Dataset Selection

By Mohammad Movahedi

Introduction

The emergence of big data has helped businesses to better understand how consumers behave, and thereby make appropriate changes to their strategies thereof. Understanding these dynamics becomes important in a highly competitive online retail store. This document presents a case of an extensive data set provided by Amazon Canada that is planned to help in the unraveling of the complex phenomenon of the interplay of consumer behavior, with emphasis on rating and customer review and pricing policy.

Problem Statement

The core question this analysis seeks to address is: "How do customer ratings and the number of reviews influence the pricing and sales success of products on Amazon Canada?" This question stems from the need to understand if and how consumer feedback impacts a product's market performance and pricing approach. In an era where customer reviews can significantly sway purchasing decisions, analyzing this correlation is vital for retailers and manufacturers to optimize their pricing strategies and product offerings.

1. The following questions can be asked as secondary questions
   How do customer ratings influence the pricing and sales success of products on Amazon Canada?
2. In what ways do the number of reviews impact the pricing strategies and sales performance of products on Amazon Canada?
3. Does consumer feedback (in the form of ratings and reviews) have a measurable impact on a product's market performance, including its pricing?
4. How does the relationship between customer feedback and product pricing vary across different product categories on Amazon.ca?
5. Can the trends in pricing and sales success be predicted based on the metrics of customer ratings and reviews for products?
6. What is the nature and strength of the correlation between customer ratings, review counts, and product pricing on Amazon Canada?
7. How can machine learning algorithms be utilized to forecast future pricing strategies based on existing customer feedback data?

Dataset Description

The dataset for this analysis is “Amazon Canada Products 2023 (2.1M products)," sourced from Kaggle, a renowned platform for big data and analytics. This dataset provides an exhaustive view of the product landscape on Amazon.ca, encapsulating over 2.1 million unique products. The data, collected through a comprehensive web scraping process in 2023, encompasses these elements :

1. Product ID (asin)
   a. Total Unique Values: 2.17 million
   b. Most Common Product ID: B07CV4L6HX
2. Product Title
   a. Total Unique Titles: 2.07 million
3. Product Image URL (imgUrl)
   a. Total Unique URLs: 1.90 million
4. Product Page URL (productURL)
   a. Total Unique URLs: 2.17 million
5. Product Rating (stars)
   a. Range: 0.00 to 5.00
   b. Majority of products have a rating of 0 (847,514 products)
6. Number of Reviews (reviews)
   a. Majority of products have between 0 to 17,377 reviews
   b. Mean Review Count: 546
7. Current Price (price)
   a. Majority of products priced between 0.00 to 818.00 CAD
   b. Mean Price: 111 CAD
8. Original Price (listPrice)
   a. Majority of products priced between 0.00 to 20.00 CAD
   b. Mean Original Price: 4.65 CAD
9. Product Category (categoryName)

    a. Most Common Categories: Baby, Luggage Travel Gear, Other (98%)
    b. Unique Category Count: 266

10. Bestseller Status (isBestSeller)

    a. 7,654 products labeled as true (bestsellers)
    b. 2,158,272 products labeled as false

11. Amount Bought in Last Month (boughtInLastMonth)
    a. Range: 0 to 20,000 units
    b. Majority of products (2,153,992) had sales between 0 to 400 units
    c. Higher sales brackets (above 400 units) have progressively fewer products
    d. Mean Quantity Sold: 9 units
    e. Standard Deviation: 98.4 units

Relevance to the Problem

This dataset is particularly suited to address the proposed question for several reasons:

- Volume and Variety: The sheer size of the dataset, covering more than 2.1 million products, offers a broad spectrum of data for analysis, ensuring that the findings are statistically significant and representative of the overall market.

- In-depth Information: The inclusion of detailed product information, pricing, and customer reviews allows for a multifaceted analysis. It facilitates an exploration of how different product categories respond to consumer feedback in terms of pricing adjustments and sales performance.

- Currency and Reliability: Given that the dataset is recent (2023) and sourced from a reliable platform, it provides up-to-date insights into current market trends and consumer preferences.

The analysis will involve several key steps:

1. Data Cleaning and Preparation: Ensuring the data is clean and structured appropriately for analysis, focusing on relevant variables.

2. Descriptive Analysis: Understanding the basic trends and patterns in the data, such as average pricing, rating distributions, and review counts across different product categories.
3. Correlation Analysis: Investigating the relationship between customer ratings, review counts, and product pricing. This would involve statistical techniques to determine the strength and nature of these relationships.
4. Predictive Modeling: Using machine learning algorithms to predict pricing strategies based on customer feedback metrics. This can help in understanding the potential impact of consumer reviews on future pricing decisions.
5. Conclusion

This project aims to use the power of big data to uncover insights into consumer behavior and its influence on pricing strategies in the online retail space. By leveraging the comprehensive dataset from Amazon Canada, the analysis will provide valuable inputs for retailers and manufacturers to strategize effectively in the ever-evolving e-commerce market.

References

asaniczka (2023). *Amazon Canada Products 2023 (2.1M Products)*. [online] Kaggle.com. Available at: https://www.kaggle.com/datasets/asaniczka/amazon-canada-products-2023-2-1m-products/data [Accessed 13 Nov. 2023].