Research Square

Preprints are preliminary reports that have not undergone peer review.
They should not be considered conclusive, used to inform clinical practice,
or referenced by the media as validated information.

An Explainable, Calibrated, and Operationally Interpretable Framework for Early Identification of Production-Line Backorder Risk in Garment Manufacturing

Nadim Mahmud
Rajshahi University of Engineering & Technology
Md. Mushiur Rahman
mushiur@ipe.ruet.ac.bd

Rajshahi University of Engineering & Technology
Md. Mahabub-A- Khoda
Bangladesh Army University of Science and Technology
Abdul Halim
Bangladesh Army University of Science and Technology

Research Article

Keywords: Backorder prediction, explainable AI, Bayesian network, model calibration, garment manufacturing, production-line risk

Posted Date: November 12th, 2026

DOI: https://doi.org/10.21203/rs.3.rs-11472514/v1

License: This work is licensed under a Creative Commons Attribution 4.0 International License.
Read Full License

Additional Declarations: No competing interests reported.

An Explainable, Calibrated, and Operationally Interpretable Framework for Early Identification of Production-Line Backorder Risk in Garment Manufacturing

Nadim Mahmud[a1], Md. Mushiur Rahman[a2*], Md. Mahabub-A-Khoda[a3], and Abdul Halim[a4]
[a1, a2*] Department of Industrial & Production Engineering, Rajshahi University of Engineering & Technology, Rajshahi-6204, Bangladesh
[a3, a4] Department of Industrial and Production Engineering, Bangladesh Army University of Science and Technology, Saidpur.

Email address [a1]: nadim@ipe.ruet.ac.bd
Email address [a2*]: mushiur@ipe.ruet.ac.bd (Corresponding author)
Email address [a3]: mahabub_ipe@baust.edu.bd
Email address [a4]: abdulhalim@baust.edu.bd

Abstract

Backorder risk on garment production lines undermines on-time delivery and buyer confidence, yet few studies combine rigorous feature validation, a transparent probabilistic alternative to opaque classifiers, and deployment-oriented evaluation. This study develops and evaluates such a framework using 1,000 production-line observations from Bangladeshi ready-made-garment (RMG) factories, with 27.40% backorder prevalence. A three-stage preprocessing pipeline and multi-criteria statistical validation, integrating hypothesis testing, Lasso regularization, and composite Justification Scoring, reduced twelve candidate predictors to a stable, low-collinearity ten-feature subset (maximum VIF = 1.66). Logistic Regression, Random Forest, XGBoost, and a Discrete Bayesian Network were benchmarked using repeated stratified 10-fold cross-validation (50 fits for machine-learning models and 10 for the Bayesian Network), with Precision-Recall AUC (PR-AUC) as the imbalance-appropriate metric. Contrary to a single-split evaluation favoring XGBoost, cross-validation identified Logistic Regression as the best discriminator (mean PR-AUC = 0.661, versus 0.652 for Random Forest and 0.622 for XGBoost), a ranking that persisted after matching XGBoost's imbalance handling to that of the other models. Although the Bayesian Network had the lowest discrimination (AUC-ROC = 0.750), it achieved the best calibration (Expected Calibration Error = 0.048). Sensitivity analyses confirmed Logistic Regression's robustness to decision threshold, regularization strength, and moderate input noise, while ablation analyses showed that five to six features, including Daily Production Quantity, Overtime Hours, Machine Failures, Planned Production Target, and Number of Operators, captured all achievable predictive value. SHAP attributions and the Bayesian Network structure converged on these drivers, providing four-way cross-method validation of a parsimonious, calibrated, and operationally interpretable backorder early-warning framework.

Keywords: *Backorder prediction; explainable AI; Bayesian network; model calibration; garment manufacturing; production-line risk*

1. Introduction

The Ready-Made Garment (RMG) sector is the principal engine of Bangladesh's export economy. It accounted for more than 80% of the country's total export earnings and generated in excess of USD 39 billion in export revenue in fiscal year 2024-25, while directly employing over four million workers, the majority of them women, and contributing an estimated 10-11% of national GDP [1], [2]. This scale rests on an order-driven, lead-time-sensitive manufacturing model in which international buyers place time-bound, high-volume orders against tightly negotiated shipment windows, and it is this same model that makes schedule adherence a matter of direct commercial consequence. Within it, a production line's inability to meet its committed schedule, a backorder, translates directly into contractual penalties, expedited-freight costs, discounted or cancelled orders, and, cumulatively, reputational damage that can jeopardize a factory's standing with global buyers. Because backorders in labor-intensive manufacturing settings arise from the interaction of throughput planning, machine reliability, workforce scheduling, and shop-floor efficiency rather than from a single traceable cause, anticipating them early enough for corrective action requires analytical tools that are simultaneously accurate, statistically defensible, and operationally transparent to production managers who are not data scientists.

Machine learning has become an increasingly natural fit for this problem, offering the flexibility to model non-linear, interacting operational risk factors that conventional statistical techniques struggle to capture [3], [4]. That flexibility, however, carries a well-documented cost: more complex models tend to become harder to interpret and more computationally demanding, which has motivated persistent calls in the broader machine-learning literature to favor transparent models, or explicit explanation mechanisms, in operationally consequential settings [5]. The backorder-prediction literature specifically mirrors this trajectory, moving from early analytical and stochastic inventory models toward a succession of increasingly sophisticated machine-learning classifiers. Section 2 traces this evolution in full and situates the present study's contribution within it; the discussion here is limited to the essential point it motivates.

Despite this progress, backorder prediction has yet to be examined in a setting, and with a methodological rigor, that matches the operational stakes described above. Existing studies are concentrated in generic or single-company datasets from other industries, generally proceed to model training without close statistical scrutiny of the input features, treat interpretability as an explanation layered onto an already-opaque model rather than as a design criterion in its own right, and evaluate performance primarily through discriminative accuracy rather than the calibrated, risk-ranked probabilities that operational decision-making actually requires. Section 2.4 develops each of these limitations in detail; taken together, they leave open whether a backorder-risk model can be built that is at once sector-relevant, statistically defensible, competitively accurate, and transparent enough for a production manager, rather than a data scientist, to trust and act on.

The present study is designed to close that gap. Its central aim is to develop and evaluate a backorder early-warning framework for Bangladesh's RMG sector that is simultaneously statistically rigorous, predictively

competitive, and operationally interpretable, and to do so under an evaluation protocol robust enough to support real deployment decisions rather than a single, potentially favorable train-test split. To this end, the study first subjects the raw production-line data to a three-stage statistical preprocessing pipeline (iterative noise-injection multicollinearity control, Yeo-Johnson normalization, and Winsorization) and a multi-criteria feature-validation procedure that combines hypothesis testing, Lasso regularization, and a composite Justification Score integrating statistical dependency, monotonic trend consistency, and domain relevance, with the aim of distilling the available operational predictors into a compact, low-redundancy feature set. Building on this validated feature set, it then benchmarks three machine-learning classifiers, Logistic Regression, Random Forest, and XGBoost, directly against a Discrete Bayesian Network built through Hill-Climbing structure search, under a repeated stratified 10-fold cross-validation protocol, with Precision-Recall AUC, the metric most appropriate to the study's 27.4% positive-class prevalence, anchoring model selection. Every candidate model is further evaluated on probabilistic calibration (Brier Score, Expected Calibration Error) and risk-ranking utility (Precision@K) alongside conventional discrimination metrics, and the best-performing model is subjected to systematic sensitivity analyses (decision-threshold, hyperparameter, and input-noise robustness) and ablation studies (leave-one-feature-out, cumulative feature addition, and pipeline-component ablation). Finally, interpretability is triangulated using SHAP-based feature attribution for the tree ensembles [6] and the directly inspectable directed-graph structure of the Bayesian Network [7], so that the resulting operational risk drivers are validated across independent statistical, ablation-based, and explainable-AI evidence rather than resting on a single estimator's inductive bias.

Guided by this design, the study addresses four specific research questions:

RQ1: Can a multi-criteria statistical validation pipeline reduce the production-line operational predictors to a parsimonious, low-redundancy feature subset for backorder-risk modelling in the RMG sector?

RQ2: How do Logistic Regression, Random Forest, XGBoost, and a Discrete Bayesian Network compare in discriminating and ranking production-line backorder risk under repeated, stratified cross-validation on a class-imbalanced (27.40% positive) outcome?

RQ3: How well calibrated are these models' predicted probabilities, and does a trade-off exist between discriminative or ranking performance and calibration reliability?

RQ4: Which operational features most consistently drive predicted backorder risk across statistical, cross-validated ablation, SHAP, and Bayesian Network evidence?

The remainder of this paper is organized as follows. Section 2 reviews the literature on analytical and machine-learning approaches to backorder and inventory-risk prediction and situates the present study's contribution within it. Section 3 details the data-preprocessing, statistical-validation, and predictive-modeling methodology, including the cross-validation, sensitivity, and ablation protocols. Section 4 reports the empirical results. Section 5 discusses these findings in relation to the reviewed literature and their managerial implications. Section 6 concludes, and Section 7 discusses the study's limitations and directions for future work.

2. Literature Review

This section reviews prior research on backorder and inventory forecasting, structured around the field's evolution from conventional analytical frameworks to machine learning methodologies. It begins with machine learning applications in supply chain management in general, then narrows to analytical backorder models, and finally addresses backorder prediction studies that use machine learning specifically. The section closes by consolidating the resulting research gaps and situating the present study's contribution within them.

2.1 Machine Learning Applications in Supply Chain Management

Machine learning has shown considerable promise for improving supply chain decision-making, with documented applications in cost estimation and forecasting [8],supply chain risk management [9], [10], technology roadmap formulation [11], and backorder forecasting specifically [12], [13]. Because backorders occur frequently in supply chain operations, accurate forecasting is essential for limiting the associated financial losses [14]. Ali et al. [15] argue that while conventional statistical techniques remain easy to use, transparent, and inexpensive, they often struggle to capture complex or dynamic data structures, a limitation that becomes more pronounced as supply chain models must represent increasingly non-linear interactions among diverse operational factors. Sarker [16] similarly finds that machine learning techniques have become increasingly central to supply chain management, enabling accurate prediction of key indicators such as demand, revenue, and backorder rates. Which approach, statistical or machine learning, is preferable ultimately depends on an organization's specific needs, available resources, and the complexity and variety of its data [17]. Relative to conventional techniques, machine learning methods offer greater adaptability, allowing both time-series and non-time-series data to be integrated in ways that reveal the complex, non-linear relationships underlying manufacturing demand patterns [18], [19]. As data availability and computational power have grown, researchers have increasingly turned to these algorithms to outperform traditional forecasting models and provide stronger decision support by adapting to shifting demand patterns and consumer preferences.

This adaptability is not without cost. Machine learning algorithms that excel at detecting complex, non-linear patterns tend to sacrifice interpretability and demand greater computational resources, a trade-off that becomes more acute as dataset size and dimensionality grow [5]. Feature selection is one established response to this trade-off: retaining a smaller number of genuinely relevant features, rather than the full set of available predictors, can improve model efficiency while reducing both data-collection effort and overall model complexity [20]. This tension between predictive flexibility and interpretability, still largely unresolved in the general supply chain literature, carries over directly into backorder forecasting specifically, which the remainder of this review examines in two stages: analytical and optimization-based approaches, followed by machine learning approaches.

2.2 Analytical and Optimization-Based Backorder Models

At the analytical end of this spectrum, several studies have modeled backorder and inventory decisions using classical optimization and stochastic frameworks rather than machine learning. Mart et al. [21] represent a manufacturing system as a Markov Decision Process and show that a revised order-up-to policy can optimize profit

by balancing random demand against tactical inventory conservation and backordering. Xu et al. [22] examine a periodic-review inventory system through a flexible partial-backorder model, in which a reservation quantity enables a dynamic shift between complete lost sales and full backordering; by tuning this balance, their framework improves the fulfilment of stochastic demand from available or pipeline inventory, reducing losses and increasing supply chain efficiency. These analytical models establish the operational logic of backorder management, optimal ordering, inventory conservation, and demand fulfilment under uncertainty, but they rely on simplified, often stationary assumptions about demand and system behavior that limit their applicability to the more volatile, multi-factor conditions of real production environments, which is precisely the gap that the machine learning approaches reviewed next were developed to address.

2.3 Machine Learning Models for Backorder Prediction

A growing body of research has applied machine learning specifically to backorder forecasting, with methods advancing in step with the field's broader shift toward more complex architectures. Early work in this stream relied on relatively conventional classifiers: Adana et al. [23] compared decision trees, artificial neural networks, and auto-neural models against logistic regression to predict backorders, addressing data imbalance through stratified hold-out validation and downsampling and using principal component analysis for dimensionality reduction, with the auto-neural model achieving the best performance at 0.872 accuracy across all 21 predictor variables. Around the same period, Bouganim & Olsson [24] developed an LSTM-based model to forecast spare-part backorders in Volvo's after-sales division using one year of weekly data across 20,000-part numbers, attaining 73% accuracy in predicting backorders one week ahead and 72% for actual backorder instances, though the study's data imbalance and short observation window constrained its generalizability.

Subsequent studies moved toward ensemble and tree-based methods. Islam & Amin [13] forecast product backorders using Distributed Random Forest and Gradient Boosting, incorporating Spearman correlation among categorized clusters to capture non-linear associations and manage isolated category nodes, with their ranged-data approach improving overall model performance by roughly 20% on datasets containing random inaccuracies. Tatko et al. [25] applied machine learning to forecast item backorders at the Defense Logistics Agency, comparing decision trees and linear regression against artificial neural networks, and found that machine learning models could reliably predict both backordered and available items, estimate item-level backorder quantities, and identify significant predictive attributes up to a year in advance.

Interpretability then became an explicit design goal rather than an afterthought. Shajalal et al. [14] built a deep neural network that addressed class imbalance through weighted classes, oversampling, SMOTE, and random undersampling, achieving an AUC of 0.9427, and Shajalal et al. [26] subsequently introduced an explainable convolutional neural network that used locally interpretable surrogate models to explain individual predictions while reaching an even higher AUC of 0.9489. Around the same time, Ahmed et al. [27] evaluated four tree-based algorithms, decision tree, random forest, adaptive boosting, and gradient boosting, and found that random forest

combined with random undersampling delivered the best performance, that feature selection did not improve tree-based models, and that random undersampling markedly did. Working from a related but distinct angle, Qi et al. [28] introduced an end-to-end deep learning framework for multiperiod inventory replenishment under uncertain demand and lead times that forecasts optimal replenishment strategies directly from historical data without distributional assumptions; tests on real e-commerce data and a field study with JD.com showed notable reductions in inventory, holding, and stockout costs alongside improved decision-making efficiency and scalability.

