Catherine Tabor
CS 5361 Project Update #1

GitHub repository for project:
https://github.com/CatTaborEP/ProjectMLUTEP/blob/main/MLProject.ipynb

Lexile score has historically been used to determine the appropriateness of level of a text. The problem is that this 'readability' score is little more than a numeric ratio between sentence length and syllables per word. As such, a book like *Fahrenheit 451* would be found in the same Lexile band as *Diary of a Wimpy Kid* (with *Diary* actually being considered to have a higher lexile score. I do not think anyone would consider those books to be generally appropriate for the same readers.

The organization commonlit.org has released a dataset that has been curated to include many of the most common lexile metrics and also include the metrics from which the lexile score are derived. In addition, they have added a Bradley-Terry Easiness score which is derived from teacher ratings of the works. This score is presented as a real number with a minimum value of -3.676267773 and a maximum value of 1.711389827. According to the documentation, the higher the score, the easier the read. MPAA (Motion Picture Association of America) ratings are also included and can be used to help assess the appropriateness of the material. The goal of this project is to attempt to closely match both the MPAA ratings and the Bradley-Terry easiness rating through the use of machine learning techniques and be able to better provide readability metrics.

Currently, I am in the process of creating an algorithm that will produce numbers that are in alignment with the Bradley-Terry Easiness score. In order to accomplish this, I have imported the data, dropped some of the columns that I feel are unnecessary for my assessment, and completed a linear regression on the data that currently exists in the set. The results are currently terrible as I have not pared down the data enough for what it is that I want to do with it. There is a visual representation which shows that the model is currently awful. However, I am in a good place to be able to continue and easily modify the training and test data for using different metrics and parameters to evaluate the readability of the text.

For the future of the project, I plan to pare down the data and run linear regressions on specific subsets of the data. After I have decided on the actual features that I will be using for my project, I will be testing other regression types to attempt to improve upon the MSE and fit. Other regression types that I have in mind are Lasso, Bayesian, and KNN-regression. Currently, I plan to use all of the different types of counts (except British words count) and the MPAA rating to attempt to align with the target value.