A Minimal Book Example

John Doe

2025-03-12

Contents

Chapter 1

About

This is a *sample* book written in **Markdown**. You can use anything that Pandoc's Markdown supports; for example, a math equation $a^2 + b^2 = c^2$.

1.1 Usage

Each **bookdown** chapter is an .Rmd file, and each .Rmd file can contain one (and only one) chapter. A chapter *must* start with a first-level heading: `# A good chapter`, and can contain one (and only one) first-level heading.

Use second-level and higher headings within chapters like: `## A short section` or `### An even shorter section`.

The `index.Rmd` file is required, and is also your first book chapter. It will be the homepage when you render the book.

1.2 Render book

You can render the HTML version of this example book without changing anything:

1. Find the **Build** pane in the RStudio IDE, and
2. Click on **Build Book**, then select your output format, or select "All formats" if you'd like to use multiple formats from the same book source files.

Or build the book from the R console:

```
bookdown::render_book()
```

To render this example to PDF as a `bookdown::pdf_book`, you'll need to install XeLaTeX. You are recommended to install TinyTeX (which includes XeLaTeX): https://yihui.org/tinytex/.

1.3 Preview book

As you work, you may start a local server to live preview this HTML book. This preview will update as you edit the book when you save individual .Rmd files. You can start the server in a work session by using the RStudio add-in "Preview book", or from the R console:

```
bookdown::serve_book()
```

Chapter 2

Rstudio

Rstudio shortcuts

keyboard combination	function
opt + _	insert assignment operator `<-`
ESC or ctrl + C	exit `+` prompt
ctrl + shift + m	add pipe operator "%>%"
ctrl + `[/]`	indent or unindent
cmd + D	delete one row
cmd + 1	move cursor to console window
cmd + 2	move cursor to editor window
ctrl + shift + S	add 80 hyphens `---` to signal a new chapter (Addin)
ctrl + shift + =	add 80 equals `===` to signal a new Chapter (Addin)
shift + cmd +N	new R script
cmd + ↑ / ↓	in console, get a list of command history
shift + ↑ / ↓	select one line up/down
fn + F2	`view()` an object, don't select the object
cmd + shift + 1	activate X11() window
ctrl (+ shift) + tab	next (last) tab in scriptor (this applies to all apps); hit ctrl first, then shift if necessary, last tab

Source

keyboard combination	function
cmd + return	Run current line/selection
opt + return	Run current line/selection (retain cursor position)

`Rmd` related

keyboard combination	function
cmd + shift + K	**Knit** rmd
cmd + opt + C	run current code chunk in `Rmd`
cmd + opt + I	insert code chunks in `Rmd`, i.e., ```` ```{r} ```` and ```` ``` ````

Q: How to print output in console rather than inline in Rmd?

A: Choose the gear in the editor toolbar and choose "Chunk Output in Console".

Set working directory

```
dir_folder <- dirname(rstudioapi::getSourceEditorContext()$path) # get the dir name of
setwd(dir_folder) # set as working dir
```

RStudio projects are associated with R working directories. You can create an RStudio project:

- In a brand new directory
- In an existing directory where you already have R code and data
- By cloning a version control (Git or Subversion) repository

Why using R projects:

1. I don't need to use `setwd` at the start of each script, and if I move the base project folder it will still work.
2. I have a personal package with a custom project, which creates my folders just the way I like them. This makes it so that the basic locations for data, outputs and analysis is the same across my work.

Double-click on a `.Rproj` file to open a fresh instance of RStudio, with the working directory and file browser pointed at the project folder.

Q: What is an **R session**? And when do I use it?

A: Multiple concurrent sessions can be useful when you want to:

- Run multiple analyses in parallel
- Keep multiple sessions open indefinitely
- Participate in one or more shared projects

Lauch a new project-less RStudio session

```
# run in console
rstudioapi::terminalExecute("open -n /Applications/RStudio.app", show = FALSE)
```

`-n` Open a new instance of the application(s) even if one is already running.

