Enterprise-Scale Machine Learning for Demand Forecasting

M. HARSHVARDHAN, CARA CURTLAND, ADAM GHOZEIL, AND CHUANREN LIU

"In theory there is no difference between theory and practice. In practice, there is."

— source unverified, although often attributed to Yogi Berra

PREVIEW *This article charts HP's voyage from simplistic statistical methods to an integrated, enterprise-grade forecasting pipeline powered by LightGBM, robust machine learning operations (MLOps), and augmented by human-in-the-loop design. The authors describe a dynamic system that learns, adapts, and delivers tangible business results. The phased implementation journey – from data infrastructure to full integration – offers valuable lessons for practitioners looking to deploy ML forecasting at scale. HP shows this is possible when machine learning is paired with human insight, not pitted against it.*

Many forecasting models work well in experimental environments but fall short of expectations when scaling up to enterprise deployments under real-world conditions. At HP, we have demonstrated a solution to this problem in our worldwide print forecasting system, which comprises over 18,000 products in over 170 countries. This article describes the key components of our solution, including not only state-of-the-art ML forecasting, but also our machine learning operations (MLOps) strategy for ongoing process management, human-in-the-loop deployment, and continuous process improvement. Because deployment is a journey, we will also share the development phases the project went through, and some of the key lessons we learned about the process of creating and integrating an ML deployment of this scale. For readers interested in the full technical methodology, implementation details, and quantitative evaluation, a comprehensive version of this work has been published in the *INFORMS Journal of Applied Analytics* (Harshvardhan et al., 2025).

ML-FORECAST HUMAN-IN-THE-LOOP SYSTEM INTEGRATION

Setting the Stage

Before our project, HP's print business forecasting included both statistical and consensus techniques. Statistical forecasting used models like ARIMA and exponential smoothing. Consensus forecasting combined the statistical modeling results, qualitative input, and human judgment to produce an 18-month forecast used for supply planning and strategic integrated business planning. These methods provided a baseline level of performance but struggled to keep up

with the increasing complexity and dynamism of global markets, especially for products with intermittent or seasonal demand. Furthermore, these approaches often lacked scalability and adaptability, requiring ongoing interventions to address evolving patterns.

Creating the End-to-End Pipeline and Measurement Foundation

While it's not as sexy as an ML algorithm, the workhorse behind successful deployments at scale is (1) getting a solid data pipeline in place, and (2) gaining alignment on the business-relevant metrics that will drive insights, actions, and continuous improvement. We began by developing an integrated data pipeline, including historical statistical and consensus forecasts along with coalescing a broad spectrum of relevant modeling features. These included historical sales, channel inventory, product details, and lifecycle stages.

The first product out of this meshed data pipeline was a prototype Tableau dashboard to get early insights and alignment on KPIs and views. Our drillable dashboards enabled executives and planners alike to have one source of truth to drive business decisions. This approach also allowed us to align all print business units to one method for calculating and managing the metrics most important to their improved performance. These metrics are bias, weighted MAPE (wMAPE), and forecast value add that measures the accuracy improvements contributed by the ML forecasting process over the consensus process. Lastly, the dashboard was embedded in the monthly S&OP workflow, where the forecasting team is held accountable to their metrics and improvement plans relative to published goals.

Developing the LightGBM ML Forecast

With our data pipeline and measure of success in place, we set our sights on improving forecast performance. Inspired by the success of LightGBM in the literature (Ke et al., 2017) and after our own investigation of various models, we created an iterative demand forecasting algorithm where the core model we use is LightGBM. This was chosen because it's fast, scalable, and still offers a degree of interpretability compared to many "black box" approaches. We train it iteratively, so each new forecast informs the next round, letting it recognize longer-term trends and nuances. We also fine-tune hyperparameters and use "warm starts" so that when we retrain, the model doesn't have to begin from scratch every time. After some engineering, our ML model includes 100-plus features, everything from broad seasonal patterns to geo-specific signals.

