The Lifecycle of a Data Science Project

Foreword

This document contains a model that describes the lifecycle of a data science project that is intended to solve a real-world problem, such as those found in public policy. Here a data science project is broken down into eight distinct phases, starting from the inception of the project all the way to its' final impact. These phases are problem identification, project scoping, data acquisition, data exploration, analytic formulation, data science analysis and modeling, modeling validation, and finally taking action. In the reading below, you will find detailed information on each of these phases, and learn both practical information about their implementation and other important considerations such as the ethical and moral implications to be taken at that phase of the project.

The following information is an expert from the paper "An Experience-Centered Approach to Training Effective Data Scientist" by Kit T. Rodolfa, Adolfo De Unanue, Matt Gee, and Rayid Ghani. The paper can be found here: https://pdfs.semanticscholar.org/ab1a/4f9a0271f184db25eb304ea8b0d6dc1cc5b9.pdf.

Figure 1 shows the inter-connectedness of the different phases of a typical data science project. These phases are described in more detail below, illustrated through a real-world example of reducing lead poisoning in children.

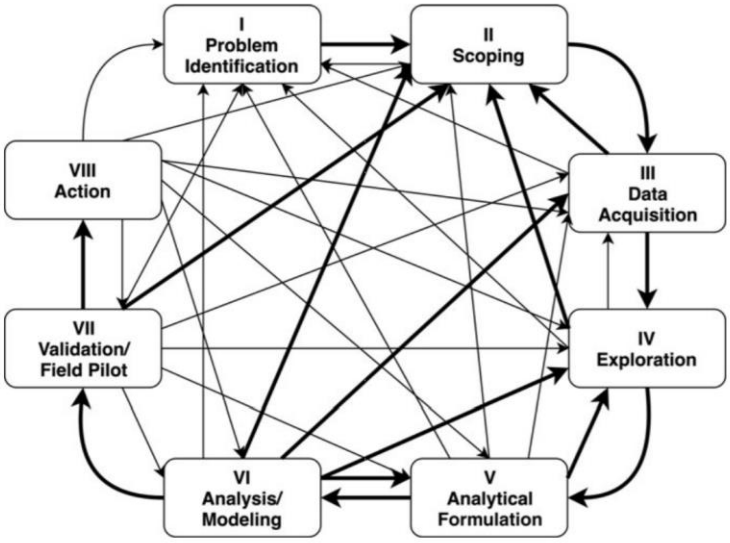

FIG. 1. Lifecycle of a data science project. The phases of real projects are highly interconnected, with insights from later phases informing and revising earlier ones, resembling more of a "web" than a typical linear pipeline or cycle. Bold arrows show more frequent transitions typically repeated multiple times in the course of a project.

Phase I: Problem identification

Even before the outset of a data scientist's work, stakeholders or policy makers may need to make a decision about pursuing a solution to a problem. This involves evaluating whether the problem is significant, whether it is feasible to solve the problem with data science, and whether there is (or will

be) commitment internally to allocate resources to addressing the problem. Data scientists have a critical role in this process, both providing a voice about what is technically feasible and why it may provide improvements over current practices, as well as an ethical duty to highlight the limitations and risks involved. Similarly, a solid understanding of laws and best practices around data privacy and sharing can be essential to helping decisionmakers understand how the data they have can be used and what other data they may or may not be able to collect for the project.

Example: Public health officials identify high rates of lead poisoning in children in their jurisdiction, but current practice only remediates issues in homes after a child has tested positive for elevated blood lead levels. They would like to reduce lead poisoning in children by proactively identifying children who may be at risk before poisoning occurs.

Phase II: Scoping

As a project begins to get underway, competencies in communication and problem definition will be particularly important in scoping the actual work. The data scientist needs to be able to evaluate what questions can be answered with the available data as well as work closely with the stakeholders to understand their needs and how any models and analyses will actually be put into use. Ethical concerns at this stage include considering how sensitive data will be handled and protected as well as establishing criteria by which analysis will be evaluated in ways that balance efficiency, effectiveness, and equity.

Example: A scoping session is held including public health officials, clinicians, lead hazard inspections teams, and data scientists to understand the data available and how risk scores would be put into use. Because of the need to work with private health information and data pertaining to children, the decision is made to restrict all analytical work to the Department of Public Health's secure server environment. Primary intervention is identified as lead hazard inspections in homes with high risk of lead hazards and presence of a child younger than 12 months. The key goal identified in the scoping phase was to effectively reduce childhood lead poisoning in an equitable manner across underserved communities.

Phase III: Data acquisition

Acquiring, storing, linking, understanding, and preparing data for analysis in a real-world project often entail an involved and iterative process, requiring working closely with the owners of various data sources to ensure any transferred data are provided in a consistent and reliable format and necessary steps are taken to protect private or sensitive information. During this phase of work, the data scientist needs to apply skills working with and structuring raw data to get it into a storage format that is appropriate for linking it with other data sources. Each of those steps requires active communication with the project's stakeholders to understand the context in which the data were collected and structured, its idiosyncrasies, and ensure data definitions actually describe the events they are supposed to reflect.

