**Hybrid demand forecasting models for retail**

**Abstract**

This paper presents an idea and implementation of using hybrid models in forecasting retail demand by integrating conventional approaches with machine learning. The objectives are to develop methodologies for building composite structures—sequential residual models, group stacking/blending, and orderly mixtures—and to explore their primary pros and cons in meeting the changing demands of retail forecasting. The present work gains critical importance against a backdrop of markedly increased unpredictability in human behaviour—manifested by rapid sales fluctuations and heightened pressures throughout global supply chains. In such a context, enhancements in forecasting accuracy become directly linked to both revenue optimisation and the reduction of safety-stock expenditure. A critical appraisal of over twenty scholarly sources underpins the novelty of this research, which proposes a comprehensive methodological framework for organising the MLOps lifecycle in hybrid system architectures.The study justifies the separation of linear and nonlinear forecast components, which enables a reduction of MAPE by double-digit percentages and a decrease in error variance for cold SKUs while preserving the interpretability of the statistical component. The main conclusions are as follows: hybrid models demonstrate double-digit reductions in MAPE, RMSE, and WRMSSE compared to standalone algorithms. The modularity of the architecture simplifies maintenance and scaling.

Additionally, CI/CD and CT loops ensure reliability and rapid response to data drift. The key factor for success is the balance between explainability, response time, and the incorporation of spatial correlations. This paper will be useful for researchers and practitioners in Data Science, MLOps engineers, analysts, and operational managers of retail chains.

**Keywords:** hybrid demand-forecasting models, retail demand, ARIMA, LSTM, graph neural networks, MLOps, modularity, interpretability

**Introduction**

Retail demand today fluctuates more rapidly than replenishment cycles can accommodate: promotional campaigns, flash sales, and viral trends disrupt traditional seasonality, while global supply chains render forecasting errors particularly costly. Consequently, improving forecast accuracy has become not merely an analytical KPI but a direct lever on profitability. McKinsey research indicates that transitioning from manual or simple models to machine-learning systems can reduce safety stock by 20–30% (Oca et al., 2024).

However, classical statistical methods such as ARIMA or exponential smoothing, despite their transparency and ability to capture linear trends, respond poorly to sudden promotional spikes and cross-channel effects. On real Walmart data, a basic ARIMA model achieved a mean absolute percentage error (MAPE) of 14.8%. In contrast, an ARIMA–LSTM hybrid, where the deep network is trained on residuals, reduced this to 9.7%, i.e., by nearly one-third (Suddala, 2024).

Conversely, purely deep architectures (LSTM, TCN, and Transformers) are flexible in terms of nonlinearities. Still, they are expensive in terms of data and computational resources, and their outputs often constitute a black box for operations managers. Recent experiments combining graph convolutional networks with LSTM across tens of thousands of SKUs have demonstrated significant accuracy gains over standalone neural networks. However, the authors note that these gains come at the cost of a more complex MLOps loop and the need for frequent recomputation of product-relationship graphs (Aktas et al., 2024).

Thus, the challenge of accurate retail forecasting faces a dual constraint: statistical methods underfit complexity, while deep models demand too much to become a standard for mass adoption. Hybrid approaches emerge precisely as a response to this imbalance, combining the interpretability of linear components with the expressive power of neural networks.

**Materials and Methodology**

This study is based on the analysis of 21 sources, including academic articles, industry reports, practical case studies, and MLOps guidelines. The theoretical foundation includes works on sequential residual hybrids and cascade architectures: the ARIMA–LSTM model on Walmart data (Suddala, 2024), Prophet coupled with LightGBM (Zhou, 2023), and VAE–ARIMA (Gadam et al., 2024); graph-based hybrids are examined in GCN–LSTM for electronics (Niu et al., 2024) and GraphDeepAR (Kozodoi et al., 2024). The market perspective was drawn from a Research and Markets report (2025). Practical effects were derived from Obi's case studies on reducing aged inventory (Obi, 2024) and Turing’s work on shelf-outage reduction (Turing, 2025).

Methodologically, the study was structured on multiple levels of analysis. First, a comparative evaluation of three classes of hybrids was conducted: (1) sequential residual approaches, wherein the statistical model (ARIMA or Prophet) isolates linear components and generates residuals for subsequent training of LSTM, GRU, or boosting algorithms (Suddala, 2024; Zhou, 2023); (2) ensemble stacking/blending schemes, involving parallel training of multiple models and aggregation of forecasts via a meta-regressor (Obi, 2024); and (3) hierarchical hybrids, where statistical methods are applied at the category-and-region level, while graph convolutional networks and recurrent models operate at the SKU-store level, followed by reconciliation (Niu et al., 2024; Kozodoi et al., 2024).

