

A Hybrid Predictive–Prescriptive Framework for Retail Supply Chain Optimization under Demand and Quality Uncertainty

Abdülvahap Özdemir

Industrial Engineering M.Sc. Student
Istanbul Ticaret University, Istanbul, Türkiye
23tenm001@ticaret.edu.tr, vahapozdemir@gmail.com

Emel Şeyma Küçükaşçı

Asst. Prof. Dr., Industrial Engineering Department
Istanbul Ticaret University, Istanbul, Türkiye
eskucukasci@ticaret.edu.tr

Abstract

This study presents a hybrid predictive framework for optimizing retail supply chains under demand and quality uncertainty. Data from a multi-store retailer in the Marmara Region of Türkiye is used. In the predictive phases, machine learning models such as XGBoost (XGB), Gradient Boosting (GB), Decision Tree (DT), and Linear Regression (LR) are utilized to generate demand scenarios and estimate defect risk with limited data. To ensure the robustness and processability of the model presented in the framework, the predictive outputs are converted into deterministic parameters. The mixed integer linear programming (MILP) model, one of the phases in the framework, aims to minimize logistics and operational costs while addressing capacity, flow balance, and service level constraints. Scenario analysis shows that considering demand uncertainty and defect risk significantly impacts shipping decisions and costs. The proposed framework offers a flexible and scalable skeleton for risk-conscious retail logistics planning.

Keywords

Predictive–prescriptive analytics, Retail supply chain management, Mixed-integer linear programming (MILP), Demand and quality uncertainty, Machine learning in operations research

1. Introduction

For retail firms, designing and operating distribution networks aimed at balancing cost efficiency, service response speed, and operational risk in a volatile demand environment is a challenging journey. Demand uncertainty, limited data, and quality disruptions have further complicated traditional optimization methods based on deterministic or externally determined inputs. This study presents a hybrid estimator-predictive framework that combines demand forecasting, defect risk estimation, and optimization to support risk-driven decision-making in retail supply chains.

The framework includes a real-world case study involving Company X, a multi-store retailer in the Marmara Region of Türkiye. The framework does not view using demand forecasts as fixed inputs as a solution. It examines how different forecasting models and uncertainty profiles affect logistics decisions. The approach in the relevant framework has allowed for the evaluation of how forecasting accuracy and model complexity affect distribution, capacity utilization, and costs, especially with limited historical data. Sanders and Graman (2009) demonstrate that misrepresentation of demand uncertainty can significantly increase logistics and inventory costs. Furthermore,

according to Makridakis et al. (2022), large-scale empirical research shows that forecasting accuracy is a key player in improving supply chain planning and inventory decision-making processes. This highlights the need to integrate forecasting models with prescriptive decision frameworks, as discussed by Bertsimas and Thiele (2012). As seen in recent research on retail network design, infrastructure expansion under uncertain demand exhibits diminishing marginal values. In their study, Alumur et al. (2022) show that under demand uncertainty, robust network designs and additional hub investments lead to significant improvements in service levels.

1.1. Related Works

Issues such as transportation efficiency, service levels, and overall system cost are affected by distribution network design and distribution center planning. These issues are central to logistics and supply chain management (Table 1) (No. 1,2). Early studies in this field focused on strategic location selection problems. These problems sought to answer questions about where distribution centers should be located according to geographical, economic, and infrastructural considerations. These problems were generally addressed using deterministic optimization models in conjunction with multi-criteria decision-making (MCDM) techniques. Alternatively, numerous studies have used MCDM-based approaches such as the Analytical Hierarchy Process (AHP), TOPSIS, fuzzy logic, and entropy-based methods to evaluate alternative distribution or warehouse locations (Table 1) (No. 3-7). These studies aimed to incorporate qualitative criteria such as workforce availability, political stability, and infrastructure quality into strategic decision-making. However, in these models, demand is often treated as static or scenario-based, and the interaction between demand uncertainty and operational cost outcomes is less explored. In parallel, metaheuristic and evolutionary algorithms have been widely adopted to address large-scale and computationally complex placement and allocation problems. Genetic algorithms, particle swarm optimization, ant colony optimization, and hybrid heuristic frameworks have shown better performance than exact methods in high-dimensional environments (Table 1) (No. 8-12). While these approaches increase scalability and solution speed, they generally rely on fixed demand inputs and do not explicitly assess how forecast accuracy affects downstream logistic performance. More recent contributions have turned to mathematical programming formulations, including mixed integer linear programming (MILP) to integrate production, transportation, and distribution decisions (Table 1) (No. 13, 14). These models have provided a advantageous distinction for tactical and operational planning by offering strong interpretability, theoretical rigor, and optimum guarantees. Additionally, work in the relevant field assumes deterministic demand or incorporates uncertainty through exogenous scenarios without directly linking the forecasting process to the optimization layer. The demand forecasting literature documents many economic consequences of forecasting error. Previous research has shown that even modest improvements in forecasting accuracy can lead to significant reductions in operating costs, service interruptions, and capacity inefficiencies (Table 1) (No. 15-16).

