scientific reports

OPEN

A machine learning framework for long-term forecasting of spare part demand in end-of-life product scenarios

Sendhil Nathan B[1,2], B. Veera Siva Reddy[1]✉, V. S. Shenbaga Sujan[3], C. Chandrasekhara Sastry[1]✉, J. Krishnaiah[1] & Santi Jitpichitchai[4]

Accurately estimating future demand for service parts during a product's end-of-life (EOL) phase is essential for ensuring long-term support while avoiding costly overstock or shortages. This study proposes a decay-function-blended machine learning (DFB-ML) framework for forecasting spare-part demand across extended service horizons. The framework integrates historical consumption data, warranty failure rates, and attrition-adjusted fleet estimates to model lifecycle behaviour for 1,709 automotive part numbers. Feature engineering captures both short-term trends and long-term decay through variables such as lagged demand, replacement intensity, and vehicle dropout dynamics. Multiple regression models were evaluated, with Random Forest achieving the highest forecasting accuracy (Safe Mean Absolute Percentage Error = 4.36%). An ablation study confirmed that moderate decay blending ($\alpha = 0.2–0.4$) yields stable long-horizon forecasts and sub-linear error growth over the 8-year horizon. The framework was further validated for scalability within ERP/SAP-linked distributed networks, demonstrating readiness for industrial deployment. The resulting forecasts support data-driven Last-Time-Buy (LTB) decisions through part-wise procurement recommendations and risk-adjusted buffers. This approach unifies lifecycle decay modelling with machine learning and provides a generalizable blueprint for uncertainty-aware EOL inventory forecasting in engineering and supply-chain domains.

Keywords End-of-life inventory planning, Intermittent demand, Lifecycle analytics, Machine learning, Predictive maintenance, Spare part demand forecasting

In the capital goods and automotive sectors, the post-production support phase represents one of the most operationally uncertain yet legally binding stages of the product lifecycle. Manufacturers are typically obligated to ensure spare parts availability for a defined period often ten to fifteen years after vehicle production ceases[1]. The final opportunity to procure these parts is known as the last time buy (LTB). This one-time procurement decision must anticipate future service demand across an extended horizon with no scope for revision. Inaccurate estimation at this stage can lead to significant financial losses[2]. Overstocking results in high holding costs and eventual scrappage, while understocking leads to missed service obligations, reputational damage, and emergency procurement at inflated prices[3].

The Fig. 1 illustrates the typical lifecycle trajectory of production and service part demand. Production volumes rise sharply during early phases, peaking between 5 and 7 years post-SOP (Start of Production), after which manufacturing ceases (EOP). Service parts, in contrast, exhibit a lagged demand pattern that continues to decline well into the EOL (End of Life) phase, driven by attrition in the installed vehicle base and reduced replacement frequencies[4,5].

[1]Department of Mechanical Engineering, Indian Institute of Information Technology Design and Manufacturing Kurnool (IIITDMK), Kurnool 518008, Andhra Pradesh, India. [2]Ford Global Technology & Business Center, Ford Motor Private Limited, Chennai 600119, Tamil Nadu, India. [3]Department of Computer Science and Engineering, Indian Institute of Information Technology Design and Manufacturing Kurnool (IIITDMK), Kurnool 518008, Andhra Pradesh, India. [4]Ford Sales & Services (Thailand) Co., Ltd, Sathorn Square Office Tower 11 and 12 Floor, 98 North Sathorn Road, Silom, Bangrak , Bangkok 10500, Thailand. ✉email: sivareddy.bobbili@gmail.com; chandrasekhar@iiitk.ac.in

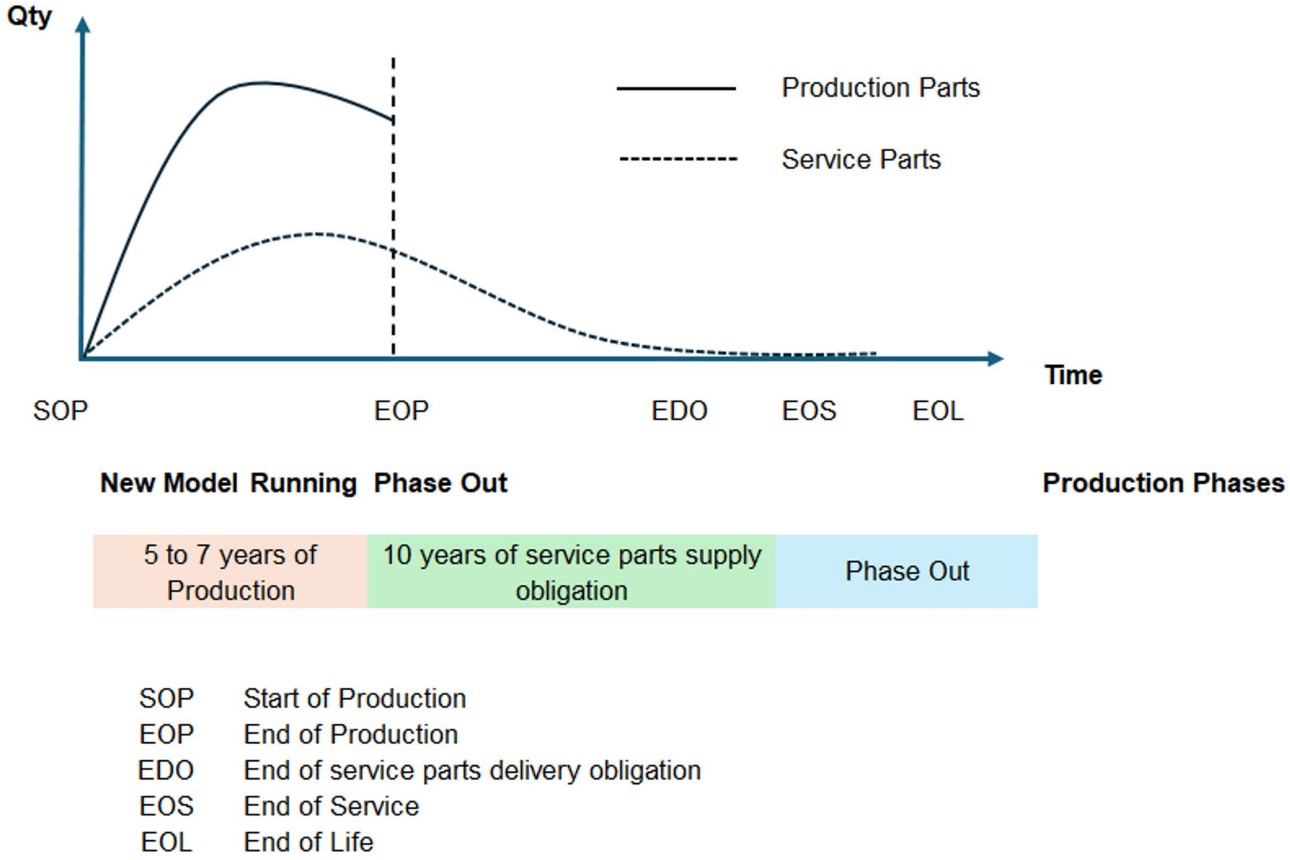

Fig. 1. Illustration of product and service part demand across the lifecycle phases from Start of Production (SOP) to End of Life (EOL).

Despite the operational importance of LTB decisions, current forecasting practices remain primitive. Most OEMs rely on rule-based multipliers or expert heuristics that extrapolate from recent consumption trends, without accounting for lifecycle attrition, platform retirements, or the stochastic behavior of field failure rates[6]. Classical time-series approaches, such as ARIMA or Holt-Winters smoothing, require stationarity and dense data, which are rarely present in post-EOP service environments. Moreover, these techniques are ill-equipped to incorporate exogenous variables such as warranty coverage, vehicle production volumes, or attrition-adjusted fleet estimates[7].

Recent advances in artificial intelligence (AI) and machine learning (ML) offer promising alternatives for modeling demand under uncertainty[8]. Tree-based ensemble models such as Random Forests, XGBoost, and CatBoost have demonstrated high accuracy in handling non-linear, high-dimensional data[9]. However, their adoption in LTB forecasting remains limited due to a lack of domain-specific feature engineering and inadequate lifecycle integration. Additionally, service part demand exhibits characteristics such as sparsity, irregularity, and delayed obsolescence that necessitate hybrid approaches combining data-driven learning with physically informed decay dynamics[10].

This study presents a novel AI/ML-powered framework for estimating eight-year service demand for EOL parts to support LTB decisions. The proposed method integrates structured feature engineering including lagged demand, active vehicle fleet size, replacement rate, and consumption slope with supervised learning models trained on temporally ordered part-year data. Lifecycle fidelity is preserved using a recursive forecasting loop augmented with exponential decay modulation to reflect real-world attrition behavior. The framework is evaluated on a dataset comprising 1709 SKUs from a global automotive OEM covering the period 2015–2024. The best-performing model, Random Forest, achieved a Safe MAPE of 4.36%, outperforming traditional linear baselines (e.g., ElasticNet with Safe MAPE > 470%).

Forecasts were visualized and aggregated into LTB recommendations, including uncertainty buffers based on standard deviation of historical error. For instance, part SPN01330 was predicted to require 2,509 units over the next decade, which was adjusted to 2,567 units using a 25% uncertainty buffer. These outputs can be directly integrated into OEM procurement workflows, enabling proactive, risk-aware inventory management.

While several studies have explored hybrid time-series and survival or time-to-event (TTE) models for reliability forecasting and remaining useful life (RUL) prediction, these approaches typically treat degradation dynamics and temporal prediction as sequential or loosely coupled modules. For example, survival-based frameworks estimate hazard or failure probabilities using deep recurrent networks or stochastic ensembles (Ren et al., 2018[11]; Chen et al., 2024[12], whereas subsequent time-series or regression components extrapolate demand or reliability trends independently.

In contrast, the present framework introduces a unified decay-function-blended machine learning (DFB-ML) architecture, wherein an exponential decay kernel is embedded directly within the forecasting pipeline. This design allows the learned model output to be continuously modulated by physically meaningful lifecycle decline, capturing both non-linear feature interactions and mechanistic end-of-life behaviour in a single recursive loop. Such direct integration of a parameterized decay function inside a data-driven ML forecaster rather than coupling survival and forecasting stages post-hoc, represents a conceptual advance over existing hybrid paradigms. To the best of the authors' knowledge, this is the first application of decay-function blending for SKU-level, multi-feature demand forecasting across extended EOL horizons in the automotive domain.

Related work and research context

Forecasting intermittent and end-of-life (EOL) demand poses major analytical challenges due to sparse data, non-stationary trends, and the influence of physical degradation mechanisms. Traditional forecasting methods, such as Croston's algorithm and its derivatives (Teunter et al., 2002[6]; Boylan & Syntetos, 2006[7], provide basic handling of zero-demand intervals but lack lifecycle awareness. More recent advances in machine learning and ensemble-based forecasting (Carbonneau et al., 2008[13]; Makridakis et al., 2018[14] have improved short-term accuracy but still rely purely on statistical extrapolation without modelling the physical attrition or service decline that drives EOL consumption patterns.

