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

It's crucial for platforms and sellers to 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. In order 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

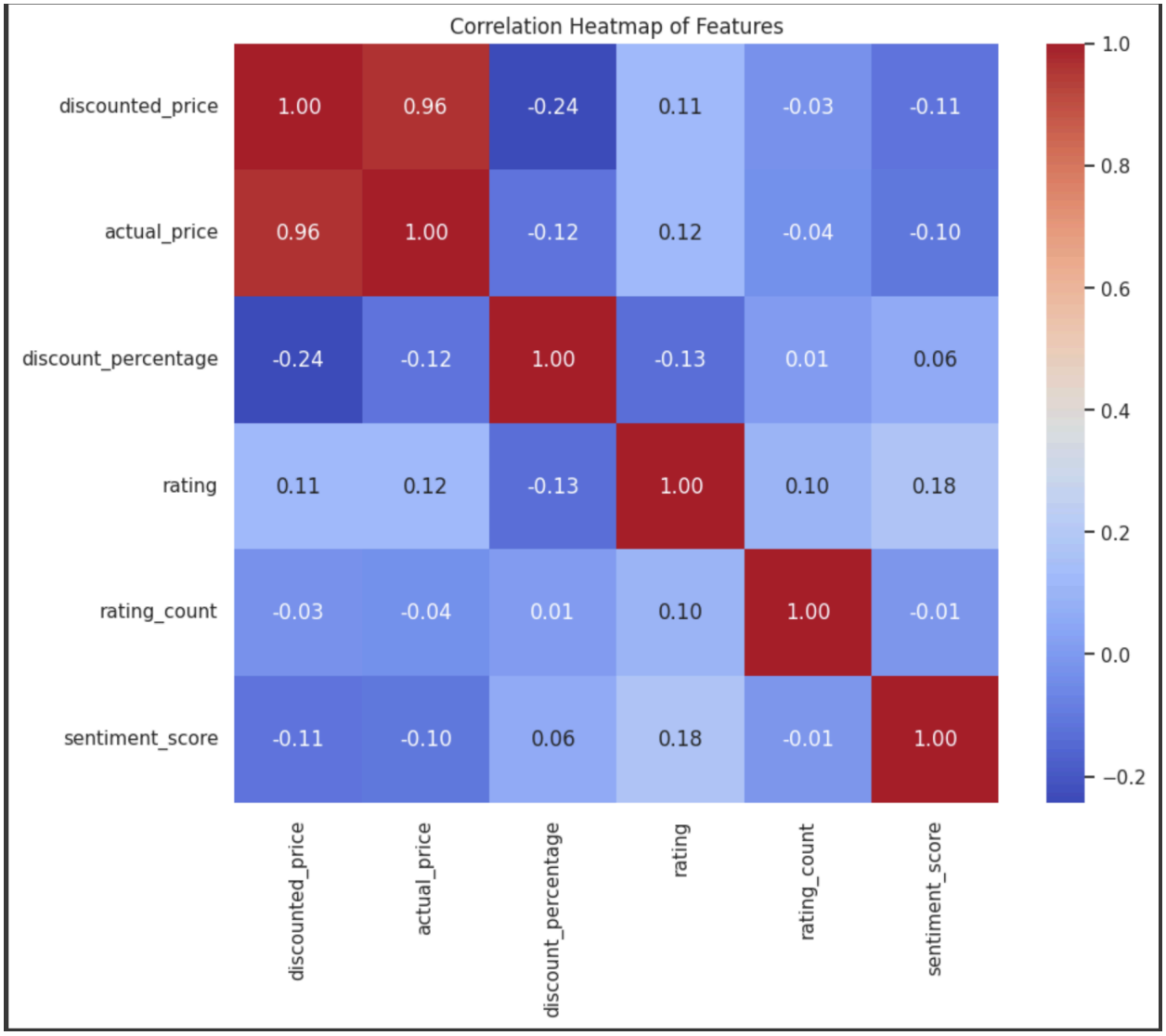
1. **Data Cleaning and Preparation**: Address missing values, duplicates, and inconsistencies.
2. **Sentiment Analysis**: Apply VADER sentiment scoring to evaluate review content.
3. **Exploratory Analysis**: Visualize relationships among attributes, such as price and review count.
4. **Feature Engineering**: 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