Systematizing with MLOps

While ML modeling is exciting, the prompt and reliable delivery of forecasts to decision makers is essential for adoption. That's where machine learning operations (MLOps) comes in (Kreuzberger et al., 2023). We automated everything: from monthly data pulls and validation, to versioning and backtesting, to

Key Points

- Build the foundation first, starting with robust data infrastructure and clear business metrics before developing models. Clean, connected data and aligned KPIs enable successful ML deployment and organizational trust
- Design for human-machine collaboration. Create systems where ML handles scale while leveraging human judgment for complex decisions. An integrated approach delivers better results than either humans or machines alone.
- Embed scalability from day one. From the first prototype, craft your system to support enterprise-wide demands using modular design, efficient algorithms, and comprehensive MLOps automation. Scalability can't be bolted on later.
- Focus on adoption, not just accuracy. Models add value only when integrated into planning systems and trusted by users. Phase implementation strategically, from visibility to manual adoption to full integration, building confidence at each stage.

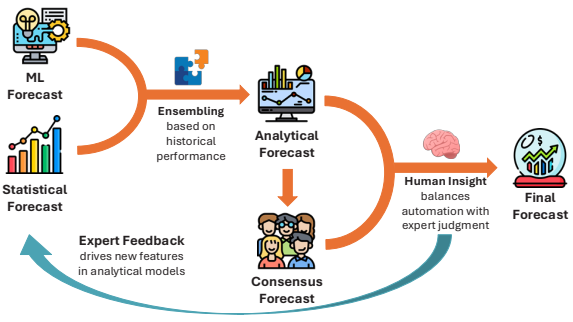
publishing final forecasts in HP's integrated business planning system. Tools like MLflow (Zaharia et al., 2018), an open-source system for ML development, help us track model experiments. Papermill, a Python tool for parametrizing Jupyter notebooks, streamlines notebook runs. And we store data with Apache Feather for speed and reproducibility. This end-to-end system ensures we can repeat our processes, compare performance over time, and adapt quickly – exactly what an enterprise solution needs to do.

Enabling Human-in-the-Loop Forecasting

Our approach integrates ML-driven insights with real-world forecasts, creating a "human-in-the-loop" process as pictured in **Figure 1** that empowers our planners while leveraging their domain knowledge. Wu et al. (2022) provides a survey of human-in-the-loop systems in ML.

Over time, we've evolved our end-to-end system to automatically ensemble our ML forecast with the statistical forecast, producing a baseline analytical forecast for the consensus forecaster to use directly or to enhance with judgment. Dashboards visualize model performance across geographies and planning horizons, driving prioritization of improvement efforts. Planner feedback continuously refines ML models through feature development and tuning – creating a winning cycle where models improve and planners maintain control. This closed-loop system lets planners focus on trickier SKUs that follow complicated patterns or benefit from information outside the model, while analytical models handle stable forecasts. This tight integration between ML and human planners ensures trust, adaptability, and uptake across HP's global print organization.

Figure 1. Human-in-the-Loop Ensembling for Deciding Final Forecasts



OPERATIONAL BENEFITS AND IMPACT

System Forecast Accuracy Improved

We evaluated our forecasting framework across multiple business-relevant horizons – monthly, three-month, and six-month – to align with HP's planning cycles using standard metrics like wMAPE, RMSE, and Bias. Our ML forecast outperformed traditional methods, especially at three- and six-month planning horizons, for a broad spectrum of products. Detailed quantitative results are available in Harshvardhan et al. (2025). The ML models delivered better accuracy with less manual intervention, while proving more scalable across diverse demand patterns. Most important, the human-in-the-loop system delivered improved performance. The enhanced ML forecast combined with human oversight remained vital – delivering superior consensus forecasts to planning for execution.

Inventory Improved

The real proof of system effectiveness is clear in the business outcomes. Since rolling out our human-in-the-loop ML forecast, we've seen a 28.5% drop in *normalized* inventory levels (over the three years from 2022 – 2024 for a select business segment of 1,484 products), while also protecting customer service levels. (Normalization means rescaling monthly inventory between 0 [minimum] and 1 [maximum] across the time period, to preserve anonymity.) This means we were able to match supply with demand at lower inventory levels – enabling savings in working capital, lower warehousing costs, and a more responsive supply chain. **Figure 2** shows how we've measured that reduction over time.

