Predict Rating by Product Reviews on Amazon

The objective of this project was to analyze product reviews of customer on amazon.com and determine what classifiers can be accurately predict the Rating given my Customers based on the text review.

Dataset:

https://www.kaggle.com/datasets/shitalkat/amazonearphonesreviews

Cleanup and Preprocessing:

Step 1: Dropped duplicate reviews (identical text)
Step 2: Dropped Empty reviews
Step 3: Remove stop words from reviews
Step 4: Stemming of the review words

Feature Engineering:

The following features were extracted from the text reviews

Polarity:
Measures the sentiment expressed in the text, ranging from -1 (negative) to 1 (positive).

Subjectivity:
Indicates how subjective or objective the text is, ranging from 0 (objective) to 1 (subjective).

Sentiment:
A sentiment score derived from the text, which reflects the overall sentiment (positive, negative, or neutral) of the review.

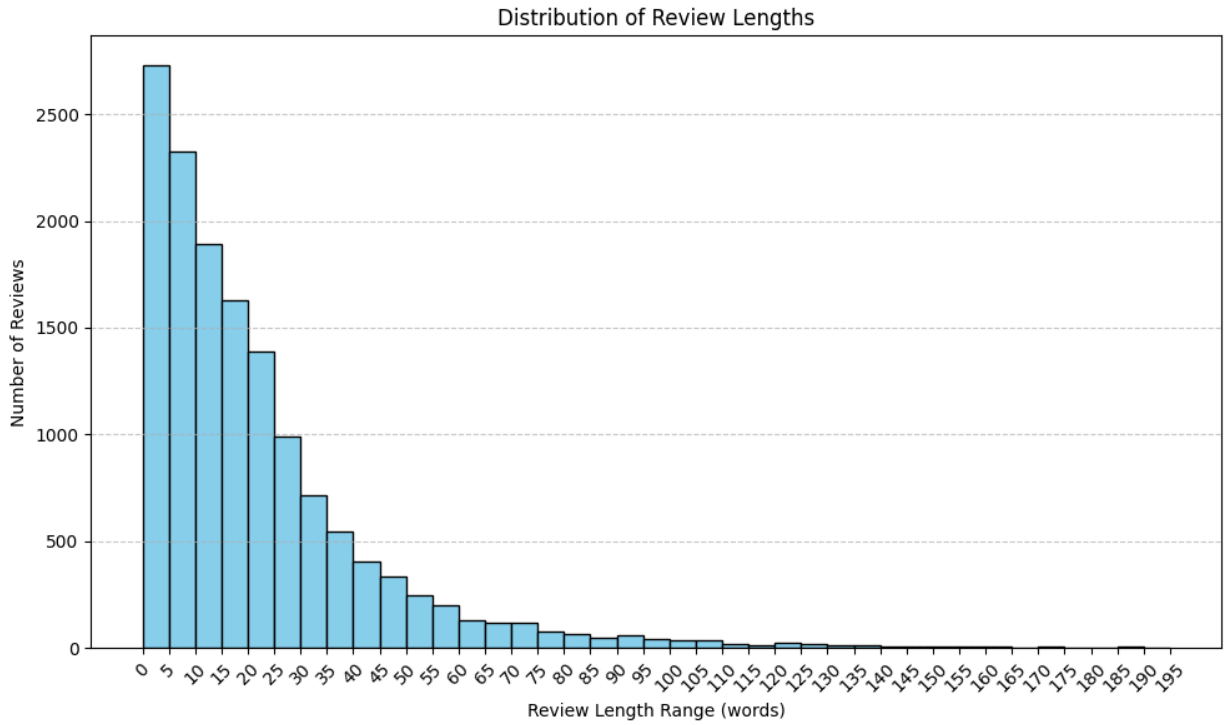
Body_len:
The length of the review body, measured in the number of words.

Reading_ease:
A score that indicates how easy or difficult the text is to read, based on the Flesch Reading Ease formula.

Reading_grade:
Represents the U.S. school grade level required to comprehend the text, calculated using the Flesch-Kincaid Grade Level formula.

