Final Project

INFO 1998 | Spring 2020

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

I. Key Dates

Assigned: April 15, 2020
Proposal Due: April 22, 2020 (11:59pm ET)
Feedback Received: April 29, 2020 (11:59pm ET)
Final Project Due: May 13, 2020 (11:59pm ET)

II. Proposal and Meeting

After your team chooses a dataset, you are required to submit a concise one-page proposal that describes the dataset, links to the dataset, the problem, and the hypothesis. This is due by **April 22, 2020**.

Once you submit the proposal, your team will get some feedback by **April 29**. An instructor would give you tips, nudge you in the right direction, and make sure the project is feasible.

III. The Project

At this point, you have learned how to clean and manipulate data, visualize data, build machine learning models, and test and optimize machine learning models. The final project is the perfect opportunity for you to bring everything together. **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?'. We've explored many more during the course, but do start thinking about whether you want to form it as a regression or classification question.
3. Clean and manipulate the data, state your hypothesis to your question, and do the following:
    a. Create (at-least) 1 meaningful visualization that adds information or context to your project.
    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

- Proposal (5): Did you submit a one-page proposal clearly stating your dataset, the problem, and the hypothesis?
- Preprocessing and Manipulation (10): Any necessary cleaning and manipulation of the dataset
- Problem and Hypothesis (10): The problem statement and the hypothesis are meaningful and are clearly stated, and are uniform throughout the project. The final results are connected back to the initial hypothesis.
- Visualizations (15): At least two visualizations of different types (i.e. you can't have merely two bar charts, for instance). Visualizations are clearly visible, clean, well-labeled, and serve a clear purpose for your question(s).
- Models (2 x 20 = 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. We recommend that you use 'Markdown' in contrast to the 'Code' on the Notebook.
- 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.