Amazon Book Review Analysis

Author: James Warsing

Books: Adored and Abhorred

Books hold a significant place in our lives, evoking a wide range of emotions from love to disdain.

This analysis is crucial for enhancing customer satisfaction, as it allows Amazon to tailor recommendations that meet diverse preferences and needs.

Business Understanding

Amazon aims to enhance customer satisfaction and improve book improve book recommendations.

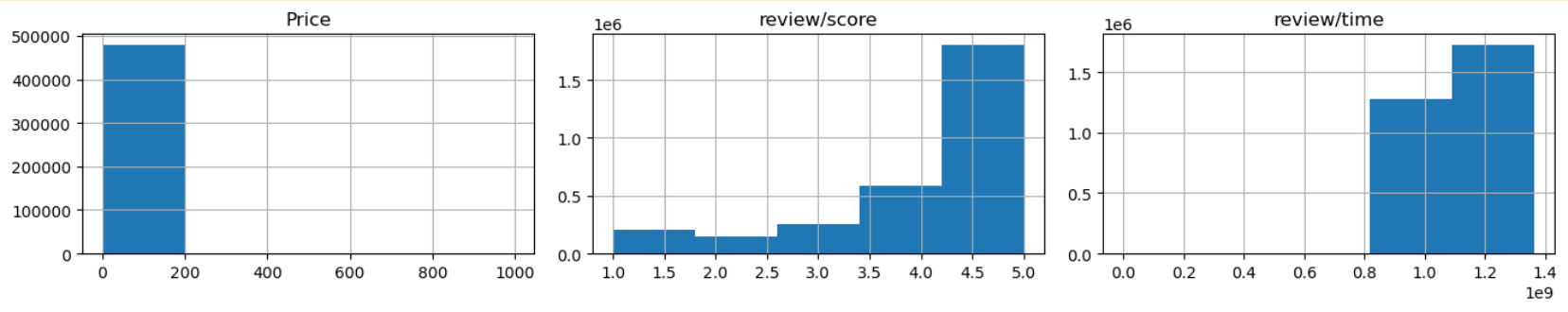
Gain insights into customer sentiment through user reviews. reviews.

Predict review scores to tailor recommendations and identify areas identify areas for product improvement.

Data Understanding

The Amazon Book Review dataset consists of two large large datasets of containing roughly 3 million rows of rows of data. This robust data allows for many different different types of models to used to make predictions. predictions.

As we can see the distribution of numerical data is heavily imbalanced. This can cause issues down the line

