How do GfK operationalise forecast processes with R?

Andrea Liotta, CMAF Workshop, 29th April 2019

Agenda

1. About me and GfK

2. Our R journey

3. What we are doing now with R

4. How to build data-driven and R-based products

5. Takeaways

About me

- I am a Senior Data Scientist at GfK

- I joined in 2015 after completing the MSc thesis' project at LUMS

- My main work focuses on time series modelling and forecasting

- My unit provides several forecast products to satisfy different clients' needs from financial hedge funds to consumer and retail brand

About GfK

- Global marketing analytics company offering multiple solutions mainly based on Point-of-Sales (POS) data

- GfK is undertaking a transformation from a traditional marketing company into a tech-enabled analytics provider

- Data Science is a key contributor on this journey, by developing and implementing the predictive and prescriptive analytics

Our R journey

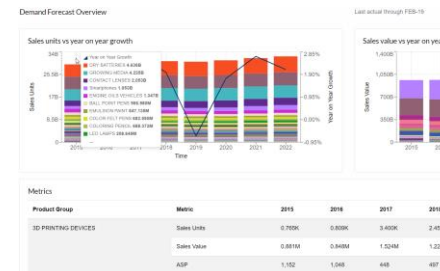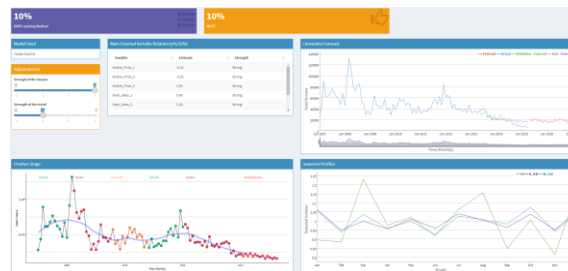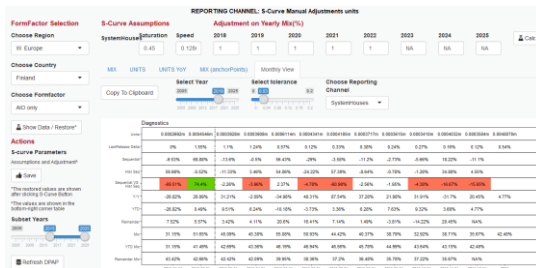## R helps us in moving from a human-driven to a data-driven approach

Human-driven

Data-driven

2. First full R-powered product

3. R & Excel based product

1. My MSc Project

4. First full R-powered automated product

What we are doing now with R

Combine market knowledge and state of the art technology and methods to continuous delivery of valuable analytics solutions

Product

- Automatic batch forecast

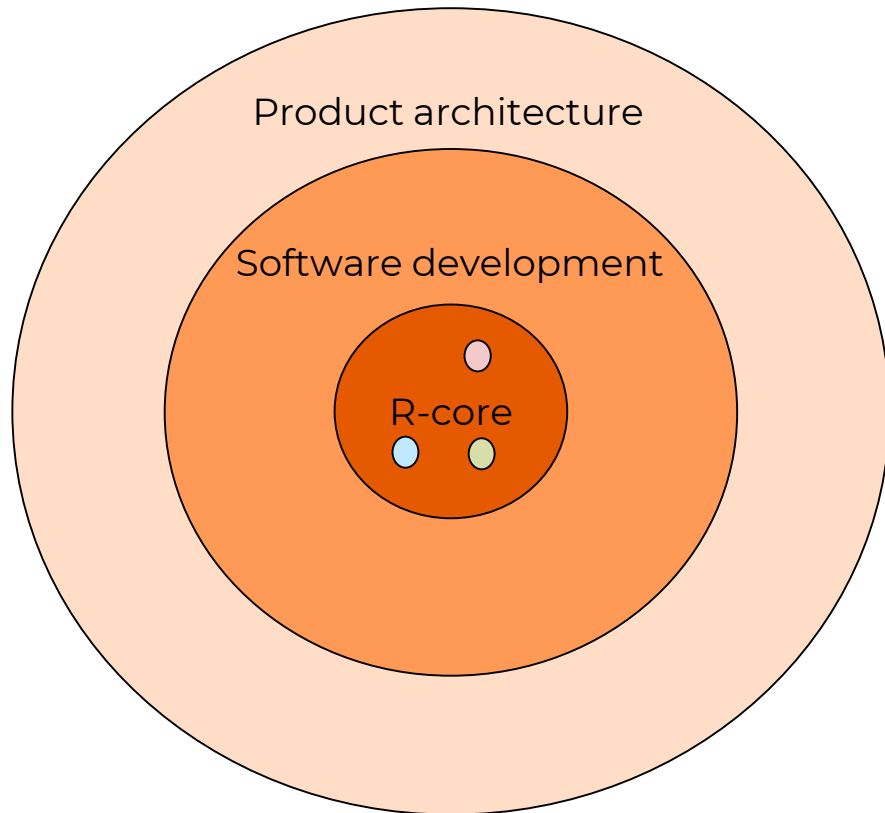 - >10K hierarchical time-series

- Forecast on Demand

Challenges

- No forecast review

- No model review

- No review on actual

Need for a full automate forecast pipeline

How to build data-driven and R-based products

Thinking in layers

R-core

The key starting points

Develop projects in R packages

- coding standard ([lintr](#) based on tidyverse)

