models will mainly be almost exact duplicates of our first example.'

Here are the steps we will follow:

- Load the library
- Set initial values to 0
- Create the function
- Set up random resampling
- Break the data into train and test
- Fit the model on the training data, make predictions and measure error on the test data
- Return the results
- Check for errors or warnings
- Test on a different data set

3.0.3 Exercise: Re-read the steps above how we will work backwards to come up with the function we need.

```
library(e1071) # will allow us to use a tuned random forest model
library(Metrics) # Will allow us to calculate the root mean squared error
library(randomForest) # To use the random forest function
#> randomForest 4.7-1.1
#> Type rfNews() to see new features/changes/bug fixes.
```

```
library(tidyverse) # Amazing set of tools for data science
#> -- Attaching core tidyverse packages ---- tidyverse 2.0.0 --
#> v dplyr      1.1.4      v readr      2.1.5
#> v forcats    1.0.0      v stringr    1.5.1
#> v ggplot2     3.5.1      v tibble     3.2.1
#> v lubridate  1.9.3      v tidyr      1.3.1
#> v purrr       1.0.2
#> -- Conflicts ----- tidyverse_conflicts() --
#> x dplyr::combine() masks randomForest::combine()
#> x dplyr::filter()  masks stats::filter()
#> x dplyr::lag()      masks stats::lag()
#> x ggplot2::margin() masks randomForest::margin()
#> i Use the conflicted package (<http://conflicted.r-lib.org/>) to force all conflicts to become
# Set initial values to 0. The function will return an error if any of these are left out.
```

```

bag_rf_holdout_RMSE <- 0
bag_rf_holdout_RMSE_mean <- 0
bag_rf_train_RMSE <- 0
bag_rf_test_RMSE <- 0
bag_rf_validation_RMSE <- 0

# Define the function

numerical_1 <- function(data, colnum, train_amount, test_amount, numresamples){

#Set up random resampling

for (i in 1:numresamples) {

# Changes the name of the target column to y
y <- 0
colnames(data)[colnum] <- "y"

# Moves the target column to the last column on the right
df <- data %>% dplyr::relocate(y, .after = last_col())
df <- df[sample(nrow(df)), ] # randomizes the rows

#Breaks the data into train and test sets
idx <- sample(seq(1, 2), size = nrow(df), replace = TRUE, prob = c(train_amount, test_
train <- df[idx == 1, ]
test <- df[idx == 2, ]

# Fit the model to the training data, make predictions on the testing data, then calcu
bag_rf_train_fit <- e1071::tune.randomForest(x = train, y = train$y, mtry = ncol(train
bag_rf_train_RMSE[i] <- Metrics::rmse(actual = train$y, predicted = predict( object =
bag_rf_train_RMSE_mean <- mean(bag_rf_train_RMSE)
bag_rf_test_RMSE[i] <- Metrics::rmse(actual = test$y, predicted = predict( object = bag
bag_rf_test_RMSE_mean <- mean(bag_rf_test_RMSE)

# Itemize the error on the holdout data sets, and calculate the mean of the results
bag_rf_holdout_RMSE[i] <- mean(bag_rf_test_RMSE_mean)
bag_rf_holdout_RMSE_mean <- mean(c(bag_rf_holdout_RMSE))

# These are the predictions we will need when we make the ensembles
bag_rf_test_predict_value <- as.numeric(predict(object = bag_rf_train_fit$best.model,

#Return the mean of the results to the user

return(bag_rf_holdout_RMSE_mean)

```

```

} # closing brace for numresamples
} # closing brace for numerical_1 function

# Here is our first numerical function in actual use. We will use 25 resamples

numerical_1(data = MASS::Boston, colnum = 14, train_amount = 0.60, test_amount = 0.40, numresamples = 25)
#> [1] 0.2555779

warnings() # no warnings, the best possible result

```

Exercise: Try it yourself: Change the values of train, test and validation, and the number of resamples. See how those change the result.

One of your own: Find any numerical data set, and make a bagged random forest function for that data set. (For example, you may use the Auto data set in the ISLR package. You will need to remove the last column, vehicle name. Model mpg as a function of the other features using the Bagged Random Forest function, but any numerical data set will work).

Post: Share on social your first results making a numerical function (screen shot/video optional at this stage, we will be learning how to do those later)

For example, “Did my first data science function building up to making ensembles later on. Got everything to run, no errors. #AIEnsembles”

Now we will build the remaining 22 models for numerical data. They are all built using the same structure, on the same foundation.

Now that we know how to build a basic function, let’s build the 22 other sets of tools we will need to make our ensemble, starting with bagging:

3.0.4 Bagging (bootstrap aggregating)

```

library(ipred) #for the bagging function

# Set initial values to 0
bagging_train_RMSE <- 0
bagging_test_RMSE <- 0
bagging_validation_RMSE <- 0
bagging_holdout_RMSE <- 0
bagging_test_predict_value <- 0
bagging_validation_predict_value <- 0

#Create the function:

bagging_1 <- function(data, colnum, train_amount, test_amount, validation_amount, numresamples){

```

```

#Set up random resampling
for (i in 1:numresamples) {

  #Changes the name of the target column to y
  y <- 0
  colnames(data)[colnum] <- "y"

  # Moves the target column to the last column on the right
  df <- data %>% dplyr::relocate(y, .after = last_col()) # Moves the target column to the

  # Breaks the data into train and test sets

  idx <- sample(seq(1, 2), size = nrow(df), replace = TRUE, prob = c(train_amount, test_
  train <- df[idx == 1, ]
  test <- df[idx == 2, ]

  # Fit the model to the training data, calculate error, make predictions on the holdout

  bagging_train_fit <- ipred::bagging(formula = y ~ ., data = train)
  bagging_train_RMSE[i] <- Metrics::rmse(actual = train$y, predicted = predict(object = 1
  bagging_train_RMSE_mean <- mean(bagging_train_RMSE)
  bagging_test_RMSE[i] <- Metrics::rmse(actual = test$y, predicted = predict(object = ba
  bagging_test_RMSE_mean <- mean(bagging_test_RMSE)
  bagging_holdout_RMSE[i] <- mean(bagging_test_RMSE_mean)
  bagging_holdout_RMSE_mean <- mean(bagging_holdout_RMSE)
  y_hat_bagging <- c(bagging_test_predict_value)

  return(bagging_holdout_RMSE_mean)

} # closing braces for the resampling function
} # closing braces for the bagging function

# Test the function:
bagging_1(data = MASS::Boston, colnum = 14, train_amount = 0.60, test_amount = 0.20, n
#> [1] 3.477499

warnings() # no warnings

```

3.0.5 BayesGLM

```

library(arm) # to use bayesglm function
#> Loading required package: MASS
#>
#> Attaching package: 'MASS'

```

```

#> The following object is masked from 'package:dplyr':
#>
#>      select
#> Loading required package: Matrix
#>
#> Attaching package: 'Matrix'
#> The following objects are masked from 'package:tidyr':
#>
#>      expand, pack, unpack
#> Loading required package: lme4
#>
#> arm (Version 1.14-4, built: 2024-4-1)
#> Working directory is /Users/russellconte/Library/Mobile Documents/com-apple~CloudDocs/Documents

# Set initial values to 0
bayesglm_train_RMSE <- 0
bayesglm_test_RMSE <- 0
bayesglm_validation_RMSE <- 0
bayesglm_holdout_RMSE <- 0
bayesglm_test_predict_value <- 0
bayesglm_validation_predict_value <- 0

# Create the function:
bayesglm_1 <- function(data, colnum, train_amount, test_amount, numresamples){

#Set up random resampling
for (i in 1:numresamples) {

#Changes the name of the target column to y
y <- 0
colnames(data)[colnum] <- "y"

#Moves the target column to the last column on the right
df <- data %>% dplyr::relocate(y, .after = last_col()) # Moves the target column to the last column

#Breaks the data into train, test and validation sets
idx <- sample(seq(1, 2), size = nrow(df), replace = TRUE, prob = c(train_amount, test_amount))
train <- df[idx == 1, ]
test <- df[idx == 2, ]

bayesglm_train_fit <- arm::bayesglm(y ~ ., data = train, family = gaussian(link = "identity"))
bayesglm_train_RMSE[i] <- Metrics::rmse(actual = train$y, predicted = predict(object = bayesglm_train_fit))
bayesglm_train_RMSE_mean <- mean(bayesglm_train_RMSE)
bayesglm_test_RMSE[i] <- Metrics::rmse(actual = test$y, predicted = predict(object = bayesglm_train_fit))
bayesglm_test_RMSE_mean <- mean(bayesglm_test_RMSE)

```

```

y_hat_bayesglm <- c(bayesglm_test_predict_value)

return(bayesglm_test_RMSE_mean)

} # closing braces for resampling
} # closing braces for the function

bayesglm_1(data = MASS::Boston, colnum = 14, train_amount = 0.60, test_amount = 0.20, n
#> [1] 4.938509

warnings() # no warnings

```

BayesRNN

{r BayesRNN model for numerical data}

library(brnn) # so we can use the BayesRNN function

Chapter 4

Set initial values to 0

```
bayesrnn_train_RMSE <- 0 bayesrnn_test_RMSE <- 0 bayesrnn_validation_RMSE  
<- 0 bayesrnn_holdout_RMSE <- 0 bayesrnn_test_predict_value <- 0  
bayesrnn_validation_predict_value <- 0
```


Chapter 5

Create the function:

```
bayesrnn <- function(data, colnum, train_amount, test_amount, validation_amount, numresamples){
```


Chapter 6

Set up random resampling

```
for (i in 1:numresamples) {
```


Chapter 7

Changes the name of the
target column to y

```
y <- 0 colnames(data)[colnum] <- "y"
```


Chapter 8

Moves the target column to the last column on the right

```
df <- data %>% dplyr::relocate(y, .after = last_col()) # Moves the target column to the last column on the right
df <- df[sample(nrow(df)), ] # randomizes the rows
```


Chapter 9

Breaks the data into train, test and validation sets

```
idx <- sample(seq(1, 3), size = nrow(df), replace = TRUE, prob =
c(train_amount, test_amount, validation_amount)) train <- df[idx ==
1,] test <- df[idx == 2, ] validation <- df[idx == 3, ]

bayesrnn_train_fit <- brnn::brnn(x = as.matrix(train), y = trainy) bayesrnntrainRMSE[i] <-
-Metrics :: rmse(actual = trainy, predicted = predict(object = bayesrnn_train_fit,
newdata = train)) bayesrnn_train_RMSE_mean <- mean(bayesrnn_train_RMSE)
bayesrnn_test_RMSE[i] <- Metrics::rmse(actual = testy, predicted =
predict(object = bayesrnntrainfit, newdata = test)) bayesrnntestRMSEmean <
-mean(bayesrnntestRMSE) bayesrnnvalidationRMSE[i] <- -Metrics ::
rmse(actual = validationy, predicted = predict(object = bayesrnn_train_fit,
newdata = validation)) bayesrnn_validation_RMSE_mean <- mean(bayesrnn_validation_RMSE)
bayesrnn_holdout_RMSE[i] <- mean(c(bayesrnn_test_RMSE_mean,
bayesrnn_validation_RMSE_mean)) bayesrnn_holdout_RMSE_mean <-
mean(bayesrnn_holdout_RMSE)

bayesrnn_test_predict_value <- as.numeric(predict(object = bayesrnn_train_fit,
newdata = test)) bayesrnn_validation_predict_value <- as.numeric(predict(object
= bayesrnn_train_fit, newdata = validation)) y_hat_bayesrnn <- c(bayesrnn_test_predict_value,
bayesrnn_validation_predict_value)

return(bayesrnn_holdout_RMSE_mean)
} # Closing brace for number of resamples } # Closing brace for the function

bayesrnn(data = MASS::Boston, colnum = 14, train_amount = 0.60,
test_amount = 0.20, validation_amount = 0.20, numresamples = 25)
warnings() # no warnings for BayesRNN function
```

Boosted Random Forest

```
{r Individual numerical model for Boosted Random Forest}
```

```
library(e1071) library(randomForest)
```

Chapter 10

Set initial values to 0

Chapter 11

Set initial values to 0

```
boost_rf_train_RMSE <- 0 boost_rf_test_RMSE <- 0 boost_rf_validation_RMSE  
<- 0 boost_rf_holdout_RMSE <- 0 boost_rf_test_predict_value <- 0  
boost_rf_validation_predict_value <- 0
```


Chapter 12

Create the function:

```
boost_rf <- function(data, colnum, train_amount, test_amount, validation_amount, numresamples){
```


Chapter 13

Set up random resampling

```
for (i in 1:numresamples) {
```


Chapter 14

Changes the name of the
target column to y

```
y <- 0 colnames(data)[colnum] <- "y"
```


Chapter 15

Moves the target column to the last column on the right

```
df <- data %>% dplyr::relocate(y, .after = last_col()) # Moves the target column to the last column on the right
df <- df[sample(nrow(df)), ] # randomizes the rows
```


Chapter 16

Breaks the data into train, test and validation sets

```
idx <- sample(seq(1, 3), size = nrow(df), replace = TRUE, prob =  
c(train_amount, test_amount, validation_amount)) train <- df[idx ==  
1,] test <- df[idx == 2, ] validation <- df[idx == 3, ]
```