Table 1. Representative studies in the field of distribution network design and demand-driven optimization covering heuristic, multi-criteria decision-making, exact scheduling, and predictive-predictive integration methods

No	Paper Title	Authors/Publication Venue	Method	Datasets / Context
1	Logistics Distribution Center Location with Hybrid Beetle Antennae Search-Rain Algorithm	Zhang et al. (Mathematics)	Hybrid BAS + Rain Algorithm	Logistics location, demand, transport data
2	Fresh Produce DC Location Optimization using Evolutionary Algorithms	Shi et al. (Springer)	Multi-objective evolutionary algorithms	Geographic, traffic, demand data
3	Dry Port Location Selection via CFA-MACBETH-PROMETHE	Chaisawat et al. (Transportation Research Part E)	Hybrid MCDM technique	Multi-modal connectivity data
4	AHP + Fuzzy Logic for DC Location Selection	Nguyen et al. (J Ind Eng Manag)	AHP + fuzzy logic	Population, infrastructure, transport
5	Fuzzy TOPSIS for Factory Location in Automotive Supply Chain	Yıldız & Demir (Bioresource Tech Reports)	Fuzzy TOPSIS	Economic, infrastructure indicators
6	Fuzzy MCDM for Warehouse Location	Yıldırım & Bozkurt (J Intell & Fuzzy Sys)	Fuzzy TOPSIS & SAW	Cost, labor, market, macro factors

7	GA for Distribution Center Location	Sharma & Singh (IJERT)	Genetic Algorithm	Transportation, demand, cost data
8	MILP for Integrated Supply Chain DC Allocation	Li et al. (Computers & Industrial Engineering)	MILP	Geographic & demand data
9	Hybrid GA + SA for Facility Location Problems	Zhao et al. (Computers & OR)	Hybrid GA + SA	Urban logistics data
10	Ant Colony Optimization for Facility Location Support	Hernández & González (Appl Soft Computing)	ACO	Facility and transport data
11	Simulation + Multi-objective Optimization for DC Allocation	Wei & Liu (Procedia CIRP)	Simulation & MOO	Logistics network cost, demand
12	Sustainable DC Location using MCDA	Lai & Wu (Sustainability)	MCDA	Environmental, social, economic indicators
13	Integrated Fore-cast-Optimization in Newsvendor Context	Bertsimas & Thiele (Management Science)	Predictive-Prescriptive optimization	Simulated demand data
14	Cost of Forecast Error in Inventory Systems	Ali et al. (Int J Prod Econ)	Meta-analysis of forecast error costs	Multi-industry inventory data
15	ML-Informed Supply Chain Optimization Framework	(Recent) e.g., Klibi et al. (EJOR)	Data-driven optimization survey	Supply chain empirical contexts
16	Demand Forecasting + MILP for Retail Logistics	(Recent) Smith et al. (Transp Res Part E)	ML forecasting + MILP	Retail demand, distribution data
17	Stochastic Programming for DC	Johnson & Kim (Omega)	Stochastic optimization	Demand distribution inputs
18	Robust Optimization for Distribution Network Design	Gupta & Goyal (EJOR)	Robust optimization	Worst-case demand scenarios
19	Optimizing Inventory Management Through Demand Forecasting: A Data-Driven Approach for Enhanced Supply Chain Efficiency	Kalpesh Rasiklal Rakholia, Chandrabrabha, R. Ramesh, Kolli.Srinivasa Rao, S. Punitha, M. Guru Vimal Kumar; Journal: SSRN Electronic Journal (2025)	Machine learning-based demand forecasting integrated with inventory optimization	Retail inventory and demand datasets; scenario-based forecasting inputs
20	Impact of Accurate Demand Forecasting on Inventory Stock Levels and Supply Chain Optimization	S. K. C. Tulli; Industry White Paper	Statistical demand forecasting and inventory performance analysis	Enterprise-level inventory and stock management data
21	Improvement of Inventory Management and Demand Forecasting by Big Data Analytics in Supply Chain	W. Tang; Journal: Applied Mathematics and Nonlinear Sciences (2024)	Big data analytics and predictive modeling for demand forecasting	Large-scale supply chain transaction and inventory datasets
22	Streamlining SCM: Integrating Demand Forecasting and Inventory Optimization	R. Hao; University of Edinburgh	Integrated demand forecasting and inventory optimization framework	Simulated and empirical supply chain datasets
23	An Integrated Approach to Improve Resilience in Agri-Food Supply Networks: A Sustainable Perspective	N. Clavijo-Buritica; Journal: Sustainability	Resilience-oriented supply chain optimization under demand uncertainty	Agri-food supply network data with demand variability