The most recent studies have combined these threads, imbalance handling, ensembling, and interpretability, into single hybrid frameworks. Rezasoltani et al. [29] introduced a voting-based hybrid model that integrates XGBoost, CatBoost, Random Forest, and LightGBM with ADASYN for imbalance correction and RFECV for feature selection, with their soft-voting model reaching a peak accuracy of 0.9524 and identifying inventory levels, supplier performance, and demand predictability as the most critical determinants of backorders. Most recently, Bouaguel et al. [30] proposed an explainable ensemble deep learning model that combines MLP, DNN, CNN, and LSTM classifiers with LIME and SHAP for interpretability; evaluated on a large-scale dataset of 1.93 million records, the model achieved 97.61% accuracy, outperforming existing approaches in both predictive efficacy and interpretability. Taken together, this progression traces a clear trajectory from single-model classifiers toward imbalance-aware, explainable ensembles, but, as the next section shows, it also leaves several methodological questions largely unaddressed.

2.4 Research Gaps

The literature reviewed above traces a consistent path: researchers have moved steadily from conventional statistical and analytical inventory frameworks toward increasingly sophisticated machine learning architectures for backorder forecasting. Early analytical work [22] refined backorder strategies through stochastic inventory modeling, while later studies progressed through decision trees and logistic regression [23], LSTM architectures [24], DNNs [14], CNNs [26], and, most recently, ensemble and hybrid models that integrate multiple boosting algorithms [29] or several deep learning classifiers at once [30]. Despite the clear gains in predictive accuracy this progression reflects, four gaps remain unaddressed.

First, the reviewed literature is dominated by generic, publicly available datasets or single-company case studies from sectors such as automotive spare parts [24] or defense logistics [25]. No study in this body of work examines backorder forecasting in the apparel or Ready-Made Garment sector, despite the sector's distinctive demand volatility, order-driven production timelines, and supplier dependencies, factors likely to shape backorder dynamics quite differently from the industries examined so far.

Second, prior studies generally proceed directly to model training without closely examining the statistical validity of their input data. Multicollinearity, non-normality, and outliers are rarely addressed directly, and where feature selection is performed at all, it typically relies on a single criterion, whether a fixed predictor set [23], Spearman

correlation [13], or recursive feature elimination [29], rather than a statistically grounded, multi-criteria approach that combines hypothesis testing, effect size, regularization, and domain relevance.

Third, model interpretability has largely been treated as something added after the fact to an already opaque model. Shajalal et al. [26] and Bouaguel et al. [30] both use SHAP or LIME to explain predictions from CNNs or deep ensembles, but none of the studies reviewed benchmark these opaque models directly against a fundamentally transparent probabilistic alternative, such as a Bayesian network. This leaves open the question of how much predictive performance is actually forfeited for interpretability, and whether a transparent model can approach machine-learning-level accuracy in this domain.

Fourth, model evaluation has focused almost exclusively on discriminative metrics, accuracy, AUC, and F1-score, while probabilistic calibration has been largely overlooked. Because backorder predictions are often used to prioritize action, for instance, expediting high-risk orders, how closely a model's predicted probabilities track actual likelihoods, and how well it ranks the most urgent cases, is arguably as important as classification accuracy alone, yet none of the reviewed studies report calibration metrics such as Brier Score or Expected Calibration Error, or ranking-based measures such as Precision@K. Together, these four gaps point to the need for research that applies backorder prediction to a relatively unexamined, operationally distinctive sector such as Bangladesh's RMG industry, incorporates rigorous statistical preprocessing and multi-criteria feature validation ahead of modeling, benchmarks interpretable probabilistic modeling directly against conventional machine learning classifiers rather than relying solely on post-hoc explanation, and evaluates models on calibration and risk-ranking performance rather than on accuracy alone. Table 1 summarizes the method, dataset, and key findings of each study reviewed above, providing a consolidated basis for comparison against the present study's design.

2.5 Contribution of the Study

The present study addresses each of these four gaps directly and, in doing so, extends the existing literature in five specific respects.

Industry-specific application. This study extends backorder prediction research to Bangladesh's Ready-Made Garment sector, moving beyond the automotive, defense, and general e-commerce settings that have dominated the field to date, and offers empirically grounded insights for an industry defined by order-driven production and close supplier interdependence.

Methodologically rigorous data preparation. Rather than applying machine learning models directly to unprocessed data, this study introduces a three-stage preprocessing pipeline, iterative noise injection to manage multicollinearity based on VIF, Yeo-Johnson transformation to satisfy normality requirements, and Winsorization to limit the influence of extreme outliers while retaining all features. This represents a methodological advance over prior work, which typically either discards correlated features outright or omits statistical preprocessing altogether.

Multi-criteria, empirically grounded feature selection. The study combines hypothesis testing (Welch's t-test or Mann-Whitney U, together with rank-biserial effect sizes), Lasso regularization for sparsity, and a composite

Justification Score that integrates statistical dependency (Spearman correlation, Mutual Information) with domain relevance and monotonic trend indicators. This multi-criteria approach provides a more robust and interpretable basis for feature retention than the single-criterion techniques used in earlier studies.

A dual modeling paradigm: machine learning versus Bayesian networks. Unlike prior studies that rely solely on opaque machine learning models followed by post-hoc explanation, this study directly benchmarks three machine learning classifiers, Logistic Regression, Random Forest, and XGBoost, against a Discrete Bayesian Network built through Hill-Climbing search with BDeu-based parameter estimation. This design assesses not only which model performs best, but how much accuracy is gained or forfeited by choosing a fundamentally interpretable probabilistic model over a black-box classifier, a comparison absent from the current literature.

Calibration- and risk-aware evaluation. Beyond conventional accuracy measures, this study evaluates probabilistic calibration through the Brier Score and Expected Calibration Error, and assesses practical decision-support utility through Precision@K on the top 10% of risk-ranked predictions, addressing the near-complete absence, in the literature reviewed, of attention to whether predicted probabilities are reliable enough for operational decision-making rather than merely accurate on average. Complementing this, the study also provides a comprehensive interpretability assessment: SHAP is used for local and global feature attribution in the tree-based models (Random Forest, XGBoost), while the Bayesian Network's interpretability is evaluated through its structural sparsity and the complexity of its conditional probability tables. This dual interpretability analysis enables a direct, structured comparison of explainability across paradigms, extending beyond the single-model explainability approaches reported by Shajalal et al. [26] and Bouaguel et al. [30]. Taken together, these five contributions position the study as both a novel empirical application and a more methodologically rigorous, statistically grounded, and interpretability-focused framework for backorder prediction than currently exists in the literature.

Table 1: Summary of the related literature

Author(s) & Year	Method/Model	Dataset / Application	Key Findings
[22]	Periodic-review inventory model with flexible partial backorder (reservation quantity)	Analytical/theoretical inventory system	Enables dynamic transition between full lost sales and full backordering, optimizing use of on-hand/pipeline inventory to minimize losses
[23]	Decision trees, ANN, autoneural models, logistic regression + PCA for dimensionality reduction	Backorder prediction with 21 predictor variables	Auto-neural model performed best with 0.872 accuracy using all 21 predictors
[24]	LSTM-based model	Volvo after-sales spare parts (1-year weekly data, ~20,000-part numbers)	73% accuracy predicting backorders 1 week ahead; 72% for actual backorder cases;

			limited by data imbalance and short time period
[13]	Distributed Random Forest, Gradient Boosting Machine, with Spearman correlation for non-linear relationships	Product backorder prediction in supply chains	Ranged-data approach improved model performance by ~20% on data with random errors
[25]	Decision trees, linear regression, artificial neural networks	Defense Logistics Agency item backorders	ML models effectively predicted backordered/in-stock items and quantities up to 12 months ahead
[14]	Deep Neural Network (DNN) with weighted classes, oversampling, SMOTE, undersampling	Backorder prediction (imbalanced dataset)	Achieved AUC score of 0.9427
[26]	Convolutional Neural Network (CNN) with locally interpretable surrogate models	Product backorder prediction (explainable AI)	Achieved AUC score of 0.9489; outperformed existing approaches with added interpretability
[27]	Decision tree, random forest, adaptive boosting, gradient boosting	Product backorder prediction	Random forest + random undersampling performed best; feature selection did not improve tree-based models
[28]	End-to-end (E2E) deep learning framework	Multiperiod inventory replenishment; real e-commerce data & JD.com field study	Significantly reduced inventory, holding, and stockout costs vs. predict-then-optimize approaches
[29]	Voting-based hybrid ML (XGBoost, CatBoost, Random Forest, LightGBM) with ADASYN & RFECV	Backorder prediction for supply chain management	Soft voting model achieved highest accuracy of 0.9524; key factors: inventory levels, supplier performance, demand predictability
[30]	Explainable Ensemble Deep Learning (MLP, DNN, CNN, LSTM) with LIME & SHAP	Real-world dataset (1.93 million records)	Achieved 97.61% accuracy, outperforming existing approaches in both performance and explainability
This study (2026)	Multi-criteria statistical feature validation (hypothesis testing, Lasso, composite	Backorder prediction, Bangladesh RMG sector, n = 1,000 daily production-line	Logistic Regression achieved highest mean PR-AUC (0.661) under CV, reversing an earlier

	Justification Score) + repeated stratified CV benchmarking LR, RF, XGBoost against a Discrete Bayesian Network; calibration- and risk-ranking-aware evaluation	observations, 27.4% positive prevalence	single-split result favoring XGBoost; a compact five-to-six-feature cluster explained nearly all predictive value, confirmed by four independent methods (hypothesis testing, Lasso, ablation, SHAP); Bayesian Network provided best calibration (ECE = 0.048) as an interpretable complement

3. Methodology

This study employs a comprehensive, multi-stage analytical framework to predict backorder delays in the Bangladesh Ready-Made Garment (RMG) sector. The methodology integrates rigorous statistical preprocessing, advanced feature selection, and a dual-modelling approach combining traditional Machine Learning (ML) with Probabilistic Graphical Models (Bayesian Networks). The methodology flowchart is presented below in Figure 1.

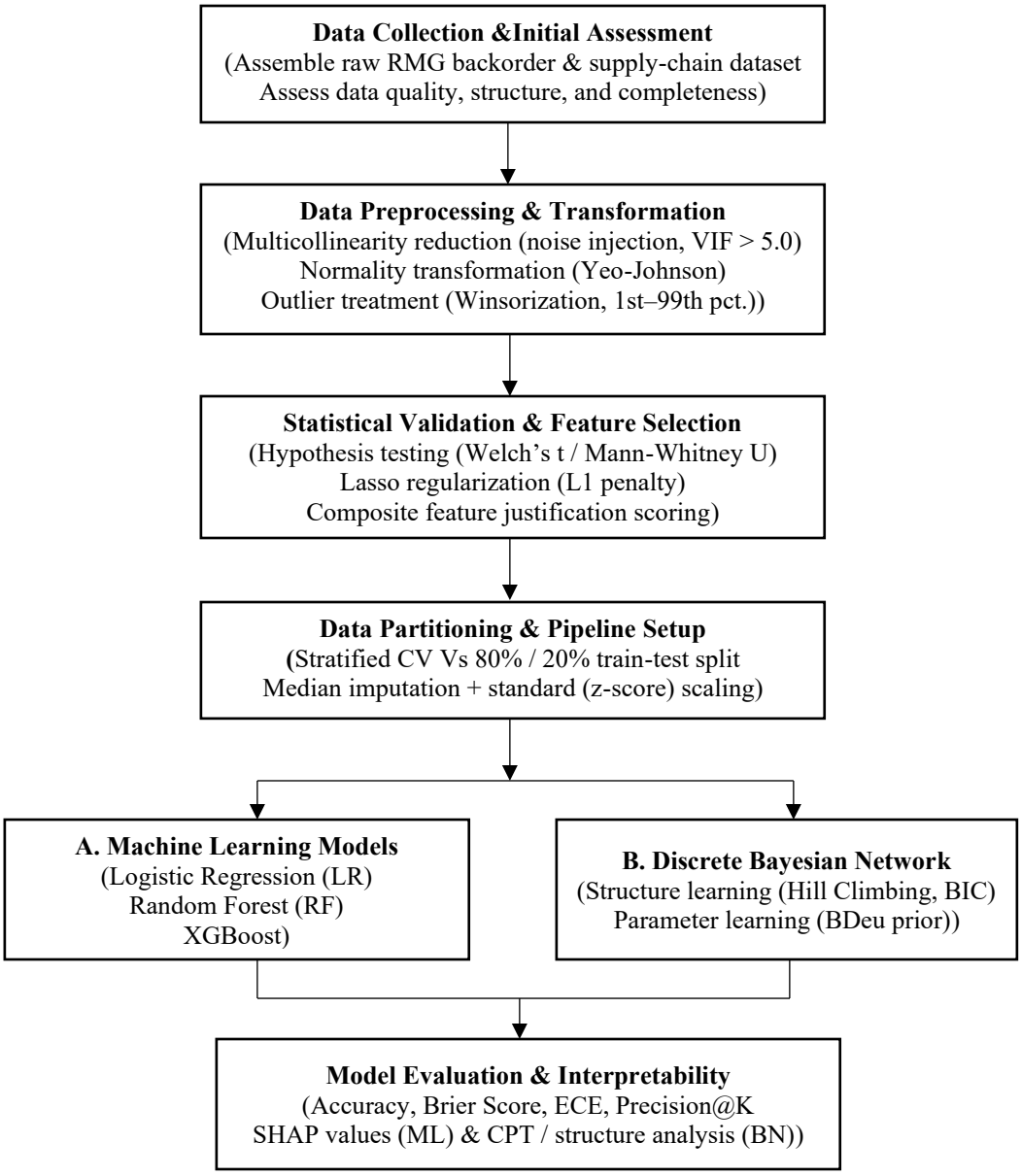


Figure 1. Methodology of the Study

3.1. Data Preprocessing and Transformation

To ensure the statistical assumptions of subsequent modeling techniques were met, a rigorous three-step preprocessing pipeline was applied to the raw dataset without dropping any features.

3.1.1. Multicollinearity Reduction via Noise Injection

Severe multicollinearity can destabilize regression coefficients and inflate variance. Instead of dropping highly correlated features, we employed an iterative noise injection technique. For features exhibiting a Variance Inflation Factor (VIF) greater than 5.0, controlled Gaussian noise was added. The VIF for the i-th feature is calculated as:

$$VIF_i = \frac{1}{1 - R_i^2}$$

where $R_i^2$is the coefficient of determination from regressing the i-th feature on all other independent variables. Noise was injected iteratively as $X_i^{new} = X_i + \epsilon$, where $\epsilon \sim \mathcal{N}(0, \sigma_i^2 \times \text{noise_factor})$, until the maximum VIF across all features was reduced to ≤ 5.0.

