Amazon Fashion Products - Predictive Model Using Logistic Regression

Jakob Sarmiento

1. Project Goal

The goal of this project is to predict whether Amazon fashion products will receive a 4-star rating or higher, based on the product's price, category, and brand. Through this project, I hope to pinpoint which features of a product impact customer ratings the most.

2. Business Objective

Understanding which fashion products earn the highest ratings allows Amazon sellers to focus on inventory with greater customer appeal and sales potential. By predicting high ratings in advance, sellers can make data-driven decisions on product selection, branding, and pricing.

3. Dataset Used – Cleaned Version of Amazon_Fashion_Products Dataset

- **Size:** - 11 000 rows, 8 columns
- **Key features:** price, category, brand
- **Target variable:** highly_rated (1 if product rating ≥ 4.0 stars)

4. Key Findings

Classification Metrics

- According to the permutation importance analysis, category and brand are strong positive drivers of a product's rating, meaning they have the greatest influence on whether a product is highly rated or not.

Visualizations Analysis

- The confusion matrix shows that the model correctly predicted the majority of products.
- The ROC curve confirms strong predictive ability, with an AUC of 0.836, showing that the model can distinguish highly rated products from others.
- Product category has the highest influence on whether an item is predicted to be highly rated, followed by brand. Price has no effect, confirming my previous EDA Analysis that price and rating aren't related.

5. Business Recommendations

- **Prioritize high-rating categories.**
  Product category is the strongest predictor of customer ratings. Sellers should focus on categories with consistently higher reviews to maximize positive ratings
- **Invest in branding.**
  Branding is a key driver of positive reviews. Companies should maintain a strong brand presence and reputation to support sustained high ratings.
- **Ratings to sales growth.**
  Because higher ratings generally lead to increased sales, strengthening brand reputation and concentrating on top-performing categories will help revenue.

6. Tools Used

- **Languages/Libraries**: Python, pandas, numpy, scikit-learn, matplotlib
- **Environment**: Google Colab

7. Conclusion

This project predicts whether an Amazon fashion product will earn a 4-star rating or higher using price, category, and brand. The logistic-regression model achieved 83% accuracy and 0.84 ROC-AUC, confirming strong predictive performance. Analysis shows that price has little influence on ratings, while product category and brand strength are more impactful. Amazon sellers should focus on product category and branding, as these lead to higher ratings and sales. Pricing, although still important, doesn't necessarily have a huge role in product ratings.