Final Project

CIS 5450: Big Data Analytics

Spring 2023

Introduction to the Final Project

The objective of this project is for you to show that you are able to apply the topics in this course in a real-world setting. You are welcome to use concepts beyond the scope of the course in addition to concepts covered in the course; however, a major portion of your project needs to be based on this course's material. You are not allowed to reuse a project from another course. This project is a good opportunity to create something that you could add to your resume, post online in your personal GitHub, or share on LinkedIn.

Projects will be done in **groups of 3 people** — no exceptions. Make sure all team members are in time zones where you all can work together without problems. If you are unable to find teammates, a good place to start is the Ed Discussion Teammate Search Thread (Spring 2023 Link)

Each group will be assigned to a **project TA.** Your project TA is your mentor for the project and will be the primary person responsible for grading your project (but not the only one). You are **strongly encouraged** to attend your TA's office hours. You should reach out to them directly (private ed or via email, emails are on the course website) if you have questions or would like to schedule a time to meet. If you are choosing deliverable option 4, you will need to schedule your final presentation with this TA. Your TA will be assigned **after the proposals are submitted.**

- Note: you may request a TA and we will try to accommodate your preferences, but note that we cannot guarantee anything especially if some TAs are requested by more groups than they can take on. In such cases, we consider them on a first-come-first-served basis so it is advantageous to submit your proposal ASAP if you really want a certain TA!

You can either choose:

(1) a staff-proposed dataset
(2) propose your own dataset/project.

Either way, each group must submit a single project proposal (more information below).

Project Deadlines & Components Overview

Your final project is expected to have the following components at minimum:

1. Introduction/ Background
2. EDA (Exploratory Data Analysis)
3. Modeling
4. Description of Challenges / Obstacles Faced
5. Potential Next Steps / Future Direction

This project has **3 deliverables** in total:

1. Proposal: **due March 22nd 2023 by 11:59 PM EST**
2. Intermediate Check-In Meeting: **between April 10th - 16th 2023 by 11:59 PM EST**
3. Final deliverable *(see options below)*: **due April 26th 2023 11:59 PM EST**

Logistical Notes:

- This is a 3 person group project (no more, no less).
    - **No collaboration outside your group is allowed.**
- Again, please use the Ed Discussion Team Search Thread (Spring 2023 Link) to find group members, taking into account potential time zone differences.
- There are **NO late days** for any components of the project. *(We need the full time to grade your project!)*

Project Proposal Guidelines (due on Gradescope by 03/22)

Each group is required to submit a **project proposal** by the date indicated above. This is required for all groups, including those using staff proposed datasets.

Your proposal should be roughly **300 words** in total and contain the following sections:

1. All Group Members' Names & Duties:
    a. Breakdown of what each member will be responsible for. **This is your best method for ensuring everyone in the group will contribute equally!**
    b. We encourage you to use the teammate search on Ed to form teams. **You need to have your team formed to submit the proposal.**
2. Data Source: regardless of whether it is one of our suggested sources, or one that you are proposing, you must still explicitly link us to it.
3. Project Plan
    a. Explain what you intend to study with your project.
    b. What is the ultimate objective?
    c. What types of models are you considering?
4. Value Proposition: why is this project interesting?
5. Anticipated Obstacles & Challenges: what challenges and obstacles might you anticipate with this project?
6. Requested TA: TA name if you would like to request a TA — note that groups who propose their own datasets, especially those who submit earlier rather than later, will have a say in which TA they would like to get assigned to. These will not be guaranteed though.

While your project can certainly vary from the proposal that you submit, the proposal component is designed to help make sure that you are on the right track and so you can receive feedback as you go.

IMPORTANT!!!

- Only one student from each group should submit this assignment on Gradescope (will be released closer to the proposal deadline), but make sure it contains the name of each person in the group.
- When you submit to Gradescope, make sure to use the Gradescope feature to add your group members to the submission as well (this can be done after submitting the assignment) otherwise your teammates will not get a grade in our gradebook!

Intermediate Check-in Meeting **(during 04/10 - 04/16)**

During the intermediate check-in week, each team will schedule a Zoom meeting with their project TA to discuss the work they have done on the project and work still left to accomplish. This can either be during the TA's Office Hours or during an agreed upon time with all group members and your assigned TA. ***It is your group's responsibility to schedule this in a timely manner, not your Project TA's!***

By this point, we expect you to have completed the EDA and data wrangling part of the project as well as have thought about the modeling process. This meeting shouldn't take more than 20-30 minutes as we will be making sure you are on the right track and answering any questions you may have about your specific project.

