CIS 5450: Big Data Analytics
Spring 2024, Online
Prof. Zachary Ives

Big Data Analytics: Final Project

Description of the 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. This project is a good opportunity to create something that you could add to your resume, post online, or share on LinkedIn.

Projects will be done in groups of 3-4 people - no exceptions. 3-member groups are preferred, however if you want to have 4 people in your group, the project should be ambitious enough to distribute the work among all of the team members. In other words, you should not be doing less work because you are in a team of 4 than the students who are on a team of 3. Although not mandatory, it would be best to have all of the team members in time zones where you all can work together without problems. **Each team member should have clearly defined and agreed-upon roles and responsibilities.**

Each group will be assigned to a **Project TA.** Your Project TA is your mentor for the project and is also responsible for grading your project. You are **strongly encouraged** to attend your TA's office hours. You should reach out to them directly (via private Ed Discussion posts) 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.**

You can either choose (1) a staff-proposed dataset or (2) to propose your own dataset/project. Either way, each group must submit a project proposal (more information below).

Your 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 4 deliverables in total:

1. Proposal: **DUE Monday of Module 12 by 11:59 PM ET**
2. Intermediate Check-in: **DUE during the week of Module 12**
3. Notebook containing all code: **DUE Saturday of the last week of modules by 11:59 ET**
4. Final deliverable (see options below): **DUE Saturday of the last week of modules by 11:59 ET**

Logistical Notes:

- This is a 3- or optionally a 4-person group project. No collaboration outside your group is allowed.
- **Your notebook should *not* be based on prior work from outside the class, by yourself or by others.**
- Please use the **designated / pinned Ed Discussion post** to find group members, taking into account potential time zone differences.
- **There are no late days for any components of the project.**

Proposal Guidelines: DUE Monday of Module 11 by 11:59 PM ET

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:

1. All group member's names
    a. Breakdown of what each member will be responsible for
    b. We encourage you to use the pinned thread on Ed Discussion to form teams. **You need to have your team formed to submit the proposal.**
2. Data source - if you are proposing your own dataset / topic.

    a. Good data sets (i.e., those where you can actually build useful and interesting models) should have:
        i. >= 50,000 rows **after cleaning and dropping null values**
        ii. A rich set of features / properties that intuitively should be useful in predicting outcomes.
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. Why is this project interesting?
5. What challenges and obstacles might you anticipate with this project?
6. 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.

Gradescope is set up to accept group submissions, so there is **no need for all** team members to submit the proposal.

Please make sure to fill out this Google form (also available on Canvas): https://docs.google.com/forms/d/e/1FAIpQLSet1ULwYQmOqCwbZLwm1Ku6TcPyUBaYrfoDeiae8vQpYUAHfA/viewform?usp=sf_link

Intermediate Check-in: DUE during the week of Module 12

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.

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.

Project Rubric

Criteria	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 (extra credit)
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, seaborn, on top of the traditional matplotlib.	10
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?	20

*****There are 105 points in total, but the final grade will be out of 100.

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

Final Project Deliverables: DUE Saturday of the Last Week of Modules by 11:59 ET

Each group must submit the following:

Proposal:

- See the guidelines above [Page 2].

Deliverable 1:

- Every group **must** submit a complete notebook or notebooks with all of the code used throughout the project.
- If you are doing an annotated notebook, you can of course submit just 1 notebook, 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.
- 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).

Deliverable 2:

Options - please pick one (or more if you'd like) of 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. Annotated notebook
    a. Think of a notebook similar to the homework assignments.
    b. Your notebook would be broken up into separate sections and should be **very readable.**
    c. 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.
    d. You would likely submit separate notebook(s) containing the remainder of your code that does not make it into your final annotated notebook.
2. Blog Post similar to https://towardsdatascience.com , Medium – Get smarter about what matters to you.
3. Recorded video presentation roughly 5-10min
4. Live video presentation (scheduled with your assigned TA and all group members) roughly 5-10min

Gradescope is set up to accept group submissions, so there is **no need for all team members** to submit the proposal.

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. Open Food Facts - The Open Food Facts database contains nutritional information from foods all over the world. Information includes ingredients, allergens, nutrition facts and all the tidbits of information we can find on product labels.
5. 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 Discussion with any questions. Good luck!