Detailed Report: Multi-Modal Analysis of Amazon Product Ratings

Surabhi Chandrakant Bhor

1. Introduction and Objectives

In this project, I aimed to predict Amazon product ratings using a multi-modal dataset containing product reviews and associated metadata. The dataset included both review text and additional features like sentiment scores, helpfulness ratios, and word counts. The goal was to apply machine learning models and feature extraction techniques, such as sentiment analysis and TF-IDF vectorization, to predict product ratings effectively.

By experimenting with various machine learning models, the objective was to evaluate their performance, select the best model, and extract meaningful insights from the dataset to understand the factors influencing product ratings.

2. Data Preprocessing

2.1 Data Cleaning

The first step in the analysis was data cleaning. This involved the following steps:

- **Chunk Processing**: Given the size of the dataset (568,454 rows), the data was processed in chunks of 50,000 rows to optimize memory usage and avoid crashes during data loading. Each chunk was cleaned and preprocessed individually before being concatenated into a single dataframe.
- **Missing Values Handling**: Missing values in the ProfileName column were replaced with 'unknown', and empty strings were filled in for missing values in the Summary column.
- **Text Preprocessing**:
  - Converted all text to lowercase.
  - Removed special characters and punctuation using regular expressions.

    - Tokenized the text and applied lemmatization to each word using the NLTK WordNetLemmatizer, excluding stop words. This ensured that the words were standardized for better consistency in analysis.

- **Feature Engineering**:
    - **Sentiment Analysis**: For sentiment analysis, I initially tried using the Hugging Face model "distilbert-base-uncased-finetuned-sst-2-english" to generate sentiment scores for each review. However, the process took too long, and it became challenging to handle large datasets efficiently. As a result, I switched to using VADER for sentiment analysis, which was faster and suitable for my needs.
    - **TF-IDF Vectorization**: I applied TF-IDF vectorization to convert the review text into numerical features. After transforming the text, I selected the top 100 features based on their importance, which were then used for training the machine learning models.

2.2 Data Transformation

Additional data transformations were applied to enhance the dataset for analysis:

- **Sentiment Score**: The sentiment score for each review was computed based on the sentiment analysis model used.
- **Day of Week**: I extracted the day of the week from the review_date column, allowing for analysis of potential temporal patterns in product ratings.
- **Helpful Ratio**: A new feature was created by calculating the helpfulness ratio, which is the ratio of helpful votes to total votes for each review. This provided insight into how useful the reviews were perceived to be.

---

3. Model Training and Evaluation

3.1 Models Used

Several machine learning models were evaluated to predict product ratings:

- **GradientBoostingRegressor**
- **RandomForestRegressor**
- **XGBoost**

- **LGBM**
- **Ridge**
- **Lasso**

3.2 Evaluation Metrics

To evaluate the performance of these models, I used the following metrics:

- **Root Mean Squared Error (RMSE)**: This measures the average magnitude of the errors between predicted and actual ratings.
- **Mean Absolute Error (MAE)**: This represents the average of the absolute errors in the predictions.
- **R^2 (R-squared)**: This metric explains the proportion of variance in the ratings that is explained by the model.

3.3 Model Results

After training and evaluating the models, the **RandomForestRegressor** performed the best. Here are the results for this model:

- **RMSE**: 0.8406
- **MAE**: 0.4971
- **R^2**: 0.5850

This model was chosen as the final model due to its superior accuracy and overall performance.

---

4. Feature Importance and Insights

4.1 Feature Importance

Using the **GradientBoostingRegressor** model, I was able to determine the feature importance of the dataset. The following features were identified as the most important for predicting product ratings:

- **Sentiment Score**
- **Word Count**
- **Helpfulness Ratio**
- **TF-IDF Features**
- **Day of Week**

- **Subjectivity Score**

The sentiment score and word count were found to be the most influential features, indicating that the review content and its length played a significant role in predicting ratings.

4.2 Visualizations

I created several visualizations to better understand the model performance and the relationships within the data:

- **Model R² Score Comparison**: This bar chart compared the R² scores of all models, showing that the RandomForest model had the highest score.
- **Feature Importance**: A bar plot displayed the most important features, with sentiment and word count being the top predictors.
- **Sentiment vs Rating**: A scatter plot was created to explore how sentiment scores correlated with product ratings, showing a clear positive correlation between higher sentiment and higher ratings.
- **Rating Distribution by Day of Week**: A box plot visualized how ratings varied across different days of the week, revealing subtle temporal trends.
- **Average Rating Over Time**: A line plot showed how average ratings fluctuated by month, uncovering seasonal trends in the data.
- **Prediction Error Distribution**: A histogram was used to compare the prediction errors from different models, revealing that RandomForest had the smallest error distribution.