Chapter 17

Fit boosted random forest model on the training data, make predictions on holdout data

```
boost_rf_train_fit <- e1071::tune.randomForest(x = train, y = trainy, mtry =
ncol(train) - 1)boost_rf_train_fit$best.model$RMSE[i] < -Metrics :: rmse(actual = trainy,
predicted = predict( object = boost_rf_train_fit$best.model, newdata =
train))boost_rf_train_fit$mean < -mean(boost_rf_train_fit$RMSE)boost_rf_test_fit$RMSE[i] <
-Metrics :: rmse(actual = testy, predicted = predict( object = boost_rf_train_fit$best.model, newdata =
test))boost_rf_test_fit$mean < -mean(boost_rf_test_fit$RMSE)boost_rf_validation_fit$RMSE[i] <
-Metrics :: rmse(actual = validationy, predicted = predict( object =
boost_rf_train_fit$best.model, newdata = validation )) boost_rf_validation_RMSE_mean
<- mean(boost_rf_validation_RMSE) boost_rf_holdout_RMSE[i] <-
mean(boost_rf_test_RMSE_mean, boost_rf_validation_RMSE_mean)
boost_rf_holdout_RMSE_mean <- mean(boost_rf_holdout_RMSE)

boost_rf_test_predict_value <- as.numeric(predict(object = boost_rf_train_fit$best.model, newdata =
test))boost_rf_validation_predict_value < -as.numeric(predict(object =
boost_rf_train_fit$best.model, newdata = validation)) y_hat_boost_rf <-
c(boost_rf_test_predict_value, boost_rf_validation_predict_value)

return(boost_rf_holdout_RMSE_mean)

} # closing brace for numresamples } # closing brace for the function

boost_rf(data = MASS::Boston, colnum = 14, train_amount = 0.60,
test_amount = 0.20, validation_amount = 0.20, numresamples = 25)
warnings() # no warnings for Boosted Random Forest function
```

Cubist

{r Individual model based on the cubist function}

library(Cubist)

Chapter 18

Set initial values to 0

```
cubist_train_RMSE <- 0 cubist_test_RMSE <- 0 cubist_validation_RMSE  
<- 0 cubist_holdout_RMSE <- 0 cubist_test_predict_value <- 0 cu-  
bist_validation_predict_value <- 0
```


Chapter 19

Create the function:

```
cubist <- function(data, colnum, train_amount, test_amount, validation_amount, numresamples){
```


Chapter 20

Set up random resampling

```
for (i in 1:numresamples) {
```


Chapter 21

Changes the name of the
target column to y

```
y <- 0 colnames(data)[colnum] <- "y"
```


Chapter 22

Moves the target column to the last column on the right

```
df <- data %>% dplyr::relocate(y, .after = last_col()) # Moves the target column to the last column on the right
df <- df[sample(nrow(df)), ] # randomizes the rows
```


Chapter 23

Breaks the data into train, test and validation sets

```
idx <- sample(seq(1, 3), size = nrow(df), replace = TRUE, prob =  
c(train_amount, test_amount, validation_amount)) train <- df[idx ==  
1,] test <- df[idx == 2, ] validation <- df[idx == 3, ]
```


Chapter 24

Fit the model on the training data, make predictions on the holdout data

```
cubist_train_fit <- Cubist::cubist(x = train[, 1:ncol(train) - 1], y =
trainy)cubist_train_RMSE[i] <- Metrics::rmse(actual = trainy, predicted =
predict(object = cubist_train_fit, newdata = train)) cubist_train_RMSE_mean
<- mean(cubist_train_RMSE) cubist_test_RMSE[i] <- Metrics::rmse(actual
= testy, predicted = predict(object = cubist_train_fit, newdata = test))cubist_test_RMSE_mean <-
mean(cubist_test_RMSE)cubist_validation_RMSE[i] <- Metrics::rmse(actual =
validationy, predicted = predict(object = cubist_train_fit, newdata = valida-
tion)) cubist_validation_RMSE_mean <- mean(cubist_validation_RMSE) cu-
bist_holdout_RMSE[i] <- mean(cubist_test_RMSE_mean, cubist_validation_RMSE_mean)
cubist_holdout_RMSE_mean <- mean(cubist_holdout_RMSE)

cubist_test_predict_value <- as.numeric(predict(object = cubist_train_fit,
newdata = test)) cubist_validation_predict_value <- as.numeric(predict(object
= cubist_train_fit, newdata = validation)) cubist_predict_value_mean <-
mean(c(cubist_test_predict_value, cubist_validation_predict_value))

return(cubist_holdout_RMSE_mean)

} # closing braces for numresamples } # closing braces for the function

cubist(data = MASS::Boston, colnum = 14, train_amount = 0.60, test_amount
= 0.20, validation_amount = 0.20, numresamples = 25) warnings() # no warn-
ings for individual cubist function
```

Elastic

{r Individual elastic model for numerical data}

library(glmnet) # So we can run the elastic model

Chapter 25

Set initial values to 0

```
elastic_train_RMSE <- 0 elastic_test_RMSE <- 0 elastic_validation_RMSE  
<- 0 elastic_holdout_RMSE <- 0 elastic_test_predict_value <- 0  
elastic_validation_predict_value <- 0 elastic_test_RMSE <- 0 elas-  
tic_test_RMSE_df <- data.frame(elastic_test_RMSE) elastic_validation_RMSE  
<- 0 elastic_validation_RMSE_df <- data.frame(elastic_validation_RMSE)  
elastic_holdout_RMSE <- 0 elastic_holdout_RMSE_df <- data.frame(elastic_holdout_RMSE)
```


Chapter 26

Create the function:

```
elastic <- function(data, colnum, train_amount, test_amount, valida-  
tion_amount, numresamples){
```


Chapter 27

Set up random resampling

```
for (i in 1:numresamples) {
```


Chapter 28

Changes the name of the
target column to y

```
y <- 0 colnames(data)[colnum] <- "y"
```


Chapter 29

Moves the target column to the last column on the right

```
df <- data %>% dplyr::relocate(y, .after = last_col()) # Moves the target column to the last column on the right
df <- df[sample(nrow(df)), ] # randomizes the rows
```


Chapter 30

Breaks the data into train, test and validation sets

```
idx <- sample(seq(1, 3), size = nrow(df), replace = TRUE, prob =  
c(train_amount, test_amount, validation_amount)) train <- df[idx ==  
1,] test <- df[idx == 2, ] validation <- df[idx == 3, ]
```


Chapter 31

Set up the elastic model

```
y <- trainy x <- data.matrix(train) elastic_model <- glmnet(x, y, alpha =
0.5) elastic_cv <- cv.glmnet(x, y, alpha = 0.5) best_elastic_lambda <- elastic_cv$lambda.min
best_elastic_model <- glmnet::glmnet(x, y, alpha = 0, lambda =
best_elastic_lambda) elastic_test_pred <- predict(best_elastic_model,
s = best_elastic_lambda, newx = data.matrix(test %>% dplyr::select(-y)))

elastic_test_RMSE <- Metrics::rmse(actual = test$y, predicted = elastic_test_pred) elastic_test_RMSE_df <-
rbind(elastic_test_RMSE, elastic_test_RMSE) elastic_test_RMSE_mean <-
mean(elastic_test_RMSE) elastic_test_RMSE[2:nrow(elastic_test_RMSE_df)]
```

31.1 Elastic using the validation data set

```
y <- trainy x <- data.matrix(train) elastic_model <- glmnet(x, y, alpha =
0.5) elastic_cv <- cv.glmnet(x, y, alpha = 0.5) best_elastic_lambda <- elastic_cv$lambda.min
best_elastic_model <- glmnet::glmnet(x, y, alpha = 0, lambda =
best_elastic_lambda) elastic_validation_pred <- predict(best_elastic_model,
s = best_elastic_lambda, newx = data.matrix(validation %>% dplyr::select(-
y))) elastic_validation_RMSE <- Metrics::rmse(actual = validation$y, predicted =
elastic_validation_pred) elastic_validation_RMSE_df <- rbind(elastic_validation_RMSE, elastic_validation_RMSE) elastic_validation_RMSE_mean <-
mean(elastic_validation_RMSE) elastic_validation_RMSE[2:nrow(elastic_validation_RMSE_df)]

elastic_holdout_RMSE <- mean(elastic_test_RMSE_mean, elastic_validation_RMSE_mean)
elastic_holdout_RMSE_df <- rbind(elastic_holdout_RMSE_df, elastic_holdout_RMSE) elastic_holdout_RMSE_mean <- mean(elastic_holdout_RMSE_df$elastic_holdout_RMSE)

elastic_test_predict_value[i] <- round(mean(elastic_test_pred), 4) elastic_test_predict_value_mean <- mean(elastic_test_predict_value)

elastic_validation_predict_value[i] <- round(mean(elastic_validation_pred), 4) elastic_validation_predict_value_mean <- mean(elastic_validation_predict_value)
```

```
elastic_test_predict_value_mean <- mean(c(elastic_test_predict_value_mean,
elastic_validation_predict_value_mean))

return(elastic_holdout_RMSE_mean)

} # closing brace for numresample } # closing brace for the elastic function

elastic(data = MASS::Boston, colnum = 14, train_amount = 0.60, test_amount
= 0.20, validation_amount = 0.20, numresamples = 25) warnings() # no warn-
ings for individual elastic function

Generalized Additive Models with smoothing splines

{r Individual model of generalized additive models with smoothng splines}

library(gam) # for fitting generalized additive models
```

Chapter 32

Set initial values to 0

```
gam_train_RMSE <- 0 gam_test_RMSE <- 0 gam_validation_RMSE <- 0  
gam_holdout_RMSE <- 0 gam_test_predict_value <- 0 gam_validation_predict_value  
<- 0
```


Chapter 33

Create the function:

```
gam1 <- function(data, colnum, train_amount, test_amount, validation_amount, numresamples){
```


Chapter 34

Set up random resampling

```
for (i in 1:numresamples) {
```


Chapter 35

Changes the name of the
target column to y

```
y <- 0 colnames(data)[colnum] <- "y"
```


Chapter 36

Moves the target column to the last column on the right

```
df <- data %>% dplyr::relocate(y, .after = last_col()) # Moves the target column to the last column on the right
df <- df[sample(nrow(df)), ] # randomizes the rows
```


Chapter 37

Breaks the data into train, test and validation sets

```
idx <- sample(seq(1, 3), size = nrow(df), replace = TRUE, prob =  
c(train_amount, test_amount, validation_amount)) train <- df[idx ==  
1,] test <- df[idx == 2, ] validation <- df[idx == 3, ]
```


Chapter 38

Set up to fit the model on the training data

```
n_unique_vals <- purrr::map_dbl(df, dplyr::n_distinct)
```


Chapter 39

Names of columns with \geq 4 unique vals

```
keep <- names(n_unique_vals)[n_unique_vals >= 4]  
gam_data <- df %>% dplyr::select(dplyr::all_of(keep))
```


Chapter 40

Model data

```
train1 <- train %>% dplyr::select(dplyr::all_of(keep))
test1 <- test %>% dplyr::select(dplyr::all_of(keep))
validation1 <- validation %>% dplyr::select(dplyr::all_of(keep))

names_df <- names(gam_data[, 1:ncol(gam_data) - 1]) f2 <- stats::as.formula(paste0("y
~", paste0("gam::s(", names_df, ") ", collapse = "+"))) gam_train_fit <-
gam::gam(f2, data = train1) gam_train_RMSE[i] <- Metrics::rmse(actual =
trainy, predicted = predict(object = gam_train_fit, newdata = train))  $gam_{train\_RMSE\_mean} <-$ 
 $-mean(gam_{train\_RMSE})$   $gam_{test\_RMSE}[i] < -Metrics :: rmse(actual =$ 
testy, predicted = predict(object = gam_train_fit, newdata = test))
gam_test_RMSE_mean <- mean(gam_test_RMSE) gam_validation_RMSE[i]
<- Metrics::rmse(actual = validation$y, predicted = predict(object =
gam_train_fit, newdata = validation)) gam_validation_RMSE_mean <-
mean(gam_validation_RMSE) gam_holdout_RMSE[i] <- mean(gam_test_RMSE_mean,
gam_validation_RMSE_mean) gam_holdout_RMSE_mean <- mean(gam_holdout_RMSE)
gam_holdout_RMSE_sd_mean <- sd(c(gam_test_RMSE_mean, gam_validation_RMSE_mean))
gam_train_predict_value <- as.numeric(predict(object = gam_train_fit,
newdata = train)) gam_test_predict_value <- as.numeric(predict(object
= gam_train_fit, newdata = test)) gam_validation_predict_value
<- as.numeric(predict(object = gam_train_fit, newdata = validation))
gam_predict_value_mean <- mean(c(gam_test_predict_value, gam_validation_predict_value))

return(gam_holdout_RMSE_sd_mean )

} # closing braces for numresamples } # closing braces for gam function

gam1(data = MASS::Boston, colnum = 14, train_amount = 0.60, test_amount
= 0.20, validation_amount = 0.20, numresamples = 25) warnings() # no warn-
ings for individual gam function
```

Gradient Boosted

```
{r Individual gradient boosted model for numerical data}
```

```
library(gbm) # to allow use of gradient boosted models
```


Chapter 41

Set initial values to 0

```
gb_train_RMSE <- 0 gb_test_RMSE <- 0 gb_validation_RMSE <- 0
gb_holdout_RMSE <- 0 gb_test_predict_value <- 0 gb_validation_predict_value
<- 0

gb1 <- function(data, colnum, train_amount, test_amount, valida-
tion_amount, numresamples){
```


