Journal of Contemporary Science and Technology Management Volume 02, Issue 03, September 2026

Machine Learning-Based Demand Forecasting and Business Optimization Using SAP S/4HANA and Azure AI

Amit Jaiswal
Net4India, NOIDA, UP, India
Email ID: Amit.jaiswalgkp1978@gmail.com

Abstract

Accurate demand forecasts are the starting point for almost every operational decision an enterprise makes, from purchasing and production to staffing, pricing, and cash planning. Yet many organizations still forecast with simple statistical methods and spreadsheets that cannot absorb the volume, variety, and volatility of modern demand. This article proposes a framework that uses machine learning on Microsoft Azure to forecast demand from SAP S/4HANA data and then converts those forecasts into optimized business decisions executed back in S/4HANA. The framework covers the data foundation, feature engineering, forecasting models, optimization of inventory, production, and pricing, a generative AI planning assistant, and an operations plane that keeps the platform reliable. Drawing on research into intelligent enterprise frameworks, large language model (LLM) integration with SAP, prompt engineering, LLM guardrails, prompt-injection security, cloud anomaly detection, and autonomous operations agents, the article describes each component, proposes evaluation metrics, and outlines a phased implementation roadmap.

DOI: https://doi.org/10.66592/40wt7e55

I. INTRODUCTION

Every supply chain decision is ultimately a bet on future demand. Purchasing teams order raw materials weeks or months ahead; production planners commit capacity; distribution teams position inventory across warehouses; finance teams project revenue and cash. When forecasts are wrong, the costs appear everywhere at once: excess inventory ties up working capital and risks obsolescence, while shortages lose sales, disappoint customers, and trigger expensive expediting.

Traditional forecasting approaches struggle with today's conditions. Product portfolios are

larger, life cycles are shorter, sales channels are more numerous, and demand reacts quickly to promotions, competitor actions, weather, and economic shifts. Statistical methods such as moving averages and exponential smoothing remain useful baselines, but they typically model each product's history in isolation and cannot easily incorporate the many external drivers that influence demand.

Machine learning offers a stronger alternative. Models can learn from thousands of related products simultaneously, incorporate promotions, prices, holidays, and external signals, and capture non-linear relationships that classical methods miss. Cloud platforms make it practical to train these models on complete transaction histories and to retrain them frequently.

A forecast, however, is not a decision. Its value is realized only when it changes how much to buy, make, stock, or charge. This article therefore treats demand forecasting and business optimization as a single, connected capability. SAP S/4HANA supplies the transactional history and executes the resulting plans. Microsoft Azure provides data processing, machine learning, and generative AI services that produce forecasts and recommendations.

The framework builds on prior research. Asuri (2022) describes an intelligent enterprise framework in which S/4HANA and SAP Analytics Cloud (SAC) deliver real-time decision intelligence, and Asuri (2023) shows how LLMs can be integrated with SAP Business Technology Platform (BTP) for intelligent automation. Sannidhanam (2024a, 2024b) addresses guardrails and prompt engineering for dependable LLM applications, Lakavath (2026b) examines prompt-injection threats, and Lakavath (2026a) and Sannidhanam (2025) address intelligent monitoring and autonomous operations on the cloud.

2. THE FORECASTING CHALLENGE IN ERP ENVIRONMENTS

Demand forecasting in an ERP context has several characteristics that shape the design of any solution.

Scale and hierarchy. A manufacturer or retailer may need forecasts for tens of thousands of products across hundreds of locations and several customer channels, at daily or weekly granularity. Forecasts must also be consistent across hierarchy levels, so that store-level forecasts add up to regional totals used by finance.

Intermittent and new demand. Many products sell sporadically, with long runs of zero demand punctuated by occasional orders. New products have little or no history at all. Both cases defeat methods that rely on smooth, long histories.

Demand versus sales. ERP systems record sales, not true demand. When products were out of stock, recorded sales understate what customers wanted. Forecasting from uncorrected sales history perpetuates past shortages.

Many drivers. Price changes, promotions, marketing campaigns, holidays, weather, competitor activity, and macroeconomic conditions all influence demand. Their effects interact and vary by product and region.