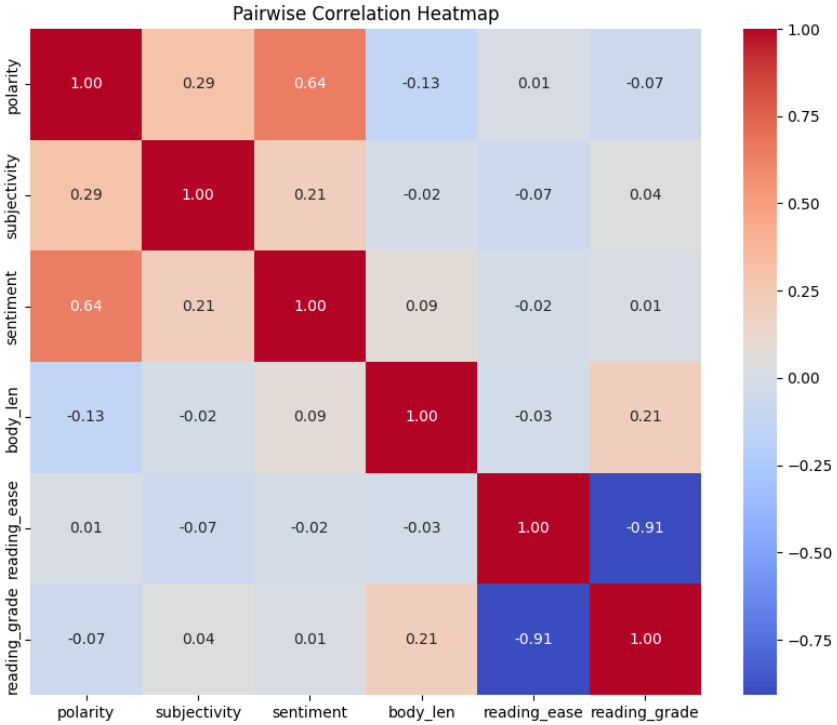
Length Analysis:

Distribution of length of review (in number of words)



Pairwise Correlation Analysis:

- Strong positive correlation between polarity and sentiment.
- Strong negative correlation between reading_ease and reading_grade
- Positive correlation between subjectivity and polarity



Scatterplot: Sentiment vs Polarity

Model Training, Testing and Evaluation
5 Star Rating System

Models:

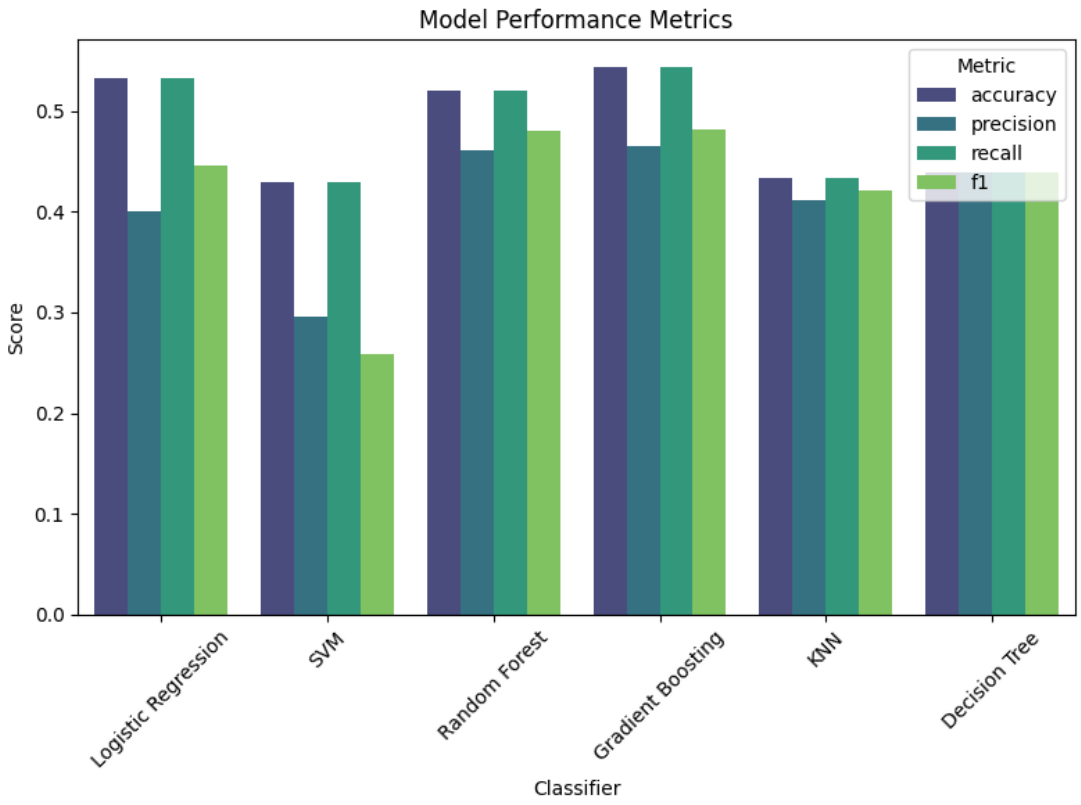
- Logistic Regression
- Gradient Boosting
- KNN
- Decision Tree
- Random Forest
- SVC

- Testing was done on 50-50 split
- Metrics measured:
  - **Accuracy:** The proportion of correct predictions (both positive and negative) made by the model out of all predictions.

