ML Frameworks for Prioritising AI Initiatives

You want to embrace machine learning in your business. How do you get started?

Version: 0.3 Dated: 2024-03-08 Authored by: CM

colum.mccoole@analect.com
colum-mccoole-746b946a
analect/technical-docs-hierarchy

Docs: > Level 2 > AI/ML Value-chain

Topic Navigation

What Problem Are We Solving For?

You're a PE firm with a strong track-record in investing with deep specialist knowledge across multiple industries. You are not quite sure how AI will affect these industries and how your holdings are best placed to navigate the inevitable changes this new technology will bring. You are keen to embrace these new technologies and weave them into your investment process, where do you start?

Identifying Low-hanging Fruit

Portfolio companies may already have well documented/mapped technology stacks (including data warehouse/lakehouse) where perhaps ML systems are embedded, and maybe those systems can be re-imagined in a cloud context, where new tooling (such as generative AI) can be used to enrich data further and unearth new insights. Maybe other companies are less sophisticated and may lack such legacy systems, or maybe there are startups, navigating this space for the first time.

These two resources (laid-out as guides and question-sets) are great for fleshing-out the problem-space. They are:

- **https://medium.com/data-science-at-microsoft/setting-up-machine-learning-projects-for-success-4cba7840d24a**
- **https://alan-turing-institute.github.io/rds-course/**

Setting up Machine Learning projects for success

Problem framing

1. Define the objective in business terms.
2. How will the solution be used?
3. What are the current solutions/workarounds (if any)? What work has been done in this area so far? Does this solution need to fit into an existing system?
4. How should performance be measured?
5. Is the performance measure aligned with the business objective?
6. What would be the minimum performance needed to reach the business objective?
7. Are there any known constraints that would have to be taken into account? (e.g., computation times, non-functional requirements)
8. Frame this problem (supervised/unsupervised, online/offline, etc.).
9. Is human expertise available?
10. How would you solve the problem manually?
11. Are there any restrictions on the type of approaches that can be used? (e.g., does the solution need to be completely explainable?)
12. List the assumptions you or others have made so far. Verify these assumptions if possible.
13. Define some initial hypothesis statements to be explored.
14. Highlight and discuss any responsible AI concerns if appropriate.

Data exploration

1. Understand and document the features, location, and availability of the data.
2. What order of magnitude is the current data (e.g., GB, TB)? Is this all relevant?
3. How does the organization decide when to collect additional data or purchase external data? Are there any examples of this?
4. What data has been used so far to analyse recent data-driven projects? What has been found to be most useful? What was not useful? How was this judged?
5. What additional internal data may provide insights useful for data-driven decision making for proposed projects? What external data could be useful?
6. What are the possible constraints or challenges in accessing or incorporating this data?
7. How was the data collected? Are there any obvious biases because of how the data was collected?
8. What changes to data collection, coding, integration, and so on, have occurred in the last two years that may impact the interpretation or availability of the collected data?

Workflow

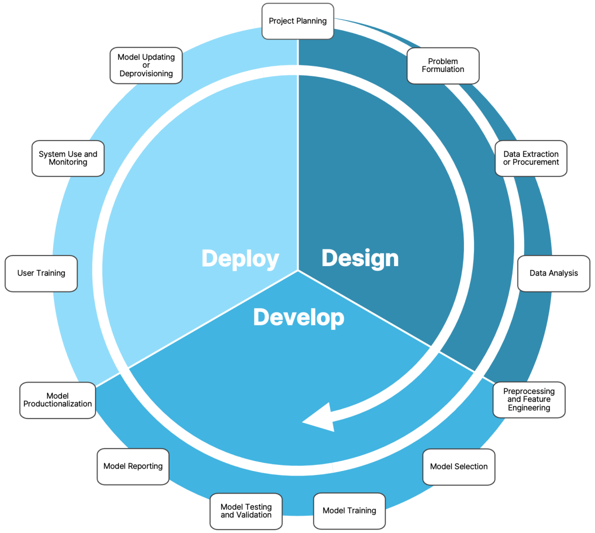
1. What data science skills exist in the organisation?
2. How many data scientists/engineers would be available to work on this project? In what capacity would these resources be available (full-time, part-time, etc.)?
3. What does the team's current workflow practices look like? Do they work on the cloud/on-prem? In notebooks/IDE? Is version control used?
4. How are data, experiments, and models currently tracked?
5. Does the team employ an Agile methodology? How is work tracked?
6. Are there any ML solutions currently running in production? Who is responsible for maintaining these solutions?
7. Who would be responsible for maintaining a solution produced during this project?
8. Are there any restrictions on tooling that must/cannot be used?

Framing an AI Strategy as a Data Science Project

The list of questions below comes from a **Research Data Science** project lifecycle, a course taught at the Alan Turing Institute, available at https://alan-turing-institute.github.io/rds-course/. It is useful as a starting point with stakeholders when scoping a project.

- **Question 1:** What is the broad problem we are trying to solve?
- **Question 2:** What is the specific research question? How does it translate to a data science problem?
- **Question 3:** Is data available and appropriate?
- **Question 4:** What are the stakeholders' expectations?
- **Question 5:** How does the output product look like and how is it going to be used?
- **Question 6:** What is the state of the art?
- **Question 7:** What is in-scope and out-of-scope?
- **Question 8:** What is the expected impact?
- **Question 9:** How do we measure the success of the project?
- **Question 10:** How do we monitor and explain output?
- **Question 11:** What computational resources are available?
- **Question 12:** How is the product going to be deployed/maintained?
- **Question 13:** What are the timelines and milestones of the project?
- **Question 14:** Ethical considerations

A typical project lifecycle can be seen in the figure below. It contains three stages - Design, Develop and Deploy. These stages initially seem sequential. But it is almost always the case that the process is iterative.


Project Planning
Problem Formulation
Data Extraction or Procurement
Data Analysis
Preprocessing and Feature Engineering
Model Selection
Model Training
Model Testing and Validation
Model Reporting
Model Productionalization
User Training
System Use and Monitoring
Model Updating or Deprovisioning
Deploy
Design
Develop