Project Methodology

Time spent/should spend: 4 hour

Method considered

- CRISP-DM
- SCRUM
- Kaban
- Domino Data Science Life Cycle
- Extreme Programming
- KDD

There is a lot of different method for data science projects. These 6 models are selected for a general review based on their popularity and relevance to data science projects. To choose a method that is suitable to my need, the following criteria are listed and considered.

Criteria:

- Experience:
    - I have no experience in software engineering and data science. So the model needs to be **easy to learn**, **well-documented** and **versatile**.
    - I want to accumulate experience for future project. So I want the model to be **widely used**.
- Goal:
    - My goal for the project is to gain some insight about sustainability using the dataset provide. The goal at this point is vague.
        - I want a model that **work well with vague goal.**
        - I want to **discover insight** from a large set of data
- Team size:
    - I am working as an individual, however, due to the nature of engineering, it is likely I will work in group later in my career.
        - The model needs to be **suitable for individual** and **can be modified** to suit group work.

Combining these considerations. I have decided to choose an Agile method, since waterfall method generally requires a understanding and clearer goal when one starts a project.

Decision Matrix

Based on this paper written by by Khan and Beg [1], I have decided to use a decision matrix to aid the decision making process. The matrix they introduced is based on qualitative descriptor and factor. I have decided to base my matrix on their's but combine it with my own criteria and give each method a quantitative score on how well I think it fit my criteria.

I also considered giving each criteria a weighing factor, to calculate a weighed score for each method. However, I soon realise all the certeria weight the same in my judgement therefore the weighing factor is scraped. The matrix they developed are attached below.

Sr. No.	Evaluation Criteria	Traditional SDLC-Models			Agile SDLC-Models	
		V-Process	*2-I (Incremental / Iterative)*	*Water-fall*	*XP*	*RUP*
1.	Business Criticality	High	Medium	Low	Low	High
2.	Customer Involvement through Life Cycle	Low	Medium	Low	High	Low
3.	Requirements Clarity	High	Medium	High	Low	Medium
4.	Requirements Volatility	Low	High	Low	High	Medium
5.	Availability of business users	Medium	High (& tapers to Medium)	Low	Medium	High (& tapers to Medium)
6.	Project Size	Large	Large	Small	Medium	Large
7.	Complexity (Business, Technical)	Low, Medium	High, High	Low, Low	Medium, High	High, Medium

Figure 1: Decision Matrix by Khan and Beg

In the process of researching, I have decided to replace KDD with SEMMA since they are similar and SEMMA is much more updated.

	CRISP-DM[2]	SCRUM [3][4]	Kaban [5]	Domino [6]	Extreme Programming [7]	SEMMA[8]
Type	agile/waterfall	Agile	Agile	Agile	Agile	life cycle
Complexity	2	2	5	3	4	3
Versatility	3	3	4	3	2	2
Popularity	5	5	4	3	4	2
Requirement Clarity	3	2	4	2	3	4
Team size	4	3	2	3	3	4
Comment	Documentation heavy	Too team-oriented	Prone to procastination	Problem first not data first	Very software engineering	Heavily Data mining oriented
Total	**17.00**	**15.00**	**19.00**	**14.00**	**16.00**	**15.00**

Complexity: Amount of managerial work. Easiness to Learn, the easier the higher the score

Versatility: Our project is both software engineering and data science in nature. The better the method is for both purpose, the higher the score.

Popularity: How commonly used the method is in industry.

Requirement Clarity: Do we have to have a clear goal in mind at the start.

Team size: Is the method suitable for individual? Can it be used for team as well

Using the matrix, I have decided to use CRISP-DM with Kanban. The reason why is although Kanban get the highest score, I feel that it is not suitable for me since I am prone to procrastination. And it was more a project management method than a life cycle. The clearly set task in CRISP-DM would help with procrastination and give me a clearer sense of direction.

Explanation

Now that I have chosen a method, I would like to explain further on it and how I can modify it to suit my project.

CRISP-DM [2]

- **CR**oss **I**ndustry **S**tandard **P**rocess for **D**ata **M**ining
- 6 phase
    - **Business understanding** – What does the business need?
    - **Data understanding** – What data do we have / need? Is it clean?
    - **Data preparation** – How do we organize the data for modeling?
    - **Modelling** – What modeling techniques should we apply?
    - **Evaluation** – Which model best meets the business objectives?
    - **Deployment** – How do stakeholders access the results?

The main advantage of this model is the popularity and it's agility. A popular and widely used model means there is a higher chance I would be able to apply it again in the future. And an agile approach would allow me to move back and forth in phases which is natural for learning processes.

As mentioned in my comment in the decision matrix, I noticed that this will be documentation heavy since 4 out of 6 of the phase are documentation heavy. Since I have no experience in software engineering or data science, and the fact that this is a coursework, I consider this as an advantage. Clearer documentation means more evidence for my work and force me to have a clearer understanding on my work. I noticed in the blog from data-science alliance[2] that students who use Crisp-dm are the most likely to start coding late. Which have been true in this coursework.

Another downside to choosing this methodology is that it is more of a data-science method than software engineering method. This might become a problem later in the year. However, after reviewing another software-engineering oriented method, extreme programming (XP), I find this method a lot more suitable and well-documentated.

Apart from the upside, I have managed the downside by limiting how many hours I spend on writing up. This is necessary because I have been working on this coursework simultaneously with my third-year design project, this limitation helps me to balance workloads.

Kanban: [5]

- **Billboard/Trello**
- **To do; WIP; Done**

This method is much more suitable for group as it allow the assignment of workload and allow everyone in the group to follow the progress. That doesn't means it is not suitable for individual as well.

As noted in my comment, I see Kanban more as a project management method than a software engineering/data science lifecycle. I have used a online billboard called Trello before. And for this project, I have created a Trello board and managing my task on the there. Task from different project/modules are colour code and longer task are split-up using to do list. Evidence can be found in tools and technique mark down file.

It have been working well. I must agree with comments that it add to managerial work [5], however, I am finding that to be minimal and it helps with stress management. Moving task to the Done column comes with sense of achievement. I have been feeling that it helps with procrastination.

Bibliography

[1]
P. M. Khan and M. M. S. S. Beg, "Extended decision support matrix for selection of SDLC-Models on traditional and agile software development projects," 2013, pp. 8–15. doi: 10.1109/ACCT.2013.12.
[2]
N. Hotz, "CRISP-DM," *Data Science Project Management*, 2021. https://www.datascience-pm.com/crisp-dm-2/
[3]
N. HOTZ, "Scrum for Data Science," *Data Science Project Management*, Jun. 06, 2022. https://www.datascience-pm.com/scrum/
[4]
I. Godfried, "Why Scrum is awful for data science," *Medium*, Aug. 28, 2020. https://towardsdatascience.com/why-scrum-is-awful-for-data-science-db3e5c1bb3b4 (accessed Nov. 04, 2022).
[5]
N. Hotz, "Kanban," *Data Science Process Alliance*, Aug. 01, 2022. https://www.datascience-pm.com/kanban/
[6]
N. Hotz, "Domino Data Science Life Cycle," *Data Science Process Alliance*, Aug. 01, 2022. https://www.datascience-pm.com/domino-data-science-life-cycle/
[7]

Lucid Content Team, “What Is the Extreme Programming Methodology? | Lucidchart Blog,” *Lucidchart.com*, Aug. 10, 2018. https://www.lucidchart.com/blog/what-is-extreme-programming
[8]
N. Hotz, “SEMMA,” *Data Science Process Alliance*, May 14, 2021. https://www.datascience-pm.com/semma/