Example: The Department of Public Health provides a database and server for analysis in their environment with an extract of individual-level blood lead test results as well as inspection reports from lead hazard inspections. Data from additional sources are imported into the environment, including census data, childhood nutrition benefit program data (to identify potentially vulnerable children), and information about buildings from the county assessor website. Address normalization and geocoding

allows data to be linked across these sources, and data scientists work closely with the owner of each data source to ensure they understand the data structures and fields.

Phase IV: Exploration

This initial phase of analysis focuses on exploring the trends and relationships in the data through summary statistics, visualization, and preliminary modeling.

Example: The data scientists use a combination of descriptive statistics, bivariate correlations, spatial and temporal analysis to begin to understand the relationships in the data and its limitations. Missing values in the childhood nutrition benefit data set identify an error in the extract, transform, load process that is corrected with a new data extract, while a sharp decrease in the number of blood tests in data older than 17 years reflects a change in policy around testing that defines the limitation in historical training data.

Phase V: Analytical formulation

This phase involves formulating the initial problem as a concrete analytical problem. In most cases, a greater understanding of the available data and its nuances will result in a greater understanding of the problem itself as well. During this phase, a data scientist will need to be able to effectively communicate preliminary results to stakeholders, including any limitations or shortcomings. At the end of this phase, the data scientists and stakeholders will have a set of design decisions to set up the technical framework for the project. From this more well-informed perspective, the project scope can be revisited and modified, which may in turn require more data collection, feature engineering, or exploratory analysis.

Example: Drawing on what they learned in exploring the data, the data scientists work with the public health officials to formulate a classification problem at the address level using blood lead levels above a specific level as a training label. Monthly risk scores will be produced for every house with a child younger than 12 months to correspond with the planning cycle of the department's housing inspection team and evaluated on the basis of precision (positive predictive value) among the top 250 highest risk addresses, consistent with their monthly capacity for lead inspections, as well as the representativeness of underserved communities in the results.

Phase VI: Analysis/modeling

Many projects will move through multiple rounds of exploration and refinement, iteratively approaching a final analytical phase as the problem definition and scope continue to evolve. While the competency domains essential to these later analytical phases are similar to those employed in the earlier exploratory work, the specific skills used here will tend to shift away from data description and more toward summarization, prediction, and/or extracting meaning. Generally, this phase involves generating a large number of models, analyses, or results followed by analysis to draw meaningful conclusions. In the case of predictive modeling, this might involve the process of model selection, balancing different performance and fairness metrics to arrive at a single model (or small menu of choices) to put into practice. For analytical projects, this phase may also involve telling a story from the available data, putting to use not only communication and data visualization skills but also ethical frameworks for how to summarize vast amounts of data in fair and meaningful ways.

Example: The data scientists run a grid of thousands of model specifications, including several families of classifiers and hyperparameters. Based on its ability to both achieve high precision in the top 250 and balance false omission rates across race and socioeconomic status, a random forest model was chosen to test in a field trial.

Phase VII: Field validation/pilot

The previous phase results in a final set of analysis results or models that are ready to be piloted or validated in a field trial. This phase of the project involves designing the trial to test the ongoing effectiveness and usability of the analysis. In some cases, this may involve developing a randomized control trial to measure the causal impact of deploying a predictive model, whereas in others it may involve collecting feedback on how a report impacts decision-making. In any case, this phase should focus on validating that the results of the analysis in fact continue to perform as anticipated when presented with truly novel data, including with respect to relevant fairness metrics. Likewise, gathering qualitative feedback from decision makers acting on the analysis is an important aspect of the field pilot.

Example: A 1-year field trial was developed, during which a random 50% of the 250 highest risk addresses were inspected for the presence of lead each month, and remediated where hazards were found. The trial confirmed the performance of the model in identifying children at risk of poisoning because of the presence of lead in their homes as well as its representativeness across communities.

Phase VIII: Taking action

Finally, for a data science project to successfully impact the decisions or actions of policy makers or stakeholders, results must be clearly and effectively communicated to these (generally nontechnical) audience along with recommended actions or a menu of choices. The ethical obligations of responsible data science practitioners at this stage reach far beyond avoiding the colloquial idea of ‘‘lying with statistics’’ to an awareness of the potential societal impact of their work. Any recommended course of action involves trade-offs (for instance, between optimizing for overall efficiency vs. fairness across affected groups) and the data scientist performing these analyses may be the best-positioned individual to articulate the trade-offs associated with any potential action.

Example: Although the number of households with lead issues remediated was too small to have a significant impact on the number of children diagnosed with lead poisoning during the trial period, calculations suggested that deploying the model could appreciably impact lead poisoning over the following decade. The Department of Public Health decided to move forward with putting it into practice, committing resources to maintain and periodically refresh and reevaluate the model.