International Journal of Computer (IJC)

ISSN 2307-4523 (Print & Online)

https://ijcjournal.org/index.php/InternationalJournalOfComputer/index

AI-Driven Capacity Forecasting for Last-Mile Logistics

Ankur Fnu*

Manager, Program Management at Amazon, Seattle, WA - USA

Email: ankurcnp@gmail.com

Abstract

The article examines the features of throughput forecasting based on artificial intelligence in the field of last-mile logistics. The relevance of the topic is driven by the fact that, under conditions of rapid growth in e-commerce and tightening consumer expectations, last-mile logistics faces unprecedented pressure, requiring hyper-accurate forecasting of operational capacity to enable a just-in-time delivery model and meet delivery deadlines without excessive costs. Traditional planning methods demonstrate their inadequacy under conditions of high demand volatility. This work describes an artificial intelligence (AI)-based framework used for forecasting daily resource requirements (personnel, transport) in large-scale logistics networks. The framework is based on a hybrid machine learning architecture that combines regression models for baseline load forecasting and the XGBoost algorithm for detecting and quantifying abnormal demand spikes. A key element of the system is a dynamic capacity buffer algorithm that, in real time, calculates the required reserve of resources to mitigate risks associated with forecast errors. The article analyzes the architecture, implementation methodology, and empirical results, and also discusses the role of such systems as a strategic tool for demand shaping.

Keywords: Last-mile logistics; capacity forecasting; machine learning; XGBoost; dynamic buffers; demand shaping; operational efficiency; predictive analytics; supply chain management; just-in-time delivery.

Received: 10/3/2025
Accepted: 12/3/2025
Published: 12/13/2025

----

* *Corresponding author.*

1.Introduction

The modern economy is characterized by the dominance of the on-demand model, catalyzed by the exponential growth of e-commerce. This shift has fundamentally changed consumer expectations, establishing new standards for delivery speed, reliability, and flexibility, effectively pushing the industry towards a just-in-time delivery paradigm at the final stage of the supply chain. The last mile — the final and crucial stage of the supply chain, responsible for delivering goods from the distribution center to the end consumer — has come under enormous pressure. This segment is not only the most logistically complex but also the most expensive. Thus, the task of last-mile optimization has transformed from a purely operational issue into a strategic one, defining companies' competitiveness in the market.

Traditional forecasting methods, such as time series analysis based on moving averages or simple statistical models, prove ineffective. They cannot adequately account for nonlinear dependencies, sudden demand spikes caused by marketing campaigns or external events, as well as local geographic and demographic specifics. This leads to a systematic mismatch between available operational capacity (number of couriers, vehicles) and actual demand. The consequences of such a mismatch are critical: excess resources lead to unjustifiably high operating costs, while shortages result in missed delivery deadlines (SLA violations), reduced customer satisfaction, reputational damage, and ultimately, loss of market share. The problem is exacerbated by the fact that managing operational capacity in logistics has shifted from an optimization task under conditions of certainty to a risk management task under deep uncertainty.

As a solution to these challenges, the academic and expert community actively proposes the use of artificial intelligence (AI) and machine learning (ML) technologies. The existing literature is extensive; however, a certain segmentation is observed. A significant part of the research is focused on theoretical models of route optimization (Vehicle Routing Problem, VRP) or on highly specialized aspects of forecasting, often without considering the complex operational environment. There is a shortage of scientific works presenting comprehensive, end-to-end frameworks that are not merely theoretically described but implemented and validated under real large-scale commercial conditions. It is precisely this gap between theoretical developments and their practical, scalable application with proven effectiveness that this study seeks to address.

The purpose of the study is to formalize and conduct a broad analysis of an innovative AI framework for forecasting and managing operational capacity, implemented in a large-scale last-mile logistics network.

- To achieve this goal, the following tasks were defined: Describe the architecture of a hybrid machine learning model combining regression analysis for forecasting baseline load and the gradient boosting algorithm (XGBoost) for detecting abnormal spikes.