Final Week Deliverables

Each group must submit the following (submission details will be posted soon):

Deliverable #1: Code Notebook(s) [DUE April 26th 2023 11:59 PM EST]

- Every group ***must*** submit a complete notebook or notebooks with all of the code used throughout the project.
- This notebook should be reasonably organized and the code should be readable (your TA should be able to understand what you are doing in any part).
- If you are doing a **presentation, article, etc.** (instead of an annotated notebook) as your Final Deliverable option, you must submit your code notebook by this date (4/26).
- If you are doing an **annotated notebook** as your Final Deliverable option, you must also submit your notebook by this date. You can just submit 1 notebook on the Final Deliverable deadline (4/26), so long as this one notebook contains all of the code that you used for your project.
    - Alternatively, you can submit multiple notebooks. i.e., have one that contains all of your code, then make a copy and in the copy, remove any uninteresting parts/ code and add analysis and clean it up.

Deliverable #2: Final Deliverable [DUE April 26th 2023 11:59 PM EST]
Please pick one (or more if you'd like) options from the following. We'd like to encourage you to create something here that you can use as part of your portfolio as a data scientist!

1. Fully annotated notebook ***(highly recommended)***
    - Think of a notebook similar to the homework assignments. Your notebook would be broken up into separate sections via Markdown sections and should be **very readable.**
    - Each section should have an introduction and some description of the findings. You should describe the analysis that you are performing and the results. Interpret any important visualizations.
2. Blog Post: similar to Towards Data Science or Medium
3. Live Presentation (5-10 min): scheduled with your assigned TA and all group members
    - *Note: If you choose this option, we can accommodate a live presentation during reading days 04/27 - 04/30, but all of your technical work will still need to be submitted to Gradescope by 04/26 so no changes can be made after that date.*

IMPORTANT!!! Like the proposal, only one student from each group should submit this assignment on Gradescope and you should make sure to use the Gradescope feature to add your group members to the submission as well.

Project Rubric *tentative, specifics may be adjusted*

Criteria	General Description	Points (/100*)
Project Proposal/ Intermediate Check-in	A brief high-level description of what you plan to do and what are your plans for the project	5
Difficulty	Have you attempted challenging analysis? How much time would have been required to complete your project?	10
Code Quality/ Readability	Is your project notebook understandable and fairly well broken into modular steps?	10
Creativity/Uniqueness	Does your project stand out from the rest? Think about how you could make your project relevant and try to perform analysis that is not obvious or has not already been done.	10
Visualization	Visualize your findings well using plots and graphs. Make sure that they are informative and appealing. We encourage exploring packages such as tensorboard, plotly, bokeh, seaborn, on top of the traditional matplotlib.	20
Modeling	Is your model useful? Is your model implemented correctly? The models that you choose should be justified. You are encouraged to explore and implement more than one modeling method.	20
Application of Course Topics	Is your project built around the topics discussed in class, recitation, or covered on the homeworks? You are welcome to go beyond the scope of the course, but you should apply a significant amount of the course topics.	20
Quality of Deliverable	Presentation matters! Is your final product clean and polished? Have you created a deliverable that is engaging and informative?	10

*There are 105 points in total, but the final grade will be out of 100 (...because we are nice 🙂)

*All students will fill out a google form at the end of the project to discuss each group member's contributions as well as any additional information we need to know.

Course Staff Proposed Datasets

1. arXiv Dataset - For nearly 30 years, ArXiv has served the public and research communities by providing open access to scholarly articles, from the vast branches of physics to the many subdisciplines of computer science to everything in between, including math, statistics, electrical engineering, quantitative biology, and economics.
2. Health Nutrition and Population - HealthStats provides key health, nutrition and population statistics gathered from a variety of international sources. Themes include population dynamics, nutrition, reproductive health, health financing, medical resources and usage, immunization, infectious diseases, HIV/AIDS, DALY, population projections and lending. HealthStats also includes health, nutrition and population statistics by wealth quintiles.
3. COVID-19 Open Research Dataset - CORD-19 is a resource of over 200,000 scholarly articles, including over 100,000 with full text, about COVID-19, SARS-CoV-2, and related coronaviruses. This freely available dataset is provided to the global research community to apply recent advances in natural language processing and other AI techniques to generate new insights in support of the ongoing fight against this infectious disease.
4. Earth Surface Temperature Data - The Berkeley Earth Surface Temperature Study combines 1.6 billion temperature reports from 16 pre-existing archives. It is nicely packaged and allows for slicing into interesting subsets (for example by country). They publish the source data and the code for the transformations they applied. They also use methods that allow weather observations from shorter time series to be included, meaning fewer observations need to be thrown away.

Please get started early, and feel free to come to office hours or post on Ed with any questions. Good luck!