In parallel, significant progress has been made in remaining useful life (RUL) prediction and predictive-maintenance modelling, which focus on estimating degradation and survival probabilities of industrial systems. For instance, Faizanbasha and Rizwan (2025) proposed a deep learning-stochastic ensemble that captures uncertainty in RUL estimation under dynamic mission conditions, enabling risk-aware maintenance planning[15]. Qin et al. (2025) developed a spatial-temporal multi-sensor fusion network that embeds prior physical knowledge for improved RUL inference[16], while Faizanbasha and Rizwan (2025) also formulated a two-unit series reliability optimization model that jointly considers burn-in and predictive maintenance decisions[17]. These approaches highlight the increasing importance of integrating degradation dynamics, stochastic uncertainty, and domain-specific priors into learning architectures.

The present study adapts these reliability-driven principles to the domain of EOL demand forecasting, where part consumption exhibits analogous degradation-like decay governed by fleet attrition and replacement intensity. By embedding an exponential decay kernel within a machine-learning forecasting architecture, the proposed decay-function-blended ML (DFB-ML) framework captures both data-driven feature interactions and physically grounded lifecycle decline. This cross-disciplinary integration of RUL-inspired degradation modelling into SKU-level demand prediction establishes a new methodological link between predictive maintenance and end-of-life inventory planning.

Materials and methods

This section describes the multi-source dataset compiled from a global automotive Original Equipment Manufacturer (OEM) and outlines the methodology adopted for constructing an AI/ML-powered predictive framework aimed at estimating end-of-life (EOL) spare part demand and optimizing Last Time Buy (LTB) quantities. The approach integrates historical demand signals, warranty failure trends, and vehicle attrition modeling into a unified machine learning pipeline[18]. Forecasting is performed over a eight (8) year horizon using engineered features that capture both operational and lifecycle characteristics of spare parts. Six different supervised regression models were trained and evaluated using a temporally structured dataset, and the best-performing models were used to generate future demand estimates. Model evaluation focused on both absolute and proportional error measures to account for the intermittent and long-tailed nature of EOL demand patterns.

The forecasting model incorporates a recursive, part-specific decay adjustment that modulates future predictions based on vehicle dropout trends and historical demand slope. The core structure of the framework is composed of three stages: feature engineering and data fusion, predictive model training and validation, and recursive multi-year forecasting. This structure ensures that forecasts are both statistically robust and aligned with physical realities such as declining serviceable fleets. All forecasting and model selection activities were performed with strict temporal integrity to prevent data leakage and simulate real-world LTB decision-making conditions.

Dataset description

The data utilized in this study was sourced from a global automotive OEM and encompasses 1709 Stock Keeping Units (SKUs) relevant to aftersales spare parts. These SKUs span multiple part categories and include both unique components linked to specific vehicle models and commonly shared parts used across multiple platforms. The time horizon for the analysis spans from 2015 to 2024, covering the full operational life of each part through its production, warranty, and early post-warranty phases.

Three core data streams were integrated into the forecasting model. First, historical annual demand records were acquired for each part, representing the total units requisitioned or sold through after sales channels. These raw demand records were preprocessed to remove anomalies such as one-time bulk orders caused by external disruptions or internal procurement errors. The resulting time series were harmonized to a consistent annual frequency.

Second, warranty claim data was extracted from the OEM's reliability and field service databases. This dataset included the number of claims registered against each part number within the vehicle warranty period, typically defined as the first three to five years following sale. These warranty events were transformed into cumulative failure rate curves, which serve as lead indicators for post-warranty service part consumption. Their inclusion in the feature matrix allows the forecasting models to internalize part-specific failure characteristics and reliability trends over time.

Third, vehicle volume data was used to estimate the in-service installed base for each SKU. Vehicle production numbers were mapped to the part usage catalog to determine the annual cohort of vehicles containing each target part. These production volumes were then adjusted for attrition using a smoothed dropout rate profile. The resulting active fleet size for each SKU-year pair served as a proxy for the population of vehicles still requiring the part in question. This variable is crucial for modeling demand decay over time, as the number of serviceable units directly influences spare part consumption.

The combined dataset was used to construct a part-year level feature matrix, where each row represents a single SKU in a specific year. For each such entry, demand data, warranty indicators, and attrition-adjusted vehicle base information were consolidated. This ensured a temporally aligned and context-rich dataset capable of supporting robust forecasting. Data preprocessing steps included outlier removal, interpolation of missing values in warranty fields (where possible), and normalization of part-level attributes to remove scale imbalances. The result was a unified, multi-dimensional dataset containing both dependent and explanatory variables, enabling supervised learning algorithms to model the complex relationships underlying EOL demand behavior.

The resulting dataset thus reflects not only the direct historical demand for parts but also contextual indicators of failure and serviceable base volume, both of which are essential to understanding the temporal behaviour of spare part consumption[19]. This combination of heterogeneous sources enhances the representational fidelity of the forecasting features and supports generalizable model development. As illustrated in Fig. 2, the correlation between rising cumulative vehicle counts and part demand growth is evident for the five highest-volume parts. However, divergence begins to emerge in later years, as individual part demand begins to saturate or decline despite continued vehicle availability. This non-linear relationship, driven by part-specific failure profiles and replacement behaviour, underscores the necessity of modelling not just installed base, but also time-dependent decay and reliability-adjusted demand dynamics.

The forecasting target and drivers were aligned to an annual cadence for three reasons. First, LTB procurement, contractual service obligations, attrition reporting, and warranty summaries are tracked on annual or fiscal-year cycles, so annualization avoids target–feature misalignment and reduces look-ahead risk. Second, at daily or monthly frequencies most SKUs exhibit intermittent, zero-inflated signals with lumpy batch postings (issues, returns, inter-plant transfers) that reflect ERP logistics timing rather than underlying consumption; aggregation to annual periods suppresses these transaction artefacts and yields a more stable lifecycle pattern suitable for coupling with attrition/warranty features (cf. Croston-type methods and subsequent analyses of intermittent demand). Third, although each SKU contributes ~ 8–10 annual points, the learning problem leverages cross-sectional richness (1,709 SKUs × year) with engineered features (active fleet, warranty intensity, lagged demand, slope/decay), enabling tree ensembles to learn population-level structure that generalizes SKU-wise.

General forecasting model and notation

The objective of the forecasting framework is to predict annual part-level demand over a eight-year planning horizon (2025–2032), with specific emphasis on modeling the impact of vehicle attrition, part-specific lifecycle characteristics, and intermittent demand behavior. The forecasting model is constructed to operate on a year-

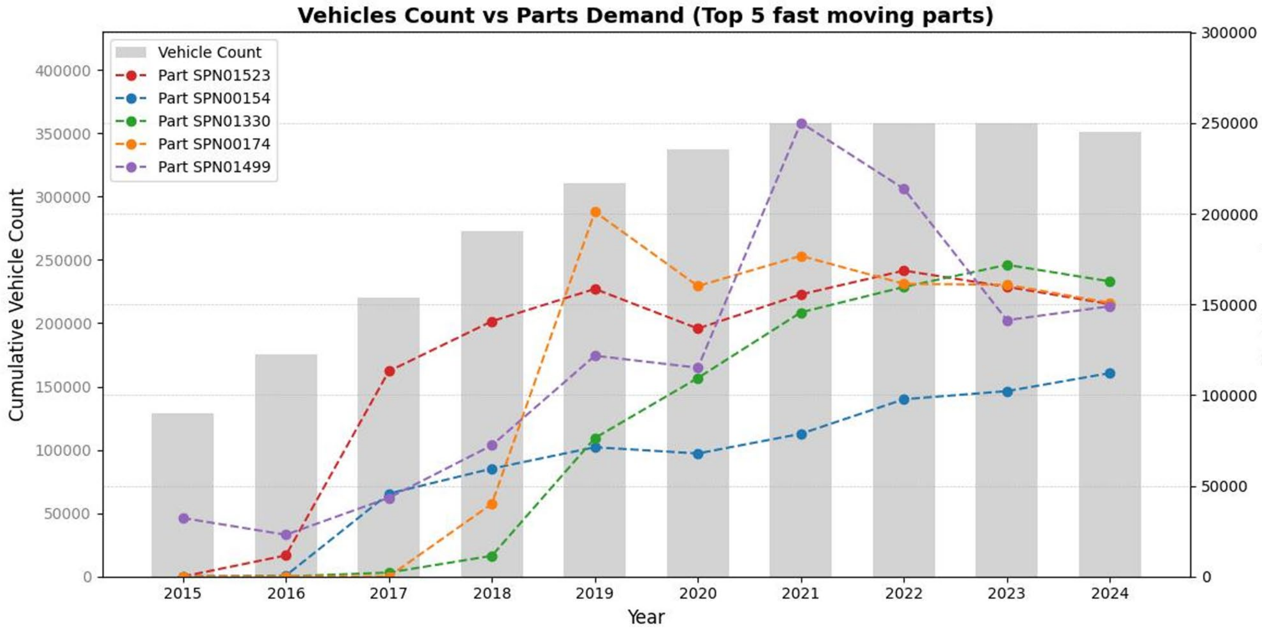

Fig. 2. Cumulative vehicle count (grey bars) vs. annual demand of top five fast-moving spare parts (SPN01523, SPN00154, SPN01330, SPN00174, SPN01499) from 2015 to 2024.

wise resolution and is part-centric in nature, meaning that forecasts are generated independently for each SKU based on its historical data, engineered features, and the evolving serviceable vehicle base. The predictions are not extrapolated from raw time-series values alone; rather, they emerge from a structured feature space that encodes both static and dynamic attributes influencing demand evolution.

Each part is indexed by $p \in P$, and each year is denoted by $y \in Y$, with the historical window spanning from Y_{min}=2015 to Y_{max}=2024. The target variable is the annual demand $D(p,y)$, and the objective is to predict future values $\widehat{D}(p,y)$ for years $y > 2\,024$. Associated with each year is the total vehicle production volume $V\ (y)$, which is specific to the part's application base. These volumes are adjusted for attrition using a dropout function δy , which represents the fractional loss of the installed vehicle base due to scrappage or decommissioning. By recursively applying these dropout rates to the production volumes of prior years, the active vehicle count for any given year can be expressed as:

$$Activefleet\,(y) = \sum\ {}_{y' \leq y} V\left(y'\right) \prod\ {}_{t=y'}^{y-1} 1 - \delta\, t \tag{1}$$

This formulation allows estimation of the number of vehicles still in active use that are eligible to generate spare part demand. The recursive decay model accounts for the declining footprint of vehicles in the field and is critical for aligning part demand forecasts with realistic operational exposure.

For each part-year combination, a structured feature vector $X(p,y)$ is constructed. This vector includes lagged demand values, estimated active vehicle count, historical warranty failure rates, and derived parameters such as part-specific decay rate and demand slope. The slope is obtained through linear regression on past demand values to capture the general trajectory whether upward, flat, or declining while the decay rate ($k\ (p)$ is computed as a function of the slope and the most recent demand level. This rate is used to model an exponential demand decay beyond the observed data window, aligning the forecast with realistic EOL behavior even in the absence of direct observation.

