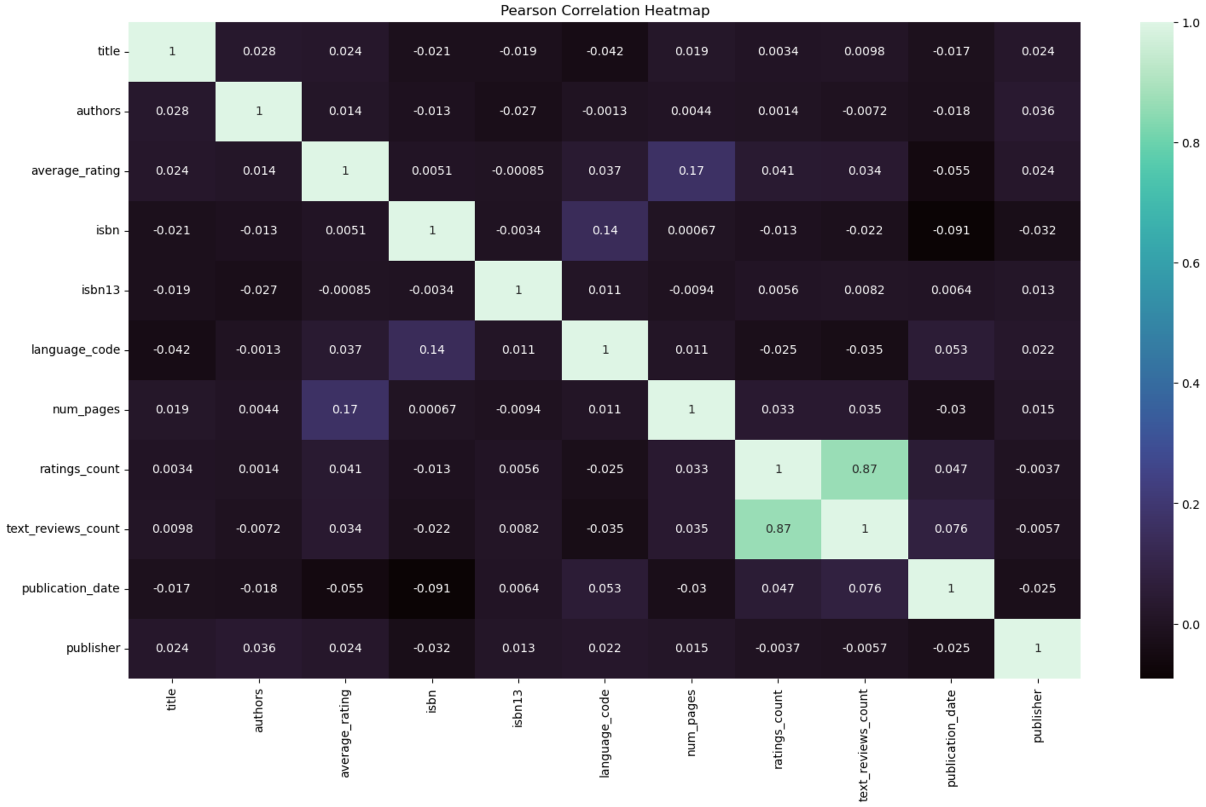
Christy Lazar
Shawna Roseaulin
Neelam Patkar
Chin Vergara

Book Rating Prediction

The goal of this study is to predict the average rating of the book according to the features given in the data set.

