2.   **CHAPTER 2**

LITERATURE REVIEW

2.1. Overview

This chapter reviews existing literature relevant to the prediction of product performance on e-commerce platforms, particularly Amazon. It begins by introducing Amazon's Best Seller Rank (BSR) as a key indicator of product success, followed by a discussion of current applications of machine learning in the e-commerce sector. The review then narrows its focus to previous efforts in sales prediction and customer analytics, emphasising the methodological limitations and research gaps that this study aims to address.

2.2. E-commerce Analytics and Product Ranking

In the digital marketplace, product visibility is an essential element for commercial success, as it directly affects consumer engagement and sales performance. Amazon, as one of the largest global e-commerce platforms, ranks products using the BSR—a score assigned based on sales volume and historical performance. Since the BSR is updated hourly, the BSR directly influences a product's visibility in category rankings and search results, thus affecting consumer behaviour and purchase decisions.

Although the commercial importance of BSR is well known, the algorithm used by Amazon to assign the BSR remains a commercial secret. Due to this, the impact of product features on the BSR has become an increasingly interesting topic for academic as well as strategic research.

2.3. Machine Learning Applications in E-commerce

The advent of big data has led to a significant increase in the application of machine learning (ML) techniques in e-commerce. Researchers have applied ML to tackle tasks, inculding customer segmentation (Rajyalaxmi M et al., 2024), return behaviour prediction (Ramirez, 2024), and sales forecasting (Sharma et al., 2022). Techniques commonly used include supervised learning (e.g., decision trees, support vector machines), unsupervised clustering (e.g., K-means, DBSCAN), and ensemble methods (e.g., Random Forest, XGBoost). Notably, the integration of deep learning architectures—particularly recurrent neural networks (RNNs)—has enabled the modelling of complex sequential behaviours such as browsing or purchase patterns, offering deeper insight into temporal consumer dynamics.

Most studies try to understand customers using their data uch as demographics, purchase logs, or session histories, but very few focus on products, like figuring out what makes one product rank higher than another on Amazon. That area is still not well-studied.

2.4. Prediction of Sales and Ranking Metrics

Several studies have attempted to model the relationship between product features and sales outcomes. Sharma et al. (2020) investigated how BSR maps to sales in Amazon's Clothing, Shoes & Jewellery category, using hourly sales and review data collected via a custom web scraper. Their linear regression model achieved moderate predictive accuracy (R^2 = 69.46%), demonstrating a clear correlation between BSR and sales. However, their approach predicts sales from BSR, not the reverse.

Other works have developed models to estimate return rates or classify customer purchasing behaviours, but rarely have these models been designed to predict BSR as the target variable. Moreover, such studies typically focus on a single category or use a limited set of features, limiting the generalisability of their findings.

2.5. Dataset Characteristics in Existing Research

The datasets used in current e-commerce research fall broadly into two types:

- Customer-Behavioural Data: Clickstreams, Transaction Logs, And Demographic Profiles, Commonly Used For Segmentation And Churn Prediction .

- Product-Level Data: Prices, Reviews, And Metadata, Which Are Less Commonly Used For Predictive Modelling But More Publicly Accessible.

While Sharma et al. (2020) collected rich time-series data on BSR, price, and reviews, their sample was limited to 10 products over 3 months. Other studies employ proprietary or platform-internal data, which limits reproducibility and open research. Furthermore, very few datasets contain textual review sentiment, category ratings, or temporal sales signals, all of which may play important roles in shaping BSR.

2.6. Limitations of Existing Studies

Although advancements in using ML for e-commerce analytics, several limitations persist in the literature:

- Narrow Task Focus: Many Studies Predict Sales Or Returns, Not Ranking Outcomes Like Bsr.

- Single-Category Constraints: Models Are Often Built On One Category (E.G., Books Or Electronics), Reducing Scalability Across Product Types.

- Lack Of Multivariate Analysis: Few Works Combine Multiple Structured Features (E.G. Price + Reviews + Ratings) To Predict Performance.

- Insufficient Interpretability: Black-Box Models Are Rarely Accompanied By Explanations, Which Limits Their Usefulness To Business Stakeholders.

- Minimal Attention To Bsr Volatility: Bsr Updates Hourly, But Time-Sensitive Models Such As Lstm Or Attention-Based Architectures Are Rarely Used.

2.7. Research Gap

Although BSR plays a critical role in shaping product visibility and success on Amazon, there is limited research that directly predicts BSR using multivariate machine learning models. Most existing works either focus on adjacent problems (e.g., customer segmentation or sales forecasting) or use simplistic methods such as linear regression with minimal feature sets. Additionally, models are often trained on limited data within a single category, limiting their generalisability. There is also a lack of attention to textual review features and time-aware model designs that could capture the dynamic nature of BSR. This gap presents an opportunity to develop interpretable, feature-rich predictive models for BSR using structured product data across diverse categories.

2.8. Summary of Literature Review

The reviewed literature demonstrates a growing interest in using machine learning to analyse e-commerce performance. However, while customer-centric problems are well-studied, product-centred outcomes such as BSR remain underexplored. Prior studies show the feasibility of mapping BSR to sales but do not attempt to forecast BSR using relevant product attributes. Moreover, few studies address cross-category generalisability, interpretability, or the incorporation of time-based behaviour. This research seeks to fill those gaps by building a predictive model for BSR using multivariate, structured product data and exploring methods that balance accuracy with explainability.