- Detail the dynamic capacity buffer algorithm as a key mechanism for operational risk management.

- Conduct an empirical assessment of the implemented system's effectiveness based on real operational data regarding planning accuracy and SLA compliance.

- Discuss the practical aspects of integrating AI solutions into operational processes and their relationship with active demand shaping strategies.

The scientific novelty of the work lies in the fact that a hybrid forecasting architecture is proposed, applied to the dualistic nature of last-mile demand (stable baseline + volatile peaks).

The author's hypothesis is that a hybrid forecasting architecture, combining regression models for stable demand, XGBoost for detecting abnormal peaks, and a dynamic power buffer algorithm, provides more accurate and cost-effective operational resource management under conditions of high demand volatility in last-mile logistics.

2.Materials and Methods

Research on capacity forecasting in the last mile forms several interconnected directions, which can be grouped by models and the position of forecasts in the decision-making chain. The first group includes works where demand forecasting and/or resource requirement forecasting are embedded in “forecast-then-optimize” formulations for routing and tactical planning: these are machine learning and hybrid methods for classical LMD (last-mile delivery) problems, including route packaging, parcel assignment, and fleet allocation [2, 5, 10, 12]. The second group comprises statistical-ML approaches to disaggregated and intermittent demand, where the main goal is to transform intermittent sequences of orders at the “address-hour/slot” level into distribution forecasts suitable for capacity calculation [3]. The third group consists of architectural and review papers on digital platforms integrating AI components into last-mile operational processes, including sustainability, mixed reality, and LLM assistants [1, 4, 6, 7, 8, 9]. The fourth group covers research on autonomous delivery means and multi-agent management, where the forecasting component is associated with estimating the load and throughput capacity of hybrid “truck-drone” and crowdsourcing systems [9, 11, 12]. In the works of the first group, the forecast serves as an input for optimization. Bruni M. E. and his colleagues [2] propose an ML-optimization scheme where the predicted volumes and spatiotemporal demand distribution are fed into VRP/VRPTW solvers. The authors emphasize joint tuning of the forecast model’s hyperparameters and the optimizer parameters, which reduces the gap between prediction and actual KPIs. Fadda E. and his colleagues [5] consider tactical capacity planning (fleet size, shifts, hubs) over a week-month horizon, combining ML for demand with stochastic programming/robust approaches for planning; the key technique is scenario branching by forecast quantiles, allowing the “capacity buffer” to be parameterized under SLA. Ghosh M. and his colleagues [10] demonstrate the “learn global, optimize local” paradigm: a global model extracts demand and traffic patterns across the network, while the local optimizer adapts the solution to district-level street micro-geometry and operational constraints; such decomposition improves the transferability of forecasts and stabilizes utilization. In crowdsourcing scenarios where capacity is partially endogenous (workers join/leave), Wang L., Xu M., Qin H. [12] combine joint parcel allocation and crowd routing: demand is modeled probabilistically, and “capacity” is modeled through an elastic supply function dependent on price and time, creating a “forecast-incentives-routes” linkage.

The second group focuses on the most “difficult” nature of the last mile — sparse demand with strong seasonality and bursts at low aggregation levels. Khan N. T., Al Hanbali A. [3] systematize methods for intermittent series:

from Croston/TSB modifications and bootstraps to gradient boosting and recurrent networks, emphasizing the importance of probabilistic inference (quantiles/predictive intervals) and proper accounting for calendar/weather/promotion factors. In the context of capacity planning, this is critical: adding penalties for under-delivery/re-routing to the loss function leads to "decision-focused" tuning, where the optimal criterion is not MAPE but service level — a thesis indirectly supported by the practices of the first group [2, 5].