Figure 2. Reduction in Inventory Levels over Three Years for a Select Business Segment

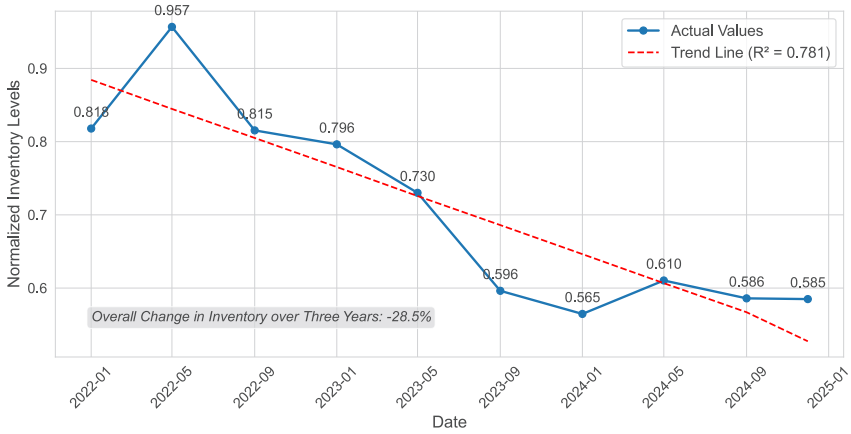
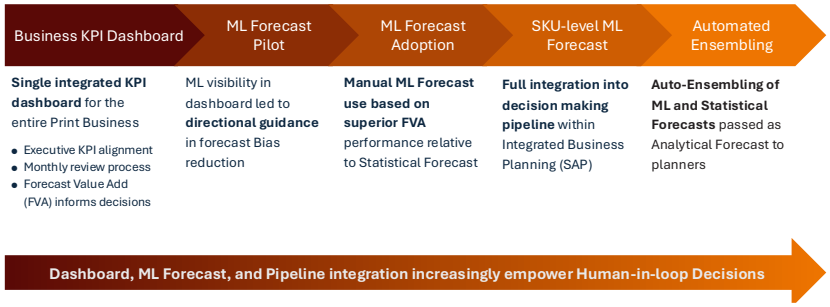


Figure 3. Staged Implementation Journey of ML Forecasting



KEY LESSONS

1. **Phased Implementation:** Our successful implementation followed a strategic phased approach. First, we built our data pipeline and an integrated KPI dashboard for all of HP's print business, creating single-source-of-truth visibility across organization levels. Next, we deployed ML forecasts and made them visible in the dashboard – providing directional guidance. When ML consistently outperformed existing methods, planner adoption accelerated naturally, through manual inclusion into models. We then fully integrated ML forecasts at the SKU level into the supply planning pipeline, culminating in automated ensembling (ML and statistical) in the forecasting pipeline. Our methodical approach fostered trust and alignment across the company. **Figure 3** details the five phases of our implementation.
2. **Start with Infrastructure, Not Models:** The best models are built on clean, connected data with robust MLOps. We prioritized building a robust data pipeline – integrating historical actuals, sell-through, and channel inventory – before modeling. Aligning on business-relevant KPIs early enabled consistent measurement and continuous improvement.
3. **Visualize Early, Align Often:** Before any ML was live, we launched interactive dashboards to align on metrics, spot anomalies, and build organizational trust. These visual tools became embedded in S&OP workflows, making insights accessible and decisions data driven.
4. **Design for Human + Machine, Not Human vs. Machine:** ML handled stable SKUs at scale; planners focused on judgment-heavy, high-impact exceptions. This human-in-the-loop

IMPACT STATEMENT

We are writing to confirm the successful implementation of Machine Learning (ML)-based forecasting within the HP Print Category Demand Management process. This solution is embedded across the entire Print worldwide operations and has led to transformational improvements in measured KPIs. For this competition, the team has focused on a specific business segment to showcase the operational performance improvements. In this segment, the published forecast improved wMAPE by 34% from Jan 2022 to Dec 2024. During the same period, inventory dollars reduced by 28% while maintaining high service levels to customers.

In 2019, we engaged HP's Strategic Planning and Modeling (SPaM) team to develop advanced analytics forecasts based on new developments in the field of machine learning. Over the past few years, the SPaM team led the project, steering from inception in 2019 through a yearlong pilot phase, to worldwide production in the summer of 2023. Collaborating with researchers from the University of Tennessee, Knoxville, the team designed, implemented, and deployed an AI-based system that significantly enhanced demand forecasting performance.