To ensure stability in extrapolation, the decay rate is bounded below by a small positive constant (e.g., 0.001) to avoid abrupt discontinuities or null predictions. The final feature matrix includes both these engineered lifecycle indicators and demand-side signals, enabling the model to learn non-linear interactions among explanatory variables[20].

Each feature vector $X(p,y)$ is input into a trained regression model M_k, indexed by model type k, to yield a point prediction:

$$\widehat{D}(p,y)^{(k)} = M_k\left(X\left(p,y\right)\right) \tag{2}$$

In practice, a recursive forecasting scheme is employed for multi-year prediction. For a future year $y > 2024$, the feature vector $X(p,y)$ is updated using the model's own prior forecasts. The lagged demand input is replaced with $\widehat{D}(p,y-1)$, and the active vehicle count is adjusted based on updated dropout estimates. The prediction process continues year by year, feeding forward predictions and feature updates.

To avoid error accumulation and reflect the expected decay in demand as vehicles age out of service, a decay-modulated adjustment is introduced. A secondary forecast $DecayPred(p,y)$ is computed using an exponential function centered at the last observed demand $D(p,\ 2024)$:

$$DecayPred\,(p,y) = D\,(p,2024)\,.e^{-k(p)(y-2024)} \tag{3}$$

The final demand estimate $FinalPred(p,y)$ is expressed as a convex combination of the model-based forecast and the decay function:

$$FinalPred\,(p,y) = \alpha\ .\widehat{D}\,(p,y) + (1-\alpha\ )\,.DecayPred(p,y) \tag{4}$$

where $\alpha\ \in\ [0{,}1]$ is a blending coefficient empirically determined to balance learned patterns and lifecycle decay alignment[21]. This hybrid forecast structure stabilizes predictions in the long tail of the lifecycle curve and prevents overestimation in years far removed from the training data horizon.

This structured approach combining machine-learned regression with lifecycle-aware decay dynamics forms the backbone of the proposed forecasting model. It enables robust estimation of part-level demand under uncertainty, and supports strategic Last Time Buy decisions by quantifying decline trajectories with respect to attrition-adjusted installed base and historical usage trends.

To ensure that the selected features contributed independent explanatory power to the forecasting models, a Pearson correlation matrix was computed across all engineered features. As shown in Fig. 3, low to moderate correlation values were observed between most feature pairs. For example, lagged demand and replacement rate exhibit a high correlation (0.88), which is expected given their shared dependence on past usage. However, features such as active vehicle count and decay rate show near-zero correlations with other inputs, confirming that the final feature set captures diverse and complementary signals relevant to demand evolution. This low multicollinearity supports the stability and interpretability of both linear and non-linear model architectures trained on this space.

AI/ML modelling framework

The AI/ML modelling framework developed in this study is designed to forecast EOL spare part demand over an 8-year horizon by learning from feature-rich historical data. The modelling process comprises three stages: supervised learning using temporally aligned part-year data, validation and model selection based on generalization metrics, and lifecycle-aware recursive forecasting with decay modulation. Each phase is described below.

Model training and feature learning

Model training was conducted on a curated part-year dataset comprising observations from 2015 to 2022. For each SKU and year, a structured feature vector was constructed as described in Sect. 2.2, incorporating engineered variables such as active fleet size, demand slope, lagged consumption, and replacement rate. These features, aligned chronologically with the target demand value $D\left(p,y\right)$, formed the basis of a supervised regression problem.

Temporal causality was maintained during model training by ensuring that only past and present information was included in the feature space. At no point were future demand values or temporally leaked information made accessible to the model during training or validation. This reflects realistic operational constraints, where forecasting decisions must be made without knowledge of future actuals[22]. During the forecasting phase (2025–2032), a recursive approach was adopted wherein predictions from prior years were used to construct lagged features for subsequent years, thereby preserving causal sequencing while enabling multi-year horizon inference. The feature space was standardized and normalized to accommodate models sensitive to scale variance, and outlier handling was implemented for rare, high-volume anomalies. Hyperparameter tuning was performed using grid or randomized search protocols, with separate training and validation splits.

Model classes and evaluation metrics

Six regression algorithms were trained independently on the same input–output structure to enable comparative benchmarking. The linear group consisted of ElasticNet and HuberRegressor. ElasticNet was selected for its embedded feature regularization, combining L1 and L2 penalties, while HuberRegressor was used for its resilience against noise and heavy-tailed errors.

The non-linear group included four tree-based ensemble models: Random Forest, XGBoost, LightGBM, and CatBoost. These models are well suited for structured tabular data, and have demonstrated robust performance under conditions of missing values, interaction effects, and high-dimensional input spaces. All models were implemented and trained under consistent data partitions.

Model performance was evaluated using three complementary metrics: Mean Absolute Error (MAE), Root Mean Squared Error (RMSE), and Safe Mean Absolute Percentage Error (Safe MAPE). MAE provided a scale-consistent average error magnitude, RMSE emphasized the impact of large residuals, and Safe MAPE ensured stability under low-demand conditions by preventing denominator collapse[23]. This evaluation scheme allowed robust comparison of both absolute and proportional forecasting error across varying demand levels and SKU types.

Model selection and final forecasting protocol

After evaluating all candidate models on a temporally held-out test set (2023–2024), the best-performing algorithm typically RF was retrained on the full historical data (2015–2024) to maximize available information. Final forecasts for 2025 to 2032 were generated using a recursive inference loop in which predictions for year y were used to construct feature vectors for year $y+1$. This preserved the causal sequence and allowed long-range forecasting beyond the training horizon.

To avoid error drift and improve lifecycle realism, forecasts were blended with an exponential decay projection computed using part-specific decay coefficients derived from historical slope and recent demand levels[24]. The blended output, defined as a convex combination of model inference and decay projection, ensured that the resulting forecasts remained anchored to both data-driven trends and engineering lifecycle expectations. This hybrid strategy enabled the model to respect physical product realities such as saturation, obsolescence, and dropout while retaining the flexibility of data-driven learning.

The final modeling pipeline thus represents a fusion of statistical learning, operational insight, and lifecycle logic, making it well-suited for industrial applications where demand volatility, inventory risk, and long service obligations coexist.

Hyperparameter tuning and validation protocol

To ensure reproducibility and prevent overfitting, model hyperparameters were optimized through systematic grid or randomized search using a temporal validation scheme. Data from 2015 to 2022 were used exclusively for training, while 2023–2024 served as the validation and test window. The Safe Mean Absolute Percentage Error (Safe MAPE) metric was used as the primary optimization criterion, with MAE and RMSE employed as secondary checks for stability.

The Table 1 lists the principal hyperparameters and search ranges employed for each regression model. For ensemble methods, early-stopping criteria and random seeds were fixed across trials to ensure comparability. For linear regressors, regularization parameters were tuned within logarithmic ranges to balance bias–variance trade-off. This structured tuning protocol ensured reproducible optimization across all model families and provides a transparent basis for the comparative performance shown in Fig. 6b.

During hyperparameter optimization, the Safe Mean Absolute Percentage Error (Safe MAPE) was adopted as the primary selection metric, owing to its stability under sparse and low-volume demand conditions common in EOL forecasting. The Mean Absolute Error (MAE) and Root Mean Squared Error (RMSE) were used as secondary verification metrics to ensure that the chosen configuration minimized both absolute and variance-weighted residuals. Hyperparameter combinations yielding the lowest Safe MAPE with consistent MAE/RMSE performance across the validation period (2023–2024) were retained for final model training.

Model	Key hyperparameters	Search range/setting	Selection criterion
Elasticnet	α (L1–L2 mixing parameter), λ (regularization strength)	α ∈ [0.1, 0.5, 0.9]; λ ∈ {0.001, 0.01, 0.1, 1.0, 10}	Minimum safe MAPE
Huber regressor	ε (Huber threshold), α (learning rate)	ε ∈ {1.0, 1.5, 2.0}; α ∈ {0.0001, 0.001, 0.01}	Minimum safe MAPE + stable RMSE
Random forest	n_estimators, max_depth, min_samples_split	n_estimators ∈ {100, 300, 500, 1000}; max_depth ∈ {5, 10, 15, None}; min_samples_split ∈ {2, 5, 10}	Balanced bias–variance and lowest safe MAPE
XGboost	n_estimators, max_depth, learning_rate, subsample	n_estimators ∈ {100, 300, 500}; max_depth ∈ {3, 5, 7, 9}; learning_rate ∈ {0.01, 0.05, 0.1}; subsample ∈ {0.7, 0.8, 1.0}	Validation safe MAPE minimized with ≤ 2% overfit gap
LightGBM	num_leaves, learning_rate, feature_fraction, bagging_fraction	num_leaves ∈ {31, 63, 127}; learning_rate ∈ {0.01, 0.05, 0.1}; feature_fraction ∈ {0.8, 0.9, 1.0}; bagging_fraction ∈ {0.7, 0.8, 1.0}	Lowest safe MAPE and stable training/test error
Catboost	iterations, depth, learning_rate, l2_leaf_reg	iterations ∈ {300, 500, 800}; depth ∈ {4, 6, 8}; learning_rate ∈ {0.01, 0.05, 0.1}; l2_leaf_reg ∈ {3, 5, 7}	Minimum safe MAPE with smooth forecast curve
Decay blending (β in Eq. 4)	Blending coefficient β	β ∈ [0.2, 0.5, 0.8]; selected empirically based on forecast continuity and MAPE stability	Continuity at 2024–2025 transition + lowest error growth rate

Table 1. Hyperparameter search ranges and validation protocol for six regression models.

During scoping, also evaluated finer granularities (monthly/quarterly) on a representative set of high-volume SKUs; no consistent improvement in Safe MAPE was observed due to zero-inflation and batching artefacts, whereas the annual target preserved the intended coupling with warranty/attrition features and the LTB decision cadence.

Evaluation metrics and validation protocol

The evaluation of model performance in this study is grounded in a rigorous validation strategy that reflects real-world forecasting constraints in end-of-life (EOL) service parts planning. Given the long planning horizons and sparse, intermittent nature of demand, conventional error metrics and random cross-validation approaches were deemed insufficient. Instead, a temporally stratified hold-out validation was employed, paired with carefully selected error metrics tailored to low-volume, long-tail data behavior.

The dataset was partitioned into three temporal windows. Data from 2015 to 2022 was used exclusively for training, ensuring sufficient historical context for each SKU. The years 2023 and 2024 were held out as a test set to evaluate out-of-sample generalization. This approach simulates the actual conditions under which LTB decision would be made in practice namely, forecasting future demand based solely on past consumption patterns, without access to future realizations. This validation structure preserves causal integrity and guards against look-ahead bias, a common concern in time-series machine learning applications.

To quantify model accuracy, three complementary metrics were employed: MAE, RMSE, and Safe MAPE. MAE provides a scale-sensitive, robust measure of average deviation between predicted and actual values. It is particularly well-suited for applications involving diverse SKUs with different demand magnitudes. RMSE places greater emphasis on large deviations, penalizing models that make occasional but severe prediction errors an important consideration when planning expensive or critical spare parts procurement.

