Reflective Report

Gabriel Medina Galicia

2022-11-17

Introduction

In last years, data science projects have been growing to an exponential level in the industry beyond the research and development, with models in action in online recommendation and detection applications for e-commerce, security, health, finance, etc. As such, projects need to scale to industrial level where changes are fast paced and software components need to be maintainable and scalable.

Methodology CRISP-DM

In software projects, we have been using different methodologies such as waterfall, agile (SCRUM, Kanban), etc. being the agile ones the most implemented in the industry. However, they are not tailored for data mining specific problems and do not address problems that an actual data process model does.

CRISP-DM gracefully address those concerns by clearly separating stages and defining what should be done on each of them. Business Understanding, in my personal opinion, is the most important one, as this is were the goals of the customers are presented, plus, all the terminology and concepts related to an area of study are defined. If this is not well understood, we would be doing the wrong things and wasting our time. But, in case things are wrong, this framework include the concept of feedback and cycles, which allow the customer and contractors to adapt to a fast-paced changing environment.

In my previous job as a software engineer, I had the oportunity to work in projects that involved data scientists on the teams, where they had their own cycles to understand data, produce analysis and defined calculations to be implemented. Although they do not were Machine Learning models, they were internally working following a subset of the CRISP-DM methodology, which was in fact combined with the agile methodology we used in the company. Most of the tasks of the business and data phases were part of the backlog of our SCRUM sprints. In the end, the deployment phase of CRISP-DM was somehow translated to the development of software components, in this case, backend APIs and data visualisation web components.

Other aspects of CRISP-DM, such as the inventory of resources and more detailed descriptions of the business understading items, were managed as part of the company internal software methodology to define resources and teams.

Based on my experience on this work and my career, I strongly belive that as a data involved person (engineer, scientist, analyst, architect, etc), it is fundamental to be acquainted with CRISP-DM to be able to lead, delegate and take control of tasks generated by data projects.

R Tools

Industry level programming environments rely on different sorts of build tools in order to produce deliverables and to set conventions of how to organised assets. ProjectTemplate fulfills this role by providing us an structure to organise our data analysis and report generation, which in fact involve code units that need to adapt to customers changes. Separation of concerns allow us easily spot where to make changes, for example, a new person joins the team and need to perform changes in the code or in the report itself.

I personally like the fact that ProjectTemplate has unit testing in mind, so that we can have some degree of confidence when developing an analysis and teams have to make changes in code. If someone submit changes

in code components, an automated CI/CD pipeline would execute these verification and will not allow any deliverable to be deployed.

Regarding tidyverse. It appears to be a robust toolset, that includes libraries that gives R environments the ability to manage datasets easily as they can be managed in SQL and other data streams libraries. Being familiar with both, it definitely boosted my productivity when it comes to manipulate data and generate results for the exploratory data analysis. Although, I had some issues when dealing with tibbles and functions that require data.frames explicitly, as I feel that at some point, the type of different datasets is juggling between tibbles and data.frame, which brakes the consistency in your code.

Finally, I consider ggplot is a very useful library that providing R with a a more defined and extensive toolset, comparable to matplotlib and seaborn in python.