Springboard

Revised: October 3rd, 2018

Capstone Project Guidelines

Data Science Career Track

Table of Contents

Capstone Project Overview

A capstone project is your time to shine. It showcases your skills and talents and represents the work you've done as a data scientist. Capstone projects are especially helpful in differentiating yourself, since you'll demonstrate your creativity and work on topics that interest you. If you have limited work experience as a data scientist, the capstone project serves as a way to illustrate your abilities to potential employers.

The capstone project helps develop a data science mindset. By working on a project from start to finish, you'll encounter all aspects of what a data scientist job entails, such as collecting a dataset, cleaning and organizing, visualizing the data, and building predictive models.

As you think about your capstone project, keep in mind that it's better to pick a relatively straightforward, “boring” project that you can test your skills on than to pick a complex idea that'll feel overwhelming. Employers also love to see a solid, completed and well-documented project than an incomplete, half-baked one. Remember to focus your project around a realistic client and problem. Lean on your mentor as a resource, sounding board, and filter, and reach out to your peers for feedback and support.

What you will complete

- **Capstone project 1**: This project focuses on a general topic to help you develop data science skills
- **Capstone Project 2**: This project focuses on your specialization and incorporates skills related to your area of focus.
- **Deliverables**: After collaborating with your mentor, you will decide whether to create a video, presentation, blog post, or report to illustrate your findings and work.

How to Pick Your Capstone Projects

When you work as a data scientist, you'll be required to deliver results that are "good enough," working under time constraints, since other teams, such as engineering, product, sales, or marketing, will be waiting on your analysis. Having a sense of the tradeoffs between different approaches will assist you in quickly selecting a method that's well-suited to a problem and available resources.

To help develop this mindset and cultivate a situation that reflects this skill, use the following guidelines to create a good capstone project:

- **Is it a real problem that someone cares about?**
  - Ideally, the results could be applied directly to a realistic situation or added to your portfolio to illustrate your skills and knowledge.
- **Is there real data available?**
  - It's important to work with real-world data, not a constructed dataset or one that's used only for academic teaching purposes, so that you can get used to the type of datasets you'll encounter in a workplace.
- **Is the data easy to acquire and clean?**
  - While you want real-world data, you don't want to spend 100 hours cleaning and wrangling with it. Pick a dataset that's relatively clean. As a rule of thumb, if you have to spend more than one week acquiring and cleaning your data, you may want to reconsider and find a cleaner dataset.

To paraphrase Einstein, *keep it as simple as possible, but no simpler.* Your mentor can help you decide if your project idea meets these guidelines and provide support and direction.

First vs. Second Capstone Project

The Data Science Career Track requires two capstone projects. Here are a few guidelines to help you determine what topics are appropriate for each project:

1. **First Capstone Project:** For your first capstone project:
    a. Propose a topic that is straightforward in both approach and complexity.
    b. *(Optional)* Choose a project that leverages any specific industry expertise you may already have from prior work experience.
    c. The goal is to get a handle on a larger project without feeling overwhelmed, so select a project that doesn't require web scraping, which is often more difficult than it appears.

2. **Second Capstone Project:** For your second capstone project:
    a. Propose a topic that requires a more complex approach.
    b. Choose a topic in the area of your specialization.
    c. Showcase more of the advanced topics you've learned.
    d. *(Optional)* Focus it in the industry you'd like to enter upon graduation.

In both projects, please work closely with your mentor to choose topics that are the right balance of challenging and attainable, given your current skill set.

You'll follow the same steps for both capstone projects, so we recommend reading through the entire document before starting.

Capstone Project Deliverables

We‘ve broken down the capstone projects into smaller parts, with prompts at appropriate points in the curriculum. Here’s a quick overview of what you’ll do for each capstone project. Collaborate with your mentor to come up with a pace and set of deliverables that makes the project manageable and helps you progress towards successful completion.

Capstone Project 1

1. **Step One: Initial Project Ideas**
   Propose up to three initial project ideas to your mentor and the Springboard community.
2. **Step Two: Project Proposal**
   Write a project proposal to help your mentor understand your idea and interest.
3. **Step Three: Data Wrangling**
   Collect Data and Apply data wrangling techniques to your project to help you analyze it.
4. **Step Four: Data Story**
   Start exploratory data analysis to find trends and identify significant features in the dataset.
5. **Step Five: Exploratory Data Analysis**
   Conduct further data analysis to identify relationships between different variables.
6. **Step Six: Milestone Report**
   Write a milestone report to show your mentor your progress.
7. **Step Seven: In-Depth Analysis**
   Perform in-depth analysis, using machine learning or another advanced technique, to draw insights from the data to answer your project question
8. **Step Eight: Final Project Deliverable**
   Complete the final deliverables to showcase what you’ve done.

