Batch 5 - Capstone - instructions

[Overview](#_v6nz4i1t47w9)

[Timeline](#_nmoqpmg6q1q5)

[Activities](#_ljdnba6dglns)

[Sending an email to your client](#_iji8ldolc3e8)

[Understanding the data you’ll receive](#_dlvjuqap7hjt)

[How to deliver the first report](#_gg88hch7bsp5)

[How to deliver your code](#_xsg6r7ee13kl)

[Success Criteria](#_7jst1vu87n9q)

[Report technical details](#_w3b7lmcph12p)

[Hints and advice](#_9z3c0ci7scgo)

# 

# Overview

In the capstone, we’re not going to treat you as a student. You will be treated as an employee.

The scenario is this: you’ve been hired by a consulting firm, and have just had the first [exchange of emails](https://docs.google.com/document/d/16FF6I4Q5jPqufDoi8HK1309ru3KZa1_ydUFNS-cvU94/edit?usp=sharing) with your new boss. You already know you have 3 deliverables:

* The analysis report
* An API endpoint
* The deployment retrospective report

In order to simulate a real-world scenario, the requirements may be ambiguous. A big part of your job as a data scientist in the real world will be turning business requirements into clear-cut data science requirements. In this specialization, the starting point is the emails. You will have to send a clarification email to your client with any follow-up questions to fully understand what you need to do.

Once you feel comfortable enough that you understand what is required of you, you produce the report, API endpoint, deal with the data as it comes, and then produce the second report.

# Timeline

[This calendar](https://ldssa.github.io/wiki/Starters%20Academy%20(LDSSA)/05-Capstone-Project/) is the source of truth for all dates. It is the full timeline. The upshot for the timeline is as follows:

* Week 1 is for understanding the dataset and asking questions, at the end of which you will receive answers disambiguating the issues;
* Weeks 2 and 3 are for training your model, writing your first report, and deploying your app - close to the end of week 3 we will do a trial round for you to test your app;

**By the end of week 3, you will need to submit the app and the first report**

* In week 4 the observations will be sent to your deployed model and instructors will be checking your reports;
* By the end of week 4, you will receive comments on your reports;
* Weeks 5 and 6 are for retraining and re-deploying your model and writing up the second report. You should also address comments made to the first report by the end of week 6;

**By the end of week 6, you will need to resubmit the app and the first report, and submit the second report**

* By the end of week 7, you will receive comments to your second report;
* Week 8 is for addressing comments to the second report;
* **By the end of week eight, you will need to resubmit the second report**

Instructors expect to grade the capstone and launch the graduates roughly one week after.

**Note:** Do not hesitate to send all of your questions our way in week 1. During that period we will still be working on solidifying the report criteria details and looking for issues that we might have missed. You will not receive a response right away but be assured that we will read them and prepare communications to go out by the end of week 1.

# Activities

1. Carefully read [**the client briefing**](https://docs.google.com/document/d/16FF6I4Q5jPqufDoi8HK1309ru3KZa1_ydUFNS-cvU94/edit?usp=sharing).
2. Disambiguate any requirements by sending an email to your client.
3. Get very familiar with the [**training dataset**](https://drive.google.com/file/d/1EJTwkK4xq3hK1e1jMgxtKL8zkr1fSTir/view). Expect to spend quite a few hours experimenting, exploring, and getting to know it.
4. Train the model that you will require for your API, and understand its limitations.
5. Produce a report that satisfies your client’s requirements, using [**this structure**](https://docs.google.com/spreadsheets/d/17qp-nz45RLLcK3KtljOLs4VEhQ0KzTaIWiddYTwgB_A/edit?usp=sharing) and [**this template**](https://docs.google.com/document/d/1gzo3ftreMr8b-8ORHwrKrS4rxBh3cFRNtJ34joDmVMA/edit?usp=sharing).
6. Deploy your model, using [**these instructions**](https://github.com/LDSSA/heroku-model-deploy)
7. Deal with the data as it arrives and ensure your API is responding successfully
8. Write the second report using [**this structure**](https://docs.google.com/spreadsheets/d/17qp-nz45RLLcK3KtljOLs4VEhQ0KzTaIWiddYTwgB_A/edit?usp=sharing) and [**this template**](https://docs.google.com/document/d/1CkDbQ5gED9jqZhvDuKnQ6hbJtPvKK60eWE05-mvaRw4/edit?usp=sharing)

# Sending an email to your client

In the real world, your requirements are never fully defined at the start of the project. You get some instructions, but it’s your responsibility to understand what is under-scoped and what information is missing.