Chapter 42

Set up random resampling

```
for (i in 1:numresamples) {
```


Chapter 43

Changes the name of the
target column to y

```
y <- 0 colnames(data)[colnum] <- "y"
```


Chapter 44

Moves the target column to the last column on the right

```
df <- data %>% dplyr::relocate(y, .after = last_col()) # Moves the target column to the last column on the right
df <- df[sample(nrow(df)), ] # randomizes the rows
```


Chapter 45

Breaks the data into train, test and validation sets

```
idx <- sample(seq(1, 3), size = nrow(df), replace = TRUE, prob =
c(train_amount, test_amount, validation_amount)) train <- df[idx ==
1,] test <- df[idx == 2, ] validation <- df[idx == 3, ]

gb_train_fit <- gbm::gbm(trainy, data = train, distribution = "gaussian", n.trees =
100, shrinkage = 0.1, interaction.depth = 10) gb_train_RMSE[i] <- Metrics::
rmse(actual = trainy, predicted = predict(object = gb_train_fit, newdata =
train)) gb_train_RMSE_mean <- mean(gb_train_RMSE) gb_test_RMSE[i]
<- Metrics::rmse(actual = testy, predicted = predict(object = gb_train_fit, newdata =
test)) gb_test_RMSE_mean <- mean(gb_test_RMSE) gb_validation_RMSE[i] <-
Metrics::rmse(actual = validationy, predicted = predict(object =
gb_train_fit, newdata = validation)) gb_validation_RMSE_mean <-
mean(gb_validation_RMSE) gb_holdout_RMSE[i] <- mean(c(gb_test_RMSE_mean,
gb_validation_RMSE_mean)) gb_holdout_RMSE_mean <- mean(gb_holdout_RMSE)

gb_train_predict_value <- as.numeric(predict(object = gb_train_fit, newdata
= train)) gb_test_predict_value <- as.numeric(predict(object = gb_train_fit,
newdata = test)) gb_validation_predict_value <- as.numeric(predict(object
= gb_train_fit, newdata = validation)) gb_predict_value_mean <-
mean(c(gb_test_predict_value, gb_validation_predict_value))

return(gb_holdout_RMSE_mean)

} # closing brace for numresamples } # closing brace for gb1 function

gb1(data = MASS::Boston, colnum = 14, train_amount = 0.60, test_amount =
0.20, validation_amount = 0.20, numresamples = 25) warnings() # no warnings
for individual gradient boosted function

K-Nearest Neighbors (tuned)
```

```
{r Individual tuned KNN model for numerical data}
```

```
library(e1071)
```

Chapter 46

Set initial values to 0

```
knn_train_RMSE <- 0 knn_test_RMSE <- 0 knn_validation_RMSE <- 0
knn_holdout_RMSE <- 0 knn_test_predict_value <- 0 knn_validation_predict_value
<- 0

knn1 <- function(data, colnum, train_amount, test_amount, valida-
tion_amount, numresamples){
```


Chapter 47

Set up random resampling

```
for (i in 1:numresamples) {
```


Chapter 48

Changes the name of the
target column to y

```
y <- 0 colnames(data)[colnum] <- "y"
```


Chapter 49

Moves the target column to the last column on the right

```
df <- data %>% dplyr::relocate(y, .after = last_col()) # Moves the target column to the last column on the right
df <- df[sample(nrow(df)), ] # randomizes the rows
```


Chapter 50

Breaks the data into train, test and validation sets

```
idx <- sample(seq(1, 3), size = nrow(df), replace = TRUE, prob =
c(train_amount, test_amount, validation_amount)) train <- df[idx ==
1,] test <- df[idx == 2,] validation <- df[idx == 3,]

knn_train_fit <- e1071::tune.gknn(x = train[, 1:ncol(train) - 1], y =
trainy, scale = TRUE, k = c(1 : 25)) knn_train_fit$best_model$MSE[i] < -Metrics ::
rmse(actual = trainy, predicted = predict( object = knn_train_fit$best_model, newdata =
train[, 1 : ncol(train) - 1], k = knn_train_fit$best_model$k)) knn_train_fit$MSE_mean <
-mean(knn_train_fit$MSE) knn_test_fit$MSE[i] < -Metrics :: rmse(actual =
testy, predicted = predict( object = knn_train_fit$best_model, k = knn_train_fit$best_model$k, newdata =
test[, 1 : ncol(test) - 1])) knn_test_fit$MSE_mean < -mean(knn_test_fit$MSE) knn_validation_fit$MSE[i] <
-Metrics :: rmse(actual = validationy, predicted = predict( object =
knn_train_fit$best_model, newdata = validation[, 1 : ncol(validation) - 1], k =
knn_train_fit$best_model$k)) knn_validation_fit$MSE_mean < -mean(knn_validation_fit$MSE) knn_holdout_fit$MSE[i] <
-mean(c(knn_test_fit$MSE_mean, knn_validation_fit$MSE_mean)) knn_holdout_fit$MSE_mean <
-mean(knn_holdout_fit$MSE) knn_holdout_fit$MSE_sd_mean < -sd(c(knn_test_fit$MSE_mean, knn_validation_fit$MSE_mean)) knn_
-as.numeric(predict(object = knn_train_fit$best_model, newdata = train[,
1:ncol(train) - 1], k = knn_train_fit$best_model$k)) knn_test_predict_value
<- as.numeric(predict( object = knn_train_fit$best_model, newdata = test[, 1 :
ncol(test) - 1], k = knn_train_fit$best_model$k)) knn_validation_predict_value <
-as.numeric(predict(object = knn_train_fit$best_model, newdata = valida-
tion[, 1:ncol(test) - 1], k = knn_train_fit$best_model$k)) knn_predict_value
<- mean(c(knn_test_predict_value, knn_validation_predict_value))
knn_predict_value_mean <- mean(c(knn_test_predict_value, knn_validation_predict_value))

return(knn_holdout_RMSE_mean)

} # closing brace for numresamples } # closing brace for knn1 function
```

```
knn1(data = MASS::Boston, colnum = 14, train_amount = 0.60, test_amount  
= 0.20, validation_amount = 0.20, numresamples = 25) warnings() # no warn-  
ings for individual knn function
```

Lasso

```
{r Individual lasso model for numerical data}
```

```
library(glmnet) # So we can run the lasso model
```

Chapter 51

Set initial values to 0

```
lasso_train_RMSE <- 0 lasso_test_RMSE <- 0 lasso_validation_RMSE <- 0
lasso_holdout_RMSE <- 0 lasso_test_predict_value <- 0 lasso_validation_predict_value
<- 0 lasso_test_RMSE <- 0 lasso_test_RMSE_df <- data.frame(lasso_test_RMSE)
lasso_validation_RMSE <- 0 lasso_validation_RMSE_df <- data.frame(lasso_validation_RMSE)
lasso_holdout_RMSE <- 0 lasso_holdout_RMSE_df <- data.frame(lasso_holdout_RMSE)
```


Chapter 52

Create the function:

```
lasso <- function(data, colnum, train_amount, test_amount, valida-
tion_amount, numresamples){

# Set up random resampling for (i in 1:numresamples) {

# Changes the name of the target column to y
y <- 0
colnames(data)[colnum] <- "y"

# Moves the target column to the last column on the right
df <- data %>% dplyr::relocate(y, .after = last_col()) # Moves the target column to the last column
df <- df[sample(nrow(df)), ] # randomizes the rows

# Breaks the data into train, test and validation sets
idx <- sample(seq(1, 3), size = nrow(df), replace = TRUE, prob = c(train_amount, test_amount, validation_amount))
train <- df[idx == 1, ]
test <- df[idx == 2, ]
validation <- df[idx == 3, ]

# Set up the lasso model

y <- train$y
x <- data.matrix(train %>% dplyr::select(-y))
lasso_model <- glmnet::glmnet(x, y, alpha = 1.0)
lasso_cv <- cv.glmnet(x, y, alpha = 1.0)
best_lasso_lambda <- lasso_cv$lambda.min
best_lasso_model <- glmnet::glmnet(x, y, alpha = 0, lambda = best_lasso_lambda)
lasso_test_pred <- predict(best_lasso_model, s = best_lasso_lambda, newx = data.matrix(test %>% dplyr::select(-y)))

lasso_test_RMSE <- Metrics::rmse(actual = test$y, predicted = lasso_test_pred)
```

```

lasso_test_RMSE_df <- rbind(lasso_test_RMSE_df, lasso_test_RMSE)
lasso_test_RMSE_mean <- mean(lasso_test_RMSE_df$lasso_test_RMSE[2:nrow(lasso_test_RMSE)]

## lasso using the validation data set
y <- train$y
x <- data.matrix(train %>% dplyr::select(-y))
lasso_model <- glmnet::glmnet(x, y, alpha = 1.0)
lasso_cv <- cv.glmnet(x, y, alpha = 1.0)
best_lasso_lambda <- lasso_cv$lambda.min
best_lasso_model <- glmnet::glmnet(x, y, alpha = 0, lambda = best_lasso_lambda)
lasso_validation_pred <- predict(best_lasso_model, s = best_lasso_lambda, newx = data.matrix(validation))
lasso_validation_RMSE <- Metrics::rmse(actual = validation$y, predicted = lasso_validation_pred)
lasso_validation_RMSE_df <- rbind(lasso_validation_RMSE_df, lasso_validation_RMSE)
lasso_validation_RMSE_mean <- mean(lasso_validation_RMSE_df$lasso_validation_RMSE[2:nrow(lasso_validation_RMSE_df)])

lasso_holdout_RMSE <- mean(lasso_test_RMSE_mean, lasso_validation_RMSE_mean)
lasso_holdout_RMSE_df <- rbind(lasso_holdout_RMSE_df, lasso_holdout_RMSE)
lasso_holdout_RMSE_mean <- mean(lasso_holdout_RMSE_df$lasso_holdout_RMSE[2:nrow(lasso_holdout_RMSE_df)])

lasso_test_predict_value[i] <- round(mean(lasso_test_pred), 4)
lasso_test_predict_value_mean <- mean(lasso_test_predict_value)

lasso_validation_predict_value[i] <- round(mean(lasso_validation_pred), 4)
lasso_validation_predict_value_mean <- mean(lasso_validation_predict_value)

lasso_test_predict_value_mean <- mean(c(lasso_test_predict_value_mean, lasso_validation_predict_value_mean))

return(lasso_holdout_RMSE_mean)
} # closing brace for numresample } # closing brace for the lasso function

lasso(data = MASS::Boston, colnum = 14, train_amount = 0.60, test_amount
= 0.20, validation_amount = 0.20, numresamples = 25) warnings() # no warn-
ings for individual lasso function

Linear (tuned)

{r Individual tuned linear model for numeric data}

library(e1071) # for tuned linear models

```


Chapter 53

Set initial values to 0

```
linear_train_RMSE <- 0 linear_test_RMSE <- 0 linear_validation_RMSE  
<- 0 linear_holdout_RMSE <- 0 linear_test_predict_value <- 0 lin-  
ear_validation_predict_value <- 0  
  
linear1 <- function(data, colnum, train_amount, test_amount, valida-  
tion_amount, numresamples){
```


Chapter 54

Set up random resampling

```
for (i in 1:numresamples) {
```


Chapter 55

Changes the name of the
target column to y

```
y <- 0 colnames(data)[colnum] <- "y"
```


Chapter 56

Moves the target column to the last column on the right

```
df <- data %>% dplyr::relocate(y, .after = last_col()) # Moves the target column to the last column on the right
df <- df[sample(nrow(df)), ] # randomizes the rows
```


Chapter 57

Breaks the data into train, test and validation sets

```
idx <- sample(seq(1, 3), size = nrow(df), replace = TRUE, prob =
c(train_amount, test_amount, validation_amount)) train <- df[idx ==
1,] test <- df[idx == 2, ] validation <- df[idx == 3, ]

linear_train_fit <- e1071::tune.rpart(formula = y ~ ., data = train)
linear_train_RMSE[i] <- Metrics::rmse(actual = train$y, predicted =
predict(object = linear_train_fit$best.model, newdata = train)) linear_train_RMSE_mean
<- mean(linear_train_RMSE) linear_test_RMSE[i] <- Metrics::rmse(actual
= test$y, predicted = predict(object = linear_train_fit$best.model, new-
data = test)) linear_test_RMSE_mean <- mean(linear_test_RMSE)
linear_validation_RMSE[i] <- Metrics::rmse(actual = validation$y, predicted =
predict(object = linear_train_fit$best.model, newdata = validation)) lin-
ear_validation_RMSE_mean <- mean(linear_validation_RMSE) lin-
ear_holdout_RMSE[i] <- mean(c(linear_test_RMSE_mean, linear_validation_RMSE_mean))
linear_holdout_RMSE_mean <- mean(linear_holdout_RMSE) linear_holdout_RMSE_sd_mean
<- sd(c(linear_test_RMSE_mean, linear_validation_RMSE_mean)) lin-
ear_train_predict_value <- as.numeric(predict(object = linear_train_fit$best.model, newdata =
train)) linear_test_predict_value <- as.numeric(predict(object = linear_train_fit$best.model,
newdata = test)) linear_validation_predict_value <- as.numeric(predict(object
= linear_train_fit$best.model, newdata = validation)) linear_predict_value_mean
<- mean(c(linear_test_predict_value, linear_validation_predict_value))

return(linear_holdout_RMSE_mean)

} # closing brace for numresamples } # closing brace for linear1 function

linear1(data = MASS::Boston, colnum = 14, train_amount = 0.60,
test_amount = 0.20, validation_amount = 0.20, numresamples = 25)
warnings() # no warnings for individual lasso function
```

