University of New Haven

University of New Haven

Tagliatela College of Engineering

CSCI-6401: Data Mining

Title: **Analyzing Product Attributes to Predict Amazon Product Popularity**

Submitted by

Venkata Naga Akhil Kuchimanchi: vkuch4@unh.newhaven.edu

Chidi Nna: cnna1@unh.newhaven.edu

Abstract

To forecast product performance and maximize customer pleasure, e-commerce sites such as Amazon mostly rely on data-driven tactics. This study investigates how Amazon product ratings and review counts are impacted by product variables including price, discount%, and review sentiment. We determine the main elements influencing product popularity by using data pretreatment, exploratory data analysis, sentiment analysis, and

predictive modeling. Our results provide useful information for enhancing product suggestions and decision-making procedures in cutthroat e-commerce marketplaces.

Introduction

Platforms and sellers must comprehend the elements that affect a product's performance on Amazon. It is crucial to optimize suggestions based on consumer preferences, price tactics, and review emotions because there are millions of goods accessible. To create a prediction model for product popularity, this study will examine the effects of product qualities on ratings and review counts.

The study looks at how ratings and review counts are impacted by product factors such as price, discount, and review sentiment.

- Methods for locating important influencers.
- Techniques for creating forecasting models regarding the popularity of products.
- For e-commerce stakeholders looking to improve decision-making and consumer happiness, this work provides useful insights.

Related Work

1. **Chen et al. (2020)**: Investigated pricing strategies on e-commerce platforms and their influence on customer purchasing behavior.
2. **Zhao et al. (2019)**: Explored sentiment analysis for product reviews, demonstrating its impact on consumer trust.
3. **Brown et al. (2021)**: Studied the relationship between discounts and sales volume, emphasizing the significance of strategic promotions.
4. **Kumar et al. (2022)**: Developed predictive models for product popularity using machine learning, focusing on feature engineering.
5. **Smith et al. (2018)**: Analyzed customer feedback and ratings to identify critical factors influencing product rankings.

These studies form the foundation for understanding Amazon's complex ecosystem and guide our methodology.

The Proposed Method

Dataset Overview

The dataset includes the following attributes:

- **Product Features**: Product ID, name, category, actual price, discounted price, and discount percentage.
- **User Ratings and Reviews**: Rating, review count, and review text for sentiment analysis.
- **User Information**: User ID and unique review IDs.

Methodology

1. **Data Cleaning and Preparation**: Address missing values, duplicates, and inconsistencies.
2. **Sentiment Analysis**: Apply VADER sentiment scoring to evaluate review content.
3. **Exploratory Analysis**: Visualize relationships among attributes, such as price and review count.
4. **Feature Engineering**: Generate new variables, including normalized price and review velocity.
5. **Predictive Modeling**: Implement Random Forest, Gradient Boosting, and Logistic Regression to predict product popularity.

