


# Capstone Project Description



[\\_ \(https://github.com/learn-co-curriculum/dsc-capstone-ml-project-enterprise\)](https://github.com/learn-co-curriculum/dsc-capstone-ml-project-enterprise)   
(<https://github.com/learn-co-curriculum/dsc-capstone-ml-project-enterprise/issues/new/choose>)

One last project to go before graduation!

In this project description, we will cover:

- Project Overview
- Deliverables
- Grading
- Getting Started

## Project Overview

For the Capstone project, you will engage in one more **supervised machine learning** process from start to finish. The choice of topic, dataset, target, and type of model is yours.

## Business Problem and Data

You are responsible for identifying an appropriate business problem and dataset. Be sure that you choose a dataset that is publicly available to ensure that you are not sharing any sensitive information. You need to be able to complete all elements of the checklist, which means that **you must utilize supervised machine learning**. Other components like unsupervised learning and data dashboards may be part of the overall process but are not sufficient to pass this project.

In other words, you need to find a business problem where you can frame the question as:

Using **data** , can we predict **target** ? This would be useful because **rationale** .

Let's break that question down further:

1. **Data:** What are the *inputs* to your model? Think about the contexts where this information would be available
2. **Target:** What is your model trying to *predict*? Is this something that would realistically be unknown in a context where the above features are known?
3. **Rationale:** Why would it be valuable to be able to predict this target? Who would find this model useful? How accurate does the model need to be in order to serve the stated purpose?

Other questions like "what is the relationship between **a** and **b** " or "how much does **c** factor into **d** outcomes" may be incidentally possible, but make sure you have that predictive framing first.

# Key Points

## Project Management

Project management is key. You have a lot of freedom in this project - this can feel liberating, but also means that you can accidentally lose a lot of time if you're not careful. Map out a rough daily project plan with key milestones and due dates for deliverables - you can adjust this as needed as you progress. Use this to make sure you're making timely progress towards successful completion. Ask for help if you find yourself struggling to keep up with your plan.

## Minimum Viable Product (MVP)

It's understandable that you want this project to represent the best of your abilities. But we also want you to graduate on time! Sometimes this means that you need to have a "done is better than perfect" mindset, to make sure that you complete all of the baseline requirements. For your MVP, just try to make sure you have completed all checklist elements and have achieved the "Meets Objective" standard for each rubric objective. You will have plenty of time afterwards to add bells and whistles, so try to stay focused on the MVP for now.

## Deliverables

At this point, the project deliverables should be familiar:

- A **non-technical presentation**
- A **Jupyter Notebook**
- A **GitHub repository**

The checklist of requirements is similar to what you saw previously as well.

## Non-Technical Presentation

As a reminder, the graded elements of the presentation are:

- Presentation Content
- Slide Style
- Presentation Delivery and Answers to Questions

## Jupyter Notebook

Feel free to make multiple notebook as you explore data and models. Just make sure that there is one final notebook that is clearly designated as such for grading.

As a reminder, the graded elements of the notebook are:

- Business Understanding



- Data Understanding
- Data Preparation
- Modeling
- Evaluation
- Code Quality

## GitHub Repository

Ensure that it contains accessible, inviting, professional content.

## Overall Repository Requirements

As a reminder, a professional GitHub repository has:

1. **README.md**
  - A file called **README.md** at the root of the repository directory, written in Markdown; this is what is rendered when someone visits the link to your repository in the browser
2. Commit history
  - Progression of updates throughout the project time period, not just immediately before the deadline
  - Clear commit messages
3. Organization
  - Clear folder structure
  - Clear names of files and folders
  - Easily-located notebook and presentation linked in the README
4. Notebook(s)
  - Clearly-indicated final notebook that runs without errors
  - Exploratory/working notebooks (can contain errors, redundant code, etc.) from all team members (if a group project)
5. **.gitignore**
  - A file called **.gitignore** at the root of the repository directory instructs Git to ignore large, unnecessary, or private files
    - Because it starts with a **.**, you will need to type **ls -a** in the terminal in order to see that it is there
  - GitHub maintains a [Python.gitignore](https://github.com/github/gitignore/blob/master/Python.gitignore)  (<https://github.com/github/gitignore/blob/master/Python.gitignore>) that may be a useful starting point for your version of this file
  - To tell Git to ignore more files, just add a new line to **.gitignore** for each new file name
    - Consider adding **.DS\_Store** if you are using a Mac computer, as well as project-specific file names
    - If you are running into an error message because you forgot to add something to **.gitignore** and it is too large to be pushed to GitHub [this blog post](#) 

(<https://medium.com/analytics-vidhya/tutorial-removing-large-files-from-git-78dbf4cf83a?sk=c3763d466c7f2528008c3777192dfb95>).(friend link) should help you address this

## README Deep Dive

The README is the "home page" of your Capstone project. Other than the *Attention to Detail* objective, all rubric objectives for this project will focus on the README.

