Drivers of E-Commerce Success: A Multi-Methodological Analysis of Amazon Product Performance

Final Integrated Project Report Date: January 22, 2026 **Prepared By:** [Vishesh Chaturvedi]

1. Executive Summary

This project presents a comprehensive data science lifecycle analysis of the Amazon e-commerce marketplace. Utilizing a web-scraped dataset of over 30,000 product listings, the study aims to demystify the factors contributing to "Best Seller" status. The project followed a structured pipeline: data cleaning (handling unstructured text and missing values), exploratory analysis (EDA) to visualize market stratification, statistical modeling to quantify elasticity, and machine learning to build a predictive engine.

Key Findings:

- **Price Sensitivity:** The market is highly price-sensitive. Regression analysis confirmed a negative price elasticity of **-0.75**, meaning a 1% increase in price correlates with a 0.75% drop in review volume.
- **The "Golden Range":** Visualization revealed a power-law distribution where "Best Seller" success is heavily concentrated in the **$20–$50** price range.
- **Social Proof vs. Quality:** While high ratings (>4.3 stars) are a prerequisite for entry, they are not a differentiator. **Sales Volume (Review Count)** was identified by the Random Forest model as the #1 predictor of Best Seller status, far outweighing star rating.
- **The Sponsorship Paradox:** Paid sponsorship showed a negative correlation with organic success metrics, suggesting it is primarily a tool for market entry rather than a characteristic of market dominance.

2. Introduction & Problem Statement

In the hyper-competitive Amazon marketplace, new sellers often struggle to identify the correct levers to pull for visibility and sales. Does cutting price drive rank? Is paying for "Sponsored" placement effective? Is a 5.0-star rating essential?

This project addresses these questions by analyzing real-world listing data to determine the statistical drivers of two key success metrics:

1. **Popularity:** Measured by Customer Review Volume.
2. **Market Dominance:** Measured by the "Best Seller" badge status.

3. Data Processing Pipeline

The raw dataset consisted of ~30,000 uncleaned product records containing unstructured text, currency symbols, and missing values.

Key Preprocessing Steps:

- **Cleaning:** Removed currency symbols (`$`, `,`) and parsed "K" notation (e.g., "10K+ bought") into integers.
- **Imputation:** Missing price values were imputed using the **Median** to mitigate the impact of high-priced outliers (right-skewed distribution). Missing "Listed Prices" were imputed using the current selling price, assuming zero discount.
- **Feature Engineering:** Boolean fields (`is_best_seller`, `is_sponsored`) were encoded as binary (0/1).
- **Transformation:** A **Log1p transformation** was applied to Price and Review Counts to normalize their power-law distributions for regression analysis.

4. Exploratory Data Analysis (EDA) & Visualization

Visual analysis was conducted to uncover the "Anatomy of a Best Seller."

4.1 Price Architecture

[Insert Figure 1: Price Distribution / Violin Plot]

- **Insight:** The analysis reveals a "Golden Price Range." While regular products are scattered across all price points, Best Sellers are heavily concentrated between **$20 and $50**. This suggests that mass-market adoption faces significant friction above the $50 mark.

4.2 The Quality-Volume Matrix

[Insert Figure 2: Rating vs. Sales Bubble Chart]

- **Insight:** There is a "4.3 Star Barrier." Best Sellers (orange bubbles) rarely drop below a 4.3 rating. However, the chart shows that *many* regular products also have 4.8 or 5.0 ratings but lack volume.

- **Conclusion:** High quality is necessary but not sufficient. Success requires the "Social Proof Engine"—massive review volume—to convert quality into sales rank.

---

5. Statistical Inference

To quantify these relationships, an **Ordinary Least Squares (OLS) Regression** was performed ($R2=0.239$).

Equation: $\log(\text{Reviews})=\beta 0+\beta 1\log(\text{Price})+\beta 2(\text{Rating})+\beta 3(\text{Sponsored})$

Results:

- **Price ($\beta=-0.745, p<0.001$):** Confirms strong negative elasticity. Cheaper products generate significantly more engagement.
- **Rating ($\beta=+0.478, p<0.001$):** Positive correlation, but the effect size is smaller than Price.
- **Sponsorship ($\beta=-0.894, p<0.001$):** Surprisingly negative. This indicates that established organic winners (Best Sellers) rarely need to pay for sponsorship, whereas newer, low-volume products rely on it heavily.

---

6. Predictive Machine Learning

A **Random Forest Classifier** was developed to predict "Best Seller" status. The model was optimized using **GridSearchCV** (5-fold cross-validation) to maximize the **F1-Score**, addressing the class imbalance.

6.1 Model Performance

- **F1-Score:** 0.96
- **Precision:** 0.98
- **Recall:** 0.95

6.2 Feature Importance

The model ranked the predictors by their influence on the decision trees:

1. **Review Count (Sales Volume):** The dominant predictor.
2. **Price:** The secondary driver.
3. **Star Rating:** Third most important.
4. **Sponsorship:** Least important predictor of organic success.

---

7. Conclusion & Recommendations

The multi-methodological analysis converges on a clear strategy for e-commerce success. The "Best Seller" status is not driven by a single metric but by a specific configuration:

1. **Aggressive Pricing Strategy:** Sellers should target the $20-$50 range to reduce friction and build initial sales velocity.
2. **Volume over Perfection:** A product with a 4.5-star rating and 5,000 reviews is statistically more likely to be a Best Seller than a 5.0-star product with 50 reviews. Strategies should prioritize review generation (Social Proof).
3. **Sponsorship Utilization:** Paid ads are effective for *launching* products (gaining initial visibility) but are not a characteristic of *mature* Best Sellers, who rely on organic rank.

---

8. Limitations & Future Work

- **Snapshot Bias:** The dataset represents a single point in time. Future work should implement time-series collection to track price elasticity over time.
- **NLP Analysis:** The current model uses numerical ratings. Integrating **Sentiment Analysis** of the review text would provide deeper insight into *why* customers rate products highly (e.g., durability vs. aesthetics).