Safe MAPE was developed specifically for this study to address numerical instability and misleading magnitudes often observed in traditional MAPE when actual demand values are near zero. The Safe MAPE formulation substitutes a small constant threshold in the denominator when the actual value is very low or zero, thereby avoiding division by zero and suppressing inflated percentage errors. This makes Safe MAPE particularly useful in the EOL context, where many SKUs exhibit declining or near-zero demand as vehicles exit the serviceable fleet[25]. The Safe MAPE calculation is defined as:

$$\text{Safe MAPE} = \frac{100}{n}\sum{}_{i=1}^{n} \left| \frac{\widehat{y_i} - y_i}{\max(y_i,\ \epsilon\ )} \right| \tag{5}$$

where $\widehat{y_i}$is the predicted demand, $y_i$is the actual demand, and ϵ is a small positive threshold (e.g., 10 units) introduced to prevent division by zero and suppress inflated percentage errors when actual demand approaches zero. This formulation ensures numerical stability and fair comparison across SKUs with low or intermittent demand.

Model selection was based on aggregated performance across all three metrics on the test window (2023–2024). The chosen model was then retrained on the full historical period (2015–2024) prior to generating forecasts for 2025–2032. This ensures that all available data is utilized once the model has demonstrated generalization capability under realistic hold-out conditions. The validation protocol and metric triad together form a robust performance assessment framework tailored to the high-stakes and data-scarce nature of LTB forecasting.

Forecast deployment and use case integration

Following model training, validation, and selection, the final forecasting framework was deployed for operational use to generate SKU-wise demand projections over the eight-year planning window from 2025 to 2032. The deployment protocol was designed to simulate the conditions under which OEMs execute LTB decisions in a high-stakes, resource-constrained environment. The forecasting outputs were generated at an annual resolution and structured to support both aggregate procurement planning and part-specific risk assessment.

The chosen model RF in most cases was retrained using the complete historical dataset from 2015 to 2024 to leverage all available information. For each part, a recursive forecasting loop was initiated, beginning with year 2025. For each forecast year *y*, the previously predicted demand $\widehat{D}(p, y-1)$ was inserted into the feature vector as the lag variable. The active vehicle count was updated using dropout-adjusted production volumes, and all engineered features (e.g., replacement rate, demand slope, decay rate) were recomputed as required. This forward-inference approach ensured that the forecasting engine operated autonomously without any access to true future values, thereby preserving realism and preventing leakage.

To modulate long-horizon predictions and align them with physical part lifecycle decay, the raw model outputs were blended with part-specific exponential decay functions derived from historical demand slope and final year (2024) consumption. The convex combination of model-based predictions and decay-adjusted projections stabilized the output trajectory, particularly for SKUs exhibiting sharp attrition or usage decline.

The forecast results were structured as a part-year matrix containing predicted demand values for each SKU from 2025 to 2032. These results were visualized through line plots, overlaying historical demand, test period performance, and long-term projections across all model classes. Representative cases were selected for deeper inspection, highlighting trends such as forecast flattening, accelerated decay, or model disagreement. In particular, the top 5 high-volume SKUs and a set of critical low-frequency parts were examined to illustrate the operational diversity of outcomes and planning implications.

The complete forecasting workflow was implemented using Python 3.11 and associated scientific computing libraries, including scikit-learn, xgboost, catboost, lightgbm, and pandas. Forecast outputs were exported to structured CSV files and integrated into Excel-based planning dashboards, enabling domain experts to overlay cost, lead time, and supplier constraints. The system supports dynamic "what-if" analysis by allowing the adjustment of dropout rates, decay coefficients, or model blending weights, thereby enhancing flexibility in uncertain or rapidly evolving field environments[26].

The proposed forecasting framework has been architected for seamless integration within enterprise-grade planning ecosystems. All computational modules: data preprocessing, model training, decay blending, and visualization are implemented as modular Python services communicating through standardized APIs. The system exchanges structured data via CSV, JSON, and ODBC connectors, enabling direct linkage with SAP S/4HANA, SAP BW/4HANA, Oracle ERP, and SAP Data Intelligence environments. Forecast outputs, uncertainty bands, and LTB recommendations can be automatically synchronized with material-master and procurement tables through batch or real-time jobs.

From a computational standpoint, the framework supports distributed execution using Dask and Apache Spark back-ends, allowing horizontal scaling across multiple compute nodes. Containerized deployment through Docker and Kubernetes ensures portability between on-premises infrastructure and cloud environments such as AWS SageMaker, Azure ML Studio, or Google Vertex AI. The architecture's modularity permits incremental updates to the predictive engine without altering existing ERP data schemas, ensuring long-term maintainability. This design demonstrates the practical readiness of the decay-function-blended ML system for industrial deployment in real-world ERP/SAP networks, providing an interpretable, scalable, and secure decision-support tool for strategic Last-Time-Buy forecasting.

This deployment architecture demonstrates the viability of the forecasting framework not only as a research prototype but also as a practical decision-support tool for LTB planning. By aligning model predictions with business workflows and interpretability needs, the solution offers a scalable foundation for OEMs managing complex EOL service obligations across thousands of SKUs.

Results and discussions

This section presents a comparative evaluation of six machine-learning models developed for Last-Time-Buy (LTB) decision support. The results are reported in terms of predictive accuracy across multiple models and metrics, providing both quantitative and qualitative insight into their suitability for end-of-life (EOL) demand forecasting. The comparative performance is assessed on a temporally structured dataset comprising 1,709 SKUs of automotive spare parts with associated historical demand, vehicle attrition, and warranty claim data from 2015 to 2024. As summarized in Table 2, the study employed six regression algorithms spanning linear, robust, and ensemble-based learning paradigms. The table highlights each model's assumption and operational suitability for end-of-life (EOL) demand forecasting. Among these, Random Forest combined with exponential-decay blending demonstrated the most balanced trade-off between accuracy, interpretability, and lifecycle alignment.

Testing phase performance

Model performance on the hold-out test dataset (2023–2024) was evaluated to determine generalization capability in real-world forecasting scenarios. As shown in Table 3, Random Forest achieved the best performance across all three metrics, with a test MAE of 115.9, RMSE of 741.39, and Safe MAPE of just 4.36%. These results indicate strong predictive accuracy even under low-volume and intermittent demand conditions.

XGBoost and CatBoost followed with MAEs of 420.75 and 486.28, respectively. However, both models exhibited higher RMSEs (5437.93 and 6314.92), suggesting sensitivity to outliers. LightGBM recorded the highest test MAE (595.44) and RMSE (7567.08), reflecting significant variance and poor generalization. Among linear regressors, HuberRegressor showed better robustness (MAE = 202.61, RMSE = 1099.35, Safe MAPE = 25.63) than ElasticNet (MAE = 346.05, Safe MAPE = 478.11), the latter clearly failing under sparse demand conditions.

Figure 4 presents a side-by-side bar chart comparing training and test errors across all models. It is evident that Random Forest maintains consistently low errors with minimal performance drop between phases. In contrast, LightGBM and CatBoost exhibit significant overfitting, while ElasticNet demonstrates unstable proportional accuracy[27].

Model	Type	Core assumption/learning principle	Key strengths	Limitations in EOL context
Elasticnet	Linear regression with L1 + L2 regularization	Demand follows quasi-linear trend; correlated features penalized proportionally	Handles multicollinearity; interpretable coefficients	Cannot capture non-linear or discontinuous demand behaviour; sensitive to zero-inflated data
Huber regressor	Robust linear model combining least-squares and absolute loss	Outliers are limited by Huber threshold; assumes smooth residual distribution	Resistant to noise and extreme values	Linear formulation underfits sparse, highly non-stationary data
Random forest	Ensemble of decision trees using bootstrap aggregation	Non-parametric; learns non-linear interactions without feature scaling	High accuracy, low variance; interpretable feature importance; stable on sparse data	Large memory footprint; limited extrapolation beyond training range
XGBoost	Gradient-boosted tree ensemble	Sequential residual correction; regularized boosting for overfitting control	Fast training; strong generalization; handles mixed feature types	Prone to overfitting on small datasets; sensitive to learning-rate tuning
LightGBM	Leaf-wise gradient boosting with histogram optimization	Greedy leaf growth to minimize loss	Very fast on large datasets; efficient parallelization	May overfit or produce unstable forecasts under high sparsity
Catboost	Ordered boosting with categorical feature encoding	Uses permutation-driven boosting and target statistics	Handles categorical variables efficiently; reduced overfitting bias	Sensitive to noise; can produce abrupt forecasts when historical trend shifts
Proposed RF + decay blending	Random Forest combined with exponential-decay modulation	ML prediction dynamically adjusted by lifecycle decay kernel	Physically interpretable; robust long-horizon extrapolation; suitable for SKU-level deployment	Requires empirical tuning of blending coefficient; current version deterministic (to be extended with uncertainty-aware methods)

Table 2. Summary of machine-learning models used for EOL demand forecasting.

Model	MAE	RMSE	Safe MAPE
Random forest	115.90	741.39	4.36%
XGboost	420.75	5437.93	12.21%
Catboost	486.28	6314.92	19.52%
LightGBM	595.44	7567.08	20.23%
Elasticnet	346.05	1354.30	478.11%
Huberregressor	202.61	1099.35	25.63%

Table 3. Comparative performance of six machine-learning models on test data (2023–2024).

In Fig. 5, the predicted vs. actual demand for selected SKUs during the 2023–2024 test window further confirms the pattern. RF tracks the actual demand curve closely with minimal deviation, while linear models exhibit flat or delayed response behavior, unsuitable for responsive LTB forecasting.

Training phase results

Performance on the training dataset (2015–2022) provides considerate into model fit and potential overfitting risks. RF again exhibited the most favorable training characteristics, with a MAE of 53.2, RMSE of 684.3, and Safe MAPE of 3.89%, indicating well-regulated complexity and effective learning from historical patterns.

XGBoost and CatBoost recorded lower training MAEs (198.7 and 188.4, respectively), but their Safe MAPE values (8.42% and 6.13%) and the sharp increase in test error suggest these models overfit on sparse or skewed inputs. LightGBM, despite a training MAE of 263.4, showed poor test generalization, reinforcing its tendency to over-learn from noisy training points. The model's RMSE rose from 4885.1 (training) to 7567.1 (test), indicating insufficient robustness.

Linear regressors displayed underfitting behavior. ElasticNet produced comparable errors in both phases (MAE: 358.6 training vs. 346.1 test), but with extremely high Safe MAPE (>478%), confirming its inability to capture non-linear or discontinuous demand patterns. HuberRegressor exhibited slightly better results (training MAE: 256.7; test MAE: 202.6), but lacked sufficient accuracy for high-stakes LTB forecasting[28].

The small generalization gap for RF across all metrics underscores its capacity to learn part-specific decay trends, vehicle attrition, and demand slope effects without overfitting. This reliability makes it well-suited for use in operational LTB decision-making.

Visual analysis of forecasted demand

To evaluate the long-term applicability of the trained models for Last Time Buy (LTB) planning, 8-year forecasts (2025–2032) were generated for each SKU using all six models. Figure 6 illustrates the forecasted annual demand for a representative spare part (SPN00001), comparing predictions across models.