These are each described below in more detail.

Step 1. Initial Project Ideas

Think of up to three project ideas that excite you and meet the guidelines. You can explore datasets from Quandl, US Government Open Data, UCI Machine Learning Repository, Kaggle competitions, Mode Analytics, or Google's public datasets directory or anywhere else. Data is Plural is an email list that sends newly released and interesting datasets. If you'd like guidance on how to select an appropriate dataset or develop a project idea, reach out to your mentor. Once you have a few ideas, narrow them down to the one idea that you're most excited to work on.

Google Dataset Search

Google has launched a tool called Google Dataset Search that makes it really easy to search publicly available data sets. We encourage all of our students to use this resource.

For your initial project ideas, you'll want to:

- Include a short description for each idea.
    - The description should briefly discuss the problem and the data you'll use to solve it. At this point, there's no need to outline specific methods and techniques.
- Post your idea, including the title and description, to the community and solicit feedback from both mentors and students.
- Pick one idea to work on based on the feedback. Discuss the idea with your mentor to ensure that they're on board.

Note: It's important that you don't work on an overly complex idea. The goal of this capstone project is to demonstrate your competence as a data scientist. It's perfectly acceptable to work on a dataset that's been worked on before and answer a question that's been answered before, as long as the work is your own.

A Word of Caution on Datasets

Kaggle Competitions

Kaggle competitions take datasets that are already cleaned and optimized for specific problems and tune machine learning algorithms for them.

Kaggle competitions are useful, but you won't be able to rely on them as a working data scientist, since you'll be responsible for collecting, wrangling, and cleaning data. If you're considering a Kaggle competition for your capstone project, here are a couple of ways you can still use the dataset:

- Can the dataset solve a different problem other than the one asked in the competition?
- Can you combine the dataset with others to solve either the same problem or a different one?

The goal of your capstone project is to demonstrate your competence with the entire data science process, so getting some experience collecting, wrangling, and cleaning data is important.

Note: Typically, recruiting competitions sponsored by top companies (e.g. Airbnb) meet the criteria for a capstone project, but your mentor has the final word on whether a Kaggle competition is appropriate for a capstone project.

Using Proprietary Data

Many students work on capstone projects that involve proprietary data (e.g. a current employer). Working with proprietary data is acceptable, and it is not required that you share the raw data with Springboard or your mentor.

However, there are a few aspects you'll need to keep in mind:

1. **Ensure you have the right permissions**: Your mentor is here to guide you through your project and can only do that effectively if they can look at your code, results, charts, etc., even if they don't have access to the actual raw data. Furthermore, if you'd like feedback on your code, your course TA will also need access.

   a. Additionally, Springboard requires that you submit a project report and slide deck based on your analysis and publicly share them on GitHub.
   b. If the owners of the proprietary raw data are not comfortable with these requirements, you may need to rethink your project topic. Please check with the appropriate legal team to determine if you need approval in the form of a legal contract or a Non-Disclosure Agreement (NDA).

2. **Start data collection early**: Even if you have the requisite permissions, please start the data collection process early and have a realistic idea of how soon you can access the data.
   a. Many companies have elaborate processes around data access and extraction to ensure security and privacy, and past students have had to wait for months to complete their project until the data became available.

If you have any questions or concerns about using proprietary data, please email your student advisor.

Step 2. Project Proposal

Once you've settled on an idea, write a 1-2 page proposal that answers the following questions:

1. What is the problem you want to solve?
2. Who is your client and why do they care about this problem? In other words, what will your client do or decide based on your analysis that they wouldn't have done otherwise?
3. What data are you using? How will you acquire the data?
4. Briefly outline how you'll solve this problem. Your approach may change later, but this is a good first step to get you thinking about a method and solution. This might include:
   a. Is this a supervised or unsupervised problem?
   b. If supervised is it a classification or regression problem?
   c. What variable is it you are trying to predict?
   d. What variables will you use as predictors?
   e. What will be your training data?
5. What are your deliverables? Typically, this includes code, a paper, or a slide deck.

The proposal will be part of a GitHub repository for your project. All code and further documentation you write will be added to this repository.

Once your mentor has approved your proposal, please share a link to your GitHub repository on the community and ask for feedback. At this point, the project proposal will be considered approved and ready.

This project will be evaluated with this rubric.