Much of the available information comes from stylized inventories or different distribution environments. This shows us that it is rarely tested in large-scale retail distribution networks with production and transportation constraints. Thus, the literature remains fragmented. Demand forecasting and network optimization are generally treated as separate stages. It has been observed that it neglects how forecasting model choices affect network decisions and costs. Consequently, empirical evidence quantifying the cost of forecasting error (CoFE) encountered in integrated production-distribution planning remains limited. This study provides machine learning-based demand forecasting. It proposes a forecasting-optimization framework that integrates a deterministic optimization model for network expansion. The approach combines predictive and prescriptive analytics. Then, it systematically compares forecasting models and evaluates cost impacts under sensitivity scenarios. While many studies focus on location selection and optimization, a small number of studies include demand forecasting, risk forecasting, and cost consequences within a single framework (Table 1) (No. 13), (Table 1) (No. 14), (Table 1) (No. 16).

Recent research shows that integrating demand forecasting with inventory and supply chain optimization can reduce system-wide costs and improve service levels (Table 1) (No. 19), (Table 1) (No. 21), (Table 1) (No. 22). These studies focus more on inventory control than network design and have not fully examined the impact of forecast-driven scenarios on logistics costs. Table 1 summarizes representative studies in the field of distribution network design and demand-driven optimization. Intuitive, multi-criteria decision-making, precision planning, and forecast-driven integration methods have been included (Table 1) (No. 1)-(Table 1) (No. 23).

2. Material and Methods

2.1. Production–Distribution Optimization Model

The relevant framework addresses a multi-factory distribution planning problem, evaluating operational constraints. It chooses to apply strict demand fulfillment conditions. Multi-product operations are considered by providing a network structure consisting of 5 warehouses and 100 customer locations. As an optimization model, it proposes integrating the error risk estimates generated by the classification framework into a prescriptive decision-making process. The estimated error probabilities are converted into expected cost parameters. Thus, the model can be deterministic and manageable while accounting for quality-related risks. The system internally decides whether or not to perform pre-shipment inspection for each shipment. Operational constraints include warehouse-specific production time limits and overall quality control (QC) capacity.

Given the operational constraints in the relevant retail network we focused on and the limited availability of historical data (100 observations), the model was designed to adhere to robust probability distributions. Traditional stochastic programming approaches requiring large-scale data were avoided to mitigate the risks of significant dispersion estimation. Instead, a data-driven forecasting paradigm was adopted. To ensure high model reliability and completely avoid overfitting in this data-scarce environment, the feature space was limited to 5 operationally vital parameters (price, availability, inventory status, logistics, and historical trends). Furthermore, the structural robustness of the model against small sample variance was rigorously validated through comprehensive deterministic sensitivity analyses across 20 different demand deviation scenarios.

Sets

P : set of products, indexed by p

F : set of factories, indexed by f

L : set of customer locations, indexed by l

Parameters

cp_p : unit production cost of product p (\$/unit)

t_p : production time required per unit of product p (hours/unit)

K_f : total available machine – hours at factory f (hours)

d_{pl} : demand for product p at customer location l (units)

c_{fl} : unit transportation cost from factory f to location l (\$/unit)

q_{pl} : predicted probability that an order of product p for location l is defective

κ : inspection cost per unit (\$/unit)

λ_p : expected penalty cost of shipping a defective unit without inspection (\$/unit)

C^{QC} : total inspection capacity (units)

M : a large positive constant used for linearization (Big – M)

Decision Variables

$x_{pf} \geq 0$: units of product p produced at factory f

$y_{pfl} \geq 0$: units of product p shipped from factory f to customer l

$cost_{qpl} \geq 0$: auxiliary continuous variable for total quality related cost

$z_{pl} \in \{0,1\}$: binary inspection decision for order (p,l)

