Print Demand Forecasting with Machine Learning at HP Inc.

M. Harshvardhan,[a]* Cara Curtland,[b] Jerry Hwang,[c] Chuck VanDam,[d] Adam Ghozeil,[e] Pedro A. Neto,[f] Frederic Marie,[g] Chuanren Liu[a]

[a]University of Tennessee, Knoxville, Tennessee; [b]HP Inc., Vancouver, Washington; [c]HP Inc., Palo Alto, California; [d]HP Inc., Boise, Idaho; [e]HP Inc., Corvallis, Oregon; [f]HP Inc., San Francisco, California; [g]HP Inc., Grenoble, France

*Corresponding author

Contact: harshvar@utk.edu, http://orcid.org/0000-0001-8086-544X (MH); cara.curtland@hp.com (CC); jhwang@alumni.stanford.edu (JH); vandam@hp.com (CV); adam.ghozeil@hp.com (AG); pedro.neto@hp.com (PAN); frederic.mr.marie@hp.com (FM); cliu89@utk.edu, http://orcid.org/0000-0001-9030-8495 (CL)

Abstract. HP Inc. manufactures and sells more than 18,000 Print-related products in over 170 countries. Accurate forecasting of the heterogeneous and dynamic demand is vital to support supply planning decisions for manufacturing, inventory management, shipment scheduling, and ultimately, customer satisfaction. Forecasting higher or lower than actual demand results in excess or shortage that reduces profitability and impacts on-time delivery to customers. Historically, the supply planning depended on (1) consensus demand forecasting approach, which requires manual collection and integration of information by the forecasting experts, and (2) statistical time-series forecasting models. The consensus forecasting approach also requires frequent corrections if some uncertainties in the demand are not accounted for when releasing the forecasting results. Traditional time-series models can work automatically without frequent correction, but their forecasting performance is unsatisfactory because of oversimplified modeling inputs and assumptions. In this project, we document the process of using machine learning (ML) techniques across all Print products at HP Inc., worldwide. Our aim is to automate the forecasting process with high accuracy and to integrate those results into a human-in-the-loop process that merges the strengths of ML, statistical, and consensus forecasting. Our tree-based (LightGBM) forecasting model reduced systematic errors in comparison with existing approaches, such as the consensus and statistical forecasting approaches, and was deployed as an integrated part of HP Inc.'s forecasting process. Furthermore, our ML framework establishes strong foundation for further methodological improvements in the ML algorithm. We report extensive empirical evidence guiding our methodology design and demonstrating the business implications of our project. We also share several important principles we have applied to manage team-based collaboration for an enterprise-scale project and to ensure the success of our ML-based demand forecasting.

Keywords: **printers and electronics, data-driven decision making, demand forecasting, machine learning**

Introduction

Background

HP Inc. manufactures and sells over 18,000 stock-keeping units (SKUs) of Print products that are sold in over 170 countries. They include home printers, office printers, ink, toner, and other services such as 3D and large-format printing. Specifically, home printers are targeted to consumers looking to buy standalone printers. They're usually sold through channel partners, including retailers like Walmart and Amazon. Office printers are usually sold via business contracts through managed account deals. The consumables, ink and toner, are sold to existing printer owners. 3D Printing offers a portfolio of additive manufacturing solutions and supplies to help customers with unique or experimental demands. Additionally, HP offers large-format printing solutions and supplies through industrial products. Beyond these five top-level categories, products are further classified based on their technology and platform, resulting in over 18,000 SKUs. Building on this portfolio breadth, HP operates on a global scale with markets organized into three world regions: Americas (AMS); Europe, Middle East, and Africa (EMEA); and Asia-Pacific (APAC). Countries in each world region are grouped by geographical proximity, and the demand forecasting is needed for each SKU in each group of countries (GOC).

Given its diverse product portfolio and extensive global reach, accurate demand forecasting is a crucial component of operational strategy for an international company like HP. Indeed, accurate forecasts are critical to planning and operational decisions such as strategically allocating resources, managing inventory, and aligning production schedules with consumer demand (Gardner 1990, Ritzman and King 1993, Lee 2002, Seifert et al. 2015). Furthermore, past studies have highlighted that effective forecasting not only can support business operations but also can lead to cost savings and improved efficiency throughout the supply chain (Simatupang and Sridharan 2005, Seifert et al. 2015, Fildes et al. 2022). With the advancement of machine learning (ML) technologies, there has been a significant interest from academics and practitioners in applying ML methods for these forecasting tasks. This paper discusses the challenges and solutions to deploy an ML-based framework to forecast product demand for a Fortune 500 technology company like HP.

Current Practices

Before implementing ML-based models, we relied on *statistical* and *consensus* forecasts for demand forecasting. The statistical forecasts leverage historical demand data and

use conventional time-series models, such as autoregressive (AR), moving averages (MA), ARMA, ARIMA, and exponential smoothing (ETS) models (Hyndman and Athanasopoulos 2018). These models are cost-effective and easy to implement, but they often lack the nuance required for accurate forecasting because of oversimplified modeling assumptions. Statistical models are also "local" in nature, training with a single time series, whereas ML-based models are "global," incorporating details from multiple time series. Local models struggle with short product life cycles, whereas global models learn from similar products. A common attempt to handle this is through predecessor-successor mapping, but such information is not always readily available to forecasters (Manary et al. 2019). In contrast, the consensus forecasts incorporate quantitative information such as historical demand and current inventory levels, as well as qualitative demand signals and contextual information, with the statistical forecast also serving as an input. Particularly, the consensus forecasters heavily leverage "soft data" like customer demand sentiments and deal progress. Soft data include qualitative knowledge on upcoming promotions offered by channel partners to their customers, deal stage for bulk corporate orders, subjective opinions from market insiders and experts, and networking insights through deep business relationships, among others (Fildes et al. 2009, Petropoulos et al. 2018). Although soft data are challenging to include and maintain, their strategic advantages in capturing transient market conditions make the data invaluable to forecasting, especially contributing to robustness of planner forecasts. Moreover, the superiority of a data-based method compared with human judgmental forecasts is not always true. Zellner et al. (2021) surveyed literature on human judgment and quantitative forecasting as well as hybrid methods that involve both humans and algorithmic approaches. They found that although quantitative methods have gotten popular over time, they are not universally superior to human judgment; the better method is subject to the availability, quality, extent, and format of data. Indeed, the two approaches can complement each other to yield more accurate and resilient models. Recent research also shows that human-based forecasts struggle to effectively filter out noise in the inputs. In fact, forecasters tend to reproduce the noise in a time series in their forecasts rather than filter it out (Petropoulos and Siemsen 2023). Khosrowabadi et al. (2022) evaluate AI-generated forecasts for a major European retailer, revealing that product attributes like price, freshness, and discounts play a crucial role in adjustment decisions. They find that

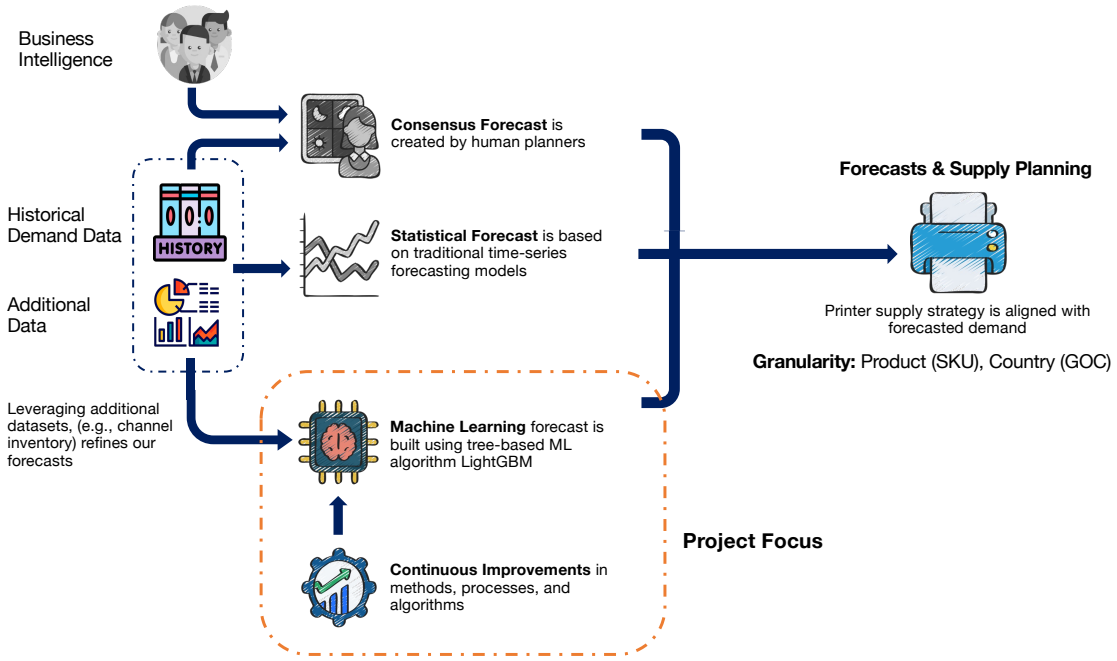

Figure 1. (Color online) Overview of the Forecasting Process

Notes. Our approach leverages historical and additional data to create robust statistical and machine learning forecasts. These forecasts are then refined by consensus planners, serving as the crucial human element in the loop, to formulate a comprehensive forecast that informs granular supply planning. The focus of this work is "ML Forecasting."

whereas large positive adjustments are more common, they tend to be less accurate. In contrast, large negative adjustments, although less frequent, are generally more precise.