Step 3. Data Wrangling

The first step in your capstone project is to collect and prepare your data for analysis. In some cases, data collection can be as simple as downloading a dataset in a zip file. During your second capstone project, or if you specialize in natural language processing, you'll need to extract data using a publicly available API or scrape a website. Please work with your mentor to ensure that the data collection process is not too difficult.

At the end of the unit on data wrangling, you'll apply some of the techniques to your collected dataset and write a 1-2 page document describing the data wrangling steps that you undertook to collect and clean your data.

As your write the document, answer these questions:

- What kind of cleaning steps did you perform?
- How did you deal with missing values, if any?
- Were there outliers, and how did you handle them?

Note: Upload the document to your GitHub repository. This document will eventually become a part of your milestone report. .

This project will be evaluated with this rubric.

Step 4. Data Story

The next step in your capstone project is to create a data story. Data storytelling is an art.
In this project submission, you will create a data story and include the questions you asked about the data, the trends you investigated, and the resulting visualizations and conclusions you made.

Note: Upload the document to your GitHub repository. This document will eventually become a part of your milestone report.

This project will be evaluated with this rubric.

Step 5. Exploratory Data Analysis

After you've obtained, cleaned, and wrangled your dataset into a form that's ready for analysis, you'll perform preliminary exploration. This exploratory data analysis (EDA) uses a combination of inferential statistics and data visualization to find interesting trends and identify significant features in the dataset. For example:

- Are there significant variables that help explain the answer to your project question?
- Are there strong correlations between pairs of independent variables or between an independent and a dependent variable?

Your findings from this phase will be summarized in a 1-2 page document. These findings will become part of your milestone report and help guide you in further analysis, especially in the machine learning phase.

This project will be evaluated with this rubric.

Step 6. Milestone Report

The milestone is a great opportunity to consolidate your work so far and practice your data storytelling skills. Your milestone will be reached when you produce an early draft of your final project paper.

The milestone report is a 3-5 page draft that should include the following:

- An introduction to the problem, including a description of the potential client and their motivation
- A deeper dive into the dataset:
  - What important fields and information does the dataset have?
  - What are its limitations? (i.e. What are some questions that you cannot answer with this dataset?)
  - What kind of cleaning and wrangling did you need to conduct?
- Any preliminary exploration you've performed and your initial findings.
- Based on these findings, what approach are you taking? If you changed your approach after submitting your proposal, how has it changed?

Add your code and milestone report to the GitHub repository. Once your mentor has approved your milestone document, please share the GitHub repository URL on the community and ask for feedback.
This project will be evaluated with this rubric.

Step 7. In-Depth Analysis

You'll perform in-depth analysis to answer your problem statement, and use machine learning, since it's a skill that you'll use often in the workplace. For your second capstone project, the techniques you use will be directly related to your specialization.

Once you complete this step, you'll have a set of insights, including visualizations and tables, that answer your original problem statement. These insights will become a part of your final deliverables.

This project will be evaluated with this rubric.

Step 8. Final Project Deliverable

The final deliverables for your project, including your code and report, are a part of your portfolio. To help you practice discussing your work with potential employers, you'll publicly share your work online and upload a video to demonstrates your technical and communication skills.

When you and your mentor agree on a stopping point for the project, you should have the following deliverables ready **before** asking your student advisor to start the completion process:

- **Code** for your project, well-documented on GitHub.
- **Final paper** in your GitHub repository explaining the problem, your approach, and your findings. Include ideas for further research, as well as up to 3 concrete recommendations on how your client can use your findings. Please give the document an appropriate title, such as Capstone_Final_Report or Capstone_Project.
- **Slide deck** for your project in your GitHub repository. As a data scientist in a company, you'll be called upon to produce and present slide decks. You can use any standard presentation tools (e.g. Powerpoint, Keynote, Google Slides) to create your deck. Please give the presentation an appropriate title, such as Capstone_Final_Slides or Capstone_Presentation.
- **Share your project** by presenting in an office hour, creating an online video, or writing a blog post. The following section describes each option in greater detail.

Before you submit, you can also send your code to the course TA to receive feedback prior to submitting your final draft. Please email your code to projects@springboard.com for a quick review.

This project will be evaluated with this rubric.

Capstone Project 2

The second capstone is similar in design to the first project, but it's more in-depth and complex, showcasing some of the additional skills that you've acquired, especially in your specialization area.

You'll follow the same steps for your second capstone project, but the process won't be tied as explicitly to your curriculum units, and you're expected to know how to complete a data science project more independently. Therefore, you'll have fewer mandated submissions. However, it's a good idea to work with your mentor and plan multiple milestones to reflect the industry practice of having several iterations with a client during a project.