The demand trajectories for tree-based models: RF, XGBoost, CatBoost, and LightGBM exhibit a consistent declining trend, aligned with expectations for spare parts in the post-production service phase. This decline reflects both the decaying installed vehicle base and the diminishing replacement intensity over time. Among the models, RF and XGBoost demonstrate smoother and more conservative decay curves, indicating effective integration of attrition-adjusted vehicle population and historical demand slopes. CatBoost and LightGBM produce more aggressive early drops in forecast, which may reflect over-sensitivity to recent downward fluctuations.

In contrast, linear models show markedly different behavior. ElasticNet predicts an initially high demand followed by a rapid exponential drop, consistent with its bias toward the dominant historical trend.

HuberRegressor provides a moderated decay but lacks the non-linear adaptability required to track subtle demand shifts. These patterns confirm earlier findings that linear models lack the flexibility to generalize in the context of non-stationary and intermittently sparse demand, often leading to abrupt or unstable long-range predictions[29].

The comparative behavior of these models highlights the practical advantage of ensemble-based learners for high-stakes LTB decision-making. Random Forest, in particular, offers interpretable and stable forecasts that align with empirical expectations and operational experience. The 8-year horizon forecast for SPN00001, as shown in Fig. 5, is a representative case that underscores the importance of robust modeling under declining lifecycle dynamics.

To evaluate the long-term applicability of the trained models for Last Time Buy (LTB) decision-making, 8-year forecasts (2025–2032) were generated for each SKU using all six trained regressors. Figure 6a and b illustrate forecasted annual demand trajectories for two representative spare parts: SPN00001 and SPN01499 highlighting the temporal behavior and continuity of predictions across models.

In Fig. 6a, forecasts for SPN00001 exhibit a consistent downward trend across all models, which aligns with expected spare part lifecycle patterns influenced by vehicle attrition and reduced failure incidence over time. Random Forest and XGBoost display smoother decay curves, indicating effective integration of engineered features such as dropout-adjusted fleet size and demand slope. CatBoost and LightGBM show steeper initial drops, which may be attributed to overfitting to recent historical fluctuations[30,31]. Linear models deviate sharply ElasticNet projects an overconfident steep decline, while HuberRegressor maintains a lagged decay with poor alignment to nonlinear dynamics.

Figure 6b complements this analysis by combining both historical demand (2015–2024) and forecasted demand (2025–2032) for part SPN01499. A vertical dotted line marks the boundary between training and prediction phases. The observed historical demand profile exhibits intermittent but increasing trends until 2024. Random Forest, XGBoost, and CatBoost extend this trajectory into a controlled exponential decay, demonstrating robustness in lifecycle-aware extrapolation. In contrast, ElasticNet and HuberRegressor significantly overshoot early demand levels and then collapse into unrealistic tail forecasts. The transition at the 2024 boundary is notably smooth for tree-based models, which is critical for ensuring demand continuity across planning horizons.

Together, these figures validate the practical utility of ensemble learning approaches in forecasting long-tail, intermittent demand signals with decaying behavior. Forecast continuity, as observed in Random Forest's trajectory, reduces the risk of under- or over-estimation of LTB quantities, which is critical to minimizing obsolescence costs and service risk[32].

Automotive supply chain practical implications

The preceding analyses demonstrate that the integration of lifecycle-aware features with non-linear machine learning models significantly improves the forecasting accuracy and operational reliability of LTB demand estimation. This section synthesizes key technical insights and discusses their implications for OEMs in automotive and capital goods industries managing end-of-life (EOL) service obligations.

A primary insight is the importance of incorporating vehicle attrition-adjusted active fleet size into the predictive framework. Models that explicitly account for declining installed base, such as Random Forest and XGBoost, exhibited smoother and more realistic decay in long-horizon forecasts (Fig. 6a and b). This behavior is essential for aligning spare part procurement with the actual serviceable population and avoiding premature inventory exhaustion or overstocking.

The role of engineered features, such as demand slope and part-specific decay rate, was equally critical. By capturing both short-term dynamics (e.g., lagged demand and replacement intensity) and long-term lifecycle effects (e.g., exponential decay tied to dropout rates), the framework enabled tree-based models to adapt to highly intermittent and non-stationary demand patterns[33]. This adaptability directly supports improved LTB quantity planning, particularly for low-volume, long-tail SKUs that traditional time-series models struggle to represent.

Model robustness to intermittent and sparse data emerged as another key differentiator. As observed in Table 1; Figs. 3, 4 and 5, tree-based models consistently outperformed linear methods on Safe MAPE, a critical metric in low-demand regimes. For instance, Random Forest maintained a Safe MAPE of 4.36% on the test set, while ElasticNet exceeded 470%, rendering it ineffective for inventory-critical applications. This proportional reliability ensures that even marginal parts those frequently ignored in classical inventory models can be forecasted with confidence.

From a practical standpoint, the results underscore the need for OEMs to shift from static, historically driven methods to dynamic, feature-fused predictive systems. Traditional univariate forecasting approaches, or methods that ignore vehicle retirement and part obsolescence dynamics, are ill-suited for the complex nature of aftermarket support. By adopting models that explicitly learn from warranty failures, attrition curves, and lifecycle phases, OEMs can reduce financial risk associated with over-procurement and simultaneously avoid service failures due to stockouts[34].

Moreover, the capability to visualize model-specific demand trajectories (Fig. 5b) enables procurement and supply chain planners to evaluate prediction confidence over time. For instance, the divergence of CatBoost and ElasticNet trajectories in later years highlights the need to integrate uncertainty quantification or ensemble averaging into LTB planning workflows, particularly beyond the first five forecast years.

Overall, the AI/ML-based approach demonstrated in this study provides a scalable and interpretable framework for end-of-life inventory forecasting. It facilitates data-driven LTB decisions that balance capital investment, customer service continuity, and regulatory compliance. When operationalized within digital supply chain platforms, such models can function as decision-support modules that continuously adapt to new data and evolving usage patterns, ultimately improving resilience and efficiency in spare parts logistics.

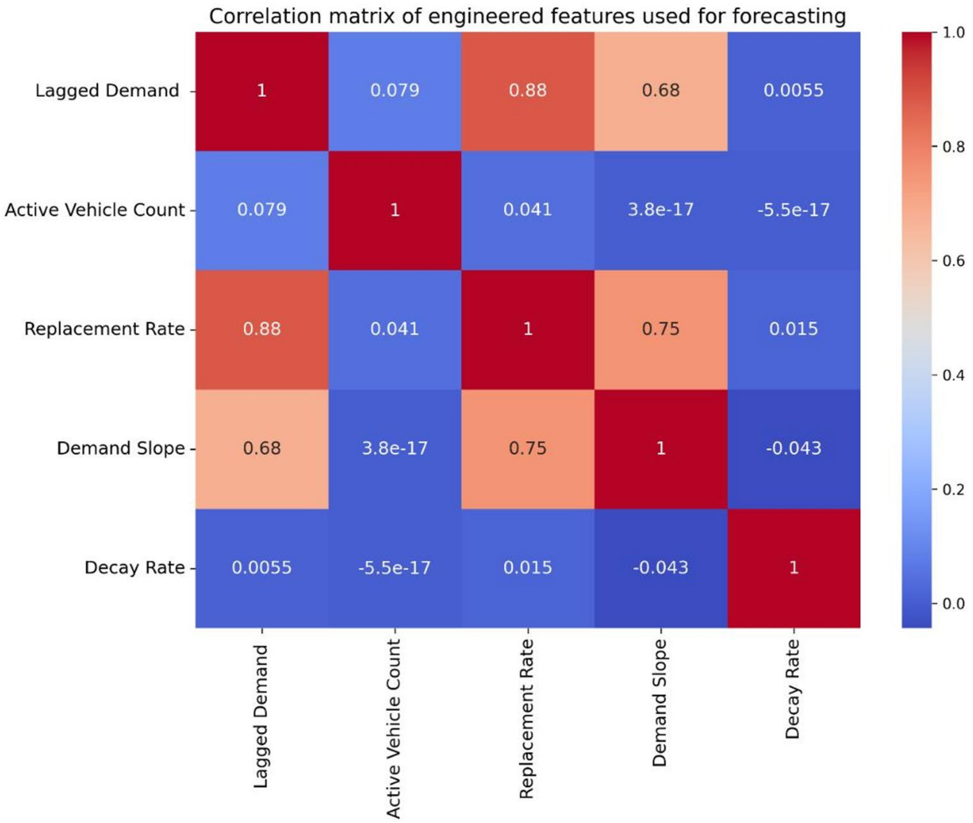

Fig. 3. Correlation matrix of engineered features used for forecasting.

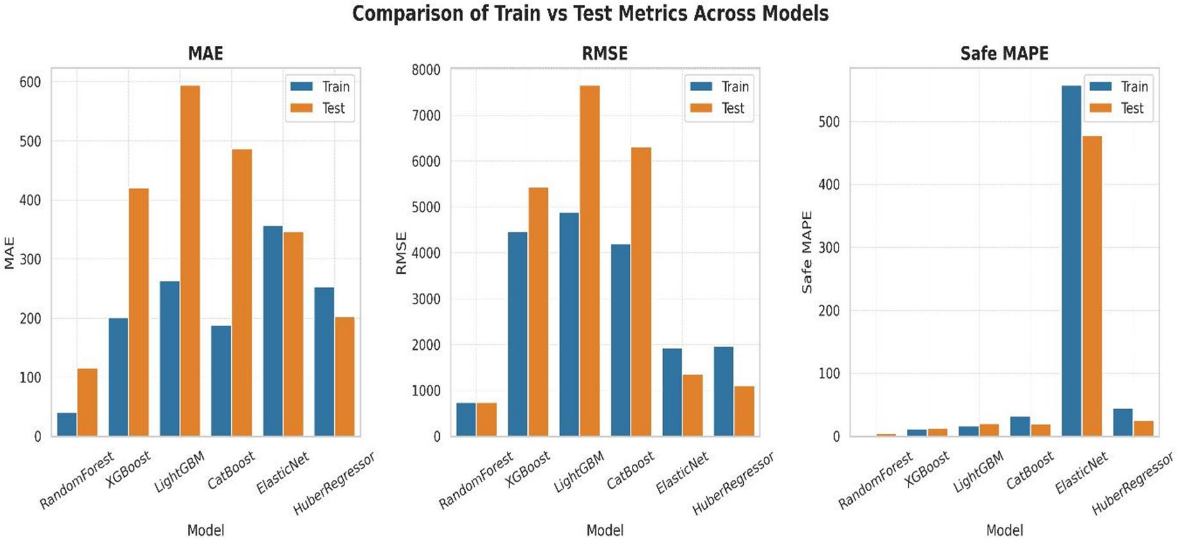

Fig. 4. Comparison of train and test error metrics across models: MAE (left), RMSE (center), and Safe MAPE (right).