Ideally, your README should include:



### 1. A project title

- Choose a title that reflects the project domain and presents you as a data scientist, not as a student. The title should not include words like "capstone" or "school"
- Some title formats to consider:
  - "Predicting **target** "
  - " **target** Prediction"
  - " **target** Detection"
  - "Classifying **target** "
  - etc.
- Feel free to add "with **data** " or "using **data** " to the end of any of those. For example, *Detecting Fake Reviews with NLP* or *Classifying Skin Lesions Using Neural Networks*
- You also might want to start the title with a catchy phrase or quote, followed by a more-standard title. For example, *Where To, First?: an Airbnb Destination Predictor* or *Stay in Your Lane! Automated Bike Lane Enforcement*
  - You can always add this element later, so don't get hung up on it if you can't think of something right away! Just start with the straightforward title

### 2. An elevator pitch

- Immediately after the title, write a very short description of the problem you are solving, the data you are using to solve it, and how well your model solves the problem
- This should be no more than a couple of sentences

### 3. A header image

- This image can be anything you want, so long as it is professional and aligns with your project
- Ideal dimensions are 1280x640 pixels, but any image with landscape orientation (wider than it is tall) will work
- Image sourcing ideas to consider:
  - Use a stock image from a source like [Wikimedia Commons](https://commons.wikimedia.org/wiki/Main_Page)  ([https://commons.wikimedia.org/wiki/Main\\_Page](https://commons.wikimedia.org/wiki/Main_Page)) or [Unsplash](https://unsplash.com/)  (<https://unsplash.com/>)
    - Make sure you double-check the usage license and attribution requirements
    - Create visualizations with code (Matplotlib, Seaborn, etc.)
- As a reminder, the Markdown notation for an image is `![alt text](path/to/image.png)`

### 4. Business Understanding and Data Understanding

- Explain the project context, using at least one citation to demonstrate your domain understanding
- Consider including visualizations here as well

## 5. Modeling and Evaluation

- What kind of model(s) did you use?
- How well did your final model perform, compared to the baseline?

## 6. Conclusion

- How would you recommend that your model be used?

## 7. Repository Navigation

- An explanation of the repository organization
- Links to the final notebook and presentation
- Reproduction instructions (or a link to them)

The best practice would be to include all of the items listed above in your README. Items 2 and 7 are particularly relevant for grading, as described below.

# Grading

***To pass this project, you must pass each rubric objective.*** The project rubric objectives for Capstone are:

1. Attention to Detail
2. Project Motivation
3. Independent Learning
4. Reproducibility

## Attention to Detail

Once again, the Attention to Detail standard has increased. ***In Capstone, you need to complete 100% (10 out of 10) or more of the checklist elements in order to pass the Attention to Detail objective.***

**NOTE THAT THE PASSING BAR IS HIGHER IN CAPSTONE THAN IT WAS PREVIOUSLY!**

## Exceeds Objective

In addition to completing 100% of the checklist items, goes above and beyond to create a compelling, accessible repository

Some things to consider doing are running a spell-checker to ensure you don't have typos, adding a description to the repository, using Markdown extensively in the README to create a skim-able document, etc.

You can also look at the [GitHub tips for post-graduation](https://docs.google.com/document/d/1cT13597jlbiVgKUFyhGC7_QFUrLQUfZViRu8kD_ErAA/edit?usp=sharing) ([https://docs.google.com/document/d/1cT13597jlbiVgKUFyhGC7\\_QFUrLQUfZViRu8kD\\_ErAA/edit?usp=sharing](https://docs.google.com/document/d/1cT13597jlbiVgKUFyhGC7_QFUrLQUfZViRu8kD_ErAA/edit?usp=sharing)) doc (which your career coach will be sending you soon!) to find more examples of enhancements

## Meets Objective (Passing Bar)

100% of the project checklist items are complete

## Approaching Objective

90% of the project checklist items are complete

## Does Not Meet Objective

80% of the project checklist items are complete

## Project Motivation

The passing bar for this element will be communicated in the **elevator pitch** portion of the README, immediately after the project title. The elevator pitch should:

- Describe the **data** used in the model
  - This does *not* mean that you should list out every single column name and data type. Instead, describe the data generally.
- Identify the **target**
  - What are you predicting?
- Communicate the **rationale** for predicting this target with this data
  - This can be very brief. The goal is to avoid a "so what?" response
- Evaluate the final model's **performance**

Here are some examples of elevator pitches, which include all of the required elements:

Searching for skincare products can be overwhelming, since there are so many options and skincare needs vary so much from person to person. This project uses product and review data scraped from SkinStore.com to make personalized recommendations to a new user based on their skin type, skin problems, and the types of products they're looking for. Using the **surprise** Python package, the final machine learning model has an RMSE of 0.96, meaning that it is able to predict the star rating (out of 5) within 1 star for a given product and user.