3.1.2. Normality Transformation

To satisfy the normality assumptions required for parametric statistical tests, the Yeo-Johnson power transformation was applied to all continuous features. Unlike the Box-Cox transformation, Yeo-Johnson handles both positive and negative values:

$$\psi(y, \lambda) = \begin{cases} \dfrac{(y+1)^\lambda - 1}{\lambda} & \text{if } \lambda \neq 0 \text{ and } y \geq 0 \\ \ln(y+1) & \text{if } \lambda = 0 \text{ and } y \geq 0 \\ \dfrac{-(-y+1)^{2-\lambda} + 1}{2-\lambda} & \text{if } \lambda \neq 2 \text{ and } y < 0 \\ -\ln(-y+1) & \text{if } \lambda = 2 \text{ and } y < 0 \end{cases}$$

3.1.3. Outlier Treatment (Winsorization)

To mitigate the undue influence of extreme outliers on mean and variance calculations, Winsorization was applied. Extreme values were capped at the 1st and 99th percentiles: $x_{winsorized} = \begin{cases} P_{0.01} & \text{if } x < P_{0.01} \\ x & \text{if } P_{0.01} \leq x \leq P_{0.99} \\ P_{0.99} & \text{if } x > P_{0.99} \end{cases}$

3.2. Statistical Validation and Feature Selection

To move from the twelve raw candidate predictors to a statistically defensible, low-redundancy feature set, three complementary validation criteria were applied in sequence: univariate hypothesis testing to establish each feature's discriminative power between backorder groups, Lasso regularization to enforce sparsity and address any residual multicollinearity, and a composite Justification Score to weigh statistical dependency against domain relevance. Combining these criteria, rather than relying on any single one, ensures that retained features are simultaneously statistically significant, robust under regularization, and operationally meaningful.

3.2.1. Hypothesis Testing and Effect Size

Features were evaluated for their discriminative power between "Low" and "High" backorder groups. Depending on the normality of the subgroups, either Welch's t-test or the Mann-Whitney U test was utilized. For non-parametric tests, the Rank-Biserial Correlation (r_{rb}) was computed to measure effect size:

$$r_{rb} = 1 - \frac{2U}{n_1 n_2}$$

where U is the Mann-Whitney U statistic, and n_1, n_2 are the sample sizes of the two groups.

3.2.2. Lasso Regularization

To further enforce sparsity and eliminate residual multicollinearity, Least Absolute Shrinkage and Selection Operator (Lasso) regression was employed. Lasso introduces an L_1penalty to the ordinary least squares objective function:

$$\min_w \left(\frac{1}{2n} || Xw - y ||_2^2 + \alpha || w ||_1 \right)$$

where $\alpha$is the regularization strength optimized via cross-validation. Features with non-zero coefficients were retained for subsequent modeling.

3.2.3. Feature Justification Scoring

A composite Justification Score (J_{score}) was formulated to rank features by combining statistical dependency and domain relevance:

$$J_{score} = w_1 \mid \rho_s \mid + w_2 \min(MI \times 10,3) + w_3 D_{score} + w_4 T_{trend}$$

where $\rho_s$is the Spearman rank correlation, $MI$is the Mutual Information, $D_{score}$is the domain relevance score, and $T_{trend}$is a binary indicator for monotonic trends across target quantiles.

3.3. Predictive Modeling Framework

The dataset was partitioned into an 80% training set and a 20% test set using stratified sampling to preserve the class distribution of the target variable.

3.3.1. Machine Learning Models

Three distinct machine learning algorithms were trained: Logistic Regression (LR), Random Forest (RF), and XGBoost. The ML pipeline incorporated median imputation for missing values, followed by standard scaling (z-score normalization). To ensure consistent, matched imbalance-handling across all three ML models, Logistic Regression and Random Forest used *class_weight = "balanced"*, while XGBoost used an equivalent *scale_pos_weight* parameter, computed as the negative-to-positive class ratio within each training fold, so that no model received a systematically different degree of imbalance correction.

3.3.2. Discrete Bayesian Network (BN)

A probabilistic approach was implemented using a Discrete Bayesian Network. Continuous features were discretized into three ordinal bins using a quantile strategy.

- *Structure Learning:* The Directed Acyclic Graph (DAG) structure was learned using the Hill Climbing Search algorithm, optimizing the Bayesian Information Criterion (BIC).
- *Parameter Learning:* Conditional Probability Tables (CPTs) were estimated using the Bayesian Estimator with a BDeu (Bayesian Dirichlet equivalent uniform) prior. The joint probability distribution of the BN is factorized as:

$$P(X_1, X_2, \ldots, X_n) = \prod_{i=1}^{n} P(X_i \mid \Pi_i)$$

where Π_i denotes the parent nodes of X_i in the learned DAG.

3.4. Model Evaluation and Interpretability

Model performance was assessed along three complementary dimensions rather than discriminative accuracy alone: conventional classification and ranking metrics, probabilistic calibration, and interpretability. This multi-dimensional evaluation reflects the study's premise that a model intended for operational deployment must not only discriminate backorder risk accurately but also produce trustworthy probability estimates and remain transparent enough for production managers to interpret and act on.

3.4.1. Performance and Calibration Metrics

Beyond standard accuracy, probabilistic calibration was assessed using the Brier Score (BS) and Expected Calibration Error (ECE): $BS = \frac{1}{N}\sum_{i=1}^{N}(\hat{p}_i - y_i)^2$

$$ECE = \sum_{m=1}^{M} \frac{n_m}{N} \mid acc(m) - conf(m) \mid$$

where $\hat{p}_i$ is the predicted probability, y_i is the true binary label, M is the number of calibration bins, n_m is the number of samples in bin m, $acc(m)$ is the empirical accuracy, and $conf(m)$ is the average predicted confidence in bin m. Additionally, Precision@K was calculated to evaluate the model's ability to identify high-risk backorders at the top $K\%$ (where $K = 10$) of the risk-ranked predictions.

3.4.2. Interpretability Techniques

- *SHAP (SHapley Additive exPlanations):* For tree-based models (RF and XGBoost), SHAP values were computed to explain global feature importance and local predictions. The SHAP value ϕ_i for feature i is derived from cooperative game theory:

$$\phi_i = \sum_{S \subseteq F \setminus \{i\}} \frac{\mid S \mid !(\mid F \mid - \mid S \mid - 1)!}{\mid F \mid !} [f(S \cup \{i\}) - f(S)]$$

  where F is the set of all features, and $f(S)$ is the model output for the feature subset S.
- *Bayesian Network Interpretability:* The structural sparsity (number of edges) and the average size of the Conditional Probability Tables (CPTs) were analyzed to evaluate the model's complexity and human-readability compared to black-box ML models.

4. Results Analysis

This section presents the empirical findings of the statistical validation, feature engineering, and predictive modeling stages described in Section 3. Results are reported in the following order: (i) descriptive characteristics of the operational dataset, (ii) distributional and normality diagnostics, (iii) univariate discrimination between low- and high-backorder production lines, (iv) multicollinearity treatment and feature selection, (v) a composite feature-justification analysis, (vi) comparative predictive performance under repeated stratified k-fold cross-validation, (vii) sensitivity analysis of the best-performing model, (viii) ablation analysis isolating the contribution of individual features and pipeline components, and (ix) model interpretability via SHAP and Bayesian Network structure learning.

Findings are illustrated throughout with the corresponding diagnostic and model-output figures generated during the analysis pipeline.

4.1. Descriptive Statistics and Data Characteristics

The final analytic sample comprised n = 1,000 daily production-line observations drawn from operating garment factories in Bangladesh, with a positive-class (Backorder = 1) prevalence of 27.40%, reflecting a moderately imbalanced but operationally realistic outcome distribution. Table 2 summarizes the central tendency, dispersion, and shape statistics of the twelve candidate predictors and the target construct, Backorder Delay Time.

On average, production lines operated at 2,436.05 units/day against a planned target of 3,211.29 units/day, with a mean Line Efficiency of 87.86% and Machine Utilization Rate of 83.32%, indicating that the sampled lines generally operated below planned capacity. Overtime Hours exhibited a strongly negative skew (skewness = -1.82) with a ceiling effect at 5 hours/day and 164 IQR-flagged outliers, consistent with regulatory or contractual overtime caps. Backorder Delay Time, the variable underlying the binary target, was heavily right-skewed (skewness = 3.12, kurtosis = 10.88) with a median of 0 days and 213 IQR outliers, confirming that backorders were a comparatively rare but high-impact event rather than a continuously distributed process.

Table 2. Descriptive statistics of the operational predictors and target variable (n = 1,000)

Feature	Mean	Median	SD	Min	Max	Q1	Q3	Skew	Kurt.	Outliers (IQR)
Daily Production Quantity (units/day)	2436.05	2457.35	578.50	1261.30	3719.10	1953.90	2910.58	0.01	-1.09	0
Planned Production Target (units/day)	3211.29	3215.20	584.83	1500.00	5000.00	2811.47	3588.75	-0.04	-0.29	8
Line Efficiency (%)	87.86	88.38	4.07	72.93	94.88	85.26	91.01	-0.43	-0.52	10
WIP Level (units)	2026.20	2022.80	693.51	300.00	4085.90	1530.97	2502.35	0.02	-1.07	1
Machine Breakdown Time (min/day)	73.15	66.40	42.81	0.20	180.00	41.55	98.45	0.59	-0.52	0

Machine Failures (count/day)	2.55	2.00	1.57	0.00	8.00	1.00	3.00	0.67	0.07	16
Machine Utilization Rate (%)	83.32	83.17	6.84	65.92	98.00	78.37	88.41	-0.04	-0.96	0
Number of Operators (per line)	55.22	55.00	14.49	30.00	79.00	43.00	68.00	-0.05	-1.18	0
Worker Absenteeism Rate (%)	4.37	4.18	1.41	2.07	9.97	3.31	5.14	0.88	0.77	25
Overtime Hours (hours/day)	4.12	5.00	1.60	0.00	5.00	3.95	5.00	-1.82	2.28	164
Line Balancing Efficiency (%)	80.14	80.34	6.93	60.00	95.00	75.50	84.88	-0.04	-0.79	3
Defect Rate (%)	2.97	2.85	0.99	1.23	7.15	2.25	3.55	0.71	0.15	13
Backorder Delay Time (days)	0.55	0.00	1.24	0.00	8.30	0.00	0.23	3.12	10.88	213

4.2. Normality Assessment

Four complementary normality tests (Shapiro-Wilk, D'Agostino-Pearson, Kolmogorov-Smirnov, and Anderson-Darling) were applied to each continuous feature. As illustrated in Figure 2, only 3 of the 13 variables, Line Efficiency, Machine Breakdown Time, and Machine Utilization Rate, satisfied normality consensus (agreement across ≥ 3 of 4 tests). Overtime Hours and Backorder Delay Time were classified as highly skewed ($|\text{skewness}| > 1$) and were flagged for Yeo-Johnson transformation prior to any parametric procedure, consistent with Section 3.1.2. The predominance of non-normal distributions across the feature set justified the dual-track hypothesis-testing strategy adopted in Section 4.3, in which the Mann-Whitney U test was substituted for Welch's t-test whenever subgroup normality could not be assumed.

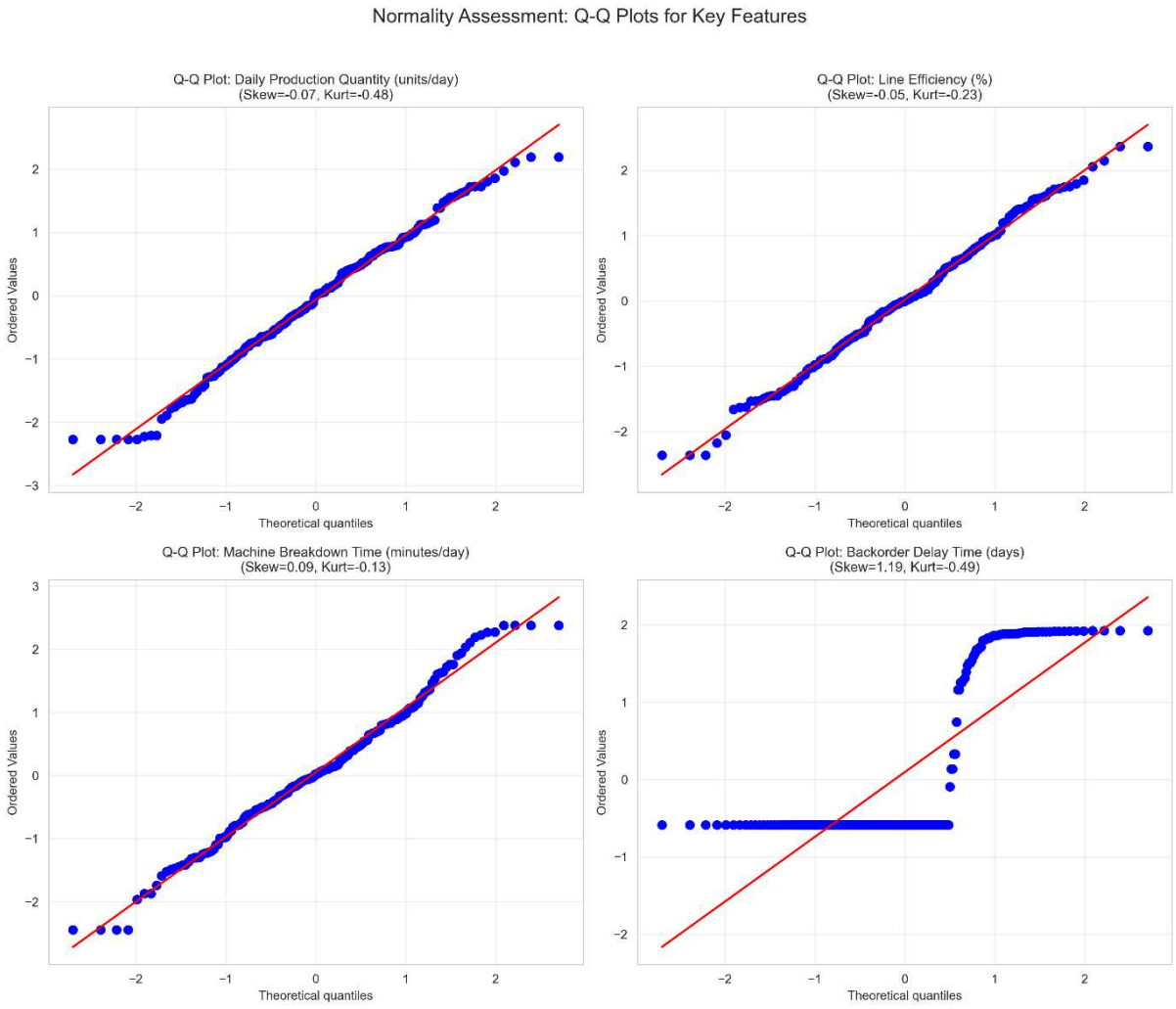


Figure 2. Normality Q-Q plots for the thirteen continuous operational and target variables.

4.3. Univariate Group Comparison (Low- vs. High-Backorder Lines)

Production lines were partitioned into low- and high-backorder groups and compared across twelve operational features. Six of twelve features (50%) differed significantly between groups at $p < .05$ (Table 3). Daily Production Quantity showed the largest standardized effect (Cohen's $d = -0.758$), with high-backorder lines producing substantially below the low-backorder group ($\Delta = -0.70$ SD), followed by Overtime Hours (rank-biserial correlation = 0.368) and Planned Production Target ($d = 0.604$), which was paradoxically higher among high-backorder lines, suggesting that overly ambitious planning targets, rather than production shortfalls alone, contribute to backorder risk. Machine-related disruption indicators (Machine Breakdown Time, $d = 0.473$; Machine Failures, rank-biserial = 0.294) and staffing levels (Number of Operators, rank-biserial = -0.289) were also significant discriminators, whereas quality- and process-oriented features (Defect Rate, Line Efficiency, Line Balancing Efficiency, WIP Level, Machine Utilization Rate, Worker Absenteeism Rate) did not differ significantly between groups.