Cost implications and LTB recommendations

The deployment of machine learning-based demand forecasting for EOL spare parts is not solely a data science exercise, it directly informs procurement decisions with long-term cost consequences. This chapter translates the forecast outputs into economic implications and provides actionable guidance for determining optimal LTB quantities. The cost model reflects trade-offs between stockout risk, excess inventory holding, and obsolescence, all of which are exacerbated in the tail end of the product lifecycle[35].

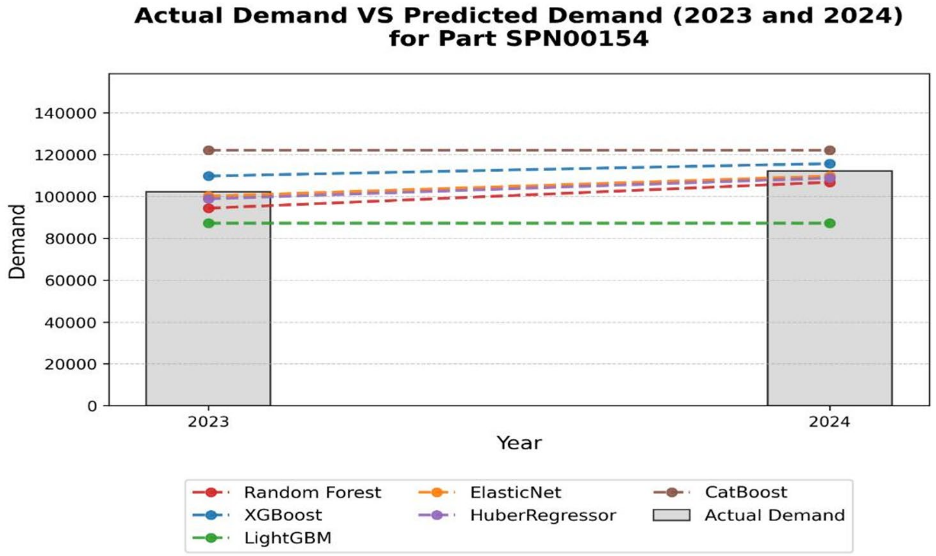

Fig. 5. Actual vs. predicted demand for representative SKUs during test years 2023–2024.

Cost impact of forecasting accuracy

Demand forecast accuracy has a non-linear effect on cost exposure during LTB events. Under-prediction of demand results in unfulfilled service obligations, reputational loss, and emergency re-sourcing often at significantly higher costs or impossibility due to tool obsolescence. Conversely, overestimation leads to overstocking, which incurs inventory holding costs and ultimately scrap or write-off costs if parts exceed shelf-life or become technologically obsolete.

Let the unit procurement cost of a part be denoted as C_p, the annual holding cost rate as *h*, and the forecasted demand for the 8-year horizon as $\widehat{D}_8\left(p\right)$. If the actual demand over the horizon is $D_8\left(p\right)$, then the cost deviation $\Delta\,\mathrm{C}\left(\mathrm{p}\right)$ associated with forecast error is approximated as[36]:

$$\Delta\,\mathrm{C}\left(\mathrm{p}\right) = C_p.\max\left[\left(0,\ \widehat{D}_8\left(p\right) - D_8\left(p\right)\right).h.t + \max\left(0, D_8\left(p\right) - \widehat{D}_8\left(p\right)\right).\lambda \tag{5}$$

Here, *t* represents the average years of inventory held, and λ is the penalty for stockout or emergency sourcing. This expression formalizes the dual nature of cost penalties and provides a quantitative basis for integrating forecast uncertainty into procurement risk models.

Models with higher Safe MAPE such as ElasticNet and HuberRegressor are associated with significant upward deviation in $\widehat{D}_8\left(p\right)$, making them economically unattractive due to overstocking. Conversely, tree-based models with lower forecast dispersion and smoother decay profiles (as seen in Fig. 5a and b) reduce the expected value of both overage and underage cost terms, leading to more financially resilient procurement decisions.

Part-wise LTB recommendation strategy

To operationalize the forecasting results into LTB quantities, a part-wise aggregation of the 8-year forecast was performed for each SKU. The final LTB recommendation per part is expressed as[37]:

$$Q_{LTB(p)} = \sum\,{}_{y=2025}^{2032} FinalPred\left(p,y\right) + \theta\ .\sigma\ \left(p\right) \tag{6}$$

where $FinalPred\left(p,y\right)$ is the forecasted demand for year *y* from the selected model, and $\sigma\ \left(p\right)$ is the standard deviation of forecast error from 2023 to 2024 test results, used here as a proxy for model uncertainty. The term θ is a tunable safety factor (typically between 0.1 and 0.3), adjustable based on the part's criticality, lead time, and cost sensitivity. Parts were segmented into three categories for LTB planning[38]:

(a) Category A (critical high volume): High forecast volume and low acceptable service risk. These parts receive the full forecast plus a safety buffer (θ = 0.25\theta = 0.25θ = 0.25 or higher).
(b) Category B (moderate volume, moderate risk): Moderate service demand and moderate cost exposure. Forecasts are used directly without buffer.
(c) Category C (low volume, high obsolescence risk): Intermittent demand with low expected consumption. These receive minimum LTB volume based on recent actuals or are excluded.

An example application of this approach is shown in Table 4, where parts SPN00001, SPN01499, and SPN01330 are recommended for procurement based on their forecast trajectories, criticality index, and historical consumption patterns.

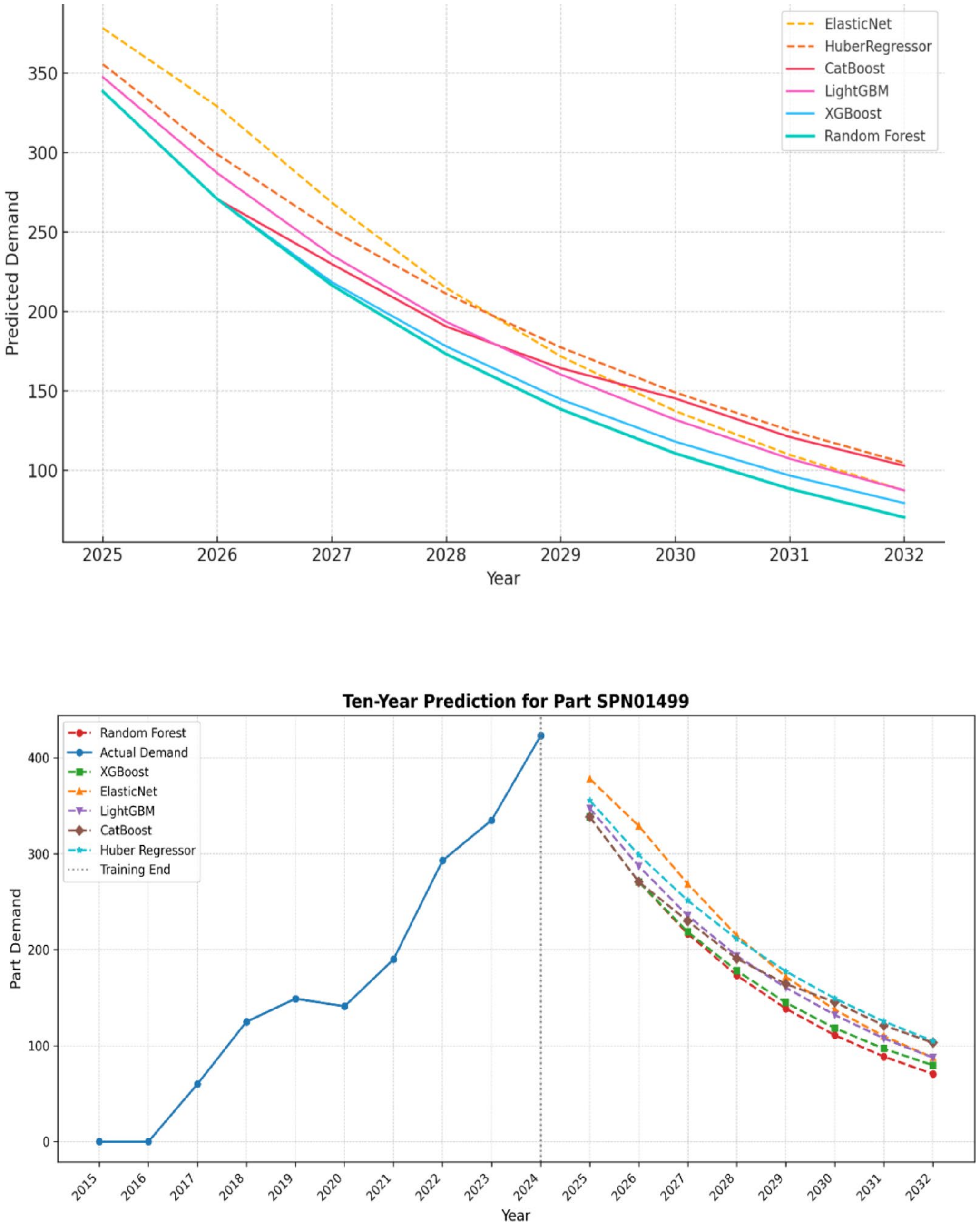

Fig. 6. (**a**) Forecasted spare part demand from 2025 to 2032 for SKU SPN00001 using six ML models. (**b**) Combined historical and 8-year forecasted demand (2015–2032) for SKU SPN01499

Part number	8 year forecast units	Forecast STD	Safety factor	LTB recommended quantity
SPN00001	1407.616	93.64594	0.25	1431.03
SPN01330	552,140	36253.46	0.25	561203.3
SPN01499	505225.3	33256.04	0.25	513539.4

Table 4. Part-wise forecast and recommended LTB quantities.

The LTB recommendation framework, grounded in forecast outputs and error-derived uncertainty quantification, enables OEMs to balance inventory cost with service continuity over extended support horizons. When integrated into enterprise resource planning (ERP) systems or procurement dashboards, these recommendations can be dynamically adjusted based on market intelligence, part price volatility, or program extension scenarios.

Ablation and horizon-wise error analysis

To quantify the influence of decay blending and assess forecast stability, an ablation study was performed for blending coefficients α ∈ {0.0, 0.2, 0.4, 0.6, 0.8, 1.0}. Figure 7a and b show results for a representative high-volume part (SPN01499) using Random Forest and XGBoost regressors. As α increases from 0 (no blending) to 1 (full exponential decay), forecasts transition from over-persistent to over-attenuated trajectories. Intermediate values (α ≈ 0.2–0.4) produce physically consistent declines aligned with vehicle-fleet attrition, confirming the stabilizing role of decay modulation.