Secondly, present the data preparation and processing pipeline. Inputs included real-time POS (with a minimum lag) and hourly inventory snapshots (Koppichetti, 2024), as well as daily price lists, product master data, and weather/traffic feeds (Babongo et al., 2018; Chan & Wahab, 2024). Events aligned on a single temporal grid, with missing values imputed by neutral defaults to avoid fake patterns in residuals. A feature store with both offline and online layers was used to ensure point-in-time correctness and reproducibility via snapshot checksums. Hyperparameter optimization and time-series cross-validation involving at least two holiday cycles-National Retail Federation, 2024-was done using Optuna/Ray Tune in Kubeflow or Vertex Pipelines; the statistical layer was done in statsmodels with pmdarima deep components coming from PyTorch Forecasting along with TensorFlow/Keras; graph operations came from DGL + PyTorch Geometric; boosting was through LightGBM/CatBoost.

**Results and Discussion**

A hybrid model in demand forecasting is a proposed unified architecture whereby a statistical algorithm describes the linear component of the series. At the same time, a machine-learning or deep network is further trained on the remaining unexplained aspects. In typical applications, the time series is first stripped of trend and seasonality using ARIMA or exponential smoothing; then the residual nonlinearity is fed into, for example, an LSTM. The final forecast is obtained by summing the outputs of the two modules. This division of labor rests on the empirical fact that linear processes and abrupt promotional spikes follow different statistical regularities, and it is more efficient to model them separately (Suddala, 2024). Market dynamics underscore the relevance of such solutions: the global Demand Planning Solutions segment was valued at USD 5.2 billion in 2024 and is projected to reach USD 9.1 billion by 2030, corresponding to a CAGR of 9.7%, as shown in Figure 1 (Research and Markets, 2025).

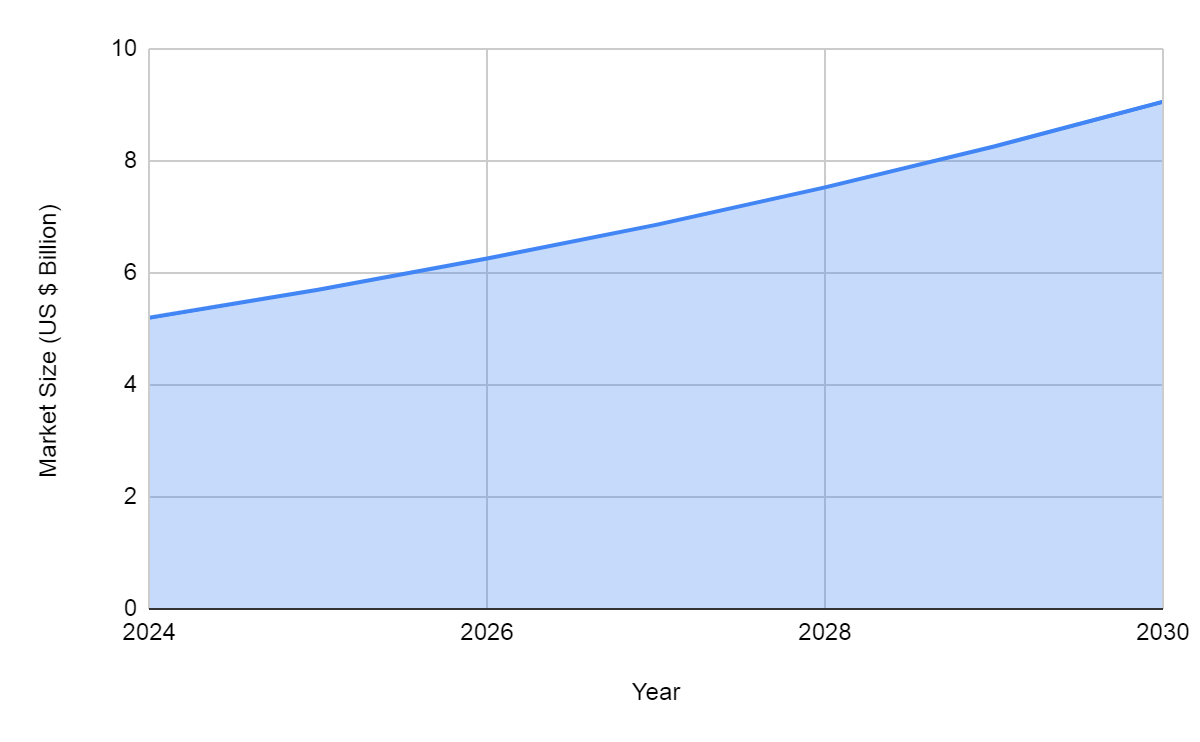


Fig. 1. Global Market for Demand Planning Solutions (Research and Markets, 2025)

