Data Science Capstone Project

GOAL

The final project should represent significant original work applying data science techniques to an interesting problem that the Practitioner may be currently experiencing their office, or would like to take a deeper dive on. The goal of this project is for the student to gain experience in understanding a substantive problem, acquiring data relevant to the problem, and applying appropriate data science techniques in an effort to address the problem. Your focus will be addressing a data-related problem in your business domain. These business problems will be published and accessible by key data stakeholders and analytics directors through the agency.

DELIVERABLES

The capstone will be submitted as 7 deliverables:

1. Project problem statement
2. Project data
3. Project proposal
4. First project presentation
5. Final project presentation
6. Executive briefing
7. Final paper

Final projects are individual attainments, but you should be talking frequently with your cohort members about them.

PROJECT PROBLEM STATEMENT

For the project, each student will identify an applied problem (or a few related problems) that could be solved using data science methods. This could be a problem you are currently facing in your everyday GSA job, or a future problem that relates to your role that you would like to spend some time discovering and deep diving, while using some of the data science tools within this curriculum.

You will submit a short write-up that answers these questions:

- What is the problem/question you hope to answer?
- What data are you planning to use to answer that question?
- What do you know about the data so far?
- Given the problem and data you know of, what do you believe is the solution? What does success look like?
- Why did you choose this topic?

Characteristics of a good project question

- Clearly defined: The question can easily be summarized in a single statement ("Can we predict A based on B?").
- As simple as possible: The question has a narrow focus rather than broad goals.
- Reasonably available data: The question depends on data that is likely to be available in a "meaningful" quantity.
- Reasonable hypothesis: The question examines factors (B) that might actually be predictive of the outcome (A).

PROJECT DATA

If you haven't already, it's time to pick a dataset. This may be a dataset that you've worked with before or one that you are not yet experienced with. The important part is that it makes sense for your capstone project, and that the data relates to your business problem you have identified. Think of the data as your evidence; this should be the backbone of your initial analysis. Ideally, you should get all the data in your hands before starting your project. Make sure you have access to the data you want to use in the *quantity* and *quality* you need.

Take time to deeply understand the data you choose, you'll be using this a lot

First, dive in and explore the dataset. Spend a lot of time going over its quirks and intricacies. You should understand how it was gathered, what's in it, and what the variables look like. Coordinate time with the data steward if you need it. Use some of the concepts you've learned so far (code and visuals) to really get to know the data, and formulate some ideas for possible testing. At our next check-in, you should be able to spend 5 minutes going over your dataset, and explaining to use how you are intending to leverage this for your capstone, with the group.

Outline your business problem, and begin testing

When you've fully explored the dataset, it's time to move on to the next step in your capstone. Using the dataset you selected, propose and outline an experiment plan. List some goals of your analysis, ideally in the form of testable hypothesis, or via well-defined success metrics. These can be tentative, and you don't need to stick to them throughout your project. But you should always approach the data with some expectations so that your efforts are focused. Begin to explore how you can leverage the tools you are learning, whether its Python or Tableau or Drupal, to solve your business problem and provide recommendations to your directors.

PROJECT PROPOSAL

Before you dive into building out your final capstone, we'd like you to craft a proposal for your project. This should allow you to clarify what you're looking to build, as well as get feedback on that objective before committing the full time it would take to build it.

The proposal should be approximately **one page long** and answer the following questions:

- ○ What is the problem you are attempting to solve? Include important background information.
- ○ How is your solution valuable?
- ○ What is your data source and how will you access it?
- ○ Give a summary of the cleaning/joining of data that you expect to do up front.
- ○ What techniques and tools from your learning do you anticipate using?
- ○ What do you anticipate to be the biggest challenge you'll face?

When answering these questions you should form a clear picture of the work you intend to do without having begun to build out the infrastructure to execute it yet.

FIRST PROJECT PRESENTATION

You'll be giving a short presentation to the cohort about the work you've done so far, as well as your plans for the project going forward. You should prepare a Slides presentation which summarizes your data science project and outcomes. Use the following format. You should end up with 6-10 slides + title.