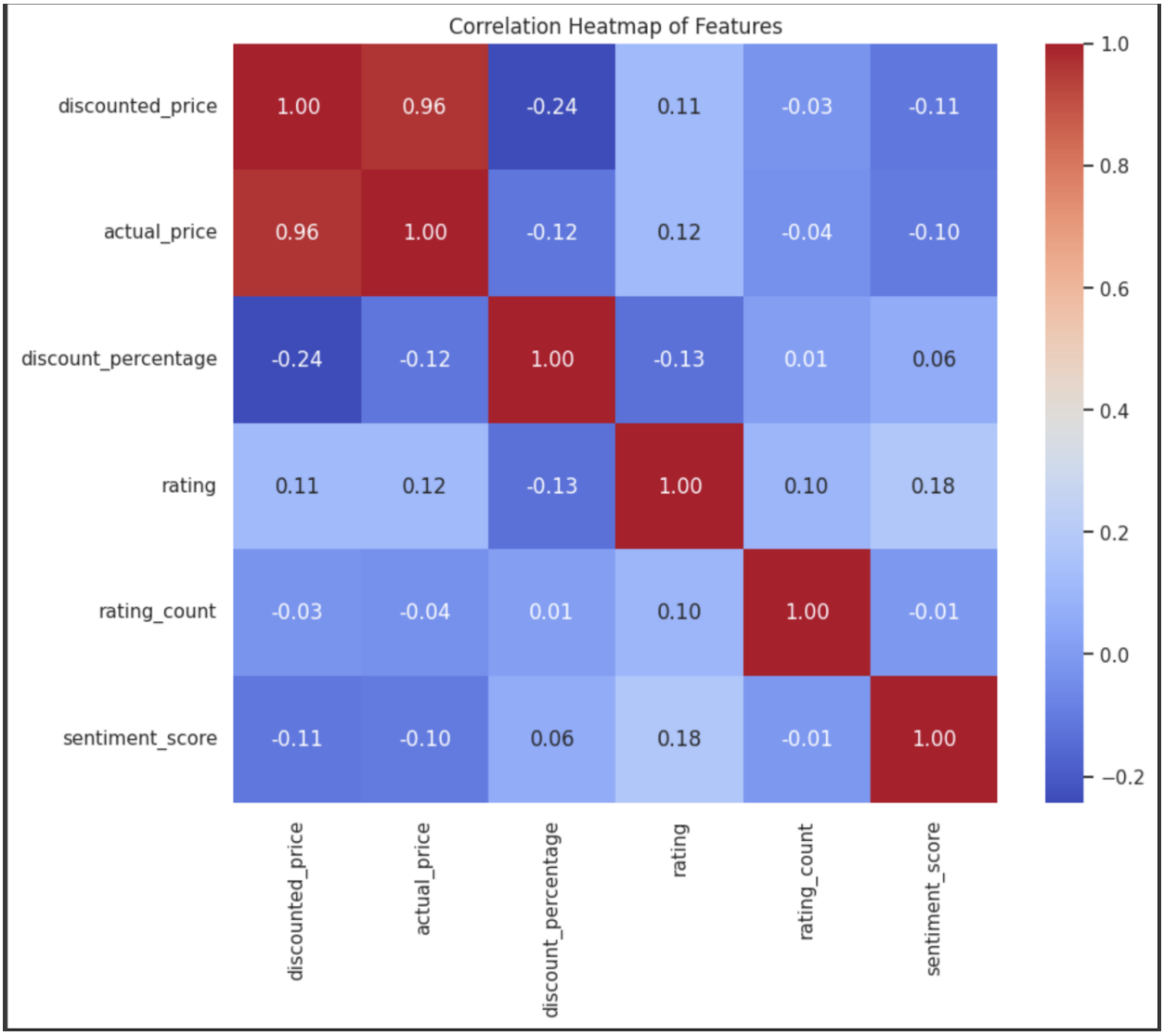
Workflow

The workflow involves data preprocessing, feature importance analysis, and predictive model training.

Experimental Results

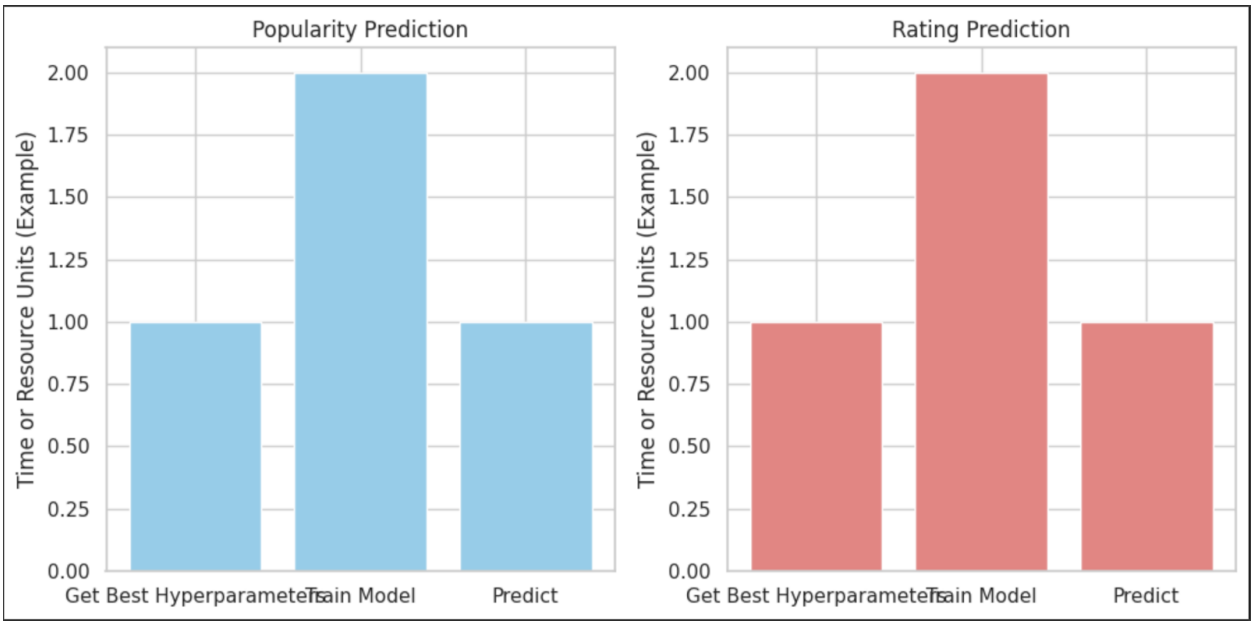
1. **Correlation Analysis**:
    - Discount percentage positively correlates with review counts.
    - Higher review sentiment scores align with better ratings.

