Capstone Project Guidelines

AI/Machine Learning Career Track

Summary

As you think about your capstone project, keep in mind that it's better to pick a relatively straightforward, "boring" project that you can deliver well than to pick a very complex, shiny idea that you'll get stuck on. Overall, here are the things that will make your capstone project a success.

- Choose a project that focuses on a realistic client and problem.
- Use your mentor as a resource, a sounding board, and as a filter.
- Reach out to your course TAs and the community for feedback at any time.

TABLE OF CONTENTS

Springboard

Introduction

How to Pick a Good Capstone Project

When you're working as a machine learning engineer in the industry, you have to deliver a 'good enough' solution that's ready for production in a limited amount of time. Unlike a typical course project, you don't have the luxury of taking the time to find the optimal or best approach, but you also have to ensure that your work is production-ready. It's very important to have a sense of the various tradeoffs between different approaches and pick one that's well-suited to the problem and resources you have.

How do you pick a capstone project that reflects this mindset? Here are some general guidelines:

- **Is this a real problem that someone would care about?**
    - Ideally, the result of your project could be something you apply directly at work, a real application that others can try out, or used as an addition to your portfolio, which can be shown to potential employers.
- **Is there real data available?**
    - Your goal should be to work on a project that has real-world data, not a toy dataset or a dataset that's used only for academic or teaching purposes. The datasets pointed to by the course material typically meet this requirement.
- **Is the data easy enough to acquire and clean?**
    - While you want real-world data, you don't want to spend hundreds of hours acquiring, cleaning, and wrangling it. Pick a dataset that's relatively clean. As a rule of thumb, if you have to spend more than two weeks acquiring and cleaning your data, you may want to reconsider using that dataset.

Basically, to paraphrase Einstein, *keep it as simple as possible, but no simpler.* :-) Your mentor will help you at this stage to decide if your project idea meets these guidelines.

Some students choose to work with a project topic that leverages specific domain expertise they've picked up from prior work experience, while others want to showcase their skills in an industry or domain they would like to enter upon graduation. In both cases, please use your best judgement and work closely with your mentor to choose projects that are the right balance of challenging and attainable given your current skillset.

We recommend reading through the entire document before starting in on your capstone project.

Springboard

Good luck!!!

Project Milestones

We have broken down the capstone project into two phases, with each phase containing several milestones. You by no means need to memorize the list found below — we've added prompts at appropriate points in the curriculum that will remind you to work on the next step of your project. With that in mind, here's a quick overview of what you'll do for the capstone project:

Phase One: Build a Working Prototype

1. **Step One:** Define your initial project ideas. You'll pick up to 3 project ideas to propose to your mentor and the Springboard community.
2. **Step Two:** Write your project proposal. This will help your mentor better understand your chosen capstone project idea.
3. **Step Three:** Collect your data. After your mentor approves your capstone project proposal, you'll collect the data you need to bring your idea to life.
4. **Step Four:** Data wrangling and exploration. You'll apply data wrangling techniques to your project to help you get ready to analyze it.
5. **Step Five:** Create a machine learning or deep learning prototype.
6. **Step Six:** Scale your prototype so that it can handle large datasets.

Phase Two: Deploy Your Prototype to Production

1. **Step One:** Create a deployment architecture.
2. **Step Two:** Run your code end-to-end to test how well it functions.
3. **Step Three:** Implement your data pipeline.
4. **Step Four:** Build a web service with an API for your application.
5. **Step Five:** Deploy your application to production.
6. **Step Six:** Test and document your API.
7. **Extra Credit Step:** Build a web interface for your application.

Springboard

Phase 1: Build a Working Prototype

Step 1. Pick Your Initial Project Ideas

Think of up to 3 project ideas that excite you and also meet the guidelines presented above. Below, you'll find a few great sources for large datasets that are appropriate for this course:

- fast.ai research datasets collection
- Google dataset search
- AWS open datasets repository
- Uber Movement
- Yelp dataset

Besides the resources listed above, you can also explore datasets from Quandl, US Government Open Data, UCI Machine Learning Repository, and Kaggle competitions, or anywhere else you like. There's a great email list called Data is Plural that lists new and interesting datasets that have been released. Ask for your mentor's help if you feel stuck at any point in time. Once you have picked your three ideas, your next goal will be to narrow your ideas down to the ONE idea that you'll focus your capstone project on.