LQS

{r LQS models for numerical data}

library(MASS) # to allow us to run LQS models

Chapter 58

Set initial values to 0

```
lqs_train_RMSE <- 0 lqs_test_RMSE <- 0 lqs_validation_RMSE <- 0
lqs_holdout_RMSE <- 0 lqs_test_predict_value <- 0 lqs_validation_predict_value
<- 0

lqs1 <- function(data, colnum, train_amount, test_amount, valida-
tion_amount, numresamples){
```


Chapter 59

Set up random resampling

```
for (i in 1:numresamples) {
```


Chapter 60

Changes the name of the
target column to y

```
y <- 0 colnames(data)[colnum] <- "y"
```


Chapter 61

Moves the target column to the last column on the right

```
df <- data %>% dplyr::relocate(y, .after = last_col()) # Moves the target column to the last column on the right
df <- df[sample(nrow(df)), ] # randomizes the rows
```


Chapter 62

Breaks the data into train, test and validation sets

```
idx <- sample(seq(1, 3), size = nrow(df), replace = TRUE, prob =
c(train_amount, test_amount, validation_amount)) train <- df[idx ==
1,] test <- df[idx == 2,] validation <- df[idx == 3,]

lqs_train_fit <- MASS::lqs(trainy ~., data = train) lqs_train_RMSE[i] <-
- Metrics::rmse(actual = trainy, predicted = predict(object = lqs_train_fit,
newdata = train)) lqs_train_RMSE_mean <- mean(lqs_train_RMSE)
lqs_test_RMSE[i] <- Metrics::rmse(actual = testy, predicted = predict(object =
lqs_train_fit, newdata = test)) lqs_test_RMSE_mean <- mean(lqs_test_RMSE)
lqs_validation_RMSE[i] <- Metrics::rmse(actual = validationy, predicted = predict(object =
lqs_train_fit, newdata = validation)) lqs_validation_RMSE_mean <-
mean(lqs_validation_RMSE) lqs_holdout_RMSE[i] <- mean(c(lqs_test_RMSE_mean,
lqs_validation_RMSE_mean)) lqs_holdout_RMSE_mean <- mean(lqs_holdout_RMSE)
lqs_train_predict_value <- as.numeric(predict(object = lqs_train_fit,
newdata = train)) lqs_test_predict_value <- as.numeric(predict(object
= lqs_train_fit, newdata = test)) lqs_validation_predict_value <-
as.numeric(predict(object = lqs_train_fit, newdata = validation))

y_hat_lqs <- c(lqs_test_predict_value, lqs_validation_predict_value)

return(lqs_holdout_RMSE_mean)

} # Closing brace for numresamples } # Closing brace for lqs1 function

lqs1(data = MASS::Boston, colnum = 14, train_amount = 0.60, test_amount =
0.20, validation_amount = 0.20, numresamples = 25) warnings() # no warnings
for individual lqs function
```

Neuralnet

```
{r Neuralnet individual model for numerical data}
```

```
library(neuralnet)
```

Chapter 63

Set initial values to 0

```
neuralnet_train_RMSE <- 0 neuralnet_test_RMSE <- 0 neuralnet_validation_RMSE  
<- 0 neuralnet_holdout_RMSE <- 0 neuralnet_test_predict_value <- 0 neu-  
ralnet_validation_predict_value <- 0  
  
neuralnet1 <- function(data, colnum, train_amount, test_amount, valida-  
tion_amount, numresamples){
```


Chapter 64

Set up random resampling

```
for (i in 1:numresamples) {
```


Chapter 65

Changes the name of the
target column to y

```
y <- 0 colnames(data)[colnum] <- "y"
```


Chapter 66

Moves the target column to the last column on the right

```
df <- data %>% dplyr::relocate(y, .after = last_col()) # Moves the target column to the last column on the right
df <- df[sample(nrow(df)), ] # randomizes the rows
```


Chapter 67

Breaks the data into train, test and validation sets

```
idx <- sample(seq(1, 3), size = nrow(df), replace = TRUE, prob =
c(train_amount, test_amount, validation_amount)) train <- df[idx ==
1,] test <- df[idx == 2, ] validation <- df[idx == 3, ]

maxs <- apply(df, 2, max) mins <- apply(df, 2, min) scaled <- as.data.frame(scale(df,
center = mins, scale = maxs - mins)) train_ <- scaled[idx == 1, ]
test_ <- scaled[idx == 2, ] validation_ <- scaled[idx == 3, ] n <-
names(train_) f <- as.formula(paste("y ~", paste(n[!n %in% "y"], col-
lapse = " + ")) nn <- neuralnet(f, data = train_, hidden = c(5, 3),
linear.output = TRUE) predict_test_nn <- neuralnet::compute(nn, test_[,
1:ncol(df) - 1]) predict_test_nn_ <- predict_test_nnnet.result * (max(dfy)
- min(dfy)) + min(dfy) predict_train_nn <- neuralnet::compute(nn, train_[,
1:ncol(df) - 1]) predict_train_nn_ <- predict_train_nnnet.result * (max(dfy)
- min(dfy)) + min(dfy) neuralnet_train_RMSE[i] <- Metrics::rmse(actual
= trainy, predicted = predict_train_nn_) neuralnet_test_RMSE_mean <-
mean(neuralnet_test_RMSE) predict_validation_nn <- neuralnet::compute(nn,
validation_[, 1:ncol(df) - 1]) predict_validation_nn_ <- predict_validation_nnnet.result *
(max(dfy) - min(dfy)) + min(dfy) neuralnet_validation_RMSE[i] <- Met-
rics::rmse(actual = validation$y, predicted = predict_validation_nn_)
neuralnet_validation_RMSE_mean <- mean(neuralnet_validation_RMSE)

neuralnet_holdout_RMSE[i] <- mean(c(neuralnet_test_RMSE, neural-
net_validation_RMSE)) neuralnet_holdout_RMSE_mean <- mean(neuralnet_holdout_RMSE)

return(neuralnet_holdout_RMSE_mean)

} # Closing brace for numresamples } # closing brace for neuralnet1 function
```

```
neuralnet1(data = MASS::Boston, colnum = 14, train_amount = 0.60,  
test_amount = 0.20, validation_amount = 0.20, numresamples = 25)  
warnings() # no warnings for individual neuralnet function
```

Partial Least Squares

```
{r Partial least squares models for numerical data}
```

```
library(pls)
```

Chapter 68

Set initial values to 0

```
pls_train_RMSE <- 0 pls_test_RMSE <- 0 pls_validation_RMSE <- 0
pls_holdout_RMSE <- 0 pls_test_predict_value <- 0 pls_validation_predict_value
<- 0

pls1 <- function(data, colnum, train_amount, test_amount, valida-
tion_amount, numresamples){
```


Chapter 69

Set up random resampling

```
for (i in 1:numresamples) {
```


Chapter 70

Changes the name of the
target column to y

```
y <- 0 colnames(data)[colnum] <- "y"
```


Chapter 71

Moves the target column to the last column on the right

```
df <- data %>% dplyr::relocate(y, .after = last_col()) # Moves the target column to the last column on the right
df <- df[sample(nrow(df)), ] # randomizes the rows
```


Chapter 72

Breaks the data into train, test and validation sets

```
idx <- sample(seq(1, 3), size = nrow(df), replace = TRUE, prob =
c(train_amount, test_amount, validation_amount)) train <- df[idx ==
1,] test <- df[idx == 2,] validation <- df[idx == 3,]

pls_train_fit <- pls::plsr(trainy ~, data = train) pls_train_RMSE[i] <-
-Metrics::rmse(actual = trainy, predicted = predict(object = pls_train_fit,
newdata = train)) pls_train_RMSE_mean <- mean(pls_train_RMSE)
pls_test_RMSE[i] <- Metrics::rmse(actual = testy, predicted = predict(object =
pls_train_fit, newdata = test)) pls_test_RMSE_mean <- mean(pls_test_RMSE)
pls_validation_RMSE[i] <- Metrics::rmse(actual = validationy, predicted = predict(object =
pls_train_fit, newdata = validation)) pls_validation_RMSE_mean <-
mean(pls_validation_RMSE) pls_holdout_RMSE[i] <- mean(c(pls_test_RMSE_mean,
pls_validation_RMSE_mean)) pls_holdout_RMSE_mean <- mean(pls_holdout_RMSE)
pls_holdout_RMSE_sd_mean <- sd(c(pls_test_RMSE_mean, pls_validation_RMSE_mean))
pls_train_predict_value <- predict(object = pls_train_fit, newdata = train)
pls_test_predict_value <- predict(object = pls_train_fit, newdata = test)
pls_validation_predict_value <- predict(object = pls_train_fit, newdata
= validation) pls_predict_value_mean <- mean(c(pls_test_predict_value,
pls_validation_predict_value))

return( pls_holdout_RMSE_mean)

} # Closing brace for numresamples loop } # Closing brace for pls1 function

pls1(data = MASS::Boston, colnum = 14, train_amount = 0.60, test_amount =
0.20, validation_amount = 0.20, numresamples = 25) warnings() # no warnings
for individual pls function
```

Principal Components Regression

```
{r Principal Components Regression models for numerical data}
```

```
library(pls) # To run pcr models
```


Chapter 73

Set initial values to 0

```
pcr_train_RMSE <- 0 pcr_test_RMSE <- 0 pcr_validation_RMSE <- 0
pcr_holdout_RMSE <- 0 pcr_test_predict_value <- 0 pcr_validation_predict_value
<- 0

pcr1 <- function(data, colnum, train_amount, test_amount, valida-
tion_amount, numresamples){
```


Chapter 74

Set up random resampling

```
for (i in 1:numresamples) {
```


Chapter 75

Changes the name of the
target column to y

```
y <- 0 colnames(data)[colnum] <- "y"
```


Chapter 76

Moves the target column to the last column on the right

```
df <- data %>% dplyr::relocate(y, .after = last_col()) # Moves the target column to the last column on the right
df <- df[sample(nrow(df)), ] # randomizes the rows
```


Chapter 77

Breaks the data into train, test and validation sets

```
idx <- sample(seq(1, 3), size = nrow(df), replace = TRUE, prob =
c(train_amount, test_amount, validation_amount)) train <- df[idx ==
1,] test <- df[idx == 2, ] validation <- df[idx == 3, ]

pcr_train_fit <- pls::pcr(trainy ~., data = train) pcr_train_RMSE[i] <-
- Metrics::rmse(actual = trainy, predicted = predict(object = pcr_train_fit,
newdata = train)) pcr_train_RMSE_mean <- mean(pcr_train_RMSE)
pcr_test_RMSE[i] <- Metrics::rmse(actual = testy, predicted = predict(object =
pcr_train_fit, newdata = test)) pcr_test_RMSE_mean <- mean(pcr_test_RMSE) pcr_validation_RMSE[i] <-
- Metrics::rmse(actual = validationy, predicted = predict(object =
pcr_train_fit, newdata = validation)) pcr_validation_RMSE_mean <-
mean(pcr_validation_RMSE) pcr_holdout_RMSE[i] <- mean(pcr_test_RMSE_mean,
pcr_validation_RMSE_mean) pcr_holdout_RMSE_mean <- mean(pcr_holdout_RMSE)
pcr_holdout_RMSE_sd_mean <- sd(c(pcr_test_RMSE_mean, pcr_validation_RMSE_mean))
pcr_train_predict_value <- predict(object = pcr_train_fit, newdata = train)
pcr_test_predict_value <- predict(object = pcr_train_fit, newdata =
test) pcr_validation_predict_value <- predict(object = pcr_train_fit,
newdata = validation) y_hat_pcr <- c(pcr_test_predict_value[, , 1],
pcr_validation_predict_value[, , 1])

return(pcr_holdout_RMSE_mean) } # Closing brace for numresamples loop
} # Closing brace for PCR function

pcr1(data = MASS::Boston, colnum = 14, train_amount = 0.60, test_amount
= 0.20, validation_amount = 0.20, numresamples = 25) warnings() # no warn-
ings for individual pls function
```

Random Forest

{r Random forest, individual model for numerical data}

library(randomForest)

Chapter 78

Set initial values to 0

```
rf_train_RMSE <- 0 rf_test_RMSE <- 0 rf_validation_RMSE <- 0
rf_holdout_RMSE <- 0 rf_test_predict_value <- 0 rf_validation_predict_value
<- 0

rf1 <- function(data, colnum, train_amount, test_amount, validation_amount,
numresamples){
```


Chapter 79

Set up random resampling

```
for (i in 1:numresamples) {
```


Chapter 80

Changes the name of the
target column to y

```
y <- 0 colnames(data)[colnum] <- "y"
```


Chapter 81

Moves the target column to the last column on the right

```
df <- data %>% dplyr::relocate(y, .after = last_col()) # Moves the target column to the last column on the right
df <- df[sample(nrow(df)), ] # randomizes the rows
```