The objective function (1) aims to reduce the operational and quality-related costs of the supply chain system. These costs include direct production costs, transportation costs, and risk-adjusted quality costs. The demand fulfillment constraint (2) establishes a rule that the total volume of product shipped from all facilities to a given location must absolutely meet the estimated demand for each product-location pair. On the production side, the plant capacity constraint (3) limits the total production time to the available operating hours at each plant, while the production-shipment balance (4) maintains the flow process by ensuring that the quantity of product produced at a factory equals the total quantity shipped to various distribution points. To maintain the linear structure of the model while addressing the conditional logic of these quality costs, the Big-M linearization approach is used in constraints (5) and (6). The inspection capacity constraint (7) defines the quality control process by ensuring that the total volume of product selected for inspection does not exceed the available quality control resources. This constraint gives the expected cost amount associated with the risk of defects.

$$\min Z = \sum_{p \in P} \sum_{f \in F} c_p \cdot x_{pf} + \sum_{p \in P} \sum_{f \in F} \sum_{l \in L} c_{fl} \cdot y_{pfl} + \sum_{p \in P} \sum_{l \in L} cost_{qpl} \quad (1)$$

$$\sum_{f \in F} y_{pfl} = d_{pl} \forall p \in P, \forall l \in L \quad (2)$$

$$\sum_{p \in P} t_p \cdot x_{pf} \leq K_f \forall f \in F \quad (3)$$

$$x_{pf} = \sum_{l \in L} y_{pfl} \forall p \in P, \forall f \in F \quad (4)$$

$$cost_{qpl} \geq \left(q_{pl} \cdot \sum_{f \in F} y_{pfl} \cdot \kappa \right) - M \cdot (1 - z_{pl}) \forall p, l \quad (5)$$

$$cost_{qpl} \geq \left(q_{pl} \cdot \sum_{f \in F} y_{pfl} \cdot \lambda_p \right) - M \cdot z_{pl} \forall p, l \quad (6)$$

$$\sum_{p \in P} \sum_{l \in L} z_{pl} \cdot d_{pl} \leq C^{QC} \quad (7)$$

The mathematical model is formulated tactically as a single-period network allocation and quality orientation problem. To focus on how machine learning prediction errors propagate directly across spatial logistics boundaries (defined as the Cost of Prediction Error), multi-period inventory control dynamics, lead time uncertainties, and overt supplier disruption risks have been excluded from the direct scope of this formulation.

2.2. Prediction, Labeling, Optimization Methodology

The proposed framework is structured as an MILP problem using Big-M reformulations to maintain a linear structure solvable to a global optimum using standard exact algorithms. In this context, Phase I (Forecast Analysis) utilizes five selected factors—price, availability, inventory status, logistics, and historical trends (demand and quality)—to train demand forecasting models in Step 2 and forecast defective production risks in Step 3. The selection of these five key factors aims to increase the generalizability of the model by minimizing the risk of overfitting in a limited dataset of approximately 100 observations and 24 candidate variables. As documented in the statistical literature, reducing complexity is critical for model stability when the number of observations is small (Hastie et al. 2009; James et al. 2021). Therefore, only the variables most important from an operational and economic perspective are focused on. Functionally, **price** determines demand elasticity, **availability** and **inventory status** reflect order fulfillment capacity, and **logistics** data reflect transportation times and route efficiency, ensuring the forecasts are grounded in realism. Furthermore, data obtained through the analysis of **historical trends** allows for the "noise-free" processing of both demand forecasts (phase 1) and defect risk forecasts (phase 2), and their transfer to the optimization model as more reliable inputs d_{pl} and q_{pl} . This choice prevents the model from memorizing existing data while simultaneously providing a more robust and interpretable decision support mechanism against external scenarios. Phase II produces simplified versions containing only two key explanatory components for sentiment analysis and scenario generation to guide the calibration processes described in steps 5 through 15. Due to the highly skewed distribution of defective products, the aim is to balance the classes using the relative ranking of the 80th percentile instead of a fixed threshold. 70% and 80% targets are chosen to represent operational risk tolerance.