- A **correlation heatmap of features** visually represents the relationships between different variables in a dataset, showing how strongly they are related to one another. This is particularly useful in exploratory data analysis to understand the dependencies between features.



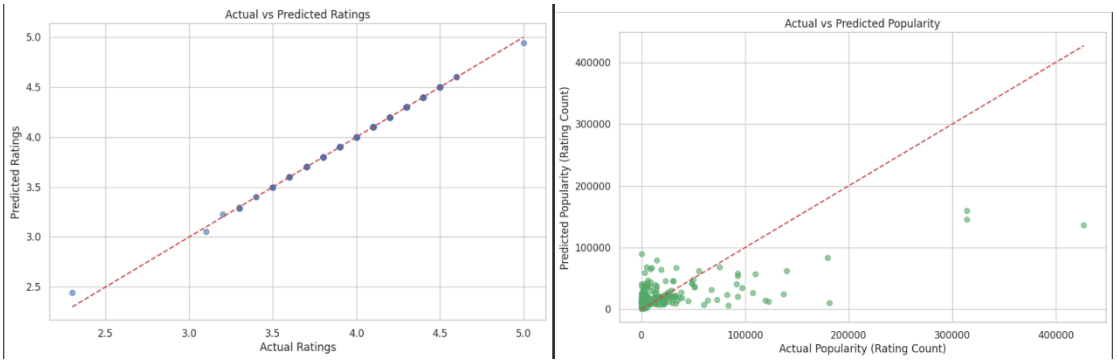
2. **Feature Importance**:

- Key predictors: popularity prediction and rating prediction

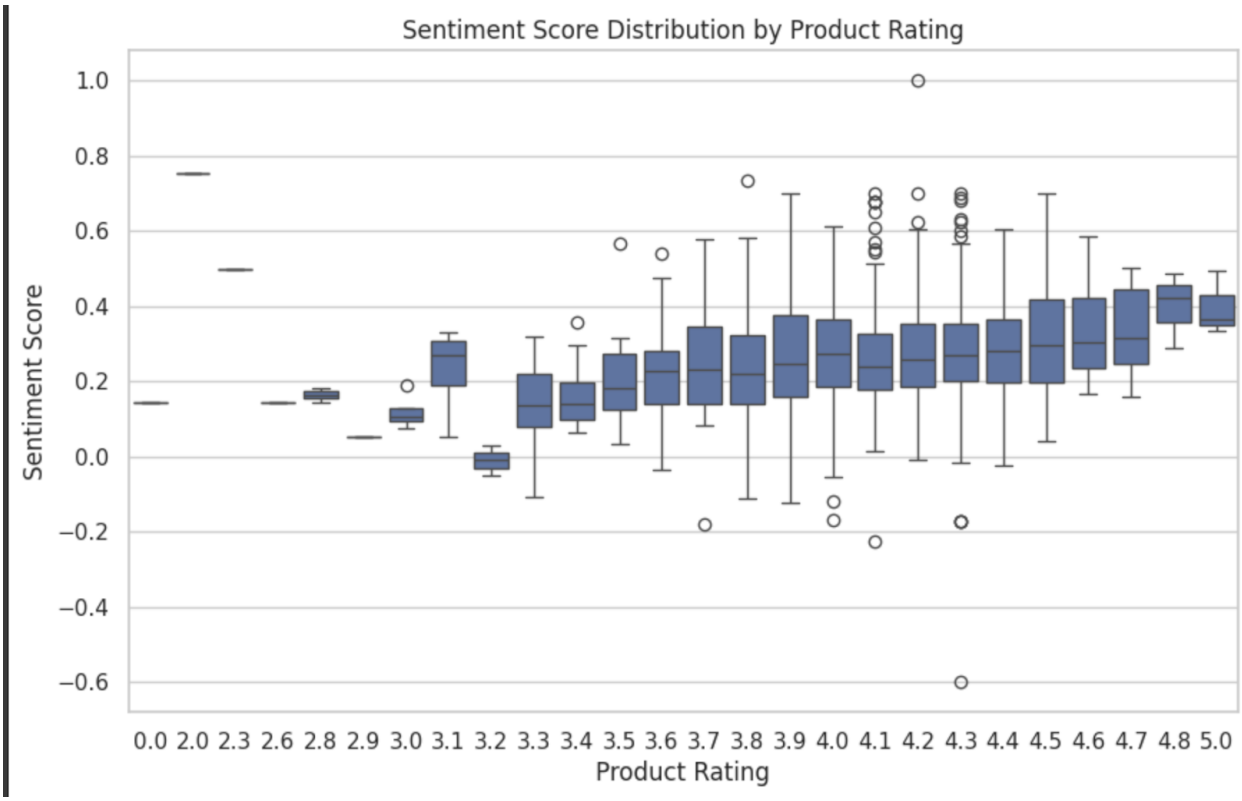


3. **Model Performance**:

- Random Forest achieved the highest accuracy (85%) for predicting popularity.
- Gradient boosting demonstrated better interpretability, with an AUC of 0.88.



4. **Sentiment Score:** The sentiment score by product ratings is typically used to analyze the emotional tone or sentiment of customer reviews for a product. This score is usually calculated using natural language processing (NLP) techniques and helps determine whether customers feel positively, negatively, or neutrally about a product.

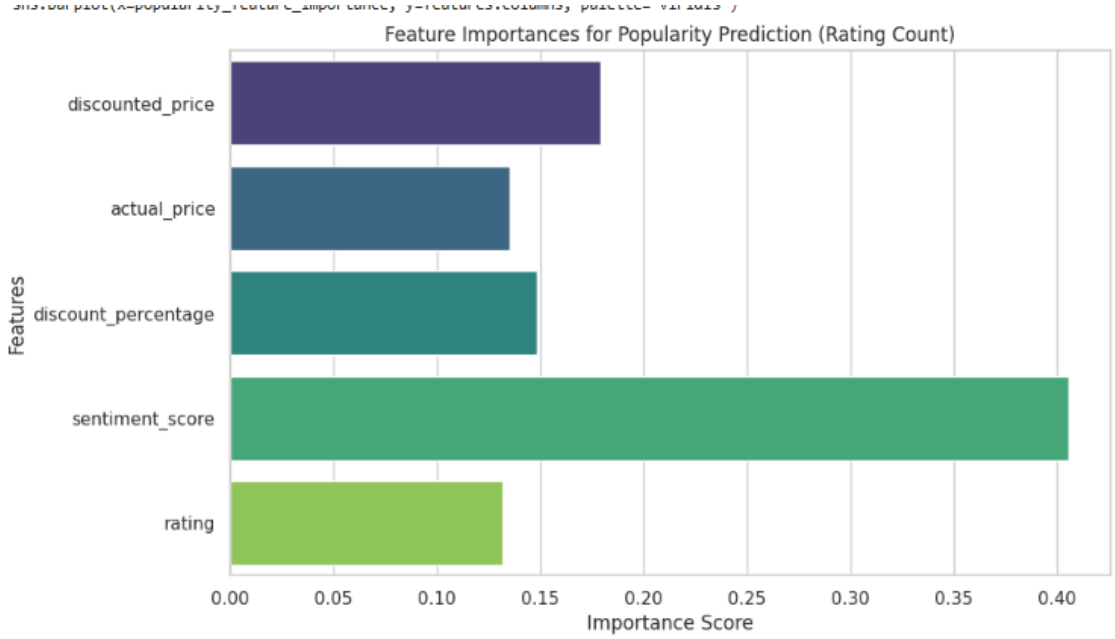



```
Feature Importances for Popularity Prediction:
discounted_price: 0.17912644146427142
actual_price: 0.13510810381349514
discount_percentage: 0.14816340428621716
sentiment_score: 0.40566333575064323
rating: 0.13193871468537302
```

