Demand Forecasting and Recommendation Engines in Retail and E-Commerce

Avinash Changdev Taskar
Assistant Professor
Department of Computer Science and Engineering
Sandip University, Nashik

Piyush Sanjay Mhatre
B-Tech CSE Student
School of Computer Science and Engineering
Sandip University, Nashik

Riya Vijay Agnihotri
B-Tech CSE Student
School of Computer Science and Engineering
Sandip University, Nashik

Moulababa Nanhesab Sheikh
B-Tech CSE Student
School of Computer Science and Engineering
Sandip University, Nashik

Atif Tasnim Shaikh
B-Tech CSE Student
School of Computer Science and Engineering
Sandip University, Nashik

ABSTRACT
Retailers and e-commerce companies are leaning more and more on data-driven methods to anticipate what customers will buy and to tailor the shopping experience around it, although the research on these two capabilities is usually reviewed in isolation even though the two now overlap heavily in practice. This survey pulls together recent work on demand forecasting and recommendation engines in retail and e-commerce, covering classical statistical models such as ARIMA and exponential smoothing, machine learning approaches such as random forests and gradient boosting, and deep learning architectures including LSTM, CNN-based, and transformer-based forecasters. On the personalization side, we look at collaborative filtering, content-based filtering, matrix factorization, and deep neural and graph-based recommenders, as well as hybrid and sentiment-aware variants. Across more than a dozen recent studies, ensemble and deep sequential models generally outperform classic baselines on large retail datasets while hybrid recommendation architectures handle cold-start and sparsity problems better than single-technique systems. We also find a convergence trend, where demand signals and behavioral embeddings are increasingly used together to sharpen both inventory planning and product rankings. Persistent challenges remain around data quality, scaling to real-time streaming environments, explainability, and evaluating business impact beyond offline accuracy metrics. The survey closes by outlining open research directions, including unified forecasting-recommendation frameworks, privacy-preserving personalization, and the use of large language models for conversational commerce.

Keywords
Demand Forecasting, Recommendation Engines, Retail Analytics, E-Commerce Personalization, Deep Learning.

1. INTRODUCTION
Retail and e-commerce businesses operate on thin margins in a demand environment that is constantly shifting, making understanding what customers will buy and helping them find it central to staying competitive. Two analytical capabilities are at the heart of this shift: demand forecasting, which predicts future sales volumes across products, stores, and channels, and recommendation engines, which personalize the assortment every shopper sees. These two were historically studied as separate disciplines, rooted in time-series econometrics and information retrieval, respectively, but have converged as retailers move to unified data platforms that feed the same customer and transaction data into both forecasting and personalization pipelines [5].

A large body of work has looked at demand forecasting from classical statistical models such as moving averages, exponential smoothing, and ARIMA, which are appealing for their interpretability and modest data requirements, while recommendation research is moving from simple item-to-item collaborative filtering to deep neural and graph-based architectures as they can capture complex, non-linear user–item interactions [1]. Most existing surveys treat these two streams of work independently — reviewing forecasting accuracy separately from personalization quality — even though both problems draw on overlapping behavioral and transactional data, and increasingly share model architectures such as recurrent and transformer networks. Not much work has systematically compared the techniques, evaluation practices, and open challenges across both domains within a single retail and e-commerce framework.

This survey aims to consolidate recent literature on demand forecasting and recommendation engines in retail and e-commerce, comparing how each field has evolved methodologically, noting where they intersect, and pointing out practical and research gaps. This intersection matters because retail businesses that align demand forecasts with personalized recommendations can better balance inventory availability with product discovery, minimizing both stockouts and unsold inventory while making what customers see more relevant [7]. The review is meant to provide a reference for researchers designing new forecasting or recommendation models, and for practitioners trying to work out what techniques fit their operational constraints.

The rest of the paper is organized as follows. Section 2 reviews the literature on demand forecasting and recommendation engines, tracing the path from classical statistical and memory-

based methods to modern deep learning approaches. Section 3 lays out a comparative taxonomy of the surveyed techniques. Section 4 discusses what this comparison reveals, including a table of representative studies. Section 5 covers the practical advantages of integrating forecasting and recommendation capabilities, Section 6 discusses the broader social and economic benefits for retail ecosystems, Section 7 points to promising directions for future research, and Section 8 concludes the paper [9].