`rstudioapi::terminalExecute(command, workingDir = NULL, env = character(), show = TRUE)` tells R to run the system command in quotes.

- `command` System command to be invoked, as a character string.
- `workingDir` Working directory for command
- `env` Vector of name=value strings to set environment variables
- `show` If FALSE, terminal won't be brought to front

The `rstudioapi` package provides an interface for interacting with the RStudio IDE with R code. Using`rstudioapi`, you can:

- Examine, manipulate, and save the contents of documents currently open in RStudio,
- Create, open, or re-open RStudio projects,
- Prompt the user with different kinds of dialogs (e.g. for selecting a file or folder, or requesting a password from the user),
- Interact with RStudio terminals,
- Interact with the R session associated with the current RStudio instance.

Set up Development Tools

https://cran.r-project.org/bin/macosx/tools/

- install Xcode command line tools

  ```
  sudo xcode-select --install
  ```

- install GNU Fortran compiler

  Using **Apple silicon** (aka arm64, aarch64, M1) Macs Fortran compiler

- Go to https://www.xquartz.org/, download the .dmg and run the installer.
- Verify that build tools are installed and available by opening an R console and running

  ```
  install.packages("pkgbuild")
  pkgbuild::check_build_tools()
  ```

Insert Code Session

To insert a new code section you can use the **Code** -> **Insert Section** command. Alternatively, any comment line which includes at least four trailing dashes (`-`), equal signs (`=`), or pound signs (`#`) automatically creates a code section.

Define your own shortcuts

https://www.statworx.com/ch/blog/defining-your-own-shortcut-in-rstudio/

https://www.r-bloggers.com/2020/03/defining-your-own-shortcut-in-rstudio/

Install the shortcut packages.

Add code session separators, `---` or `===`.

```
install.packages(
    # same path as above
  "~/Downloads/shoRtcut_0.1.0.tar.gz",
  # indicate it is a local file
  repos = NULL)
install.packages(
    # same path as above
  "~/Downloads/shoRtcut2_0.1.0.tar.gz",
  # indicate it is a local file
  repos = NULL)
```

Now go to Tools > Modify Keyboard Shortcuts and search for "dashes". Here you can define the keyboard combination by clicking inside the empty Shortcut field and pressing the desired key-combination on your keyboard. Click Apply, and that's it!

2.1 Dark Theme

https://community.rstudio.com/t/fvaleature-req-word-background-highlight-color-in-find-and-spellcheck/18578/3

https://rstudio.github.io/rstudio-extensions/rstudio-theme-creation.html

https://docs.posit.co/ide/user/ide/guide/ui/appearance.html#creating-custom-themes-for-rstudio

`.ace_marker-layer .ace_selection` Changes the color and style of the highlighting for the currently selected line or block of lines.

`.ace_marker-layer .ace_bracket` Changes the color and style of the highlighting on matching brackets.

Recommended highlight color: `rgba(255, 0, 0, 0.47)`

`RStudio` editor theme directory on Mac:

right click `RStudio.app`, "Show Package Contents" to navigate to the application folder.

`/Applications/RStudio.app/Contents/Resources/resources/themes/ambiance.rstheme`

Custom theme (user-defined) folder:

- `~/.config/rstudio/themes/idle_fingers_2.rstheme` on mac
- viridis-theme

```
/* yaml tag */
.ace_meta.ace_tag {
  color: #2499DA;
}
/* quoted by $...$ and code chunk options */
.ace_support.ace_function {
  color: #55C667;
}
```

2.2 Packages Management

Install R pakages from source

```
# From local tarball
install.packages(
  # indicate path of the package source file
  "~/Documents/R/UserPackages/shoRtcut2_0.1.0.tar.gz",
  # indicate it is a local file
  repos = NULL)

# From github
install.packages("Rcpp", repos="https://rcppcore.github.io/drat")
```

