Institute and Faculty of Actuaries

B1: Approaches to Building Models in R

Speaker: Simon Tam, EY

05 June 2017

Institute and Faculty of Actuaries

Agenda

- Introduction to R
- Key tools
- Simplified modelling process demonstration
- Benefits of using R for building models
- R limitations
- Resources

Introduction to R

- Open source statistical programming language based upon “S”
- R is one of the most popular data science tools (along with Python)
- The base functionality can be expanded using “packages”
- The usage of R has dramatically increased over recent years:
  - Popular with educational and research communities (e.g. LondonR)
  - Known to be used at many of the leading tech firms (Airbnb, Facebook, Google, Twitter, Uber, etc.)
  - R Consortium support from Google, IBM, Microsoft, Oracle, etc.
  - Microsoft have invested significantly in R after their purchase of Revolutions Analytics (R Open, R Server, SQL Server, AzureML)
  - Insurance applications (e.g. `library(actuar)`, `library(ChainLadder)`, R in Insurance)

Institute and Faculty of Actuaries

Key tools (1/2) – RStudio, R Markdown, R Notebooks

- RStudio is a popular Integrated Development Environment (IDE) for R
- R Markdown documents can be “knit” into HTML, PDF, Word documents or even PowerPoint slides
  - *“R Markdown documents are fully reproducible. Use a productive notebook interface to weave together narrative text and code to produce elegantly formatted output. Use multiple languages including R, Python, and SQL.”* (http://rmarkdown.rstudio.com/)
- R Notebooks are an extension of R Markdown documents that allow outputs to be inline with code (similar concept to Jupyter Notebooks)