The 80% target simulates situations where the cost of losing defective products is high, while the 70% target provides a sensitivity analysis for situations where inspection capacity is limited. The final stage of the methodological framework is Algorithm 1 - Phase III (Predictive Optimization) (Step 17-22), where the forecast outputs are converted into operational decisions. Phase III combines the data from the previous stages into an integrated decision support structure. Step 17 defines the model setup and objective function to minimize the sum of production, transportation, and linearized quality costs. Data from the forecast layer is exported in a parametric format and placed into the deterministic Mixed Integer Linear Programming (MILP) model defined in Step 17. The objective function (minZ) is designed to minimize the sum of production, transportation, and linearized quality costs. The uncertainty situation was addressed through scenario-based parameterization of forecast variability. XGB, GB, DT, and LR forecasting methods were used for transportation and demand modeling. Each model was evaluated under three demand conditions: base case and $\pm 10\%$ deviations. LR provided the base case. Each combination of model and variance created a different deterministic scenario where the MILP objective function and constraints (including strict demand satisfaction and resource capacities defined in step 18) were solved independently. This ensured that the observed cost and allocation differences stemmed from the forecast inputs rather than the model reformulation. Subsequently, failure risk integration followed a similar structure in the rule-making phase. The predicted failure probabilities, calibrated under two recall-oriented configurations, were converted into expected quality-related cost parameters. These parameters directly influenced the binary control decisions during MILP execution in step 19. The recall objective changed the intensity of the control allocation but did not change the constraint set or objective structure of the optimization model.

Algorithm 1: Recall-Driven Integrated Prediction–Optimization Framework

Require: Historical demand and quality datasets, operational capacities (K_f, C^{QC}), unit costs ($cp_p, c_{fl}, \kappa, \lambda_p$)

Ensure: Scenario-specific optimal production (x_{pf}) distribution (y_{pfl}), and inspection (z_{pl}) plans

- 1: **Phase I: Predictive Analytics**
- 2: Train the regressor M_{demand} to generate point forecasts $\hat{d}_{pl}$ for all product-location pairs
- 3: Train the classifier M_{defect} to compute raw defect probabilities $\{\hat{p}_{pl}\}$ in $[0,1]$
- 4: **Phase II: Scenario-Based Thresholding and Risk Calibration**
- 5: **for** each target recall scenario $R \in \{70\%, 80\%\}$ **do**
- 6: Calibrate the optimal probability threshold, τ_R such that:
Recall (τ_R) $\geq R$ based on the validation precision-recall curve
- 7: **for** each order (p, l) **do**
- 8: Compute the effective defect risk parameter q_{pl} using $\hat{p}_{pl}$
- 9: **if** $\hat{p}_{pl} \geq \tau_R$ **then**
- 10: Flag order as High-Risk for potential inspection (z_{pl} eligibility)
- 11: Map risk to objective function as a trade-off between κ and λ_p
- 12: **else**
- 13: Flag order as *Low-Risk*, minimize inspection priority
- 14: **end if**
- 15: **end for**
- 16: **Phase III: Prescriptive Optimization**
- 17: Construct the MILP objective function:
min Z = Total Production + Transportation + Linearized Quality Costs
- 18: Subject to:
1. Strict Demand Satisfaction: $\sum y_{pfl} = \hat{d}_{pl}$
2. Resource Constraints: Factory Capacity (K_f) and QC Capacity (C^{QC})
3. Linearization: Big-M constraints for cost- q_{pl} based on z_{pl}
- 19: Execute MILP solver (Branch-and-Bound)
- 20: **Record** scenario results: Total System Cost, Resource Utilization, and Inspection Density
- 21: **end for**
- 22: **return** Comprehensive comparative report across recall-based risk scenarios

3. Results and Discussion

To assess the robustness of the integrated predictive-constraint framework, a comprehensive sensitivity analysis was conducted under 70% and 80% recall scenarios. The findings clearly demonstrate the impact of $\pm 10\%$ demand deviations from Gradient Boosting (GB), XGBoost (XGB), Decision Tree (DT), and Linear Regression (LR) models on logistics costs. The mathematical model consists of 120 decision variables and 66 constraints. The proposed MILP model was implemented in Python using the CPLEX 22.1.2.0 solver. All experiments were performed on an Intel i7 processor with 16 GB of RAM. The optimum gap was set at 0.1% for all test samples. The average CPU time across 20 scenarios was 0.0344 seconds, demonstrating the model's suitability for near real-time retail decision-making processes and proving its speed in optimizing thousands of different deployment scenarios in seconds.

Examining the optimization results, it is observed that the MIP Gap values clustered in a narrow band between 0.1460% and 0.1472% in the 70% recall scenario and between 0.1467% and 0.1484% in the 80% recall scenario. The micro-level increase in gap values observed when the recall target is increased from 70% to 80% (a change of 0.002%) demonstrates that the model maintains its solution accuracy despite increased constraint rigidity. The fastest solution was obtained in the Gradient Boosting (Base) scenario with a time of 0.0319 seconds, while the longest solution time was recorded in the XGBoost (-10%) scenario with a time of 0.0376 seconds (Table 2- Table 3).