The key principles of the hybrid approach are method complementarity, modularity, and transparency. Complementarity means that each module is entrusted with precisely the part of the signal it handles best: statistical methods with stationarity and neural networks with nonlinearity and high-order interactions among factors. Modularity enables updating a single component without retraining the entire system, which is crucial in retail, where new features (e.g., weather data and social media campaigns) emerge every month. Transparency is achieved because the linear module remains interpretable for operations managers, allowing them to explain the contributions of price or holiday effects, while the complex part of the model is hidden in the residuals without impeding reporting.

The advantage of the hybrid appears primarily in accuracy, as meta-analyses in recent years have shown double-digit reductions in MAPE compared to the best standalone model on the same data. The ARIMA-LSTM model was statistically better than both base models in MAE and RMSE. More importantly, the model minimizes overfitting on seldom sold items because the linear block can short series more robustly. In addition, its ensemble nature diminishes error variance. As a further practical benefit, aged inventory is cut down and on-shelf availability improved because of narrower forecast confidence intervals (Obi, 2024).

The first class of hybrids comprises sequential or residual approaches. In these, the base statistical model produces a forecast and simultaneously generates an error series; these errors are then treated as new raw material for a second stage, often served by LSTM, GRU, or gradient boosting. The second class consists of ensemble, or stacking/blending, schemes. Here, several heterogeneous models are trained in parallel, and their forecasts are subsequently aggregated by a meta-regressor (commonly a linear model, XGBoost, or a lightweight neural network). The third class comprises hierarchical hybrids. Different aggregation levels are handled by their models, and then the forecasts are reconciled. At higher levels (category, region), typically, budget and channel consistencies are ensured using interpretable statistical methods. On the other hand, graph networks or recurrent models that can capture local promotional elasticity are used at the SKU-store level.

Hybrid demand-forecasting architectures have evolved from simple cascaded linkages to fully fledged multi-module systems, in which each component is responsible for a clearly defined type of signal. At this level of abstraction, the hybrid serves as a means to partition complexity: linear statistical models extract the coarse trend, recurrent networks capture local temporal anomalies, and graph computations model spatial correlations among points of sale and product groups. Such modularity resolves the tension between interpretability and predictive power, allowing the solution to be described as the sum of narrowly specialized experts rather than as a black box.

The classical ARIMA-LSTM linkage illustrates this idea in its most direct form: ARIMA first isolates the linear trend and seasonality, after which the residual is trained on an LSTM—a model sensitive to nonlinear spikes, promotions, and sudden assortment changes. A practical detail of importance is that both parts can be updated asynchronously, which simplifies operation under conditions of regular retraining.

When inference speed and extensive use of regressors are more important than an exact residual model, it is more convenient to rely on Prophet as a fast base component and complement it with gradient boosting on tabular features. In the study where Prophet was combined with LightGBM, a smoothed trend-seasonality forecast was first constructed, and then boosting distributed the total volume across twenty-eight-day horizons based on price, promotion flags, and calendar indicators. The WRMSSE that resulted decreased to 0.614, compared to 0.767 for the version that was not decomposed, with the improvement remaining constant across all levels of the hierarchy (Zhou, 2023). The primary understanding is that Prophet offers a loosely joined but time-consistent path, adjusting for spikes and troughs by utilizing a rich set of features without requiring strict stationarity.

If the sales organization is decentralized across regions, stores, or online groups, then linear decomposition will no longer suffice: explicit spatial relationships must be modeled. Here, hybrids based on graph convolutional networks are in demand. The Graph CNN–LSTM variant first passes a temporal slice of sales through several GCN layers, aggregating information from adjacent nodes (stores or SKUs linked by substitution), and then feeds the resulting embeddings into an LSTM for time-series forecasting. Verification on electronics data showed that this architecture consistently outperformed four traditional benchmarks in RMSE and WAPE, particularly during volatile demand periods (Niu et al., 2024). A similar concept was scaled up for million-item catalogs in GraphDeepAR/GEANN: adding a graph encoder to a sequence-to-sequence decoder yielded a double-digit relative improvement in CRPS, with particularly strong effects on cold-start SKUs with short sales histories (Kozodoi et al., 2024).