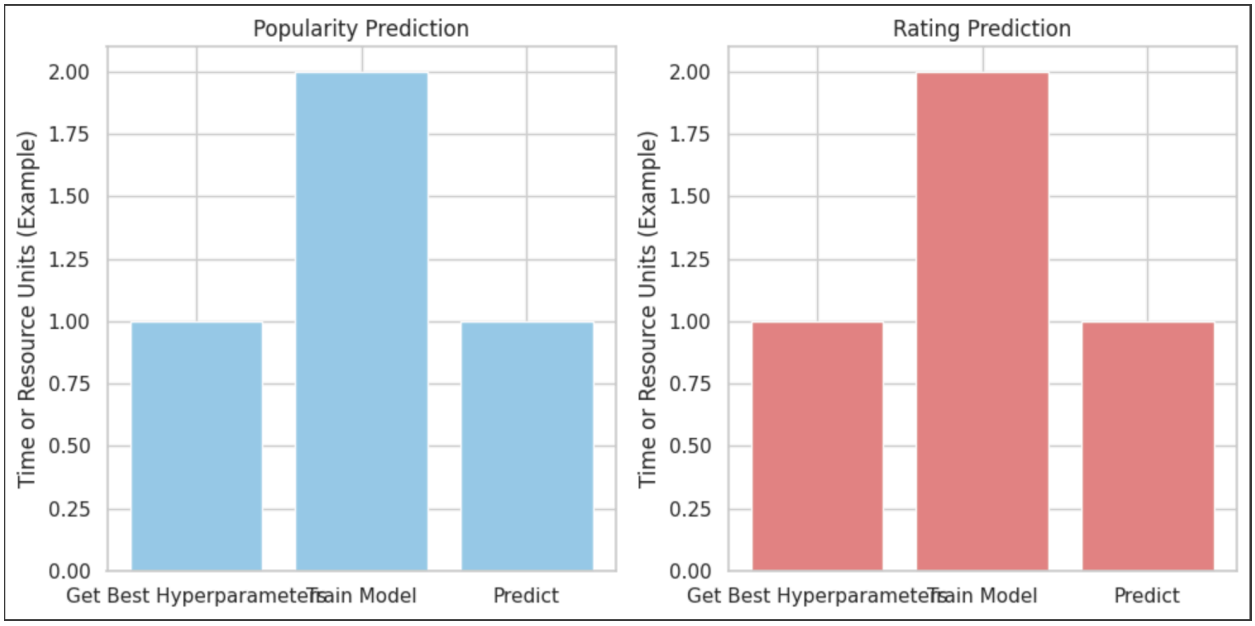
The workflow involves data preprocessing, feature importance analysis, and predictive model training.

Experimental Results

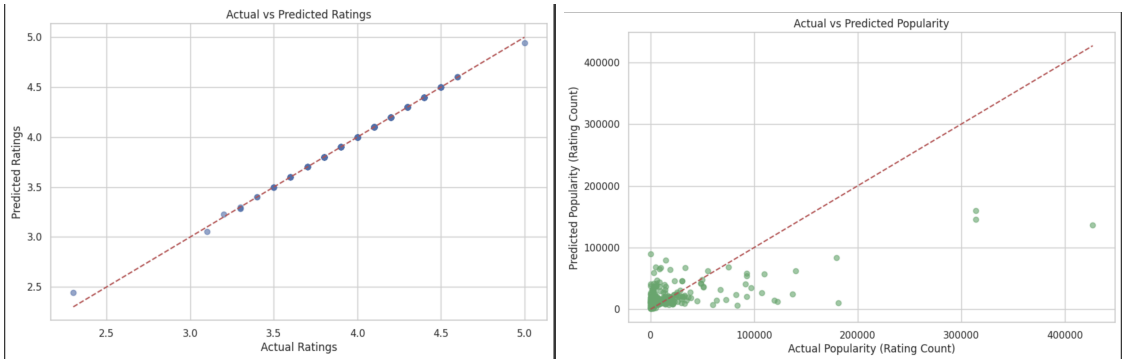
1. **Correlation Analysis**:
    - Discount percentage positively correlates with review counts.
    - Higher review sentiment scores align with better ratings.
    - A **correlation heatmap of features** visually represents the relationships between different variables in a dataset, showing how strongly they are related to one another. This is particularly useful in exploratory data analysis to understand the dependencies between features.



