The Four Phases of a Data Science Project

Justus Perlwitz

2018-03-02

Contents

Imagine this: You're pitching an idea for an interesting data science problem that you can solve for your client. The client is sold on the idea and wants to immediately know how fast you can get it done, and more importantly, what the project milestones will look like.

The first observation that I have made is that packaging a project into neat little pieces makes adjustments and time estimates much easier. Typical adjustments in a project are

- Changing the scope of the project or the way it is implemented due to technical reasons
- Changing focus because a more interesting problem to solve has been identified
- Simplifying the question asked to solve, e.g. *predict a person's height at age X*, instead of *predict a person's height throughout their life.*

Adjustments are inevitable and will improve the outcome of a project. The changes that are necessary become much more obvious if the project has been divided into small chunks. And of course, when breaking down work into piecemeal milestone estimating how long each step will take becomes much easier.

Roughly speaking, a project can be divided in four phases. Each phase serves a specific purpose, as will be explained later. Individual phases can be revisited throughout the project as it changes. Knowing in which phase you are can help you decide easily what to do next. It's important to know whether you are done with the work required for the current phase and whether you can move on. Sometimes the current phase requires more work. Understanding this only comes with experience and this knowledge can only be based on a lot of trial-and-error. The four phases for data science projects are:

- **Strategy**: Understanding what the client needs
- **Exploration**: Knowing what can and can't be done
- **Development**: Creating a shippable solution
- **Maintenance**: Maintaining the shipped solution

I will now describe each phase in further detail.

1 Strategy (Phase 1)

The **strategy** phase is about getting to know what your client does, what his business problems are that you can specifically solve and whether they can be solved by you. This will typically last 1 week. In this phase, the following activities will take place:

- Sitting down together with the client and understanding their business
- Interviewing domain experts

- Identifying problems that can be optimized using AI/ML techniques
- Identifying business value to ensure that "value" is being added by AI/ML
- Understanding the current data warehousing solution that the client has in place
- Determining benchmark, performance criteria for a good problem solution, and seeing if the client already has established a baseline for good solutions
- Verifying data source quality, get access to everything needed (S3 buckets, SQL database, and so on)

After this phase, you should have a very good understanding of the client's business. You should exactly know what problem you can solve for them and how a best case solution will look like. Ideally, you can also estimate how much value you are creating for your client. Roughly speaking, you want to look at how much you can increase the efficiency or their business, or by how much you decrease inefficiency. For example, this can be an improvement in their conversion rate, revenue, or productivity loss. At this point, you can already tell your client what you are trying to solve and what the expected outcome of the project will be.

2 Exploration (Phase 2)

After having established what particular problem you want to solve, the feasibility of solving it needs to be understood further in the **exploration** phase. This can take up to two weeks, as techniques and methods need to be evaluated and presented to the client so that a sensible choice for developing the actual solution can be made. In this phase a data source that can help solve the problem and create training data should already have been found. The phase will typically last 2 weeks. The following activities will take place:

- Creating data extraction and processing workflows
- Using off the shelf ML tools (scikit-learn, xgboost, etc.), try to get within 90% of agreed upon benchmark or baseline
- Exploring 2-3 different machine learning models that are appropriate for solving the problem at hand to have more options
- Coming up with 1 candidate model that performs very well and presenting it to the client
- Defining performance goal for next phase

After this phase, you can tell your client exactly how well you can solve their problem and how the problem solution will be implemented. You can already demo a small application or share a notebook with them that can give them a feel of how the end solution can look like.

3 Engineering (Phase 3)

The **engineering** phase is dedicated to the nitty-gritty of software development. After having established that you are working on the right solution, you will now concentrate your efforts on making this solution as good as possible. This can range from simple API integration work to performing a grid search to find the best parameters for a machine learning model. Typically, this phase can last from 1 week to up to 1 month. The activities in this phase vary a lot from project to project but typically involve some of these activities:

- Going from 90 % prediction performance to 99 % prediction performance
- Creating model creation or training batch jobs for daily re-evaluation on new client data
- Creating an API or a prediction service that can be used by other services
- Creating a dashboard

After this, the project is usable by the client and can start bringing real value into their company.

4 Maintenance (Phase 4)

The maintenance phase will follow any solution that needs to adjust to new demands and requirements. Quite often, client requirements evolve over time and the original solution will be less and less appropriate for solving the current problem. That's why being able to provide maintenance over a long time is essential. The duration of this phase is open-ended, but can last as long as the solution you have developed is being used by the client.

Some of the activities include:

- Ensuring ML/AI solution works as client migrates and changes data schemata
- Keeping up good performance when nature of data changes
- Adding or enhancing features as client requirements change

5 Summary

Dividing data science projects into four distinct phases better reasoning and understanding how a project should be organized. They give the client well-deserved peace of mind as they make a projects timeline transparent. I would be excited to learn how you structure your typical project and what has worked well for you.