In practice, all of the architectures above are built on a standard technology stack. Data processing and quick prototyping is with pandas/Polars; the statistical layer happens in stats models or pmdarima; deep parts come from PyTorch + PyTorch Forecasting or TensorFlow/Keras; DGL and PyTorch Geometric assist GCNs run. For boosting, one typically selects LightGBM or CatBoost, often in conjunction with Hyperopt/Optuna for Bayesian TPE tuning. Training runs via MLflow Tracking, Airflow DAGs, and Kubernetes Jobs; deployment is using Docker containers wrapped in gRPC/ONNX servers. This type of integrated pipeline enables smooth configuration switching: for instance, saving ARIMA parameters in a metadata store while LSTM and GCN weights are stored in a central S3, updating model parts on separate schedules, and rolling out updates automatically through CI/CD.

Thus, linear–recurrent, boosting, and graph hybrids tackle the same prediction issue, however, at different degrees of data difficulty. The choice of a specific architecture is dictated by the balance between explainability, allowable inference latency, and the necessity to account for spatial distortions. Adhering to the modular principle, one signal type, one expert, remains a universal recommendation, enabling the incremental enhancement of predictive accuracy in line with the expanding informational horizon of a retail network while preserving the operational reliability of models.

Accurate forecasting is impossible without a rich, multichannel data stream that reflects both the retailer’s internal operations and the external context. In retail, POS transactions form the core; they are delivered almost in real-time and processed by modern data marts with a lag of only a few minutes, enabling analysts to capture instantaneous responses to price or promotion (Koppichetti, 2024). At the other end of the frequency spectrum are product master files and daily price lists. Between these lie hourly inventory snapshots, weather and traffic feeds, and media metrics, which are delivered several times per day. In designing the data warehouse, it is crucial to ensure temporal alignment. All events are mapped onto a unified time axis, and missing values are imputed with neutral defaults, ensuring that neural residuals do not learn non-existent demand.

On the cleansed streams, a feature store with offline and online layers is constructed. Physically, this consists of columnar tables annotated with valid from/valid to timestamps, guaranteeing point-in-time correctness during both training and scoring. The extraction of the training dataset is implemented as an idempotent query: the job generates a snapshot for the selected date. Then, it records its checksum, thereby enabling the exact reproduction of the experiment even after intermediate files have been purged. The resulting features are passed to an orchestrator (Kubeflow or Vertex Pipelines), where hyperparameter optimization (Optional / Ray Tune) and cross-validation on temporal folds run in parallel; artifacts and metrics are logged in the MLflow Model Registry.

Model delivery into production follows a classic CI/CD pattern, augmented by a CT—continuous training—loop. A practical scheme, described in Google Cloud’s recommendations, comprises three independent pipelines: code and data validation, automated deployment of the model container, and a retraining trigger activated upon drift detection or on a schedule (Kazmierczak et al., 2023), as illustrated in Figure 2.