The third, platform-architectural line describes how forecasting services are embedded in operational loops. Rosendorff A., Hodes A., Fabian B. [8] propose a procedural LML architecture where the capacity forecast module supplies the shift scheduler and router via a unified data layer and event bus — reducing the latency between signal appearance and reaction. Jucha P. [7] provides an overview of the spectrum of AI applications from forecasting to dynamic slot pricing. Oršič J., Jereb B., Obrecht M. [4] link ML processes with sustainability: load forecasting enables balancing the fleet between ICE/EV, minimizing emissions while maintaining SLA; the idea of multi-criteria planning (carbon, noise, hub overload) is introduced. Ieva S. and his colleagues [1] combine AI fleet optimization with mixed reality and LLM assistants in warehouses: MR reduces picking errors, stabilizing the input flow for the last mile, while LLM agents reduce operators' cognitive load; the result is less variability in the output flow, making capacity forecasting easier. Finally, Serkan Özarık S., da Costa P., Florio A. M. [6] discuss "data-driven LMD" as a pipeline: from feature engineering to online model updating. Capacity forecasting is viewed as a platform-level service supporting A/B testing in decision-making.

The fourth group reveals the role of autonomous and multi-agent systems, where capacity is a property of carrier interaction. Bi Z. and his colleagues [11] use multi-agent reinforcement learning to coordinate truck and drone operations: the policy distributes tasks considering stochastic order arrivals and battery constraints, which in fact coincides with the "online capacity" problem in a heterogeneous fleet. The review by Shuaibu A. S., Mahmoud A. S., Sheltami T. R. [9] systematizes strategies for integrating drones, robots, and human couriers and emphasizes that sustainable gains appear only when reliable short-term forecasts of arrivals and delivery windows are available; otherwise, systems lose efficiency. In crowd scenarios Wang L., Xu M., Qin H. [12] capacity arises endogenously, and thus forecasting requires joint consideration of workers' behavioral response to incentives — a bridge between econometrics and stochastic control.

Overall, the approaches used cover: probabilistic forecasting of intermittent demand with calibrated quantiles [3]; ML+optimization hybrids with explicit pricing of forecast error in capacity plans [2, 5]; multi-level "platform-as-a-model" architectures with online updating and integration into operational cycles, including MR and LLM assistants [1, 6, 8]; multi-agent RL approaches and hybrid fleets [9, 11] and (v) mechanisms for joint optimization of demand-supply in crowd platforms [12]. A cross-cutting theme in Ghosh M., Kuiper A., Mahes R., Maragno D. [10] is the decomposition into global learning and local optimization, which is practical for transferable capacity forecasts in diverse urban topologies.However, contradictions remain. First, there is still a gap between "forecast for accuracy" and "forecast for decision": some works focus on time-series error metrics, whereas operational KPIs require decision-focused learning and explicit accounting for penalty asymmetry [2, 3, 5]. Second, there is a contradiction between centralized and decentralized capacity management schemes: global schedulers improve consistency, but crowd- and multi-agent systems gain in adaptability at the cost of variability and requirements for online forecasting [11, 13]. Third, sustainability is often declared as a separate goal but rarely integrated as a hard constraint or loss function component in capacity forecasting Oršič J., Jereb B., Obrecht M. Reference [4] point the way, but operational implementations are limited). Poorly covered are: joint “forecast→optimization” learning with differentiable solvers and uncertainty transfer to KPIs; calibration of probabilistic forecasts for extreme events (promotion/weather peaks) in last-mile micro-geography; integration of MR/LLM assistants with predictive control to smooth flow variability (Ieva S. and his colleagues [1] outline this but do not close the question of impact metrics); long-term sustainability assessment with carbon budgets in multi-period capacity planning; econometrics of worker response to dynamic incentives in crowd models and its connection to robust supply forecasting. These directions set the agenda for the next wave of research on AI-driven capacity forecasting in the last mile.

3.Results and Discussion

The presented AI framework is built on a multi-level hybrid architecture designed for the efficient modeling of the complex and multi-component nature of last-mile logistics demand. Such an architecture is a deliberate design decision that allows leveraging the strengths of different classes of models to address specific sub-tasks [6, 7].

The levels of this framework are shown below in Figure 1.

