20211025

Monday, October 25, 2021 4:32 PM

- Midterms
  - You may be getting frustrated that you are getting good results from the models you are applying
    - Data science is exploratory
    - Data science is difficult
    - You've learned a lot so far
    - Feature engineering
    - Understanding the dataset
    - Ask interesting questions
    - Get interesting answers about the data
    - 80% report
      - More important that you can explain and makes sense than one you can't explain but performs better
      - Make sure you focus on this report
      - Why you are doing what you are doing
      - You should have a good understanding of the tools and techniques
      - This is more important than the kaggle competition
      - You should mention what is relevant to the general result that it's
      - Don't need all the code you did, but should have everything that is relevant
    - 20% competition
      - Hopefully you are viewing it as fun

- Follow rubric that is on Piazza
- No word count
  - Try to limit to 2? pages with an appendix
  - Concise is better
  - Make sure you are clear
  - Takes about 2 pages to address those things in the rubric
- Report and leader board are independent from each other
- 30% seen on kaggle board now, will be graded on remaining 70%
  - Already technically done, you just don't see it right now
- You all have a lot of opportunity still to get an A in this class
- You can use kaggle kernels, google co-lab, cloud resources
  - You shouldn't need to , you can do a lot with very little
- Validation
  - This reduces the amount you have for training
  - More data isn't necessarily better
  - If you skip this, you may have a workflow problem
  - There are other methods of validation we can discuss
    - Cross-validation
    - K-fold cross validation
  - Look at the test set

        - Look at what we are asking you to predict
            - **Are there particular movies or users that is better for predicting on the testing set**
        - Worth knowing what you are being asked to make predictions on
            - This may be overfitting, but that is kind of the task you have
        - You can do non-machine learning to predict in this way
            - Recommendation systems
- If TFDF vectorizer is having a problem
    - Pull out specific words that relevant to each group
    - Top words per score
    - Top words overall
    - Top words that are different
- Another little trick:
    - If you look at the data
    - A lot of words are used across all the scores
    - That confuses the model
        - That makes it struggles between 2 and 4 and 1 and 5 since a lot of the same words are used
        - Maybe first try to predict whether the review is positive or negative, THEN use the other words to predict the score