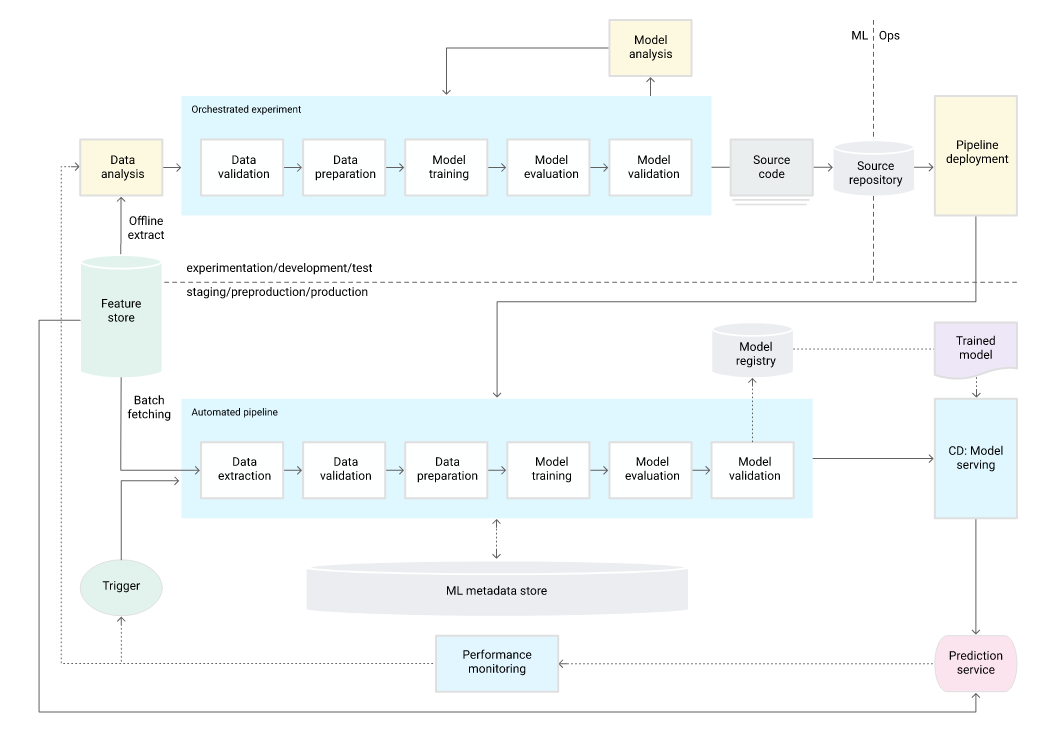


Fig. 2. ML pipeline automation for CT (Kazmierczak et al., 2023)

A model that passes all checks is signed with a data checksum; the container is then deployed in Kubernetes, and a control sample is fed into the monitoring service. Such cyclicality is especially important for hybrids: the statistical module can be retrained monthly, whereas the neural network residual is updated less frequently but with fine-tuning on the latest percentile data to capture fresh promotional shocks.

Finally, monitoring consolidates data, metadata, and predictions into a unified dashboard: latency, feature-missing rates, rolling-window MAPE, and the distribution of predicted values. Alerts on feature-distribution shifts or error spikes beyond confidence intervals trigger the model to switch into shadow mode and activate a fallback layer—typically a simple linear forecast of the most recent period. This MLOps loop closes the logical chain of the preceding sections: it is precisely the continuous cycle data → features → model → monitoring → retraining that renders hybrid schemes not only accurate but also operationally reliable in the dynamic context of real-world retail.

Classical metrics—MAPE, RMSE, and MAE—remain the foundation for evaluating retail forecasts because they provide the business with clear absolute and relative error measures. Nonetheless, they are often sufficient at the candidate-comparison stage to justify adopting hybrids: a recent example is the implementation of a two-stage VAE-ARIMA, where a generative network produces probabilistic scenarios, and ARIMA fine-tunes seasonal fluctuations; as a result, MAPE dropped by 14.3% compared to standard ARIMA on the same retail-sales dataset (Gadam et al., 2024).

To meet inventory management requirements, retailers augment baseline metrics with quantile measures. In practice, the 70th percentile (p70) is most commonly used; the model must ensure that actual demand falls below the forecast in 70% of cases. This quantile is officially supported in the AWS Supply Chain and serves as an SLA target metric for procurement teams (AWS, 2025). In forecasting competitions, hybrid ensembles demonstrate dramatic improvements in aggregate error: in the M5 contest, the winning solution reduced WRMSSE to 0.520, 22.4% better than the strongest statistical benchmark, and one hybrid team’s entry achieved an aggregated-level error of 9 at WRMSSE = 0.614, placing in the top 15 (Makridakis et al., 2022). Such figures provide developers with benchmarks when selecting learning curves and recursion depths.

The choice of validation strategy is determined not only by the forecast horizon but also by the weight of seasonal peaks. Winter holidays in November and December account for approximately 19% of the industry’s annual revenue, so rolling window or back-testing procedures must include at least two such cycles; otherwise, the model will underestimate out-of-stock risks during the year’s highest-value weeks (National Retail Federation, 2024). For cold-start SKUs with limited historical data, it is recommended to combine cross-validation across items and overtime to avoid overfitting noise.

Exogenous features prove their worth when their influence is quantitatively validated. For seasonal products (such as skis and snowboards), incorporating weather time series resulted in a reduction in overall forecast error of up to 45%—equivalent to roughly 2% of annual category sales—an economic benefit sufficient to rapidly recoup the cost of integrating meteorological data (Babongo et al., 2018). In practice, this implies that the feature store should retain not only observed but also forecasted temperature anomalies with granularity at least down to the ZIP code level.

Once deployed in production, interpretability becomes the key to procurement decisions. SHAP decomposition enables managers to receive explanations of each promotion’s or weather indicator’s contribution to a given forecast point, thereby facilitating manual adjustments. These values can be embedded directly in the BI dashboard, updated alongside batch inference, and wired into the CI/CD loop to trigger an alert as soon as the overall distribution of feature importances begins to drift.

Finally, change management. A rapid proof of concept (POC) whose benefits are confirmed by the metrics described shifts the conversation from proving accuracy to restructuring processes. Two agreements are critical here: first, who owns the feature lifecycle so that features are refreshed immediately after any change to the promotion calendar; and second, what degradation threshold triggers automated retraining. Clearly defining these rules transforms a hybrid model from an experimental prototype into a reliable operational planning tool.