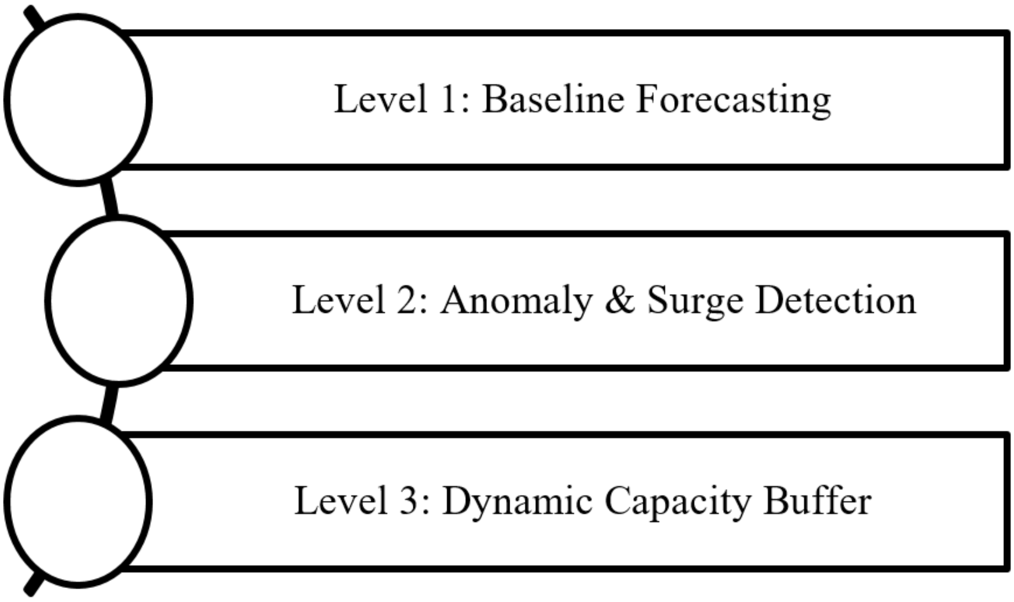
Level 1: Baseline Forecasting
Level 2: Anomaly & Surge Detection
Level 3: Dynamic Capacity Buffer

Figure1: Levels of an AI framework based on a multi-level hybrid architecture [6, 7]

At the first level, robust time series regression models are used to forecast the baseline, predictable component of demand. These models effectively capture stable patterns such as seasonality (e.g., pre-holiday peaks), intra-week cycles (higher demand on weekends), and long-term growth trends. The relative simplicity and high interpretability of these models make them an ideal tool for building a stable forecasting foundation, which aligns with the practice of using baseline models as a starting point in more complex systems [1, 9].

The second level of the architecture addresses the modeling of the nonlinear, hard-to-predict component of demand. For this purpose, the XGBoost (Extreme Gradient Boosting) model is used. The XGBoost algorithm was selected due to its proven high efficiency in handling complex, nonlinear, and non-monotonic dependencies, as well as its scalability and robustness when working with large datasets. The model is trained on a set of specially engineered features that may signal potential anomalies [5, 11]. Such features include data on planned marketing campaigns, demand dynamics over the past few hours, weather anomalies, and other external factors. It is important to note that this model does not simply forecast the deviation but provides a quantitative estimate of the probability and expected magnitude of the demand surge. This information is a critically important input for the next level of the system [4, 10].

The third level represents the core of the resource management strategy and is the key innovative element of the framework. It departs from traditional approaches with static reserves and implements an algorithm for dynamic calculation of the capacity buffer. This mechanism allows for a much more flexible and economically sound allocation of resources, moving operational processes closer to the principles of just-in-time delivery, where capacity is aligned with real-time demand rather than pre-planned averages [7, 12].

Below, Table 1 presents a comparative analysis of the models used in the hybrid architecture, which justifies the feasibility of such a design.