Institute and Faculty of Actuaries

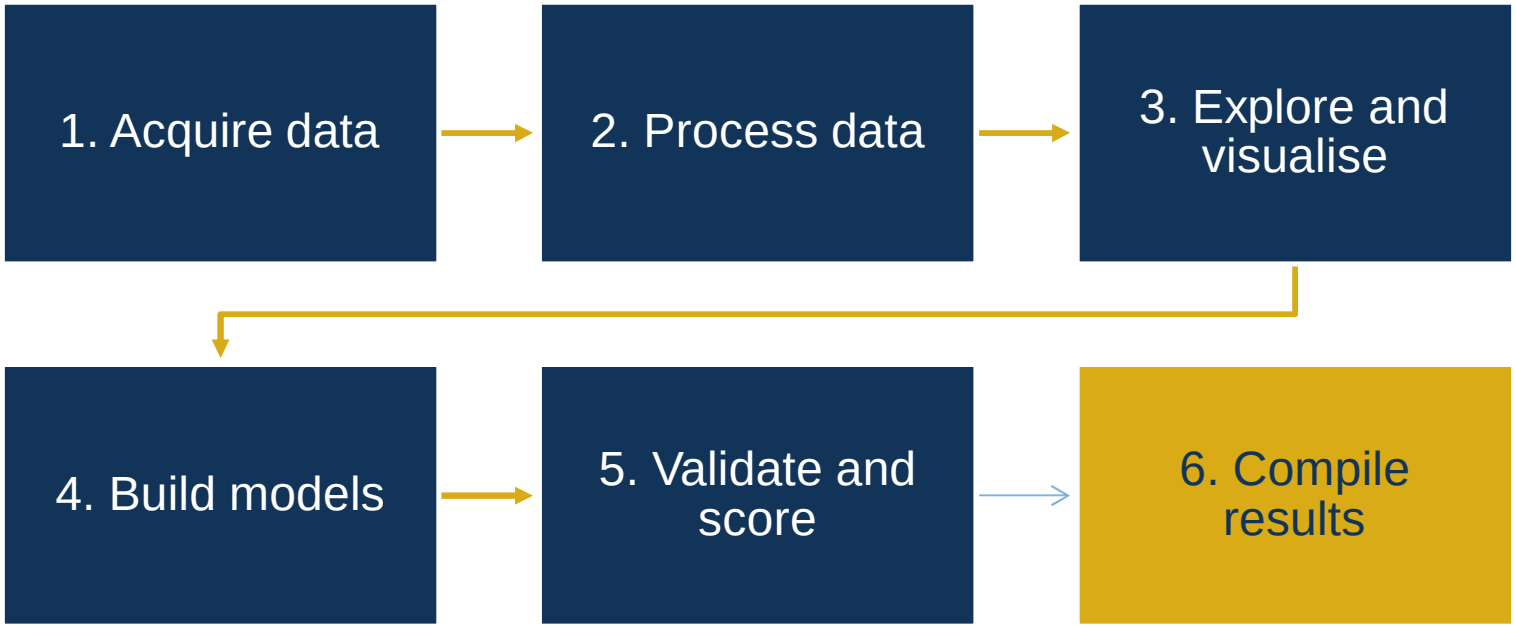
Key tools (2/2) – `library(checkpoint)`

- The successful execution of an R project relies on numerous packages
- However the packages installed by one user may differ from another
- To ensure reproducibility, the checkpoint package downloads the packages as of a particular date for use within an R project

```
library(checkpoint)
checkpoint("2017-06-06")
```

Institute and Faculty of Actuaries

Simplified modelling process



Institute and Faculty of Actuaries

1. Acquire data

- Common file formats are easily read into R
  - `library(data.table)`, `fread(…)` for CSV (as an alternative to `read.csv(…)`)
  - `library(readxl)` for Excel
  - `library(haven)` for SAS datasets
- Access and submit SQL queries using ODBC and `library(dplyr)`
- Other packages can be used to access online APIs or scrape information from the internet
- For testing purposes `library(insuranceData)`may be useful
- Use `readRDS(…)` and `saveRDS(…)` to read and save R objects

Institute and Faculty of Actuaries

2. Process data

- Data is usually stored in a `data.frame` object
  - Consider using `tibble` which makes some improvements in the default `data.frame` behaviour
- Two main packages are used for processing data in R
  - `library(dplyr)` uses action verbs to act upon data frames
  - `library(data.table)` is faster and more powerful however the syntax is more challenging to learn
- `library(feather)` can be used to datasets in a faster binary format
  - `feather` was co-created by the authors of `dplyr` and Python package `pandas`
  - This package was meant to be cross compatible between R and Python

Institute and Faculty of Actuaries

3. Explore and visualise

- `library(ggplot2)` is a very popular graphics package for R
  - "The Grammar of Graphics" is written by Leland Wilkinson and defines a structure to the way data is presented and visualised
  - Graphs are built up by defining `aesthetics` (assign data to `x`, `y`, fill, etc.), `geoms` (types of plots), and other components such as labels, axes, titles, themes etc. `facet_grid` can create panels by dimension.
  - Commonly used in professional publications (newspapers, data journalism, etc.)
  - Functionality can also be expanded using extensions
  - `library(ggvis)` is the next iteration of "gg" graphics with interactive components on the `Shiny` platform
- Alternatives include the Base R graphics and `library(lattice)`which focuses on displaying multivariate relationships
- Note: two axes charts are not easy to implement in R

Institute and Faculty of Actuaries

4. Build models

- `glm(…)` is already included within the included `library(stats)` package
  - Use `step(…)` to execute stepwise regression
  - Use `drop1(glmfit, test="Chisq")` to test factor significance
  - `library(broom)` makes it easier to process `glm(…)` output using the verbs `tidy(…)`, `glance(…)`, `augment(…)`
  - Note: `glm` objects should be pared down to save memory (http://www.win-vector.com/blog/2014/05/trimming-the-fat-from-glm-models-in-r/)
- A selection of popular machine learning packages is listed below:

Package	Description
`earth`	Multivariate Adaptive Regression Splines
`gam`	Generalized additive models with smoothness estimation
`gbm`	Gradient Boosted Regression models
`glmnet`	Lasso and Elastic-Net Regularized GLMs
`lme4`	Linear Mixed-Effects Models using 'Eigen' and S4
`xgboost`	Extreme Gradient Boosting