Chapter 82

Breaks the data into train, test and validation sets

```
idx <- sample(seq(1, 3), size = nrow(df), replace = TRUE, prob =
c(train_amount, test_amount, validation_amount)) train <- df[idx ==
1,] test <- df[idx == 2,] validation <- df[idx == 3,]

rf_train_fit <- tune.randomForest(x = train, y = trainy, data = train)rf_train_fit$MSE[i] <
-Metrics::rmse(actual = trainy, predicted = predict(object = rf_train_fit$best.model, newdata =
train))rf_train_fit$MSE_mean < -mean(rf_train_fit$MSE)rf_test_fit$MSE[i] <
-Metrics::rmse(actual = testy, predicted = predict(object = rf_train_fit$best.model, newdata =
test))rf_test_fit$MSE_mean < -mean(rf_test_fit$MSE)rf_validation_fit$MSE[i] <
-Metrics::rmse(actual = validationy, predicted = predict(object =
rf_train_fit$best.model, newdata = validation))rf_validation_fit$MSE_mean <
-mean(rf_validation_fit$MSE)rf_holdout_fit$MSE[i] < -mean(c(rf_test_fit$MSE_mean, rf_validation_fit$MSE_mean))rf_holdout_fit$MSE_mean <
-mean(rf_holdout_fit$MSE)rf_holdout_fit$MSE_sd_mean < -sd(c(rf_test_fit$MSE_mean, rf_validation_fit$MSE_mean))rf_train_predict_value
<- predict(object = rf_train_fit$best.model, newdata = train)rf_test_predict_value
<- predict(object = rf_train_fit$best.model, newdata = test)rf_validation_predict_value <-
predict(object = rf_train_fit$best.model, newdata = validation)y_hat_rf <-
c(rf_test_predict_value, rf_validation_predict_value)

return(rf_holdout_RMSE_mean)

} # Closing brace for numresamples loop } # Closing brace for rf1 function

rf1(data = MASS::Boston, colnum = 14, train_amount = 0.60, test_amount =
0.20, validation_amount = 0.20, numresamples = 25) warnings() # no warnings
for individual random forest function

Ridge Regression

{r Individual ridge model for numerical data}
```

```
library(glmnet) # So we can run the ridge model
```

Chapter 83

Set initial values to 0

```
ridge_train_RMSE <- 0 ridge_test_RMSE <- 0 ridge_validation_RMSE <- 0
ridge_holdout_RMSE <- 0 ridge_test_predict_value <- 0 ridge_validation_predict_value
<- 0 ridge_test_RMSE <- 0 ridge_test_RMSE_df <- data.frame(ridge_test_RMSE)
ridge_validation_RMSE <- 0 ridge_validation_RMSE_df <- data.frame(ridge_validation_RMSE)
ridge_holdout_RMSE <- 0 ridge_holdout_RMSE_df <- data.frame(ridge_holdout_RMSE)
```


Chapter 84

Create the function:

```
ridge1 <- function(data, colnum, train_amount, test_amount, valida-
tion_amount, numresamples){

# Set up random resampling for (i in 1:numresamples) {

# Changes the name of the target column to y
y <- 0
colnames(data)[colnum] <- "y"

# Moves the target column to the last column on the right
df <- data %>% dplyr::relocate(y, .after = last_col()) # Moves the target column to the last column
df <- df[sample(nrow(df)), ] # randomizes the rows

# Breaks the data into train, test and validation sets
idx <- sample(seq(1, 3), size = nrow(df), replace = TRUE, prob = c(train_amount, test_amount, validation_amount))
train <- df[idx == 1, ]
test <- df[idx == 2, ]
validation <- df[idx == 3, ]

# Set up the ridge model

y <- train$y
x <- data.matrix(train %>% dplyr::select(-y))
ridge_model <- glmnet::glmnet(x, y, alpha = 0)
ridge_cv <- cv.glmnet(x, y, alpha = 0)
best_ridge_lambda <- ridge_cv$lambda.min
best_ridge_model <- glmnet::glmnet(x, y, alpha = 0, lambda = best_ridge_lambda)
ridge_test_pred <- predict(best_ridge_model, s = best_ridge_lambda, newx = data.matrix(test %>% dplyr::select(-y)))

ridge_test_RMSE <- Metrics::rmse(actual = test$y, predicted = ridge_test_pred)
```

```

ridge_test_RMSE_df <- rbind(ridge_test_RMSE_df, ridge_test_RMSE)
ridge_test_RMSE_mean <- mean(ridge_test_RMSE_df$ridge_test_RMSE[2:nrow(ridge_test_RMSE_df)])

## ridge using the validation data set
y <- train$y
x <- data.matrix(train %>% dplyr::select(-y))
ridge_model <- glmnet::glmnet(x, y, alpha = 0)
ridge_cv <- cv.glmnet(x, y, alpha = 0)
best_ridge_lambda <- ridge_cv$lambda.min
best_ridge_model <- glmnet::glmnet(x, y, alpha = 0, lambda = best_ridge_lambda)
ridge_validation_pred <- predict(best_ridge_model, s = best_ridge_lambda, newx = data.matrix(validation))
ridge_validation_RMSE <- Metrics::rmse(actual = validation$y, predicted = ridge_validation_pred)
ridge_validation_RMSE_df <- rbind(ridge_validation_RMSE_df, ridge_validation_RMSE)
ridge_validation_RMSE_mean <- mean(ridge_validation_RMSE_df$ridge_validation_RMSE[2:nrow(ridge_validation_RMSE_df)])

ridge_holdout_RMSE <- mean(ridge_test_RMSE_mean, ridge_validation_RMSE_mean)
ridge_holdout_RMSE_df <- rbind(ridge_holdout_RMSE_df, ridge_holdout_RMSE)
ridge_holdout_RMSE_mean <- mean(ridge_holdout_RMSE_df$ridge_holdout_RMSE[2:nrow(ridge_holdout_RMSE_df)])

ridge_test_predict_value[i] <- round(mean(ridge_test_pred), 4)
ridge_test_predict_value_mean <- mean(ridge_test_predict_value)

ridge_validation_predict_value[i] <- round(mean(ridge_validation_pred), 4)
ridge_validation_predict_value_mean <- mean(ridge_validation_predict_value)

ridge_test_predict_value_mean <- mean(c(ridge_test_predict_value_mean, ridge_validation_predict_value_mean))

return(ridge_holdout_RMSE_mean)
} # closing brace for numresample } # closing brace for the ridge function

ridge1(data = MASS::Boston, colnum = 14, train_amount = 0.60, test_amount
= 0.20, validation_amount = 0.20, numresamples = 25) warnings() # no warnings
for individual ridge function

Robust Regression

{r}

library(MASS) # To run rlm function for robust regression

```


Chapter 85

Set initial values to 0

```
robust_train_RMSE <- 0 robust_test_RMSE <- 0 robust_validation_RMSE  
<- 0 robust_holdout_RMSE <- 0 robust_test_predict_value <- 0 ro-  
bust_validation_predict_value <- 0  
robust1 <- function(data, colnum, train_amount, test_amount, valida-  
tion_amount, numresamples){
```


Chapter 86

Set up random resampling

```
for (i in 1:numresamples) {
```


Chapter 87

Changes the name of the
target column to y

```
y <- 0 colnames(data)[colnum] <- "y"
```


Chapter 88

Moves the target column to the last column on the right

```
df <- data %>% dplyr::relocate(y, .after = last_col()) # Moves the target column to the last column on the right
df <- df[sample(nrow(df)), ] # randomizes the rows
```


Chapter 89

Breaks the data into train, test and validation sets

```
idx <- sample(seq(1, 3), size = nrow(df), replace = TRUE, prob =
c(train_amount, test_amount, validation_amount)) train <- df[idx ==
1,] test <- df[idx == 2,] validation <- df[idx == 3,]

robust_train_fit <- MASS::rlm(x = train[, 1:ncol(df) - 1], y = train$y)robust_train_RMSE[i] <-
-Metrics::rmse(actual = train$y, predicted = robust_train_fit$fitted.values)robust_train_RMSE_mean <-
-mean(robust_train_RMSE)robust_test_RMSE[i] <- -Metrics::rmse(actual =
test$y, predicted = predict(object = MASS::rlm(y ~ ., data = train), new-
data = test)) robust_test_RMSE_mean <- mean(robust_test_RMSE)
robust_validation_RMSE[i] <- Metrics::rmse(actual = validation$y, predicted
= predict( object = MASS::rlm(y ~ ., data = train), newdata = validation
)) robust_validation_RMSE_mean <- mean(robust_validation_RMSE)
robust_holdout_RMSE[i] <- mean(c(robust_test_RMSE_mean, ro-
bust_validation_RMSE_mean)) robust_holdout_RMSE_mean <- mean(robust_holdout_RMSE)

robust_train_predict_value <- as.numeric(predict(object = MASS::rlm(y
~ ., data = train), newdata = train)) robust_test_predict_value <-
as.numeric(predict(object = MASS::rlm(y ~ ., data = train), newdata = test))
robust_validation_predict_value <- as.numeric(predict(object = MASS::rlm(y
~ ., data = train), newdata = validation)) robust_predict_value_mean <-
mean(c(robust_test_predict_value, robust_validation_predict_value))
y_hat_robust <- c(robust_test_predict_value, robust_validation_predict_value)

return(robust_holdout_RMSE_mean)

} # Closing brace for numresamples loop } # Closing brace for robust1 function

robust1(data = MASS::Boston, colnum = 14, train_amount = 0.60,
test_amount = 0.20, validation_amount = 0.20, numresamples = 25)
```

```
warnings() # no warnings for individual robust function
Rpart
{r Individual model using Rpart}
library(rpart)
```

Chapter 90

Set initial values to 0

```
rpart_train_RMSE <- 0 rpart_test_RMSE <- 0 rpart_validation_RMSE <- 0
rpart_holdout_RMSE <- 0 rpart_test_predict_value <- 0 rpart_validation_predict_value
<- 0

rpart1 <- function(data, colnum, train_amount, test_amount, valida-
tion_amount, numresamples){
```


Chapter 91

Set up random resampling

```
for (i in 1:numresamples) {
```


Chapter 92

Changes the name of the
target column to y

```
y <- 0 colnames(data)[colnum] <- "y"
```


Chapter 93

Moves the target column to the last column on the right

```
df <- data %>% dplyr::relocate(y, .after = last_col()) # Moves the target column to the last column on the right
df <- df[sample(nrow(df)), ] # randomizes the rows
```


Chapter 94

Breaks the data into train, test and validation sets

```
idx <- sample(seq(1, 3), size = nrow(df), replace = TRUE, prob =
c(train_amount, test_amount, validation_amount)) train <- df[idx ==
1,] test <- df[idx == 2,] validation <- df[idx == 3,]

rpart_train_fit <- rpart::rpart(trainy ~., data = train) rpart_train_RMSE[i] <-
- Metrics::rmse(actual = trainy, predicted = predict(object = rpart_train_fit,
newdata = train)) rpart_train_RMSE_mean <- mean(rpart_train_RMSE)
rpart_test_RMSE[i] <- Metrics::rmse(actual = testy, predicted = predict(object =
rpart_train_fit, newdata = test)) rpart_test_RMSE_mean <- mean(rpart_test_RMSE) rpart_validation_RMSE[i] <-
- Metrics::rmse(actual = validationy, predicted = predict(object =
rpart_train_fit, newdata = validation)) rpart_validation_RMSE_mean <-
mean(rpart_validation_RMSE) rpart_holdout_RMSE[i] <- mean(c(rpart_test_RMSE_mean,
rpart_validation_RMSE_mean)) rpart_holdout_RMSE_mean <- mean(rpart_holdout_RMSE)
rpart_holdout_RMSE_sd_mean <- sd(c(rpart_test_RMSE_mean, rpart_validation_RMSE_mean))
rpart_train_predict_value <- as.numeric(predict(object = rpart::rpart(y
~ ., data = train), newdata = train)) rpart_test_predict_value <-
as.numeric(predict(object = rpart::rpart(y ~ ., data = train), newdata
= test)) rpart_validation_predict_value <- as.numeric(predict(object =
rpart::rpart(y ~ ., data = train), newdata = validation)) y_hat_rpart <-
c(rpart_test_predict_value, rpart_validation_predict_value)

return(rpart_holdout_RMSE_mean)

} # Closing loop for numresamples } # Closing brace for rpart1 function

rpart1(data = MASS::Boston, colnum = 14, train_amount = 0.60, test_amount
= 0.20, validation_amount = 0.20, numresamples = 25) warnings() # no warn-
ings for individual rpart function
```

Support Vector Machines

{r Individual models using Support Vector Machines}

library(e1071)

Chapter 95

Set initial values to 0

Chapter 96

Set initial values to 0

```
svm_train_RMSE <- 0 svm_test_RMSE <- 0 svm_validation_RMSE <- 0
svm_holdout_RMSE <- 0 svm_test_predict_value <- 0 svm_validation_predict_value
<- 0

svm1 <- function(data, colnum, train_amount, test_amount, valida-
tion_amount, numresamples){
```


Chapter 97

Set up random resampling

```
for (i in 1:numresamples) {
```


Chapter 98

Changes the name of the
target column to y

```
y <- 0 colnames(data)[colnum] <- "y"
```


Chapter 99

Moves the target column to the last column on the right

```
df <- data %>% dplyr::relocate(y, .after = last_col()) # Moves the target column to the last column on the right
df <- df[sample(nrow(df)), ] # randomizes the rows
```


