Analyzing E-Commerce Product Trends Using Shopify and Social Commerce Data

Business Problem

The rapid expansion of e-commerce and social commerce platforms has intensified competition among digital retailers and entrepreneurs. While platforms such as Shopify enable rapid product launches and low barriers to entry, they also increase the risk of poor product selection, inefficient pricing strategies, and unsustainable inventory decisions. Many sellers rely on anecdotal evidence, short-lived viral trends, or influencer signals rather than structured analytical approaches when selecting products to sell.

The business problem addressed in this study is the challenge of identifying high-performing e-commerce products using data-driven methods rather than intuition-based decision making. Specifically, this project seeks to determine how trend scores, pricing strategies, and product categories relate to estimated revenue performance. By applying predictive modeling and exploratory analysis to Shopify trend data, this study aims to provide actionable insights that support informed product selection and pricing decisions for digital retailers and product managers.

Background and History

Social commerce has transformed the way consumers discover, evaluate, and purchase products. Platforms such as Instagram, TikTok, and Shopify-powered storefronts have significantly shortened product life cycles, while simultaneously increasing competition across nearly every product category. As a result, sellers must continuously adapt to shifting consumer preferences and rapidly evolving market trends.

Trend scores, estimated revenue metrics, and unit sales estimates are frequently used as indicators of consumer demand. However, these metrics are often interpreted in isolation without deeper validation of their predictive value. Prior research in digital retail analytics suggests that combining trend indicators with pricing and category-level analysis can improve understanding of sustainable product performance. This study builds upon that foundation by applying applied data science techniques to real-world Shopify and social commerce data, emphasizing business relevance and interpretability rather than purely theoretical modeling.

Data Explanation and Preparation

Dataset Overview

Several data preparation steps were conducted to ensure analytical reliability. Duplicate product records were removed to prevent overrepresentation of individual products. Category and subcategory labels were standardized to ensure consistency across observations. Price range values, originally represented as text ranges, were converted into numerical midpoint values to enable quantitative analysis. Missing values were addressed through exclusion where appropriate, and additional feature engineering was performed to support predictive modeling.

These preprocessing steps reduced noise within the dataset and ensured that downstream analysis accurately reflected underlying relationships among trend, pricing, and revenue variables.

Methods

This study follows an applied data science workflow consisting of exploratory data analysis, visualization, and predictive modeling.

Exploratory Data Analysis

Exploratory data analysis was used to examine revenue distributions across product categories, relationships between trend score and estimated revenue, and price range patterns among high-performing products. Visualizations and summary statistics were used to identify early patterns and guide model selection.

Predictive Modeling

Two predictive modeling approaches were employed. Linear Regression was used as a baseline model due to its interpretability and ability to quantify linear relationships between variables. Random Forest Regression was applied to capture nonlinear relationships and interaction effects among trend score, pricing, and product category. Model performance was evaluated using standard metrics including Root Mean Squared Error (RMSE), Mean Absolute Error (MAE), and the coefficient of determination (R^2), with an emphasis on interpretability and business relevance rather than model optimization.

Analysis and Results

Figure 1: Estimated Revenue by Product Category

Figure 1 presents a bar chart illustrating total estimated revenue aggregated by product category. The visualization highlights clear differences in revenue concentration across categories, indicating that a small subset of categories accounts for a disproportionate share of total estimated revenue.
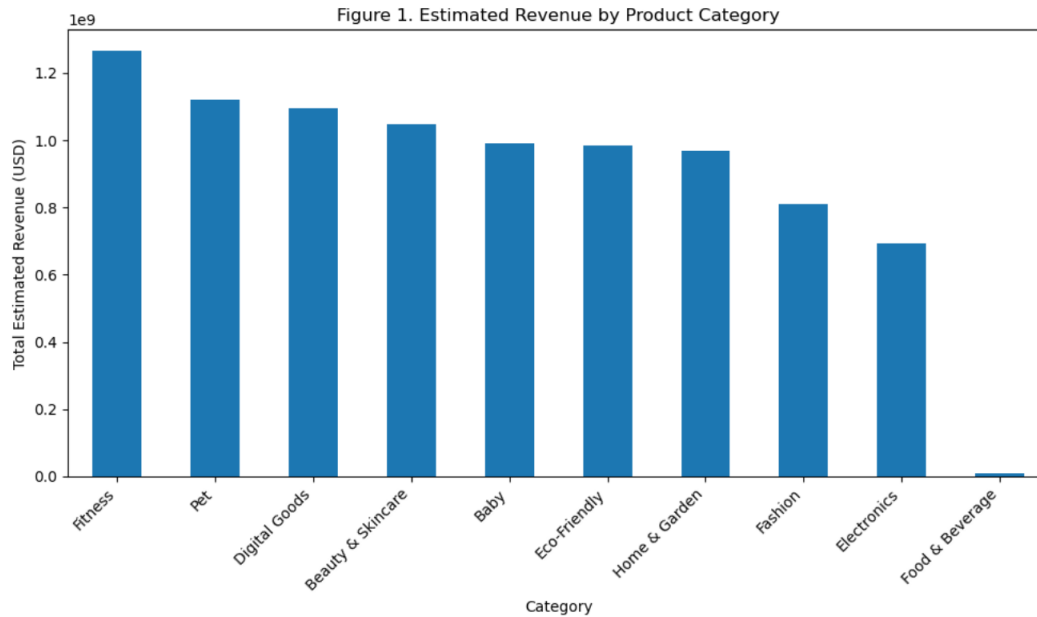
Figure 1. Estimated Revenue by Product Category