1. **Step One**: Propose up to three initial project ideas to your mentor and the Springboard community.
2. **Step Two**: Write a project proposal to help your mentor understand your idea and interest. This project will be evaluated with this rubric.
3. **Step three**: Write an initial milestone report to show your mentor your progress.
    a. Collect and wrangle data
    b. Tell a data story
    c. Conduct exploratory data analysis

    This project will be evaluated with this rubric.
4. **Step four**: Write a second milestone report to show your mentor your progress.
    a. Include in-depth analysis to draw insights from the data
    b. Follow best practices to organize and present your code and report

    This project will be evaluated with this rubric.
5. **Step Five**: Complete the final deliverables to showcase what you've done.
    a. Include advanced techniques acquired through your specialized learning track.

This project will be evaluated with this rubric. After you submit your final project, your TA will automatically receive your project to provide detailed feedback.

Course Completion

For Springboard to consider your course complete and issue a certificate of completion, your mentor needs to approve all your project submissions, including the final projects according to the rubric described below. Meeting the expectations on your intermediate project submissions and milestone reports will greatly improve the quality of your work, and prevent surprises and last-minute difficulties. Please regularly discuss the feedback with your mentor to determine what improvements must be made for resubmission at any time. You can also ask your course TA for a review before submitting anything to your mentor.

Understanding the rubrics

Each submission has a rubric attached to allow you and your mentor to be on the same page about the expectations from the deliverable. The rubric consists of three sections: Completion, Process & Understanding, and Presentation. Your submission "*Meets Expectations*" if you meet expectations on **all** of the criteria listed in these sections.

Please review the rubrics and discuss with your mentor what success looks like for each criterion.

Because your mentor assesses your work, it's vital that you collaborate with your mentor at every stage of the project to understand and agree on what the expectations are, and to incorporate their feedback during the intermediate stages.

View the Data Science Course Project List which links to the rubrics for each part of Capstone Project 1 and 2.

Note: Your student advisor won't be able to process your course completion until your project is approved by your mentor.

Presenting and Sharing Your Projects

Sharing your project is an important step towards building your brand as a data scientist. We highly encourage you to share your project with the world, and we'd love to support you in that effort. Here are some options you may consider:

- **Present in an office hour:** Your student advisor can help you schedule your presentation as part of the data science office hour once you complete the course.
- **Create an online video of your presentation:** Record yourself presenting your project with a screenshare, upload it to YouTube or Vimeo, and share the link with your student advisor. Zoom offers free accounts to record video.
- **Write a blog post:** Blogs are a great way to generate awareness of your work and your brand as a data scientist. Here are Springboard's guidelines for blog posts on our official blog. If you'd like your blog post to be considered for the Springboard blog, please write a draft according to the guidelines and share it with your student advisor. You're also welcome to post on your own blog, Medium, LinkedIn, or other platforms. Please share the link to your post with your student advisor.

Note: Remember that both your paper and slides should be targeted to the client that you picked in your proposal.

Idea: Some students have shared their report/slide deck with their proposed client and found the experience to be rewarding. If you share or present your project with a client, please share the response with the Springboard community.

What Makes a Good Presentation?

A good presentation lasts about 20 minutes, with 10 minutes for Q&A. Here are some suggested guidelines for the structure of the presentation:

- A clear explanation of the problem, client, and motivation (i.e. why does the client need to know this information?) (1-2 slides)
- The formulation of the project as a data science problem and a description of the dataset (1-2 slides)
- Suggested data wrangling steps (1-2 slides)
- Exploratory analysis and findings (3-4 slides)
- In-depth analysis (e.g. machine learning) at a high level (i.e. What method did you choose and why?) (1-2 slides)
- Analysis results (1-2 slides)
- Recommendations for your client based on your results (1-2 slides)
- Practical considerations and suggestions for improvement (1 slide)

Note: To get an understanding of the final presentation, please look at the sample presentations provided at the end of this document for inspiration.

Your Professional Portfolio

Your capstone projects will be included in a professional portfolio you develop as you work through the curriculum. Your portfolio consists of all of your data science projects, including the code and documentation, usually in your GitHub account. Typically, a data scientist who looks at your portfolio wants to see evidence of both your technical and communication skills. Having a well-organized and easy-to-navigate portfolio will help potential employers understand your skill set and knowledge.