2. LITERATURE SURVEY

2.1 Demand Forecasting in Retail

Classical statistical models — moving averages, exponential smoothing, ARIMA — formed the backbone of early retail forecasting research, valued for being interpretable and cheap to compute, although struggled with irregular demand, promotions, and multiple interacting product categories [3]. Tree-based ensemble methods have since become a strong practical alternative: a comparison of extra tree regressors against long short-term memory networks on more than five million daily demand records found the ensemble approach outperformed the recurrent model on most error metrics, particularly for perishable categories, although the gap narrowed for less volatile products [1]. Recurrent architectures still hold their place when longer historical context and external covariates are available; combining LSTM networks with random forests in a multi-channel retail setting produced forecasts that adapted better to channel-specific demand patterns than either model on its own [2]. More recent work has explored convolutional and hybrid CNN-BiLSTM designs that capture both local demand spikes and longer sequential dependencies, reporting accuracy gains over single-architecture baselines at the store-item level [4]. Complementing these technique-focused studies, broader reviews caution that overloading model complexity does not always translate into better business outcomes, arguing for simpler, more frugal forecasting pipelines unless the added complexity earns its keep through measurable accuracy or cost improvements [5]. A recent bibliometric review of the demand-forecasting literature also documents a sharp rise in publication activity between 2021 and 2024, reflecting growing interest in applying machine learning and deep learning to supply chain and retail demand problems [6].

2.2 Recommendation Engines in E-Commerce

Item-to-item collaborative filtering, popularized through Amazon's early recommendation architecture, set the template of computing similarity between products from co-purchase patterns rather than relying purely on user-to-user comparisons — a practice still cited as the foundation of modern collaborative filtering systems [7]. As catalogs and interaction data grew, memory-based collaborative filtering gave way to model-based and deep learning variants; a systematic review covering more than a hundred studies published between 2019 and 2023 sorts these approaches into memory-based and model-based families, finding that deep neural collaborative filtering consistently improves ranking quality over traditional matrix factorization, with reduced interpretability and higher training overhead [9]. Hybrid architectures that jointly learn from wide, memorized feature interactions and deep, generalized embeddings were shown to improve recommendation relevance in large-scale production settings compared with either a purely linear or purely deep model on its own [8]. Comparative studies across e-commerce big data further show that no single recommendation algorithm dominates across all product categories, and that hybrid approaches combining collaborative and content-based signals outperform single-technique baselines, particularly under sparse interaction data [10]. Retrieval-based personalization strategies and sentiment-aware recommenders that pull in review text through transformer-based language models have also emerged as promising extensions, reporting better relevance and cold-start handling relative to interaction-only models [11][12].

2.3 Convergence and Research Gap

Taken together, the forecasting and recommendation literatures reveal a similar trajectory — from simple, interpretable statistical or similarity-based methods toward deep, data-hungry architectures that sacrifice transparency for accuracy. Yet the two are rarely reviewed together, even though both increasingly draw from the same transaction logs, session data, and sequential modeling techniques such as LSTM and transformer networks. Few studies look at how forecast outputs could inform recommendation ranking, or, the other way around, how personalization signals could sharpen demand estimates at the product-store level. This gap is what motivates the comparative taxonomy and discussion in the sections that follow.

3. COMPARATIVE TAXONOMY OF TECHNIQUES

Pulling together the studies reviewed in Section 2, the surveyed techniques can be organized along two axes: model family (statistical, classical machine learning, or deep learning) and task (demand forecasting or recommendation). On the forecasting axis, statistical methods (ARIMA, exponential smoothing) are at the interpretable, low-data end; ensemble methods such as random forests, extra tree regressors, and gradient boosting are in the middle, offering strong accuracy with moderate data and tuning requirements; and sequential deep learning models (LSTM, CNN-BiLSTM hybrids, transformer-based forecasters) are at the high-data, high-accuracy end, needing larger, cleaner historical datasets and more computational investment. On the recommendation axis, a similar progression runs from neighborhood-based collaborative filtering, through matrix factorization and content-based filtering, to deep neural collaborative filtering, wide-and-deep architectures, and graph- or transformer-based recommenders that incorporate review text and knowledge graphs.

A third, cross-cutting category is emerging at the intersection of both axes: models that jointly exploit sequential transaction data for forecasting and behavioral embeddings for personalization, usually built on shared recurrent or transformer encoders that feed separate forecasting and ranking heads. Table 1 summarizes representative studies against this taxonomy, comparing their technique family, target task, and reported strengths, and serves as the basis for the comparative discussion in Section 4. By laying out the studies this way, it becomes evident where the literature is dense — deep learning applied to single-task forecasting or single-task recommendation — and where it is sparse, namely, joint or transferable forecasting-recommendation architectures for retail and e-commerce [6].

4. RESULTS AND DISCUSSION

Applying the taxonomy from Section 3 to the surveyed literature highlights three consistent patterns. First, for demand forecasting, ensemble and deep sequential models outperform classical statistical baselines on large, high-frequency retail datasets, but the margin narrows for smoother, less volatile product categories, where simpler models remain competitive at a fraction of the computational cost. Second, for recommendation engines, hybrid architectures that combine collaborative signals with content, sentiment, or knowledge-

graph features consistently outperform single-technique systems, particularly under sparse or cold-start conditions. Third, very few of the reviewed studies evaluate forecasting and recommendation jointly, even though both tasks increasingly run on shared sequential encoders — which is an open opportunity rather than a settled research question.