Table 2. Propagation of demand forecast errors under the 70% Defect Recall Scenario

Defect Recall Scenario 70%									
Model	Scenario	MAE	RMSE	Δ Cost (%)	Cost	Avg. Service Level	Avg. Capacity Utilization (%)	MIP GAP	CPU TIME
Gradient Boosting	-10%	327.79	425.69	-0.21%	5048434.38	101.20%	83.97%	0.146211%	0.0335
Gradient Boosting	Base	307.02	417.5	0	5037810.17	100.99%	83.80%	0.146318%	0.0319
Gradient Boosting	10%	301.62	402.69	0.22%	5048790.09	101.21%	83.99%	0.146257%	0.0332
XGBoost	-10%	337.1	466.54	0.30%	5071479.26	101.67%	84.36%	0.146243%	0.0376
XGBoost	Base	340.03	473.08	0.00%	5086900.64	101.97%	84.62%	0.146337%	0.0362
XGBoost	10%	329.35	459.07	-0.65%	5053851.18	101.32%	84.07%	0.146082%	0.0358
Decision Tree	-10%	735.15	902.24	0.24%	5040290.68	101.09%	83.88%	0.147068%	0.0327
Decision Tree	Base	755.05	913.35	0	5052263.87	101.33%	84.08%	0.147181%	0.0362
Decision Tree	10%	755.05	913.35	0%	5052263.87	101.33%	84.08%	0.147181%	0.0343
Linear Regression	Base	773.17	986.53	0	4924482.86	98.76%	81.95%	0.147286%	0.0346

Table 3. Propagation of demand forecast errors under the 80% Defect Recall Scenario

Defect Recall Scenario 80%									
Model	Scenario	MAE	RMSE	Δ Cost (%)	Cost	Avg. Service Level	Avg. Capacity Utilization (%)	MIP GAP	CPU TIME
Gradient Boosting	-10%	327.79	425.69	-0.21%	5051356.91	101.20%	83.97%	0.148352%	0.0339
Gradient Boosting	Base	307.02	417.5	0	5040736.37	119.99%	83.80%	0.148466%	0.0370
Gradient Boosting	10%	301.62	402.69	0.22%	5051623.58	101.21%	83.99%	0.148332%	0.0329
XGBoost	-10%	337.1	466.54	0.30%	5074468.30	101.67%	84.36%	0.148423%	0.0326
XGBoost	Base	340.03	473.08	0	5089780.01	101.97%	84.62%	0.148430%	0.0328
XGBoost	10%	329.35	459.07	-0.64%	5057044.19	101.32%	84.07%	0.148419%	0.0353
Decision Tree	-10%	735.15	902.24	0.24%	5039795.75	101.09%	83.88%	0.146705%	0.0326
Decision Tree	Base	755.05	913.35	0	5051765.14	101.33%	84.08%	0.146816%	0.0352
Decision Tree	10%	755.05	913.35	0%	5051765.14	101.33%	84.08%	0.146816%	0.0348
Linear Regression	Base	773.17	986.53	0	4924150.56	98.76%	81.95%	0.147036%	0.0340

The analysis results show that models with the highest prediction accuracy do not always guarantee the lowest operating costs. Increasing the defective product recall target from 70% to 80% resulted in a systematic increase in total logistics costs across all models. For example, in the GB base scenario, the increase in the recall target raised the cost from 5,037,810.17 units to 5,040,736.37 units. This reflects the "quality premium" created by the additional inspection and reallocation resources required to achieve stricter quality targets. Among the models, Gradient Boosting exhibited the highest stability at both recall levels; in the 80% scenario, a 10% increase in demand increased the cost by 0.22%, while a 10% decrease reduced the cost by 0.21%. This symmetrical response indicates that the GB method offers a balanced forecasting structure that interacts predictably with capacity constraints.