Figure 2: Relationship Between Trend Score and Estimated Revenue

Figure 2 displays a scatter plot showing the relationship between product trend score and estimated revenue. The results indicate a positive association between higher trend scores and increased revenue, although the relationship is not strictly linear. This finding suggests that while trend score is an important indicator of performance, additional factors such as pricing and category context also play a significant role.
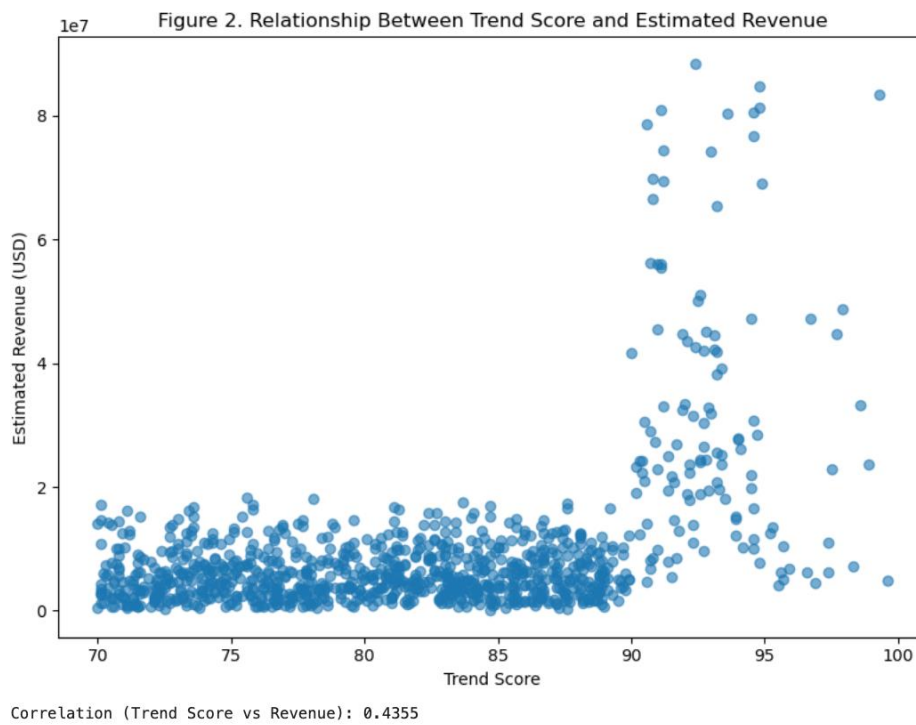