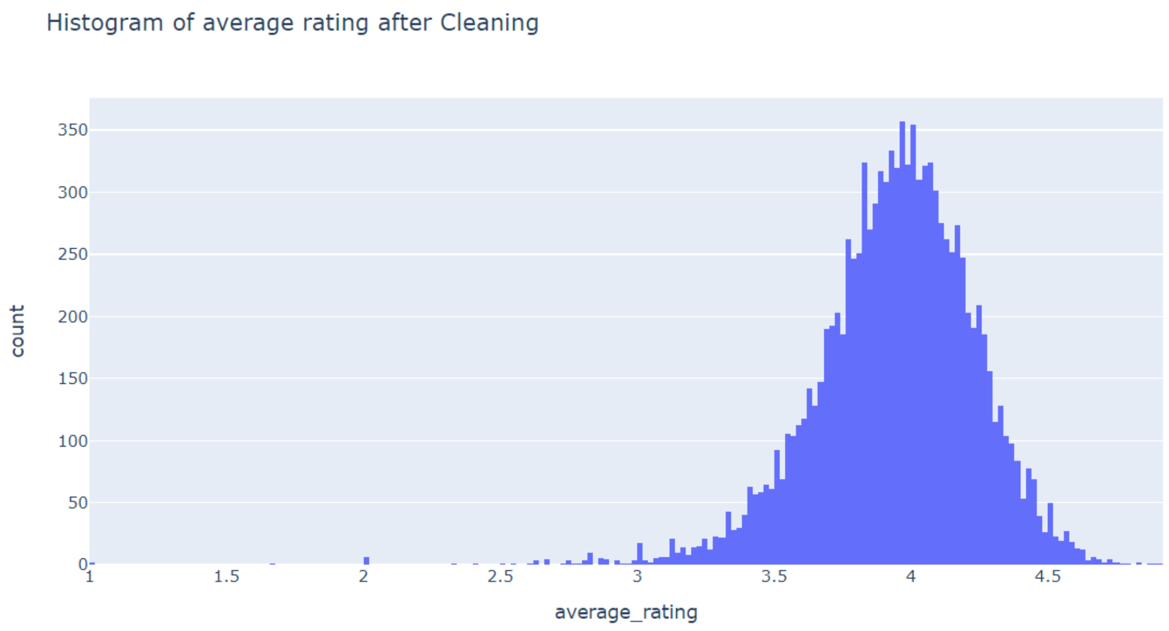
The book data set has 12 attributes that ranges from authors to publishers. We removed suspected outliers such as number of pages and average rating equals to 0, unknown publication date and certain points that are out of the norm like number of pages higher than 2000 and ratings count higher than 10^6. Some non-numerical features were label encoded.

After looking at the possible relationships between the target value and the features, we observed that the highest correlation among the features is only 0.17 (book_num).



Graph 1: Correlation Matrix

Features having low correlation on the average rating could hinder the algorithm in producing great results. Nevertheless, the features with the highest correlation were selected and some were dropped to avoid redundancy.



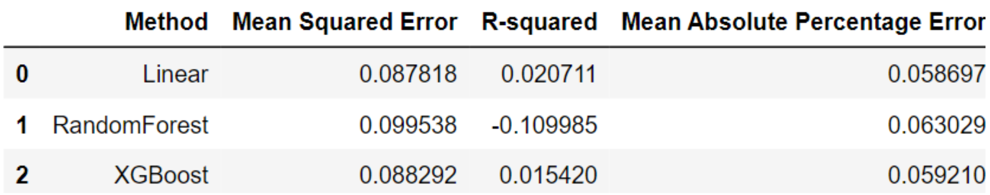
Graph 2: Average Rating Distribution after removing outliers

The graph shows us a gaussian distribution where most of the average rating is between 3 and 4. This could lead to an imbalance data problem where the algorithm favours learning the majority range.

To address this problem, a data augmentation was done using SMOGN for the algorithm to learn more on the minority class.

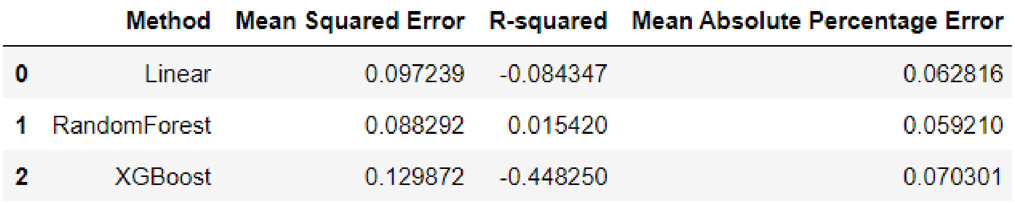
The problem was tackled with both regression and classification, binning the average rating into bad, good, and excellent.