Check insatlled packages

```
# print all installed packages
rownames(installed.packages())
# check if `ggplot2` is installed
"ggplot2" %in% rownames(installed.packages())
```

Update packages

```
packageVersion("ggplot2") # check package version
install.packages("ggplot2") # update one specific package

## update all installed packages in a stated library location, default to `.libPaths()`
update.packages(lib.loc = .libPaths())
```

Which will ask you for every package if you want to update, to just say yes to everything use `ask = FAlSE`.

```
update.packages(ask = FALSE)
```

Unfortunately this won't update packages installed by `devtools::install_github()`

Updating all Packages after R update

R packages are missing after updating. So have to save the installed packages and re-install them after updating.

```
## get packages installed
packs <- as.data.frame(installed.packages(.libPaths()[1]), stringsAsFactors = F)
# Save to local
f_name <- "~/Documents/R/packages.csv"
rownames(packs)
write.csv(packs, f_name, row.names = FALSE)
packs <- read_csv(f_name)
packs
## Re-install packages using install.packages() after updating R
install.packages(packs$Package)
```

R library path `/Library/Frameworks/R.framework/Versions/4.2-arm64/Resources/library`

- use `find.package("ggplot2")` to find the location of the source file.
- alternatively, `.libPaths()`
  - returns the directory within which packages are looked for.

`library(package)` returns an error if the package doesn't exist.

`require(package)` returns `FALSE` if the package doesn't exist. `require` is designed for use inside other functions.

Put your R package on GitHub

https://jennybc.github.io/2014-05-12-ubc/ubc-r/session2.4_github.html

- Change to the package directory
- Initialize the repository with `git init`
- Add and commit everything with
  1. `git add .` stage changes;
  2. `git status` optional check staged changes, but yet to submit;
  3. and `git commit` submit staged changes.
- Create a new repository on GitHub
- Connect your local repository to the GitHub one

```
git remote add origin https://github.com/username/reponame
```

- Push everything to github

```
git branch -M main
git push -u origin main
```

Chapter 3

Knit Rmd

R Markdown is a powerful tool for combining analysis and reporting into the same document. R Markdown has grown substantially from a package that supports a few output formats, to an extensive and diverse ecosystem that supports the creation of books, blogs, scientific articles, websites, and even resumes.

Nice documentations

- R markdown: The definitive guide. provides detailed references
- R markdown cookbook concise and covers essential functions, with examples.

Quick takeaways:

- Can still use horizontal separator ctrl + shift + S for dashed lines and ctrl + shift + = for equals
- Headers must have one empty line above and below to separate it from other text

YAML metadata

Q: What is YAML?

A: YAML is a human-friendly data serialization language for all programming languages.

Q: What does YAML do?

A: It is placed at the very beginning of the document and is read by each of Pandoc, **rmarkdown**, and **knitr**.

- Provide metadata of the document.
- located at the top of the file.

- adheres to the YAML format and is delimited by lines containing three three dashes (`---`).

It can set values of the template variables, such as `title`, `author`, and `date` of the document.

- The `output` field is used by rmarkdown to apply the output format function `rmarkdown::html_document()` in the rendering process.

  There are two types of output formats in the **rmarkdown** package: documents (e.g., `pdf_document`), and presentations (e.g., `beamer_presentation`).

  Supported output format examples: `html_document`, `pdf_document`.

  R Markdown documents (`html_documents`) and R Notebook documents (`html_notebook`) are very similar; in fact, an R Notebook document is a special type of R Markdown document. The main difference is using R Markdown document (`html_documents`) you have to knit (render) the entire document each time you want to preview the document, even if you have made a minor change. However, using an R Notebook document (`html_notebook`) you can view a preview of the final document without rendering the entire document.