Table 1: Comparative analysis of models in the hybrid architecture [4, 7, 10, 12]

Parameter	Baseline Model (Linear Regression)	Anomaly Detection Model (XGBoost)
Objective	Forecasting stable, seasonal patterns	Identification and quantification of nonlinear surges and anomalies
Data type	Aggregated time series	Granular data, engineered features (promotions, etc.)
Complexity	Low, high interpretability	High, black-box model
Role in the system	Formation of the main demand forecast	Forecast adjustment, risk assessment for buffer calculation

Empirical validation of the framework was carried out based on a comparison of key operational indicators before and after its full-scale implementation in more than 50 fulfillment centers. The key result of the system implementation was an increase in the accuracy of forecasting personnel requirements by more than 20%. This indicator was measured as the mean absolute deviation between the forecasted and the actually required number of man-hours. The improvement of this indicator has a direct economic effect: costs associated with excessive staffing during periods of low workload are reduced, and operational disruptions and urgent measures caused by a shortage of resources during peak hours are minimized.The success of an operational AI system is determined not only by the accuracy of its algorithms but also by its ability to be effectively integrated into human workflows and serve as a foundation for strategic decision-making [2, 3].

A critical factor in the success of the described framework was its seamless integration into the daily work of operational, product, and regional teams. This was achieved through the development of intuitive dashboards and an automated alert system. These tools visualize in real time the forecasts, the current load level, and the recommended buffer capacity, providing local managers with actionable insights for decision-making. This approach aligns with the modern paradigm of human–machine interaction, where AI does not replace the human but augments human capabilities (human augmentation), freeing them from routine calculations and enabling focus on more complex and creative tasks [2, 8].

Further in Table 2, the advantages, limitations, and future trends in AI-based throughput forecasting for last-mile logistics will be described.

Table 2: Advantages, limitations, and future trends in AI-based throughput forecasting for last-mile logistics [2, 3, 8]

Advantages	Limitations	Future Trends
1. Enhanced accuracy in predicting demand fluctuations.	1. Dependence on high-quality, large-scale datasets for training.	1. Increased adoption of hybrid AI models combining machine learning with domain expertise.
2. Real-time adaptation to changes in traffic, weather, and customer behavior.	2. Potential bias in AI models due to incomplete or skewed data.	2. Greater use of edge computing for faster, decentralized decision-making.
3. Reduction of operational costs through optimized routing and scheduling.	3. Limited explainability of complex AI decision-making processes.	3. Integration with autonomous delivery systems (drones, robots).
4. Improved customer satisfaction via reliable delivery time windows.	4. High initial investment in infrastructure and skilled personnel.	4. Expansion of predictive analytics to include sustainability metrics.
5. Ability to integrate multiple data sources (IoT, GPS, ERP systems).	5. Vulnerability to cyberattacks and data breaches.	5. Development of explainable AI frameworks for transparent forecasting.

The availability of reliable forecasts and controllable buffer capacities allows the company to shift from reactive demand response to proactive demand shaping. This is the key strategic conclusion of the present study. By knowing in advance about potential capacity constraints or surpluses in a specific hub at a specific time, the system can initiate automatic adjustments to smooth demand peaks. For example, when high load is forecasted during evening hours, the platform can dynamically adjust the cost or availability of express delivery options, offering customers a discount for selecting a less congested morning or daytime slot. This mechanism directly links operational forecasting to advanced commercial strategies such as dynamic pricing and optimization of service options, which represents a cutting-edge direction in supply chain management. Thus, a causal chain is established: accurate technical model (forecast) → integration layer (dashboards) → enhanced managerial capabilities (augmented decision-making) → strategic action (demand shaping).

4.Conclusion

This article has presented and analyzed a comprehensive AI framework for forecasting and managing operational capacity in last-mile logistics. The study has confirmed that in conditions of high demand volatility, which is characteristic of the modern on-demand economy, effective management requires the application of sophisticated hybrid AI solutions that go beyond traditional forecasting methods.