To bridge these gaps and develop a unified approach, an expert group was tasked with developing and deploying an ML-based framework for demand forecasting. Armed with ML knowledge and domain expertise, the Strategic Planning and Modeling (SPaM) group at HP Inc. utilized a tree-based ML model using LightGBM for forecasting Print demand and deployed the model for forecasting at scale. Figure 1 depicts and compares the different demand forecasting solutions, where our focus is to develop the new ML forecasts as shown in the orange box.

Strategic Planning and Modeling Group (SPaM)

Formed in 1994, SPaM is a team of operations research specialists, data scientists, and external collaborators who provide internal support to HP product divisions to improve their efficiency, cost-effectiveness, and profitability (Laval et al. 2005). SPaM has developed and adapted many supply chain models for specific applications at HP (Cargille and

Branvold 2000). For example, Ward et al. (2010) documents the team's work in transforming product portfolio management: developing a new framework for screening new products using custom return-on-investment calculators, and a revenue-coverage-optimization tool to manage product variety after introduction. Similarly, Billington et al. (2004) documents how efforts from SPaM helped HP create a standard process for analyzing and designing supply chain networks.

Challenges

Implementing machine learning forecasting methods at the scale required for predicting demand for 18,000 SKUs across 170+ countries is a complex and resource-intensive endeavor, despite the accuracy improvements and efficiency they offer. Here, we outline some of the key challenges in adopting these techniques for product demand forecasting at HP. First, demand for products in different markets can be impacted by the complex interplay of various factors, such as economic conditions, seasonal trends, and regional variations. As shown in Figure 1, our approach incorporates both historical demand data and additional data sets, such as channel inventory, to refine predictions. Given the scope of the problem involving a wide range of products being sold across numerous market regions, it is a nontrivial endeavor to develop one versatile model to incorporate all the factors that can generalize well while still being tailored to individual products and regions. Second, adaptability to market fluctuations and external factors is essential for accurate predictions in the face of demand shifts, supply chain disruptions, or unforeseen events. Our continuous improvement process, depicted in Figure 1, allows us to refine our methods and algorithms to better handle these changes, but ensuring real-time adaptability remains an ongoing challenge. Third, the availability and quality of historical demand data also play a crucial role in the prediction performance (Cortes et al. 1994). Addressing data quality issues, such as inconsistent, inaccurate, outdated, and missing information, is crucial to ensure that the forecasting model is robust and reliable to support planning and operational decisions.

In addition to the three technical challenges in developing the ML models for demand forecasting, a robust *project management strategy* is pivotal for the successful deployment of such project, discussed later in detail. To achieve that, our team coordinated efforts from data scientists, production planners, and external experts, in addition to the consensus and statistical forecasting teams. By refining the model through iterative design and experimental validation, we can deliver more accurate forecasts while complementing existing statistical

and consensus forecasting methods. Our goal is to create analytical forecasts using more advanced models than simplistic statistical models, which demonstrates higher accuracy than planners' forecasts. However, it is not imperative to predict everything more accurately. As shown in Figure 1, the novelty of our system lies in its dynamic switching between ML and statistical forecasts on a product-by-product and geography-specific basis, ensuring the most accurate method is always applied across global markets. When the accuracy of the ML model outperforms the traditional statistical forecasts in terms of accuracy, it is beneficial to use ML forecasts as the basis for the consensus forecasts—that is, choosing the best analytical forecast by product and geography. Combining analytical forecasts with planner forecasts is known to increase forecast accuracy (Lawrence et al. 1986, Armstrong 2001). Our system, akin to human-in-the-loop, allows planners to automate usage of best performers as final forecasts, freeing them to focus on critical products. This collaborative approach enables all three forecasts to continue to improve in parallel while shifting the work to higher-value-add analytics as the ML forecast improves.

Rigor in our approach is demonstrated through extensive experimentation, in which we tested multiple ML algorithms (e.g., XGBoost, Random Forests) before selecting LightGBM for its superior speed and accuracy. Using Hyperopt for hyperparameter optimization, we fine-tuned the model for optimal performance across various markets, validated through rigorous back-testing over years of historical data and benchmarked against statistical and planner forecasts.

Contributions

Addressing the limitations described in the previous section in traditional demand forecasting by creating an accurate, automated ML model is a challenging yet valuable endeavor. Implementing this at HP—that is, forecasting 18,000+ products across 170 countries—provides a scalable case study for other businesses. Our key contributions include the following:

1. **Scalable ML-based forecasting framework:** We demonstrate the effectiveness of tree-based models, specifically LightGBM, in addressing enterprise-scale product demand forecasting challenges across diverse products and countries.

2. **ML operations (MLOps) for forecasting:** We stress the need for robust project management and maintenance strategies, specifically aligning with principles outlined by Curtland et al. (2022). Our framework values reproducible analysis with parameterized

notebooks and the use of advanced experiment tracking for sustaining the performance and reliability of models over time with MLFlow (Zaharia et al. 2018).

3. **Case study at HP Inc.:** Our work serves as a comprehensive guide for both practitioners and researchers attempting to tackle similar enterprise-scale forecasting challenges in other industries or contexts. This work establishes a strong foundation for further ML model improvements and operationalizing them at scale.

Literature Review

Demand forecasting models are pivotal for managing production and inventory (Gardner 1990, Kremer et al. 2016, Dodin et al. 2023). Many aspects of forecasting are well studied, especially around model learning and selection. However, details on model deployment are scant. In this section, we first provide some related methodology papers, then compare direct and iterative forecasting, discuss feature selection, and conclude with papers on the implementation of demand forecasting in organizations. A summary of works in demand forecasting like ours is provided in Table 1.

Table 1. Summary of Related Research Papers with a Focus on Demand Forecasting

Reference	Input	Model	Evaluation Metric
Dodin et al. (2023)	Lagged demands, demand statistics, seasonality components, region and month index, average age of shipped products	Improved LightGBM, Elastic Net	RMSSE
Qi et al. (2023)	Lagged demand, inventory	End-to-end Model (Dynamic Programming, RNN, MLP)	Stockout rate, turnover rate, total inventory management, holding, and stockout costs
Deng et al. (2023)	Lagged demand, inventory, among others	DeepAR, N-BEATS, Prophet	WMAPE
Makridakis et al. (2018)	M-3 data	MLP, BNN, RBF, GRNN, KNN, CART, SVR, GP, RNN, LSTM, SES, ETS	sMAPE, MASE
Sagaert et al. (2018)	Lagged demand, macroeconomic indicators	LASSO Regression	MAPE
Hamzaçebi et al. (2009)	Lagged demand	Artificial Neural Networks (ANN)	SAE, SSE
Marcellino et al. (2006)	Lagged demand	Linear models	MSFE
Gardner (1990)	Lagged demand	Exponential-smoothing Model (ETS)	Investment and delay time

Forecasting Models

The methodology frameworks for demand forecasting have significantly evolved over the last few decades. There is a large body of literature on demand pattern recognition and prediction. Traditionally, classic time-series models such as AR, MA, ARMA, ARIMA, and ETS were used for demand forecasting tasks, which only use lagged demands as the input (Hyndman and Athanasopoulos 2018). Today, ML models can accommodate nonlinearity and handle a broader range of inputs, such as unstructured and high-dimensional data of various types. In recent years, we have seen huge potential of ML algorithms in demand forecasting tasks because of their better data-fitting capabilities. Some recent works that are similar to our goals are as follows: Deng et al. (2023) outlined a comprehensive omnichannel retail infrastructure by Alibaba, which was a 2022 Edelman Award finalist. The infrastructure integrates demand forecasting with inventory management and price optimization, driven by product recommendations. Their implementation leverages deep learning models like DeepAR (Salinas et al. 2017), Prophet (Taylor and Letham 2018), Wavenet (Oord et al. 2016), and N-BEATS (Oreshkin et al. 2019) to generate demand forecasts. Dodin et al. (2023) showcased a pragmatic application of LightGBM models in forecasting the demand of parts at Bombardier. Ferreira et al. (2016) used a regression tree–based model for demand forecasting in the pipeline for price optimization.

The Makridakis (M-series) competition has been a key test bed for evaluating different forecasting models, such as multilayer perceptron, Bayesian neural networks, radial basis functions, generalized regression neural networks (also called kernel regression), K-nearest neighbor regression, classification and regression trees, support vector regression, and Gaussian processes (Makridakis and Hibon 2000; Ahmed et al. 2010; Makridakis et al. 2018, 2021). LightGBM (Ke et al. 2017), which is an advanced tree-based model, is notable for its fast and efficient training and prediction and was used by all of the top-50 performers in the M-5 competition (Makridakis et al. 2022). LightGBM's accuracy has been validated by several other research studies for predictive modeling (Bandara et al. 2020, Zhang et al. 2020, Deng et al. 2021). Motivated by these studies, results from M-5 competition, and our own experiments, we adopted the LightGBM algorithm for our task.

