Final Project Overview

Purpose:

To answer a data driven question by deploying the data science lifecycle. At the end of the process, you will have an appreciation for what it is like to develop your own questions, match these questions with data, test the ability of the data via a model to answer your question, develop a final solution and the articulation of the weakness and benefits of the approach. The goal is for you to pull all the skills together that we have worked on all semester, essentially living the life of a practicing data scientist! This is also a great way to start developing a data project portfolio on your repo or website.

Tasks:

Work with your lab groups work to answer a discrete question related to a dataset of your choosing. Some dataset resources are list below for you to potential use, but you are also welcomed to use a dataset you have used in the past or is not a part of the listed resources (not from class) Present a cleanly knitted final presentation that walks the reader through your project step by step. This means you need to reference the data science lifecycle and work through each stage deliberately. Make sure to use a Github repo for the project, with each member contributing components of the process.

General topics we will/have covered that can be a focus of the final project (feel free to combine topics or extend them)

* Data Visualization – interactive or static
* Text Mining
* kNN
* Clustering - Kmeans
* Decision Trees
* Ensemble – Random Forrest or other Ensemble Tree Method

Generate a publishable Rmarkdown document with the following sections:

* Question and background information on the data and why you are asking this question(s). References to previous research/evidence generally would be nice to include. – You must present your question to me during office hours, either next week on 26th or the following week on the 3rd
* Exploratory Data Analysis – Initial summary statistics and graphs with an emphasis on variables you believe to be important for your analysis.
* Methods – Techniques you are using to address your question and the results of those methods.
* Evaluation of your model – Select appropriate metrics and explain the output as it relates to your question.
* Fairness assessment – if necessary, should you happen to have any protected classes.
* Conclusions – What can you say about the results of the methods section as it relates to your question given the limitations to your model.
* Future work – What additional analysis is needed or what limited your analysis on this project.

Publish the final html to Rpubs or create a github page (website) that sits “on top” of your repo using Github’s internal “git pages” tool.

Submit:

* .rmd file or files
* Link to your Github repo
* Data used for the project
* Link to the Rpubs or “github page”

Criteria for evaluation:

* Come to office hours and let’s talk about your project, if able
* A documentation of each step in the process
* Creativity of the approach
* Correct application and interpretation of the methods, evaluation and fairness assessment
* Conclusion that talks through your question while utilizing your model output.

Potential Data Sources:

Google Dataset Search: <https://datasetsearch.research.google.com/>

Covid 19 - <https://github.com/XinerNing/CGDV.github.io/blob/master/dataSource/index.md>

data.world - <https://data.world/>

UCI ML Repository - <http://archive.ics.uci.edu/ml/index.php>

Data is Plural - <https://docs.google.com/spreadsheets/d/1wZhPLMCHKJvwOkP4juclhjFgqIY8fQFMemwKL2c64vk/edit#gid=0>