Capstone Report 1 - instructions

Overview

You were given [this briefing](#) by your company. After that, there was [this exchange of emails](#) between you and your boss. Now it is time to write the first of two reports, before the data starts flowing.

You will notice that the report structure contains a section called "reader". That is because every sub-section has an intended audience. We do not want to have the chief of police reading about the depth of your random forest, but the model decisions need to be well documented. At each section, bear in mind who the reader is, and address them at the correct technical level.

Finally, remember that we want to test your skills as a professional data scientist. As in real life, some of these questions are slightly subjective. Answer as you would answer in a professional environment, rather than in an academic one. As long as your report is professional, does not contain glaring errors, and covers all the topics we expect you to catch, you will pass.

Note that you can pass even if you don't achieve every requirement your client asked of you, as long as you've documented that this is not a reasonable requirement, and proposed informed alternatives. In real life, this will happen a lot.
If you have questions feel free to ask them, as would occur in a professional setting.

Activities

1. Conduct analysis of the dataset, to answer [the client briefing](#)
2. Train a model, and understand its limitations
3. Deploy it, using [these instructions](#)
4. Write the report using [this structure](#)

How to deliver the 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. Share it with pedro@lisbondatascience.org, giving edit access
5. 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)

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 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

Really tricky questions

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, as the data is biased by design.

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.

Timeline

[Here](#)

Important links

- [Deployment instructions](#)
- [Client briefing](#)
- [Q&A](#)
- [Report structure](#)