Direct vs. Iterative Forecasting

Conventionally, two methods exist for regression-based time series prediction: (i) direct and (ii) iterated forecasting method. The direct method uses separate models for each forecast

horizon, whereas the iterated method predicts the next period and uses that estimate for subsequent forecasts. The choice between methods involves a bias-variance trade-off and depends on the unknown population projection (Findley 1983). Theoretically, the direct method yields a lower mean squared error, but its superiority in practice is not guaranteed (McElroy 2015). Empirical evidence in literature is conflicting: Marcellino et al. (2006) found the iterative method superior for long-lag specifications and longer horizons, whereas Hamzaçebi et al. (2009) observed better performance with the direct method using artificial neural networks. For more related works, we refer interested readers to these publications' literature reviews. With experimentation, we discovered the superiority of the iterated method in our case and thus use forecasted demand as a lagged input for subsequent predictions.

Feature Selection

Incorporating additional data into ML-based forecasting models is beneficial to improve forecasting performance. For instance, Sagaert et al. (2018) leverage a broad set of macroeconomic indicators from the Federal Reserve Economic Data (FRED) in a LASSO regression model to improve tactical forecasting accuracy. In supply chain, private data create information asymmetry; lack of information sharing hinders the ability to adequately harmonize manufacturers' activities to align with customers (Simatupang and Sridharan 2002). Information shared by suppliers and customers can also improve accuracy of demand forecasting. Hartzel and Wood (2017) show that demand forecasts benefit heavily from point-of-sale reporting. Kurtuluş et al. (2012) show that such forecast (called "collaborative forecast") can be helpful for customers as well as suppliers, depending on the contractual obligations of both parties. Under the Newsvendor model setting, Taylor and Xiao (2010) show that the manufacturer benefits from selling to a better-forecasting retailer if and only if the retailer is already a good forecaster. These studies guide us to use demand and inventory information reported by our supply chain partners as part of the input to our forecasting model to further improve the forecasting performance.

Although some studies report on ML-based implementation of demand forecasting models in companies (Ferreira et al. 2016, Dodin et al. 2023), there are few detailed discussions on project management, deployment pipelines, and continuous performance monitoring, specifically in the domain of demand forecasting. Based on our previous works (Curtland et al. 2022) and concepts of MLOps (Zaharia et al. 2018), our work will share generalizable

lessons on management strategies for the enterprise-scale implementation of ML-based demand forecasting. We believe our project management strategy will be valuable to readers, as such issues are nontrivial in practice.

Project Management Strategy

Large-scale projects with numerous collaborators and users necessitate robust coordination and maintenance tools. To enhance value creation and streamline the entire ML project life cycle, data scientists and managers repurposed several DevOps (Development and Operations) concepts (Chen et al. 2020) as MLOps. As outlined by John et al. (2021), the MLOps framework proves indispensable for tracking of data for ML development, validation of ML models, release of ML models, and storage of serialized models for replication and future applications.

Most ML enhancements are driven by *experimentation*. This involves exploring multiple datasets, variable transformations, model architectures, software libraries, and more. These experiments not only have diverse inputs and outputs but must also be efficiently timed. Given the reliance of model performance on input data and training, *reproducibility* becomes paramount. In our project, before each month starts, one model gets selected for *deployment* and producing ML forecasts to support operational decisions. Yet, experimentation persists to further refine our models for future months. We show our project management strategy through a flowchart in Figure 2.

We adopted various open-source tools in our project management strategy:

1. **Experimentation and reproducibility:** *MLflow* is an open-source ML platform (Zaharia et al. 2018) that tackles challenges linked to experimentation, reproducibility, and deployment. It provides extensive experiment tracking, covering parameters, metrics, code, and data, which are accessible through an API and an interactive dashboard. We opted for MLflow because of its self-hosting capabilities, which streamlined our workflow at no additional cost to HP.

2. **Documenting results:** Jupyter Notebooks aid reproducibility, allowing detailed annotations on processes, inputs, and outputs using markdown cells. These notebooks can be parameterized, turning their execution into function calls with the `papermill` library, a tool that enables operating one notebook from another notebook similar to a function call. We incorporated a keyword, such as "Monkey," into our codebase to facilitate quick navigation for necessary adjustments before rerunning routine scripts. This approach

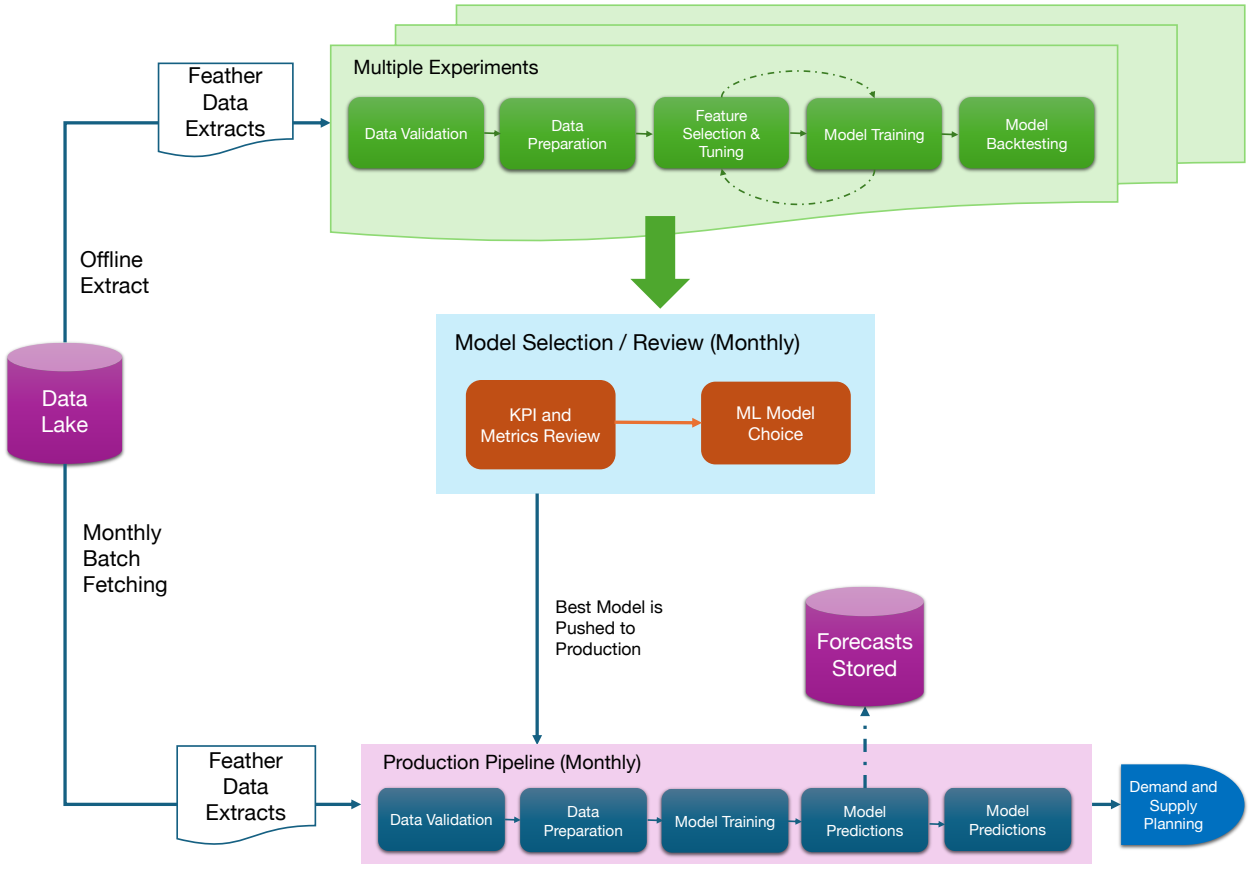

Figure 2. (Color online) Project Management for Continuous Deployment Pipeline of Our ML Forecasting Efforts

simplifies the identification of key areas for updates, making recurring tasks like monthly time-series forecasting more efficient and automated. By combining this with parameterized notebooks, we were able to expedite the early stages of coding and experimentation. Once operational workflows moved to production, these processes were fully automated using the same parameterized notebooks.

3. **Model serialization:** Once we move experimental models to production, serializing and storing saved models and parameters as artifacts for future reference becomes essential, which is where MLflow becomes indispensable again. The serialized models can be used later for warm-starting future training, which reduces computational time and effort. They can also be used for comparing accuracy between experimentation and production. Furthermore, results from the serialized models can be reproduced when necessary.

4. **Data storage:** Data storage demanded substantial disk space. Initially, we used Python's Pickle for data snapshots. However, because of fundamental issues with Pickle, such as corruption from version changes and ballooning file sizes, we transitioned to Apache

Feather. Feather boasts a powerful compression algorithm resulting in vastly reduced file sizes compared with CSV and native compatibility with libraries like `pandas`. Crucially, Feather maintains forward and backward compatibility, ensuring hassle-free file accessibility across versions.