Ideally, your dataset should have at least 15K-20K samples **at a minimum**. We'd like you to see you build large-scale applications while working through this course. We encourage you to work on larger datasets; something that's at least 8GB in size or has at least 1 million samples.

For your initial project ideas, please:

- Include a short blurb for each of your ideas.
    - The blurb should, at a high level, describe the problem and the data you'll be using to solve it. At this point, there's no need to talk about specific methods and techniques.
- Post your idea (title and blurb) to your online community and solicit feedback from both the mentors and other students.

Pick one idea to work on based on the feedback you get from your community. Discuss the idea with your mentor to ensure that they think you should move forward with your idea.

Note: The goal of this project is NOT to do something novel (You're not building the next Facebook or Airbnb just yet — though we'd be really proud if you did that someday!) Instead, the goal of this project is to demonstrate your competence as a machine learning engineer. It's

perfectly acceptable to work on a dataset that's been worked on before and even answer a question that's been answered before, as long as the work you submit is your own.

How complex does a project need to be?

The goal of this course is to give you the skills to not only design and create a Machine Learning (ML) or Deep Learning (DL) application but to also deploy it to production using the latest engineering tools and techniques.

ML/DL contains a wide swath of techniques and it's important to note that not all of the techniques you'll cover in this course will be able to be applied to the problem you have chosen. Obviously, the more complex and 'cool' the techniques you use, the more attractive your project will be to employers. But it's important to balance that lure with a practical approach that allows you to produce a real, scalable application. Some questions you can ask yourself include:

- What's the technique that best applies to this problem?
- How easily can this technique be deployed to production?
- Are the results of this technique 'good enough' to meet the business requirements of the solution?

In the real world, it can be better to use a simple logistic regression approach that's good enough and can be deployed quickly and scalably as a production application than a really complex deep learning approach that's extremely difficult to deploy or maintain. You'll have to make similar decisions and trade-offs about your capstone project.

Work with your mentor to determine a problem and approach that meets the requirements of this course and is "cool enough" that you feel excited to work on it. Ultimately, you'll work together to identify a problem and approach that you both feel will set you up for success.

Some example ideas for inspiration

Here are a few ideas that might help spark inspiration for your capstone project. Many of these ideas come from natural language processing or computer vision, as these two fields are currently some of the hottest in the world of AI. You're also welcome to come up with your own ideas that have nothing to do with these.

- **Inventory tracking and compliance using object recognition:** A company wants to track inventory in its warehouses automatically using a camera with an object recognition algorithm. You can also think of this as a home application; for example, a smart fridge recognizes what's in it based on pictures.

- **Language translation:** Also called neural machine translation, this uses AI to translate one human language to another, whether through text or speech. You can also work between the two formats e.g. speech-to-text transcription or text-to-speech generation.
- **QA systems and chatbots:** More and more companies are using automated chatbots to address their customer service workloads. These bots can produce human-like responses to questions (within limits) and are getting better every day.
- **Text summarization:** Imagine an application that can digest the daily news and produce a coherent summary for a consumer. You can apply summarization to different domains, such as an application that can automatically produce a personalized 'Cliff's notes' for a student who's trying to research a large amount of material.
- **Fraud/spam detection:** Detect "bad" transactions or items in a dataset. This could take the form of detecting fraudulent credit card transactions, fake news on social media, spam in email, doctored images or video, abusive behavior on Twitter, and so on. Depending on the problem, you can use a variety of techniques ranging from "traditional" machine learning to the latest in deep learning.

A Word of Caution about Datasets

Kaggle Competitions

It's perfectly fine to use a dataset from Kaggle for your project. However, many Kaggle competitions are about taking a dataset that's already clean and optimized for a specific problem and tuning a machine learning algorithm to produce the highest accuracy (or similar metric).

While that's an important skill for a machine learning engineer to have, it's not all that you should be focusing on. In real-world scenarios, you'll be the person who has to collect, wrangle, and clean that data. If a Kaggle competition you're considering falls into that category, here are a couple of ways you could still use the dataset:

- Could you use that dataset to solve a different problem than the one asked in the competition?
- Could you combine it with other datasets to solve the same problem asked in the competition or to solve a different problem altogether?

