The Data Science Method (DSM)

Getting started doesn't have to mean going around in circles.

Guy Maskall Jul 15, 2020 · 4 min read ★

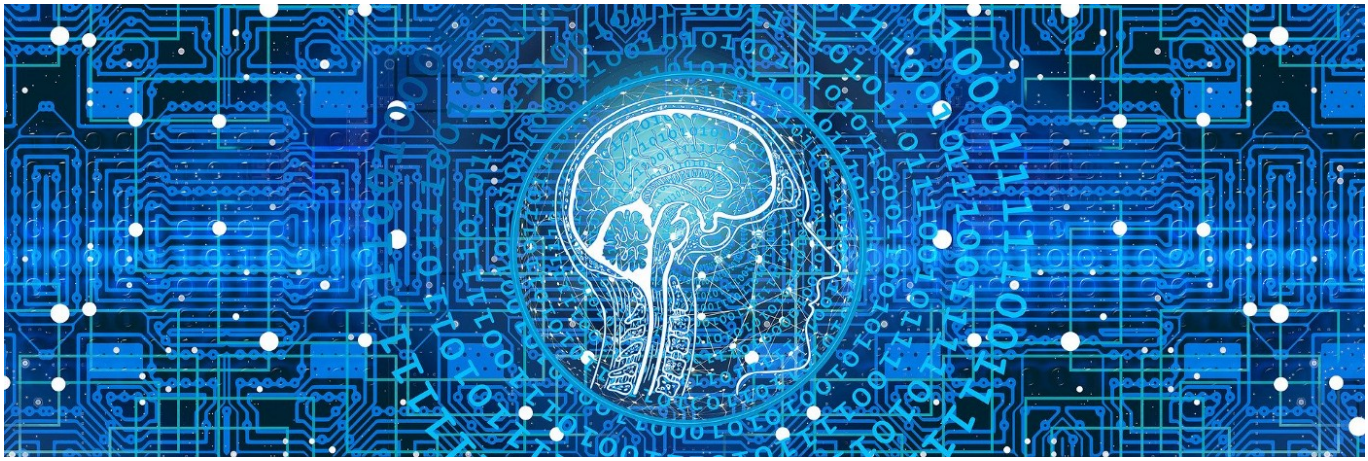

Image by Gerd Altmann from Pixabay

Introduction

This is the start of a short series of articles on the data science method. They are designed, specifically, to complement Springboard's Data Science Career Track course. Writing about the method is a hard task, because there is no such single thing as *the* data science method. What I mean is that there is no process you can robotically follow that is guaranteed to yield ideal results. There are, however, guiding principles that will steer you in the right direction. Think of them together, if you will, as forming a North Star for you to follow.

CRISP-DM

A worthy industry standard process is the cross-industry process for data mining (CRISP-DM). There are many articles describing this process, so I won't go into it in detail here. Broadly, CRISP-DM recognizes some six stages:

1. Business understanding
2. Data understanding
3. Data preparation
4. Modelling
5. Evaluation
6. Deployment

Business understanding

Yes, we want to gain a clear understanding of the business context. How else could we know what problem we're really tackling and what a usable solution might look like?

Data understanding

Yes, we want to understand the data. Does it seem useful for addressing the business problem? Do we trust it (or can we ring fence our concerns over it, at least)?

Data preparation

We will inevitably need to prepare the data, because data are rarely gathered for the purpose of any analysis or the convenience of the data scientist.

Modelling

After all that, we can finally move on to the more interesting modelling stage. This does not necessarily mean machine learning! It may be sufficient to gain a thorough understanding of certain relationships within the data; that alone may be the end goal of "modelling" for a project. Yes, often it will involve machine learning, but a vital precursor to any machine learning work is generating insight into patterns in the data via exploratory data analysis.

Evaluation

Then we want to evaluate how good our understanding, or our model, is. Is it good enough to meet the needs of the business? Are we confident enough to present some clear conclusions? Does a model offer an advantage over the previous solution?

Deployment

If our data product is good enough, then we'll want to "deploy" it. In some cases, deploying the product may mean passing on the knowledge gained. This needs to get to the appropriate recipient in an appropriately digestible format. In other cases, deployment may mean passing a machine learning model into a production process where it will generate predictions that drive autonomous business processes.

Springboard's Data Science Method

But if there are already industry standard processes for doing data science, why define a new one? Well, to start with, CRISP-DM is not the only game in town for data science project management. And, as this Smart Vision article highlights, CRISP-DM

> *is an idealised sequence of events. In practice many of the tasks can be performed in a different order and it will often be necessary to backtrack to previous tasks and repeat certain actions. The model does not try to capture all possible routes through the data mining process.*

This can present substantial challenges to entrants to the data science field, not to mention teachers. Do you always start at the same point? When do you backtrack? How far do you backtrack? How many times do you iterate until you declare the entire endeavour "done"?

With this in mind, we can strip the process down to a more linear path that largely ducks the nuances of an iterative back and forth workflow but still keeps faith with the scientific method. We'll just point out here that you should be constantly re-evaluating your assumptions and never be afraid to revisit earlier steps in light of lessons learned.

And so here we outline a sequence that inevitably makes some assumptions and simplifications, but a sequence that you, with experience, will be able to build upon and bend to your needs.

1. Problem identification

2. Data wrangling
3. Exploratory data analysis
4. Pre-processing and training data
5. Modelling
6. Documentation

Hopefully you will see a broad agreement of the above with the six CRISP-DM steps. As we've discussed, the real process inevitably involves cycles of iteration. Much of the differences between these steps and those of CRISP-DM lie in where you apply the cuts in the process and how you label them.

Problem identification

You want to adequately identify the real problem to be solved, otherwise you're heading off in the wrong direction from the start! This is a direct correspondence with the CRISP-DM step.

Data wrangling

Data wrangling is a broad term that easily covers the acquisition and any initial bashing of data into shape, as well as the ongoing transformations of the data to suit. To understand the data, you do first need to acquire it, quite likely wresting it from some original source or sources. This step crosses the boundaries of CRISP-DM's *data understanding* and *data preparation*. Indeed, data wrangling is arguably an activity that persists throughout the entire process.

Exploratory data analysis

This has a strong connection to CRISP-DM's *data understanding* step. We put it here because you generally have to wrangle data to really explore it and, thus, understand it.

Pre-processing and training data

Pre-processing and training data development can also be regarded as part of *data preparation* but yet also form an aspect of *modelling*. Because the pre-processing can be so inextricably linked to the modelling, there's inevitably considerable crossover here with the CRISP-DM *modelling* step.

Modelling

CRISP-DM has *evaluation* as a distinct step, but it is arguably an implicit part of any model development effort.

Documentation

Finally, CRISP-DM has *deployment* as the ultimate step. Here we focus on documentation as a deployment artifact. An entire course curriculum can be dedicated to putting machine learning models into production…

Data Science