- automatic tests ([testthat](#))

Standardise the delivery of the solution

- automatic report (RMarkdown, [officer](#))

- clear interface

  - web API ([plumber](#), [OpenCPU](#))

  - wrap R functions within custom layers

Software development

No full-proof solution yet

Same pipeline for research and production

- double costs

- increase testing

Controlled workflow (git-based)

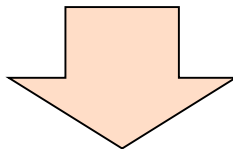- shared ownership with responsibilities

Consistent infrastructure (Docker)

Product architecture

Be able to deliver a robust solution without human intervention

Quality control by design

- evaluate results from "get the data" to deliverable

- modularisation, decouple IO and core functionalities

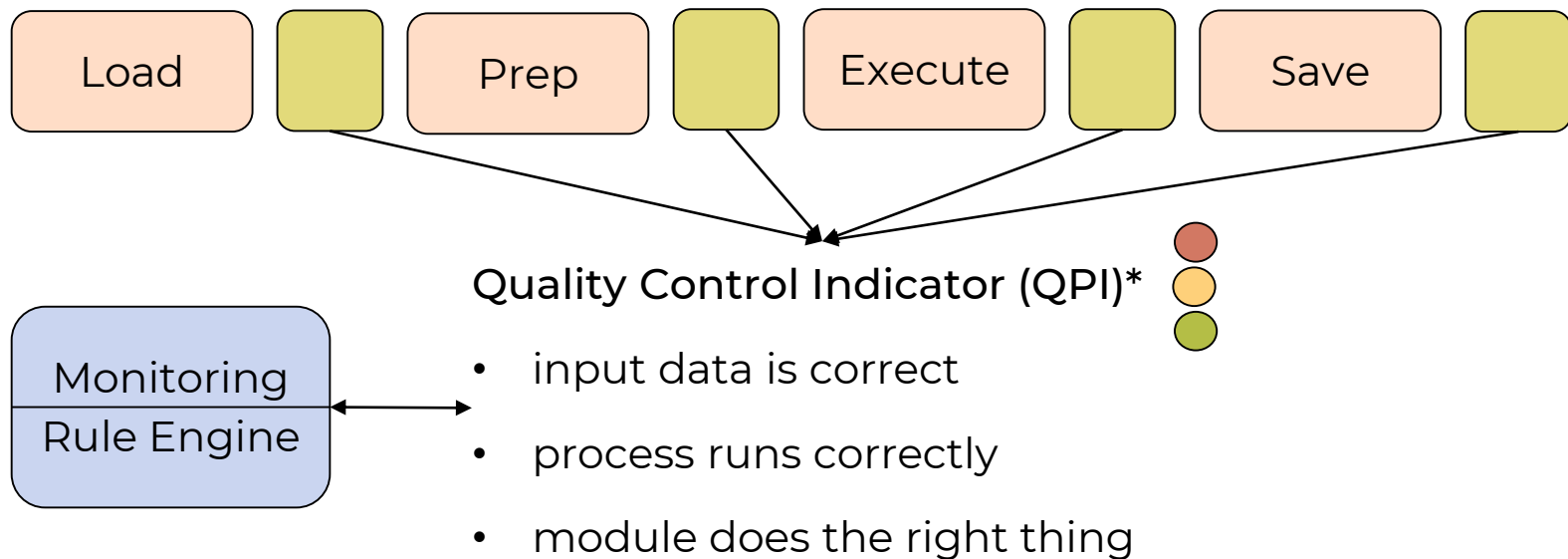- tracking of the module behaviour

Paradigm Shift

From - Develop a sophisticated solution and scale it

To - Build a simple solution that scales, control it and then improve it on scale

Product architecture

Quality control in action

Load

Prep

Execute

Save

Quality Control Indicator (QPI)*

- input data is correct

- process runs correctly

- module does the right thing

Monitoring

Rule Engine

Takeaways

- R is ready to be used in business application

- BUT it is not enough to create a bullet-proof product

- Data scientists need to work closely with software and data engineers in order to ensure the delivery of successful solutions

Q&A

Thank you!

Andrea Liotta, CMAF Workshop, 29th April 2019