4.3 Additional Insights

Additional visualizations provided deeper insights into the data:

- **Review Subjectivity vs Rating**: A scatter plot illustrated the relationship between the subjectivity of reviews and their ratings, suggesting that more subjective reviews tended to have higher ratings.
- **Word Count Distribution by Rating**: A box plot demonstrated that longer reviews typically received higher ratings, indicating that more detailed feedback was associated with better product ratings.
- **Helpful Ratio vs Rating**: A scatter plot was used to visualize how the helpfulness ratio of reviews related to their ratings, showing that reviews with higher helpfulness were generally rated higher.

- **Weekend vs Weekday Ratings**: A box plot compared ratings between reviews posted on weekends and weekdays, indicating that ratings were slightly higher on weekends.

---

5. Predictive Modeling and Implications for Product Strategies

5.1 Predictive Insights

The RandomForestRegressor model was effective in predicting product ratings, with sentiment analysis, word count, and helpfulness ratio being the most influential features. This suggests that review content, especially in terms of sentiment and detail, plays a significant role in shaping customer perceptions and overall ratings.

The model's ability to predict ratings can have several practical applications for businesses. For example, by analyzing sentiment and identifying key patterns in the reviews, businesses can proactively address customer concerns, improve products, or modify marketing strategies. The analysis of word count also indicates that longer, more detailed reviews tend to be associated with higher ratings. This insight can inform how companies encourage detailed customer feedback, which may lead to higher engagement and more favorable ratings.

5.2 Recommendations for Product Strategies

- **Focus on Sentiment**: Given that sentiment analysis was a key predictor of product ratings, businesses should monitor sentiment trends in customer reviews. Using this insight, companies can adjust their products or services based on customer sentiment, addressing negative feedback and enhancing features that are well-received.
- **Encourage Detailed Reviews**: Longer reviews correlate with higher ratings. Companies should consider strategies to encourage more comprehensive customer feedback, which may include offering incentives for detailed reviews or creating prompts that guide customers to provide more specific feedback.
- **Leverage Helpfulness Ratio**: The helpfulness ratio has an important impact on ratings. Products that generate helpful reviews (i.e., reviews that are marked as useful by other customers) tend to receive higher ratings. By understanding what makes reviews helpful, companies can improve the quality of their product feedback and enhance overall customer satisfaction.

5.3 Future Directions

While the model performed well, there are opportunities to improve its accuracy and interpretability:

- **Incorporate More Features**: Other factors like product category, brand reputation, and customer demographics could provide additional insights into the factors influencing ratings.
- **Deep Learning Models**: Although RandomForestRegressor performed well, more sophisticated models like **neural networks** or **ensemble methods** could further enhance predictive performance.

---

6. Challenges Faced

During the project, I encountered several challenges:

1. **Sentiment Analysis Model Selection**: Initially, I used the Hugging Face "distilbert-base-uncased-finetuned-sst-2-english" model for sentiment analysis. However, the model took a significant amount of time to process each review, making it inefficient for large datasets. As a result, I switched to **VADER**, which provided faster results and was more suitable for my analysis.
2. **Handling Large Dataset**: The dataset contained over half a million rows, which posed challenges in terms of memory and processing time. To address this, I processed the data in chunks, which helped optimize memory usage and prevented crashes during the data preprocessing stage.
3. **Model Tuning and Optimization**: Fine-tuning the hyperparameters of the models for optimal performance was time-consuming. It required testing multiple combinations and evaluating model performance based on different metrics, which added complexity to the analysis.
4. **Computational Resource Limitations**: Due to limited computational resources, I had to balance the complexity of the models with the available processing power, leading to compromises in terms of model selection and training time.

Despite these challenges, I successfully completed the project and gained valuable insights into the factors influencing product ratings and customer satisfaction.

---

7. Conclusion and Future Work

The final model, **RandomForestRegressor**, provided reliable predictions of Amazon product ratings, outperforming other models in terms of accuracy. Key features like sentiment scores, word count, and helpfulness ratio were found to be most predictive of ratings.

Going forward, there are a few areas for improvement:

- **Hyperparameter Tuning**: Further hyperparameter tuning of the **RandomForestRegressor** could potentially improve the model's accuracy even further.
- **Incorporating More Complex Models**: Although I switched to simpler methods due to time constraints, more complex models such as BERT could be revisited once computational resources allow.

Ultimately, this analysis demonstrates that combining multiple features and using machine learning models can provide an effective approach to predicting product ratings. The insights derived from the features will also aid in improving product feedback mechanisms and understanding customer preferences more clearly.

Saved Files:

- **Processed Data**: The processed dataset with sentiment and TF-IDF features was saved as processed_data_sentiment_tfidf_features.csv.
- **Feature Importance**: The top 10 important features were saved to feature_importance.csv.
- **Best Model**: The final trained model was saved as best_model.pkl.