Stat 426 Project – Fall 2020

Part 1

Walk through the data science process:

1. Pose a prediction question that can be answered with data and a machine learning model
2. Collect data to answer your question via webscraping, APIs and/or combining several readily available dataset (i.e. kaggle, uci ML repo, etc.)
3. Clean / wrangle your data
4. Create features
5. Explore the data through EDA
6. Analyze the data with multiple machine learning approaches
7. Evaluate each model
8. Answer the original question
9. Understand and explain potential sources of bias in how your data/model answers your question of interest
10. Communicate the highlights of your work in a markdown report (this should be the Readme file of a Github repository)
11. Post all your work (including clean, well-documented, and reproducible code) in a public Github repository

Part 2

Record a 5 minute video presentation of your project. Your presentation should be self contained, meaning that someone could watch your video and know the main purpose and conclusions of your work without having seen your Github repository. You should clearly define the question you are answer / purpose of the project as well as the main highlights and/or conclusions.

Rules

- Your project should be original for this class
- Your project should be individual work
  - You can work with a group (of no more than 4 people) for data collection, but you must pose your own research question
- Your project should be original. You can be inspired by something you have seen on kaggle, Github, another class, work, etc. but should be original work.
- For full points, don't just use a readily available dataset but either:
  - combine two or more readily available sources of data, OR
  - collect your own data
- #1 and #2 will be due by November 20
- The final project will be due December 10