5. **Rapid testing with `FLAML`:** Our expansive data set made it impractical to run full experiments each time. Thus, preliminary assessments were vital. We relied on `FLAML`'s efficient search and evaluation mechanisms, leveraging its automatic Bayesian hyperparameter search and cross-validation. It is designed to minimize computational costs while gradually transitioning from cheap, inaccurate trials to more expensive, accurate ones by iteratively optimizing learner selection, hyperparameters, and sample size. This allowed for targeted improvements within our time budgets, ensuring only the most promising strategies proceeded to in-depth testing. Such rapid tests were foundational; more exhaustive experimentation followed once a direction was determined, culminating in integrating findings into our primary model, as presented in our *Iterative Forecasting Algorithm* (Algorithm A.1).

Problem Formulation and Methodology

We now describe our problem setting and solution methodology. Our work addresses the problem of predicting demand for a product p in a specific country c at time t. Given a data set of historical demand data among others, our goal is to train a model that can forecast the demand for future time periods. The historical data include information about the actual demand $y_{t,c,p}$ and a set of associated features $\mathbf{X}_{t,c,p}$. These features represent various aspects of the time, market, and product, as well as lagged demand for up to 15 months before the forecasting month. Complete model formulation is provided in the appendix. Details of the model inputs are provided in Table 2 later in this section.

To select our model, we rigorously evaluated many algorithms, including XGBoost, LightGBM, Prophet, ARIMAX, ETS, and multilayer perceptrons, using the Python `darts` library for a unified and methodologically consistent comparisons (Herzen et al. 2022). Our empirical evaluations, emphasizing predictive accuracy and computational efficiency, demonstrated clear superiority of tree-based models, specifically LightGBM (Ke et al. 2017). These models excel at capturing nonlinear relationships and complex data structures, making them effective for demand forecasting, while offering interpretability that outperforms other algorithms, with straightforward parameter optimization and a reduced memory

footprint that simplify generalization and expedite training at scale. Results from the M-5 competition discussed previously affirmed our choice of LightGBM because of its proven effectiveness with data sets similar to ours in structure and complexity.

Iterative Forecasting Algorithm

Our *Iterative Forecasting Algorithm* is outlined in Algorithm A.1 in the appendix. It uses the LightGBM model as its core predictive engine, although it is adaptable to other algorithms. The model begins by preprocessing the data, which includes data cleaning and feature engineering. It is designed to forecast demand iteratively over a time window T, which allows for dynamically updating forecasts. We optimize our LightGBM model's hyperparameters using Hyperopt, a library that efficiently explores both discrete and continuous parameter spaces (Bergstra et al. 2013). Typical hyperparameters and their suggested ranges are provided in the appendix of this paper.

Using the last month's data for validation, we use Hyperopt's Tree of Parzen Estimators (TPE) algorithm to navigate this parameter space. This Bayesian hyperparameter optimization allows for faster convergence to optimal configurations by focusing on hyperparameter values that maximize performance on the validation set. By leveraging Hyperopt's capabilities, we can balance exploration of the search space with the exploitation of promising configurations, ensuring our LightGBM model is finely tuned for optimal performance.

Our approach allows us to capture both the seasonality and trends in the demand while benefiting from the efficiency and scalability of LightGBM. Moreover, the iterative nature of this algorithm allows for frequent model updating, leveraging the most recent one-month data for cross-validation. This ensures that the model stays responsive to any significant changes in the underlying data patterns. Storing the serialized model in MLFlow, we are able to ensure repeatability and continuity for future efforts, detailed previously in the *Project Management Strategy* section.

Model Input Features

Our ML models surpass conventional time-series approaches by integrating diverse features—categorical, numeric, and beyond—that not only capture historical demand but also illuminate the complex dynamics of demand generation and fulfillment. These features are listed in the following section and summarized in Table 2 for ease of reference.

Types of Features.

1. *Lag demands:* Demands from the previous m months are factored in, with $m = 15$ for products with intermittent demand and annual buying cycles.

2. *Rolling demand features:* These are statistical measures—mean, coefficient of variation, and outlier counts—computed over rolling windows of 3, 6, and 12 months, capturing both recency and variability in demand.

3. *Product- and geography-based statistics:* Summary statistics are categorized by product and geography to model unique trends and attributes within these dimensions.

4. *Seasonal fluctuations:* Binary indicators for each fiscal quarter are included to capture seasonal demand patterns. A monthly integer representing month of the quarter is also included.

5. *Product life cycle:* Calculated as $(M - m)/M$, where M is the total expected lifetime of product and m is the current forecasting month, this feature considers a product's remaining lifespan, enriching the model's temporal context. Typically, products introduced to the market experience a surge in demand initially, attributable to their innovative features and promotional efforts, followed by a gradual decline in sales as they progress through their product life cycle.

6. *Channel metrics:* Features such as "channel partner inventory" and "sell-through" provide a nuanced understanding of real-time market demand and potential future orders with direct inputs from our distribution channel partners (customers in a B2B setting). Channel partner inventory refers to the SKU-level inventory that our channel partners report monthly, and sell-through represents the sales by our partners to their customers.

Feature Selection. We considered two algorithms for our feature selection strategy: the Fast AI method based on Howard (2019) and the Quadratic Programming Feature Selection (QPFS) technique as proposed by Rodriguez-Lujan et al. (2010). Fast AI's approach involves generating a correlation matrix followed by a dendrogram of all features. This guides the systematic pruning of correlated features, thus honing the feature set down to those that are most informative. QPFS, on the other hand, uses quadratic programming to balance feature importance against redundancy. From our comparative analysis between these methods, we discovered that QPFS produced high variance in each cycle's feature importance results, whereas Fast AI method led to a stable set of features. Given this, we chose the Fast AI method for our production code.

Table 2. The Forecasting Model Incorporates over 100 Input Features, Including Various Calculated Statistics

Feature name	Description	Granularity	Utility for forecasting
Lagged demand	Size of demand from previous m months; m varies per product group	Month (t)	Captures influence of past on future trends
Rolling demand features	Statistics of demands within n-month rolling window (mean, coefficient of variation, outliers)	Month (t)	Assesses recent trend, variability
Product-based statistics	Mean and coefficient of variation of lagged demand and rolling features, per product category	SKU (p)	Specific trends in product categories
Geography-based statistics	Mean and coefficient of variation of lagged demand and rolling features, per country	Country (c)	Location-specific trends
Seasonal fluctuation	Binary indicator for each fiscal quarter and integer month within a quarter	Month (t)	Captures seasonal effects
Product life cycle	Proportion of product life cycle left, calculated as $(M-m)/M$	SKU, country (p,c)	Stage of the product in its life cycle
Channel partner inventory	Inventory reported by distribution channel partners	SKU, country, month (p,c,t)	Indicates potential reordering
Sell-through	Sales to distribution channel partners	SKU, country, month (p,c,t)	Reflection of downstream demand (to channel partners' customers)

Performance Evaluation and Results

The ultimate adoption of a new ML forecasting pipeline hinges on its accuracy. We validate the performance of ML-based forecasts' performance against existing statistical and consensus forecasts, serving two critical purposes. First, before enterprise-wide deployment across products and geographies, we must demonstrate that the ML pipeline's accuracy and reliability meet or exceed the accuracy and reliability of current methods. Second, we must also evaluate the judicious use of the additional project management machinery, which requires significant investment (see Figure 2). Successfully achieving the first goal justifies the allocation of these additional resources.

Evaluation Metrics

We use three key metrics to evaluate our forecasts: bias, weighted mean absolute percentage error (wMAPE), and root-mean-squared error (RMSE), all defined in the appendix. RMSE, our preferred metric for ML model training, is symmetric and continuously differentiable. It balances sensitivity to larger errors with scale dependency, making it valuable for emphasizing significant deviations. However, because of RMSE's sensitivity to outliers, models trained with this metric may prioritize minimizing larger errors, which in our case, has occasionally resulted in underforecasting. Planners and managers primarily use bias

and wMAPE as key performance indicators (KPIs) because of their ease of interpretation and actionability. For a comprehensive comparison of these and other accuracy metrics, including their application in M-3 forecasting, we refer readers to Hyndman and Koehler (2006).

These metrics are calculated over a specified number of months, denoted as CM-k, where k represents the number of months. For a given month t, the k-month cumulative actuals are calculated as $\sum_{i=t}^{t+k-1} y_i$, whereas the cumulative forecasts are $\sum_{i=t}^{t+k-1} \hat{y}_i$. For example, three-month cumulative forecast (CM3) starting in January would sum the forecasts for January, February, and March. The choice of cumulative forecast horizons depends on specific supply chain lengths and decision-making requirements. Measuring and improving the forecast over different lead time horizons is important for practical business reasons. Supply chains have specific lead times for manufacturing and shipping products, and businesses maintain inventory close to customers to manage demand variability during these periods. CM1, CM3, and CM6 forecasts are commonly reported, with CM3 often being the most critical because of its alignment with the typical three-month production lead time. On the other hand, CM1 provides immediate feedback on short-term operations, whereas CM6 offers a longer-term outlook. Cumulative forecasts are preferred over point forecasts also because they more effectively manage lead time variability. In an optimized supply chain, this approach allows for better inventory pooling and more accurate adjustment of factory capacity based on appropriate lead times and forecast performance.