Decision coupling. Forecasts feed material requirements planning (MRP), production scheduling, replenishment, and financial planning. A forecast that is accurate on average but biased in specific product groups can still cause systematic shortages or overstock in those groups.

These characteristics point toward an approach that learns across many related series, incorporates external drivers, quantifies uncertainty, and integrates tightly with the planning processes in S/4HANA.

3. FRAMEWORK OVERVIEW

The framework links S/4HANA and Azure in a closed loop from data to decision and back.

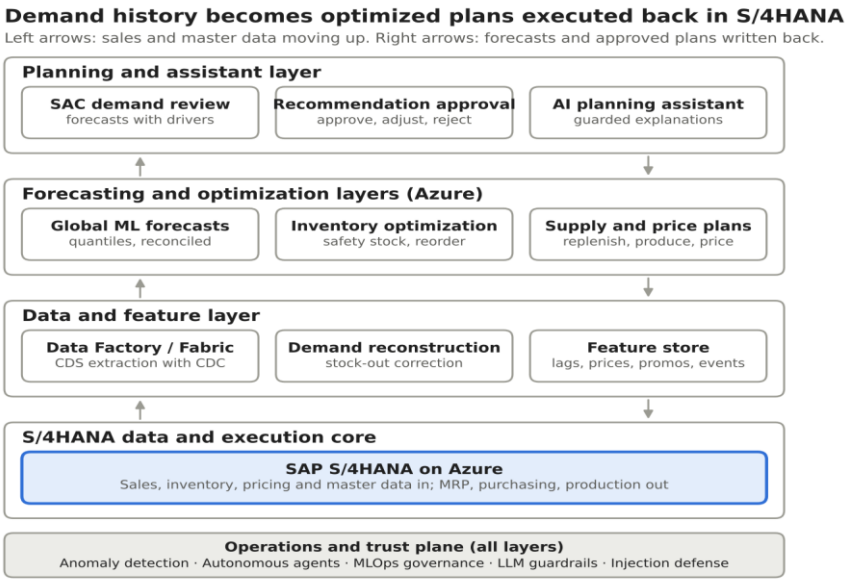

Figure 1. Machine learning-based demand forecasting and business optimization framework using SAP S/4HANA and Azure AI

Sales history and master data flow from S/4HANA into Azure, where forecasts and optimized

plans are produced and then written back to SAC and S/4HANA for approval and execution.

The framework has five layers:

1. S/4HANA data and execution core. S/4HANA provides sales orders, deliveries, inventory, purchasing, production, pricing, and master data, and it executes the final plans through MRP, purchasing, and production orders.
2. Data and feature layer. Azure pipelines extract S/4HANA data, combine it with external signals, correct for stock-outs, and produce reusable forecasting features.
3. Forecasting layer. Azure Machine Learning trains and serves demand models that produce point and probabilistic forecasts at multiple hierarchy levels.
4. Optimization layer. Optimization models turn forecasts into recommended inventory targets, replenishment quantities, production plans, and pricing actions.
5. Planning and assistant layer. Planners review forecasts and recommendations in SAC, supported by a generative AI assistant that explains them in natural language.

An operation and trust plane spans all layers, providing monitoring, autonomous remediation, governance, and AI safety controls.

4. BUILDING THE DATA FOUNDATION FROM S/4HANA

S/4HANA is an excellent source for forecasting data because its in-memory architecture and simplified data model provide current, consistent records. Asuri (2022) highlights the value of real-time decision intelligence built on S/4HANA, in which analytics operate on live transactional data rather than stale extracts. For forecasting, this means models can incorporate yesterday's orders and today's inventory position rather than last month's.

Data extraction. Pipelines built with Azure Data Factory or Microsoft Fabric extract relevant Core Data Services (CDS) views covering sales order items, deliveries, billing documents, stock levels, pricing conditions, promotions, and product, customer, and plant master data. Change data capture moves only new or updated records after the initial load.

Cleansing and demand reconstruction. Raw sales history is cleansed of returns, cancellations, and one-off bulk orders that would distort patterns. Periods when products were out of stock are identified from inventory records, and true demand in those periods is estimated, so models learn what customers wanted rather than what was available.