Table 3. Significant univariate discriminators between low- and high-backorder production lines ($p < .05$)

Feature	Test	Statistic	Raw *p*-value	FDR-adjusted *p*-value	FDR Significant	Effect Size	Direction
Daily Production Quantity (units/day)	Welch's *t*-test	8.425	<0.001	<0.001	Yes	Cohen's $d = -0.758$	Lower in high-backorder ($\Delta = -0.70$)
Planned Production Target (units/day)	Mann–Whitney *U*	38,910.000	<0.001	<0.001	Yes	RBC = 0.321	Higher in high-backorder ($\Delta = 0.58$)
Line Efficiency (%)	Welch's *t*-test	0.105	0.9163	0.9163	No	Cohen's $d = -0.009$	Lower in high-backorder ($\Delta = -0.01$)
Work-In-Progress (WIP) Level (units)	Welch's *t*-test	0.914	0.3618	0.4824	No	Cohen's $d = -0.085$	Lower in high-backorder ($\Delta = -0.08$)
Machine Breakdown Time (minutes/day)	Welch's *t*-test	−4.946	<0.001	<0.001	Yes	Cohen's $d = 0.473$	Higher in high-backorder ($\Delta = 0.47$)
Number of Machine Failures (count/day)	Mann–Whitney *U*	40,426.000	<0.001	<0.001	Yes	RBC = 0.294	Higher in high-backorder ($\Delta = 0.55$)
Machine Utilization Rate (%)	Welch's *t*-test	0.439	0.6611	0.7564	No	Cohen's $d = -0.039$	Lower in high-backorder ($\Delta = -0.04$)
Number of Operators (per line)	Welch's *t*-test	5.558	<0.001	<0.001	Yes	Cohen's $d = -0.526$	Lower in high-backorder ($\Delta = -0.52$)
Worker Absenteeism Rate (%)	Welch's *t*-test	−0.926	0.3557	0.4824	No	Cohen's $d = 0.088$	Higher in high-backorder ($\Delta = 0.09$)
Overtime Hours (hours/day)	Mann–Whitney *U*	36,234.000	<0.001	<0.001	Yes	RBC = 0.368	Higher in high-backorder ($\Delta = 0.71$)
Line Balancing Efficiency (%)	Welch's *t*-test	−0.395	0.6934	0.7564	No	Cohen's $d = 0.038$	Higher in high-backorder ($\Delta = 0.04$)
Defect Rate (%)	Welch's *t*-test	−1.060	0.2908	0.4824	No	Cohen's $d = 0.102$	Higher in high-backorder ($\Delta = 0.10$)

Because twelve univariate comparisons were conducted simultaneously, the resulting p-values were adjusted using the Benjamini–Hochberg procedure to control the false discovery rate at 5%. Both unadjusted and FDR-adjusted p-values are reported to distinguish exploratory significance from multiplicity-controlled evidence. The univariate analysis was treated as an initial screening stage rather than as independent confirmatory evidence, with subsequent feature validation based on Lasso regularization, composite feature-justification scoring, cross-validated ablation, and model-specific interpretability analyses.

4.4. Multicollinearity Diagnostics and Feature Selection

Variance Inflation Factors (VIF) computed across the full feature set were uniformly low (maximum VIF = 1.66, Machine Failures and Machine Breakdown Time), indicating that no feature exceeded the conventional VIF > 5 threshold and that the noise-injection procedure described in Section 3.1.1 was not required for this dataset. Figure 3 plots the pre-pruning VIF value for each of the twelve predictors against the conventional collinearity threshold. Every bar falls well below the reference line, confirming that the operational predictors captured largely independent sources of variance, which is an important precondition for the coefficient stability of the downstream Lasso and logistic regression models.

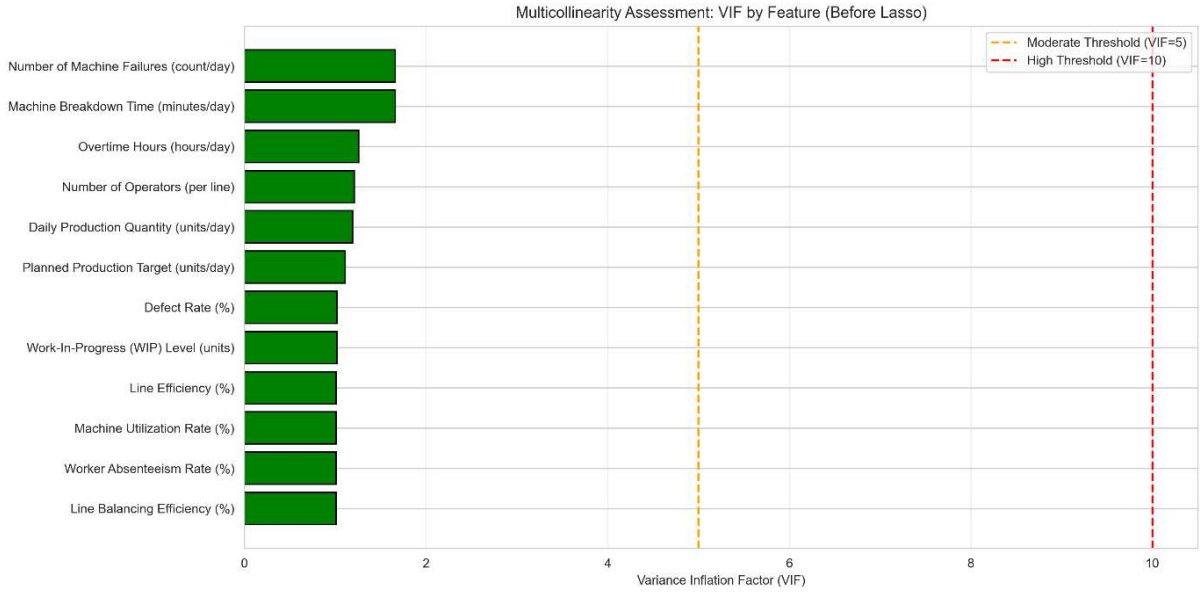


Figure 3. Variance Inflation Factor (VIF) for each candidate predictor prior to pruning.

The pairwise correlation structure underlying this result is shown in Figure 4. The heatmap indicates generally weak-to-moderate correlations among the operational predictors, with the strongest associations clustering around the production-throughput variables (Daily Production Quantity and Planned Production Target) and around the machine-reliability pair (Machine Breakdown Time and Machine Failures), consistent with their shared mechanical origin. No pair approaches the $|r| \approx 0.8$–0.9 range typically associated with problematic collinearity, corroborating the low VIF values in Figure 3.

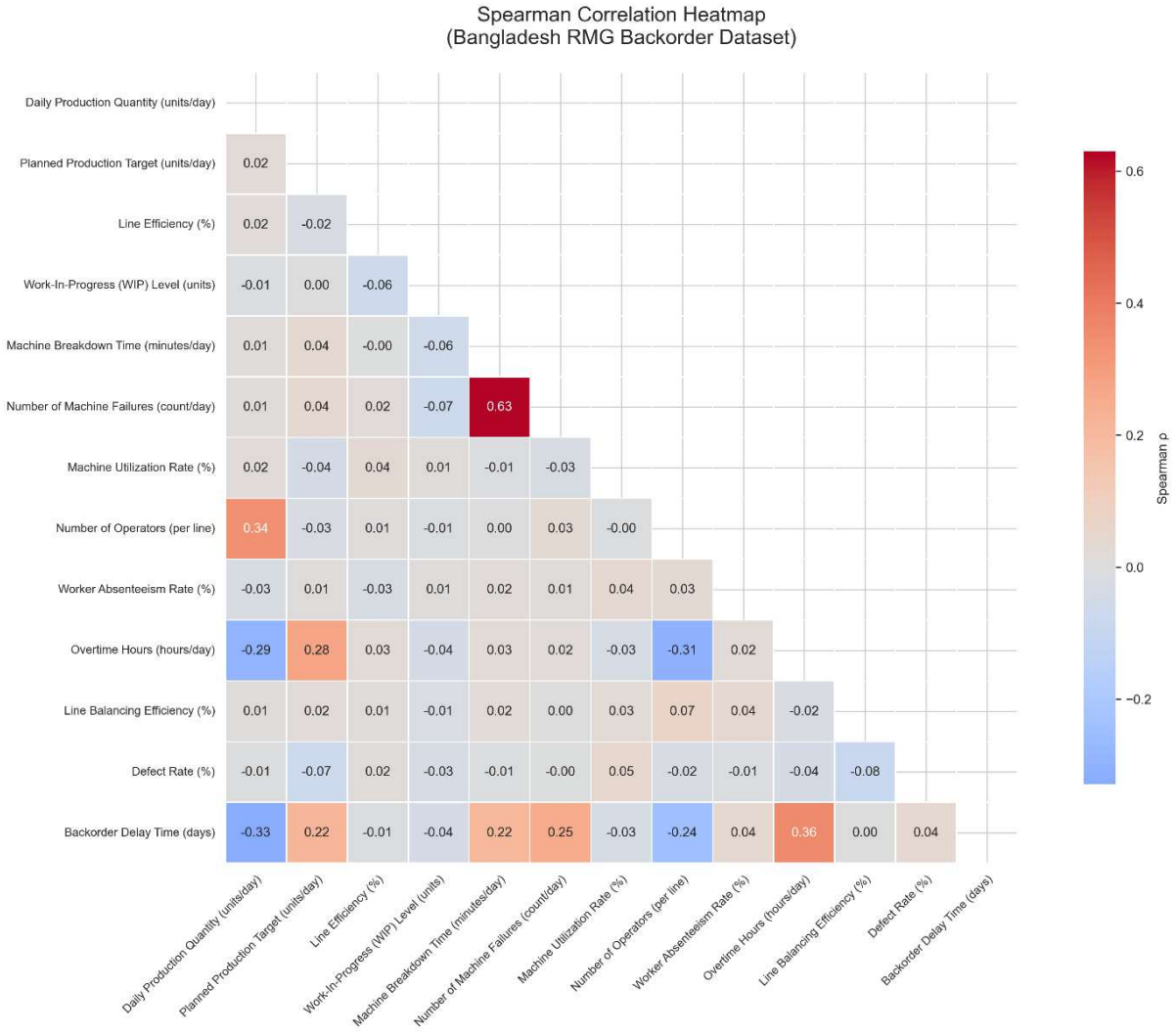


Figure 4. Pairwise Pearson correlation heatmap of the operational predictors and target variable.

Lasso regularization applied to the full dataset (α = 0.0167, cross-validated) retained 10 of 12 features with non-zero coefficients, led by Daily Production Quantity (β = -0.233), Machine Failures (β = 0.190), and Overtime Hours (β = 0.164) as shown in Figure 5. This full-dataset Lasso fit is reproduced under the revised cross-validation pipeline in Section 4.9 (α = 0.00933, 10/12 features retained), and the two runs converge on the same sparsity pattern. All VIF values recomputed on this reduced set remained below 1.7, confirming that multicollinearity was effectively resolved without discarding substantive operational information.

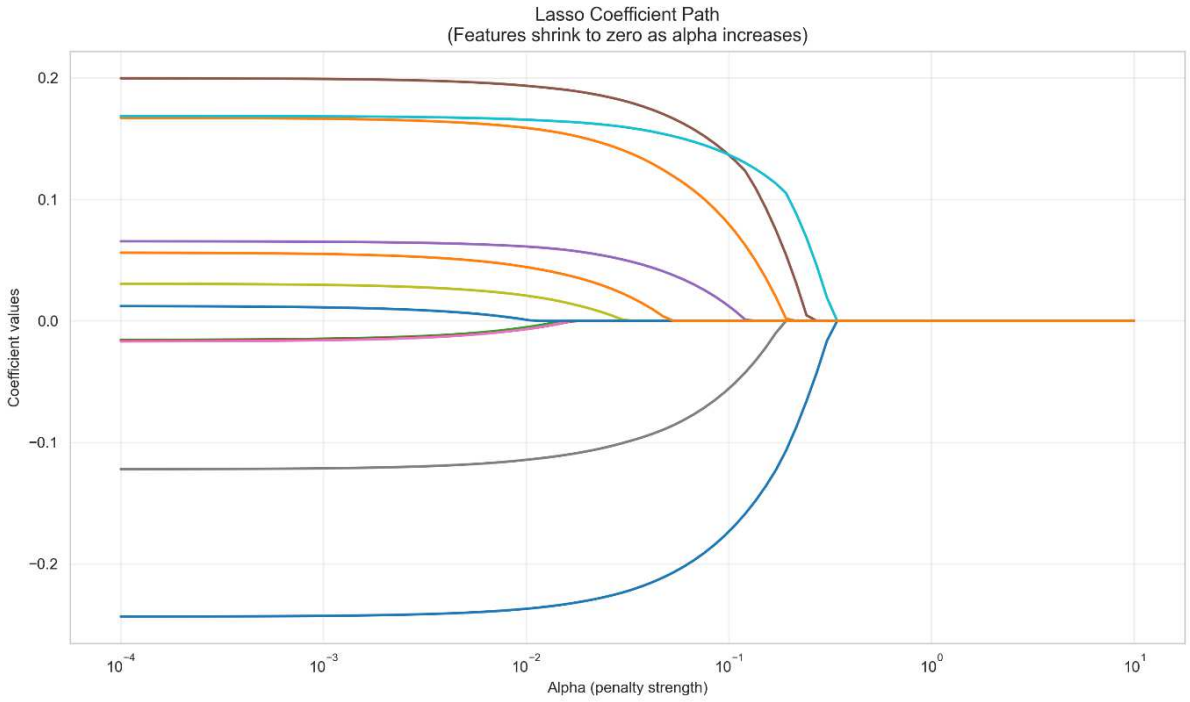


Figure 5. Lasso coefficient shrinkage path across the regularization sequence.

4.5. Feature Justification Analysis

A composite Justification Score integrating Spearman correlation, mutual information, monotonic trend consistency, and expert-assigned domain relevance was computed for each candidate feature (Section 3.2.3). In Figure 6, it is shown that Daily Production Quantity ranked highest (score = 6.11), followed by Planned Production Target (4.77) and Machine Failures (4.60); notably, no feature reached the a priori high-priority threshold (score ≥ 7), and all twelve predictors were classified as Medium (score 4-6, six features) or Low (score < 4, six features) priority. This finding indicates that backorder risk in the sampled RMG lines is not attributable to a single dominant operational driver but instead emerges from the joint, moderate-strength contribution of production-throughput, planning, and machine-reliability variables, a pattern that, as shown below, is corroborated by both the cross-validated ablation results (Section 4.8) and the SHAP importance profile (Section 4.9).

