CSCI - 6409 - The Process of Data Science - Summer 2021

Final Project

The submission must be done through Brightspace. Due date and time as shown on Brightspace under Assignments.

- The detailed requirements for your writing and code can be found in the evaluation rubric document on Brightspace.
- Questions will be marked individually with a letter grade. Their weights are shown in parentheses after the question.
- The project can be done in a team with a maximum of 5 students per team. Only one of the students of the team must submit the project on Brightspace.
- Create a repository on GitHub to share the code with your teammates and include the link of your code in the PDF report. At the end of the course, you can make it public if desired to showcase your work .
- Your submission is a single PDF with the report. The file names should be:
  - **FP-<name1>-<name2>-<name3>-<name4>-<name5>.pdf**

The final project is a team project that gives you the opportunity to apply what you have learned during this course to your domain of interest. You are free to work on any topic as long as the contribution is commensurate with a graduate course and the content has not been taken from elsewhere. Please follow the steps below and make sure to do each step for your final project (Due Date: July 26th, 2022 at 11:59 pm).

The project purpose (progress report) should include (part of steps 1 and 2):

1- A description of the business/research problem. Include a brief motivation (why the problem is important) and what questions/problems are meant to be answered with this work. You are allowed to work on any topic.

2- A brief description of the datasets involved and how you have acquired such data.

3- References of prior work/solutions that tackle the problem either partially or completely.

4- A description of your plan in how to collect and select features, the pre-processing, and potential data science solutions that can be applied based on the references you found and/or your own ideas.

5- Provide a schedule planning of your project with the milestones that will allow you to finish the project in time.

6- The design and development of your project must follow the steps present in CRISP-DM and you must make them explicit in your report.

7- Include some preliminary visualizations if possible.

Due Date: June 12th, 2022 at 11:59 pm. The progress report should be no longer than 4 pages.

1. Data understanding and preprocessing (0.2)

In this step, you will focus on collecting and preprocessing the data needed to address the business/research. You should take into account the following points:

1- Data must be big enough to address your data analysis needs. In case you are planning to apply supervised learning, make sure that you have a target variable (or labeled data instances).
2- You should provide visualizations for each step to facilitate understanding of the data and identify data issues.
3- Describe any data cleansing/transformation that was applied to your data.
4- To explain your data/datasets, make sure that the selected datasets are covering the objectives of your research. If you need more than one dataset, feel free to add new datasets, but do not forget to explain how you integrate the datasets.

2. Literature Review (0.2)

After selecting your topic, you must conduct a research literature review to find available/possible solutions that are related to your problem. Make a list of available/possible solutions and report on the following points:

1- A brief discussion of the prior work applied to address the problem. Do not forget to include any references that you are discussing in your project.
2- List the strengths and weaknesses of each solution. You can provide evidence from their reports (references) or based on your own judgment.
3- Select those approaches that you will use to address your problem. These can come from the references you have found, your own innovative approach or a combination of the two.
4- Provide a justification of why each of the selected solutions addresses the problem set for your project.

3. Description of the solution (0.2)

Either if you selected an approach from your literature review, you came up with your own solution, or a combination of the two, you must explain it in detail here. Pay attention to the following items:

1- Explain the solution in a general way and draw a big picture of the solution for your reader. You can include some high-level diagrams, such as a data flow diagram.
2- Explain each component of the big picture. Make sure to explain the objectives and rationale for each component to be part of the solution.
3- Use some sample data and explain how your solution addresses your problem using that sample data.

4. Data Analysis, Results and Evaluation (0.2)

Here you explain your data features, and any pattern or model that is inferred from the data based on your model. Also, show how you evaluate your model and discuss the results obtained.

Then, evaluate your model, providing all the analysis performed, insights gained and conclusions found in your report. For each experiment, design it properly to find fair ways for evaluating your model. Discuss how your model compares to other models. Also, discuss the election of features and hyperparameters. Explain the main metric(s) used for evaluating your approach and justify their selection. Important notes:

1- Check if the data preprocessing was suitable for the dataset and models used.
2- Check if the target feature is balanced or imbalanced.
3- Check if the evaluation metric is suitable for the data and models.
4- Do a list of what are you analyzing and the conclusions you can draw from that analysis. This will help organize your thoughts.
5- You can create a section for data analysis and another for model analysis.
6- Don’t forget to include visualizations to improve the understanding of the model and/or any insights found.

5. Conclusions (0.2)

In this step, you have to explain the problem and the solutions employed, but from a business perspective. Discuss your conclusions, insights, and findings as if you were presenting them to your client/stakeholders, showing how your approach addresses the business problem. Feel free to add future work that can be done to improve your methods or to solve issues faced during your development, i.e., potential solutions for weaknesses that you found in your selected approach. Notes about future work:

1- State the limitations of your approach.
2- Find potential solutions for these limitations.
3- Discuss other possible improvements for your approach.

Write a report based on your work. Include all the items mentioned above in your paper.