External signals. Internal data is joined with calendars of public holidays and school terms,

weather observations and forecasts, marketing calendars, competitor pricing where available, and macroeconomic indicators. Each signal is aligned to the product, location, and time granularity of the forecast.

Feature engineering. The data layer computes features that capture demand dynamics: lagged sales and rolling averages, seasonality indicators, price relative to recent averages, promotion flags and depth, product life-cycle stage, and similarity to other products. Features are stored in a governed feature store so that training and scoring use identical definitions.

5. MACHINE LEARNING FORECASTING MODELS

The forecasting layer uses Azure Machine Learning to develop, train, deploy, and monitor demand models.

Global models across many series. Rather than fitting a separate model to each product and location, the framework favors global models that learn from all series at once. Gradient-boosted decision trees and deep learning architectures for time series can share patterns across related products, which is especially valuable for intermittent and new items with little history of their own. Classical statistical models remain in the portfolio as baselines and for stable, high-volume items where they perform well.

Probabilistic forecasts. Optimization decisions depend on uncertainty, not just expected values. A product with an expected weekly demand of 100 units but a plausible range of 60 to 180 requires different safety stock than one ranging from 95 to 105. The framework therefore produces quantile forecasts, such as the 10th, 50th, and 90th percentiles, alongside point forecasts.

Hierarchical reconciliation. Forecasts are generated at the most useful granularity, typically product by location by week, and then reconciled so that they add up consistently to product family, region, and total company levels used in sales and operations planning and financial plans.

New product forecasting. For products with no history, models rely on attributes such as category, price point, and launch channel, and on the observed launch curves of similar past products. As actual sales arrive, the forecast quickly shifts weight toward the product’s own data.

Explainability. Planners need to understand why the forecast changed. The framework attaches driver explanations to each forecast, showing, for example, that a 20 percent increase is driven

mainly by planned promotion and a public holiday. Explanations are generated with Azure Machine Learning's responsible AI tools and displayed beside the numbers in SAC.

MLOps. Models are versioned, registered with their training data and performance metrics, and deployed through automated pipelines. Batch scoring produces forecasts on a schedule, typically weekly or daily, while monitoring compares forecasts with actual demand and tracks drift in input data. Retraining is triggered when accuracy deteriorates beyond agreed thresholds.

6. FROM FORECASTS TO BUSINESS OPTIMIZATION

The optimization layer converts forecasts into recommended decisions. Several decision areas benefit directly.

Inventory optimization. Safety stock protects against forecast error and supply variability. A common formulation sets safety stock in proportion to the variability of demand over the replenishment lead time:

$SS = z \times \sigma d \times \sqrt{L}$

where SS is safety stock, z reflects the target service level, σd is the standard deviation of demand per period, and L is the lead time in periods. Probabilistic forecasts improve on this formula by estimating uncertainty directly for each product and location, allowing safety stock to be set precisely where it is most needed and reduced where demand is predictable.

Replenishment and purchasing. Using forecasts, current stock, open orders, and lead times, the framework recommends order quantities and timing for each item and supplier, respecting minimum order quantities, pack sizes, and budget limits. Approved recommendations become purchase requisitions or stock transport orders in S/4HANA.

Production planning. Forecasts feed S/4HANA MRP and production planning, while optimization models balance capacity, changeover costs, and inventory carrying costs to propose production sequences and lot sizes.

Pricing and promotion planning. Demand models that capture price elasticity allow planners to simulate the effect of price changes and promotions on volume, revenue, and margin, and to choose actions that meet commercial objectives without creating supply shortfalls.

Financial planning. Reconciled forecasts flow into SAC financial plans, aligning revenue, cost, and cash projections with the same demand view used by operations.

The table below summarizes the main optimization levers.

Decision area	Forecast input	Optimization objective	S/4HANA execution
Inventory	Quantile demand forecasts by location	Meet service targets at minimum inventory	Safety stock and reorder point parameters
Replenishment	Forecasts, stock, open orders, lead times	Minimize stock-outs and ordering costs	Purchase requisitions and stock transfers
Production	Reconciled forecasts by plant	Balance capacity, changeovers, and inventory	Planned and production orders
Pricing and promotions	Price-elastic demand forecasts	Maximize margin within supply limits	Pricing conditions and promotion records
Financial planning	Forecasts aggregated to company level	Align revenue, cost, and cash plans	Plan data shared with finance

Human review and closed loop. Recommendations are presented in SAC with their expected impact on service levels, inventory value, and margin. Planners approve, adjust, or reject them, and approved actions are executed in S/4HANA through governed workflows. Actual outcomes flow back into the data layer to improve both forecasts and optimization parameters, completing the decision loop that Asuri (2022) identifies as central to sustainable business value.