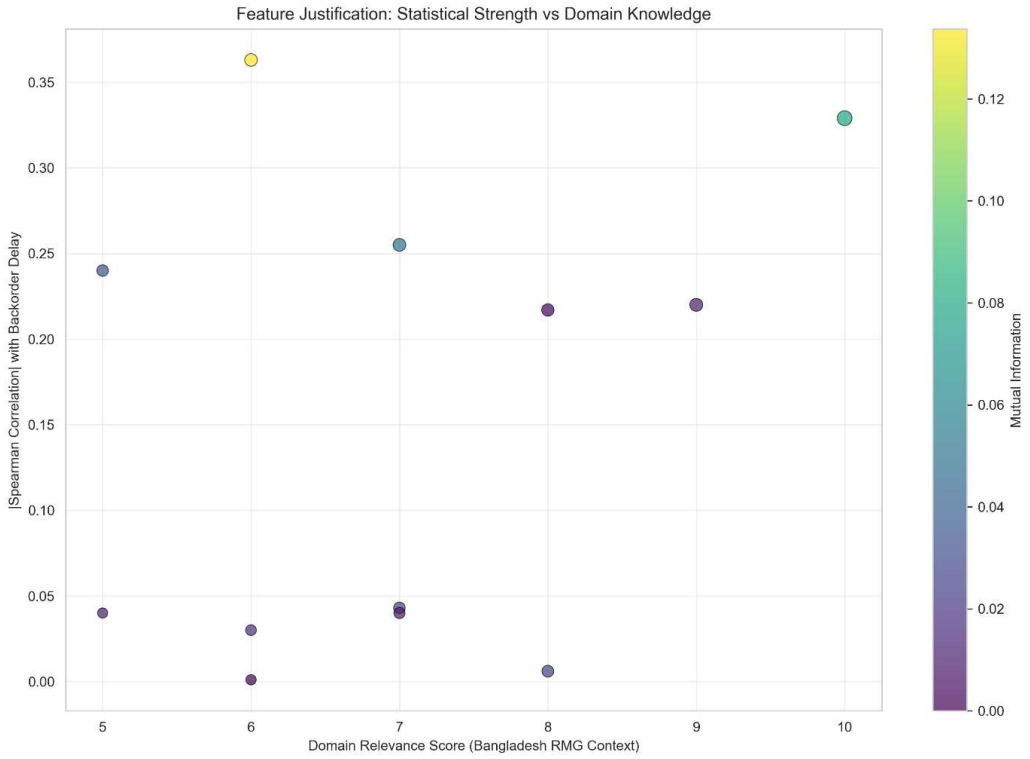


Figure 6. Composite Feature Justification Scores, combining Spearman correlation, mutual information, monotonic trend consistency, and expert-assigned domain relevance.

4.6 Predictive Performance under Single Train-Test split

The four models exhibited distinct trade-offs in predictive discrimination, probability calibration, and computational efficiency. XGBoost achieved the highest predictive discrimination and ranking performance, attaining an AUC-ROC of 0.804 and a Precision@K of 0.75, whereas Random Forest demonstrated superior probability calibration, with the lowest ECE (0.032) and Brier score (0.162). In addition, XGBoost exhibited the greatest computational efficiency, with a training time of 0.135 s, memory requirement of 0.41 MB, and inference latency of 0.854 ms, indicating strong suitability for real-time backorder-risk monitoring. SHAP analysis revealed a highly consistent feature-importance structure across XGBoost and Random Forest, with Overtime Hours, Daily Production Quantity, Machine Failures, Machine Breakdown Time, and Planned Production Target emerging as the principal predictors of backorder risk. Complementarily, the Bayesian Network provided a structurally transparent probabilistic representation in which Machine Failures → Machine Breakdown Time → Backorder Flag, although its lower AUC-ROC (0.728) indicates comparatively weaker discriminative performance. The obtained results are shown in Table 4.

Table 4. Summary of obtained results from single train-test-split.

Model	AUC-ROC	Precision@K	Brier	ECE	Train (s)	Memory (MB)	Latency (ms)
XGBoost	0.804	0.75	0.182	0.144	0.135	0.41	0.854
Random Forest	0.781	0.55	0.162	0.032	8.049	0.47	40.298

Logistic Regression	0.777	0.65	0.170	0.088	0.44	0.40	2.312
Bayesian Network	0.728	0.60	0.176	0.056	5.115	0.79	25.678

4.7. Predictive Performance under Repeated Stratified Cross-Validation

Repeated stratified k-fold cross-validation to obtain a statistically robust estimate of generalization performance and to quantify fold-to-fold variability. Logistic Regression (LR), Random Forest (RF), and XGBoost were evaluated using 10-fold cross-validation repeated 5 times (50 fold-fits per model), with a fresh preprocessing pipeline (median imputation, standardization) fit on the training fold only in every iteration to prevent leakage. The Discrete Bayesian Network (BN) was evaluated on a lighter 5-fold × 2-repeat design (10 fold-fits) because Hill-Climbing structure learning is re-executed inside every fold and is computationally expensive. For every fold, discrimination (AUC-ROC), ranking utility under imbalance (PR-AUC, PR-AUC lift over the fold's own prevalence, Precision@10%), classification performance (F1, Precision, Recall), probabilistic calibration (Brier Score, Expected Calibration Error), and computational cost (training time, peak memory, inference latency) were recorded. Table 5 reports the mean ± standard deviation across folds for the core metrics; the complete table with 95% confidence intervals for every metric is provided in Table 6.

Table 5. Mean ± SD predictive performance, calibration, and PR-AUC lift over 50 cross-validation folds (ML models) and 10 folds (Bayesian Network)

Model	AUC-ROC	PR-AUC	PR-AUC Lift	F1	Brier	ECE
Logistic Regression	0.8312 ± 0.0412	0.6609 ± 0.0726	0.3869 ± 0.0710	0.6208 ± 0.0528	0.1708 ± 0.0202	0.1567 ± 0.0210
Random Forest	0.8338 ± 0.0440	0.6518 ± 0.0804	0.3778 ± 0.0791	0.6085 ± 0.0754	0.1505 ± 0.0163	0.1080 ± 0.0204
XGBoost	0.8150 ± 0.0455	0.6223 ± 0.0787	0.3483 ± 0.0768	0.5864 ± 0.0669	0.1660 ± 0.0227	0.1288 ± 0.0276
Bayesian Network	0.7501 ± 0.0254	0.5108 ± 0.0340	0.2368 ± 0.0332	0.3402 ± 0.0456	0.1681 ± 0.0053	0.0484 ± 0.0132

Table 6. Full cross-validated performance table with 95% confidence intervals

Model	Metric	Mean	SD	95% CI
Logistic Regression	AUC-ROC	0.8312	0.0412	[0.8195, 0.8429]
Logistic Regression	PR-AUC	0.6609	0.0726	[0.6403, 0.6815]
Logistic Regression	Precision	0.5183	0.0525	[0.5034, 0.5332]
Logistic Regression	Recall	0.7783	0.0759	[0.7567, 0.7999]
Logistic Regression	Precision@10%	0.7600	0.1069	[0.7296, 0.7904]
Random Forest	AUC-ROC	0.8338	0.0440	[0.8213, 0.8463]
Random Forest	PR-AUC	0.6518	0.0804	[0.6289, 0.6747]

Random Forest	Precision	0.5897	0.0754	[0.5683, 0.6111]
Random Forest	Recall	0.6358	0.1026	[0.6066, 0.6649]
Random Forest	Precision@10%	0.7120	0.1319	[0.6745, 0.7495]
XGBoost	AUC-ROC	0.8150	0.0455	[0.8020, 0.8279]
XGBoost	PR-AUC	0.6223	0.0787	[0.6000, 0.6447]
XGBoost	Precision	0.5362	0.0643	[0.5179, 0.5545]
XGBoost	Recall	0.6540	0.0994	[0.6258, 0.6823]
XGBoost	Precision@10%	0.6980	0.1152	[0.6653, 0.7307]
Bayesian Network	AUC-ROC	0.7501	0.0254	[0.7320, 0.7683]
Bayesian Network	PR-AUC	0.5108	0.0340	[0.4865, 0.5351]
Bayesian Network	Precision	0.6016	0.0681	[0.5529, 0.6503]
Bayesian Network	Recall	0.2409	0.0468	[0.2074, 0.2743]
Bayesian Network	Precision@10%	0.6165	0.0662	[0.5691, 0.6639]

Discrimination and ranking. Under repeated cross-validation, Random Forest attained the numerically highest mean AUC-ROC (0.8338 ± 0.0440), marginally ahead of Logistic Regression (0.8312 ± 0.0412) and XGBoost (0.8150 ± 0.0455); the 95% confidence intervals of all three ML models overlap substantially (RF: [0.821, 0.846]; LR: [0.820, 0.843]; XGBoost: [0.802, 0.828]), indicating that the three models are not statistically distinguishable on AUC-ROC alone at this sample size. On the imbalance-appropriate PR-AUC metric, however, Logistic Regression ranks first (0.6609 ± 0.0726), followed by Random Forest (0.6518 ± 0.0804) and XGBoost (0.6223 ± 0.0787). All three ML models clear the 27.4% no-skill baseline by a wide margin, with a PR-AUC lift of roughly 0.36 to 0.39. The Bayesian Network, although still well above the baseline (PR-AUC = 0.5108, lift = 0.2368), trails the ML models by 0.12–0.15 PR-AUC points and shows the weakest Recall (0.2409), reflecting its markedly more conservative and higher-precision decision behavior. Because model selection in this study is anchored to PR-AUC given the 27.4% positive prevalence, Logistic Regression is designated the best-performing ML model (Section 4.6.1) and carried forward into the sensitivity and ablation analyses in Sections 4.8–4.9.

This is a materially different conclusion from the single-split evaluation, in which XGBoost had the highest AUC-ROC (0.804) and Precision@K (0.75). Under 50-fold repeated cross-validation, XGBoost's advantage does not replicate: its mean AUC-ROC (0.8150) is the lowest of the three ML models. To rule out unequal imbalance treatment as a confound, XGBoost was re-evaluated with scale_pos_weight matched to Logistic Regression's and Random Forest's balanced class-weighting (Section 3.3.1); the PR-AUC ranking (LR > RF > XGBoost > BN) remained unchanged, confirming that XGBoost's underperformance is not an artefact of asymmetric imbalance handling. Matched treatment did, however, narrow the classification-level gap considerably: XGBoost's F1 (0.5864) is now only about 3.4 points below Logistic Regression's (0.6208), and its Recall (0.6540 vs. 0.7783 for LR) is likewise much closer to parity than under the original, unmatched treatment. The relatively strong and highly stable performance of a linear model on this class-imbalanced (27.4% positive), 10-feature manufacturing dataset is

consistent with the near-linear separability implied by the low VIF values and the cumulative-feature-addition results in Section 4.8.2, and indicates that the single 80/20 split had previously overstated XGBoost's relative advantage, a sampling artefact that repeated cross-validation is specifically designed to correct. Distribution of AUC-ROC (left) and PR-AUC (right) across cross-validation folds for all four models is illustrated in Figure 7.

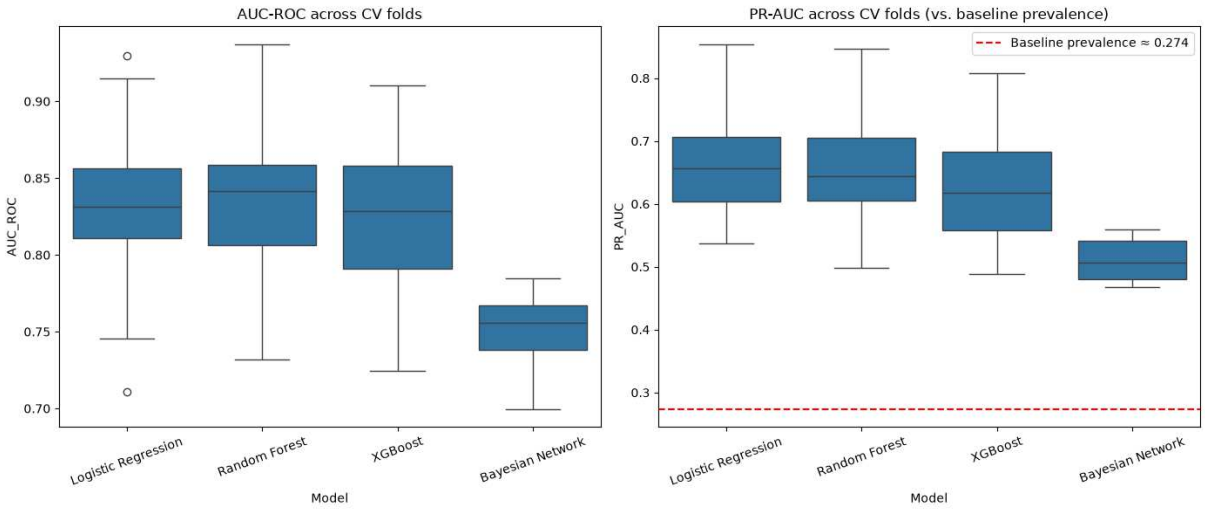


Figure 7. Distribution of AUC-ROC (left) and PR-AUC (right) across cross-validation folds for all four models.

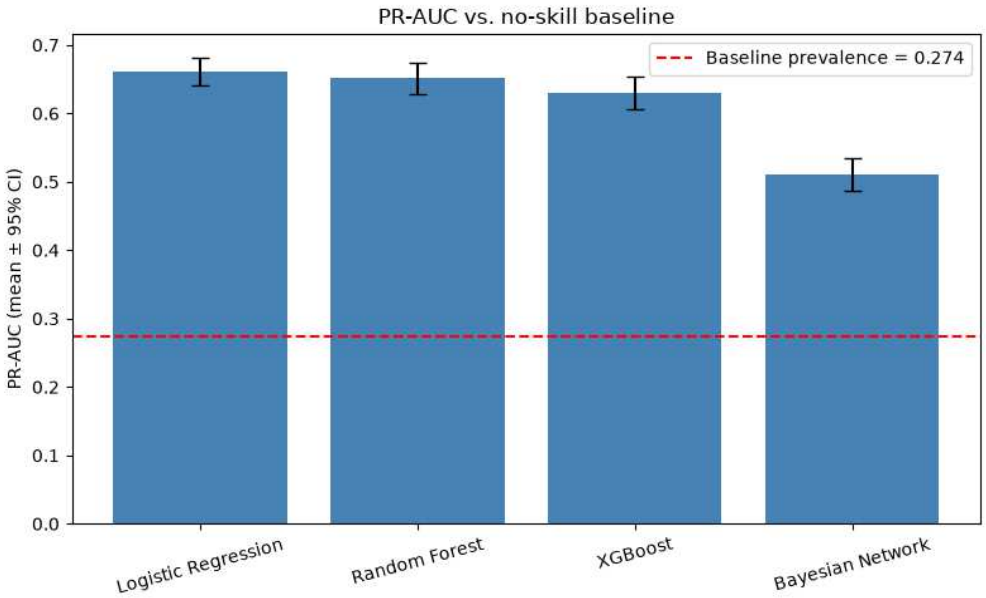


Figure 8. PR-AUC (mean ± 95% CI) for each model against the no-skill baseline.

