AGILE DATA SCIENCE WITH R

eRum 2020

Edwin Thoen

A RECIPE FOR DISASTER

Ingredients:

- *1 statistics MSc, with no software engineering skills*
- *1 consulting firm*
- *1 customer site*
- *A bunch of project stakeholders*
- *1 ill-defined project*
- *A handful of undocumented databases*

A RECIPE FOR DISASTER

Preparations:

1. *Use the consulting firm to sell the statistics MSc as an experienced data scientist.*
2. *Place the "data scientist" at the customer site.*
3. *Have the stakeholders convincingly explain the project and set a deadline in the far future.*
4. *Be sure the "data scientist" starts trying a great number of models and algorithms.*
5. *Let gradually simmer for a couple of months.*
6. *As soon as the project implodes you know it is done.*

AI Sponsored

Why do 87% of data science projects never make it into production?

VB STAFF JULY 19, 2019 4:10 AM

Failure rates for analytics, AI, and big data projects = 85% – yikes!

July 23, 2019 by Brian T. O'Neill

Why most big data analytics projects fail

Executive Edge: Why data science projects fail

Delivering brilliant ideas is the easy part; execution is often where things go awry.

TechRepublic. SEARCH 5G Developer

85% of big data projects fail

Nearly all big data projects end up in failure, despite all the mature technology available. Here's how to make big data efforts actually succeed

WHOAMI

Data Scientist @ funda

(main site for real estate in NL)

~7 years applied experience

Mainly focuses on quick research and large ML projects

Requirements
Design
Implementation
Verification
Maintenance

We are uncovering better ways of developing
Software by doing it and helping others do it
Through this work we have come to value:
Individuals and interactions
over Process and tools.
Working Software
over Comprehensive documentation.
Customer Collaboration
over Contract negotiations.
Responding to Change
over Following a Plan
That is, while there is value in the items on the right
we value the items on the left more.

edwinth.github.io/ADSwR

ESSENCE OF AGILE (IMO)

Deliver working software early, let users interact with it and then continuously improve.

from David Robinson's keynote @ rstudio::conf 2019

DELIVER EARLY ?!

But, but R squared is only 0.39…

But, I still have to make six more tabs in my Shiny….

But, my third data source is not yet connected …

But, my AUC is only 0.67…

But, I don't want to deliver a low-quality product!

DELIVER EARLY ?

product quality

product completeness

DELIVER EARLY ?

DELIVER EARLY !

DELIVER EARLY - WHY ?

- remain interest of your stakeholder / customer
- find out quickly if you are solving the right problem
- avoid weaving a web of complexity

THE DATA SCIENCE CHALLENGE

A TWO-WORLD WORKFLOW

- A software product
- Well-organised
- High quality code
- Complete reproducibility

- Testing hypotheses
- Get insight
- Speed, speed, speed
- Low reproducibility is ok

“Ok, but we are at an R conference here …”

–This talk’s audience

R PACKAGE STRUCTURE

R PACKAGE STRUCTURE

RMARKDOWN

```
---
title: "Test some hypothesis"
author: "Edwin Thoen"
date: "6/14/2020"
output: html_document
---

```{r setup, include=FALSE}
knitr::opts_chunk$set(echo = TRUE)
devtools::load_all(.)
```

Taking the log

We currently have a model with 17 predictors and an R sq of 0.57.
Here we will check if we can improve the model with the existing
features, by taking their log.
```

WRAP-UP

- Early delivery and continuous updating lowers the risk of failure
- Try not to blend in data science research with the product
- R package structure is a great way to work Agile

THANK YOU

Grazie mille Milano!

@edwin_thoen

edwinthoen@gmail.com

edwinth.github.io/ADSwR