- Identify yourself on a title slide.
- (1-2 slides) Problem statement

    - What problem you are trying to solve. Should include quality metrics you use to measure performance/accuracy.
    - Should *not* describe the algorithm or method you're using to solve the problem.
- (2) Methods you explored
    - May include some data preparation/featurization, and then the algorithms you tried, and possibly visualization or interaction methods.
- (1-2 slides) Tools
    - The tools you used, and a rationale for their use. Can cover data preparation, learning, visualization, performance measurement etc.
- (1-2 slides) Results
    - Any results you have to report so far. What insights have you gained from your exploration? May also be a report of unexpected challenges.
- (1-2 slides) Lessons learned
    - Lessons learned and/or plans to mitigate challenges.
    - Will you be able to answer your question with this data, or do you need to gather more data/adjust your question?

FINAL PROJECT PRESENTATION

When working with data science projects, you will frequently have to present your findings to business partners and other interested parties - many of whom won't know anything about data science. That's why it's important to practice communicating your work clearly and effectively - for any audience.

Note: Prior to your final presentation in front of GSA's executive leaders, we are offering an extensive prep session with GSA's CDO to prepare you for your 5 min presentation. In this session, we will go over expectations, and how best to deliver your project.

Your goal is to create a 5 minute presentation that guides a non-technical, business heavy audience through your problem or initial challenge, the approach you took to solve it, your findings, and results/impacts. You should already have the analytical work complete, so now it's time to clean up and clarify your findings (in a non-technical way).

Create a detailed 10-20 slide deck or interactive demo that explains your data, visualizes your model, describes your approach, articulates strengths and weaknesses, and presents specific recommendations. Plan for a 5 minute presentation with 1-2 minutes of QA; be prepared to explain and defend your model to an inquisitive audience!

Goal: A detailed 10-20 slide presentation deck that relates your data, model, and findings to a non-technical audience.

Requirements:

- Show off your work to what would be a less technical, more business oriented audience
- Summarize the work you've completed from earlier deliverables into a clean presentation, including:
  - Your project's background, problem and hypothesis
  - Descriptions of the datasets you used
  - Data exploration with summary and charts
  - An explanation of your model (for non-technical audiences)
  - Recommendations based on your findings
  - An appendix that includes all of your work and technical terminology
- Review next steps with your audience; what could you do beyond the scope of this course?

Detailed Breakdown: A 10 to 20 slide deck consisting of:

- (1) Outline Slide
    - What is your project about?
    - What is its history?
    - What relevant information is required for a colleague to jump in to understand your project?
- (2-3) Summary Slides (including data and problem statement)
    - What were you trying to accomplish?
    - What steps did your project take?
    - Where did the data come from? What does a sample look like? Was there data you considered but decided to remove?
- (3-4) Modeling Insight Slides
    - What is the visualization explaining?
    - What do the x and y axes mean?
    - How does the visualization help either prove or disprove your work?
    - What caveats have to be explained to best understand it?
- (2-3) Modeling Approach Slides
    - What was your model trying to optimize for? Why was it the right metric for optimization?
    - What algorithm did you try? How does it work?
- (2-3) Results Slides
    - What worked? What didn't? Why?
- (1-2) Conclusion Slides
    - What had the most impact on your work?
    - What can you confirm? What can you suggest? What is still to be determined?
- (1-2) Next Steps Slides

    - What should this project do moving forward?
    - What would be the next two or three things you want to try? What impact might they have?
    - What might your conclusions enable others to do?

Bonus:

- You might also include a Citations Slide, if necessary

EXECUTIVE BRIEFING

You will be invited to an executive briefing session where you will report out on your project. You should be able to articulate your business problem and results within 3 mins, and then give a high level overview of your results, whether through a dashboard, report or other format, in 2 mins. The most important outcome is to be able to present very technical work to a very non-technical Executive on how your work can solve key business decisions.

A prep session with the GSA CDO will be setup for you to review/prep for this session.

FINAL PAPER

Your final report will build off of the initial project proposal you wrote, but will give the final project details instead of anticipated outcomes. This should be a 1-page document to include a summary of problem statement, challenges faced, approach taken to solve, what tools were used, and impacts. This will be published & sent over to the Data Governance Boards for your organization.

Goal: A 1-page non-technical document that relates your project, approach, findings, and impact to a non-technical audience.