Figure 2. Relationship Between Trend Score and Estimated Revenue

Correlation (Trend Score vs Revenue): 0.4355

Figure 3: Price Range Distribution of High-Performing Products

Figure 3 presents a boxplot comparing price range midpoints among high-performing products across top revenue-generating categories. The visualization shows that mid-priced products tend to dominate high-performing segments, outperforming both low-price and premium-price extremes. This pattern supports the hypothesis that pricing strategy is a critical determinant of revenue performance.
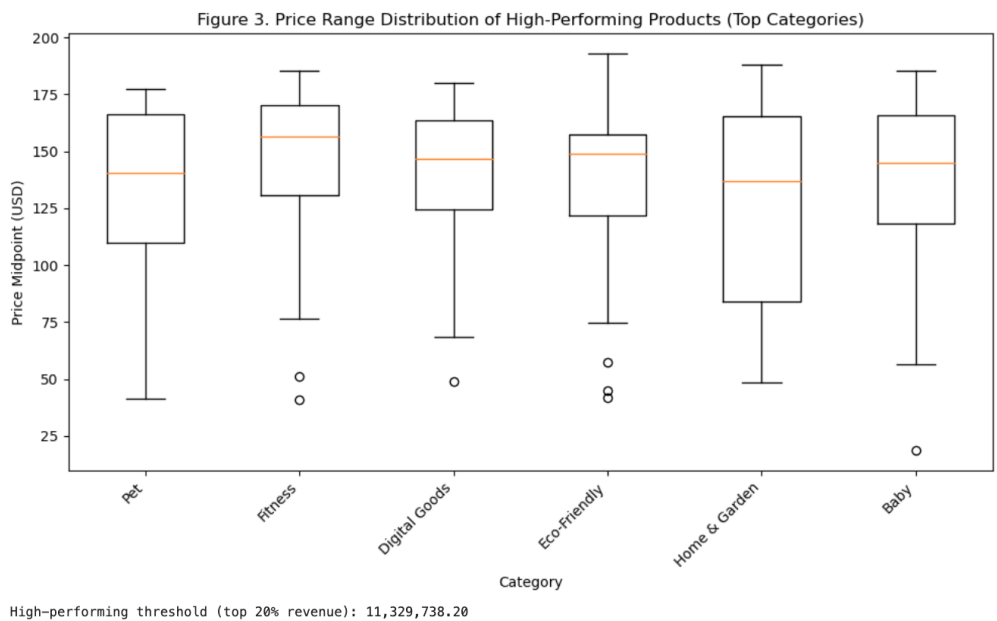

Figure 3. Price Range Distribution of High-Performing Products (Top Categories)

High-performing threshold (top 20% revenue): 11,329,738.20

Figure 4: Random Forest Feature Importance

Figure 4 illustrates the relative importance of predictor variables derived from the Random Forest model. Trend score emerges as the most influential feature, followed by price midpoint and product category indicators. These results confirm that trend-based metrics, when combined with pricing information, provide meaningful predictive value in explaining revenue outcomes.
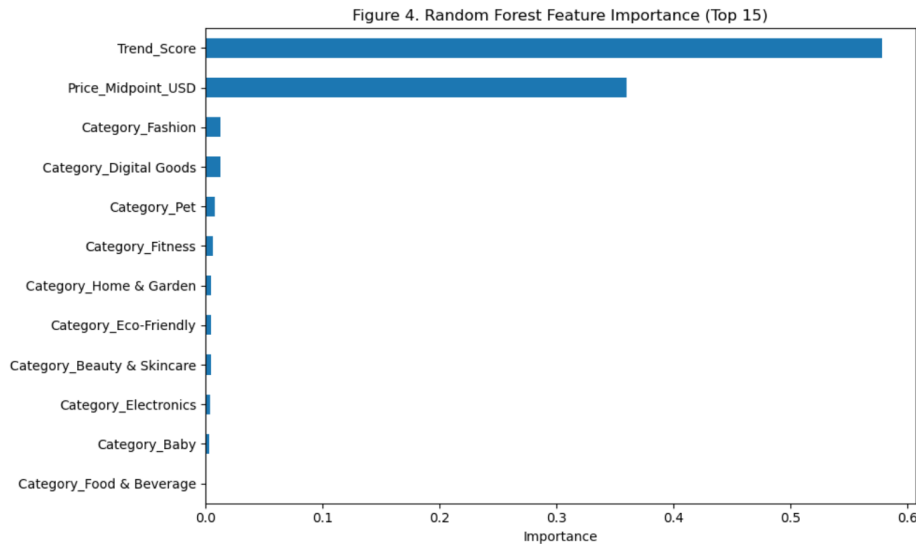
Figure 4. Random Forest Feature Importance (Top 15)

Figure 5: Predicted vs. Actual Revenue Comparison

Figure 5 presents a model validation plot comparing predicted revenue values against observed estimates for the Random Forest model. The clustering of points around the reference line indicates reasonable predictive accuracy, demonstrating the model's ability to generalize beyond the training data. While prediction error remains present, the model performs well as a decision-support tool rather than a precise forecasting mechanism.