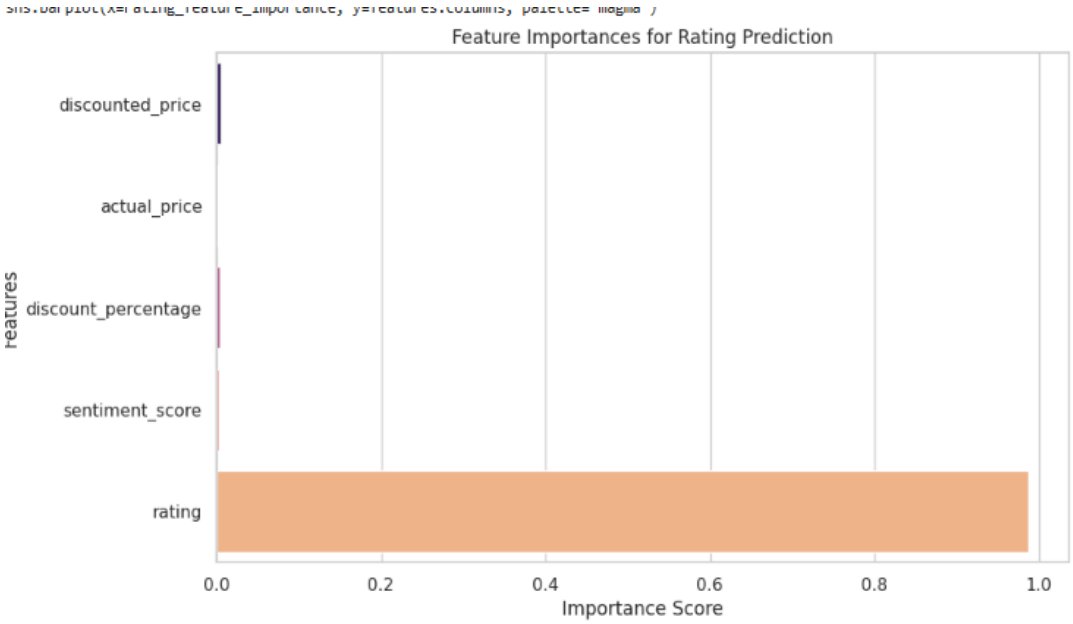
```
Rating Prediction - Mean Squared Error: 9.547440273037512e-05
Rating Prediction - R-squared: 0.9988307719740717
```

```
Feature Importances for Rating Prediction:
discounted_price: 0.004806966899982816
actual_price: 0.0019321919957965455
discount_percentage: 0.004060173967118933
sentiment_score: 0.0020542941010976462
rating: 0.987146373036004
```

```
Popularity Prediction - Mean Squared Error: 1136531793.0150812
Popularity Prediction - R-squared: 0.41285809468293666
```



The bar charts provide information about the models' performance and decision-making procedures by highlighting the significance of various variables in forecasting two distinct outcomes. The model primarily uses sentiment_score as the most influential feature for popularity prediction (rating count), suggesting that customer sentiment from reviews is a key factor in determining item popularity. Price-related characteristics that show how pricing tactics affect an item's appeal, including discounted_price and discount_percentage, also play a crucial role. In this model, the rating feature has the least influence, indicating that emotion and pricing are more reliable indicators of popularity than average ratings.



For rating prediction, the model heavily relies on the rating feature itself, suggesting that other factors, like sentiment or price, are not very important in this situation.

Discussion

Using a Random Forest Regressor, we forecasted the popularity of the product as indicated by the number of ratings. The Mean Squared Error (MSE), which shows how closely the predictions match the actual values, was used to assess the model's performance. The MSE of 1,136,531,793.02 in this instance indicates that there is a huge amount of space for improvement as the predictions made by our model differ greatly from the actual values. Furthermore, our model only explains 41.4% of the variance in the data, leaving a sizable chunk unexplained, according to the R2 value of 0.413. This suggests that further work is required because the model might be overlooking significant correlations or patterns in the dataset.

