We got your last project submission and will review it soon. In the meantime, you can already start thinking about your capstone project by accessing the instructions below. Please note that all your projects have to be completed (i.e. submitted, reviewed and accepted) before being eligible for the capstone.

Project Pending: Applied Machine Learning 2 project (last submitted 12 December 2019 21:12)

01. Proposal

Content Resources 1 Questions 5

In the capstone project, you will propose a data science project based on a real-world problem. Your goal is to showcase the knowledge and skills that you have developed throughout the four courses of this program. At the same time, you are encouraged to choose something that truly interests you. If you are stuck for ideas, we provide a list of some public data sets at the end of this page to provide inspiration. Feel free to book a 1-1 to discuss your ideas and get feedback on them.

Capstone project requirements

The capstone project will assess whether you have successfully met the learning objectives of the course

DATA PREPARATION

- Collect and import the required data sets
- Demonstrate good understanding of the data and its structure (values, encoding, format, etc.)
- Prepare the data appropriately for the analysis and modeling (data cleaning, manipulation, feature engineering and encoding)

EXPLORATORY DATA ANALYSIS

- Demonstrate a deep understanding of the data specific to the project goals by performing detailed exploratory data analysis including all necessary descriptive statistics and visualization methods
- Discuss insights and potential difficulties to justify the proposed implementation
- Discuss data quality and completeness with respect to the intended project goals

MACHINE LEARNING MODELS

- Identify and discuss suitable machine learning models, baselines, metrics and evaluation strategies
- Tune the different models, analyze their performance
- Discuss the results and potential trade-offs (complexity, interpretability, computational resources) using the appropriate terminology

COMMUNICATION

- Use data to tell a story and discuss the value of the results. Highlight interesting insights through discussion and visualizations
- Document and discuss each step of the analysis to support the approach and implementation
- Present and evaluate findings, discuss results in the context of the project goals

The submission is done in two stages: first the proposal, and second the actual project. Below we outline the details for submitting your project proposal.

Overview of the Capstone Project process

To help you better plan, here is an overview of the capstone project process

1. **Write and submit the proposal (now)** as described on this page. Don't hesitate to reach out to us to check that it fits the project requirements and that the proposal document is detailed enough. Finalising the proposal often requires multiple submissions. So don't worry if we return the proposal with a request for modifications
2. Once your proposal is reviewed and accepted by our team, **work on the project itself**. Again, don't hesitate to reach out to us to discuss your choices and intermediate results. However, note that we cannot provide direct help (code related, debugging) during the implementation as this is your final graded project
3. **Submit the project**. Note that you can only submit your project once, so make sure it's ready!
4. Once your project is reviewed by our team: **book a 1-1 video call for the defense** via the invitation link that we will send you by email
5. Finally, **present the project** during the 1 hour-long final defense 1-1

Project proposal

For your capstone project you will select the data and define the data science problem you intend to solve with this data. In your proposal you will provide a detailed roadmap for the execution of your capstone project. It will outline the analysis that you plan to perform using the data science methods and tools covered in the courses of this program.

Your proposal will include what kind of data cleaning and preprocessing will be required, what data analysis will be carried out, what kind of machine learning methods you plan to apply, and how you will assess the performance of your methods or models. Moreover, your proposal will identify the main challenges, outline how you plan to address them and demonstrate a clear awareness of the feasibility of your project goals.

Communication is a key component of the project proposal and the final capstone project. For both submissions you should

- Use Jupyter Notebooks
- Structure your work using sections
- Document your code, workflow and results with detailed comments in Markdown cells
- Use appropriate means to communicate your insights and results ex. tables, figures ...
- Communicate your observations, identify problems and obstacles and discuss potential approaches to resolve these issues
- Support your planned strategy with insight gained into the data

In the **review of your proposal**, our aim is to minimize the chances that you encounter major problems down the line and to ensure that everything is in place for you to successfully complete your capstone project. Hence it may take a few iterations with us to finetune your proposal.