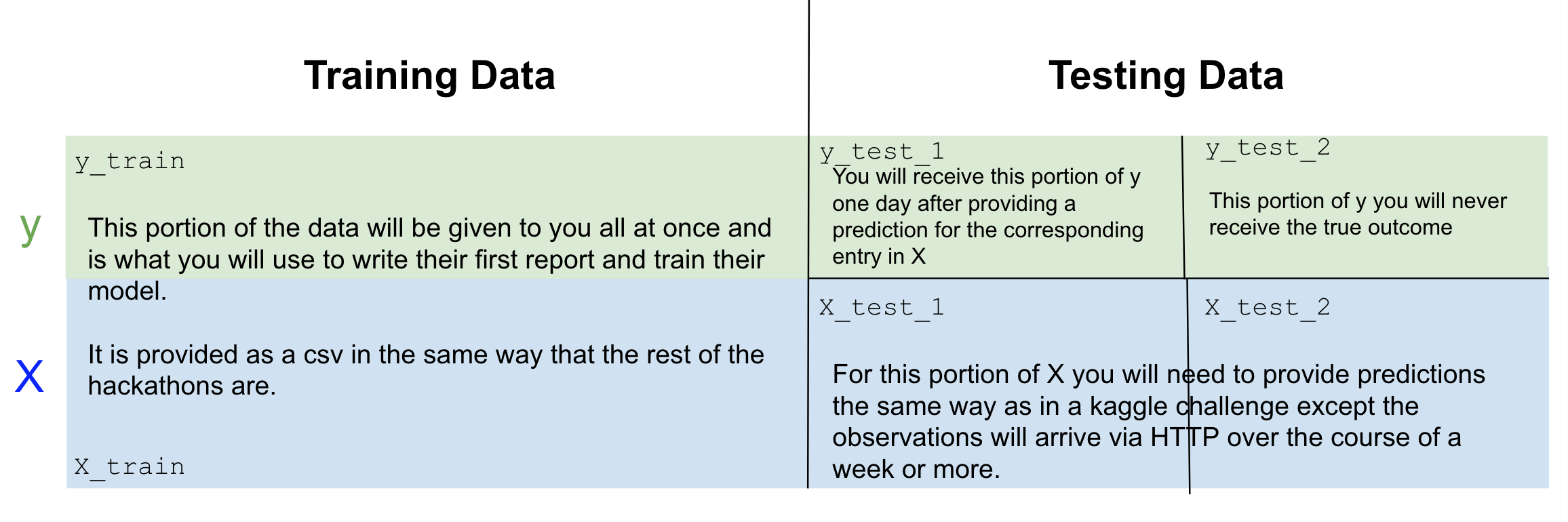
In this part of the capstone, you will understand the instructions carefully, do some exploration to understand the data, and realize where you have questions and where information is missing. You will then compile all of this into a professional “email”, and send it to your client. You will consider who you are talking to (what hints have been given about them) and make sure that you use language at the appropriate tech level.

Hint: *Do you think they will know machine learning jargon, or do you have to break down what you need any further?*

To “send” your email, please use [this form](https://forms.gle/yieQvTDWmDwm3Kby7). Your client will answer most questions on Jan 25th. Your last date for sending them an email is Jan 24th. Please send only 1 email.

# 

# Understanding the data you’ll receive



There are 3 moments when you will receive data.

**Receiving the train set:**

The first is when you receive these instructions. The email from your client will already link to your training dataset. You will, however, have to build your own target, as it’s not already in clean binary (0’s and 1’s) form.