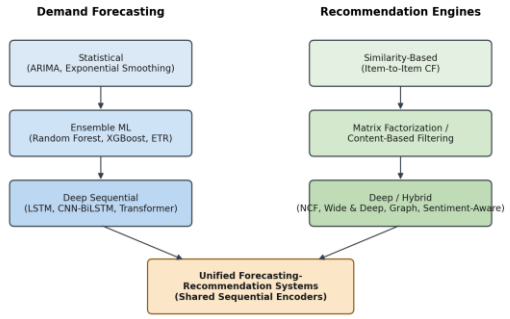

Fig. 1. Taxonomy of Demand Forecasting and Recommendation Engine Techniques Reviewed in this Survey.

These patterns have practical implications: retailers choosing a forecasting technique should weigh data volume and volatility against interpretability needs rather than defaulting to the most complex model available, and recommendation teams facing sparse interaction data should lean toward hybrid designs rather than pure collaborative filtering. Three limitations of this review are worth acknowledging. First, most reported accuracy gains come from offline benchmarks on proprietary or single-retailer datasets, which limits how comparable the studies are. Second, few studies report inference latency or infrastructure cost, both of which matter a great deal for real-time retail deployment. Third, business-level outcomes such as revenue or waste reduction are rarely measured, so offline accuracy improvements do not always translate to demonstrated commercial value. These gaps are what motivate the future research directions discussed in Section 7 [10].

Table 1. Representative Studies on Demand Forecasting and Recommendation Engines in Retail/E-Commerce.

Sl. No.	Study	Technique	Task Domain	Key Reported Outcome
1	Punia et al. (2020) [2]	LSTM + Random Forest	Demand Forecasting	Better accuracy in multi-channel retail forecasting
2	Kharfan et al. (2023) [1]	Extra Tree Regressor vs. LSTM	Demand Forecasting	ETR beats LSTM on most error metrics
3	Joseph et al. (2022) [4]	CNN + BiLSTM Hybrid	Demand Forecasting	Better store-item level accuracy than single models
4	Linden et al. (2003) [7]	Item-to-Item Collaborative Filtering	Recommendation	Established a scalable, similarity-based recommendation approach
5	Cheng et al. (2016) [8]	Wide & Deep Learning	Recommendation	Better relevance than linear- or deep-only models
6	Bobadilla et al. (2023) [9]	Deep Neural Collaborative Filtering	Recommendation	Better ranking quality than matrix factorization

5. ADVANTAGES OF AN INTEGRATED FORECASTING-RECOMMENDATION APPROACH

Across the literature reviewed, four practical advantages stand out for retailers that adopt data-driven demand forecasting and recommendation engines together rather than running them as siloed systems. First, forecasting models built on ensemble or deep sequential architectures consistently reduce stockouts and overstock relative to classical statistical baselines, particularly for high-velocity or promotion-sensitive products, which directly improves inventory turnover [1][2]. Second, hybrid recommendation engines that combine collaborative, content-based, and sentiment-aware signals ease the cold-start and sparsity problems that limit single-technique systems, improving how relevant recommendations are for new products and infrequent shoppers [9][10]. Third, sharing sequential encoders and transaction embeddings between forecasting and recommendation pipelines reduces engineering duplication and allows demand signals to inform product rankings, so items likely to be understocked aren't aggressively promoted, and vice versa [6]. Fourth, both capabilities scale to real-time, high-throughput retail environments built on modern deep learning frameworks, supporting dynamic pricing, personalized promotions, and same-day fulfillment decisions that static, rule-based systems cannot support. Taken together, these advantages suggest the value of demand forecasting and recommendation technology in retail lies not just on how accurate each capability is on its own, but in the operational coherence gained from integrating them.

6. SOCIAL AND ECONOMIC IMPACT

The technologies reviewed in this survey carry benefits that extend beyond individual retailers. More accurate demand forecasting reduces overproduction and food waste in grocery and perishables categories, lowering the environmental footprint of unsold inventory and its disposal, while also reducing stockouts that affect lower-income shoppers most severely, since they have limited flexibility to travel between stores. Personalized recommendation engines, when designed responsibly, can improve access to relevant products for shoppers with limited time or research capacity, and support small or niche sellers on e-commerce marketplaces by surfacing their products to interested buyers who might not have found them through generic search. These benefits are not automatic, however: opaque personalization can narrow the range of products shoppers see, and demand forecasts built on historical data can perpetuate existing biases in who a retailer's supply chain actually serves well. Realizing the social value of these systems requires deliberate design choices — diversity-aware recommendation ranking, fairness checks in forecasting pipelines, and transparency on how personalization decisions are made — rather than accuracy improvements alone.