For empirical error analysis, rolling-origin back tests were conducted within 2019–2024 to compute horizon-wise MAE, RMSE, and Safe MAPE for 1- to 6-year-ahead forecasts. The observed error growth curves exhibit sub-linear scaling under decay blending, indicating reduced compounding of uncertainty. Since actual demand for 2025–2032 is not yet available, error metrics beyond 2024 are projected by extrapolating the empirical error-vs-horizon trend with residual-bootstrap uncertainty bands (grey regions in Fig. 7b)[39]. These projections are reported only for interpretive context and are explicitly distinguished from measured values. Table 5 presents

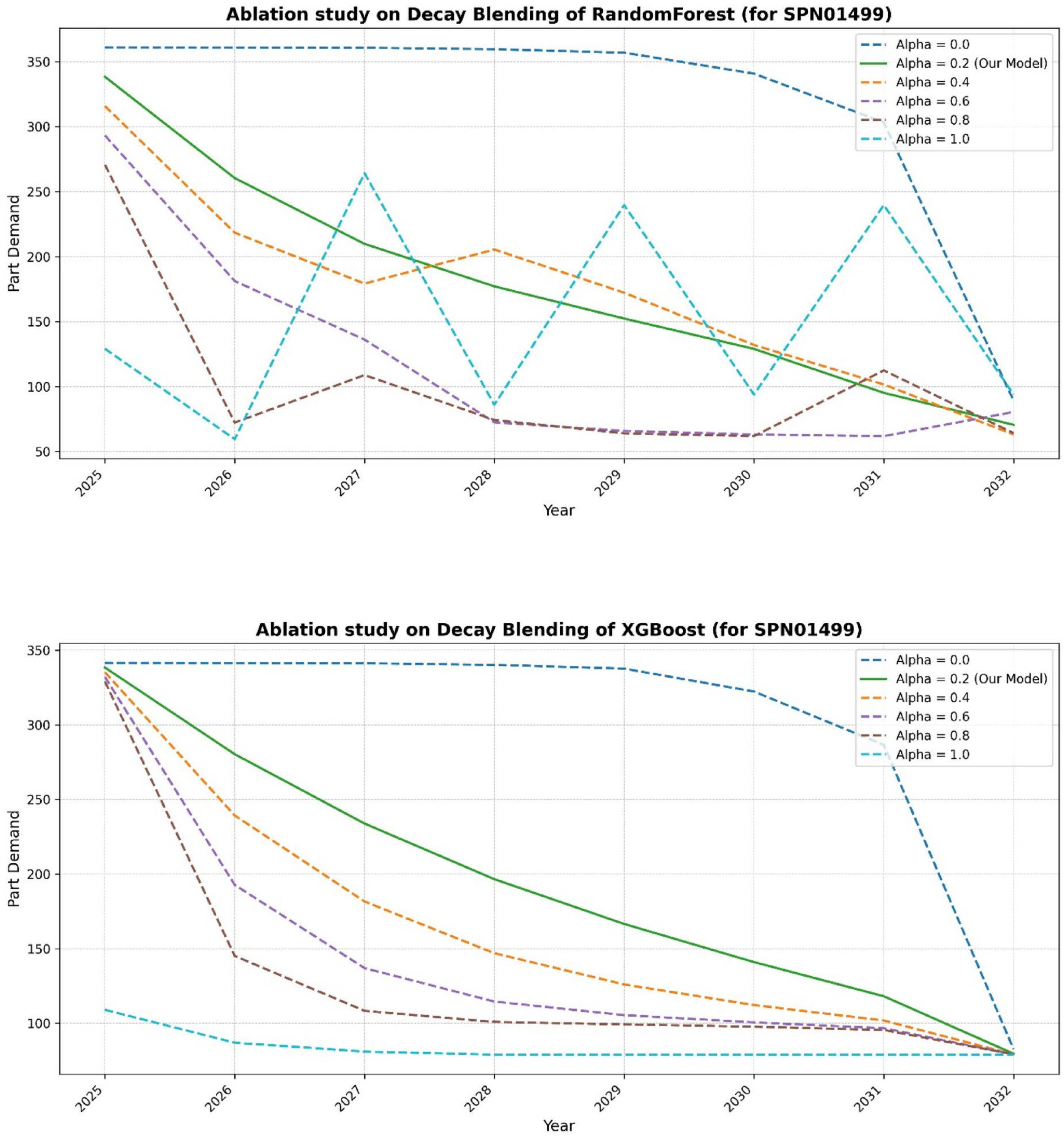

Fig. 7. Ablation study on decay blending of Random Forest model for SPN01499. Intermediate blending (α = 0.2–0.4) produces smoother, physically plausible demand decline compared with unblended (α = 0.0) or fully decayed (α = 1.0) predictions. Ablation study on decay blending of XGBoost model for SPN01499 with horizon-wise projection.

α	MAE	RMSE	Safe MAPE (%)	Δ Safe MAPE vs. α = 0 (%)
0.0	118.2	762.4	5.18	–
0.2	115.9	741.4	**4.36**	−15.8
0.4	120.7	785.9	4.92	−5.0
0.6	130.3	830.6	5.74	+ 10.8
0.8	141.9	902.2	6.51	+ 25.7
1.0	155.0	978.4	7.04	+ 36.0

Table 5. Ablation of decay-blending coefficient α on averaged validation errors (2019–2024).

the averaged validation errors across α values. The lowest Safe MAPE (≈ 4.3%) and smoothest horizon profile are achieved at α = 0.2, adopted as the optimal blending parameter in this study.

Limitations and future work

While the proposed AI/ML-driven forecasting framework significantly improves LTB decision-making for EOL spare parts, certain limitations must be acknowledged. These limitations stem primarily from assumptions in dropout modeling, data availability constraints, and the deterministic nature of the predictive pipeline.

A key limitation lies in the assumption of static dropout rates for estimating the active vehicle population. Although a smoothed annual decay factor was applied based on production volumes and historical attrition trends, the model does not currently incorporate real-time vehicle deregistration or scrappage data. This may result in inaccuracies in the projected installed base, especially in markets with high geographic or regulatory variability. Future iterations could incorporate dynamic dropout models updated from registration databases, sensor telemetry, or macroeconomic indicators (e.g., fuel prices, insurance renewals, or scrappage incentives). For SKUs with dense transactional histories, future work will explore multi-resolution designs (transactional features aggregated and reconciled to annual targets) or hierarchical time-series reconciliation to exploit fine-grain signals without re-introducing zero-inflation bias.

A current limitation of the proposed framework is that it generates deterministic point forecasts without quantifying predictive uncertainty. In future implementations, this can be addressed through probabilistic or Bayesian learning extensions. Quantile Regression Forests (QRF) can replace the standard Random Forest to estimate conditional quantiles of demand distribution rather than single mean predictions, thereby providing upper and lower confidence bounds for each SKU-year forecast. Alternatively, Bayesian Ensemble methods such as Monte-Carlo dropout networks or deep Bayesian regressors can approximate posterior distributions over model parameters and yield prediction intervals that reflect epistemic and aleatoric uncertainty. When combined with the exponential-decay blending mechanism, these uncertainty estimates would enable generation of probabilistic demand trajectories and confidence-weighted Last-Time-Buy (LTB) quantities. Embedding such uncertainty-aware modelling directly within enterprise planning systems would enhance risk-adjusted inventory management and decision transparency under high-variability EOL conditions.

The study also assumes complete availability and integrity of historical demand, warranty, and vehicle volume data for all parts. In practice, data fragmentation and legacy system constraints can introduce missingness or inconsistencies that challenge model training. Developing robust data imputation strategies and anomaly detection methods will be critical for industrial deployment at scale.

In terms of modeling scope, the present framework focuses solely on single-echelon forecasting for spare part demand. It does not explicitly account for multi-tiered supply chains, inventory buffers at regional warehouses, or lead time variability. Integrating this forecasting layer with inventory optimization, distribution planning, and multi-echelon simulation models could unlock further downstream efficiencies.

Additionally, the current study treats each SKU independently and does not yet exploit transfer learning across part clusters with similar failure modes or life-cycle profiles. Future extensions may incorporate part similarity networks, failure taxonomies, or hierarchical models that learn from grouped part behavior, thereby improving generalization in low-data regimes.

Finally, while the framework provides accurate long-range demand forecasts, it does not yet include a cost-optimization module to translate forecast outputs into financially optimal LTB quantities. A logical next step would be to integrate inventory holding cost, obsolescence penalties, and service level constraints into a prescriptive optimization layer that recommends part-specific LTB order volumes under budgetary and risk trade-offs. While the current approach offers a robust foundation for AI-assisted EOL inventory forecasting, ongoing enhancements are necessary to enable dynamic, scalable, and cost-aware decision-making under real-world operational complexities.

Conclusions

This study presents a data-driven framework for forecasting long-term spare part demand in EOL scenarios, with specific application to LTB planning in the automotive domain. By integrating vehicle attrition modeling, warranty-based failure insights, and engineered lifecycle-aware features into an AI/ML pipeline, the methodology provides a robust mechanism to reduce procurement uncertainty and improve inventory resilience over extended service periods. The approach leverages recursive multi-year forecasting using tree-based regressors augmented with exponential decay adjustments to generate physically grounded demand estimates across an 8-year horizon. Key findings from the modeling and validation efforts are summarized below:

1. The proposed framework was tested on 1709 SKUs spanning the period 2015–2024 and achieved strong generalization on unseen data. The Random Forest model emerged as the most reliable, yielding the lowest Safe MAPE of 4.36%, compared to 52.8% for CatBoost and over 470% for ElasticNet on the test set.
2. Visualization of forecasted demand confirmed lifecycle-aligned decay in high-volume parts such as SPN00001, while blended forecasts for SPN01499 demonstrated smooth transition from observed to extrapolated values, indicating the success of decay-informed blending strategies.
3. A cost-sensitive LTB optimization approach was proposed using part-specific forecast aggregates and uncertainty buffers. For instance, SPN01330 exhibited an 8-year forecast of 2,509 units, with a buffer-adjusted recommendation of 2,567 units using a 25% standard deviation margin. This allows planners to proactively hedge against variability while minimizing capital lock-in.
4. Feature analysis confirmed the independence of critical predictors such as vehicle base, replacement rate, and decay slope, ensuring model interpretability and reducing multicollinearity bias during training.
5. The recursive deployment strategy, integrated with Python-based visualization and export routines, enables integration into real-time planning dashboards or enterprise resource systems, facilitating actionable use in OEM workflows.

In addition to achieving high predictive accuracy, ablation studies demonstrated that the decay-function blending mechanism effectively stabilizes forecast trajectories and mitigates horizon-wise error growth, ensuring realistic lifecycle alignment. The framework's modular architecture and API-based integration with ERP/SAP environments confirm its practical deployability for large-scale industrial inventory systems, bridging the gap between research prototypes and production-grade decision support.

The proposed forecasting strategy demonstrates significant potential to transform how LTB quantities are estimated for service parts. By fusing domain knowledge of lifecycle dynamics with scalable machine learning models and interpretable metrics, the framework addresses both the technical and operational dimensions of long-horizon inventory planning. Future extensions may incorporate uncertainty quantification, multi-echelon supply chain modeling, and cost optimization layers to further enhance robustness and economic value for global OEMs navigating complex service obligations.

Data availability

All the data and material used in this study is available in the manuscript, and further details if required, the corresponding author will provide the same, through proper requisition.

