This is the main document for the Capstone, everything links from here

Capstone - instructions

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 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 the 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 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.

Success Criteria

The passing criteria is 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.

Activities

1. Carefully read **the client briefing**.
2. Disambiguate any requirements, by **sending an email** to your client.
3. Get very familiar with the training set. 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
6. Deploy your model, using **these instructions**
7. Deal with the data as it arrives and ensure your API is responding successfully
8. Write the second report using **this structure**

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. 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.

How to deliver the first report

1. Write the report as a Google Doc.
2. Name the document ***Report 1 <your email>***
3. Go to File → Version History → Name current version (name it Report 1)
4. Ensure you've read and complied with the report technical details
5. Share it with capstone@lisbondatascience.org, giving edit access
6. **Do not make any alterations after the delivery date** unless explicitly requested. Answer any comments that any instructor may make on your document.
    a. (note: Google Docs timestamps versions, so it will be obvious if changes are made after the hour)

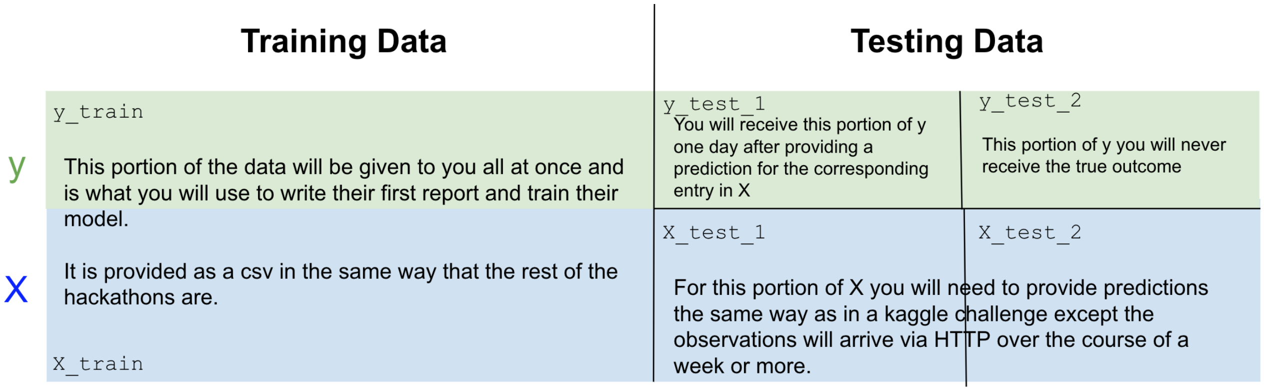
How to deliver your code

Your code should be delivered using this form by the deadline of report 2.
Note: if you are having trouble try incognito mode, seems to fix it.

Timeline

This calendar is the source of truth for all dates. It is the full timeline. The upshot for the timeline is as follows:

- The first week is for understanding the dataset and asking questions.
- At the end of the first week you will receive answers disambiguating the issues.
- Weeks two and three are for training your model, writing your first report, and deploying your model.
- At the end of week three you need to submit the first report
- In week 4 the observations will be sent to your deployed model
- At the end of week 4 you may retrain your model if you choose
- In week 5 the second set of observations are send to your deployed model
- In week 6 and part of week 7 you write and deliver your second report
- Instructors will then be given 3 weeks to grade the entire capstone which includes both reports and your code



Training Data

Testing Data

y

X

y_train

This portion of the data will be given to you all at once and is what you will use to write their first report and train their model.

It is provided as a csv in the same way that the rest of the hackathons are.

X_train

y_test_1

You will receive this portion of y one day after providing a prediction for the corresponding entry in X

y_test_2

This portion of y you will never receive the true outcome

X_test_1

X_test_2

For this portion of X you will need to provide predictions the same way as in a kaggle challenge except the observations will arrive via HTTP over the course of a week or more.

Data

There are 3 moments when you will receive data.

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 0 and 1 form.

Test set 1:
Later (see timeline) 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.

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.

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.

You will be building your own target, which is new in the Academy, but very frequent in the real world. There is of course an objective truth, but here is a hint: 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.

Report technical details

In real life, your company will probably have a template or an older report that you can follow. Here you will have a few more degrees of freedom, so common sense is the rule. We ask that you follow the following:

- Keep to the sequence and titles as indicated in the model report structure.

- 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.

- Use the Google Docs titles, making sections Title 1 and sub-sections Title 2.

- When you are done auto-generate a table of contents in the first page (Insert → Table of contents). Pro-tip:
    - Title 1: command-alt-1
    - Title 2: command-alt-2

Miscellaneous Comments

We are still working on benchmarking and solidifying the report criteria details during the first week that you have the dataset. Please keep in mind that during the first week we will also be looking for issues that we might have missed so send all of the questions that you've got our way. You will not receive a response to them right away but be assured that we are reading them and preparing communications to go out at the end of the first week.