In order for us to be able to provide you with constructive feedback and evaluate your proposal, please provide detailed information on the following points in the **proposal template notebook** that you can find in the resource section.

1) THE PROBLEM

Provide a non-technical but sufficiently clear formulation of your problem

- What is the wider context of your problem and what story you would like to tell with the data?
- What problem would you like to address?

2) THE DATA

Each data set is different and each one of you has different questions you would like to address for a given data set

a) Please provide a detailed overview of the data including

- Source and context of the data set, provide the relevant links

- A small sample of the entries, features, values
- Number of features and samples
- Information captured by the (groups of) features
- Encoding of the features
- Granularity/details of the data
- Quality of the data: missing, incorrrect values, data completeness
- If you are using data close to your domain of expertise, please give us some insight into the use of the features
- If your are collecting the data yourself (ex. web scraping, personal recordings) or don't have access to the full data yet, please provide already a representative sample

b) Discuss your plan for

- Managing your data ex. storing, file formats, database usage
- Data cleaning and data manipulation
- Feature engineering
- If you are combining multiple data sets please comment on the following
    - How do you plan to combine the data sets
    - Compatibility of the different data sets: encodings, standards, units, etc.
    - The features you will select from each data set

3) EXPLORATORY DATA ANALYSIS (EDA)

You should demonstrate that you are familiar with the data and its properties

a) Include a preliminary EDA

- Provide descriptive statistics and informative plots on your features
- Explore relationships amongst the features themselves and with the target
- Check for missing values and outliers

b) Discuss how the EDA informs your project plan

- What interesting patterns do you observe in the data? What problems do you identify?
- What insight can you gain to inform your data processing and your modelling?
- Clearly state your observations and discuss their impact on your project
- Discuss possible approaches and your resulting decisions for the project plan

c) Based on this preliminary EDA, you should

- Outline your plan for further EDA in the project

4) MACHINE LEARNING

Remember that there are many interesting ways to approach a machine learning question

- Compare different approaches to a problem
- Use different features spaces
- Use different model types
- Fine tune specific models

a) Phrase your project goal as a clear machine learning question

- What is your intended outcome in machine learning terms?
- What are the features and the target variable that you are using?
- Is your question a regression or a classification problem?

b) The models you plan to use

- Which models are you planning to use and why you are choosing them (model interpretability, suitability, scalability, diversity, ...)?
- A ranking of your approaches: priority, optional or "nice to add"

c) Your proposal should clearly communicate your strategy for the machine learning part

- Preprocessing steps of your data for each machine learning model
- Methodologies to be used to train and finetune your models
- Your baseline model
- The metrics and methodologies you are considering to evaluate and compare your models
- You should clearly justify any of the above

Feel free to use models that have not be covered in the course. We might ask you to implement a simple version of the model to make sure that you are suitably familiar with this model ahead of starting with the project.

Scope of the project

You might need to use additional tools or libraries not covered in this program during your project - which is fine, we encourage you to continue learning about each step of the data science pipeline on your own, read more about data science and machine learning, discover new tools and build upon the solid foundations that you have acquired during this program. However, keep in mind that your capstone project should demonstrate that you have clearly **mastered the skills, tools, and techniques from the course** at every level of the data science pipeline. Also, note that we might not be able to provide support for topics outside the scope of the program.

Support during the project

Feel free to book a 1-1 with our team at any stage of the proposal to discuss your ideas, problems and plan. Finetuning the project proposal usually takes a few iterations and 1-1s.

However, note that we cannot provide direct help (code related, debugging) during the implementation as this is your final graded project.

Project submission and project defense

The final project will be a well documented and well structured Jupyter notebook containing all your discussions, analyses and models and telling the story of your project. Note that you can use your work from your project proposal as the starting point for your project.

The capstone defense will be done online via video conference. Once we have reviewed your capstone project we will send you a link to book the 1-1. During the project defense you will be given around 30 minutes to present your project and your results and explaining your choices and approaches. This will be followed by questions and more detailed discussions with our team.