Basically, we'd like your capstone project to demonstrate your competence with the entire process of creating machine learning systems, not just one aspect of it.

That being said, **your mentor has the final word** on whether a Kaggle competition is appropriate for a capstone project or not. Typically, we've found that recruiting competitions sponsored by top companies (e.g. Airbnb) meet the criteria for a capstone project.

Using Proprietary Data

Many of our students work on capstone projects that involve proprietary data, which may, for example, come from their employer. This is perfectly fine. **We don't require that you share the raw data** with Springboard or your mentor. However, there are a few items you'll need to pay attention to:

1. **Ensure you have the right permissions:** Your mentor is here to guide you through your project. They can only do that effectively if they can look at your code, summarized results, charts etc, even if they don't have access to the actual data.
    a. In addition, Springboard requires that you turn in a project report and a slide deck based on your analysis and place it publicly on GitHub.
    b. If your employer or the people who are providing you the raw data are not comfortable with these requirements, you may need to rethink your project topic.

Please check with the legal team at your company to see 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 make sure to start the data collection process early, and have a realistic idea of how soon you can actually get the data.
    a. Many companies have elaborate processes around data access and extraction (with good reason!), so sometimes students have become stuck for weeks or months waiting for their project data to become available.
    b. Ensure that you follow good privacy and security practices, such as anonymizing the data where appropriate. In some cases, (such as if you're using healthcare data,) you may be legally required to anonymize it. Please work with the legal and security teams at your company to ensure you're always in compliance with their codes.

If you have any questions about whether or not you can use proprietary data for your capstone project, feel free to email your student advisor!

Step 2. Write Your Project Proposal

Once you've decided on your final capstone project idea, we'd like you to write a proposal. A project proposal is a short (1-2 page) document that answers the following questions:

1. What is the problem you want to solve? Why is it an interesting problem?
2. What data are you going to use to solve this problem? How will you acquire this data?
3. In brief, outline your approach to solving this problem (knowing that you may not know everything in advance and this might change later). This might include information like:
    a. Is this a supervised or unsupervised problem?
    b. If supervised, is it a classification or regression problem?
    c. What are you trying to predict?
    d. What will you use as predictors?
    e. Will you try a more "traditional" machine learning approach, a deep learning approach, or both?
4. What will be your final deliverable? This is typically an application deployed as a web service with an API or (for extra credit) a web/mobile app.

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 the GitHub repository URL on the community and ask for feedback.

At this point, the project proposal will be considered approved and ready.

Step 3. Collect Your Data

To kick-start your capstone project, the first thing you'll need to do is collect your data. In some cases, it can be as simple as downloading a dataset in a zip file or a tarball. In other cases, it can require extracting data using a publicly available API or scraping a website. We urge you to work closely with your mentor to ensure that the data collection process is not too onerous for a capstone project. Also, **if your data collection requires you to write code, it's important that you start early.**

At the end of this step, you'll submit a link to your your GitHub repository that contains the following:

1. Code for how you collected the data if applicable

2. The actual dataset: if your dataset is small enough to fit in a CSV, then feel free to include it in the repository. If it's a big dataset or has a lot of binary files (graphics, audio), consider using the Git Large File Storage extension.

Step 4. Clean and Wrangle Your Data

After you've worked through all of the resources on data wrangling, you'll apply some of the data wrangling techniques you've learned to your capstone dataset. As you're working in your Jupyter notebook, take notes documenting the data wrangling steps that you followed to clean your capstone project dataset. Consider the following as you do so:

- What kind of cleaning steps did you perform?
- How did you deal with missing values, if any?
- Were there outliers? If so, how did you handle them?
- If your dataset is too large to work with, does it make sense to build your prototype on a smaller subset of the data?

Optional: Explore Your Data

After you've obtained the dataset for your capstone project, cleaned, and wrangled it into a form that's ready for analysis, you will perform a preliminary exploration of the data. 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 variables that are particularly significant in terms of explaining the answer to your project question?
- Are there strong correlations between pairs of independent variables or between an independent and dependent variable?

At the end of this step, you'll submit a link to your Jupyter notebook in your Github repository that shows how you cleaned, wrangled, and (optionally) explored the data.