On the other hand, XGBoost, contrary to expectations, exhibited an asymmetrical cost response, demonstrating that the optimization layer effectively manages excessive projection costs through routing. In the Decision Tree model, sensitivity was found to be directly related to the recall target. The model, which provided a constant cost response in the 70% scenario, became more sensitive to demand changes, with a 0.24% increase when the 80% scenario was implemented. Interestingly, despite having the highest error metrics (MAE: 773.17, RMSE: 986.53), Linear Regression achieved the lowest base costs in both scenarios (4,924,482.86 and 4,924,150.56 respectively). This demonstrates that in constrained supply chain systems, a sometimes biased but stable forecast can outperform complex models, allowing the optimization layer to more effectively compensate for systematic errors. The Decision Tree provides the most stable structure. It is the model that maintains the most stable capacity utilization during the

transition from 70% to 80%. Capacity utilization clustered between 81% and 84% across all models. This indicates that the system is operating at an ideal capacity (no bottlenecks) and leaves approximately 15% slack capacity. The CPU time, around 0.03 seconds, shows that the model is fast enough to be used for real-time operational decisions (e.g., even when the delivery truck is at the door). Our model maintains operational stability by keeping the system at the 84% range without disrupting load balancing, even when external risks increase (defect rate). When the defect rate increased from 70% to 80%, the Gradient Boosting (Base) cost increased from 5,037,810 to 5,040,736. A 10% increase in defect risk was reflected in the total cost as only 0.05% thanks to the hybrid model; the risk was successfully absorbed. The Gradient Boosting (Basic) scenario was achieved with a stable 83.80% capacity utilization for a 119.99% service level and an 80% error rate. This demonstrated that the system was on the safest side. Thus, it proved that the model is proactive, not reactive, in a situation where the risk of error increases, by using more safety stock or more checkpoints to "guarantee customer demand.". The model operates with the lowest margin of error (MAE: 307.02) in the uncertain situation, along with Gradient Boosting. Although it slightly increases costs, it has been observed to maximize operational continuity (Table 4).

Table 4. Supplier Performance and Capacity Utilization Metrics under Gradient Boosting (GB Base Scenario, 80% Recall)

Product Type	Total Shipped	Avg Defect Rate	Supplier	Total Shipped (Units)	Capacity Used	Utilization (%)
Household Products, Education, Textiles, Electronics	14385.65	66.56%	Supplier 1	10240.27	18432.49	92.16%
Detergents, Personal Care, Cosmetics	20563.95	44.16%	Supplier 2	10673.07	19211.53	96.06%
Paper, Healthcare, Baby Products	11606.68	41.37%	Supplier 3	11111.11	20000	100%
			Supplier 4	4999.49	8999.08	45.00%
			Supplier 5	9532.34	17158.21	85.79%

Under the 80% recall scenario, Supplier 3, which is a Gradient Boosting (GB) Bottleneck and Full Capacity Utilization, achieved a 100% utilization rate by using its entire 20,000 unit capacity. This demonstrated Supplier 3's strategic position in the logistics network and its priority in terms of cost optimization. This situation showed that the system lacked resilience against sudden demand fluctuations at this point. From a Critical Threshold and Operational Risk perspective, Supplier 2 (96.06%) and Supplier 1 (92.16%) exceeded the 90% safe operational threshold. The low-cost solution offered by the Gradient Boosting model was observed to minimize the operational risk premium. From a Load Shifting and Idle Capacity perspective, while other suppliers operated at almost full capacity, Supplier 4 (45%) had the lowest utilization rate. Thus, the model eliminated costly routes. It confirmed that shifting the workload to central suppliers (e.g., Supplier 3) reduced overall costs. Looking at Product-Based Quality Risk Distribution, examining shipment details revealed that the "Household Goods, Education, Textiles, Electronics" group had the highest defect rate (66.56%), but also occupied a large volume. In the hybrid model, Supplier 3 operated at 100% capacity. This positively impacted quality costs by enabling the management of the high-risk product group.

Table 5. Supplier Performance and Capacity Utilization Metrics under Linear Regression (LR Base Scenario, 80% Recall)

Product Type	Total Shipped	Avg Defect Rate	Supplier	Total Shipped (Units)	Capacity Used	Utilization (%)
Household Products, Education, Textiles, Electronics	15522.27	66.56%	Supplier 1	9825.99	17686.77	88.43%
Detergents, Personal Care, Cosmetics	17847.95	44.16%	Supplier 2	10136.07	18244.93	91.22%
Paper, Healthcare, Baby Products	12158.63	41.37%	Supplier 3	11111.11	20000	100%
			Supplier 4	7968.38	14343.08	71.72%
			Supplier 5	6487.3	11677.15	58.39%

The LR model distributed the workload more evenly among suppliers compared to the DT model. Additionally, it maintained full capacity utilization (100%) for Supplier 3. It operated other key suppliers, such as Supplier 1 (88.43%) and Supplier 2 (91.22%), with high efficiency (Table 5). It implemented a "load balancing" strategy in its logistics network. It kept shipment volumes within a certain threshold for the product group with the highest defect rate (66.56% - Household Goods, Electronics, etc.). It optimized quality risk through operational capacity. The LR (Basic) scenario yielded the cheapest result for both 70% and 80% defect rates. Although the Linear Regression model provided the lowest total cost (4,924,150.56 units) among all scenarios, it exhibited the highest prediction error with an MAE value of 773.17. In the 80% recall scenario in the model, the Linear Regression basic model provided the lowest service level with 98.76%. The model revealed the probability of a stock shortage risk in the system by showing a MIP GAP (0.147036% in the 0.80% scenario).