Guidelines for a Good Portfolio

1. Every project should be in a separate, appropriately titled folder or repository.
2. For each project and folder:
    a. Include a README page:
        i. The README page should provide an executive summary of the project, which includes a summary of the problem, your approach, and the final results.
        ii. The README should also include a list of the important files that the reader should view. The files should be clearly named and organized.
    b. Clean your code and document:
        i. Your approach and methodology should be clear to any technical reader. You don't need to document every line of your code, but include comments or text explaining important decision points and why you chose them.
    c. Include any other documents that you've created, such as a report, slide deck, or video in the same repository as the code.
        i. The README should identify these documents the reader.
3. Ensure that your portfolio is not cluttered with repositories or folders that are incomplete, irrelevant, or undocumented. The goal is to ensure that the portfolio is clean, well-organized, and easy-to-navigate.

To see these guidelines in action, check out this great example of a portfolio by one of our past students, Aashish Jain.

Note: To help you get into the mindset of how you should organize your portfolio, imagine working as an experienced data scientist who has a limited amount of time to look at your portfolio. How can you make it easy to get a good idea of your skills and abilities? Your mentor, course TA, and the community can give you feedback on your portfolio, so please use them as resources.

Addendum:

Sample Data Science Career Track Capstone Projects

Here are some sample projects completed by past students:

1. Aashish Jain: “Predicting the Likelihood of Flight Cancellations”
   a. Report
   b. Presentation
   c. GitHub

2. Pablo Guevara: Research Market Analysis
   a. Github

3. Pablo Guevara: Loan Default Predictive Modeling
   a. Github

A Deep Dive into a High Quality Capstone Project

In this section, we'll go through the capstone project titled *"Predicting the Likelihood of Flight Cancellations"* by Aashish Jain. We'll look at each step of the project and discuss why this is an effective project to present to an employer.

Note: This is one of the best projects completed by a Springboard student. Even if your project doesn't turn out like this, it's important to go through Aashish's project and understand why it works and how it can inspire you.

Proposal

As you review Aashish's proposal, keep in mind the following:

1. The project proposal clearly follows the guidelines, breaking down the content into sections:
   a. Problem
   b. Client
   c. Dataset
   d. Approach
   e. Deliverables
2. The project uses an existing dataset. However, the dataset is still raw, leaving plenty of opportunities for wrangling, cleaning, and other real-world techniques.

In each case, Aashish has outlined his thoughts clearly and concisely, making it easy for his mentor (or future employer) to follow along.

Note: You may not have your data or the approach solidified at this stage, but that's perfectly fine. Getting the rest of the proposal down goes a long way to creating a successful project.

Milestone Report

Aashish's milestone report includes the following sections:

1. The milestone report begins with a clear statement of the problem and the client.
2. A data and its acquisition process that describes the challenges faced in acquiring and merging the datasets involved.
3. A review of the exploratory data analysis in a systematic manner, posing various hypotheses and using simple visualizations, such as bar charts and scatterplots, to identify predictor variables that may have interesting relationships with the target variable (e.g. In Aashish's project, the target variable is flight cancellation rates).

   a. Each chart is clearly labeled and captioned, making each chart self-explanatory.
   b. Each chart is accompanied by an explanation of how it “advances the plot” (i.e. contributes to our understanding of flight cancellations.)
4. The “Limitations” section shows the Aashish’s awareness of how the available data limits his understanding of the problem. In the real world, this is a critical skill for data scientists to have, because you’ll constantly have to make tradeoffs between the quality and quantity of data at your disposal vs. the results you can deliver.
5. Finally, a clear “Conclusions” section ties in observations into a concise, clear story.

Final Report

While this project report is longer than the recommendation, you can learn a lot about its structure.

1. Similar to the milestone report, the final report includes the problem, client, data gathering, data cleaning, and EDA steps.
2. Aashish explains the data pre-processing steps required before applying machine learning algorithms, and provides a clear explanation and justification of the metric that’s used to evaluate the machine learning techniques.
3. The project covers a series of machine learning techniques, each more complex than the last. For each step, Aashish summarizes the relevant results and explains how changing various parameters would affect the results and why.
4. At the end of the modeling section, Aashish compares the various models and makes a case for picking a specific model.
5. The report ends with a clear summary of the conclusions and ideas for future work.

Note: While this report does have a stronger academic slant, it’s important to pay attention to the structure and clarity of writing.

Presentation

The presentation slide deck is a great example of an engaging way to present a complex topic. This presentation is also longer than the recommendation, but it has several elements that are worth emulating, including:

1. An explanation of the problem and its importance to clients.
2. An emphasis on visuals rather than text. Aashish used visual diagrams and charts to make the presentation more engaging.
3. A summary of recommendations (slide 31) that’s explained in a way that non-technical clients could understand.

Note: This project and your GitHub portfolio are great examples of what impresses potential data science employers. Please use the resources to see what has worked for past students.