Overview

I built several machine learning models to effectively predict the average rating a book would receive based on characteristics of the books such as the authors, number of pages, e.t.c. My best performing model was the Gradient Boosting model, among other models such as Decision Trees, Linear Regression, and Adaptive Boosting. I based my model selection criterion on the best R2 Score which measures the the proportion of the variance in the dependent variable that is predictable from the independent variables.

Approach

The book prediction machine learning model was built in Python, using libraries such as pandas, matplotlib, scikit-learn, and numpy to name a few. The project passed through the following phases;

1. **Data Loading & Cleansing**: Here I loaded the data from the csv file into the Jupyter Notebook using the pandas library. It was also with pandas that I munged the data into the final form I need for analysis.
2. **Exploratory Data Analysis**: I examined the relationship between the dependent variable and several independent variables, as well as the relationship between independent variables. Also I checked the presence of data artefacts and outliers which can be damaging while building models. Matplotlib was the major library used here to create the various charts.
3. **Feature Engineering**: I derived new features from existing features for the purpose of increasing the predictive power of the final machine learning models.
4. **Model Training, Testing, and Selection**: I split the data into training and testing set with the ratio 3:1, and trained 4 different models; Linear Regression, Gradient Boosting, Adaptive Boosting, and Decision Trees using the scikit-learn package. The best parameters were selected for these models usign k-fold cross validation. Following the training, I tested the performance of the model on ‘unseen’ test data to get my model metrics, of which the R2 score is the metric with the highest weight in my selection process. Following this, I computed model metrics for each model I built, and the Gradient Boosing model was the best model with a R2 score of 24.4%.