Taxi drivers in New York City have the flexibility to choose when and where to start their workdays, but the best time and location varies. This project uses data about taxi trips from the NYC Open Data portal as well as weather and CitiBike data to forecast demand for taxi pickups across

different neighborhoods. The model is able to explain about 70% of the variance in demand and would be especially useful for a taxi company to orchestrate their distribution of resources.

**Note:** There is NO threshold for model performance you are required to meet. So long as you explain what you tried and properly evaluate the outcome, it's fine to have a model with poor performance. We recognize that your time is limited during Capstone and we want you to focus more on the process and communication, not worrying about model performance.

## Exceeds Objective

Weaves a clear narrative throughout the project explaining the scope, methods, and specific use cases for the final model

## Meets Objective (Passing Bar)

Describes data, identifies a target, and communicates the rationale for predicting the target with this data and an evaluation of the final model's performance

## Approaching Objective

Does not adequately explain why these features allow you to predict this target, why this target is useful to predict, or how well the project performs

## Does Not Meet Objective

Does not include an elevator pitch

# Independent Learning

Learning the specific content we teach in the program is important, but learning how to learn is even more important! This objective is asking you to flex that skill.

The baseline requirement is that you do some domain research, rather than just diving into the dataset. Sometimes a clear understanding of the domain will help you collect a key feature or perform a key preprocessing step that will massively improve your model performance. Work smarter, not harder!

## Exceeds Objective

Demonstrates a "deep dive" beyond the course content of the program, such as extensive domain knowledge or extensively using additional Python packages

Ideally, your notebook would teach your *instructor* something new about a domain or a technique!

## Meets Objective (Passing Bar)

Cites at least one external source to demonstrate domain understanding of the project topic

The source is up to you. It can be an academic whitepaper, an industry publication, a newspaper article, or even just the Wikipedia page on a topic.

## Approaching Objective

Cites at least one source, but it is not related to domain understanding or not incorporated into the project

For example, if you cite a blog post or StackOverflow comment because you reused their code snippet, that is not a sufficient citation. You need to cite something that demonstrates domain understanding and is used in some way to direct your decision-making.

## Does Not Meet Objective

Does not cite any external sources

## Reproducibility

While you may explore additional forms of *deployment* such as creating an API or web app, the main way that projects are deployed is through a reproducible notebook on GitHub.

The explanation of how to reproduce your analysis can either be directly in your README in the Repository Navigation section, or you can create a separate Markdown file and link it in the README.

Be sure to describe the operating system used (including cloud systems such as Kaggle or Google Colab), the packages, and how to get the data.

## Exceeds Objective

Writes a script so that reproducing the project is seamless on a designated platform

This can be a Python script or a terminal script. It might involve downloading images and moving them into subfolders, installing packages with `conda` or `pip`, etc.

## Meets Objective (Passing Bar)

Incorporates the basic elements of a reproducible project: a description of software used + how to get the data

Ideally you would list the specific versions of packages used through an `environment.yml` or `requirements.txt` file. If you are using a platform like Kaggle, describe the additional `pip install` commands required beyond the base Kaggle environment.



If your data is proprietary or otherwise not accessible to someone trying to reproduce your analysis, make sure you explain this clearly. If possible, include a sample or anonymized version of the data in the repo.

## Approaching Objective


Attempts to describe how to reproduce the project, but elements are missing or not understandable

## Does Not Meet Objective

Does not describe how to reproduce the project

## Getting Started

Please start by reviewing the contents of this project description. If you have any questions, please ask your instructor ASAP.

To get started with project development, create a new repository on GitHub. For this project, we recommend that you do not fork the template repository, but rather that you make a new repository from scratch, starting by going to [github.com/new](https://github.com/new)  [.\(https://github.com/new\)](https://github.com/new).


## Summary

This is your final project in the DS program, the "crown jewel" of your portfolio. Let's do this!

How do you feel about this lesson?



Have specific feedback?

[Tell us here!](https://github.com/learn-co-curriculum/dsc-capstone-ml-project-enterprise/issues/new/choose)  [.\(https://github.com/learn-co-curriculum/dsc-capstone-ml-project-enterprise/issues/new/choose\)](https://github.com/learn-co-curriculum/dsc-capstone-ml-project-enterprise/issues/new/choose)