The Data Science Method — Problem Identification

Aiden V Johnson Follow
Dec 10, 2019 · 4 min read ★

In the previous article, we presented the Data Science Method (DSM) as a structured thinking approach to solving data science problems. Next, we will address the first and most important step, problem identification, in more detail. The additional DSM steps will be described in detail in future articles.

The Data Science Method

1. Problem Identification (this article)

2. Data Wrangling

3. Exploratory Data Analysis

4. Pre-processing and Training Data Development

5. Modeling

6. Documentation

1. Problem Identification

Problem identification is the very first, essential step to a well-positioned data science project.

Start by identifying the goal of the data science project. Ask the question: Is this an exploratory project or a predictive modeling project? If the answer is exploratory, then less planning may be needed at the outset to ensure interesting and meaningful outcomes. You might have questions about how you can tell if a project is exploratory or predictive, so let's work through some examples. You may be given a data set for a project and asked questions such as:

- Process the data — what are the important findings you can glean?
- What can you tell me about sales in the last year?
- What type of customers do we have?

All of the above three questions indicate that you are working on an exploratory data project — you're not explicitly predicting any response variable to apply to a future dataset. For the first question, you have the potential to spend countless days looking at the data a thousand different ways. In order to apply some necessary bounds to the analysis, you can reframe the open-ended question into a few more specific questions that are actionable using SMART principles. The other two questions are equally difficult to answer without following structured thinking and framing the context, criteria for success, and stakeholders. It helps to identify what the expected use of the final product is. For an exploratory project, try to hypothesize the kind of findings that are of value *before* you get started. Let's work on rephrasing the three questions above to be actionable.

Original question: Process the data and tell us what important findings you can glean.

Revised question: What are the summary statistics of this data set and what do we know about the context of the data that we can investigate further for business impact?

Original question: What can you tell me about sales in the last year?

Revised question: What is the most common product we sell, and how much did we sell every quarter over the last year?

Original question: What type of customers do we have?

Revised question: What are the average ages, incomes, and home locations of our customers?

As you look at the differences between the original question and the revised question, you can hopefully see that the revised questions are now problem statements that you can use data science analysis to answer. Developing revised questions might take some effort on your part — you might need to return to the stakeholders for feedback before fully identifying the problem and the core focus of the analysis. Be sure to ask yourself if the data you have access to supports the question you're trying to answer. If it doesn't, ask!!

If your goal is to evaluate the variable correlations and multi-dimensional interactions of the data set, then the initial motivations of the data science project must be more firmly defined.

Outlined below is a step-by-step approach to Problem Identification, the first step in the DSM. Defining each one of these bullets at the project outset will guide your project to a fruitful outcome.

Problem Identification Steps:

1. Problem statement formation
2. Context
3. Criteria for success
4. Scope of solution space
5. Constraints
6. Stakeholders
7. Data sources

Here is a list of general questions to help you get started in defining the above-listed steps for problem identification.

- Is the goal of this project exploratory or predictive?
- Identify what the completed model will be used for and/or the expected outcome of the exploratory work — consider supervised or unsupervised methods.
- Does the data you have access to answer #2 above, or do you need more or different data?
- What is the data timeline and/or temporal scale of interest?
- What is the modeling response variable? How is it described and defined?
- Is this a classification or regression problem?
- What deliverables will be provided after this modeling project?

As you develop answers to these questions and the steps outlined in problem identification, you will not only gain a focused trajectory of work, but you also will get at the key details needed for model documentation. Further, you will connect your data analysis to a business need, which may have motivated the work in the first place. If you clearly define your data science work, you will have a framework for successful implementation that works within any industry.

To receive updates about DSM sign up here.

Data Science Project Management Machine Learning AI Springboard