The transition from describing hybrid-model quality to outlining practical steps begins with agreeing on what level of accuracy is commercially significant. It is optimal to tie the metric to direct economic impact: McKinsey research shows that implementing AI algorithms reduces average forecasting errors in supply chains by 20–50% and lost sales due to stockouts by up to 65% (Amar, 2022). The global AI in the retail market will grow from USD 11.61 billion in 2024 to USD 40.74 billion by 2030 (CAGR 23%), underscoring that improving forecast accuracy is not merely an academic exercise but part of a rapidly expanding market, as shown in Figure 3 (Grand View Research, 2024). Therefore, when evaluating models, it is useful to record not only MAPE or RMSE but also their monetary equivalents, such as aged-inventory costs, lost revenue, and service-level metrics.

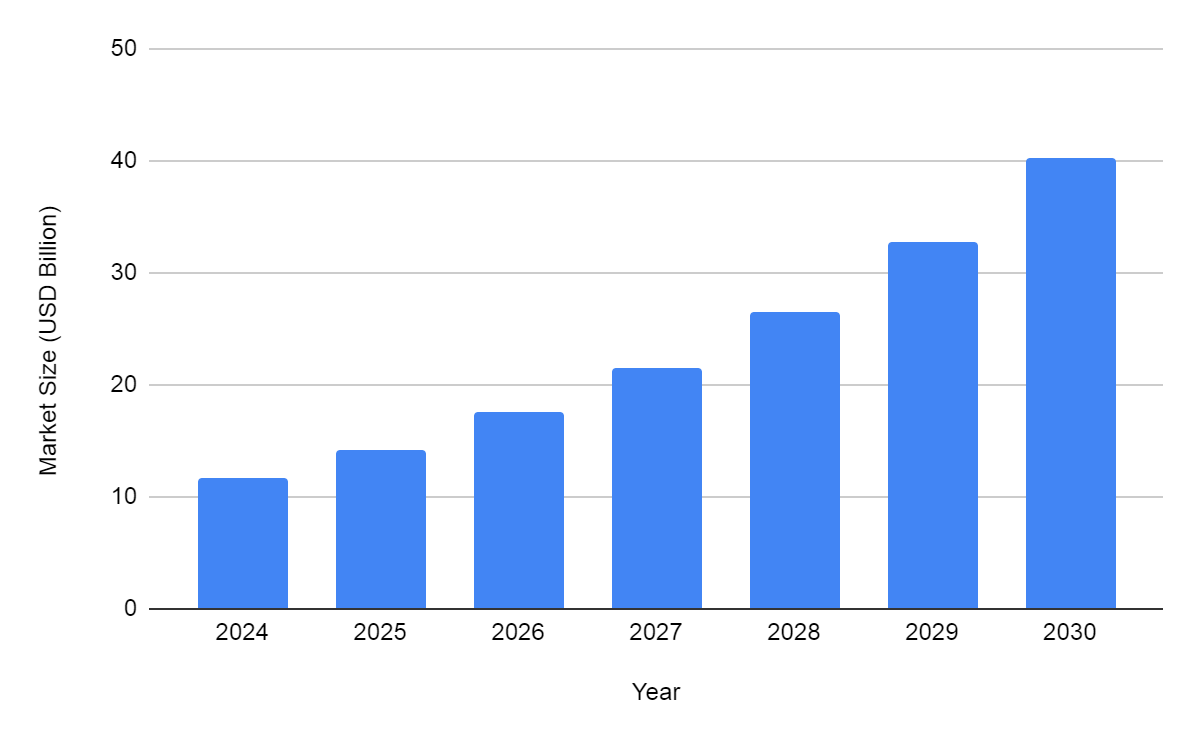


Fig. 3. Artificial Intelligence In Retail Market Size (Grand View Research, 2024)

The following recommendation is to adopt a short prototype-phased ROI cycle. Retailers that have piloted a limited assortment confirm that an eight-week PoC is sufficient to deploy a working hybrid setup and observe the first live improvements. In a similar case of AI forecasting implementation, the company reduced the number of out-of-stock occurrences by 30% while maintaining its standard order infrastructure (Turing, 2025). Such a brief interval demonstrates the model’s value before budgets are exhausted or seasonal processes shift.

The third recommendation concerns expanding the feature set. Laboratory measurements on an FMCG sample revealed that incorporating weather time series can explain up to 47% of the additional sales variance compared to a purely historical model (Chan & Wahab, 2024). Therefore, for each new external factor—whether meteorological data, a sports events calendar, or a competitor price index—it is essential to precisely measure the accuracy gain and maintain feature versions in a feature store.