Data sets

We encourage you to **combine** several **data sets** from different sources to enrich your analysis and demonstrate that you have acquired the skills to complete each step of the data science pipeline (cleaning, merging, analyzing, presenting the data).

WORKING WITH SENSITIVE DATA

If you are working with sensitive data, e.g. from your employer, please consider the necessary anonymization steps ahead of using the data for your project. If you need a non-disclosure agreement (NDA) please contact our team at **nda@extensionschool.ch**

EXAMPLES OF PUBLIC DATA SETS

Here is a list of a few publicly available and local data sources - feel free to suggest others

- The opendata.swiss data portal

- Data from the Swiss Federal Statistical Office (FSO)
- Google Dataset Search
- The World Bank Open Data portal
- Geo data - ex. openstreetmap.org, openflights.org ..
- Gov. portals - ex. data.gov, data.gov.uk, NYC opendata, Chicago data portal ..
- Agencies portals - ex. NASA's data portal, api.nasa.gov, open.fda.gov, ..
- Medias - ex. Swiss SRG SSR, theguardian.com, ..
- Wikipedia.org and dbpedia.org
- Google Earth Engine's public data catalog
- The Metropolitan Museum of Art Open Access
- TCdata360 - Open Trade and Competitiveness Data
- Other platforms: discogs.com, goodreads.com, ..

We encourage you to apply what you've learned to your own data - ex. data loggers, connected devices and so on.

You can also use data sets from open source distributors like kaggle.com. However, clearly explain in your proposal how your analysis is different from the other analyses available online.

RESOURCES

project-proposal-yourname.ipynb.zip 1.02 KB

QUESTIONS

Ask a Question

Joker · Learner · 2 months ago

✓ 2 Answers

Hi teacher, I got trapped with this statment: pd.set_option('display.max_rows', None) !!!

One other line of my code call one giant df and now my file.ipynb is up to 130MB, then it crashes anytime I try to open it. My computer runs out of memory and the Jupyter session crashes.

I have converted .ipynb to a .txt. And I open it but it's an infinite file. more than 10^6 lines. Reason why it crashes I guess.

Any suggestion to recover that notebook ? Thx

POST ANSWER

Michael Notter · Teacher · 2 months ago

✓ 1

Yes, `pd.set_option('display.max_rows', None)` can be very useful if you want to see the full extend of your dataset. But, I would rather use it with `max_columns` . For the full number of rows you could otherwise always use `.head(-1)` or just `.head(1000)` for the first 1000 of rows.

Concerning the issues at hand. Yes, jupyter has a lot of problem opening such huge files. My first appoach would also be to open the notebook as a text file and just remove a lot of these "spamming" lines to reduce the size of the notebook.

Alternatively, you can also try to use some code to clear the output of the cells. But I cannot guarantee that this will work on your notebook, as I'm not sure if it will not also lead to memory issues. Having said so, try out the following code:

```
# Import nbformat (if missing, install with 'conda
import nbformat

# Path to borken jupyter notebook
nb_filename = 'notebook_broken.ipynb'
nb_filename_out = 'notebook_fixed.ipynb'

# Read Notebook
with open(nb_filename, 'rb') as nb_file:
    txt = nb_file.read()
nb_content = nbformat.reads(txt, nbformat.NO_CONVER

# Clear Notebook (delete output of cells)
for cell in nb_content['cells']:
    if 'code' == cell['cell_type']:
        if 'outputs' in cell:
            cell['outputs'] = []
        if 'execution_count' in cell:
            cell['execution_count'] = None

# Rewrite Notebook
nbformat.write(nb_content, nb_filename_out)
```

Joker · Learner · 2 months ago

✓ 1

A BIG CONGRATS !!!

It worked ! Impressive, the NB is back to 29.7kB. Well done ! Thanks a lot.

