QUANTUP

How R helps us deliver Machine Learning projects

Examples from QuantUp's past and present.

Introduction

QUANTUP

- Examples of projects
- Useful R features / packages
- Conclusions

QUANTUP

Debt portfolio pricing

Development of debt pricing models for a collection agency

About the project

QUANTUP

- Debt collection agency buys debts
- They need to be priced to make an offer
- Pricing is based on an accurate prediction of recoveries
- Usually, a kind of reference population is used

Details

- Pricing must be accurate
- Analysis of representativeness of reference populations
- Measuring uncertainty
- Flexibility very important
- „Almost real time pricing” to allow analysts to think and make the final touch
- Custom goal functions

Rating for a rating agency

Development of rating models for one of national rating agencies

About the project

QUANTUP

- One of national rating agencies

- Building of quantitative models (including some ML)

- Five parties involved: client, data provider 1 (predictors), data provider 2 (target), QuantUp, software company

Details

QUANTUP

- Non-standard techniques applied, including:
  - Multidimensional optimization
  - Approximation
- Iterating over data generations: reproducibility!
- Generation of documentation
- Reproducibility for audit purposes

QUANTUP

Bacteria species recognition

Recognition of bacteria species basing on spectra of colonies

About the project

QUANTUP

- A sample (urine, saliva, blood, …) is taken
- Bacteria colonies grow on a dish
- Their spectra are generated using an optical device and pictures are taken
- Basing on these pictures bacteria species are recognized

Details

- Image processing with R to keep the single tool (!)

- Documentation generation and rapid reporting

- Prototype applications

- Hardware / server software integration

- Some form of R&D project

QUANTUP

R features & packages

What features and packages of R help us?

What's most important

QUANTUP

- Fast prototyping: everything relies on data
- Early integration: integration can be long
- Building complete products from the start: risk reduction
- Presentations for clients (including interactive prototypes): earning trust is the key + engaging them
- Fast and simple reporting: usually there are standard working reports but it is a lot of ad-hoc analyses
- Easy upgrade of documentation: there are constantly changing data
- Easy and fast experiments: for harder projects we test a lot of approaches
- Short computing times: not always

R features & packages 1

- RStudio. A good IDE for Data Science.

- `knitr` /Rmarkdown. Just simple. Excellent.

- `shiny`, whose API is enabled by pass-by-expression semantics.

- `ggplot2` for simple and good visualization

R features & packages 2

- Tons of classical machine learning algorithms. `caret` supports 238 classes of models while including many popular preprocessing methods and validation schemes.
- Best support for classical statistics. Because sometimes what you really need is a well-executed hypothesis test, not a prediction model.
- Parallel processing: `foreach` is really easy to use
- Rcpp. Extremely easy to use compared to alternatives in other very high-level languages. Case study: needing to implement a custom variant of a Levenstein distance literaly took ~10 minutes to speed up 10-fold.
- Hadley. He's a smart guy.

QUANTUP

Conclusions

What features and packages of R help us?

Conclusions

QUANTUP

- Project execution = software development
- Good tools, libraries, simple code & good processes needed
- Take an advantage of programmatic approach over point-and-click
- One-click generation of everything, including working reports and documentation
  - Data constantly changing
  - Requirements and directions changing (often projects close to R&D)
- Usually preferred R over Python
  - R was designed with data analytics in mind and many tasks are much easier (less code)
  - We use R even for simpler Deep Learning projects because data preparation part is always big
  - *I'm not going to discuss in details what is better: R or Python*

Final conclusions

- All of the above make R put less cognitive load on a data scientist,

- freeing his / her brain for thinking on actual data analysis.

- The processes are important!

Member of

iT CORNER

QUANTUP

Contact

Artur Suchwałko, PhD
m: +48 506 564 841
e: artur@quantup.pl
w: quantup.pl