Results

We present forecasting performance for a select business segment (1,484 products) from all three methods: consensus (ConsFcst), statistical (StatFcst), and ML (MLFcst), evaluated at cumulative horizons of one (CM1), three (CM3), and six (CM6) months. Although the scales have been adjusted for anonymity, the observed trends remain the same. Results from all product lines are not presented because of data sensitivity, and accuracy results vary across business segments.

A summary of accuracy results is provided in Table 3. These metrics are also presented as a dumbbell plot in Figure 3 with center points being 12-month averages and whiskers indicating one standard deviation. Additionally, Figure 4 visualizes these metrics over all 12 months, highlighting the monthly accuracy trends for each method. Finally, a statistical comparison of metrics over 12 months using paired t-test is presented in Table 4.

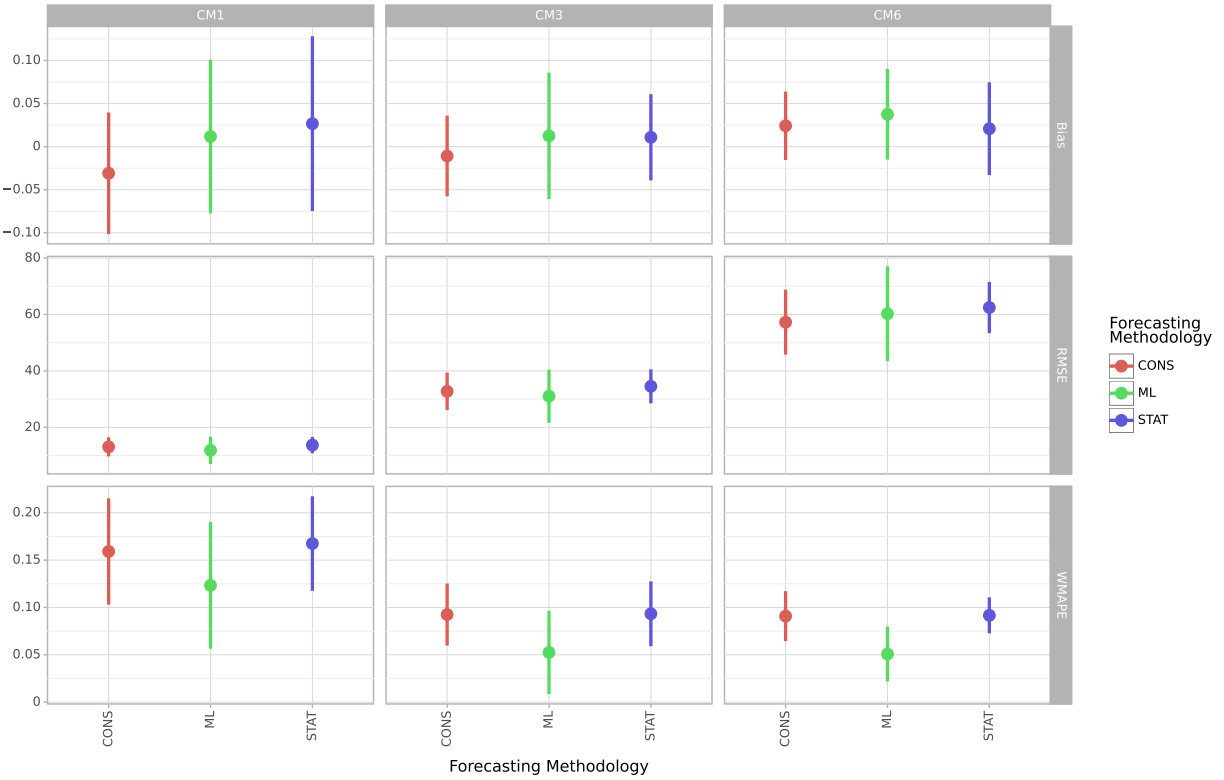

Figure 3. (Color online) Dumbbell Plot Visualizing the Mean (Center Point) and One Standard Deviation (Vertical Lines) of Bias, RMSE, and wMAPE for Three Forecasting Methods (Consensus, Machine Learning, and Statistical) over Cumulative Forecast Horizons of One Month (CM1), Three Months (CM3), and Six Months (CM6)

Note. Each color represents a different forecasting method, illustrating the variability and central tendency of the forecast accuracy metrics across different periods.

Table 3. Forecasting Accuracy Metrics (Bias, RMSE, and wMAPE) for Cumulative Forecast Horizons (CM1, CM3, CM6) with Mean (Standard Deviation)

Model	CM1			CM3			CM6		
Metric	Bias	RMSE	wMAPE	Bias	RMSE	wMAPE	Bias	RMSE	wMAPE
Consensus	−3.08%	13.09	15.92%	−1.08%	32.76	9.25%	2.42%	57.29	9.08%
	(7.05%)	(3.38)	(5.62%)	(4.68%)	(6.66)	(3.28%)	(3.96%)	(11.51)	(2.64%)
ML	1.17%	11.87	12.33%	1.25%	31.03	5.25%	3.75%	60.28	5.08%
	(8.92%)	(4.87)	(6.69%)	(7.34%)	(9.43)	(4.39%)	(5.26%)	(16.83)	(2.91%)
Statistical	2.67%	13.71	16.75%	1.08%	34.55	9.33%	2.08%	62.47	9.17%
	(10.14%)	(2.89)	(4.99%)	(5.00%)	(6.03)	(3.42%)	(5.38%)	(9.08)	(1.90%)

The ML forecast method demonstrates considerable strengths in its forecasting accuracy as compared with the statistical method, particularly in the metrics of wMAPE and RMSE. We observe that wMAPE for ML forecast is better than the other two in all three cumulative periods. In fact, at CM3 and CM6 (i.e., for longer-range forecasts), our model has a

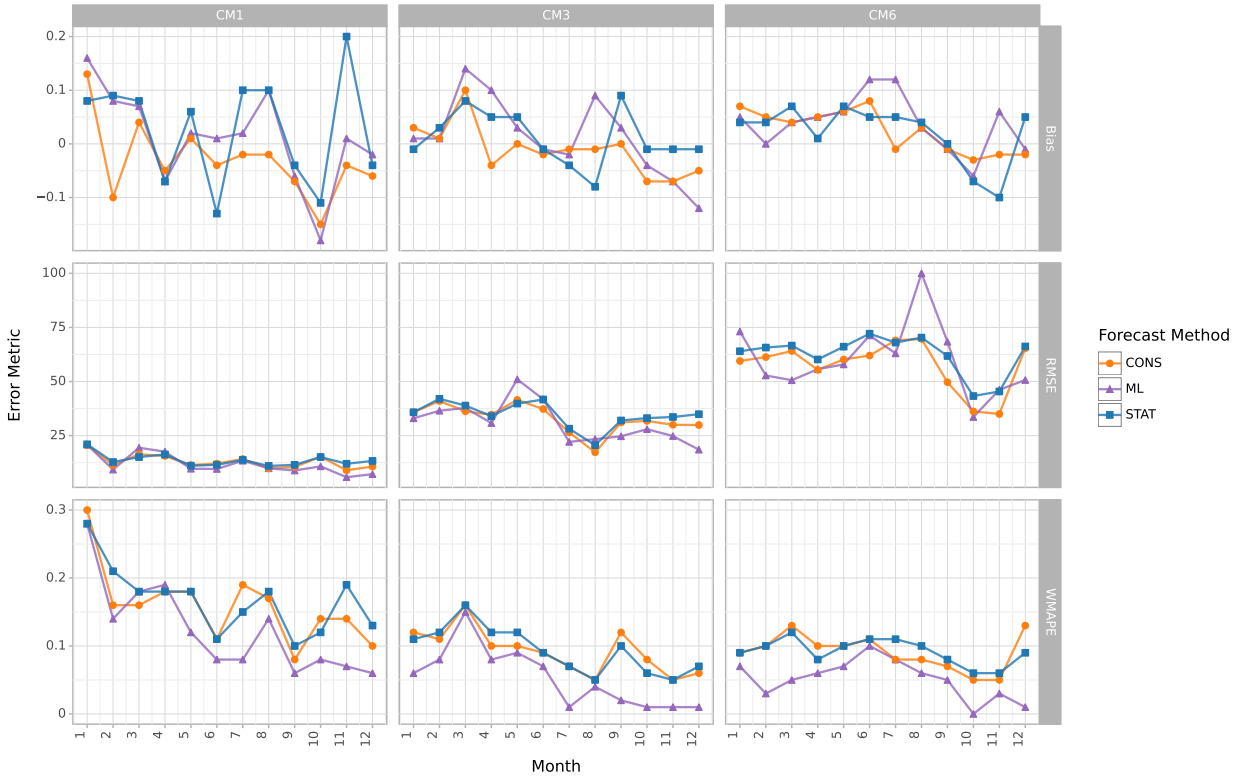

Figure 4. (Color online) Bias, WMAPE, and RMSE Metrics over 12 Months Show That ML Model Is Consistently Among the Top Performers of the Three Models

Note. CM1 is point forecast, whereas CM3 and CM6 are three- and six-month cumulative forecasts, respectively.