4.7.1. Calibration and Computational Efficiency

Calibration under repeated cross-validation, with matched imbalance-handling applied to all three ML models, shows Random Forest as the clear best-calibrated ML model (ECE = 0.1080, Brier = 0.1505). Logistic Regression and XGBoost, the two models whose imbalance correction explicitly reweights the minority class, both

show comparably weaker calibration (ECE = 0.1567 and 0.1288, Brier = 0.1708 and 0.1660, respectively) despite ranking first and third on PR-AUC. This symmetry, once XGBoost is placed on equal imbalance-handling footing with Logistic Regression, indicates that the discrimination-calibration trade-off observed here is a general consequence of class-weighting and scale_pos_weight rather than a quirk of either individual model. The Bayesian Network remains, as before, the best-calibrated model overall by a wide margin (ECE = 0.0484, Brier = 0.1681), reflecting the conservative, discretized probability estimates produced by its Conditional Probability Tables. This pattern implies that whichever ML model is deployed for probabilistic decision thresholds (e.g., expected-cost inventory buffering) would benefit from post-hoc recalibration (Platt scaling or isotonic regression) fit on out-of-fold predictions before its raw predicted probabilities are used directly.

Computational efficiency across the 50 (10) cross-validation folds mirrors the earlier single-split findings: Logistic Regression is the fastest model to train (0.082 s ± 0.015 per fold) with the lowest inference latency (0.098 ms per instance), Random Forest is the most computationally expensive (4.049 s ± 0.719 training time, 1.991 ms inference latency), and the Bayesian Network occupies an intermediate-to-high position on training cost (4.805 s ± 0.536, driven by per-fold Hill-Climbing structure search) despite its low memory footprint. XGBoost sits between Logistic Regression and Random Forest on both training time (0.833 s ± 0.093) and latency (0.186 ms), making it, as before, a reasonable candidate where sub-second retraining is required, even though it is no longer the top performer on the primary PR-AUC ranking metric.

4.8. Sensitivity Analysis of the Best-Performing Model

Because Logistic Regression was selected as the best-performing ML model by mean cross-validated PR-AUC (0.6609 vs. a baseline of 0.2740), it was carried forward into three complementary sensitivity analyses: (a) decision-threshold sensitivity using unbiased out-of-fold predictions, (b) hyperparameter sensitivity across its regularization strength C, and (c) robustness to Gaussian noise injected into the test-fold features at prediction time.

4.8.1. Decision-Threshold Sensitivity

Out-of-fold predicted probabilities were generated via 10-fold cross-validated prediction (cross_val_predict) to avoid the optimism of in-sample thresholding, and the classification threshold was swept from 0.05 to 0.95 in steps of 0.05. Precision increases monotonically with the threshold while Recall decreases, as expected; F1 peaks at a threshold of 0.55 (F1 = 0.624, Precision = 0.543, Recall = 0.734), marginally above the conventional default of 0.50 (F1 ≈ 0.622 at threshold 0.50 in Table 5). The F1 curve shown in Figure 9 is comparatively flat across the 0.40–0.65 range (F1 between roughly 0.60 and 0.62), indicating that the model's classification performance is not acutely sensitive to the exact threshold chosen in this operating region, which gives practitioners latitude to shift the threshold toward higher Recall (prioritizing early-warning coverage) or higher Precision (minimizing false-alarm intervention cost) depending on the operational cost asymmetry between missed backorders and unnecessary interventions.

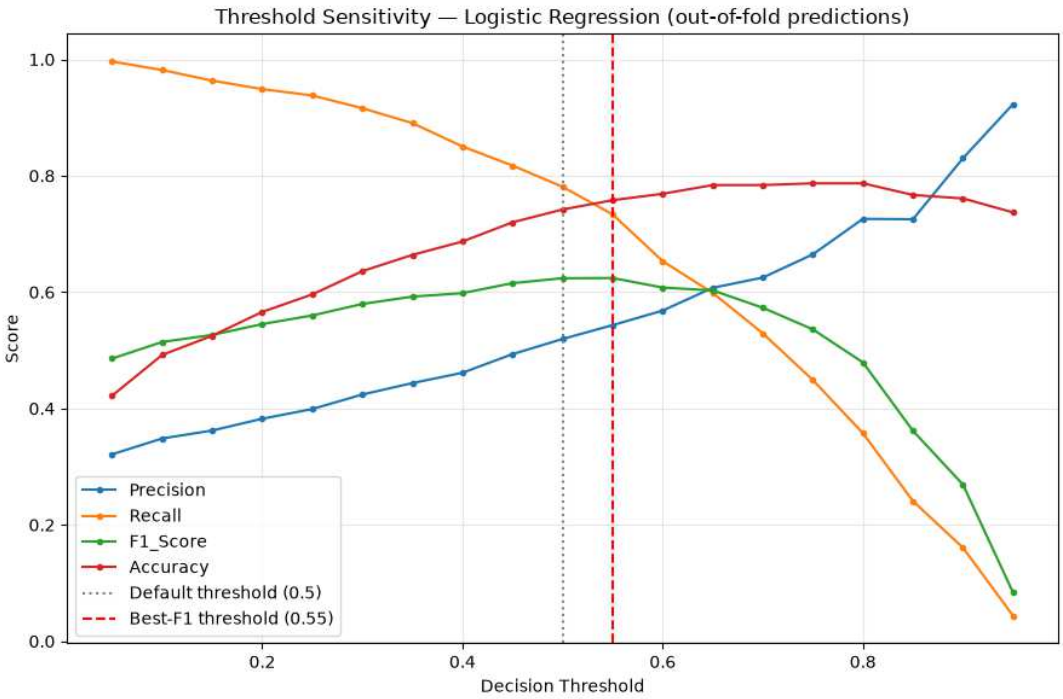


Figure 9. Decision-threshold sensitivity for Logistic Regression using out-of-fold predictions.

4.8.2. Hyperparameter Sensitivity

The regularization strength C was varied across five orders of magnitude (0.01, 0.1, 1.0, 10.0, 100.0) under the same repeated stratified cross-validation design (10-fold × 3 repeats). As shown in Figure 10, both AUC-ROC (≈ 0.830 ± 0.043) and PR-AUC (≈ 0.658–0.660 ± 0.071) remain essentially flat across the entire range of C, with mean values shifting by less than 0.002 between the most restrictive (C = 0.01) and least restrictive (C = 100) settings and fold-to-fold standard deviation dwarfing any between-setting difference. This near-total insensitivity to C indicates that, on this ten-feature, VIF-screened dataset, Logistic Regression's performance is governed primarily by the underlying feature set and class-imbalance weighting rather than by the strength of L2 regularization, and that the default C = 1.0 used throughout this study is not a meaningfully sub-optimal choice.

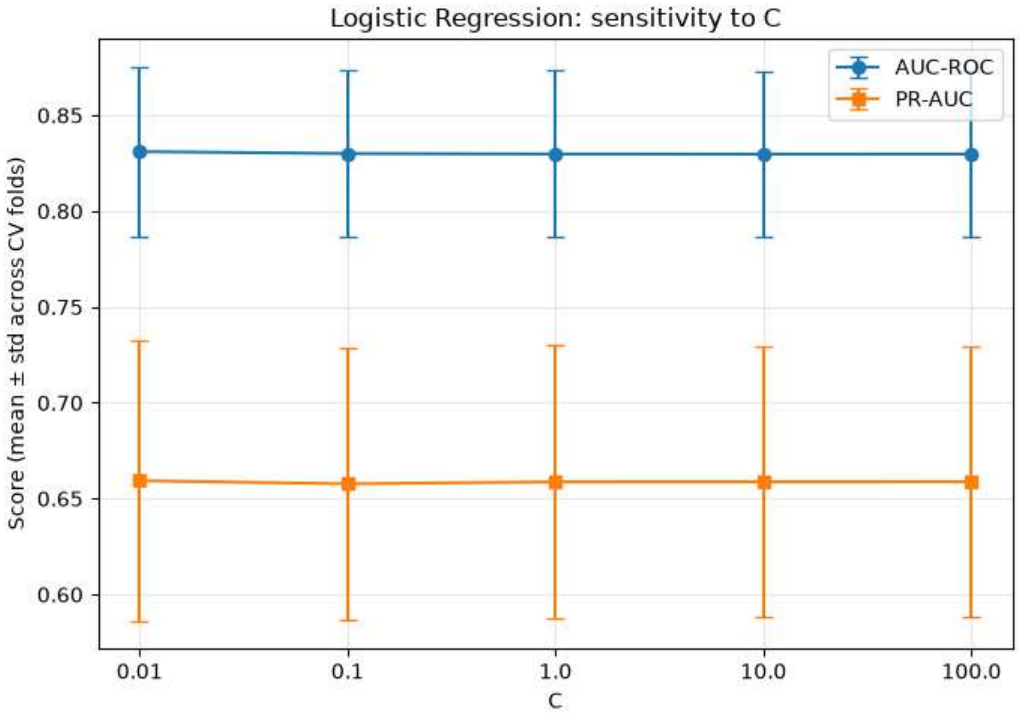


Figure 10. Hyperparameter sensitivity of Logistic Regression to the regularization strength C.

4.8.3. Robustness to Input Noise

Zero-mean Gaussian noise, scaled to a fraction of each feature's own standard deviation, was injected into the test-fold features only (the model was trained on clean data) at five noise levels: 0, 0.10, 0.25, 0.50, and 1.00 × feature SD. Performance degrades gracefully: at 10% and 25% noise, AUC-ROC and PR-AUC are statistically indistinguishable from the noise-free baseline (AUC-ROC 0.8294 and 0.8258 vs. 0.8298; PR-AUC 0.6557 and 0.6512 vs. 0.6588). Figure 11 depicts that degradation becomes material only at 50% noise (AUC-ROC = 0.8097, −0.024 vs. baseline; PR-AUC = 0.6238, −0.035) and pronounced at 100% noise (AUC-ROC = 0.7660, −0.064; PR-AUC = 0.5549, −0.104), at which point the injected noise is comparable in magnitude to the natural spread of each predictor. This pattern suggests the model is reasonably robust to the moderate measurement error typical of manually logged production data, and only becomes materially unreliable under substantially larger, likely non-representative levels of input corruption.

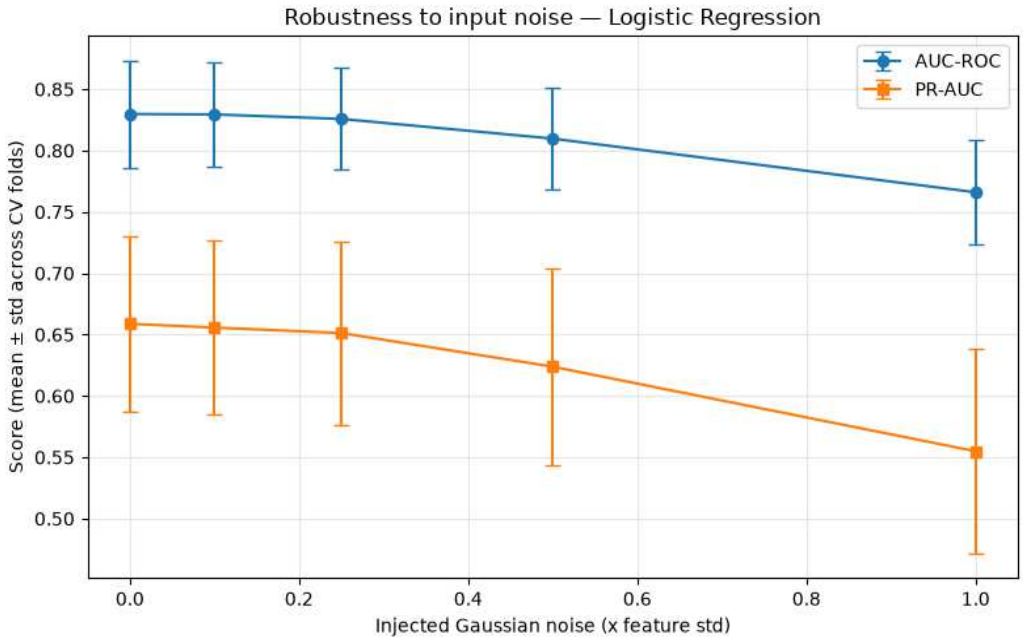


Figure 11. Robustness of Logistic Regression to Gaussian noise injected into test-fold features at prediction time.

4.9. Ablation Analysis of the Best-Performing Model

Feature importance for the ablation analysis was drawn from the Lasso coefficient ranking computed on the full dataset in Section 4.4 (model-agnostic, independent of the specific ML model). Three complementary ablations were performed on Logistic Regression under the same 10-fold × 3-repeat cross-validation design: (a) leave-one-feature-out, (b) cumulative feature addition, and (c) pipeline-component ablation.

4.9.1. Leave-One-Feature-Out (LOFO)

Starting from the full-feature baseline (AUC-ROC = 0.8298 ± 0.0435, PR-AUC = 0.6588 ± 0.0711), each of the eight top-ranked Lasso features was individually removed and the model re-evaluated. Removing Daily Production Quantity produced by far the largest degradation (ΔAUC-ROC = −0.0293, ΔPR-AUC = −0.0432), confirming its status as the single most load-bearing predictor identified throughout Sections 4.3–4.5. Number of Machine Failures (ΔPR-AUC = −0.0265) and Overtime Hours (ΔAUC-ROC = −0.0235) were the next most consequential, followed by Planned Production Target and Number of Operators. Removing Machine Breakdown Time, Defect Rate, or WIP Level produced negligible or slightly positive changes (ΔAUC-ROC between −0.0001 and +0.0006), indicating these three variables contribute little unique, non-redundant signal once the higher-ranked predictors are already in the model, consistent with the moderate correlation between Machine Breakdown Time and Machine Failures noted in Figure 4. Table 7 and Figure 12 show the results of the ablation test.