Chapter 100

Breaks the data into train, test and validation sets

```
idx <- sample(seq(1, 3), size = nrow(df), replace = TRUE, prob =
c(train_amount, test_amount, validation_amount)) train <- df[idx ==
1,] test <- df[idx == 2,] validation <- df[idx == 3,]

svm_train_fit <- e1071::tune.svm(x = train, y = trainy, data = train)svm_train_fit$MSE[i] <
-Metrics::rmse(actual = trainy, predicted = predict(object = svm_train_fit$best.model, newdata =
train))svm_train_fit$MSE_mean < -mean(svm_train_fit$MSE)svm_test_fit$MSE[i] <
-Metrics::rmse(actual = testy, predicted = predict(object = svm_train_fit$best.model, newdata =
test))svm_test_fit$MSE_mean < -mean(svm_test_fit$MSE)svm_validation_fit$MSE[i] <
-Metrics::rmse(actual = validationy, predicted = predict(object =
svm_train_fit$best.model, newdata = validation))svm_validation_fit$MSE_mean <
-mean(svm_validation_fit$MSE)svm_holdout_fit$MSE[i] < -mean(c(svm_test_fit$MSE_mean, svm_validation_fit$MSE_mean))
-mean(svm_holdout_fit$MSE)svm_holdout_fit$MSE_sd_mean < -sd(svm_validation_fit$MSE)svm_train_fit$predict_value <
-as.numeric(predict(object = svm_train_fit$best.model, newdata = train))
svm_test_fit$predict_value <- as.numeric(predict(object = svm_train_fit$best.model, newdata =
test))svm_validation_fit$predict_value < -as.numeric(predict(object = svm_train_fit$best.model,
newdata = validation)) svm_predict_value_mean <- mean(c(svm_test_fit$predict_value,
svm_validation_fit$predict_value)) y_hat_svm <- c(svm_test_fit$predict_value,
svm_validation_fit$predict_value)

return(svm_holdout_fit$RMSE_mean)

} # Closing brace for numresamples loop } # Closing brace for svm1 function

svm1(data = MASS::Boston, colnum = 14, train_amount = 0.60, test_amount
= 0.20, validation_amount = 0.20, numresamples = 25) warnings() # no warn-
ings for individual Support Vector Machines function
```

Trees

```
{r Individual models using trees}
```

```
library(tree)
```


Chapter 101

Set initial values to 0

```
tree_train_RMSE <- 0 tree_test_RMSE <- 0 tree_validation_RMSE <- 0
tree_holdout_RMSE <- 0 tree_test_predict_value <- 0 tree_validation_predict_value
<- 0

tree1 <- function(data, colnum, train_amount, test_amount, valida-
tion_amount, numresamples){
```


Chapter 102

Set up random resampling

```
for (i in 1:numresamples) {
```


Chapter 103

Changes the name of the
target column to y

```
y <- 0 colnames(data)[colnum] <- "y"
```


Chapter 104

Moves the target column to the last column on the right

```
df <- data %>% dplyr::relocate(y, .after = last_col()) # Moves the target column to the last column on the right
df <- df[sample(nrow(df)), ] # randomizes the rows
```


Chapter 105

Breaks the data into train, test and validation sets

```
idx <- sample(seq(1, 3), size = nrow(df), replace = TRUE, prob =
c(train_amount, test_amount, validation_amount)) train <- df[idx ==
1,] test <- df[idx == 2,] validation <- df[idx == 3,]

tree_train_fit <- tree::tree(trainy ~., data = train) tree_train_RMSE[i] <-
- Metrics::rmse(actual = trainy, predicted = predict(object = tree_train_fit,
newdata = train)) tree_train_RMSE_mean <- mean(tree_train_RMSE)
tree_test_RMSE[i] <- Metrics::rmse(actual = testy, predicted = predict(object =
tree_train_fit, newdata = test)) tree_test_RMSE_mean <- mean(tree_test_RMSE)
tree_validation_RMSE[i] <- Metrics::rmse(actual = validationy, predicted = predict(object =
tree_train_fit, newdata = validation)) tree_validation_RMSE_mean <-
mean(tree_validation_RMSE) tree_holdout_RMSE[i] <- mean(c(tree_test_RMSE_mean,
tree_validation_RMSE_mean)) tree_holdout_RMSE_mean <- mean(tree_holdout_RMSE)
tree_holdout_RMSE_sd_mean <- sd(c(tree_test_RMSE_mean, tree_validation_RMSE_mean))
tree_train_predict_value <- as.numeric(predict(object = tree::tree(y ~ ., data
= train), newdata = train)) tree_test_predict_value <- as.numeric(predict(object
= tree::tree(y ~ ., data = train), newdata = test)) tree_validation_predict_value
<- as.numeric(predict(object = tree::tree(y ~ ., data = train), newdata = valida-
tion)) y_hat_tree <- c(tree_test_predict_value, tree_validation_predict_value)

return(tree_holdout_RMSE_mean)

} # Closing brace for numresamples loop } # Closing brace for tree1 function

tree1(data = MASS::Boston, colnum = 14, train_amount = 0.60, test_amount
= 0.20, validation_amount = 0.20, numresamples = 25) warnings() # no warn-
ings for individual tree function

XGBoost
```

```
{r Individual XGBoost model}
```

```
library(xgboost)
```

Chapter 106

Set initial values to 0

```
xgb_train_RMSE <- 0 xgb_test_RMSE <- 0 xgb_validation_RMSE <- 0
xgb_holdout_RMSE <- 0 xgb_test_predict_value <- 0 xgb_validation_predict_value
<- 0

xgb1 <- function(data, colnum, train_amount, test_amount, valida-
tion_amount, numresamples){
```


Chapter 107

Set up random resampling

```
for (i in 1:numresamples) {
```


Chapter 108

Changes the name of the
target column to y

```
y <- 0 colnames(data)[colnum] <- "y"
```


Chapter 109

Moves the target column to the last column on the right

```
df <- data %>% dplyr::relocate(y, .after = last_col()) # Moves the target column to the last column on the right
df <- df[sample(nrow(df)), ] # randomizes the rows
```


Chapter 110

Breaks the data into train, test and validation sets

```
idx <- sample(seq(1, 3), size = nrow(df), replace = TRUE, prob =  
c(train_amount, test_amount, validation_amount)) train <- df[idx ==  
1,] test <- df[idx == 2, ] validation <- df[idx == 3, ]  
train_x <- data.matrix(train[, -ncol(train)]) train_y <- train[, ncol(train)]
```


Chapter 111

define predictor and response variables in test set

```
test_x <- data.matrix(test[, -ncol(test)]) test_y <- test[, ncol(test)]
```


Chapter 112

define predictor and response variables in validation set

```
validation_x <- data.matrix(validation[, -ncol(validation)]) validation_y <- val-  
idation[, ncol(validation)]
```


Chapter 113

define final train, test and validation sets

```
xgb_train <- xgboost::xgb.DMatrix(data = train_x, label = train_y) xgb_test  
<- xgboost::xgb.DMatrix(data = test_x, label = test_y) xgb_validation <-  
xgboost::xgb.DMatrix(data = validation_x, label = validation_y)
```


Chapter 114

define watchlist

```
watchlist <- list(train = xgb_train, validation = xgb_validation) watchlist_test  
<- list(train = xgb_train, test = xgb_test) watchlist_validation <- list(train  
= xgb_train, validation = xgb_validation)
```


Chapter 115

fit XGBoost model and display training and validation data at each round

```
xgb_model <- xgboost::xgb.train(data = xgb_train, max.depth = 3,
watchlist = watchlist_test, nrounds = 70) xgb_model_validation <-
xgboost::xgb.train(data = xgb_train, max.depth = 3, watchlist = watch-
list_validation, nrounds = 70)

xgboost_min <- which.min(xgb_model$evaluation_log$validation_rmse) xg-
boost_validation_min <- which.min(xgb_model$evaluation_log$validation_rmse)

xgb_train_RMSE[i] <- Metrics::rmse(actual = train_y, predicted = predict(object =
xgb_model, newdata = train_x)) xgb_train_RMSE_mean <- mean(xgb_train_RMSE) xgb_test_RMSE[i] <-
Metrics::rmse(actual = test_y, predicted = predict(object = xgb_model,
newdata = test_x)) xgb_test_RMSE_mean <- mean(xgb_test_RMSE)
xgb_validation_RMSE[i] <- round(Metrics::rmse(actual = validation$y,
predicted = predict(object = xgb_model, newdata = validation_x)), 4)
xgb_validation_RMSE_mean <- mean(xgb_validation_RMSE)

xgb_holdout_RMSE[i] <- mean(xgb_test_RMSE_mean, xgb_validation_RMSE_mean)
print(xgb_holdout_RMSE) xgb_holdout_RMSE_mean <- mean(xgb_holdout_RMSE)
xgb_holdout_RMSE_sd_mean <- sd(c(xgb_test_RMSE_mean, xgb_validation_RMSE_mean))

y_hat_xgb <- c(predict(object = xgb_model, newdata = test_x), pre-
dict(object = xgb_model, newdata = validation_x))

return(xgb_holdout_RMSE_mean)
```

```
} # Closing brace for numresamples loop } # Closing brace for xgb1 function  
xgb1(data = MASS::Boston, colnum = 14, train_amount = 0.60, test_amount  
= 0.20, validation_amount = 0.20, numresamples = 2500) warnings() # no  
warnings for individual XGBoost function
```

Chapter 116

Building weighted ensembles to model numerical data

In the last chapter we learned how to make 23 individual models, including calculating the error rate (as root mean squared error), and predictions from the holdout data (test and validation).

This chapter will show how to use those results to make weighted ensembles that can be used to model numerical data.

Let's start at the end. Let's imagine the finished product. A list of ensembles and individual models, with error rates sorted in decreasing order (best result on the top of the list)

Therefore we're going to need a weighted ensemble. But what weight to use?

It turns out there is an excellent answer available for virtually no work on our part (which is great!). Each model has a mean error score. The weight we will use will be the reciprocal of the error.

Let's say we have two models. One has an error rate of 5.0 and the other has an error rate of 2.0. Clearly the model with the error rate of 2.0 is superior to the model with the error rate of 5.0.

What we will do in building our ensemble is multiply the values in the ensemble by $1/(\text{error rate})$. This will give higher weights to models with higher accuracy.

Let's see how this works with an extremely simple ensemble.

```
library(tree) # Allows us to use tree models
library(MASS) # For the Boston Housing data set library(Metrics)
```

```

library(tidyverse)
#> -- Attaching core tidyverse packages ---- tidyverse 2.0.0 --
#> v dplyr      1.1.4      v readr      2.1.5
#> v forcats    1.0.0      v stringr   1.5.1
#> v ggplot2    3.5.1      v tibble    3.2.1
#> v lubridate  1.9.3      v tidyr     1.3.1
#> v purrr      1.0.2
#> -- Conflicts ----- tidyverse_conflicts() --
#> x dplyr::filter() masks stats::filter()
#> x dplyr::lag()     masks stats::lag()
#> x dplyr::select() masks MASS::select()
#> i Use the conflicted package (<http://conflicted.r-lib.org/>) to force all conflicts.

# Set initial values to 0
linear_train_RMSE <- 0
linear_test_RMSE <- 0
linear_RMSE <- 0
linear_test_predict_value <- 0

tree_train_RMSE <- 0
tree_test_RMSE <- 0
tree_RMSE <- 0
tree_holdout_RMSE <- 0
tree_test_predict_value <- 0

ensemble_linear_RMSE <- 0
ensemble_linear_RMSE_mean <- 0
ensemble_tree_RMSE <- 0
ensemble_tree_RMSE_mean <- 0

numerical_1 <- function(data, colnum, train_amount, test_amount, numresamples){

# Move target column to far right
y <- 0
colnames(data)[colnum] <- "y"

# Set up resampling
for (i in 1:numresamples) {
  idx <- sample(seq(1, 2), size = nrow(data), replace = TRUE, prob = c(train_amount, test_amount))
  train <- data[idx == 1, ]
  test <- data[idx == 2, ]

# Fit linear model on the training data, make predictions on the test data
linear_model <- lm(y ~ ., data = train)
linear_predictions <- predict(object = linear_model, newdata = test)

```



```

linear_RMSE[i] <- Metrics::rmse(actual = test$y, predicted = linear_predictions)
linear_RMSE_mean <- mean(linear_RMSE)

# Fit tree model on the training data, make predictions on the test data
tree_model <- tree(y ~ ., data = train)
tree_predictions <- predict(object = tree_model, newdata = test)
tree_RMSE[i] <- Metrics::rmse(actual = test$y, predicted = tree_predictions)
tree_RMSE_mean <- mean(tree_RMSE)

# Make the weighted ensemble
ensemble <- data.frame(
  'linear' = linear_predictions / linear_RMSE_mean,
  'tree' = tree_predictions / tree_RMSE_mean,
  'y_ensemble' = test$y)

# Split ensemble between train and test
ensemble_idx <- sample(seq(1, 2), size = nrow(ensemble), replace = TRUE, prob = c(train_amount, test_amount))
ensemble_train <- ensemble[ensemble_idx == 1, ]
ensemble_test <- ensemble[ensemble_idx == 2, ]

# Fit the ensemble data on the ensemble training data, predict on ensemble test data
ensemble_linear_model <- lm(y_ensemble ~ ., data = ensemble_train)

ensemble_linear_predictions <- predict(object = ensemble_linear_model, newdata = ensemble_test)

ensemble_linear_RMSE[i] <- Metrics::rmse(actual = ensemble_test$y, predicted = ensemble_linear_predictions)

ensemble_linear_RMSE_mean <- mean(ensemble_linear_RMSE)

# Fit the tree model on the ensemble training data, predict on ensemble test data
ensemble_tree_model <- tree(y_ensemble ~ ., data = ensemble_train)

ensemble_tree_predictions <- predict(object = ensemble_tree_model, newdata = ensemble_test)

ensemble_tree_RMSE[i] <- Metrics::rmse(actual = ensemble_test$y, predicted = ensemble_tree_predictions)

ensemble_tree_RMSE_mean <- mean(ensemble_tree_RMSE)

results <- data.frame(
  'Model' = c('Linear', 'Tree', 'Ensemble_Linear', 'Ensemble_tree'),
  'Error_Rate' = c(linear_RMSE_mean, tree_RMSE_mean, ensemble_linear_RMSE_mean, ensemble_tree_RMSE_mean)
)

results <- results %>% arrange(Error_Rate)

```

```

return(list(results))

} # Closing brace for numresamples
} # Closing brace for the function

numerical_1(data = MASS::Boston, colnum = 14, train_amount = 0.60, test_amount = 0.40,
#> [[1]]
#>           Model Error_Rate
#> 1 Ensemble_Linear  2.583655
#> 2  Ensemble_tree  3.500350
#> 3      Linear     4.591160
#> 4      Tree       4.701966

warnings()

```

What we will be doing in this section is making a weighted ensemble using 17 models of numerical data, then using that ensemble to measure the accuracy of the models on the holdout (test) data.