Table 4. Forecasting Accuracy Metrics: Bias, wMAPE, and RMSE Comparison for CONS, ML, and STAT Methods

Cumulative	Comparison	Bias	RMSE	WMAPE
CM1	CONS vs ML	−1.295 (0.209)	0.716 (0.482)	1.421 (0.169)
	STAT vs ML	0.385 (0.704)	1.128 (0.272)	1.832 (0.080)
CM3	CONS vs ML	−0.929 (0.363)	0.518 (0.610)	2.528 (0.019)
	STAT vs ML	−0.065 (0.949)	1.089 (0.288)	2.541 (0.019)
CM6	CONS vs ML	−0.701 (0.490)	−0.507 (0.617)	3.526 (0.002)
	STAT vs ML	−0.767 (0.451)	0.399 (0.694)	4.074 (0.001)

Note. The accompanying table presents t-statistics and p-values (in parentheses) for an in-depth assessment across various cumulative forecast horizons.

wMAPE almost half of the wMAPE of the other two methods. When looking at statistically significant differences, we find statistically significant difference between ML and STAT models with positive t-statistic and p-values less than 0.05. These findings strongly suggest the statistical superiority of the ML forecast in wMAPE, further demonstrating the model's alignment with HP's business objectives, as wMAPE is a business KPI. The higher accuracy of the ML model in wMAPE is particularly surprising because it was trained with RMSE

as the loss function. In the case of RMSE, which is sensitive to large forecast errors, the ML forecast again proves to be more adept than others, although not statistically significant.

However, the ML forecast does not consistently dominate across all metrics and comparisons. When considering bias, which reflects the systematic error in forecasts (either as overestimation or underestimation), the ML method does not exhibit a statistically significant difference from the statistical or consensus forecasts in any of the cumulative periods (CM1, CM3, and CM6), as evidenced by p-values greater than 0.05. Our model exhibits higher bias compared with the consensus and statistical models. We observed a strong tendency for the ML model to underforecast, particularly over longer time horizons. This issue appears to be influenced by the intermittent demand of many products, where the model occasionally learns to forecast zero incorrectly. Although this may explain the underforecasting, further investigation is required to definitively identify the root cause.

These results suggest that, in certain scenarios—particularly those involving longer-term predictions—the Consensus forecast may provide more accurate outcomes than our method. This contrast underscores the ML forecast's strengths in specific contexts, guiding the modeling team in targeting improvements and enabling the business team to select the best-performing model for each product and country. By acting as a "human in the loop," the business team plays a crucial role in validating and verifying forecasts generated by the automated model. The data in Table 4 and the trends in Figure 4 collectively bolster the case for adopting the ML model alongside the statistical and consensus models at HP, contributing to an integrated effort aimed at improving overall forecasting performance.

Dashboard of Results

As an essential advancement in disseminating forecasting analytics, the incorporation of analytical dashboards facilitates sharing results with a wider audience, including planners and decision makers. This powerful tool not only exhibits the performance of various models but also provides an avenue for scrutinizing their historical accuracies and pertinent details. Constructed with customizable KPIs, the dashboard extends the ability to inspect product hierarchies from different lenses, thereby promoting informed business strategies and policies. The dashboard presents bias, wMAPE, and RMSE to compare historical performance of algorithms. To assist planners in selecting the best model, we create a heat map of best forecast as measured by wMAPE. This heat map covers all HP Print product categories by time period, enabling planners to visualize the relative performance of different methods over time.

Lessons Learned and Business Implications

Implementing a global-scale ML-based demand forecasting system at HP revealed critical lessons and challenges, transforming the way forecasts are integrated into business processes. Work on the project started in 2019 and remained in pilot phase for a year. In 2020, the results were published in the standard KPI dashboards and available for manual use within the Statistical and Consensus forecasting modeling processes depicted in Figure 1. By summer 2023, SKU-level forecasts for all Print products across geographies were fully integrated into the data pipeline for business forecasting. Inclusion of the analytical forecast in the business KPI dashboards led to wide-scale adoption of our work.

While previous efforts to implement ML-based models were scattered and unsuccessful, our solution was implemented at scale because of its inclusive approach. First, our solution performed well across HP's Print portfolio. Second, early collaboration with the business team helped preemptively address change management challenges that often hinder large-scale projects. A unified, holistic approach that used only one model architecture with various data sources proved essential in overcoming the complexities of enterprise-scale forecasting. Data quality is critical in machine learning, and HP's simultaneous digital transformation presented both challenges and opportunities. Key issues, such as missing data and unlinked data sets, required meticulous reconciliation to ensure the accuracy and reliability of our forecasts. A key challenge was the integration of "soft data," which involved manual analysis. Successfully overcoming these hurdles was key to improving the accuracy and reliability of our forecasting models.

MLOps, enabled by MLFlow, streamlined the development life cycle by automating key processes, facilitating experiment tracking, and ensuring consistent deployment of the latest and most accurate models. This enabled linking analysis to outputs and ensured comprehensive documentation. Because of the scale of implementation, computational resource optimization became necessary. This involved using high-performance workstations and adopting efficient data storage and retrieval methods, like the Apache Feather format, which provided significant improvements in data handling and processing efficiency.

Adopting LightGBM was crucial for its ability to handle large data sets and complex patterns at speed. It demonstrated remarkable adaptability in handling market fluctuations, including during the pandemic's supply-constrained environment. Incorporating our signal

into the standard business process in conjunction with effective dashboard visualizations and KPIs enabled successful implementation.

An important aspect of our solution is that our final forecast is not confined to being either human- or machine-produced. Our human-in-the-loop architecture ensures that human forecasters can apply contextual knowledge, adjusting plans independent of modeling when necessary to reflect market nuances. This synergy between machine precision and human insight has improved forecast accuracy and decision making. As of August 2024, some downstream users are also using our ML forecasts for ensembling with their own forecasts. Over time, we expect our solution to improve by supporting the human with prescriptive drift and anomaly detection, along with AI-enhanced dashboards. This will build upon the existing explainability and causality capabilities of the solution, creating better insight generation and enhancing model itself.

In summary, our experience at HP underscores the importance of a well-integrated, adaptive, ML-based approach in demand forecasting. Addressing these challenges was pivotal in optimizing the ML models for supply chain management, leading to more efficient decision making and operational management. The insights gleaned offer a valuable template for business leaders facing similar challenges in large-scale demand forecasting. Our collaborative, agile development model is expected to further improve accuracy as we implement our backlog of modeling ideas.

Concluding Remarks

In this paper, we detailed our implementation of an ML-based demand forecasting system at HP, implemented for all Print products worldwide. Our iterative forecasting algorithm and project management strategy are modular and adaptable to different industries in which demand forecasting and supply chain optimization are crucial. Our approach, combining machine learning with a human-in-the-loop framework, presents a novel and scalable solution to demand forecasting challenges that traditional time-series models could not adequately address. Our work improved forecast accuracy and provided a robust system capable of adapting to market fluctuations and supply chain disruptions. Key insights highlight the importance of computational resources, robust data management, and a proactive stance toward market changes and data quality. Downstream forecasters are now incorporating our ML forecasts into their ensemble models.

Our work at HP has resulted in tangible improvements in supply chain management and inventory optimization, reducing forecast errors and leading to cost savings and more accurate production planning. Our experience serves as a blueprint for other companies in the technology and manufacturing sectors facing similar challenges in demand forecasting. This approach not only facilitates more accurate demand predictions but also fosters an agile and responsive business environment.

Acknowledgments

The authors acknowledge Barrett Crane, former Director of SPaM, for his leadership in championing this project at HP. The authors thank Steve Radosevich and Alok Kumar for their facilitation and organizational support. The authors also thank Kevin Kacmarynski for assistance with periodic execution of production models and Karen Nguyen for continuing to extend their work. The authors thank Dr. Caroline Johnston, their colleague during the Summer 2022 internship at SPaM, for her valuable collaboration and insights. Finally, the authors thank the two anonymous reviewers, whose thorough feedback and constructive suggestions substantially improved the quality of the manuscript.

Appendix

Problem Formulation

We formulate the problem as a supervised learning task in which we aim to minimize the forecasting loss over the data set D, consisting of pairs of input features $\mathbf{X}_{t,c,p}$ and corresponding demands values $y_{t+1,c,p}$. That is,

$$\mathcal{D} = \{(\mathbf{X}_{t,c,p}, y_{t+1,c,p}) : \forall c, p, t_{first} \leq t < t_{now}\}, \tag{A.1}$$

where t_{first} is the first period when we have enough observations to create all features, especially the lagged features. The training process minimizes the forecasting loss (RMSE):

$$\ell(f|\mathcal{D}) = \sqrt{\mathbb{E}_{X,y\in\mathcal{D}}\left(f(X) - y\right)^2}, \tag{A.2}$$

in addition to necessary regularization terms.