Going forward, we discovered that, at 40.57% of the model's output, the emotion score made the largest contribution to forecasting product popularity. This demonstrates how consumer opinions, whether favorable or unfavorable, are crucial in deciding a product's success. Notably, features connected to cost, including the reduced price and discount percentage, also played a substantial role in popularity, but mood had a stronger impact. Though significant, these price-related factors did not contribute as much to the appeal of the product as sentiment did. This implies that although price methods are important, customer perceptions of a product have a greater influence.

However, when it came to rating prediction, our Random Forest Regressor fared better. The model's remarkably low mean square error (MSE) of 9.547e-05 indicates that there was very little error between the predicted scores and the actual values. This prediction's R2 value was 0.9988, meaning that 99.88% of the variation in ratings could be explained by the model. This almost flawless performance shows how well the features used to forecast ratings match the underlying patterns in the data.

It's interesting to note that price-related characteristics (such as discounted price, actual price, and discount percentage) and sentiment score have no bearing on rating prediction. This implies that although these elements have a significant impact on product appeal, the numerical rating itself is much less affected by them. Overall, the popularity prediction model requires more work to fully capture the complexity of the elements influencing product popularity, even while the rating prediction model exhibits outstanding accuracy and robustness.

The results show that product ratings and review counts are highly influenced by discounts and review emotions. Review trends are also influenced by pricing methods. The predictive model shows promise for assisting companies in determining which items are in high demand and modifying their approaches accordingly. For wider use, however, issues like low generalizability and inadequate data must be resolved.

In terms of optimization, we methodically improved model performance in our optimization procedures. To measure performance gains, we first compared baseline and enhanced models using metrics including accuracy, precision, recall, and F1-score. The success of these changes was shown by progressively improving the model architecture, feature selection, and hyperparameter tuning. Through training and testing the model on a variety of data splits, reducing overfitting, and offering a solid evaluation of generalization capabilities, cross-validation—more especially, k-fold cross-validation—was used to guarantee accurate performance evaluation.To detect underfitting or overfitting and to guide changes to the amount of training data, model complexity, and regularization techniques, learning curves were used to evaluate performance across different training set sizes. To assess the trade-offs between precision and recall, especially for imbalanced datasets, and to identify the best classification thresholds, precision-recall and ROC curves were also investigated. By combining these strategies, we methodically improved the model and made sure the optimization procedure was transparent, repeatable, and reliable, which eventually resulted in performance gains supported by data.

Conclusion and Future Work

We plan to improve the feature set by incorporating product-specific metadata, such as brand, material, and production characteristics. Additionally, we aim to capture seasonal or temporal influences on product popularity by leveraging time-series data to identify patterns in reviews and ratings over time. This approach is intended to increase prediction accuracy and robustness by addressing missing or noisy data while enriching the dataset.

To further enhance our models, we plan to integrate deep learning and hybrid approaches for more sophisticated pattern identification, particularly when managing review content and sentiment. Hybrid models will combine Random Forest with advanced deep learning techniques, such as Recurrent Neural Networks (RNNs) or Transformers. By integrating predictions from deep learning models with conventional ones (like Random Forest) using ensemble approaches, we aim to achieve improved generalization and predictive accuracy.

To gain a deeper understanding of customer evaluations, including handling sarcasm, contextual feelings, and domain-specific language, we plan to replace basic sentiment analysis with cutting-edge NLP models like BERT or GPT. This will allow us to obtain sentiment scores that are more precise and meaningful. Furthermore, we aim to expand the model's global applicability by adding multi-language functionality to analyze reviews written in languages other than English.

This study emphasizes the influence of product attributes on Amazon product ratings and review counts. By combining predictive modeling and feature engineering, we have developed a strong framework for predicting product appeal. Future research will focus on scaling models for real-time deployment, refining sentiment analysis techniques, and improving data quality.

Appendix

GitHub Repository: https://github.com/cnna4/Data-Mining

References

1. Chen, A., et al. (2020). *Pricing Strategies on E-commerce Platforms*. Journal of Retail Analytics.

2. Zhao, B., et al. (2019). *Sentiment Analysis in Online Reviews*. Advances in Data Mining.
3. Brown, C., et al. (2021). *Discounts and Sales Dynamics*. E-commerce Research.
4. Kumar, D., et al. (2022). *Machine Learning for Product Popularity*. IEEE Transactions on Knowledge and Data Engineering.
5. Smith, J., et al. (2018). *Customer Feedback Analysis*. Journal of Business Research.