Without SMOGN



	Method	Mean Squared Error	R-squared	Mean Absolute Percentage Error
0	Linear	0.087818	0.020711	0.058697
1	RandomForest	0.099538	-0.109985	0.063029
2	XGBoost	0.088292	0.015420	0.059210

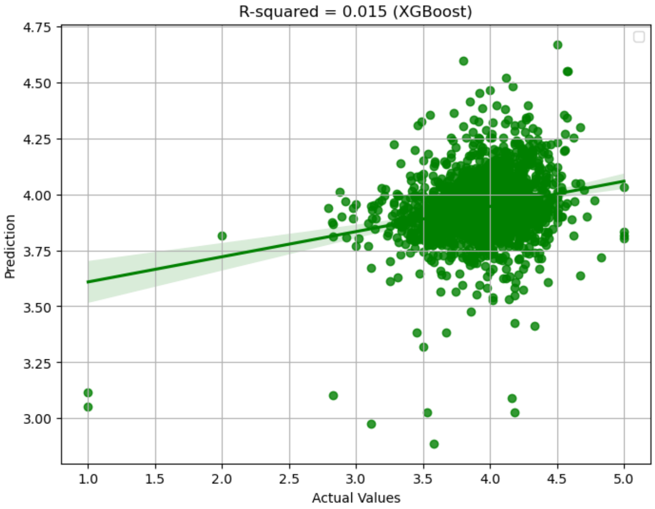
With SMOGN



	Method	Mean Squared Error	R-squared	Mean Absolute Percentage Error
0	Linear	0.097239	-0.084347	0.062816
1	RandomForest	0.088292	0.015420	0.059210
2	XGBoost	0.129872	-0.448250	0.070301

Table 1: Regression Results

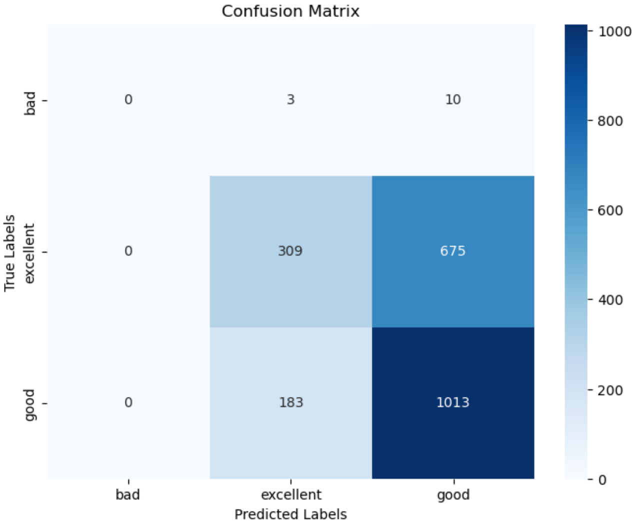
We could observe that MSE and MAPE are low as the target value (average rating) ranges only from 0 to 5. But the R^2 score is very low. This depicts that the algorithm needs more information to predict accurately.



We could observe that certain points were well predicted on the range of 3.75 and 4. Underestimation and Overestimation were also observed due to lack of correlation from features and the target value.

Graph 3: XGBoost without SMOGN

We obtained an accuracy score of 0.603 using the Random Forest Classifier. 60,3 % of the test data set are well predi cted on their respective classes.



Graph 4: Classification Confusion matrix

To conclude, we treated the book rating prediction as a regression and classification problem. Classification results yielded higher than the regression. Our features were not well correlated on the target variable where the highest is only at 0.17. To yield better results, other features such as book price, sales, etc. could be incorporated and inspected.