In this context, our model $f(\cdot)$ learns to predict future demand based on the input features. Once trained, the model can be applied to forecast demand for future time periods $t \geq t_{now}$.

We use $\mathbf{F}_{t,c,p} \in \mathbb{R}^T$ to represent forecasts for T periods starting with t_{now}:

$$\mathbf{F}^T_{t,c,p} = (\hat{y}_{t+1,c,p}, \cdots, \hat{y}_{t+T,c,p}). \tag{A.3}$$

Iterative Forecasting Algorithm

Here, we describe our iterative forecasting algorithm. For each time step t_α, the algorithm constructs a training data set $\mathcal{D}_\alpha$ using all available data up to that point in time. Identified hyperparameters are used with $\mathcal{D}_\alpha$ to train the LightGBM model $f(\cdot)$, which is optimized to minimize the RMSE. Once trained, the model generates T future forecasts for each time step t_α. The LightGBM model is then either incrementally updated (i.e., warm-started from best results from last month) or retrained from scratch, providing flexibility in handling significant changes in underlying data distribution.

Optimizing LightGBM Hyperparameters with Hyperopt

Hyperopt (https://github.com/hyperopt/hyperopt) is a powerful Python library for hyperparameter optimization, supporting several machine learning models, including LightGBM. In this work, we leverage Hyperopt to fine-tune key parameters of our LightGBM model for enhanced performance. Specifically, we tune key parameters including

1. **Learning rate:** controls how much to adjust the model with each step, with a range between 0.1 and 1;
2. **Maximum tree depth:** dictates the maximum depth of each decision tree, explored between 10 and 100;
3. **Regularization parameters:** L_1 and L_2 regularization terms help prevent overfitting, with values explored between 0 and 1;
4. **Minimum child weight:** specifies the minimum sum of instance weights needed in a child, ranging from 1 to 50;
5. **Subsample and column-sample proportion:** controls the fraction of samples and features used per tree, ranging from 0.5 to 1.

Algorithm A.1 Enhanced Training and Forecasting Algorithm with LightGBM

1: **Preprocess the data:** Data cleaning and feature engineering.

2: **Determine optimal hyperparameters:** Use grid search or random search for the LightGBM model.

3: **Initialize forecast horizon** T (e.g., 7).

4: **for** t_α in $(t_{\text{first}} : t_{\text{now}})$ **do**

5: Create the training data:

$$\mathcal{D}_\alpha = \{(\mathbf{X}_{t,c,p}, y_{t,c,p}) : \forall c, p, t_{\text{first}} \le t \le t_\alpha\}$$

6: Perform time-series cross-validation on $\mathcal{D}_\alpha$ and train the LightGBM model $f(\cdot)$ with optimal hyperparameters, minimizing loss (RMSE):

$$\ell(f \mid \mathcal{D}) = \sqrt{\mathbb{E}_{\mathbf{X},\, y \in \mathcal{D}} \big(f(\mathbf{X}) - y\big)^2}$$

7: With the fitted model, create T forecasts for $t_\alpha + 1$ to $t_\alpha + T$:

$$\mathbf{F}^T_{t_\alpha, c, p} = \Big(f(\hat{\mathbf{X}}_{t_\alpha+1,c,p}), f(\hat{\mathbf{X}}_{t_\alpha+2,c,p}), \cdots, f(\hat{\mathbf{X}}_{t_\alpha+T,c,p})\Big)$$

8: Update the LightGBM model incrementally by warm starting from last month's best results if possible, or retrain it from scratch.

9: **end for**

10: **Perform backtesting:** Apply the trained model to a historical data set $\mathcal{D}_{\text{historical}}$ to simulate past predictions. Evaluate its performance using appropriate metrics (e.g., RMSE).

11: **Store forecasts:** Save the generated forecasts $\mathbf{F}^T_{t_\alpha, c, p}$ to a dedicated database or file storage for future evaluation, comparison, or direct usage.

12: **Log model:** Serialize the LightGBM model, hyperparameters, and performance metrics for future reference or retraining using MLFlow.

Evaluation Metrics

Let n represent the number of data points; y_i, the actual value; and $\hat{y}_i$, the predicted value.

Bias measures the weighted percentage error in forecasts, signified by a positive or negative value indicating over- or underforecasting, respectively. Bias is calculated using the formula

$$Bias = \sum_{i=1}^{n} \frac{y_i - \hat{y}_i}{y_i}.$$

wMAPE represents the weighted mean of absolute percentage errors, a metric easily understood even by nontechnical stakeholders as percentage deviation from actuals. It is expressed as

$$wMAPE = \frac{\sum_{i=1}^{n} |\hat{y}_i - y_i|}{\sum_{i=1}^{n} y_i}.$$

RMSE is defined as

$$RMSE = \sqrt{\frac{\sum_{i=1}^{n} (y_i - \hat{y}_i)^2}{n}}.$$

References

Ahmed NK, Atiya AF, Gayar NE, El-Shishiny H (2010) An empirical comparison of machine learning models for time series forecasting. *Econometric Reviews* 29(5–6):594–621, ISSN 0747-4938, URL http://dx.doi.org/10.1080/07474938.2010.481556.

Armstrong JS (2001) Combining forecasts. *Principles of Forecasting*, 417–439 (Springer).

Bandara K, Bergmeir C, Smyl S (2020) Forecasting across time series databases using recurrent neural networks on groups of similar series: A clustering approach. *Expert Systems with Applications* 140:112896.

Bergstra J, Yamins D, Cox D (2013) Making a science of model search: Hyperparameter optimization in hundreds of dimensions for vision architectures. *International Conference on Machine Learning*, 115–123 (PMLR).

Billington C, Callioni G, Crane B, Ruark JD, Rapp JU, White T, Willems SP (2004) Accelerating the profitability of hewlett-packard's supply chains. *Interfaces* 34(1):59–72.

Cargille B, Branvold D (2000) Diffusing supply chain innovations at hewlett-packard company: Applications of performance technology. *Performance Improvement Quarterly* 13(4):6–15.

Chen A, Chow A, Davidson A, DCunha A, Ghodsi A, Hong SA, Konwinski A, Mewald C, Murching S, Nykodym T, et al. (2020) Developments in mlflow: A system to accelerate the machine learning lifecycle. *Proceedings of the Fourth International Workshop on Data Management for End-to-End Machine Learning*, 1–4.

Cortes C, Jackel LD, Chiang WP (1994) Limits on learning machine accuracy imposed by data quality. *Advances in Neural Information Processing Systems* 7.

Curtland C, Neto P, Ghozeil A (2022) Hp inc.. advanced analytics powers technology in the service of humanity. URL https://pubsonline.informs.org/do/10.1287/orms.2022.02.18/full/.

Deng T, Zhao Y, Wang S, Yu H (2021) Sales forecasting based on lightgbm. *2021 IEEE International Conference on Consumer Electronics and Computer Engineering (ICCECE)*, 383–386 (IEEE).

Deng Y, Zhang X, Wang T, Wang L, Zhang Y, Wang X, Zhao S, Qi Y, Yang G, Peng X (2023) Alibaba realizes millions in cost savings through integrated demand forecasting, inventory management, price optimization, and product recommendations. *INFORMS Journal on Applied Analytics* 53(1):32–46.

Dodin P, Xiao J, Adulyasak Y, Alamdari NE, Gauthier L, Grangier P, Lemaitre P, Hamilton WL (2023) Bombardier aftermarket demand forecast with machine learning. *INFORMS Journal on Applied Analytics* .

Ferreira KJ, Lee BHA, Simchi-Levi D (2016) Analytics for an online retailer: Demand forecasting and price optimization. *Manufacturing & Service Operations Management* 18(1):69–88.

Fildes R, Goodwin P, Lawrence M, Nikolopoulos K (2009) Effective forecasting and judgmental adjustments: An empirical evaluation and strategies for improvement in supply-chain planning. *International Journal of Forecasting* 25(1):3–23.

Fildes R, Ma S, Kolassa S (2022) Retail forecasting: Research and practice. *International Journal of Forecasting* 38(4):1283–1318.

Findley DF (1983) On the use of multiple models for multi-period forecasting. *Proceedings of Business and Economic Statistics, American Statistical Association*, 528–531.

Gardner ES (1990) Evaluating forecast performance in an inventory control system. *Management Science* 36(4):490–499.

Hamzaçebi C, Akay D, Kutay F (2009) Comparison of direct and iterative artificial neural network forecast approaches in multi-periodic time series forecasting. *Expert Systems with Applications* 36(2):3839–3844.

Hartzel KS, Wood CA (2017) Factors that affect the improvement of demand forecast accuracy through point-of-sale reporting. *European Journal of Operational Research* 260(1):171–182.