7. A GENERATIVE AI PLANNING ASSISTANT

Planners spend much of their time investigating why forecasts changed, explaining plans to colleagues, and preparing for planning meetings. A generative AI assistant can reduce that effort.

Asuri (2023) presents a strategic framework for integrating LLMs with SAP BTP to enable intelligent automation and operational excellence. Following that model, the assistant runs on Azure OpenAI Service and accesses forecasts, explanations, and S/4HANA data through BTP, within each planner's authorizations. Planners can ask why the forecast for a product family rose, which items are at highest risk of stock-out next month, or what would happen to inventory if a supplier's lead time doubled. The assistant can also draft summaries for sales and operations planning meetings.

Because planning decisions are consequential, the assistant must be reliable, safe, and secure.

Reliable outputs. Sannidhanam (2024b) examines prompt engineering patterns for reliable LLM outputs in production environments. The assistant uses structured, versioned prompts and

explicit output formats, and every number it cites is retrieved from the forecasting system rather than generated by the model. Responses that propose changes, such as adjusted order quantities, are validated against business rules before being shown.

Guardrails. Sannidhanam (2024a) describes guardrails and safety mechanisms for LLM-powered enterprise applications. The assistant applies input and output filters, restricts answers to approved data sources, declines questions outside its scope, and escalates uncertain cases to human experts.

Prompt-injection defense. Lakavath (2026b) analyzes prompt injection threats in Azure-based LLM applications, where hidden instructions in user input or retrieved content try to manipulate a model. Planning assistants may read customer comments, supplier emails, or promotion descriptions that could contain such instructions. The framework isolates retrieved content from system instructions, grants the assistant read-only access by default, and requires explicit human approval before any plan changes are executed in S/4HANA.

8. OPERATING THE FORECASTING PLATFORM

A forecasting and optimization platform runs on a schedule that the business depends on. If a data pipeline fails before the weekly forecast run, planners may unknowingly work with outdated forecasts. Reliable operations are therefore essential.

Anomaly detection. Lakavath (2026a) shows how machine learning can detect anomalies in Azure cloud infrastructure by learning normal patterns in metrics and logs and flagging deviations that static thresholds miss. The framework applies this to pipeline runtimes, data volumes, data freshness, scoring job durations, and endpoint latency. The same idea applies to forecasts themselves: sudden, unexplained shifts in forecast distributions are flagged for review, since they often signal data problems rather than real demand changes.

Autonomous operations agents. Sannidhanam (2025) describes autonomous AI agents for cloud infrastructure operations that observe system state, diagnose issues, and take corrective action. In this framework, agents rerun failed extraction jobs, scale compute for large training runs, trigger retraining when drift is detected, and notify planners when forecasts are delayed. Agents operate within defined policies and least-privilege permissions, and every action is logged.

Governance. Each forecast is traceable to its model version, features, and source data in S/4HANA. Access to forecasts and plans follows the same role-based controls as the underlying

ERP data, enforced through centralized identity management.

9. EVALUATING PERFORMANCE

Forecasting and optimization should be judged by both statistical accuracy and business outcomes.

Category	Metric	What it shows
Forecast accuracy	Weighted mean absolute percentage error (WMAPE)	Overall accuracy, weighted toward high-volume items
Forecast accuracy	Bias by product group	Systematic over- or under-forecasting
Forecast value	Forecast value added versus baseline	Whether machine learning beats simpler methods
Service	Fill rate and stock-out frequency	Customer impact of plans
Inventory	Inventory value, turns, and obsolescence	Working capital efficiency
Adoption	Share of recommendations accepted; override frequency	Planner trust in the system
Operations	Forecast run on time; data freshness	Platform reliability

Baselines should be measured before deployment, and results reported by product segment, because improvements are rarely uniform across fast movers, slow movers, and new products.