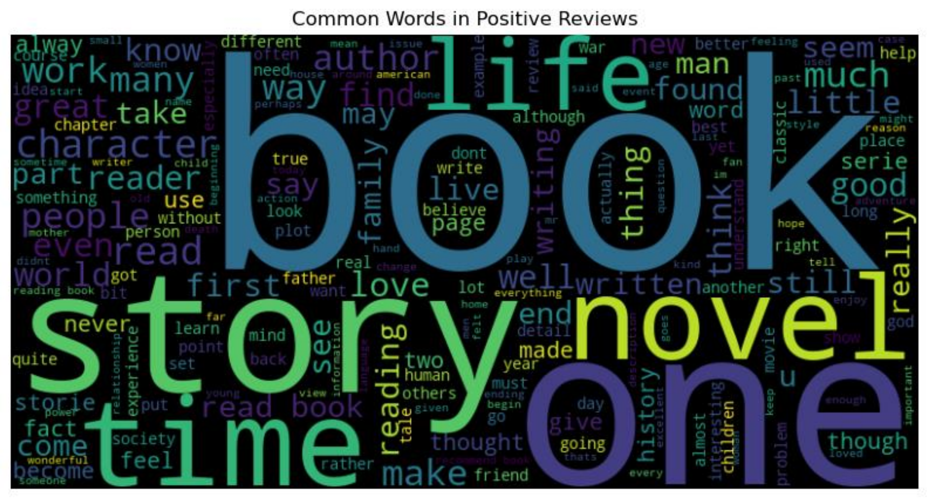


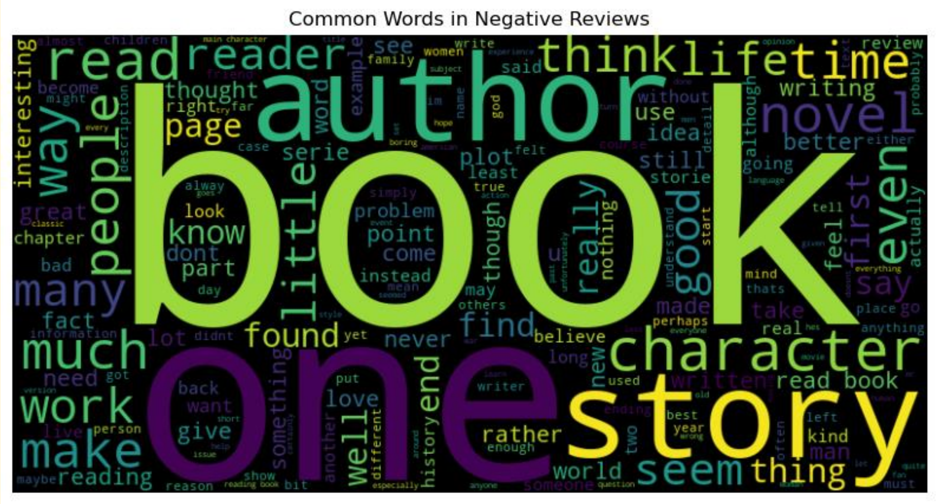
Data Preparation

- Combined the two large datasets then sampled it into a smaller set with balanced classes focusing on the scores of reviews.

- Feature Extraction: Use TF-IDF to convert text data into numerical features suitable for machine learning models.