Package	Description
`caret`	(C)lassification And (RE)gression (T)raining
`h2o`	R scripting functionality for H2O, open source math engine
`sparklyr`	R Interface to Apache Spark

Institute and Faculty of Actuaries

5. Validate and score (1/2)

- Models can be easily scored on different datasets using the associated `predict(…)` functions
- Deviance can be extracted from `glance(…)`
- `library(Hmisc)`, `rcorr.cens(…)` used to determine the Gini coefficient
- `library(ggplot2)` can be used to create coefficient and standard error graphs using `geom_pointrange(…)`
- Other techniques to consider
  - k-fold cross-validation can be implemented using `cv.glm(…)` although difficult to interpret and perhaps better done using a `library(purrr)` approach
  - `library(broom)` has functionality to bootstrap models

Institute and Faculty of Actuaries

5. Validate and score (2/2)

- `data.frames` are not constrained to holding “data”; they can also hold lists of other objects
- For example, the structure below could be useful in examining different model types

Index	Description	Model_Object	Gini_Train …	Gini_Val …	Graphs …
1	Gradient Boosted Tree	model_gbm	0.342342	0.27373	list…
2	Lasso	model_glmnet	0.340994	0.2699	list…
3	Mixed Model	model_mm	0.330509	0.2238	list…
4	XGBoost	model_xgboost	0.350329	0.28882	list…

- `library(purrr)` facilitates computing metrics and diagnostics across multiple models
- This structure also makes it easy to extract all relevant information for one model iteration

Institute and Faculty of Actuaries

6. Compile results

- Previously discussed R Markdown and R Notebook
- R Markdown also supports Python code chunks
- Government Digital Services (GDS) is using R Markdown for a project
  - Working on a project with the Department for Culture, Media, and Sport (DCMS) on the production of "Economic Estimates for DCMS Sectors Statistical First Release (SFR)"
  - *"At any point in the future we should be able to look back at this work and be able to reproduce everything that we have done today - something that is difficult with manual/semi-manual processes."* (GDS)
  - https://gdsdata.blog.gov.uk/2017/03/27/reproducible-analytical-pipeline/

Institute and Faculty of Actuaries

Benefits of using R for building models

- In “Assessing data analysis and programming” (Hadley Wickham, Garret Grolemund), three properties of good data analysis are noted:
  - [Reproducibility] Rather than using multiple pieces of software which require user intervention, the modelling process can be completed more efficiently from end to end exclusively in R
  - [Automation] With a robust process in R, changes to data and models are easily managed within the processing framework, e.g. updating models with new data and creating comparisons against existing models
  - [Communication] Description of the modelling process is improved with Markdown and Notebook documents
- Packages can be used to enhance and support every phase of the modelling process
- New techniques can be quickly tested within the modelling framework
- Very easy to get started
- Community support and established knowledge base online

Institute and Faculty of Actuaries

R limitations

- All objects are stored in memory
- For larger scale projects, professionally supported versions or cloud-based solutions should be considered
- Python offers functionality that in many cases exceeds R and is perhaps preferred by data scientists and computer programmers https://www.datacamp.com/community/tutorials/r-or-python-for-data-analysis
- Newer packages and developments may not be as thoroughly vetted as commercial software options

Institute and Faculty of Actuaries

Resources

- R help function, e.g. `help(glm)`
- Package vignettes which provide examples
- Hadley Wickham is a well known author/contributor of many of the packages discussed today and Chief Scientist at RStudio http://www.tidyverse.org/
- RStudio Cheat Sheets (`dplyr`, `ggplot2`, `rmarkdown`, etc.) https://www.rstudio.com/resources/cheatsheets/
- Cross Validated (Stack Overflow) https://stats.stackexchange.com/
- YouTube

Institute and Faculty of Actuaries

Questions

Comments

Institute and Faculty of Actuaries

Contact

Presenter:

Simon Tam STam@uk.ey.com

Alternate contacts:

Andy Saunders ASaunders1@uk.ey.com

Ben Wilson BWilson2@uk.ey.com

Theeban Kuganesan TKuganesan@uk.ey.com

Institute and Faculty of Actuaries