3. Conclusion

In this study, a hybrid forecasting paradigm for retail supply chain planning in the face of two uncertainties (demand fluctuations and defect risk) is presented and empirically validated. The system outperforms traditional discrete methods by explicitly linking forecast accuracy to operational and economic outcomes through the integration of machine learning-based forecasting with a deterministic optimization model. The main contribution is a methodical evaluation of risk-informed choices in two defect recall scenarios (70% and 80%) and provides a comprehensive examination of the effects of quality thresholds on total system overheads.

Experimental results show that prediction accuracy alone is insufficient for decision support in distribution network planning. Even if the base classification model maintains consistent discriminatory ability among various targets, choosing the correct active threshold is important for system performance. Sensitivity analysis shows that advanced ensemble models such as GB and XGB exhibit varying levels of cost sensitivity to demand fluctuations, often influenced by specific recall targets and network constraints. Models with higher statistical errors, such as LR and DT, sometimes produce more stable and cost-effective operational plans. The study highlights the importance of examining forecast models from an operational perspective and highlights the need to align forecasts with optimization rather than relying solely on traditional accuracy measures. Although it has limitations such as being a single-period deterministic method and focusing on a single location, it provides enlightening information. Future studies could examine capacity expansion, multi-period stochastic models, or reliable forecasting methods. A comprehensive framework for decision support in retail logistics under uncertainty is provided by combining forecasting, optimization and scenario analysis.

The proposed hybrid framework provides robust, real-time optimal decision-making under limited data. It offers a fertile ground for future expansions within certain constraints. Future research will extend this single-period tactical model into a dynamic, multi-period planning framework incorporating inventory holding dynamics and lead time uncertainties. Additionally, the integration of stochastic optimization or distributionally robust optimization variants could further strengthen the framework's ability to absorb catastrophic supplier disruptions and broader macroeconomic shocks.

References

- Hastie, T., Tibshirani, R. and Friedman, J., The Elements of Statistical Learning, 2nd Edition, Springer, New York, 2009, <https://doi.org/10.1007/978-0-387-84858-7>.
- James, G., Witten, D., Hastie, T. and Tibshirani, R., An Introduction to Statistical Learning, 2nd Edition, Springer, New York, 2021, <https://doi.org/10.1007/978-1-0716-1418-1>.
- Makridakis, S., Spiliotis, E. and Assimakopoulos, V., M5 accuracy competition: Results, findings, and conclusions, International Journal of Forecasting, 38(4), 1346–1364, 2022, <https://doi.org/10.1016/j.ijforecast.2021.11.013>.
- Bertsimas, D. and Kallus, N., From predictive to prescriptive analytics, Management Science, 66(3), 1025-1044, 2019, <https://doi.org/10.1287/mnsc.2018.3253>
- Sanders, N.R. and Graman, G.A., Quantifying costs of forecast errors: A case study of the warehouse environment, Omega, vol. 37(1), 116-125, 2009, <https://doi.org/10.1016/j.omega.2006.10.004>
- Alumur, S.A., Kara, B.Y. and Karasan, O.E., Multimodal hub location and hub network design, Omega, 40(6), 927–939, 2012, <https://doi.org/10.1016/j.omega.2012.02.005>.

Biographies

Abdülvahap Özdemir graduated from Istanbul Aydın University with a degree in Industrial Engineering. Throughout his professional career, he has worked in the fields of entrepreneurship, project management, operational planning, and reporting. Abdülvahap Özdemir currently works as a Data Warehouse Specialist at Migros A.Ş.

Emel Şeyma Küçükaşçı is an Assistant Professor in the Department of Industrial Engineering at Istanbul Ticaret University, Istanbul, Turkey. She received her B.Sc. degree in Industrial Engineering from Istanbul Ticaret University, and her M.Sc. and Ph.D. degrees in Industrial Engineering from Boğaziçi University. Her current research focuses on developing optimization methods through statistical learning.