- Many aspects of the LaTeX template used to create PDF documents can be customized using *top-level* YAML metadata (note that these options do not appear underneath the `output` section, but rather appear at the top level along with `title`, `author`, and so on). For example:

```
---
title: "Crop Analysis Q3 2013"
output: pdf_document
fontsize: 11pt
geometry: margin=1in
---
```

  A few available metadata variables are displayed in the following (consult the Pandoc manual for the full list):

Variable	Description
`lang`	Document language code
`fontsize`	Font size (e.g., `10pt`, `11pt`, or `12pt`)
`documentclass`	LaTeX document class (e.g., `article`)
`classoption`	Options for documentclass (e.g., `oneside`)
`geometry`	Options for geometry class (e.g., `margin=1in`)
`mainfont`, `sansfont`, `monofont`, `mathfont`	Document fonts (works only with `xelatex` and `lualatex`)

Variable	Description
`linkcolor`, `urlcolor`, `citecolor`	Color for internal links (cross references), external links (link to websites), and citation links (bibliography)
`linestretch`	Options for line spacing (e.g. 1, 1.5, 3).

– In PDFs, you can use code, typesetting commands (e.g., `\vspace{12pt}`), and specific packages from LaTeX.

1. The `header-includes` option loads LaTeX packages.

```
---
output: pdf_document
header-includes:
- \usepackage{fancyhdr}
---

\pagestyle{fancy}
\fancyhead[LE,RO]{Holly Zaharchuk}
\fancyhead[LO,RE]{PSY 508}

# Problem Set 12
```

2. Alternatively, use `extra_dependencies` to list a character vector of LaTeX packages. This is useful if you need to load multiple packages:

```
---
title: "Untitled"
output:
  pdf_document:
    extra_dependencies: ["bbm", "threeparttable"]
---
```

f you need to specify options when loading the package, you can add a second-level to the list and provide the options as a list:

```
---
title: "Untitled"
output:
  pdf_document:
    extra_dependencies:
      caption: ["labelfont={bf}"]
      hyperref: ["unicode=true", "breaklinks=true"]
      lmodern: null
---
```

Here are some examples of LaTeX packages you could consider using

within your report:

  * pdfpages: Include full PDF pages from an external PDF document within your document.
  * caption: Change the appearance of caption subtitles. For example, you can make the figure title italic or bold.
  * fancyhdr: Change the style of running headers of all pages.

- Some options are passed to Pandoc, such as `toc`, `toc_depth`, and `number_sections`. You should consult the Pandoc documentation when in doubt.

```
---
output:
  pdf_document:
    toc: true
        keep_tex: true
---
```

  * `keep_tex: true` if you want to keep intermediate TeX. Easy to debug. Defaults to `false`.

We can include variables and R expressions in this header that can be referenced throughout our R Markdown document. For example, the following header defines `start_date` and `end_date` parameters, which will be reflected in a list called `params` later in the R Markdown document.

Thus, if we want to use these values in our R code, we can access them via `params$start_date` and `params$end_date`.

Should I use quotes to surround the values?

- Whenever applicable use the unquoted style since it is the most readable.
- Use quotes when the value can be misinterpreted as a data type or the value contains a `:`.

```
# values need quotes
foo: '{{ bar }}' # need quotes to avoid interpreting as `dict` object
foo: '123'       # need quote to avoid interpreting as `int` object
foo: 'yes'           # avoid interpreting as `boolean` object
foo: "bar:baz:bam" # has colon, can be misinterpreted as key

# values need not quotes
foo: bar1baz234
bar: 123baz
```

3.1 Chunk Options

If you want to set chunk options globally, call `knitr::opts_chunk$set()` in a code chunk (usually the first one in the document), e.g.,

````
```{r, label="setup", include=FALSE}
knitr::opts_chunk$set(
  comment = "#>", echo = FALSE, fig.width = 6
)
```
````

Full list of chunk options: https://yihui.org/knitr/options/

Chunk options can customize nearly all components of code chunks, such as the source code, text output, plots, and the language of the chunk.