**Receiving the test set 1:**

Later (see [timeline](#_nmoqpmg6q1q5)), the data will start flowing from the client via HTTP. You will only receive the labels (not the target).

After this data has stopped flowing, you will receive the respective targets. At this time you will be able to adjust your model and re-deploy if you feel that it’s worth updating. This model update is optional.

**Receiving the test set 2:**

Finally, the data will restart flowing, and the second test set will arrive via HTTP. You will never receive the true labels for this dataset.

# 

# How to deliver the first report

1. Write the report
2. Name the document ***Report 1 <your email>***
3. Ensure you’ve read and complied with the [Report technical details](#_w3b7lmcph12p)
4. Export the report as PDF
5. Go to the Portal → Capstone → Report 1
   1. Provisory if it’s the first submission
   2. Final if it’s final delivery with fixed comments

# 

# How to deliver your code

Your code should be delivered using [this form](https://docs.google.com/forms/d/e/1FAIpQLSfC7nXABJgogrVx5WWiNQYCGKORxocU7ZCfRX35lflU8oxD3Q/viewform) by the [deadline of report 2](https://ldssa.github.io/wiki/Starters%20Academy%20(LDSSA)/05-Capstone-Project/).

*Note: if you are having trouble filling the form, try incognito mode (which seems to fix it).*

# Success Criteria

The passing criteria are also similar to that in the professional world. We expect you to deliver something that would be acceptable by a client. There isn’t a single number we are expecting you to hit, nor is there a grader to tell you if you are right.

That will lead to a bit of subjectivity. In general, if you deliver on all the requirements with an acceptable level of quality you will pass. If you deliver something that would get you a bad performance review, you won’t.

# 

# Report technical details

In real life, your company will probably have a template or an older report that you can follow. We will be providing this template for [report 1](https://docs.google.com/document/d/1gzo3ftreMr8b-8ORHwrKrS4rxBh3cFRNtJ34joDmVMA/edit) and [report 2](https://docs.google.com/document/d/1CkDbQ5gED9jqZhvDuKnQ6hbJtPvKK60eWE05-mvaRw4/edit) to you. You can choose to directly use it or adapt it, but we ask that you respect the following:

* Keep to the sequence and titles as indicated in the model [report structure](https://docs.google.com/spreadsheets/d/17qp-nz45RLLcK3KtljOLs4VEhQ0KzTaIWiddYTwgB_A/edit#gid=0). Note that any ignored section will be an automatic fail.
* The number of pages (listed in the report structure) is a guideline, not a hard rule, but please don’t deviate too much from it. Knowing what to leave out is an important skill. In the annexes, however, feel free to go much more overboard.
* Don’t include code in the report. You will deliver the code separately.
* Size 11, Arial or some other normal font.
  + Comic sans will be the reason for immediate fail.
* Don’t forget to provide a table of contents on the first page and before delivering and make sure it is up to date.

# Hints and advice

This is a capstone. It contains data science, data engineering, and project management. Don’t worry if it feels a bit overwhelming at first, take a breath and read everything twice. Make a plan for how you will approach each challenge. Ask questions. This is going to be difficult, but you can do it!

You may find that part of this assignment contains some pretty tricky questions. For instance, you may find that every model you train discriminates against some protected group. You will most likely find it impossible to completely remove this effect. That’s how the real world works.

You may also discover that there are trade-offs where diminishing one type of discrimination actually increases another. Or that your model performance would go down on some metrics as you attempt to fix others. You may also find that as you attempt to fix true positive rates, your true negative rates will become unequal. To be clear, there is no perfect solution.

Any solution will be subjective, and we are not expecting you to find the “right one”. What we are expecting is that you are able to do your best to deal with this, and then support your decisions in an informed way.

Here is a hint for you: make sure that you always answer either true or false and that you aren’t caught off and answer np.nan. Look for edge cases. Be skeptical of assuming things will work, and look hard at your predictions on the training set, not just as aggregate numbers.