The fourth point relates to model governance after production deployment. A case study of a global consumer-goods manufacturer demonstrated that integrating automated parameter re-estimation reduced safety stocks by 20 percentage points while simultaneously improving forecast accuracy by six percentage points (Ghandour, 2021). Achieving such results requires a rigorous CI/CD cycle: the static component is retrained more frequently. In contrast, the neural component is retrained less frequently, but always with mandatory data drift checks.

Finally, a careful retraining schedule conserves resources. An experiment on two large retail datasets found that infrequent, periodic updates preserve the global model’s accuracy while reducing computational costs and carbon footprint (Zanotti, 2023). This confirms that not every new observation should immediately trigger a full retraining; instead, thresholds must be set so that quality gains justify the expense.

In summary, the recommendations are as follows: measure effects in monetary terms; launch compact PoC cycles; incrementally incorporate exogenous factors, automate quality control; and learn to retrain models less frequently but more thoughtfully. This sequence ensures that hybrid forecasts are not only theoretically accurate but also robust within the live operational environment of a retail business.

**Conclusion**

The review demonstrates that hybrid retail models estimate demand by combining the best attributes of traditional statistics with new machine learning techniques. This provides understanding and the skill to notice more complex connections. Statistical models, such as ARIMA or Prophet, draw out baseline movements along with seasonality, making it easier for operational managers to identify obvious reasons for contributions from pricing and calendar factors. The nonlinear portion—supplied via LSTM, gradient boosting, or graph convolutional networks—assists in modeling abrupt, large increments due to promotions and interactions among various channels and nearby selling spots related to specific product categories. This division of labor yields double-digit reductions in MAPE, RMSE, and WRMSSE compared with standalone models and also reduces forecast-error variance on cold SKUs thanks to the linear module’s more stable performance on short time series.

A key advantage of hybrid approaches is their modularity. Each component can be updated independently, simplifying system maintenance and scalability in the face of constantly changing external factors and the emergence of new regressors. Built-in continuous integration and continuous training (CI/CD and CT) loops permit rapid responses to data drift and preserve forecast confidence intervals within predefined SLAs. Model-quality monitoring—via latency metrics, feature-dropout checks, rolling MAPE, and error-distribution control—ensures timely activation of fallback layers and automatic retraining only when a justified accuracy degradation occurs.

The real use of hybrid systems — based on having one set of tools, starting with getting and matching time series data in Pandas or Polars, moving to statistical libraries like StatsModels and Prophet, then to deep-learning frameworks such as PyTorch or TensorFlow, followed by specialized graph-computation libraries like DGL and PyTorch Geometric. MLflow, Airflow, and Kubernetes handle training and organizing hyperparameters , making sure experiments can be repeated and the platform can scale. This approach supports cascading residual hybrids, stacking/blending ensembles, and hierarchical schemes, delivering forecast accuracy at levels of granularity ranging from category and region to SKU store.

Beyond technical aspects, the successful deployment of hybrid models requires rigorous change management, including defining feature lifecycle owners, setting degradation threshold agreements, and establishing automatic retraining protocols. Brief PoC cycles of approximately eight weeks and linking forecast-accuracy metrics to economic indicators (reduction of expired stock, improvement in service level, decrease in lost revenue) enable rapid demonstration of business impact and facilitate the transition from experimental phases to regular operational practice.

Thus, hybrid models represent the optimal solution for modern retail, as they combine interpretability, modularity, and high predictive power. They not only enhance forecast accuracy but also ensure operational reliability in dynamic omnichannel sales environments, making the predictive analytics tool an effective lever for increasing profitability and reducing costs.

**COMPETING INTERESTS DISCLAIMER:**

Authors have declared that they have no known competing financial interests OR non-financial interests OR personal relationships that could have appeared to influence the work reported in this paper.

**References**

Aktas, M. Y., Ji, T., & Lu, C.-T. (2024). Time Series Forecasting with GCN-LSTM Based Unified Model for Product Demand Prediction. *2024 IEEE International Conference on Big Data (BigData)*, 5901–5906.<https://doi.org/10.1109/bigdata62323.2024.10825969>

Amar, J. (2022, February 15). *AI-driven Operations Forecasting in Data-Light Environments*. McKinsey & Company.<https://www.mckinsey.com/capabilities/operations/our-insights/ai-driven-operations-forecasting-in-data-light-environments>

AWS. (2025). *AWS Supply Chain*. Amazon.<https://docs.aws.amazon.com/aws-supply-chain/latest/userguide/forecast-forecast-entity.html>

