MAIS 202 - PROJECT DELIVERABLE 1

Due: October 8th, 11:59 PM

Over the course of MAIS202, you will be completing a machine learning-based project of your choice for the final project. You will also demo your project by integrating it into a webapp.

Submission

The deliverables may be completed individually or in teams of 3-4. Keep in mind that the project scope and grading will take into consideration the size of the team. All deliverables should be electronically submitted on Github. They should also be completed with the same academic integrity and standards expected at McGill University. Include appropriate citations.

To submit:

- Create a repository on Github (one per team),
- Push your deliverable there as a pdf file: “Project Proposal.pdf”
- List your/your team’s repository link in this spreadsheet Final Project Teams.

Note: All of the code and deliverables for this project should be found in this repository. Make sure to maintain it with properly documented README and structured code.

Expected length: 1 page

Forming a Project Idea

1. **Choose the dataset that you want to work with and propose your project idea.**

   Suggestion: Kaggle is an excellent platform with thousands of datasets you can use. You can use its search bar to search for datasets that are related to your interests. (For example, if you’re a foodie, you can type in “Food”, and you’ll see various datasets on food items, cuisines, etc.). Use its filter feature to sort by “Usability” to get datasets that are well-curated and easy to use. Google Dataset Search is also a great tool!

   Furthermore, you can also look into creating your own custom dataset by scraping websites. If so, explain what kind of data you will be scraping in your deliverable.

   A list of pre-approved project ideas can be found here: MAIS202 Pre-approved Project List.pdf . Note that the ideas on the pre-approved project list are first come, first served (no duplicates). We strongly encourage bootcampees to first explore datasets and project ideas that fit their personal interests before turning to pre-approved project ideas because the original project ideas that bootcampees come up with are almost always better than our pre-approved ideas.

2. **Get in touch with your assigned TPM.** Once you have an idea, discuss it with your assigned TPM (you can find out who's your assigned TPM on the spreadsheet once you have signed up). The two of you will make sure that the project is doable. If you're working in a team and your teammates have different assigned TPMs, don't be shy to reach out to them as well to get different points of view.

3. Once your project has been approved, **you can begin writing up your deliverable**. Don't forget to sign-up at Final Project Teams!

Deliverable Description (Content of your Deliverable)

1. **Choice of dataset**: Explain the reasons why you choose this dataset. If you are going to create your own custom dataset, explain what kind of data you will be scraping.

2. **Methodology**: Describe how you plan on approaching the project. This should be a high level overview of your plans, and this will allow us to judge the feasibility of your project. Be as thorough as you can, so we can give you critical feedback if necessary.

    a. Data Preprocessing: Is the dataset you chose feasible? What information provided is/are the most useful? How are you planning on preprocessing the dataset to extract this information? You can take a look at these F2019 slides on data preprocessing.

    b. Machine learning model: What do you want to predict/estimate from this dataset? Propose a machine learning model/algorithm for it, and explain your reasoning. Have you considered other alternative models? What are the pros and cons?

    Note: We are aware that at the point of Deliverable 1, many machine learning models have yet to be covered in lectures. To find out what models are generally used by the ML/AI community to tackle your problem, search online or ask your assigned TPM!

    Suggestion: Each Kaggle dataset has a "Code" tab that contains previous work done using the dataset. It can be an excellent source of inspiration for this section!

c. Evaluation Metric: Analysis requirements differ in every field, but some things to consider reporting include but should not be limited to:

   i. Confusion matrix and accuracy/precision-recall/logistic loss (classification problems).
   ii. Mean squared error (regression problems)
   iii. Rand index (unsupervised models)
   iv. BLEU score with brevity penalty (text generation)
   v. Variance of the dimension reduced set vs variance of the initial dataset (dimensionality reduction/PCA)

   If you are not sure, ask your assigned TPM.

   Furthermore, you should be able to explain the specific problem's accepted metrics. Keep track of the average baseline results which you hope to beat (For example, predict X with at least Y % accuracy).

3. **Application:** We hope you integrate your model in a simple landing page webapp. For those of you who have experience, you are welcome to integrate your model in more sophisticated technologies (eg. mobile, hardware, webapps).

   In this section, give the general idea of your application:
   - What does the user input? How does the user provide inputs? (Is there a webcam? A way for users to submit images? text?)
   - What does the user receive as output, and how will the output be displayed?

   Helpful: We hosted a workshop on webapp deployment in Fall 2021. You might find the materials under the webapp deployment app within mycourses useful!