Other languages are supported in `Rmd`

You can list the names of all available engines via:

```
names(knitr::knit_engines$get())
##  [1] "awk"          "bash"         "coffee"
##  [4] "gawk"         "groovy"       "haskell"
##  [7] "lein"         "mysql"        "node"
## [10] "octave"       "perl"         "php"
## [13] "psql"         "Rscript"      "ruby"
## [16] "sas"          "scala"        "sed"
## [19] "sh"           "stata"        "zsh"
## [22] "asis"         "asy"          "block"
## [25] "block2"       "bslib"        "c"
## [28] "cat"          "cc"           "comment"
## [31] "css"          "ditaa"        "dot"
## [34] "embed"        "eviews"       "exec"
## [37] "fortran"      "fortran95"    "go"
## [40] "highlight"    "js"           "julia"
## [43] "python"       "R"            "Rcpp"
## [46] "sass"         "scss"         "sql"
## [49] "stan"         "targets"      "tikz"
## [52] "verbatim"     "theorem"      "lemma"
## [55] "corollary"    "proposition"  "conjecture"
## [58] "definition"   "example"      "exercise"
## [61] "hypothesis"   "proof"        "remark"
## [64] "solution"     "marginfigure"
```

The engines from `theorem` to `solution` are only available when you use the **bookdown** package, and the rest are shipped with the **knitr** package.

To use a different language engine, you can change the language name in the chunk header from `r` to the engine name, e.g.,

````
```python
x = 'hello, python world!'
print(x.split(' '))
```
````

For engines that rely on external interpreters such as `python`, `perl`, and `ruby`, the default interpreters are obtained from `Sys.which()`, i.e., using the interpreter found via the environment variable `PATH` of the system. If you want to use an alternative interpreter, you may specify its path in the chunk option `engine.path`.

For example, you may want to use Python 3 instead of the default Python 2, and we assume Python 3 is at `/usr/bin/python3`

````
```{python, engine.path = '/usr/bin/python3'}
import sys
print(sys.version)
```
````

- All outputs support markdown syntax.
- If the output is html, you can write in html syntax.

The **chunk label** for each chunk is assumed to be unique within the document. This is especially important for cache and plot filenames, because these filenames are based on chunk labels. Chunks without labels will be assigned labels like `unnamed-chunk-i`, where `i` is an incremental number.

- Chunk label doesn't need a `tag`, i.e., you only provide the `value`.
- If you prefer the form `tag=value`, you could also use the chunk option `label` explicitly, e.g.,

  ````
  ```{r, label='my-chunk'}
  # one code chunk example
  ```
  ````

You may use `knitr::opts_chunk$set()` to change the default values of chunk options in a document.

Commonly used chunk options

- Complete list here. Or `?opts_chunk` to get the help page.

Options	Definitions
`echo=TRUE`	Whether to display the **source code** in the output document.Use this when you want to show the output but not the code itself.
`eval=TRUE`	Whether to evaluate the code chunk.
`include=TRUE`	Whether to include the chunk **output** in the output document. If `FALSE`, nothing will be written into the output document, but the code is still evaluated and plot files are generated if there are any plots in the chunk, so you can manually insert figures later.
`message=TRUE`	Whether to preserve messages emitted by `message()`

Options	Definitions
`warning=TRUE`	Whether to show warnings in the output produced by `warning()`.
`results='markup'`	Controls how to display the text results. When `results='markup'` that is to write text output as-is, i.e., write the raw text results directly into the output document without any markups.Useful when priting `stargazer` tables.
`comment='##'`	The prefix to be added before each line of the text output. Set `comment = ''` remove the default `##`.
`fig.keep='high'`	How plots in chunks should be kept. `high`: Only keep high-level plots (merge low-level changes into high-level plots). `none`: Discard all plots. `all`: Keep all plots (low-level plot changes may produce new plots). `first`: Only keep the first plot. `last`: Only keep the last plot. If set to a numeric vector, the values are indices of (low-level) plots to keep.If you want to choose the second to the fourth plots, you could use `fig.keep = 2:4` (or remove the first plot via `fig.keep = -1`).
`fig.align="center"`	Figure alignment.
`fig.pos="H"`	A character string for the figure position arrangement to be used in `\begin{figure}[]`.
`fig.cap`	Figure caption.