Table 7. Leave-one-feature-out ablation for Logistic Regression (10-fold × 3-repeat CV)

Removed Feature	AUC-ROC	PR-AUC	ΔAUC-ROC	ΔPR-AUC
(none – full model)	0.8298	0.6588	0.0000	0.0000
Daily Production Quantity	0.8005	0.6155	−0.0293	−0.0432
Number of Machine Failures	0.8180	0.6322	−0.0118	−0.0265
Overtime Hours	0.8063	0.6365	−0.0235	−0.0222
Planned Production Target	0.8194	0.6356	−0.0104	−0.0232
Number of Operators	0.8247	0.6492	−0.0051	−0.0095
Machine Breakdown Time	0.8300	0.6566	+0.0002	−0.0022
Defect Rate	0.8300	0.6567	+0.0002	−0.0021
WIP Level	0.8304	0.6592	+0.0006	+0.0005

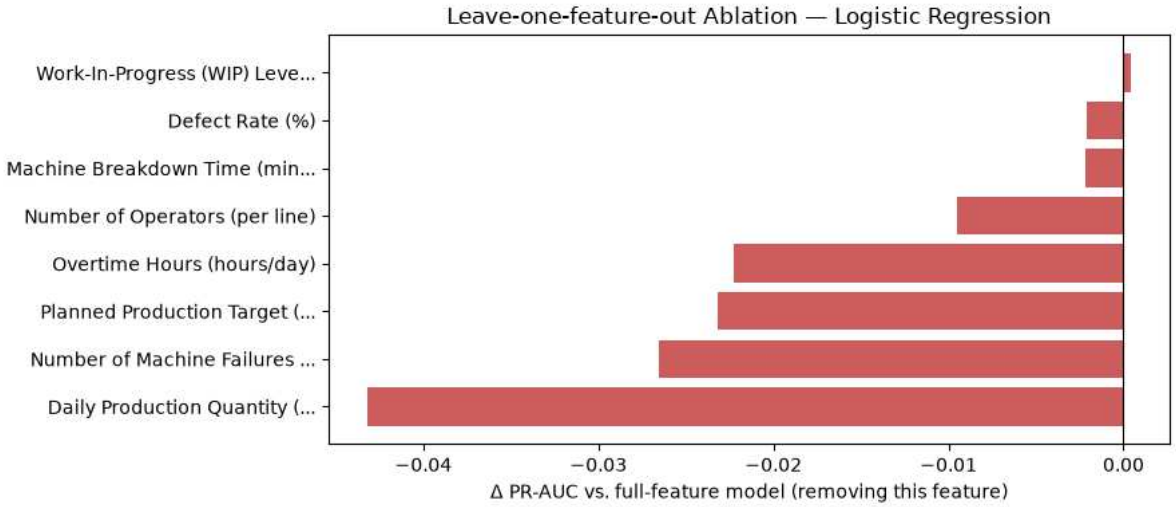


Figure 12. Leave-one-feature-out ablation for Logistic Regression, showing the change in PR-AUC relative to the full-feature model when each top-ranked feature is individually removed.

4.9.2. Cumulative Feature Addition

Features were added one at a time in descending order of Lasso importance, and the model was re-evaluated at each step. Table 8 and Figure 13 show the cumulative feature-addition ablation for Logistic Regression. Performance rises sharply over the first three features (AUC-ROC: 0.708 → 0.767 → 0.819; PR-AUC: 0.479 → 0.569 → 0.627) and continues to improve, at a decelerating rate, through the fifth and sixth features (AUC-ROC peaking at 0.8357, PR-AUC at 0.6660 with 6–7 features). Beyond six features, performance plateaus and even dips fractionally with the seventh and eighth features added, indicating that the marginal predictive value of the lowest-ranked features (Defect Rate, WIP Level) is effectively zero and that a five-to-six-feature model captures essentially all of the achievable discrimination and ranking performance on this dataset. This saturation pattern reinforces the Section 4.5

conclusion that backorder risk is driven by a moderate-sized, jointly-acting cluster of production-throughput, overtime, and machine-reliability variables rather than requiring the full twelve-feature set.

Table 8. Cumulative feature-addition ablation for Logistic Regression

N Features	AUC-ROC	PR-AUC
1 (Daily Production Quantity)	0.7076	0.4793
2 (+ Machine Failures)	0.7666	0.5692
3 (+ Overtime Hours)	0.8192	0.6266
4 (+ Planned Production Target)	0.8283	0.6561
5 (+ Number of Operators)	0.8340	0.6620
6 (+ Machine Breakdown Time)	0.8357	0.6639
7 (+ Defect Rate)	0.8352	0.6660
8 (+ WIP Level)	0.8350	0.6655

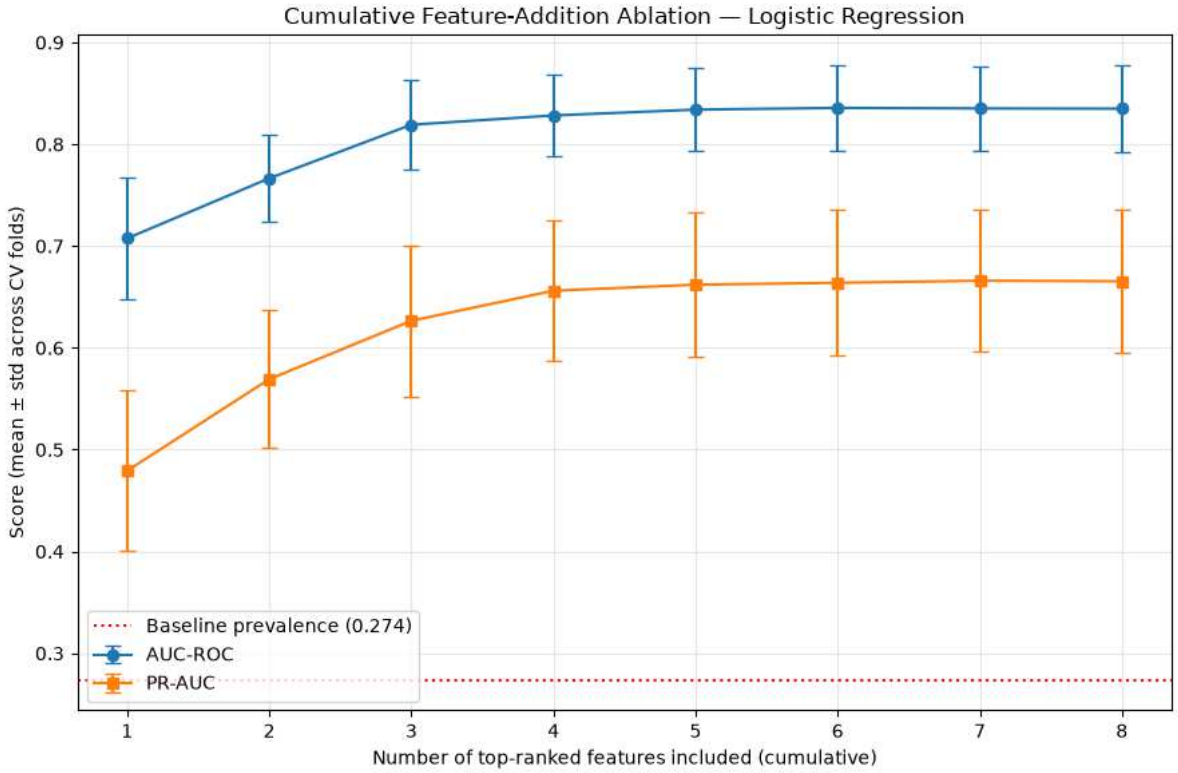


Figure 13. Cumulative feature-addition ablation for Logistic Regression.

4.9.3. Pipeline-Component Ablation

The full preprocessing-and-modeling pipeline (median imputation, standard scaling, and class_weight = "balanced") was compared against two ablated variants: one without class-imbalance weighting and one without feature scaling. Removing class-imbalance weighting left AUC-ROC and PR-AUC essentially unchanged (0.8294 vs.

0.8298; 0.6558 vs. 0.6588) but produced a substantial drop in F1 (0.5456 vs. 0.6218, a decline of nearly 8 points), driven by the classifier defaulting toward the majority class and sacrificing Recall for the minority (backorder) class. This confirms that class-imbalance weighting is not necessary for ranking performance (AUC-ROC, PR-AUC) but is operationally important for threshold-based classification decisions at the default 0.50 cutoff. Removing feature scaling had no measurable effect on any metric (AUC-ROC = 0.8298, PR-AUC = 0.6588, F1 = 0.6218, identical to the full pipeline to four decimal places), which is consistent with scikit-learn's lbfgs solver converging to the same optimum regardless of feature scale on this well-conditioned (low-VIF), moderately-sized feature set; scaling therefore contributes to numerical stability and faster convergence rather than to predictive performance per se in this setting. Table 9 shows pipeline-component ablation for Logistic Regression.

Table 9. Pipeline-component ablation for Logistic Regression

Variant	AUC-ROC	PR-AUC	F1
Full pipeline (impute + scale, balanced class weight)	0.8298 ± 0.0435	0.6588 ± 0.0711	0.6218 ± 0.0551
Without class-imbalance weighting	0.8294 ± 0.0434	0.6558 ± 0.0720	0.5456 ± 0.0771
Without feature scaling	0.8298 ± 0.0435	0.6588 ± 0.0711	0.6218 ± 0.0551

4.10. Model Interpretability: SHAP and Bayesian Network Structure

Consistent with standard practice, interpretability artefacts (Lasso sparsity, SHAP values, Bayesian Network CPDs) were computed from final models fit on the full n = 1,000 dataset after cross-validation had already established generalization performance (Sections 4.7–4.9); this full-data Lasso fit ($\alpha = 0.00933$) retained 10 of 12 features, closely reproducing the sparsity pattern reported in Section 4.4. SHAP analysis of the two tree-based models converged on a consistent global importance ranking (Table 10, Figure 14-15), led in both cases by Overtime Hours, Daily Production Quantity, Machine Failures, and Machine Breakdown Time, the same features identified as significant univariate discriminators in Section 4.3 and as the most consequential predictors in the leave-one-feature-out ablation in Section 4.8.1, followed by Planned Production Target. The convergence between the SHAP rankings of XGBoost and Random Forest, the independent hypothesis-testing and Lasso-selection results (Sections 4.3–4.4), and the cross-validated ablation results (Section 4.8) provides four-way, cross-method validation that these operational constructs constitute the primary drivers of backorder risk in the sampled lines.

Table 10. Top-5 SHAP feature importances (mean |SHAP value|) for Random Forest and XGBoost, full-dataset fit

Rank	XGBoost - Feature	Mean |SHAP|	Random Forest - Feature	Mean |SHAP|
1	Overtime Hours	0.1185	Overtime Hours	1.1376
2	Daily Production Quantity	0.0793	Daily Production Quantity	0.4796
3	Machine Breakdown Time	0.0606	Machine Failures	0.3807
4	Machine Failures	0.0540	Machine Breakdown Time	0.3573
5	Planned Production Target	0.0433	Planned Production Target	0.3290

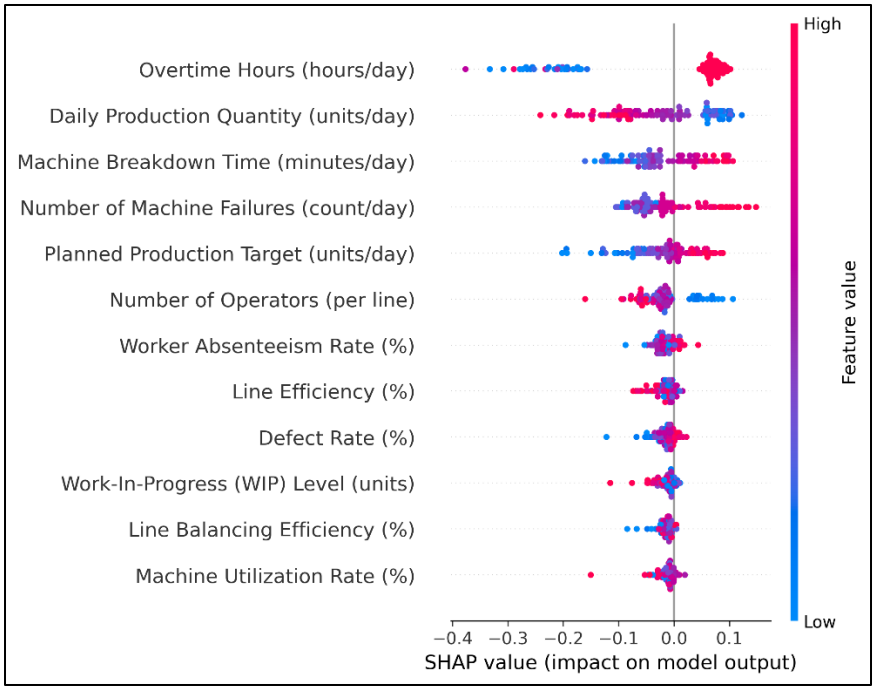


Figure 14. SHAP feature importance (mean |SHAP value|) for XGBoost, full-dataset fit.

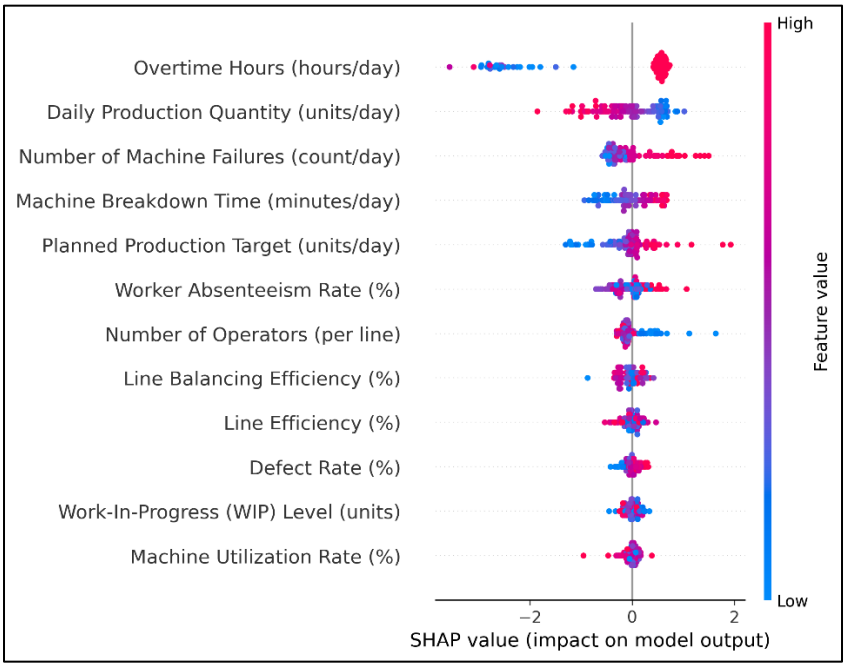


Figure 15. SHAP feature importance (mean |SHAP value|) for Random Forest, full-dataset fit.

The Discrete Bayesian Network offers a complementary, structurally transparent form of interpretability. Under the revised cross-validation pipeline, the final full-data BN fit again converged on a sparse DAG using only 4 of the available features, with an average Conditional Probability Table (CPT) size of 6.50 states, essentially identical

to the 6.5-state average reported previously, and small enough to be directly inspected by domain experts. The learned structure places Daily Production Quantity, Number of Operators, and Machine Breakdown Time as intermediate nodes conditioning the Backorder_Flag node, with Machine Failures entering as a root-level driver of Machine Breakdown Time. For example, the estimated conditional probability of Backorder_Flag = 1 varied from approximately 0.41 to 0.59 across the discretized states of Machine Breakdown Time, giving domain experts a directly interpretable, human-readable probabilistic chain (Machine Failures → Machine Breakdown Time → Backorder_Flag) that mirrors the causal narrative implied by the univariate, SHAP, and ablation results, at the cost of the lower discrimination and PR-AUC reported under cross-validation in Section 4.6.

5. Discussion

This section interprets the key findings of the study in relation to the research objectives and existing literature. The results are discussed from methodological, statistical, and practical perspectives, with particular attention to model performance, feature importance, and the implications of the findings for real-world manufacturing applications. The discussion also highlights the consistency of the results across different validation approaches and identifies the key limitations and directions for future research.