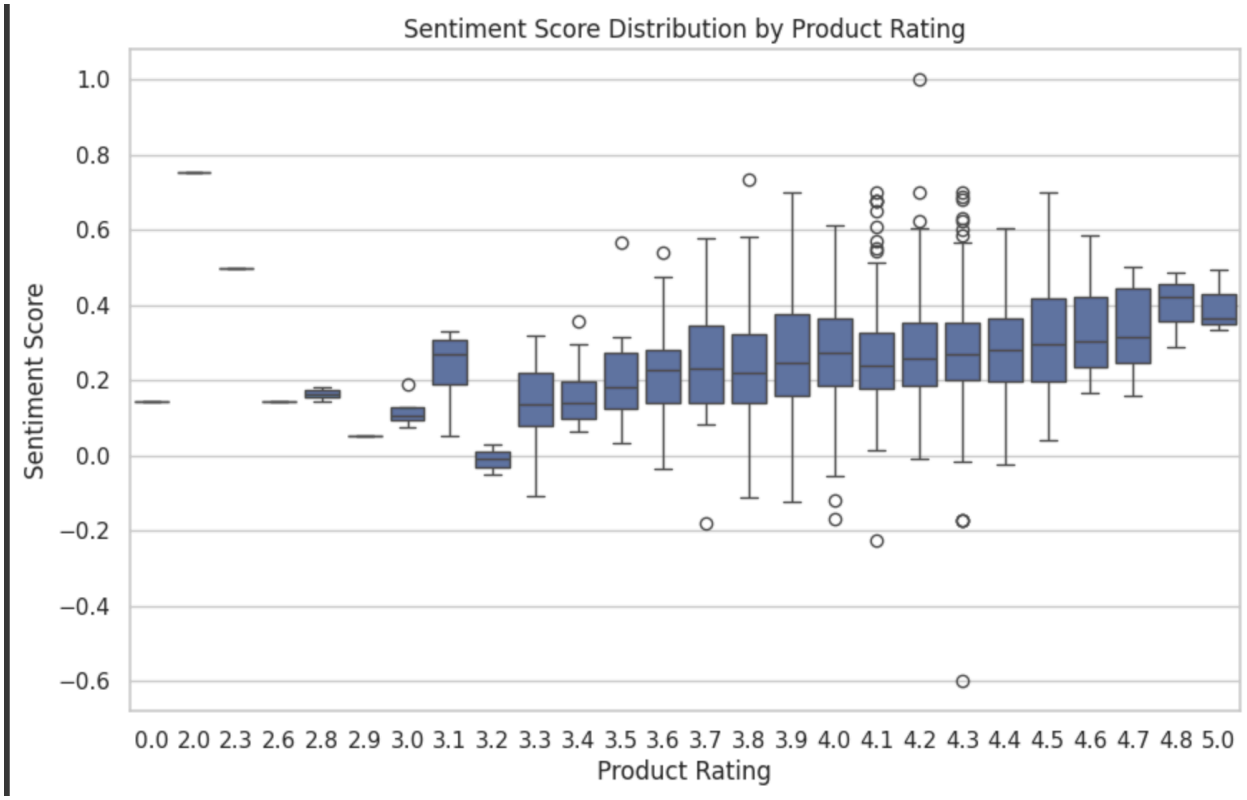
2. **Feature Importance**:
    - Key predictors: popularity prediction and rating prediction



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.



Discussion

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.

Conclusion and Future Work

The influence of product qualities on Amazon product ratings and review counts is highlighted by this study. A strong framework for predicting product appeal was developed by fusing predictive modeling and feature engineering. Future research will concentrate on scaling models for real-time deployment, refining sentiment analysis methods, and improving data quality.

Appendix

GitHub Repository: https://github.com/kvnakhil/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.