`results='markup'` note plural form for result**s**.

- `markup`: Default. Mark up text output with the appropriate environments depending on the output format. For example, for R Markdown, if the text output is a character string `"[1] 1 2 3"`, the actual output that **knitr** produces will be:

  ```
  ```
  [1] 1 2 3
  ```
  ```

  In this case, `results='markup'` means to put the text output in fenced code blocks (“‘).

- `asis`: Write text output as-is, i.e., write the raw text results directly into the output document without any markups.

  ```
  ```{r, results='asis'}
  cat("I'm raw **Markdown** content.\n")
  ```
  ```

  Sometime, you encounter the following error messages when you have R codes within `enumerate` environment.

> You can't use `macro parameter character #` in horizontal mode.

By default, knitr prefixes R output with `##`, which can't be present in your TeX file.

Solution:

- specify `results="asis"` in code chunks.

- `hold`: Hold all pieces of text output in a chunk and flush them to the end of the chunk.
- `hide` (or `FALSE`): Hide text output.

3.2 Print Verbatim R code chunks

Including verbatim R code chunks inside R Markdown

One solution for including verbatim R code chunks (see below for more) is to insert hidden inline R code (`` `r ''` ``) immediately before or after your R code chunk.

- The hidden inline R code will be evaluated as an inline expression to an empty string by knitr.

Then wrap the whole block within a markdown code block. The rendered output will display the verbatim R code chunk — including backticks.

R code generating the four backticks block:

```
output_code <-
"````markdown
```{r}
plot(cars)
``` \n````"
cat(output_code)
```

Write this code in your R Markdown document:

````
````markdown
`r ''````{r}
plot(cars)
```
````
````

or

````
````markdown
```{r}`r '''
````
````

````
plot(cars)
```
````

Knit the document and the code will render like this in your output:

````
```{r}
plot(cars)
```
````

References:

https://yihui.org/en/2017/11/knitr-verbatim-code-chunk/

https://support.posit.co/hc/en-us/articles/360018181633-Including-verbatim-R-code-chunks-inside-R-Markdown

https://themockup.blog/posts/2021-08-27-displaying-verbatim-code-chunks-in-xaringan-presentations/
````

Chapter 4

Machine Learning

Parametric models such as generalized linear regression and logistic regression has advantages and disadvantages.

Strength:

- The effects of individual predictors on the outcome are easily understood
- Statistical inference, such as hypothesis testing or interval estimation, is straightforward
- Methods and procedures for selecting, comparing, and summarizing these models are well-established and extensively studied

Disadvantages in the following scenarios:

- Complex, non-linear relationships between predictors and the outcome
- High degrees of interaction between predictors
- Nominal outcome variables with several categories

In these situations, non-parametric or algorithmic modeling approaches have the potential to better capture the underlying trends in the data.

Here we introduce three models: classification and regression trees (CART), random forests, k-nearest neighbors.

- Classification and regression trees (CART) are "trained" by recursively partitioning the -dimensional space (defined by the explanatory variables) until an acceptable level of homogeneity or "purity" is achieved within each partition.
- A major issue with tree-based models is that they tend to be high variance (leading to a high propensity towards over-fitting). Random forests are a non-parametric, tree-based modeling algorithm that is built upon the idea that averaging a set of independent elements yields an outcome with lower variability than any of the individual elements in the set.

This general concept should seem familiar. Thinking back to your introductory statistics course, you should remember that the sample mean, $\overline{x}$, of a dataset has substantially less variability ($\frac{\sigma}{\sqrt{n}}$) than the individual data-points themselves (σ).

