Milestones for Deep Dive Project

The goal of the Deep Dive Project is for you to think through how to implement Deep Learning in a real setting (similar to what you might do in a job). This means using your judgment in making decisions, and articulating why.

- Milestone 0:
  - Form groups (of at most 4, using the existing canvas *Deep Dive Project* Groups) and submit the URL of the project (one of https://drive.google.com/drive/folders/1Z5_Cd4aN-qWmOuf-GKj7DNlw1Q5bbrm0?usp=drive_link) you will be working on.
- Milestone 1::
  - Construct Google Folder (and give TA's, graders, and Sowers access and URL)
  - Download some data
  - Make a **README** file
    - Listing the team members
    - Explaining the problem (as well as you understand at this point)
    - Stating a license
  - Make a **Data Extraction** notebook.
    - Debugging dataset: small enough to test code with; reasonable code should run in 2 minutes
    - Working dataset: large enough to do the problem on (training should run no more than 40 minutes)
    - Convert these datasets to pandas
      - I suggest that you convert datetime to pandas timestamps (allows for time deltas and time manipulation)
      - Pickle (https://pandas.pydata.org/docs/reference/api/pandas.DataFrame.to_pickle.html) the data. That converts it to a binary file which can be loaded directly (must faster) into the correct datatypes
- Milestone 2::
  - Make a **Data Exploration** notebook giving
    - some visualization of the data
    - some descriptive statistics (including biases in labels)
    - explain what you are doing in text cells.
  - Discuss missing, imbalanced, or sparse data problems.
  - Make a **Baseline learning** notebook carrying, some sort of linear or logistic regression (to be used as a benchmark; feel free to use sklearn). Details left to you, but explain what you are doing in text cells in the notebook.
- Milestone 3:
  - Build a **Deep Learning** notebook (or notebooks).
    - Build a deep learning model for the dataset
    - Investigate effects of mini-batch learning

        - Investigate effects of different optimizers
            - Tune hyperparameters (training testing and validation). Explain conclusions about hyperparameters in colab markdown cells.
- Milestone 4:
    - Build a **Feature Importance** notebook discussing feature importance
    - Conclusions
- Milestone 5:
    - Documentation and cleanup of files
    - Make **Conclusions** document for the entire project (use format of your choice)
    - Conversion to repo
    - Video summary of project. NOTE: We will watch this first as a way to organize our grading.

        - Should be between 5 and 7 minutes long. Note: we won't watch the video beyond 7 minutes.
        - Should motivate problem
            - Discuss technical challenges or lessons learned in project.
        - Should discuss conclusions (feature importance?), particularly for possible stakeholders
        - One slide should give explicit sample data
        - **Each slide should be labelled with list of group members who contributed to that slide.**
        - Each page of video should have page numbers ("I have a question about slide 5")
        - Use UIUC template (https://publicaffairs.illinois.edu/resources/powerpoint-templates/)