Figure 5. Predicted vs. Actual Revenue (Random Forest)
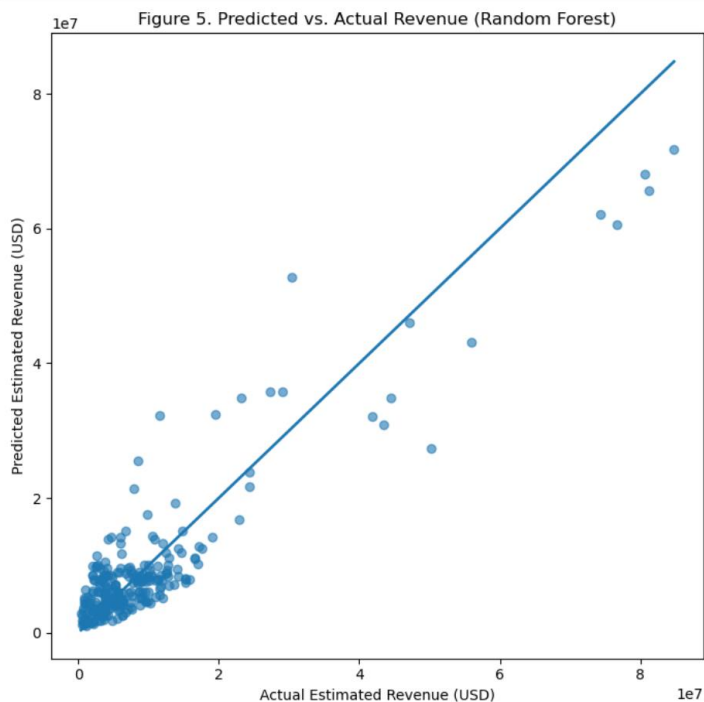
Model Performance (Test Set):
RMSE: 5,241,103.86
MAE: 3,648,391.16
R^2: 0.8353

Key findings indicate that trend score and price range jointly influence revenue performance, with certain mid-priced categories outperforming both low- and high-price extremes. Random Forest results suggest that trend score is the strongest predictor, followed by price midpoint and product category.

Conclusion

The findings of this study demonstrate that data-driven trend analysis provides valuable insights into e-commerce product performance. Trend score, pricing strategy, and product category jointly influence estimated revenue outcomes, with mid-priced products and high trend scores associated with stronger performance. The Random Forest model outperforms linear regression by capturing nonlinear relationships and interactions among variables.

Rather than relying on intuition or short-lived viral trends, digital retailers can use structured analytics to identify sustainable product opportunities. These results highlight the importance of combining trend-based indicators with pricing context to support informed decision-making in competitive e-commerce environments.

Assumptions

This analysis assumes that estimated revenue values reasonably reflect relative product performance and that trend scores accurately represent consumer interest. It is also assumed that product categories are consistently defined across data sources and that observed relationships remain stable during the period represented by the dataset.

Limitations

Several limitations must be acknowledged. Revenue and sales figures are estimates rather than verified financial records. The dataset represents a snapshot in time and does not capture longitudinal trends or seasonality effects. Additionally, external factors such as marketing spend, brand recognition, and supply chain constraints are not explicitly modeled.

Challenges

Key challenges encountered include translating trend-based metrics into actionable financial insights, balancing model interpretability with predictive accuracy, and managing noise inherent in third-party aggregated data. Careful preprocessing and validation were required to mitigate these challenges.

Future Uses and Applications

Future extensions of this project could incorporate time-series data, social media engagement metrics, or advertising spend variables. These enhancements would improve forecasting accuracy and allow the framework to support enterprise-level product planning, inventory optimization, and real-time trend monitoring.

Ethical Assessment

Ethical considerations include transparent communication of data limitations, avoidance of overstatement regarding revenue implications, and responsible interpretation of predictive results. The analysis is intended as a decision-support tool rather than a guarantee of financial outcomes. Because the dataset contains no personal or identifiable information, privacy risks are minimal.

Appendix A: Supporting Documentation

- Data dictionary
- Feature engineering summary
- Model configuration parameters

Audience Questions

1. How reliable are the estimated revenue figures for real-world investment decisions?
2. Can trend score predict long-term product success or only short-term demand?
3. How often should this analysis be updated to remain actionable?
4. What role does seasonality play in trend-driven products?
5. How transferable are these findings across geographic markets?
6. Can this framework be automated for continuous monitoring?
7. How would marketing spend alter these results?
8. What risks exist in over-optimizing trend-based products?
9. How can small businesses implement this analysis with limited resources?
10. How does this approach compare to traditional market research methods?

References (APA Style)

Shopify. (2025). Trending products and social commerce insights.

McKinney, W. (2022). Python for data analysis. O'Reilly Media.

Provost, F., & Fawcett, T. (2013). Data science for business. O'Reilly Media.