HOW TO CREATE A WINNING

PROJECT ANALYSIS PLAN

The Artists of Data Science
with Harpreet Sahota

PROBLEM STATEMENT

Provide a clear description (3 to 5 sentences, maximum) that summarizes your project. **You can look to your problem statement and purpose statement from the scientific method to include here.**

A good way to summarize your project is to start with a sentence that clearly defines the problem, e.g., **"This project will develop a machine learning model to do Y…."**.

You can look to your research question to include here, and state your hypothesis here.

You should follow this with a summary of your planned technical approach, e.g., "**The approach I will take to address this problem is n-fold:(1), use method A to do X (2), use method B to do Y,(3) …..**" (as an example)

Some questions to get you thinking about what to write here:

- What is the data domain?
- What is the goal of your project?
- What is the motivation for rigorous data analytics?
- What are the questions you want to answer?
- Why the analysis is important?
- What are a few potential applications?

TECHNICAL APPROACH

Write 2 or 3 paragraphs with a clear, more detailed description of the methods and algorithms that you plan to use on the project. **You should look at your research design here.**

If the system you are building can be thought of as a pipeline with multiple components, **a useful approach is to provide a figure that illustrates the pipeline** (with blocks for different components) along with brief descriptions of each component (e.g., the names of algorithms or methods you plan to evaluate).

Make sure it is clear what your pipeline or system is doing, i.e., what each component will do in terms of taking inputs and producing outputs.

This section should include a **brief description** of each of the database, machine learning/statistics, and visualization/interface components in your project.

DATA

Briefly describe what data sets you plan to use in the project.

Include **specific references** to the data (e.g., a URL) if you can. If, for example, you are planning to work with Web/text data,

it would be good to **do some preliminary assessment** of how much is available, what the fields/attributes/metadata are, if there are labels for the data (if you are doing supervised machine learning), and so on.

You can change/update your data sets during the project if you need to, but you should have identified at least one data set to work with by the time you start working on this analysis plan.

For data engineering projects, think about: Provide a draft of your overall data analytics architecture. Describe the data collection/ingestion process, data storage, data processing, and data serving.

Provide a draft figure that depicts the overall architecture and data flow in your system. Describe anticipated limitations or difficulties with your approach.

EXPERIMENTS & EVALUATION

Provide a **brief and clear description** of how you plan to evaluate the results of your project.

Discuss your proposed analysis from the **scientific method.**

For example, if you are doing classification you should consider metrics such as classification accuracy and precision-recall.

Will you use cross-validation, or does your data set(s) come with a fixed train-test partition?

For unsupervised learning tasks like clustering or topic modeling you may have to do some research to see how evaluation is done on these tasks.

Or your evaluation may be more qualitative in that you hope to generate insights about a particular problem.

Some questions to get you thinking about this section:

- How will you evaluate your system and architecture?
- What results do you plan to obtain?
- What type of data analysis will you perform?
- How is this type of analysis adequate for the data, problems and questions posed?
- What other datasets can be used?
- What are the steps you need to take to scale your solution?

SOFTWARE

Provide a list of the major pieces of project software that you expect to use, divided into 2 sets:

(1) publicly available code, and

(2) code you will write yourself.

The list of what public software you will use will probably be incomplete at this point (which is fine) since you may not know yet about all the software that might be relevant to your project.

You will want to state this explicitly here as well as including the appropriate requirements or environment file.