10. IMPLEMENTATION ROADMAP

1. Prepare data and baselines. Establish real-time reporting on S/4HANA and SAC (Asuri, 2022), build pipelines to Azure, reconstruct true demand, and measure current forecast accuracy and inventory performance.
2. Pilot machine learning forecasts. Select a product segment with clear business value, train global models in Azure Machine Learning, and run them alongside existing forecasts to measure forecast value added.
3. Connect to optimization and execution. Use probabilistic forecasts to set safety stocks and recommend replenishment, with planner approval in SAC and execution in S/4HANA.
4. Industrialize operations. Automate retraining and monitoring, apply anomaly detection (Lakavath, 2026a), and introduce autonomous operations agents (Sannidhanam, 2025).
5. Add the planning assistant. Deploy the generative AI assistant through BTP (Asuri, 2023), with reliable prompts (Sannidhanam, 2024b), guardrails (Sannidhanam, 2024a), and

prompt-injection defenses (Lakavath, 2026b).

6. Scale. Extend to further product segments, production planning, pricing, and financial planning.

11. CHALLENGES

Data quality remains the most common barrier, particularly inconsistent product hierarchies and missing stock-out records. Demand shocks, such as pandemics or sudden supply disruptions, can invalidate learned patterns, so human judgment must remain central during unusual periods. Planner adoption depends on transparency, visible accuracy improvements, and the ability to override recommendations. Integration with existing S/4HANA planning parameters must be managed carefully to avoid conflicting signals. Finally, compute and AI costs must be monitored so that the value of better forecasts exceeds the cost of producing them.

12. CONCLUSION

Machine learning has made it possible to forecast demand with greater accuracy, granularity, and awareness of uncertainty than traditional methods allow. The framework presented here connects that capability to business outcomes by combining SAP S/4HANA's transactional data and execution with Azure's machine learning, optimization, and generative AI services. Forecasts become optimized inventory, replenishment, production, and pricing decisions, which planners review in SAC and execute in S/4HANA, while outcomes continuously improve the models.

The research drawn on throughout this article shows how each element can be realized: real-time decision intelligence on S/4HANA and SAC (Asuri, 2022), LLM integration through SAP BTP (Asuri, 2023), reliable and safe generative AI (Sannidhanam, 2024a, 2024b; Lakavath, 2026b), and intelligent, self-managing cloud operations (Lakavath, 2026a; Sannidhanam, 2025). Organizations that adopt the framework step by step, measure results honestly, and keep planners accountable for final decisions can turn demand forecasting from a periodic exercise into a continuous source of competitive advantage.

REFERENCES

1. Asuri, R. (2022). Leveraging SAP S/4HANA and SAP Analytics Cloud to build an intelligent enterprise framework for real-time decision intelligence, digital innovation, and

sustainable business value creation. International Journal of Science, Technology & Society, 6(1), 85–100.

2. Asuri, R. (2023). A strategic enterprise framework for integrating large language models with SAP Business Technology Platform to enable intelligent automation and operational excellence. International Journal of Humanities and Information Technology, 5(01), 87–103. https://doi.org/10.21590/ijhit.05.01.7
3. Lakavath, S. R. (2026a). Anomaly detection in Azure cloud infrastructure using machine learning. Stout in Computer Science and Technology Studies, 2(1), 109–116. https://doi.org/10.61424/f9w38090
4. Lakavath, S. R. (2026b). Prompt injection threats in Azure-based large language model applications. Journal of Computer Science and Information Technology, 3(2), 91–98. https://doi.org/10.61424/jcsit.v3i2.1077
5. Sannidhanam, A. H. (2024a). Guardrails and safety mechanisms for LLM-powered enterprise applications. International Journal of Engineering Science & Humanities, 14(3), 273–285. https://www.ijesh.com/j/article/view/933
6. Sannidhanam, A. H. (2024b). Prompt engineering patterns for reliable LLM outputs in production environments. Journal of Multidisciplinary Knowledge, 4(1), 29–39.
7. Sannidhanam, A. H. (2025). Autonomous AI agents for cloud infrastructure operations. Journal of Contemporary Science and Technology Management, 1(01), 83–100. https://jcstm.info/index.php/jcstm/article/view/20