Joker · Learner · 3 months ago

✓ 1
Answer

Hi teacher, about the topic source, would EPFL help to provide a real case from a company?

It is nice to solve a local issue as we may know better the data. But it is also a nice opportunity to get known by an other company and demonstrate our new skills. This could lead to a internship or a job.

If it is not yet the case I would suggest it to the EPFL E. S. team. Thx.

POST ANSWER

Michael Notter · Teacher · 2 months ago

✓ 1

Hello. No, we don't provide any real case dataset from companies for the capstone project. This is due to many things. One of them is that there are already many open source and freely available datasets online to be found. Another one is that some datasets from companies might be very noisy or complex which means even more data preparation, cleaning and requires in some cases advanced expert knowledge. We encourage you to use a dataset from your company, hobby, environment, … as the best motivation comes from a personal interest in a challenge.

Thank you for your suggestion. I agree, creating a closer relationship between you and companies would be a great idea. The obvious pros of such a relationship would be great, and we will certainly think about such a setup. But the question would be, in what way does such a setup differ from something like kaggle.com? Why is it lucrative for a company to give such a dataset to only a few EXTS learners and not the whole internet? Who is curating these datasets? How can we make sure that these datasets are not too complex or too distant from what was taught in the course? Etc.

mitch.buchannon · Learner · 8 months ago

✓ 1
Answer

Suggestions for datasets for the final capstone project

Hello, I've been thinking about the final capstone project and didn't come up with a good idea even after browsing Kaggle. Have you got any suggestions? I'm interested in topics like economics, energy systems, climate change, life science, logistics, supply chain management. If possible I'd like to answer some meaningful questions with the project rather than to do it just as an exercise.

POST ANSWER

ChristianLuebbe · Teacher · 8 months ago

✓ 0

Hi Mitch,

We have a list of data sets on the proposal page. Scroll down to the bottom.

These days many countries and NGO host open data platform or provide access to their data sets on their websites.

For finding questions: Think about what really interests you about these topics? What would you like to know in more detail? Start with a few broad questions. Then check the data sets available and how you can combine them. As you explore the data sets you can narrow down your questions. Don't be afraid that your first idea has to be the one.

I hope this helps

BRZ · Learner · 9 months ago

✓ 1 Answer

Documentation of the project

Hello, Am I allowed to present the documentation of the project in a separate file (PDF) ? Of course the code will also be locally commented but the analysis/explanations and the files structure would be presented in the PDF file.

POST ANSWER

Frederic Ouwehand · Teacher · 9 months ago

✓ 0

Hello. Yes, you can include additional documentation in PDF format or markdown files to the GitHub repository.

Ghazaleh dandelion · Learner · a year ago

✓ 1 Answer

Dear sir/madam I've uploaded the fourth project of the data science/machine learning course and just the "capstone project" is remained. As you know I have to pay a monthly fee for the course and I want to know if I should pay that fee now that I've finished the courses. What's the deadline for the upload of "Capstone project" to not to be required to pay the fee for the coming month? Regards Ghazaleh Eslami

POST ANSWER

linda.farczadi · Teacher · a year ago

✓ 1

Hello, the capstone project is a very time-significant part of the program. The general timeline is as follows: you upload a proposal, we review the proposal and either accept it directly or ask for some updates (it is quite common at this stage to have several 1-1s in order to really clarify the scope of the project and ensure that everyone is on the same page). Once this has been accomplished and we are confident that you have a strong basis for starting the project we give you the go ahead. Then once you complete the project you upload it, we review it, and we invite you to book a date for a defense of the project. In this final step we ask you to give a 30min presentation of the project and then we have another 30min or so discussion where we can ask you to explain some of your choices and demonstrate the knowledge you acquired in the project. Once this final defense has successfully occurred, we can then accept your capstone project. This usually happens the same day as the defense. At this point, (when the capstone is accepted) you would no longer need to pay any monthly fees. As you can see it is hard to put a direct timeline at the moment since there are still quite a few steps that remain.