The scientific and practical contribution of this work lies in the formalization and empirical validation of an innovative architecture that successfully combines predictive modeling (a hybrid of regression models and XGBoost) with algorithmic resource management through a dynamic capacity buffer mechanism. This approach enables not only accurate demand forecasting but also proactive operational risk management.

The empirical analysis, based on large-scale system deployment data, has demonstrated its high practical value. The achieved improvement in workforce planning accuracy by more than 20% and the reduction in SLA violations by 15% serve as compelling evidence of the commercial efficiency and return on investment of such technological initiatives.

In conclusion, it should be emphasized that AI capacity forecasting systems, such as the one described, are not merely operational tools but strategic assets. They ensure efficiency gains, increased customer satisfaction, and create a technological foundation for the implementation of advanced commercial strategies such as dynamic demand shaping, ultimately enabling a true just-in-time delivery operational model.

References

[1]. Ieva, S., Bilenchi, I., Gramegna, F., Pinto, A., Scioscia, F., Ruta, M., & Loseto, G. (2025). Enhancing Last-Mile Logistics: AI-Driven Fleet Optimization, Mixed Reality, and Large Language Model Assistants for Warehouse Operations. Sensors, 25(9), 2696. https://doi.org/10.3390/s25092696

[2]. Bruni, M. E., Beraldi, P., Guerriero, F., & Pinto, E. (2023). A machine learning optimization approach for last-mile delivery and third-party logistics. Computers & Operations Research, 157, 1-14. https://dx.doi.org/10.1016/j.cor.2023.106262

[3]. Khan, N. T., & Al Hanbali, A. (2025). Machine learning approaches for disaggregated and intermittent demand forecasting for last-mile logistics. Transportation Research Procedia, 84, 307–314. https://doi.org/10.1016/j.trpro.2025.03.077

[4]. Oršič, J., Jereb, B., & Obrecht, M. (2022). Sustainable Operations of Last Mile Logistics Based on Machine Learning Processes. Processes, 10(12), 2524. https://doi.org/10.3390/pr10122524

[5]. Fadda, E., Meloni, C., & Vancroonenburg, W. (2021). Mixing machine learning and optimization for the tactical capacity planning in last-mile delivery. In 2021 IEEE 45th Annual Computers, Software, and Applications Conference (COMPSAC), 1291-1296. https://doi.org/10.1109/COMPSAC51774.2021.00180

[6]. Serkan Özarık, S., da Costa, P., & Florio, A. M. (2022). Machine learning for data-driven last-mile delivery optimization. Machine Learning for Data-Driven Last-Mile Delivery Optimization, 1-22.

[7]. Jucha, P. (2021). Use of artificial intelligence in last mile delivery. SHS Web of Conferences, 92, 1-9. https://doi.org/10.1051/shsconf/20219204011

[8]. Rosendorff, A., Hodes, A., & Fabian, B. (2021). Artificial intelligence for last-mile logistics—Procedures and architecture. Online Journal of Applied Knowledge Management, 9(1), 46–61. https://doi.org/10.36965/OJAKM.2021.9(1)46-61

[9]. Shuaibu, A. S., Mahmoud, A. S., & Sheltami, T. R. (2025). A Review of Last-Mile Delivery Optimization: Strategies, Technologies, Drone Integration, and Future Trends. Drones, 9(3),158. https://doi.org/10.3390/drones9030158

[10]. Ghosh, M., Kuiper, A., Mahes, R., & Maragno, D. (2023). Learn global and optimize local: A data-driven methodology for last-mile routing. Computers & Operations Research, 159, 106312.

[11]. Bi, Z., Guo, X., Wang, J., Qin, S., & Liu, G. (2024). Truck-Drone Delivery Optimization Based on Multi-Agent Reinforcement Learning. Drones, 8(1), 27. https://doi.org/10.3390/drones8010027

[12]. Wang, L., Xu, M., & Qin, H. (2023). Joint optimization of parcel allocation and crowd routing for crowdsourced last-mile delivery. Transportation Research Part B: Methodological, 171, 111-135. https://doi.org/10.1016/j.trb.2023.03.007