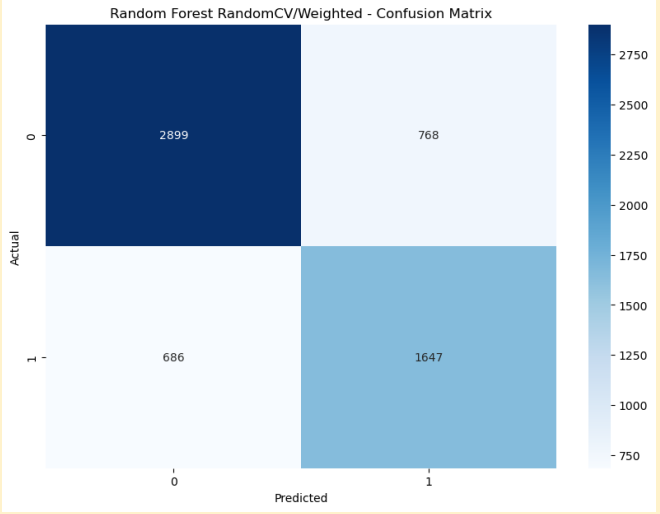
Word Cloud

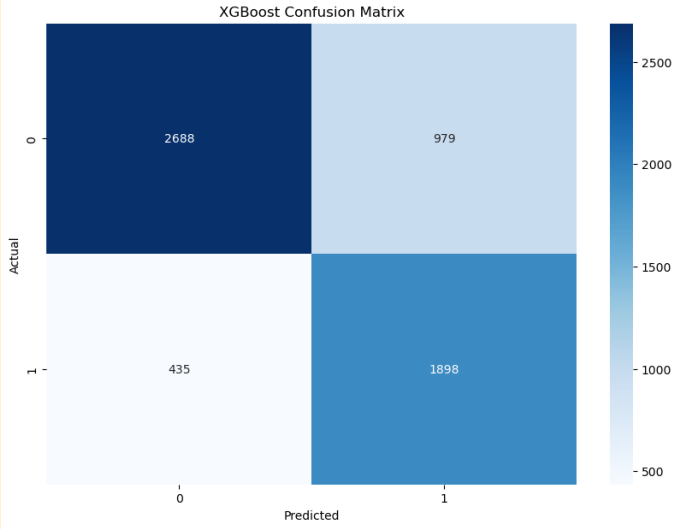

Common Words in Positive Reviews
book
story
one
life
time
novel

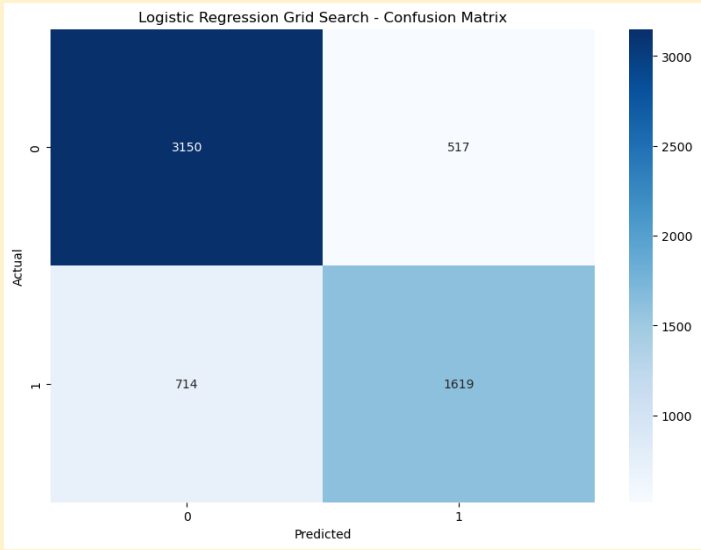

Common Words in Negative Reviews
book
one
author
story
character
life
time

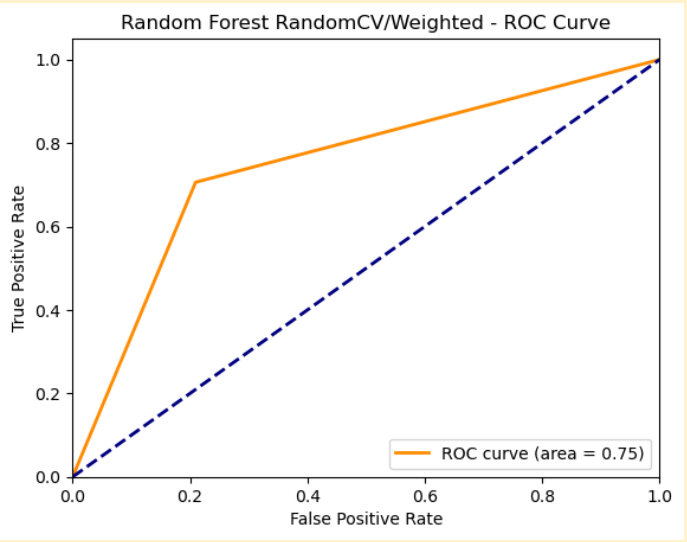
Modeling

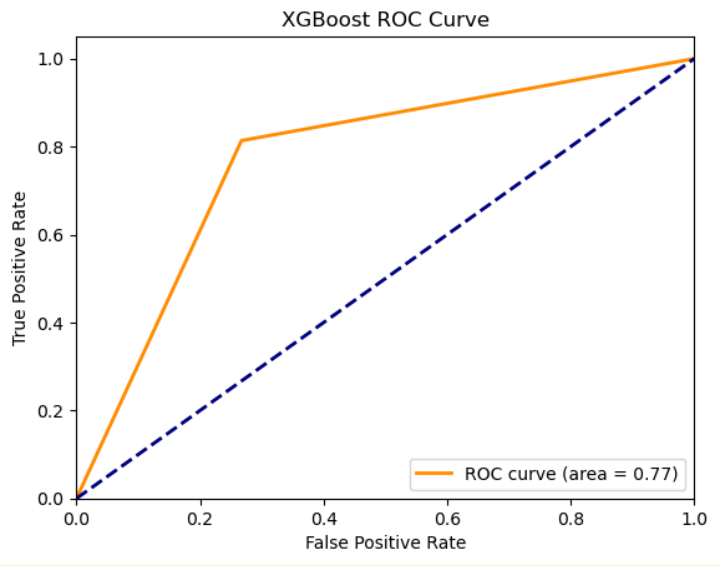
- Logistic Regression: Simple and interpretable, helps predict positive ratings for better recommendations.
- Random Forest: Robust, identifies important features influencing customer satisfaction.
- XGBoost: High performance, handles imbalanced datasets, identifies highly satisfied customers.