116.1 Think before you do something. This will help when we start at the end and work backwards toward the beginning.

We are going to make an ensemble. The ensemble is going to be made of predictions from numerical models. You've already seen a couple of ensembles. This one will be extremely similar, but will involve more models.

Starting at the end, we want the error rate (root mean squared error) for the ensemble and prediction models.

That means we'll need to make an ensemble. That means we'll need individual model predictions, because that's how an ensemble is made. If an ensemble is made of individual model predictions, that means we'll need individual models. We already know how to do that, because we did it in the last chapter.

We're going to build a simple ensemble with seven models, and then use that ensemble with four very different methods. The Ensembles package actually works with a total of 40 different models. The process is exactly the same, whether we are working with seven individual models or 23 individual models, five ensemble models or 17 ensemble models. The structure and methods are the same.

So let's get started!

The seven individual models we will be building are:

116.2. ONE OF YOUR OWN: ADD ONE MODEL TO THE LIST OF SEVEN INDIVIDUAL MODELS, SEE HOW IT IMPACTS RESULTS.

- Linear (tuned)
- Bayesglm
- Bayesrnn
- Gradient Boosted
- RandomForest
- Trees

It's important to understand that many other options are possible. You are encouraged to add at least one more modeling method to the ensemble, and see how it impacts the results.

116.2 One of your own: Add one model to the list of seven individual models, see how it impacts results.

116.3 Plan ahead as much as you can, that makes the entire model building process much easier.

Here is the code to build ensembles. I very strongly recommend doing this yourself, and checking every 5-10 lines to make sure there are no errors.

```
#Load packages we will need

library(arm) # Allows us to run bayesglm
#> Loading required package: Matrix
#>
#> Attaching package: 'Matrix'
#> The following objects are masked from 'package:tidyr':
#>
#> expand, pack, unpack
#> Loading required package: lme4
#>
#> arm (Version 1.14-4, built: 2024-4-1)
#> Working directory is /Users/russellconte/Library/Mobile Documents/com-apple~CloudDocs/Documents

library(brnn) # Allows us to run brnn
#> Loading required package: Formula
#> Loading required package: truncnorm
```

```
library(e1071) # Allows us to run several tuned model, such as linear and KNN
library(randomForest) # Allows us to run random forest models
#> randomForest 4.7-1.1
#> Type rfNews() to see new features/changes/bug fixes.
#>
#> Attaching package: 'randomForest'
#> The following object is masked from 'package:dplyr':
#>
#>     combine
#> The following object is masked from 'package:ggplot2':
#>
#>     margin
```

```
library(tree) # Allows us to run tree models
```

116.3.1 A few other packages we will need to keep everything running smoothly

```
library(tidyverse) # Amazing set of tools for data science
library(MASS) # Gives us the Boston Housing data set
library(Metrics) # Allows us to calculate accuracy or error rates
```

116.3.2 Build the function that will build the individual and ensemble models

```
numerical <- function(data, colnum, numresamples, train_amount, test_amount){
  # Make the target column the right most column, change the column name to y:

  y <- 0
  colnames(data)[colnum] <- "y"

  df <- data %>% dplyr::relocate(y, .after = last_col()) # Moves the target column to the
  # Set initial values to 0 for both individual and ensemble methods:

  bayesglm_train_RMSE <- 0
  bayesglm_test_RMSE <- 0
  bayesglm_validation_RMSE <- 0
  bayesglm_sd <- 0
  bayesglm_overfitting <- 0
```

116.3. PLAN AHEAD AS MUCH AS YOU CAN, THAT MAKES THE ENTIRE MODEL BUILDING PROCESS MUCH EASIER

```
bayesglm_duration <- 0
bayesglm_duration_mean <- 0
bayesglm_holdout_mean <- 0
bayesglm_holdout_RMSE <- 0
bayesglm_holdout_RMSE_mean <- 0

bayesrnn_train_RMSE <- 0
bayesrnn_test_RMSE <- 0
bayesrnn_validation_RMSE <- 0
bayesrnn_sd <- 0
bayesrnn_overfitting <- 0
bayesrnn_duration <- 0
bayesrnn_duration_mean <- 0
bayesrnn_holdout_mean <- 0
bayesrnn_holdout_RMSE <- 0
bayesrnn_holdout_RMSE_mean <- 0

gb_train_RMSE <- 0
gb_test_RMSE <- 0
gb_validation_RMSE <- 0
gb_sd <- 0
gb_overfitting <- 0
gb_duration <- 0
gb_duration_mean <- 0
gb_holdout_mean <- 0
gb_holdout_RMSE <- 0
gb_holdout_RMSE_mean <- 0

linear_train_RMSE <- 0
linear_test_RMSE <- 0
linear_validation_RMSE <- 0
linear_sd <- 0
linear_overfitting <- 0
linear_duration <- 0
linear_holdout_RMSE <- 0
linear_holdout_RMSE_mean <- 0

rf_train_RMSE <- 0
rf_test_RMSE <- 0
rf_validation_RMSE <- 0
rf_sd <- 0
rf_overfitting <- 0
rf_duration <- 0
rf_duration_mean <- 0
rf_holdout_mean <- 0
```

```

rf_holdout_RMSE <- 0
rf_holdout_RMSE_mean <- 0

tree_train_RMSE <- 0
tree_test_RMSE <- 0
tree_validation_RMSE <- 0
tree_sd <- 0
tree_overfitting <- 0
tree_duration <- 0
tree_duration_mean <- 0
tree_holdout_mean <- 0
tree_holdout_RMSE <- 0
tree_holdout_RMSE_mean <- 0

ensemble_bayesglm_train_RMSE <- 0
ensemble_bayesglm_test_RMSE <- 0
ensemble_bayesglm_validation_RMSE <- 0
ensemble_bayesglm_sd <- 0
ensemble_bayesglm_overfitting <- 0
ensemble_bayesglm_duration <- 0
ensemble_bayesglm_holdout_RMSE <- 0
ensemble_bayesglm_holdout_RMSE_mean <- 0
ensemble_bayesglm_predict_value_mean <- 0

ensemble_bayesrnn_train_RMSE <- 0
ensemble_bayesrnn_test_RMSE <- 0
ensemble_bayesrnn_validation_RMSE <- 0
ensemble_bayesrnn_sd <- 0
ensemble_bayesrnn_overfitting <- 0
ensemble_bayesrnn_duration <- 0
ensemble_bayesrnn_holdout_RMSE <- 0
ensemble_bayesrnn_holdout_RMSE_mean <- 0
ensemble_bayesrnn_predict_value_mean <- 0

ensemble_gb_train_RMSE <- 0
ensemble_gb_test_RMSE <- 0
ensemble_gb_validation_RMSE <- 0
ensemble_gb_sd <- 0
ensemble_gb_overfitting <- 0
ensemble_gb_duration <- 0
ensemble_gb_holdout_RMSE <- 0
ensemble_gb_holdout_RMSE_mean <- 0
ensemble_gb_predict_value_mean <- 0

ensemble_linear_train_RMSE <- 0

```

116.3. PLAN AHEAD AS MUCH AS YOU CAN, THAT MAKES THE ENTIRE MODEL BUILDING PROCESS M

```
ensemble_linear_test_RMSE <- 0
ensemble_linear_validation_RMSE <- 0
ensemble_linear_sd <- 0
ensemble_linear_overfitting <- 0
ensemble_linear_duration <- 0
ensemble_linear_holdout_RMSE <- 0
ensemble_linear_holdout_RMSE_mean <- 0

ensemble_rf_train_RMSE <- 0
ensemble_rf_test_RMSE <- 0
ensemble_rf_test_RMSE_mean <- 0
ensemble_rf_validation_RMSE <- 0
ensemble_rf_sd <- 0
ensemble_rf_overfitting <- 0
ensemble_rf_duration <- 0
ensemble_rf_holdout_RMSE <- 0
ensemble_rf_holdout_RMSE_mean <- 0

ensemble_tree_train_RMSE <- 0
ensemble_tree_test_RMSE <- 0
ensemble_tree_validation_RMSE <- 0
ensemble_tree_sd <- 0
ensemble_tree_overfitting <- 0
ensemble_tree_duration <- 0
ensemble_tree_holdout_RMSE <- 0
ensemble_tree_holdout_RMSE_mean <- 0

#Let's build the function that does all the resampling and puts everything together:

for (i in 1:numresamples) {

  # Randomly split the data between train and test
  idx <- sample(seq(1, 2), size = nrow(df), replace = TRUE, prob = c(train_amount, test_amount))
  train <- df[idx == 1, ]
  test <- df[idx == 2, ]

  # Bayesglm

  bayesglm_train_fit <- arm::bayesglm(y ~ ., data = train, family = gaussian(link = "identity"))
  bayesglm_train_RMSE[i] <- Metrics::rmse(actual = train$y, predicted = predict(object = bayesglm_train_fit))
  bayesglm_train_RMSE_mean <- mean(bayesglm_train_RMSE)
  bayesglm_test_RMSE[i] <- Metrics::rmse(actual = test$y, predicted = predict(object = bayesglm_train_fit))
  bayesglm_test_RMSE_mean <- mean(bayesglm_test_RMSE)
  bayesglm_holdout_RMSE[i] <- mean(bayesglm_test_RMSE_mean)
  bayesglm_holdout_RMSE_mean <- mean(bayesglm_holdout_RMSE)
```

```

bayesglm_test_predict_value <- as.numeric(predict(object = bayesglm_train_fit, newdata =
y_hat_bayesglm <- c(bayesglm_test_predict_value)

# Bayesrnn

bayesrnn_train_fit <- brnn::brnn(x = as.matrix(train), y = train$y)
bayesrnn_train_RMSE[i] <- Metrics::rmse(actual = train$y, predicted = predict(object =
bayesrnn_train_RMSE_mean <- mean(bayesrnn_train_RMSE)
bayesrnn_test_RMSE[i] <- Metrics::rmse(actual = test$y, predicted = predict(object = b
bayesrnn_test_RMSE_mean <- mean(bayesrnn_test_RMSE)
bayesrnn_holdout_RMSE[i] <- mean(c(bayesrnn_test_RMSE_mean))
bayesrnn_holdout_RMSE_mean <- mean(bayesrnn_holdout_RMSE)
bayesrnn_train_predict_value <- as.numeric(predict(object = bayesrnn_train_fit, newdata
bayesrnn_test_predict_value <- as.numeric(predict(object = bayesrnn_train_fit, newdata
bayesrnn_predict_value_mean <- mean(c(bayesrnn_test_predict_value))
y_hat_bayesrnn <- c(bayesrnn_test_predict_value)

# Gradient boosted

gb_train_fit <- gbm::gbm(train$y ~ ., data = train, distribution = "gaussian", n.trees
gb_train_RMSE[i] <- Metrics::rmse(actual = train$y, predicted = predict(object = gb_tr
gb_train_RMSE_mean <- mean(gb_train_RMSE)
gb_test_RMSE[i] <- Metrics::rmse(actual = test$y, predicted = predict(object = gb_train
gb_test_RMSE_mean <- mean(gb_test_RMSE)
gb_holdout_RMSE[i] <- mean(c(gb_test_RMSE_mean))
gb_holdout_RMSE_mean <- mean(gb_holdout_RMSE)
gb_train_predict_value <- as.numeric(predict(object = gb_train_fit, newdata = train))
gb_test_predict_value <- as.numeric(predict(object = gb_train_fit, newdata = test))
gb_predict_value_mean <- mean(c(gb_test_predict_value))
y_hat_gb <- c(gb_test_predict_value)

# Tuned linear models

linear_train_fit <- e1071::tune.rpart(formula = y ~ ., data = train)
linear_train_RMSE[i] <- Metrics::rmse(actual = train$y, predicted = predict(object = l
linear_train_RMSE_mean <- mean(linear_train_RMSE)
linear_test_RMSE[i] <- Metrics::rmse(actual = test$y, predicted = predict(object = lin
linear_test_RMSE_mean <- mean(linear_test_RMSE)
linear_holdout_RMSE[i] <- mean(c(linear_test_RMSE_mean))
linear_holdout_RMSE_mean <- mean(linear_holdout_RMSE)
linear_train_predict_value <- as.numeric(predict(object = linear_train_fit$best.model,
linear_test_predict_value <- as.numeric(predict(object = linear_train_fit$best.model,
y_hat_linear <- c(linear_test_predict_value)

# RandomForest

