Julissa Mijares
Prof. Galletti
28 October 2024

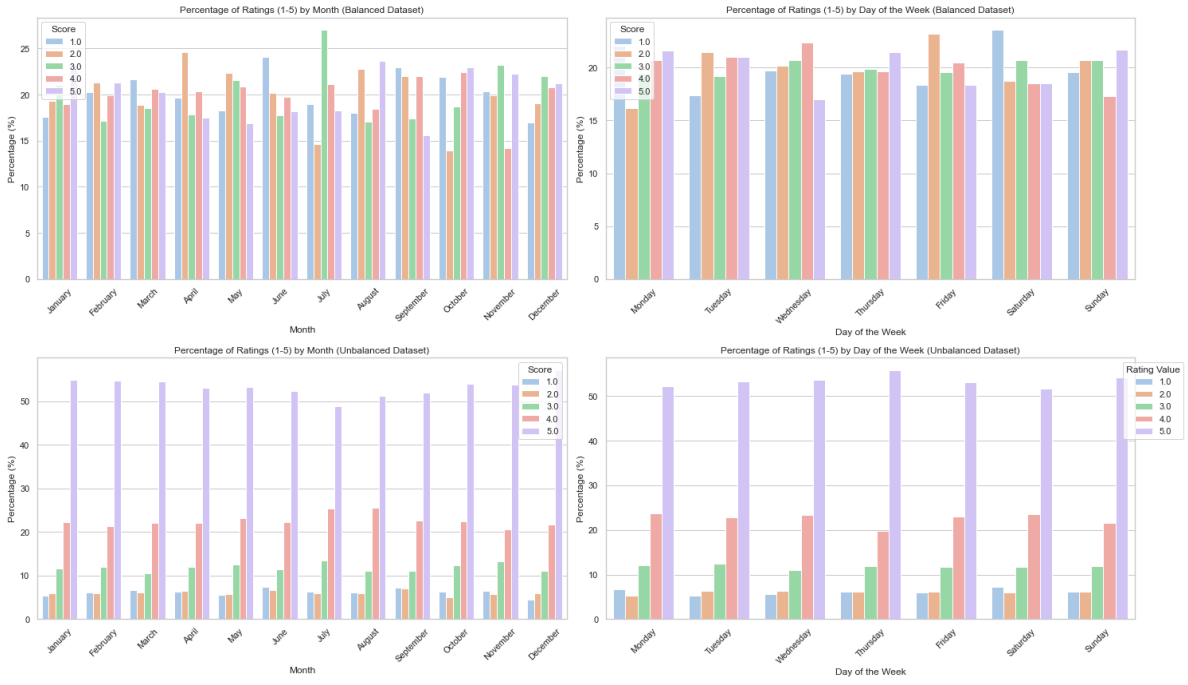
CS 506 Midterm Report – Kaggle Competition

The goal for this model was to identify significant features influencing movie rating predictions by examining and processing key columns from the dataset, including review frequency, timestamps, and textual data. The analysis also explores the impact of balanced versus unbalanced training sets and evaluates various machine learning models to determine which best captures patterns in user sentiment for rating prediction.

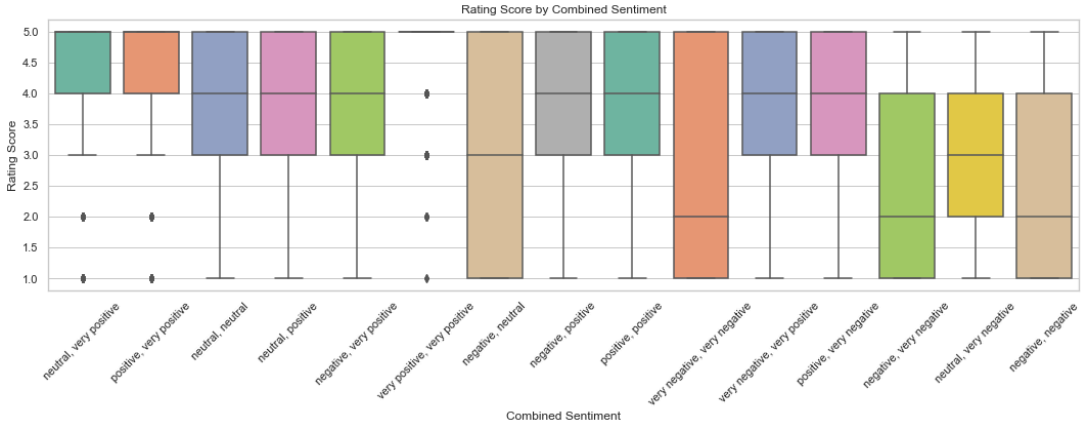
Analyze Columns

To begin assessing the dataset, I examined the frequency of two primary columns: “UserID” and “ProductID.” The objective was to determine if users who submit more frequent reviews exhibit distinct rating patterns—perhaps being more critical—or if movies with a higher number of reviews tend toward higher ratings. Unfortunately, no significant trends emerged from this initial exploration. With more in-depth analysis, these columns might reveal more significance.