Herzen J, Lässig F, Piazzetta SG, Neuer T, Tafti L, Raille G, Pottelbergh TV, Pasieka M, Skrodzki A, Huguenin N, Dumonal M, Kościsz J, Bader D, Gusset F, Benheddi M, Williamson C, Kosinski M, Petrik M, Grosch G (2022) Darts: User-friendly modern machine learning for time series. *Journal of Machine Learning Research* 23(124):1–6, URL http://jmlr.org/papers/v23/21-1177.html.

Howard J (2019) Practical deep learning. Online Course.

Hyndman RJ, Athanasopoulos G (2018) *Forecasting: Principles and Practice* (OTexts).

Hyndman RJ, Koehler AB (2006) Another look at measures of forecast accuracy. *International Journal of Forecasting* 22(4):679–688.

John MM, Olsson HH, Bosch J (2021) Towards mlops: A framework and maturity model. *2021 47th Euromicro Conference on Software Engineering and Advanced Applications (SEAA)*, 1–8 (IEEE).

Ke G, Meng Q, Finley T, Wang T, Chen W, Ma W, Ye Q, Liu TY (2017) Lightgbm: A highly efficient gradient boosting decision tree. *Advances in Neural Information Processing Systems* 30:3146–3154.

Khosrowabadi N, Hoberg K, Imdahl C (2022) Evaluating human behaviour in response to ai recommendations for judgemental forecasting. *European Journal of Operational Research* 303(3):1151–1167.

Kremer M, Siemsen E, Thomas DJ (2016) The sum and its parts: Judgmental hierarchical forecasting. *Management Science* 62(9):2745–2764.

Kurtuluş M, Ülkü S, Toktay BL (2012) The value of collaborative forecasting in supply chains. *Manufacturing & Service Operations Management* 14(1):82–98.

Laval C, Feyhl M, Kakouros S (2005) Hewlett-packard combined or and expert knowledge to design its supply chains. *Interfaces* 35(3):238–247.

Lawrence MJ, Edmundson RH, O'Connor MJ (1986) The accuracy of combining judgemental and statistical forecasts. *Management Science* 32(12):1521–1532.

Lee HL (2002) Aligning supply chain strategies with product uncertainties. *California Management Review* 44(3):105–119.

Makridakis S, Hibon M (2000) The m3-competition: Results, conclusions and implications. *International Journal of Forecasting* 16(4):451–476.

Makridakis S, Spiliotis E, Assimakopoulos V (2018) Statistical and machine learning forecasting methods: Concerns and ways forward. *PLoS ONE* 13.

Makridakis S, Spiliotis E, Assimakopoulos V (2021) The m5 competition: Background, organization, and implementation. *International Journal of Forecasting* .

Makridakis S, Spiliotis E, Assimakopoulos V (2022) M5 accuracy competition: Results, findings, and conclusions. *International Journal of Forecasting* 38(4):1346–1364.

Manary MP, Wieland B, Willems SP, Kempf KG (2019) Analytics makes inventory planning a lights-out activity at intel corporation. *INFORMS Journal on Applied Analytics* 49(1):52–63.

Marcellino M, Stock JH, Watson MW (2006) A comparison of direct and iterated multistep ar methods for forecasting macroeconomic time series. *Journal of Econometrics* 135(1-2):499–526.

McElroy T (2015) When are direct multi-step and iterative forecasts identical? *Journal of Forecasting* 34(4):315–336.

Oord Avd, Dieleman S, Zen H, Simonyan K, Vinyals O, Graves A, Kalchbrenner N, Senior A, Kavukcuoglu K (2016) Wavenet: A generative model for raw audio. *arXiv preprint arXiv:1609.03499* .

Oreshkin BN, Carpov D, Chapados N, Bengio Y (2019) N-beats: Neural basis expansion analysis for interpretable time series forecasting. *arXiv preprint arXiv:1905.10437* .

Petropoulos F, Kourentzes N, Nikolopoulos K, Siemsen E (2018) Judgmental selection of forecasting models. *Journal of Operations Management* 60:34–46.

Petropoulos F, Siemsen E (2023) Forecast selection and representativeness. *Management Science* 69(5):2672–2690.

Qi M, Shi Y, Qi Y, Ma C, Yuan R, Wu D, Shen ZJ (2023) A practical end-to-end inventory management model with deep learning. *Management Science* 69(2):759–773.

Ritzman LP, King BE (1993) The relative significance of forecast errors in multistage manufacturing. *Journal of Operations Management* 11(1):51–65.

Rodriguez-Lujan I, Elkan C, Santa Cruz Fernández C, Huerta R, et al. (2010) Quadratic programming feature selection. *Journal of Machine Learning Research* .

Sagaert YR, Aghezzaf EH, Kourentzes N, Desmet B (2018) Temporal big data for tactical sales forecasting in the tire industry. *Interfaces* 48(2):121–129.

Salinas D, Flunkert V, Gasthaus J (2017) Deepar: Probabilistic forecasting with autoregressive recurrent networks. arxiv 2017. *arXiv preprint arXiv:1704.04110* .

Seifert M, Siemsen E, Hadida AL, Eisingerich AB (2015) Effective judgmental forecasting in the context of fashion products. *Journal of Operations Management* 36:33–45.

Simatupang TM, Sridharan R (2002) The collaborative supply chain. *The International Journal of Logistics Management* 13(1):15–30.

Simatupang TM, Sridharan R (2005) An integrative framework for supply chain collaboration. *The International Journal of Logistics Management* 16(2):257–274.

Taylor SJ, Letham B (2018) Forecasting at scale. *The American Statistician* 72(1):37–45.

Taylor TA, Xiao W (2010) Does a manufacturer benefit from selling to a better-forecasting retailer? *Management Science* 56(9):1584–1598.

Ward J, Zhang B, Jain S, Fry C, Olavson T, Mishal H, Amaral J, Beyer D, Brecht A, Cargille B, et al. (2010) Hp transforms product portfolio management with operations research. *Interfaces* 40(1):17–32.

Zaharia M, Chen A, Davidson A, Ghodsi A, Hong SA, Konwinski A, Murching S, Nykodym T, Ogilvie P, Parkhe M, et al. (2018) Accelerating the machine learning lifecycle with mlflow. *IEEE Data Eng. Bull.* 41(4):39–45.

Zellner M, Abbas AE, Budescu DV, Galstyan A (2021) A survey of human judgement and quantitative forecasting methods. *Royal Society Open Science* 8(2):201187.

Zhang Y, Zhu C, Wang Q (2020) Lightgbm-based model for metro passenger volume forecasting. *IET Intelligent Transport Systems* 14(13):1815–1823.

M. Harshvardhan is a PhD candidate at the Haslam College of Business at the University of Tennessee, Knoxville, advised by Dr. Chuanren Liu. He is broadly interested in developing and deploying machine learning algorithms for useful applications. During his PhD, he also worked with the Strategic Planning and Modeling (SPaM) team at HP for over a year in bringing this project to action. Harsh earned his BA and MBA at the Indian Institute of Management Indore.

Cara Curtland is a supply chain data science strategist in the Strategic Planning & Modeling (SPaM) team at HP Inc. A 28-year HP veteran, Cara is a senior advisor to executive leaders focused on enabling end-to-end processes and solutions that deliver profitable growth. Her experience spans manufacturing, R&D, planning, forecasting, supply chain design, complexity management, and inventory and cash flow optimization. Cara earned a BS and MS in industrial engineering from Purdue University.

Jerry Hwang is a data scientist with over 20 years of experience in supply chain analytics. He has contributed to demand forecasting initiatives at HP and Uber. He has also worked on projects encompassing procurement risk management, inventory optimization, and supply chain network design. He holds an MSc degree in management science and engineering from Stanford University.

Chuck VanDam is a supply chain data scientist in the Strategic Planning & Modeling (SPaM) team at HP Inc. Over the past 25 years, Chuck has held operations management, engineering, and consulting roles within HP, Verigy, Agilent Technologies, and Cisco Systems, with a focus on forecasting, planning, inventory and working capital management, network design, and complexity management. Chuck holds MBA and manufacturing systems engineering MS degrees from Stanford University.

Adam Ghozeil is a principal data scientist in the Digital and Transformation Office at HP Inc. His focus is 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.

Pedro A. Neto is a supply chain data scientist in the SPaM team at HP Inc. With >12 years' industry experience, Neto leads advanced projects in data science and supply chain analytics and creates models to facilitate complex decision-making processes and provide actionable insights. He is currently on the Board of Directors at the Association for Supply Chain Management (ASCM). He has a BS in industrial engineering and PhD in operations research and industrial engineering, both from Penn State.

Frederic Marie is a supply chain data scientist in the Strategic Planning & Modeling (SPaM) team at HP Inc. He enjoys using operations research and artificial intelligence to answer business questions. He loves to create intuitive visualizations and explore fusion cuisine. Frederic holds an advanced engineering degree in computer science, an MS in applied mathematics from Joseph Fourier University in Grenoble, and an MBA from the Wharton School of the University of Pennsylvania.

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 journals and conference proceedings, such as *IEEE Transactions on Knowledge and Data Engineering*, *INFORMS Journal on Computing*, and *European Journal of Operational Research*.