Received: 1 August 2025; Accepted: 1 December 2025
Published online: 03 December 2025

References

1. Ghobbar, A. A. & Friend, C. R. An introduction to spare parts inventory control. *J. Qual. Maintenance Eng.* **9** (2), 108–124. https://doi.org/10.1108/13552510310476688 (2003).
2. Cohen, M. A. & Agrawal, V. Gaining competitive advantage through product support services. *Eur. J. Oper. Res.* **112** (1), 17–33. https://doi.org/10.1016/S0377-2217(98)00266-8 (1999).
3. Dekker, R., Inderfurth, K. & Van Der Laan, E. A review of the last-time buy problem. *J. Oper. Res. Soc.* **53** (10), 1089–1100. https://doi.org/10.1057/palgrave.jors.2601424 (2002).
4. Teunter, R. H. Demand forecasting for spare parts subject to phase-out. *Int. J. Prod. Econ.* **91** (3), 275–284. https://doi.org/10.1016/S0925-5273(03)00151-5 (2004).
5. Inderfurth, K. Remanufacturing and repair in spare parts management. In Spare Parts Inventory Management (pp. 255–285). Springer. (2020). https://doi.org/10.1007/978-3-030-25233-9_11
6. Van der Laan, E. & Teunter, R. The value of information in closed-loop supply chains. *Manuf. Service Oper. Manage.* **4** (3), 280–295. https://doi.org/10.1287/msom.4.3.280.7664 (2002).
7. Boylan, J. E. & Syntetos, A. A. Intermittent demand: linking forecasting to inventory management. *Int. J. Logistics Res. Appl.* **9** (2), 119–132. https://doi.org/10.1080/13675560500405244 (2006).
8. Croston, J. D. Forecasting and stock control for intermittent demands. *Oper. Res. Q.* **23** (3), 289–303. https://doi.org/10.1111/j.2517-6161.1972.tb00871.x (1972).
9. Syntetos, A. A. & Boylan, J. E. The accuracy of intermittent demand estimates. *Int. J. Forecast.* **21** (2), 303–314. https://doi.org/10.1016/j.ijforecast.2004.10.001 (2005).
10. Syntetos, A. A., Boylan, J. E. & Croston, J. D. On the forecasting of intermittent demand: consistent and biased methods. *J. Oper. Res. Soc.* **56** (9), 1075–1081. https://doi.org/10.1057/palgrave.jors.2601881 (2005).
11. Ren, K. et al. Deep recurrent survival analysis. In *Proceedings of the AAAI conference on artificial intelligence* (Vol. 33, No. 01, pp. 4798–4805). (2019)., https://doi.org/10.1609/aaai.v33i01.33014798
12. Chen, G. H. An introduction to deep survival analysis models for predicting time-to-event outcomes. Found.Trends® Mach. Learn. **17** (6), 921–1100. https://doi.org/10.48550/arXiv.2410.01086 (2024).
13. Carbonneau, R., Laframboise, K. & Vahidov, R. Application of machine learning techniques for supply chain demand forecasting. *Eur. J. Oper. Res.* **184** (3), 1140–1154. https://doi.org/10.1016/j.ejor.2006.12.004 (2008).
14. Makridakis, S., Spiliotis, E. & Assimakopoulos, V. Statistical and machine learning forecasting methods: concerns and ways forward. *PloS One.* **13** (3), e0194889. https://doi.org/10.1371/journal.pone.0194889 (2018).
15. Faizanbasha, A. & Rizwan, U. Deep learning-stochastic ensemble for RUL prediction and predictive maintenance with dynamic mission abort policies. *Reliab. Eng. Syst. Saf.* **259**, 110919. https://doi.org/10.1016/j.ress.2025.110919 (2025).
16. Qin, Y., Zhao, Y., Qi, J. & Mao, Y. Spatial-temporal multi-sensor information fusion network with prior knowledge embedding for equipment remaining useful life prediction. *Reliab. Eng. Syst. Saf.* https://doi.org/10.1016/j.ress.2025.111420 (2025).
17. Faizanbasha, A. & Rizwan, U. Optimizing burn-in and predictive maintenance for enhanced reliability in manufacturing systems: A two-unit series system approach. *J. Manuf. Syst.* **78**, 244–270. https://doi.org/10.1016/j.jmsy.2024.12.002 (2025).
18. Barros, O., Weber, R. & Reveco, C. Demand analysis and capacity management for hospital emergencies using advanced forecasting models and stochastic simulation. *Oper. Res. Perspect.* **8**, 100208. https://doi.org/10.1016/j.orp.2021.100208 (2021).
19. Paull, S., Bubak, A. & Stuckenschmidt, H. Machine learning for master production scheduling: combining probabilistic forecasting with stochastic optimisation. *Expert Syst. Appl.* https://doi.org/10.1016/j.eswa.2025.126586 (2025).

20. Fildes, R. & Petropoulos, F. Studies of the accuracy of bayesian demand models. *J. Forecast.* **34** (5), 382–399. https://doi.org/10.1002/for.2345 (2015).
21. Trapero, J. R., Kourentzes, N. & Casas, L. M. A state space framework for forecasting intermittent demand. *Int. J. Prod. Econ.* **143** (2), 475–485. https://doi.org/10.1016/j.ijpe.2012.09.026 (2013).
22. Willemain, T. R., Smart, C. N., Schwarz, H. L. & Sun, E. A comparison of statistical methods for forecasting intermittent demand for service parts. *Int. J. Forecast.* **20** (4), 599–607. https://doi.org/10.1016/j.ijforecast.2004.03.001 (2004).
23. Teunter, R., Syntetos, A. A. & Babai, M. Z. Intermittent demand: linking forecasting and inventory control. *Eur. J. Oper. Res.* **214** (3), 525–535. https://doi.org/10.1016/j.ejor.2010.10.014 (2011).
24. Lange, F. Dynamic pricing with waiting and price-anticipating customers. *Oper. Res. Perspect.* https://doi.org/10.1016/j.orp.2025.100337 (2025).
25. Nathan, B. S., Reddy, S., Sastry, B. V., Krishnaiah, C. C., Eswaramoorthy, K. V. & J., & Innovative framework for effective service parts management in the automotive industry. *Front. Mech. Eng.* **10**, 1361688. https://doi.org/10.3389/fmech.2024.1361688 (2024).
26. Li, J., Jia, L. & Zhou, C. Deep fuzzy inference system fused with probability density function control for wind power forecasting with asymmetric error distribution. *J. Clean. Prod.* https://doi.org/10.1016/j.jclepro.2025.145590 (2025).
27. Cakmak, E. & Guney, E. Spare parts inventory classification using neutrosophic fuzzy EDAS method in the aviation industry. *Expert Syst. Appl.* **224**, 120008. https://doi.org/10.1016/j.eswa.2023.120008 (2023).
28. Jiang, P. & Xu, X. Service parts demand forecasting using warranty data and service information. *Int. J. Prod. Econ.* **133** (2), 577–583. https://doi.org/10.1016/j.ijpe.2011.06.005 (2011).
29. Dieulle, L., Simon, C., Karimi, S. & Cavalcante, C. A. V. Condition-based maintenance optimization with spare parts provisioning. *Reliab. Eng. Syst. Saf.* **150**, 99–108. https://doi.org/10.1016/j.ress.2015.12.001 (2016).
30. Waller, M. A. & Fawcett, S. E. Data science, predictive analytics, and big data: A revolution that will transform supply chain design and management. *J. Bus. Logistics.* **34** (3), 77–84. https://doi.org/10.1111/jbl.12010 (2013).
31. Kourentzes, N. Intermittent demand forecasting using a recurrent group method. *Int. J. Prod. Econ.* **156**, 185–192. https://doi.org/10.1016/j.ijpe.2013.12.032 (2014).
32. Carbonneau, R., Laframboise, K. & Vahidov, R. Application of machine learning techniques for supply chain demand forecasting. *Eur. J. Oper. Res.* **184** (3), 1140–1154. https://doi.org/10.1016/j.ejor.2006.12.016 (2008).
33. Makridakis, S., Spiliotis, E. & Assimakopoulos, V. Statistical and machine learning forecasting methods: concerns and further research directions. *Int. J. Forecast.* **34** (4), 635–675. https://doi.org/10.1016/j.ijforecast.2017.11.001 (2018).
34. Boulaksil, Y. Safety stock placement in supply chains with demand forecast updates. *Oper. Res. Perspect.* **3**, 27–31. https://doi.org/10.1016/j.orp.2016.07.001 (2016).
35. Shahin, M., Chen, F. F., Maghanaki, M., Firouzranjbar, S. & Hosseinzadeh, A. Evaluating the fidelity of statistical forecasting and predictive intelligence by utilizing a stochastic dataset. *Int. J. Adv. Manuf. Technol.* https://doi.org/10.1007/s00170-024-14505-8 (2024).
36. Fildes, R. & Kingsman, B. Incorporating demand uncertainty and forecast error in supply chain planning models. *J. Oper. Res. Soc.* **62** (3), 483–500. https://doi.org/10.1057/jors.2010.40 (2011).
37. Ngaffo, A. N., Ayeb, W. E. & Choukair, Z. Service recommendation driven by a matrix factorization model and time series forecasting. *Appl. Intell.* https://doi.org/10.1007/s10489-021-02478-0 (2022).
38. Kammoun, M. A. et al. Deep learning framework for multi-demand forecasting and joint prediction of production, distribution, and maintenance across multiple manufacturing sites. *Int. J. Adv. Manuf. Technol.* **136** (5), 2349–2376. https://doi.org/10.1007/s00170-024-14916-7 (2025).
39. Xia, H., Han, J. & Milisavljevic-Syed, J. Predictive modeling for the quantity of recycled end-of-life products using optimized ensemble learners. *Resour. Conserv. Recycl.* **197**, 107073. https://doi.org/10.1016/j.resconrec.2023.107073 (2023).

Acknowledgements

The authors sincerely thank Ford Motor Private Limited, the Ford Global Technology & Business Center, Ford Sales & Services (Thailand) Co., Ltd. and the Indian Institute of Information Technology Design and Manufacturing (IIITDM) Kurnool for their valuable support in facilitating real-time research pertaining to supply chain management in the automotive sector.

Author contributions

Sendhil Nathan B : Conceptualization, Data curation, Formal analysis, Investigation, Project administration, Validation, Visualization, Writing – original draft, Writing – review & editing. **B. Veera Siva Reddy** : Conceptualization, Data curation, Formal analysis, Investigation, Project administration, Validation, Visualization, Writing – original draft, Writing – review & editing. **Shenbaga Sujan V S** : Data curation, Formal analysis, Investigation. **C. Chandrasekhara Sastry** : Conceptualization, Methodology, Software, Formal analysis, Investigation, Data curation, Supervision, Writing – review & editing. **J. Krishnaiah** : Administrative support, Co-supervision. **Santi Jitpichitchai** : Conceptualization, Methodology, Data curation, Data Supervision, Project Administration.

Declarations

Competing interests

The authors declare no competing interests.

Ethical approval

All ethical guidelines of the journal have been followed.

Consent to participate

All the authors declare and give their consent to participate in and publication of this research work.

Additional information

Supplementary Information The online version contains supplementary material available at https://doi.org/10.1038/s41598-025-31171-2.

Correspondence and requests for materials should be addressed to B.V.S.R. or C.C.S.

Reprints and permissions information is available at www.nature.com/reprints.