The success of their efforts resulted from the creation of an innovative forecasting workflow framework, an efficient and scalable ML pipeline, and a visibility and control KPI reporting solution coupled with full process integration. The ML model is now incorporated into the forecasting process used by all HP Print Category Forecasters in our organization, and we continue to work cross-functionally on the end-to-end analytics and process to drive better performance.

Steve Radosevich
Head of Global Demand Management
HP Inc, Print Category

Dele Oladeji
VP, Print Supply Chain Planning
HP Inc, Print Category

setup preserved domain expertise while unlocking ML's leverage – resulting in better forecasts and higher trust.

5. **Aim for Forecast Adoption, Not Just Accuracy:** ML forecasts can't add value to the organization if they aren't used. By working with planners as collaborators to develop the models, embedding them directly into HP's S&OP system, using performance metrics the business cares about, and making performance visible through dashboards, we ensured model output became actionable, accountable, and continuous.

6. **Make Scalability a Design Constraint:** Forecasting 18,000 SKUs in 170-plus countries wasn't an afterthought. Even early prototypes were designed for scale – leveraging efficient data structures, warm starts, and fast algorithms like LightGBM.

CONCLUSION

At a time when global supply chains are more unpredictable than ever, HP's experience highlights the value of an integrated, human-in-the-loop approach to forecasting. Instead of aiming for a "perfect model" in a vacuum, we built an ongoing, evolving system – one that leverages cutting-edge ML and invites human expertise at just the right touchpoints. With the updated decision-making system, we can reduce our inventory while keeping our customers happy. Because of our flexible structure, we can refresh our ML models, features, and decision making modularly. Looking ahead, this modular, human-plus-ML framework positions us to extend these forecasting gains across HP, while inspiring other organizations to rethink how people and algorithms can forecast smarter, not harder.

REFERENCES

Harshvardhan, M. et al. (2025). Print demand forecasting with machine learning at HP Inc. *INFORMS Journal of Applied Analytics*.

Ke, G. et al. (2017). LightGBM: A highly efficient gradient boosting decision tree. *Advances in Neural Information Processing Systems*, 30.

Kreuzberger, D., Kühl, N., & Hirschl, S. (2023). Machine learning operations (MLops): Overview, definition, and architecture. *IEEE Access*, 11, 31866-31879.

Wu, X. et al. (2022). A survey of human-in-the-loop for machine learning. *Future Generation Computer Systems*, 135, 364-381.

Zaharia, M. et al. (2018). Accelerating the machine learning lifecycle with MLflow. *IEEE Data Engineering Bulletin*, 41(4), 39-45.

M. Harshvardhan recently completed his PhD at the Haslam College of Business, University of Tennessee, Knoxville. He focuses on developing and deploying machine learning algorithms for practical applications. During his PhD, he worked with HP's Strategic Planning and Modeling team for over a year. He is currently an Assistant Professor of Information Systems and Analytics at the American University of Sharjah, UAE.

harshvar@utk.edu

Cara Curtland recently retired from HP as Supply Chain Data Science Strategist in the Strategic Planning and Modeling team. Her experience spans manufacturing, R&D, planning, forecasting, supply chain design, complexity management, and inventory optimization. Cara earned BS and MS degrees in industrial engineering from Purdue University. She is now the Director of Integrated Business Planning at NETGEAR.

cara.curtland@netgear.com

Adam Ghozeil served as a Principal Data Scientist in the Digital and Transformation Office at HP Inc. His focus was on developing data pipelines, AI models, and digital tools to drive efficiency and accuracy. Adam earned a BS degree in electrical engineering from UC San Diego and has 28 years of experience across R&D, manufacturing, and business functions. He currently works as Senior Product Manager at Microsoft.

adam.ghozeil@gmail.com

Chuanren Liu is an Associate Professor and Melton Faculty Fellow at the Haslam College of Business, University of Tennessee, Knoxville. He holds a PhD from Rutgers University. His research interests include data mining and knowledge discovery. He has published papers in various journals and conference proceedings, such as *IEEE Transactions on Knowledge and Data Engineering*, *INFORMS Journal on Computing*, and the *European Journal of Operational Research*.

cliu89@utk.edu