Q: What is Bias-Variance Trade-Off in Machine Learning?
A:

- Bias refers to error caused by a model for solving complex problems that is over simplified, makes significant assumptions, and misses important relationships in your data.
- Variance error is variability of a target function's form with respect to different training sets. Models with small variance error will not change much if you replace couple of samples in training set. Models with high variance might be affected even with small changes in training set. High variance models fit the data too well, and learns the noise in addition to the inherent patterns in the data.

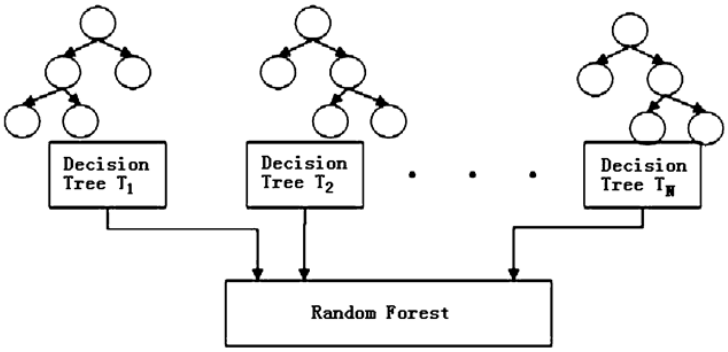
4.1 Random Forest

Averaging of independent trees

The goal of bagging is to produce B separate training datasets that are independent of each other (typically is in the hundreds). The model of interest (in this case classification and regression trees) is trained separately on each of these datasets, resulting in B different estimated "models". These are then averaged to produce a single, low-variance estimate.

Bagging is a general approach, but its most well-known application is in the random forest algorithm:

1. Construct B bootstrap samples by sampling cases from the original dataset with replacement (this results in B unique datasets that are similar to the original)
2. Fit a classification and regression tree to each sample, but randomly choose a subset of m variables that can be used in the construction of that tree (this results in B unique trees that are fit to similar datasets using different sets of predictors)
3. For a given data-point, each of the B trees in the forest contributes a prediction or "vote", with the majority (or average) of these votes forming the random forest's final prediction, $\widehat{y}_i$



A downside of both the CART and random forest algorithms (as well as many other algorithmic modeling approaches) is an inability to clearly quantify the roles played by individual variables in making predictions. However, the importance of individual variables in a random forest can still be expressed using a measure known as variable importance.

The random forest algorithm requires the following tuning parameters be specified in order to run:

- `ntree` - the number of bagged samples, B, onto which trees will be grown
- `mtry` - the number of variables that are randomly chosen to be candidates at each split
- Some sort of stopping criteria for individual trees, this can be:
  - `nodesize`, which sets the minimum size of terminal nodes
    - larger `nodesize` leads to shallower trees
    - smaller node size allows for deeper, more complex trees
  - `maxnodes`, which sets the maximum number of terminal nodes an individual tree can have.

Applications of Random Forest

Some of the applications of Random Forest Algorithm are listed below:

- Banking: It predicts a loan applicant's solvency. This helps lending institutions make a good decision on whether to give the customer loan or not. They are also being used to detect fraudsters.
- Health Care: Health professionals use random forest systems to diagnose patients. Patients are diagnosed by assessing their previous medical history. Past medical records are reviewed to establish the proper dosage for the patients.
- Stock Market: Financial analysts use it to identify potential markets for stocks. It also enables them to remember the behaviour of stocks.
- E-Commerce: Through this system, e-commerce vendors can predict the preference of customers based on past consumption behaviour.

When to Avoid Using Random Forests?

Random Forests Algorithms are not ideal in the following situations:

- Extrapolation: Random Forest regression is not ideal in the extrapolation of data. Unlike linear regression, which uses existing observations to estimate values beyond the observation range.
- Sparse Data: Random Forest does not produce good results when the data is sparse. In this case, the subject of features and bootstrapped sample will have an invariant space. This will lead to unproductive spills, which will affect the outcome.