Next I analyzed the “Time” column by transforming the Unix format into a more readable format, YYYY:MM:DD. This allowed me to explore whether reviews were influenced by time-based factors, such as the month or day of the week. While there was no significant correlation with rating outcomes overall, I observed a tendency in the balanced training set for certain months– like July– to yield mid-range scores (3 or 4). However, when I incorporated the month value into my model, it did not enhance the prediction accuracy.



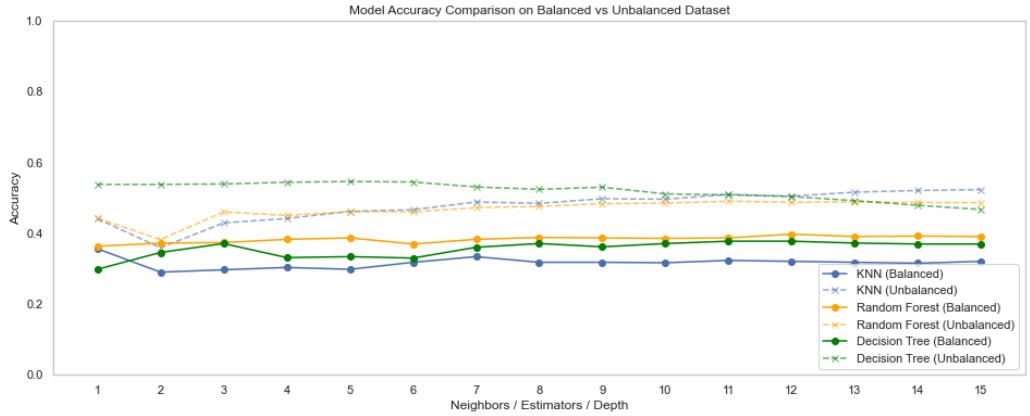
The primary columns driving the model's success were "Text" and "Summary," which I analyzed using a sentiment analyzer. These features showed the highest correlation with review scores and thus became the model's main features. After processing the text data by removing filler words through TF-IDF, I created two additional columns that calculated a percentage sentiment score and categorized these scores into "Very Negative," "Negative," "Neutral," "Positive," and "Very Positive." This categorization helped identify specific combinations with a smaller IQR, suggesting certain textual patterns that could influence the score distribution.



In particular, I tried to focus on identifying patterns in negative reviews (scores of 1 or 2) and differentiating between scores of 4 and 5, as 4 was the second most common rating. This proved challenging due to the subtle nuances in the text.

Build Model

Given the dataset's left-skewed distribution, I split the training data into balanced and unbalanced versions, each containing 10,000 samples. The balanced set included roughly 2,000 samples for each class, while the unbalanced set comprised a random sample from the full dataset. To assess model suitability, I tested KNN, Random Forest, and Decision Tree models, plotting the accuracy for each model with varying neighbor counts, depths, and other key hyperparameters. This provided an initial view of model performance across balanced and unbalanced sets. It was a simple model assessment, likely too simple to accurately predict the complex dataset since it was only changing one of the hyperparameters.



Among the models tested, Linear Regression yielded the highest accuracy, improving approximately 6% when using the “Text” sentiment score. Adding the “Summary” and “Helpfulness” scores resulted in a further 2% increase. I ultimately selected Linear Regression for its consistent performance across both training and test sets, which showed minimal variance. Attempts to combine models, such as using an ensemble of Linear Regression, Random Forest, and Naive Bayes, resulted in significant overfitting, even after adjusting hyperparameters. Running a grid search for optimal hyperparameters proved challenging, as the limited memory capacity of Google Colab resulted in frequent crashes, so I had to manually change the parameters.

Future Improvement

For improving this model or future projects of a similar nature, several changes could enhance model performance and workflow efficiency. Starting earlier would allow more time to explore additional features and perform deeper analysis on potential correlations within columns like “UserID” and “ProductID.” Access to a cloud platform with greater computational resources would enable more extensive model training and hyperparameter tuning without the memory constraints encountered in Google Colab.

One significant limitation of this project was the restriction on using deep learning models and neural networks. As a senior in data science, I initially assumed I would be able to leverage more advanced techniques for this midterm project. However, the need to rely solely on simpler models and clarify what techniques were allowed in class restricted the model’s performance and limited the potential I was able to utilize. Allowing students to use the full extent of their knowledge would likely yield better models and more accurate results, as the real world doesn’t place everyone on an equal footing in terms of skills and resources. Future projects would benefit from the flexibility to apply more advanced methods, reflecting more realistic industry standards.

On the other hand, implementing a more robust feature engineering process to include n-grams or word embeddings might further improve model accuracy by capturing nuanced sentiment shifts in the textual data. Exploring ensemble methods that incorporate different types of models, such as gradient boosting, could provide a more comprehensive approach to sentiment prediction. Additionally, automating model selection and tuning through grid or randomized search on a cloud platform would help maximize accuracy while minimizing manual intervention.

Overall, I gained valuable insights into data processing and utilized several new libraries throughout this project. While the model's accuracy did not meet my initial expectations, I feel satisfied with the results I achieved given the constraints and resources available. The experience has deepened my understanding of sentiment analysis and reinforced the importance of adapting to challenges.