Step 5. Build Your Machine Learning (or Deep Learning) Prototype

The goal of this step is to find a machine learning or deep learning approach that works for your problem, and then show that the approach you choose is a viable one. Since the application has not been deployed to production yet, we'll call it a *prototype*.

For this step, you'll build your prototype in a Jupyter notebook. Depending on your problem, you'll be using a more 'traditional' machine learning (ML) technique or a deep learning (DL) technique. Your goal is to come up with a working implementation of your prototype in a Jupyter

notebook. This prototype could work on a subset of the data but demonstrates that your approach to solving the problem is a viable one based on the following criteria:

- The data has been reasonably split into training, validation, and test sets.
- You have used the correct metric(s) to evaluate the performance of your algorithm.
- The performance of your algorithm is 'good enough' as determined by your mentor.

At this point, you'll submit a link to your Jupyter notebook with the ML/DL algorithm coded and your results well-documented in a way that your mentor (or a potential employer) can easily follow.

Step 6. Scale Your Prototype with Large-Scale Data

In this step, your goal is to ensure that your ML/DL approach, which has proved to be viable, can work with large volumes of data. Please work with your mentor to determine what that means for your problem.

Using scikit-learn, SparkML, Keras, TensorFlow, PyTorch or some of the other technologies you have learned, implement your prototype at scale.

In case your earlier prototype was working with a subset, ensure that this scaled-up prototype can handle your complete dataset.

Think about what your capstone problem would look like in the real world:

- How much data would you need to handle?
- Can you scale your prototype to handle that volume of data using the approach and tools you have selected?

In a Jupyter notebook, implement the scaled version of your prototype and document clearly what trade-offs and implementation decisions you have to make to scale your algorithm. Submit the GitHub link to this notebook.

Phase 2: Deploy Your Prototype to Production

[Phase 2 details coming soon]

Guidelines for a Good Portfolio

Your portfolio consists of all of your projects, including the code and documentation contained in your Github account. Typically, a hiring manager who looks at your portfolio wants to see evidence of both your technical skills and your communication skills. It is your responsibility to ensure that your portfolio is clear and easy to navigate.

Tips to Help Your Portfolio Stand Out

1. Every project should ideally be in a separate repository that is clearly named.
2. For each project:
    a. Have a README page:
        i. Make sure you have a README page that provides an executive summary of the project i.e. summarizes the problem, approach, and final results.
        ii. The README should also include a list of the important files that the reader should look at. The files themselves should be as clearly named and organized as possible.
    b. Clean up your code and document:
        i. Your approach and methodology are clear to any technical reader. You do not need to document every line of your code, but you should include comments or text explaining important decision points and why you chose them.
    c. Include any other documents that you have created e.g. a report or slide deck in the same repository as the code.
        i. Make sure the README points them out to the reader.
3. Ensure that your portfolio is not cluttered with “junk” i.e. repositories or folders that are incomplete, irrelevant, or undocumented.

Overall, try to put yourself in the shoes of an experienced machine learning engineer or hiring manager who has a limited amount of time to look at your portfolio. How can you ensure that you make it easy for them to get a good idea of your skills and abilities? Your course TA, mentor, and the community should also be able to provide good feedback on your portfolio, so please use them as resources.

Project Evaluation

For Springboard to consider your workshop complete and issue a certificate of completion, your mentor needs to approve your final project submissions based on the rubric listed below. If the

project is not approved, please discuss the feedback from your mentor and resubmit your capstone with any improvements that have been identified as necessary for approval. Your student advisor will not be able to process your course completion until your project is approved by your mentor!

Capstone Project Rubric

We use the following rubric for evaluating final capstone projects. Please take a good look at it and take a moment to talk with your mentor about success criteria.

View the Capstone Project Rubric here

Your capstone project is evaluated based on two main criteria: 1) Completion and 2) Process and Understanding. Within each criteria, specific benchmarks and expectations are listed to help ensure the overall quality and mentor approval of your capstone project.

Because your mentor approves each step and, later, the entirety of your capstone, it's vital that you work with your mentor throughout the course to determine and agree on what the bar is for each criterion, as well as to incorporate timely feedback during the intermediate stages.