FAQ

Q: Is RF a linear or non-linear model?
A: RF can capture complex, non-linear relationships.

Q: Is RF sensitive to Imbalanced Data?
A: It may perform poorly if the dataset is highly imbalanced like one class is significantly more frequent than another.

Q: What is the loss function?
A: Entropy/gini or any other loss function you want.

Q: Difference btw RF and a linear model?
A: A major difference is that a decision tree does not have "parameters", whereas the linear models need to create a functional form and find the optimal parameters.

Implementation in R

`ranger` package offers a computation efficient function for RF.

```
RF_ranger <- ranger(formula = formula,
                    data = data_before[idx,],
                    probability = TRUE,
                    importance = "permutation",
                    scale.permutation.importance = TRUE,
                    )
    # print(RF_ranger)

rf.pred.test <- predict(RF_ranger, data=data_before[-idx,])$predictions
```

Parameters controlling the general process of RF:

- `probability=FALSE`: Whether to forecast a probability forest.

The hyperparameters `mtry`, `min.node.size` and `sample.fraction` determine the degree of randomness, and should be tuned.

- `mtry=500`: Number of variables to possibly split at in each node in one tree. In plain language, it indicates how many predictor variables should be used in each tree.
  - Default is the (rounded down) square root of the number variables. Alternatively, a single argument function returning an integer, given the number of independent variables.
  - Range btw 1 to the number of predictors.
  - If all predictors are used, then this corresponds in fact to bagging.
- `min.node.size`: The number of observations a terminal node should at least have.
  - Default 1 for classification, 5 for regression, 3 for survival, and 10 for probability. For classification, this can be a vector of class-specific values.
  - Range between 1 and 10
- `sample.fraction`: Fraction of observations to be used in each tree. Default is 1 for sampling with replacement and 0.632 for sampling without replacement. For classification, this can be a vector of class-specific values.
  - Smaller fractions lead to greater diversity, and thus less correlated trees which often is desirable.
  - Range between 0.2 and 0.9

Parameters controlling what and how intermediate results are saved:

- `keep.inbag = FALSE`: Whether to save how often observations are in-bag in each tree.

  Set to `TRUE` if you want to check sample composition in each tree.

- `importance = 'none'|'impurity'|'impurity_corrected'|'permutation'`: Variable importance mode.

- `scale.permutation.importance = FALSE`: Whether to scale permutation importance by standard error as in (Breiman 2001). Only applicable if `'permutation'` variable importance mode selected.

- `write.forest = TRUE`: Whether to save `ranger.forest` object, required for prediction. Set to `FALSE` to reduce memory usage if no prediction intended.

  - Set to `FALSE` when you do parameter tuning.

Q: How to tune hyperparameters?
A: Check out `mlr3` package. Here is an example.

Imbalance Classification

You can balance your random forests using case weights. Here's a simple example:

```r
library(ranger)

# Make a dataste
set.seed(43)
nrow <- 1000
ncol <- 10
X <- matrix(rnorm(nrow * ncol), ncol=ncol)
CF <- rnorm(ncol)
Y <- (X %*% CF + rnorm(nrow))[,1]
Y <- as.integer(Y > quantile(Y, 0.90))
table(Y)

# Compute weights to balance the RF
w <- 1/table(Y)
w <- w/sum(w)
weights <- rep(0, nrow)
weights[Y == 0] <- w['0']
weights[Y == 1] <- w['1']
table(weights, Y)

# Fit the RF
data <- data.frame(Y=factor(ifelse(Y==0, 'no', 'yes')), X)
model <- ranger(Y~., data, case.weights=weights)
print(model)
```

Code Source: https://stats.stackexchange.com/a/287849

References:

https://remiller1450.github.io/m257s21/Lab10_Other_Models.html