7. FUTURE RESEARCH DIRECTIONS

While the reviewed literature shows clear progress in demand forecasting and recommendation engines individually, several directions remain underexplored. The most immediate is building unified forecasting-recommendation architectures that share sequential representations of transaction and browsing data, allowing forecast uncertainty to inform recommendation ranking and promotional strategies instead of treating the two as independent downstream tasks [6]. A second direction is real-time and streaming adaptation: most surveyed studies evaluate models on static, historical datasets, but production retail systems need incremental learning that can adapt to sudden demand shifts — a viral trend, a supply disruption — without full retraining. A third direction is privacy-preserving personalization, using techniques such as federated learning or on-device inference so recommendation quality does not depend on centralizing sensitive shopping histories. A fourth direction is the use of large language models for conversational commerce, where natural-language product search and dialogue-based recommendation could subsume and extend traditional collaborative and content-based methods, building on the early sentiment-aware and retrieval-based recommenders already reported in the literature [11][12]. Finally, future work should focus more on business-relevant evaluation — inventory cost, waste reduction, revenue impact — alongside standard accuracy and ranking metrics, so reported technical gains connect more closely to retail outcomes.

8. CONCLUSION

This survey set out to bring together recent research on demand forecasting and recommendation engines in retail and e-commerce, and the literature reviewed shows both fields following a similar arc — from classical statistical or similarity-based methods toward ensemble and deep learning architectures that sacrifice some interpretability for measurable gains in accuracy and relevance. Critically, the review shows the two are rarely studied together despite drawing from overlapping transactional and behavioral data and increasingly similar sequential model architectures, which is the central gap in the current literature. This suggests retailers and researchers alike stand to gain from unified forecasting-recommendation frameworks that allow inventory signals to inform personalization and vice versa, rather than optimizing each capability in isolation. This survey did not run new empirical experiments and instead synthesizes published results that vary in dataset scale, evaluation protocol, and how they report business-level outcomes; even so, future work bridging these two domains — supported by real-time adaptation, privacy-preserving personalization, and conversational, language-model-driven interfaces — is well positioned to make retail and e-commerce systems more accurate, relevant, and responsive to what customers actually want.

9. REFERENCES

[1] Kharfan, M., Chan, V. W. K., & Firdolas Efendigil, T. (2023). Applying machine learning in retail demand prediction-a comparison of tree-based ensembles and long short-term memory-based deep learning. Applied Sciences, 13(19), 11112.

[2] Punia, S., Nikolopoulos, K., Singh, S. P., Madaan, J. K., & Litsiou, K. (2020). Deep learning with long short-term memory networks and random forests for demand forecasting in multi-channel retail. International Journal of Production Research, 58(16), 4964-4979.

[3] Tsoumakas, G. (2018). A survey of machine learning techniques for food sales prediction. Artificial Intelligence Review, 52(1), 441-447.

[4] Joseph, R. V., Mohanty, S. N., et al. (2022). A hybrid deep learning framework with CNN and Bi-directional LSTM for store item demand forecasting. Computers and Electrical Engineering, 103, 108358.

[5] Petropoulos, F., Grushka-Cockayne, Y., Siemsen, E., & Spiliotis, E. (2022). Wielding Occam's razor: Fast and frugal retail forecasting [Working paper]. University of Bath.

[6] Ghobakhloo, M., et al. (2024). Machine learning and deep learning models for demand forecasting in supply chain management: A critical review. Analytics, 7(5), 93.

[7] Linden, G., Smith, B., & York, J. (2003). Amazon.com recommendations: Item-to-item collaborative filtering. IEEE Internet Computing, 7(1), 76-80.

[8] Cheng, H. T., Koc, L., Harmsen, J., Shaked, T., Chandra, T., Aradhye, H., ... & Shah, H. (2016). Wide & deep learning for recommender systems. In Proceedings of the 1st Workshop on Deep Learning for Recommender Systems (pp. 7-10).

[9] Bobadilla, J., González-Prieto, Á., Ortega, F., & Lara-Cabrera, R. (2023). Deep learning-based collaborative filtering recommender systems: A comprehensive and systematic review. Neural Computing and Applications, 35, 1-30.

[10] Jin, Z., Ye, F., Nedjah, N., & Zhang, X. (2024). A comparative study of various recommendation algorithms based on E-commerce big data. Electronic Commerce Research and Application, 68, 101461.

[11] Nguyen, D-N., Nguyen, V.-H, Trinh, T., Ho, T., & Le, H.-S. (2024). A personalized product recommendation model in e-commerce based on retrieval strategy. Journal of Open Innovation: Technology, Market, and Complexity, 10(2), 100303.

[12] Thomas, R., & Jeba, J.R. (2024). A novel framework for an intelligent deep learning based product recommendation system using sentiment analysis (SA). Automatika, 65(2), 410424.