```



```

rf_train_fit <- e1071::tune.randomForest(x = train, y = train$y, data = train)
rf_train_RMSE[i] <- Metrics::rmse(actual = train$y, predicted = predict(object = rf_train_fit$best.model, newdata = train))
rf_train_RMSE_mean <- mean(rf_train_RMSE)
rf_test_RMSE[i] <- Metrics::rmse(actual = test$y, predicted = predict(object = rf_train_fit$best.model, newdata = test))
rf_test_RMSE_mean <- mean(rf_test_RMSE)
rf_holdout_RMSE[i] <- mean(c(rf_test_RMSE_mean))
rf_holdout_RMSE_mean <- mean(rf_holdout_RMSE)
rf_train_predict_value <- predict(object = rf_train_fit$best.model, newdata = train)
rf_test_predict_value <- predict(object = rf_train_fit$best.model, newdata = test)
y_hat_rf <- c(rf_test_predict_value)

# Trees

tree_train_fit <- tree::tree(train$y ~ ., data = train)
tree_train_RMSE[i] <- Metrics::rmse(actual = train$y, predicted = predict(object = tree_train_fit, newdata = train))
tree_train_RMSE_mean <- mean(tree_train_RMSE)
tree_test_RMSE[i] <- Metrics::rmse(actual = test$y, predicted = predict(object = tree_train_fit, newdata = test))
tree_test_RMSE_mean <- mean(tree_test_RMSE)
tree_holdout_RMSE[i] <- mean(c(tree_test_RMSE_mean))
tree_holdout_RMSE_mean <- mean(tree_holdout_RMSE)
tree_train_predict_value <- as.numeric(predict(object = tree::tree(y ~ ., data = train), newdata = train))
tree_test_predict_value <- as.numeric(predict(object = tree::tree(y ~ ., data = train), newdata = test))
y_hat_tree <- c(tree_test_predict_value)

# Make the weighted ensemble:

ensemble <- data.frame(
  "BayesGLM" = y_hat_bayesglm * 1 / bayesglm_holdout_RMSE_mean,
  "BayesRNN" = y_hat_bayesrnn * 1 / bayesrnn_holdout_RMSE_mean,
  "GBM" = y_hat_gb * 1 / gb_holdout_RMSE_mean,
  "Linear" = y_hat_linear * 1 / linear_holdout_RMSE_mean,
  "RandomForest" = y_hat_rf * 1 / rf_holdout_RMSE_mean,
  "Tree" = y_hat_tree * 1 / tree_holdout_RMSE_mean
)

ensemble$Row_mean <- rowMeans(ensemble)
ensemble$y_ensemble <- c(test$y)
y_ensemble <- c(test$y)

# Split the ensemble into train and test, according to user choices:

ensemble_idx <- sample(seq(1, 2), size = nrow(ensemble), replace = TRUE, prob = c(train_amount, test_amount))
ensemble_train <- ensemble[ensemble_idx == 1, ]
ensemble_test <- ensemble[ensemble_idx == 2, ]

```

Ensemble BayesGLM

```

ensemble_bayesglm_train_fit <- arm::bayesglm(y_ensemble ~ ., data = ensemble_train, fa
ensemble_bayesglm_train_RMSE[i] <- Metrics::rmse(actual = ensemble_train$y_ensemble, p
ensemble_bayesglm_train_RMSE_mean <- mean(ensemble_bayesglm_train_RMSE)
ensemble_bayesglm_test_RMSE[i] <- Metrics::rmse(actual = ensemble_test$y_ensemble, pre
ensemble_bayesglm_test_RMSE_mean <- mean(ensemble_bayesglm_test_RMSE)
ensemble_bayesglm_holdout_RMSE[i] <- mean(c(ensemble_bayesglm_test_RMSE_mean))
ensemble_bayesglm_holdout_RMSE_mean <- mean(ensemble_bayesglm_holdout_RMSE)

```

Ensemble BayesRNN

```

ensemble_bayesrnn_train_fit <- brnn::brnn(x = as.matrix(ensemble_train), y = ensemble_t
ensemble_bayesrnn_train_RMSE[i] <- Metrics::rmse(actual = ensemble_train$y_ensemble, p
ensemble_bayesrnn_train_RMSE_mean <- mean(ensemble_bayesrnn_train_RMSE)
ensemble_bayesrnn_test_RMSE[i] <- Metrics::rmse(actual = ensemble_test$y_ensemble, pre
ensemble_bayesrnn_test_RMSE_mean <- mean(ensemble_bayesrnn_test_RMSE)
ensemble_bayesrnn_holdout_RMSE[i] <- mean(c(ensemble_bayesrnn_test_RMSE_mean))
ensemble_bayesrnn_holdout_RMSE_mean <- mean(ensemble_bayesrnn_holdout_RMSE)

```

Ensemble Graident Boosted

```

ensemble_gb_train_fit <- gbm::gbm(ensemble_train$y_ensemble ~ ., data = ensemble_train
ensemble_gb_train_RMSE[i] <- Metrics::rmse(actual = ensemble_train$y_ensemble, predict
ensemble_gb_train_RMSE_mean <- mean(ensemble_gb_train_RMSE)
ensemble_gb_test_RMSE[i] <- Metrics::rmse(actual = ensemble_test$y_ensemble, predicted
ensemble_gb_test_RMSE_mean <- mean(ensemble_gb_test_RMSE)
ensemble_gb_holdout_RMSE[i] <- mean(c(ensemble_gb_test_RMSE_mean))
ensemble_gb_holdout_RMSE_mean <- mean(ensemble_gb_holdout_RMSE)

```

Ensemble using Tuned Random Forest

```

ensemble_rf_train_fit <- e1071::tune.randomForest(x = ensemble_train, y = ensemble_tra
ensemble_rf_train_RMSE[i] <- Metrics::rmse(actual = ensemble_train$y_ensemble, predict
ensemble_rf_train_RMSE_mean <- mean(ensemble_rf_train_RMSE)
ensemble_rf_test_RMSE[i] <- Metrics::rmse(actual = ensemble_test$y_ensemble, predicted
ensemble_rf_test_RMSE_mean <- mean(ensemble_rf_test_RMSE)
ensemble_rf_holdout_RMSE[i] <- mean(c(ensemble_rf_test_RMSE_mean))
ensemble_rf_holdout_RMSE_mean <- mean(ensemble_rf_holdout_RMSE)

```

Trees

```

ensemble_tree_train_fit <- tree::tree(ensemble_train$y_ensemble ~ ., data = ensemble_t
ensemble_tree_train_RMSE[i] <- Metrics::rmse(actual = ensemble_train$y_ensemble, predi
ensemble_tree_train_RMSE_mean <- mean(ensemble_tree_train_RMSE)

```

116.3. PLAN AHEAD AS MUCH AS YOU CAN, THAT MAKES THE ENTIRE MODEL BUILDING PROCESS MU

```
ensemble_tree_test_RMSE[i] <- Metrics::rmse(actual = ensemble_test$y_ensemble, predicted = predicted)
ensemble_tree_test_RMSE_mean <- mean(ensemble_tree_test_RMSE)
ensemble_tree_holdout_RMSE[i] <- mean(c(ensemble_tree_test_RMSE_mean))
ensemble_tree_holdout_RMSE_mean <- mean(ensemble_tree_holdout_RMSE)

summary_results <- data.frame(
  'Model' = c('BayesGLM', 'BayesRNN', 'Gradient_Boosted', 'Linear', 'Random_Forest', 'Trees', 'Ensemble_BayesGLM', 'Ensemble_BayesRNN', 'Ensemble_Random_Forest', 'Ensemble_Gradient_Boosted', 'Ensemble_Trees', 'Ensemble_Gradient_Boosted_Trees'),
  'Error' = c(bayesglm_holdout_RMSE_mean, bayesrnn_holdout_RMSE_mean, gb_holdout_RMSE_mean, linear_holdout_RMSE_mean, rf_holdout_RMSE_mean, trees_holdout_RMSE_mean, ensemble_bayesglm_holdout_RMSE_mean, ensemble_bayesrnn_holdout_RMSE_mean, ensemble_rf_holdout_RMSE_mean, ensemble_gb_holdout_RMSE_mean, ensemble_trees_holdout_RMSE_mean, ensemble_gb_trees_holdout_RMSE_mean))

summary_results <- summary_results %>% arrange(Error)

return(summary_results)

} # closing brace for numresamples

} # closing brace for numerical function

numerical(data = MASS::Boston, colnum = 14, numresamples = 100, train_amount = 0.60, test_amount = 0.40)
#> Number of parameters (weights and biases) to estimate: 32
#> Nguyen-Widrow method
#> Scaling factor= 0.7016032
#> gamma= 31.337      alpha= 5.385      beta= 13821.09
#> Using 100 trees...
#>
#> Using 100 trees...
#>
#> Using 100 trees...
#>
#> Using 100 trees...
#> Number of parameters (weights and biases) to estimate: 20
#> Nguyen-Widrow method
#> Scaling factor= 0.7039884
#> gamma= 13.689      alpha= 2.0059      beta= 5888.377
#> Using 100 trees...
#>
#> Using 100 trees...
#>


|      | Model                     | Error     |
|------|---------------------------|-----------|
| #> 1 | Ensemble_BayesGLM         | 0.1577687 |
| #> 2 | BayesRNN                  | 0.1909165 |
| #> 3 | Ensemble_BayesRNN         | 0.2074324 |
| #> 4 | Ensemble_Random_Forest    | 0.7931544 |
| #> 5 | Random_Forest             | 1.6955217 |
| #> 6 | Ensemble_Gradient_Boosted | 1.9315225 |
| #> 7 | Ensemble_Trees            | 2.0538994 |
| #> 8 | Gradient_Boosted          | 3.3657072 |


```

```
#> 9                Trees 4.9186294
#> 10               Linear 4.9423946
#> 11               BayesGLM 5.3420949
```

```
warnings()
```

Here's the very cool part of setting it up this way. If you have a totally different data set, all you need to do is put the information into the function, and everything runs. Check this out:

```
numerical(data = ISLR::Auto[, 1:ncol(ISLR::Auto)-1], colnum = 1, numresamples = 250, t
#> Number of parameters (weights and biases) to estimate: 20
#> Nguyen-Widrow method
#> Scaling factor= 0.7024061
#> gamma= 19.0224    alpha= 3.1831    beta= 9309.53
#> Using 100 trees...
#>
#> Using 100 trees...
#>
#> Using 100 trees...
#>
#> Using 100 trees...
#> Number of parameters (weights and biases) to estimate: 20
#> Nguyen-Widrow method
#> Scaling factor= 0.7052367
#> gamma= 12.2626    alpha= 2.3239    beta= 4077.013
#> Using 100 trees...
#>
#> Using 100 trees...
#>
#> Model      Error
#> 1          BayesRNN 0.1338067
#> 2      Ensemble_BayesGLM 0.1338222
#> 3      Ensemble_BayesRNN 0.2492269
#> 4      Ensemble_Random_Forest 0.9258752
#> 5          Random_Forest 1.3439267
#> 6      Ensemble_Gradient_Boosted 1.6029827
#> 7          Ensemble_Trees 1.6641738
#> 8          Gradient_Boosted 2.6607100
#> 9          BayesGLM 3.1566656
#> 10         Linear 3.3890816
#> 11         Trees 3.3890816
```

116.4. ONE OF YOUR OWN: ADD A MODEL TO THE INDIVIDUAL MODELS, AND A MODEL TO THE ENSEMBLE

116.4 One of your own: Add a model to the individual models, and a model to the ensemble of models

One of your own: Change the data, run it again, comment on the results

116.5 Post your results on social media in a way that a non-technical person can understand them. For example:

“Just ran six individual and six ensemble models, very easy to do, no errors or warnings. I plan to do ensembles with other data sets soon. #AIEnsembles

116.6 Exercises to help you improve your skills:

Build an individual numerical model using each of the following model methods (it's perfectly OK to check prior sections of the book, this is an example of delayed repetition):

Gradient Boosted (from the gmb library)

Rpart (from the rpart library)

Support Vector Machines (tuned from the e1071 library)

One model method of your own choosing

Build an ensemble using those four methods, test it using the Boston Housing data set. Compare the results of this ensemble to the one made in the text of this chapter.

Apply the function you made to a different numerical data set. This can be done in one line of code, once the ensemble is set up.

116.7 Post the results of your new ensemble on social media in a way that helps others understand the results or methods.

Chapter 117

4. Classification data: How to make 14 individual classification models

Introduction Photo of City of Chicago snow plow stuck in snow: By Victorgrigas at English Wikipedia - I (t3xt (talk)) created this work entirely by myself., CC0, Link