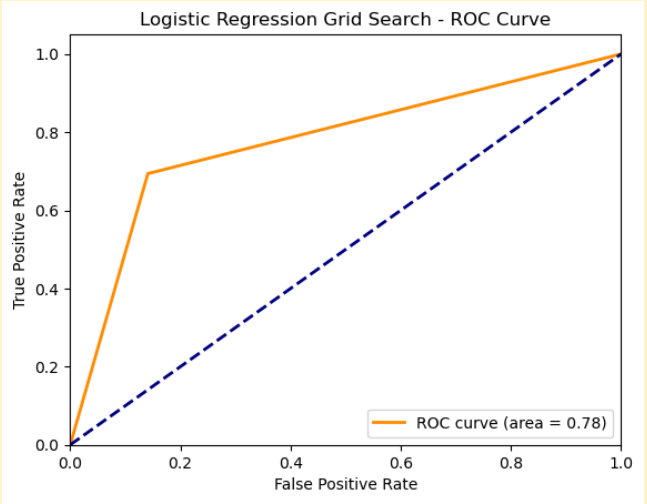

Random Forest RandomCV/Weighted - Confusion Matrix
2899
768
686
1647
Actual
Predicted
0
1
2750
2500
2250
2000
1750
1500
1250
1000
750


XGBoost Confusion Matrix
2688
979
435
1898
Actual
Predicted
0
1
2500
2000
1500
1000
500


Logistic Regression Grid Search - Confusion Matrix
3150
517
714
1619
Actual
Predicted
0
1
3000
2500
2000
1500
1000


Random Forest RandomCV/Weighted - ROC Curve
True Positive Rate
False Positive Rate
ROC curve (area = 0.75)
0.0
0.2
0.4
0.6
0.8
1.0


XGBoost ROC Curve
True Positive Rate
False Positive Rate
ROC curve (area = 0.77)
0.0
0.2
0.4
0.6
0.8
1.0


Logistic Regression Grid Search - ROC Curve
True Positive Rate
False Positive Rate
ROC curve (area = 0.78)
0.0
0.2
0.4
0.6
0.8
1.0

Evaluation

- **Confusion Matrices:** Visualize model performance, showing showing true positives, false positives, true negatives, and and false negatives, which are crucial for understanding model understanding model errors and areas for improvement. improvement.
- **ROC Curves:** Illustrate the trade-off between sensitivity and sensitivity and specificity, helping in selecting the optimal optimal model threshold.

Scores

- **Logistic Regression:** Accuracy: 0.79, Precision: 0.76, Recall: 0.69, F1-Score: 0.72
- **Random Forest:** Accuracy: 0.76, Precision: 0.68, Recall: 0.71, F1-Score: 0.69
- **XGBoost:** Accuracy: 0.76, Precision: 0.66, Recall: 0.81, F1-Score: 0.73

Accuracy:
- Indicates the overall correctness of the model.
- Important for understanding the model's general reliability but should be considered alongside other metrics to get a full picture of performance.

Precision:
- The proportion of true positive predictions out of all positive predictions made by the model.
- Crucial for making accurate recommendations, as it minimizes the number of incorrect positive predictions.

Recall:
- The proportion of actual positive cases correctly identified by the model.
- Essential for identifying promoters, as it minimizes the number of missed positive reviews.

