Project

INFO 1998 | Spring 2024

Encouraged: Groups of 2-3 | Allowed: Individual Submissions

I. Key Dates

Due: May 1, 2024 (11:59pm EST)

II. Mid-Semester Check-In

There will be one meeting that your team will have with TAs during their office hours. This is so that the TA can check in on your progress, give you tips, nudge you in the right direction, and help to answer any questions that your team may have.

More details will be provided via Ed Discussion.

III. The Project

This project is the perfect opportunity for you to bring everything together that you will learn in this class. **For this project, you will conduct predictive analytics on a dataset (or datasets!) of your choice and share your code and inferences through a well-documented Jupyter Notebook.**

1. Feel free to find a dataset of your choice online - Kaggle, Data.gov, and Dataverse are some incredible resources but feel free to explore. Since you will build a predictive model, it is important that your dataset of choice has enough samples (rows) and useful features (columns). Usually, the more, the better.
2. After choosing a dataset, come up with a question that you want to answer. For example, a question relevant to the Titanic Dataset could be 'Will a person survive or not?'.
3. Clean and manipulate the data, state your hypothesis to your question, and do the following:
    a. Create (at-least) 2 meaningful visualizations that add information or context to your project. These should be different types of visualizations from one another (i.e. don't only do two scatter plots).
    b. Build (at-least) 2 machine learning models: These should be different from each other.
    c. Try to optimize these models further, and track if the accuracies increase.

d. If applicable, compare your models and infer what worked well and what didn't. In the past, students have depicted this comparison as a visualization (this would count for your 1 required visualization).

IV. Rubric and Submission

- Mid-semester check-in (5): Check-in to make sure the project is on track and check to see if any help or questions can be provided by staff. Complete it any time between lecture 6 and lecture 8
- Preprocessing and Manipulation (10): Any necessary cleaning and manipulation of the dataset
- Visualization (30): At least two visualizations. Visualizations are clearly visible, clean, well-labeled, and serve a clear purpose for your question(s).
- Models (40): At least 2 machine learning models that are chosen wisely, implemented correctly, and give meaningful results. For example, you won't get points if you run a linear regression for a classification problem. If applicable, the results of the models are compared.
- Write-Up (10): The methodologies and inferences are properly explained. Walk us through the steps and thought process you took for each step of the project. We recommend that you use 'Markdown' in contrast to the 'Code' on the Notebook. Additionally, please pre-run all the cells so we can see the output. It makes it much easier to grade on our end.
- Creativity (10): Did you go above and beyond just satisfying the requirements?

Please submit your Jupyter Notebook (that includes the link to the dataset source) and dataset in a zip file through CMS. If your dataset is too big a file, provide a link to the dataset in the notebook.

V. Other Notes

- Start early - finding a suitable dataset and cleaning it takes more time than you'd expect. Additionally, you may sometimes preprocess it only to find that the dataset does not reveal sufficient insights and will have to find another dataset.
- Seek help - The instructors would be happy to help you with dataset selection and/or any other components of the assignment. Feel free to work at office hours and ask questions as they arise.

- Be authentic - the datasets you find online would likely have multiple other projects stemming from them that you'd easily find online. We encourage you to skim through these for some inspiration but also warn against copying those or conducting identical analyses. We'll cross-check all the submissions and this has led to academic integrity violations in the past.