  - **Precision:** The proportion of true positive predictions out of all positive predictions made by the model. It measures how many selected items are relevant.

  - **Recall:** The proportion of true positive predictions out of all actual positive instances. It measures how many relevant items are selected.

  - **F1**: The harmonic mean of precision and recall. It provides a balance between precision and recall, especially useful when the class distribution is imbalanced.



Binary Rating System

5 Star Rating system was converted to binary. Ratings above 3 considered `positive` and ratings below 3 considered `negative`. Ratings on 3 ignored to keep binary intact.

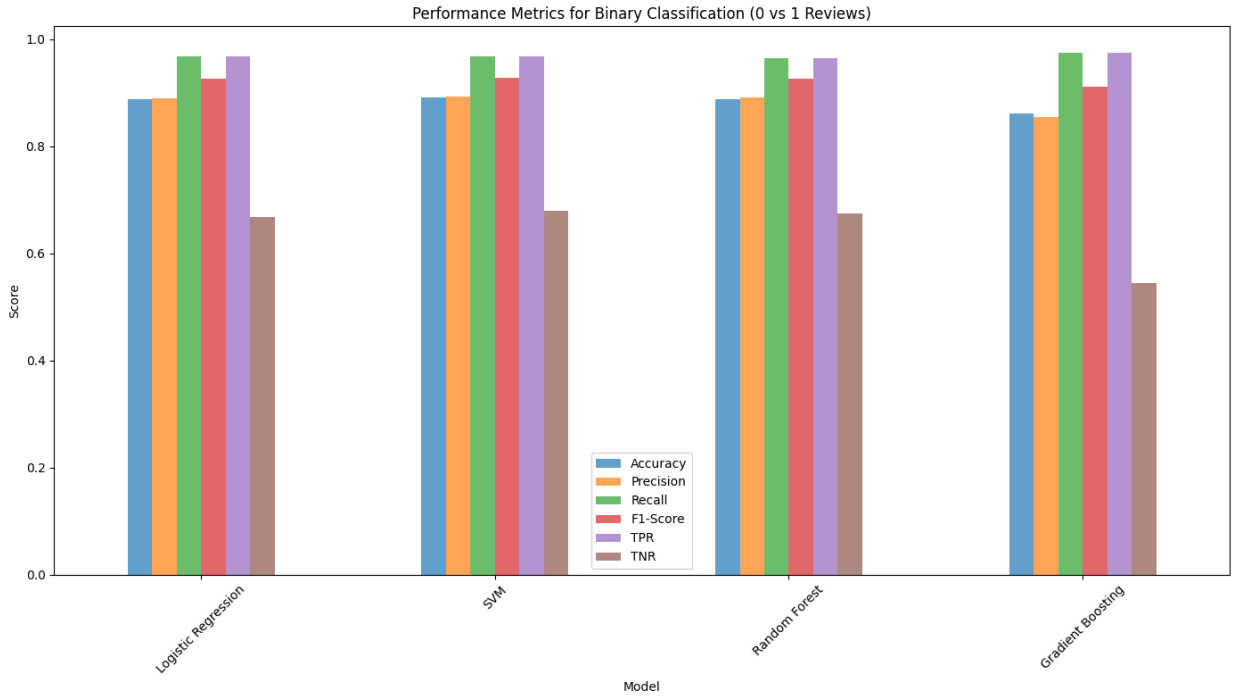
1 = `positive`
0 = `negative`

Models:

- Logistic Regression
- Gradient Boosting
- Random Forest
- SVC

- Testing was done on 50-50 split
- Metrics measured:
  - **Accuracy**
  - **Precision**
  - **Recall**
  - **F1**
  - **TPR:** Also known as sensitivity or recall, it is the proportion of actual positive instances that were correctly identified by the model.

  - **TNR:** Also known as specificity, it is the proportion of actual negative instances that were correctly identified by the model.



Conclusion

- Classification models are more performant when predicting binary reviews as compared to predicting 5 star rating systems.
- Binary predictions are around 90% accurate where as 5 star predictions are in low 50%
- Gradient Boosting and Logistic Regression are the most performant in 5 star rating predictions
- In binary predictions, all four have similar accuracies with SVM being the most accurate