Reproducible Data Science: What Can We Learn From Other Professions?

Philip Darke FIA
Dr Matthew Forshaw
Case study
Official statistics in UK Government
Official statistics in UK Government

In-house software

https://ukgovdatascience.github.io/rap_companion/why.html#the-current-statistics-production-process
Manual processes add risk

- Errors in spreadsheets are common
- Manual processes risk introducing human error
- Checking and peer review are not embedded in the process
- Challenging to reproduce previous work
Reproducible analytical pipelines

https://ukgovdatascience.github.io/rap_companion/why.html#desired-reproducible-analytical-pipeline
The potential **time savings** for analysts are enormous, freeing them up to focus on the interpretation of the results. The other huge benefit comes from building a process that is fully **transparent, auditable and verifiable** – reducing risk and improving quality.

Matt Upson and Mat Gregory, Government Digital Service

How critics and filmgoers disagree
Difference in average score from critics and audience for 2017's Oscar-nominated films

- The Square
- All The Money In The World
- The Insult
- Molly's Game
- Guardians Of The Galaxy, Vol.2
- Wonder
- Marshall
- Beauty and the Beast
- Loving Vincent
- Kong: Skull Island
- Victoria & Abdul
- Roman J. Israel, Esq.
- Ferdinand
- The Boss Baby
- The Greatest Showman

Source: IMDB

https://www.bbc.co.uk/news/entertainment-arts-43146027 (edited to fit)
[This approach] **saves a huge amount of time and effort**, in particular when working with data that needs updating regularly, with **reproducibility** a key requirement of our workflow. In short, it was a game changer…

BBC Visual and Data Journalism team

Reproducible work
Reproducibility is the process of making code and data available so that others can easily replicate, verify and build on your analysis.
Building blocks of a reproducible workflow

Data
Analytic code and automated checks
Documentation
Computational environment
Packaged in a standard way
Why is this important for actuaries?

• Enables more efficient working
• Allows analysts to focus on the bigger picture
• Easier collaboration
• Helps meet compliance requirements – internal and TAS
• A step towards automation
Reproducible actuarial pipeline

An example using R
Why R?

- Stable, up-to-date and free
- Open source with an active support community
- Well suited to building reproducible pipelines and reporting
- Next generation of actuaries will learn R under the 2019 curriculum
Where does that figure come from?
Hashflow analysis

Organisation name

September 2019

... the present value of the projected cashflows is £8,851,000.

... a present value in the range £8,696,000 to £9,007,000.

... 95% of model outcomes have a present value in the range £8,696,000 to £9,007,000.
Update the report using 2.75%
Cashflow analysis

Organisation name

September 2019

Summary

Using a discount rate of 3% p.a., the present value of the project is projected. 95% of model outcomes have a present value in the range

Cashflow analysis

Lorem ipsum dolor sit amet, ad mea sumo vocibus graviora. Quo offendisse quos, quemque pecunia est, нaer longeaeque qui deserunt, vix disce deterrimi. Nisi magnis atque, velit autem nasciuntur ut, pri esse armatis delenit.


Dicit alterum positionem qui. Quod quod consul est. Emi mei accusam officii, habemus accusam officii, nam quid quid. Quod quid quid.

12 month cashflow projection

<table>
<thead>
<tr>
<th>100000</th>
<th>150000</th>
<th>200000</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

# Load project
library("ProjectTemplate")
load.project()

# Model cashflows as an ARIMA(2,1,0) time series
cashflow_model <- arima(cashflows, order = c(2,1,0))

# Create a 12 month forecast
forecast <- forecast(cashflow_model, 12, level = c(80, 90, 95, 99))

# Plot the forecast
cf_plot <- autoplot(forecast) +
  xlab("Year") +
  ylab("Cashflow") +
  title("12 month cashflow projection")

# Set discount rate
disc <- 0.03

# Discount cashflows
pv_lower <- discount(forecasts["lower"], disc, 12)
pv_central <- discount(forecasts["central"], disc, 12)
pv_upper <- discount(forecasts["upper"], disc, 12)
Collaboration and keeping an audit trail
Collaboration and keeping an audit trail

Update exercises (#3) · phillip

Fix unit test bug, simplify assumptions table in report and minor wording changes

phillipdarke committed on 3 Mar

1 parent 2902f16  commit 34a5948963e846b5ba5966db5824a75e8f155130

Showing 4 changed files with 60 additions and 38 deletions.

---

```r
### Assumptions

```---

- `summarise_assumptions_used` table = data.frame(c("Discount rate", "", "", ""),
- `c(paste(100+disc, "% p.a.", sep="", "", "")))`
- `kable(table, col.names=c("Assumption", "Value"))`
- `Assumption | Value`
- `Discount rate |
- `r paste(100+disc, "% p.a.", sep="", "", "")`
Challenges

- Relies on open source software
- Timing consuming to set up
- Training requirements
Build a simple reproducible pipeline at
philipdarke.com/reproducible-actuarial-work
Applying these techniques in your work

- Take an existing process
- Develop a minimal viable solution (see the exercises)
- Pilot it and let others contribute
- Share what you learn
The views expressed in this presentation are those of invited contributors and not necessarily those of the IFoA. The IFoA do not endorse any of the views stated, nor any claims or representations made in this presentation and accept no responsibility or liability to any person for loss or damage suffered as a consequence of their placing reliance upon any view, claim or representation made in this presentation.

The information and expressions of opinion contained in this publication are not intended to be a comprehensive study, nor to provide actuarial advice or advice of any nature and should not be treated as a substitute for specific advice concerning individual situations. On no account may any part of this presentation be reproduced without the written permission of the authors.
Useful tools for building a reproducible workflow

**RStudio** is a free and widely used development environment for R that integrates with the tools below.

**ProjectTemplate** automates the menial parts of statistical analysis and provides a standard way of working in R.

**R Markdown** is a notebook interface that allows code to sit alongside narrative text and can be used for reporting as part of a reproducible framework with **ggplot2** for creating charts and visualisations.

**Git** is a version control system for managing code and audit trails – it can be used privately in an organisation or with a web-based service such as **GitHub**.

**testthat** is a formal automated testing (“unit testing”) package for R.

**TravisCI** integrates with GitHub to automatically run your tests when code is updated.

**roxygen2** automates the production of documentation for your code in R.

**Docker** packages dependencies inside a container which can run consistently on any infrastructure (also see **checkpoint/packrat** or consider creating a R package).
References and resources

- RAP companion [https://ukgovdatascience.github.io/rap-website/](https://ukgovdatascience.github.io/rap-website/)
- RAP Udemy video course [https://www.udemy.com/course/reproducible-analytical-pipelines/](https://www.udemy.com/course/reproducible-analytical-pipelines/)
- Accompanying exercises [https://philipdarke.com/reproducible-actuarial-work/](https://philipdarke.com/reproducible-actuarial-work/)
- Icons made by Smashicons and Dimitry Mirolubov from [www.flaticon.com](http://www.flaticon.com)
Get in touch

Dr Matthew Forshaw is a Lecturer in Data Science at Newcastle University, and Data Skills Policy Leader at The Alan Turing Institute working on the Data Skills Taskforce. He is the Programme Director of Newcastle’s Industrial MSc in Data Science. mattforshaw.com

Philip Darke is an actuary with over 10 years’ consulting experience at Mercer and a PhD researcher in data science at the EPSRC Centre for Doctoral Training in Cloud Computing for Big Data at Newcastle University. philipdarke.com