Babongo, F., Appelqvist, P., Demoulin, V. C., Hameri, A. P., & Niemi, T. (2018). Using weather data to improve demand forecasting for seasonal products. *International Journal of Services and Operations Management*, *31*(1), 53.<https://doi.org/10.1504/ijsom.2018.094183>

Chan, H., & Wahab, M. I. M. (2024). A Machine Learning Framework for Predicting Weather Impact on Retail Sales. *Supply Chain Analytics*, 100058–100058.<https://doi.org/10.1016/j.sca.2024.100058>

Gadam, H., Upadhyay, A., & Panda, S. (2024). Leveraging Generative AI for Real-Time Financial Forecasting Accuracy in Cloud ERP Environments. *Original Research Paper International Journal of Intelligent Systems and Applications in Engineering IJISAE*, *2024*(21s).<https://ijisae.org/index.php/IJISAE/article/download/7496/6509/12841>

Ghandour, J. (2021, November 24). Consumer goods companies must transform their planning from end to end. *McKinsey & Company*.<https://www.mckinsey.com/industries/consumer-packaged-goods/our-insights/consumer-goods-companies-must-transform-their-planning-end-to-end>

Grand View Research. (2024). *Global AI In Retail Market Size & Share Report*. Grand View Research.<https://www.grandviewresearch.com/industry-analysis/ai-retail-market-report>

Kazmierczak, J., Salama, K., & Huerta, V. (2023, May 18). *MLOps: Continuous delivery and automation pipelines in machine learning*. Google Cloud.<https://cloud.google.com/architecture/mlops-continuous-delivery-and-automation-pipelines-in-machine-learning>

Koppichetti, R. K. (2024). Real-Time Data Processing for Retail Insights. *International Journal of Multidisciplinary Research and Growth Evaluation*, *5*(4), 1378–1386.<https://doi.org/10.54660/.ijmrge.2024.5.4.1378-1386>

Kozodoi, N., Zinovyeva, E., Valentin, S., Pereira, J., & Agundez, R. (2024). Probabilistic Demand Forecasting with Graph Neural Networks. *ArXiv*.<https://doi.org/10.48550/arxiv.2401.13096>

Makridakis, S., Spiliotis, E., & Assimakopoulos, V. (2022). M5 accuracy competition: Results, findings, and conclusions. *International Journal of Forecasting*, *38*(4).<https://doi.org/10.1016/j.ijforecast.2021.11.013>

National Retail Federation. (2024). *Winter Holiday FAQs*. National Retail Federation.<https://nrf.com/research-insights/holiday-data-and-trends/winter-holidays/winter-holiday-faqs>

Niu, T., Zhang, H., Yan, X., & Miao, Q. (2024). Intricate Supply Chain Demand Forecasting Based on Graph Convolution Network. *Sustainability*, *16*(21), 9608–9608.<https://doi.org/10.3390/su16219608>

Obi, C. (2024). Demand Forecasting in Retail Business Using the Ensemble Machine Learning Framework - A Stacking Approach. *American Scientific Research Journal for Engineering, Technology, and Sciences*, *98*(1), 309–329.<https://www.researchgate.net/publication/384869869_Demand_Forecasting_in_Retail_Business_Using_the_Ensemble_Machine_Learning_Framework_-_A_Stacking_Approach>

Oca, A., Panikkar, R., Sampat, C., & Brown, T. (2024, November 15). *Harnessing the power of AI in distribution operations*. McKinsey & Company.<https://www.mckinsey.com/industries/industrials-and-electronics/our-insights/distribution-blog/harnessing-the-power-of-ai-in-distribution-operations>

Research and Markets. (2025). *Demand Planning Solutions Market Size & Forecast*. Research and Markets.<https://www.researchandmarkets.com/report/demand-planning-solution?srsltid=AfmBOoopSpo1liuNy97s79KQTSn_qteoNfq-Sl3Vk74DCjAY1_p1L0e1>

Suddala, S. (2024). Dynamic Demand Forecasting In Supply Chains Using Hybrid ARIMA-LSTM Architectures. *International Journal of Advanced Research*, *12*(10), 1167–1171.<https://doi.org/10.21474/ijar01/19738>

Turing. (2025). *30% Fewer Stockouts: AI-Powered Demand Forecasting*. Turing.<https://www.turing.com/case-study/ai-demand-forecasting-solutions>

Zanotti, M. (2023). *Do global forecasting models require frequent retraining?* Arxiv.<https://arxiv.org/html/2505.00356v1>

Zhou, T. (2023, May 26). *Improved Sales Forecasting using Trend and Seasonality Decomposition with LightGBM*. ArXiv.<https://doi.org/10.48550/arXiv.2305.17201>