F1-Score:
- The harmonic mean of precision and recall.
- Helps in minimizing both false positives and false negatives, ensuring a balanced performance.

Best Model

With accuracy being our main metric, the Logistic Regression model would be the best bet to use.

Accurate recommendations maintain customer trust by consistently providing relevant book suggestions.

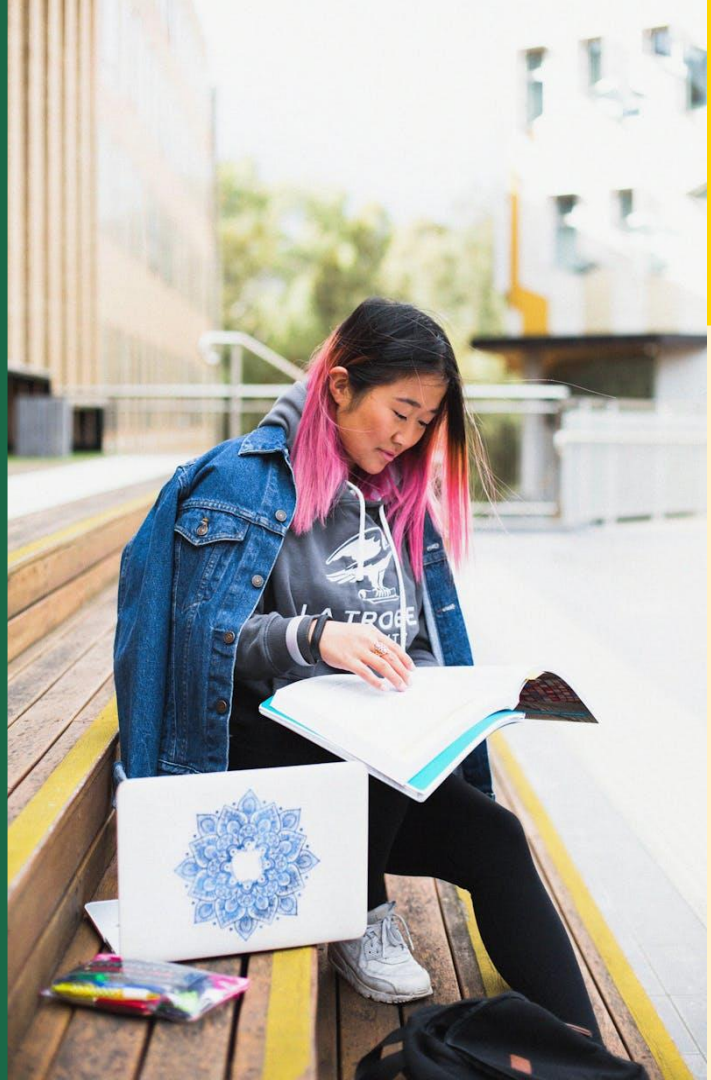
Sentiment Analysis

- **TextBlob:** High recall indicates better identification of positive sentiments, crucial for highlighting satisfied customers.
- **VADER:** Handles informal language well, suitable for social media text, and captures nuanced sentiments, helping in understanding customer emotions more accurately.

- **TextBlob:** Accuracy: 0.43, Precision: 0.39, Recall: 0.81, F1-Score: 0.52
- **VADER:** Accuracy: 0.45, Precision: 0.39, Recall: 0.71, F1-Score: 0.50

Business Recommendations

- Improve Product Recommendations: Use Logistic Regression to predict positive ratings.
- Target Satisfied Customers: Use XGBoost to identify highly satisfied customers for testimonials and promotions.
- Identify Customer Issues: Use TextBlob and VADER to analyze reviews for common issues or complaints.



Next Steps

- Invest in Advanced Text Analysis Tools: Explore models like BERT for better sentiment analysis.
- Continuous Model Improvement: Regularly update and fine-tune predictive models with new data.
- Potential use of the cloud due to the sheer size of the data

Photo by Pexels

Questions?

Thank you for your time, I will now take any questions!