5.1. Reinterpreting Predictive Performance under a Robust Evaluation Protocol

The findings of this study are that repeated cross-validation changes the model ranking obtained from the original single-split analysis. Under repeated cross-validation, XGBoost's mean AUC-ROC (0.8150) is the lowest of the three ML models, and more importantly for a 27.4%-prevalence target, its mean PR-AUC (0.6223) trails both Logistic Regression (0.6609) and Random Forest (0.6518); this ranking persists even after XGBoost's imbalance handling (scale_pos_weight) was matched to the balanced class-weighting used by the other two ML models. Averaging over 50 independently resampled folds shrinks this sampling noise and reveals that Logistic Regression's apparent underperformance in the original single-split analysis (AUC-ROC = 0.777, the lowest of the four models) was itself partly an artefact of that particular 800/200 partition rather than a stable property of the model.

The practical implication is that model-selection decisions for this class of moderately-sized (n = 1,000), moderately imbalanced (27.4% positive) manufacturing datasets should not be made based on a single train–test split, however large that split appears; the 95% confidence intervals reported in Table 6 (e.g., LR PR-AUC 95% CI = [0.640, 0.682]; XGBoost PR-AUC 95% CI = [0.600, 0.645]) show that while the three ML models' AUC-ROC intervals overlap almost completely, their PR-AUC intervals are more clearly separated, which is precisely the metric this study argues is the operationally correct one given the class imbalance. This reinforces the broader methodological point that PR-AUC, not AUC-ROC, should anchor model selection whenever the positive class represents a minority of cases and false negatives (missed backorders) and false positives (unnecessary interventions) carry asymmetric operational costs.

AUC-ROC treats the true-negative rate as informative, but with only 27.4% of production-line-days experiencing a backorder, the majority class dominates the ROC curve's denominator and can mask meaningfully different levels of ranking quality among the minority (backorder) cases. PR-AUC, by contrast, is referenced directly

against the positive-class prevalence; a random or constant-score classifier achieves a PR-AUC equal to that prevalence (0.274 here), not the 0.5 that would be implied by AUC-ROC. Every model in this study clears that no-skill baseline by a substantial margin (lift of 0.237 to 0.387 PR-AUC points; Figure 8), confirming genuine predictive signal, but the ranking among models under PR-AUC (LR > RF > XGBoost > BN) differs from the ranking under AUC-ROC (RF ≈ LR > XGBoost > BN), and it is the PR-AUC ranking that most directly reflects each model's ability to correctly prioritize the relatively rare backorder-risk lines that a capacity-constrained supply-chain intervention would actually act upon. This distinction matters operationally: a monitoring dashboard built around Precision@K or expected-cost thresholds inherits whichever model's minority-class ranking is most reliable, and that model, under this more rigorous evaluation protocol, is Logistic Regression rather than XGBoost.

The earlier single-split analysis identified a discrimination-calibration trade-off centred on XGBoost: highest AUC-ROC, weakest calibration (ECE = 0.144). Under cross-validation, with matched imbalance-handling applied to all three ML models, the same qualitative trade-off persists but broadens: Logistic Regression has the best PR-AUC among ML models yet the weakest calibration (ECE = 0.1567), while XGBoost, now reweighted with scale_pos_weight rather than left unbalanced, shows a comparably degraded calibration (ECE = 0.1288) relative to Random Forest (ECE = 0.1080). This is a coherent and expected pattern: both Logistic Regression's balanced class-weighting and XGBoost's matched scale_pos_weight (Section 3.3.1) deliberately inflate predicted probabilities for the minority class to correct for its rarity, which improves ranking-based metrics (PR-AUC, Recall) at the direct expense of the numerical accuracy of the probabilities themselves; that this cost appears for both reweighted models, and not for Random Forest, indicates the trade-off is a property of the imbalance-correction mechanism itself rather than of any one algorithm. Practically, this means the recommended deployment model (Logistic Regression, by PR-AUC) should not have its raw predicted probabilities used directly for probabilistic decision rules such as expected-cost inventory buffering; a post-hoc recalibration step (Platt scaling or isotonic regression), fit on out-of-fold predictions to avoid re-introducing optimism, is a low-cost, high-value addition before operational deployment. The Bayesian Network's persistently superior calibration (ECE = 0.0484), now confirmed under a proper cross-validated evaluation rather than a single split, further substantiates its usefulness as a complementary, well-calibrated but lower-discrimination companion to the primary ranking model, particularly in settings where domain experts require directly interpretable conditional probabilities (Section 4.9) rather than the highest possible Recall.

The threshold-sensitivity analysis (Section 4.7.1) shows that Logistic Regression's classification performance is comparatively flat across a wide operating region (F1 between roughly 0.60 and 0.62 for thresholds 0.40–0.65), which gives supply-chain practitioners real latitude to tune the decision threshold to the specific cost asymmetry of their intervention programme without materially sacrificing overall accuracy, for example, lowering the threshold below 0.55 to prioritize Recall (catching more true backorders at the cost of more false alarms) in a context where a missed backorder is far more costly than an unnecessary quality check. The hyperparameter-sensitivity analysis (Section 4.7.2) is, if anything, reassuring from a robustness standpoint: performance is essentially invariant to the regularization strength C across five orders of magnitude, meaning the reported results are not an artefact of a fortuitously tuned hyperparameter and that the model would be simple to maintain in production without an elaborate

re-tuning pipeline. The noise-robustness analysis (Section 4.7.3) provides a further, practically important reassurance: performance is stable under noise levels up to roughly 25% of each feature's own standard deviation, which likely exceeds the measurement error present in manually logged daily production data (e.g., rounding in reported Overtime Hours or Machine Breakdown Time), suggesting the model is unlikely to degrade materially due to routine data-entry noise, though it should not be expected to remain reliable if entire fields become systematically corrupted or miscoded.

The leave-one-feature-out and cumulative-feature-addition ablations (Sections 4.9.1–4.9.2) provide convergent, cross-validated confirmation of the feature-importance narrative established independently through univariate testing (Section 4.3), Lasso regularization (Section 4.4), the composite Justification Score (Section 4.5), and SHAP (Section 4.9): Daily Production Quantity, Overtime Hours, Number of Machine Failures, and Planned Production Target are the four features whose removal most degrades predictive performance, and whose sequential addition accounts for the great majority of the achievable PR-AUC. This four-method agreement, involving statistical testing, regularization, ablation, and post-hoc explanation, provides stronger validation than any single method alone because each approach has different assumptions and blind spots. Univariate testing ignores feature interactions, Lasso assumes linear-additive effects, and SHAP is model-specific. In contrast, ablation directly measures the model-agnostic effect of removing a feature from the fitted pipeline under the same cross-validated protocol used for the headline performance results. The plateau beyond five to six features (Section 4.8.2, Figure 13) further suggests that a parsimonious, five-feature operational monitoring model consisting of Daily Production Quantity, Machine Failures, Overtime Hours, Planned Production Target, and Number of Operators could recover the great majority of the achievable predictive value while reducing the data-collection burden on factory-floor reporting systems. This represents a practically important simplification for real-world deployment. The pipeline-component ablation (Section 4.8.3) adds a further operational insight not previously available: class-imbalance weighting is essential for usable classification behavior (its removal costs nearly 8 F1 points) even though it barely moves ranking metrics, meaning it must not be dropped from any production implementation despite appearing unimportant if only AUC-ROC or PR-AUC were monitored.

5.2. Managerial and Operational Implications

The cross-validated results strengthen the operational findings of the original study. Backorder risk is mainly associated with a compact group of throughput, overtime, and machine-reliability factors. Three actions follow. First, production planners should closely monitor gaps between Daily Production Quantity and Planned Production Target. Second, increases in Machine Failures and Machine Breakdown Time should trigger early maintenance attention. Third, the Logistic Regression threshold of 0.55 can serve as an initial operating point for a monitoring dashboard. However, its probabilities should be recalibrated before cost-based decisions, and a parsimonious five-feature version may be sufficient for practical deployment.

6. Conclusion

This study developed a robust approach for predicting backorder risk in Bangladeshi ready-made-garment production lines using daily operational data. Repeated stratified cross-validation, PR-AUC, sensitivity analysis, and

feature ablation provided a stronger basis for model selection than a single train-test split. All four models clearly outperformed the no-skill PR-AUC baseline, confirming reliable predictive signal. Logistic Regression achieved the highest mean PR-AUC (0.6609), outperforming Random Forest (0.6518) and XGBoost (0.6223), a ranking that remained stable after XGBoost's imbalance treatment was matched to the other ML models. This reverses the earlier single-split finding and supports Logistic Regression as the preferred model for minority-class ranking. Its performance was also stable across different regularization strengths, decision thresholds, and moderate input noise.
The ablation results identified five key operational features: Daily Production Quantity, Overtime Hours, Machine Failures, Planned Production Target, and Number of Operators. These features captured most of the predictive value, while class-imbalance weighting remained important for practical classification. Overall, the findings support a simple, robust, and operationally practical backorder early-warning model for the sampled garment-production context.

7. Limitations & Future Work

Several limitations should be considered when interpreting these findings. First, the study used 1,000 daily observations from a single production context. Although this supported repeated cross-validation, the sample remains modest for ensemble models, and some confidence intervals are still relatively wide. Future studies should validate the models using larger, multi-factory datasets across different regions and ownership structures. This would test the generalizability of both the identified key features and the Logistic Regression-based model selection. The Bayesian Network should also be evaluated with a larger and more comparable cross-validation design.
Second, the analysis treated each line-day as an independent observation. Possible temporal dependence between consecutive production days was not explicitly considered. Future work should use blocked or time-series cross-validation, such as rolling-origin validation, to better assess forecasting performance and reduce the risk of temporal leakage.

Finally, the study identified a recurring trade-off between discrimination, calibration, and interpretability. Future research could investigate hybrid or stacked models that combine the strong ranking performance of Logistic Regression or ensemble models with the superior calibration and interpretability of the Bayesian Network. Such approaches may provide a better balance between predictive accuracy and practical decision support.

Data Availability Statement

The data analysed in this study were obtained from operating garment factories in Bangladesh and are confidential. They are not publicly available because of commercial and privacy restrictions. The data are available from the corresponding author, Md. Mushiur Rahman (mushiur@ipe.ruet.ac.bd), on reasonable request and with the permission of the participating factories.

Code Availability

The code used for data preprocessing, feature validation, model development, and evaluation is available from the corresponding author on reasonable request.

Clinical Trial Number

Clinical trial number: not applicable.

Ethics, Consent to Participate, and Consent to Publish Declarations

Ethics, Consent to Participate, and Consent to Publish declarations: not applicable.

Author Contributions

Nadim Mahmud contributed to conceptualization, methodology, software, formal analysis, investigation, data curation, visualization, and wrote the original draft. **Md. Mushiur Rahman** contributed to conceptualization, methodology, software, data curation, and supervision, wrote the original draft, and reviewed and edited the manuscript. **Md. Mahabub-A-Khoda** contributed to conceptualization, methodology, software, investigation, data curation, and visualization, and wrote the original draft. **Abdul Halim** contributed to formal analysis, investigation, and data curation, and wrote the original draft. All authors read and approved the final manuscript.

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgment

The authors thank the participating garment factories and their production management teams for providing the operational data used in this study and for their cooperation throughout the research.

Declaration of Generative AI and AI-Assisted Technologies in the Writing Process

During the preparation of this work, the authors used QuillBot and Grammarly to improve the language and check for grammatical errors. The authors reviewed and edited the output as needed and take full responsibility for the content of the published article.

Funding Statement

The authors declare that no funds, grants, or other support were received during the preparation of this manuscript, and this research received no specific grant from any funding agency in the public, commercial, or not-for-profit sectors.

Reference

[1] "READY-MADE GARMENT & TEXTILE SECTOR: THE CHANGE MAKER OF BANGLADESH ECONOMY." Accessed: Aug. 22, 2026. [Online]. Available: https://www.ficci.org.bd/ficci-stories/ready-made-garment-textile-sector-the-change-maker-of-bangladesh-economy

[2] "Bangladesh – Working towards a sustainable export future," World Trade Organization. Accessed: Aug. 22, 2026. [Online]. Available: https://www.wto.org/english/tratop_e/devel_e/bangladeshssldc_e.htm

[3] I. H. Sarker, "Machine Learning: Algorithms, Real-World Applications and Research Directions," *SN Computer Science 2021 2:3*, vol. 2, no. 3, pp. 160-, Mar. 2021, doi: 10.1007/S42979-021-00592-X.

[4] A. Ali, R. Jayaraman, E. Azar, and M. Maalouf, "Maximizing supply chain performance leveraging machine learning to anticipate customer backorders," *Comput. Ind. Eng.*, vol. 194, p. 110414, Aug. 2024, doi: 10.1016/J.CIE.2024.110414.

[5] C. Rudin, "Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead," *Nature Machine Intelligence 2019 1:5*, vol. 1, no. 5, pp. 206–215, May 2019, doi: 10.1038/s42256-019-0048-x.

[6] S. M. Lundberg and S.-I. Lee, "A unified approach to interpreting model predictions," in *NIPS'17: Proceedings of the 31st International Conference on Neural Information Processing Systems*, ACM Digital Library, Dec. 2017, pp. 4768–4777. Accessed: Aug. 22, 2026. [Online]. Available: https://proceedings.neurips.cc/paper_files/paper/2017/hash/8a20a8621978632d76c43dfd28b67767-Abstract.html

[7] D. Koller and N. Friedman, *Probabilistic Graphical Models: Principles and Techniques - Adaptive Computation and Machine Learning*, no. 3. The MIT Press, 2009.

[8] F. Bodendorf, P. Merkl, and J. Franke, "Intelligent cost estimation by machine learning in supply management: A structured literature review," *Comput. Ind. Eng.*, vol. 160, p. 107601, Oct. 2021, doi: 10.1016/J.CIE.2021.107601.

[9] M. Er Kara, S. Ü. Oktay Fırat, and A. Ghadge, "A data mining-based framework for supply chain risk management," *Comput. Ind. Eng.*, vol. 139, p. 105570, Jan. 2020, doi: 10.1016/J.CIE.2018.12.017.

[10] M. Yang, M. K. Lim, Y. Qu, D. Ni, and Z. Xiao, "Supply chain risk management with machine learning technology: A literature review and future research directions," *Comput. Ind. Eng.*, vol. 175, p. 108859, Jan. 2023, doi: 10.1016/J.CIE.2022.108859.

[11] L. Feng, W. Zhao, J. Wang, J. Feng, and Y. Guo, "Combining machine learning with a pharmaceutical technology roadmap to analyze technological innovation opportunities," *Comput. Ind. Eng.*, vol. 176, p. 108974, Feb. 2023, doi: 10.1016/J.CIE.2022.108974.

[12] H. Gao and Q. Ren, "Supply Chain Management and Backorder Products Prediction Utilizing Neural Network and Naive Bayes Machine Learning Techniques in Big Data Area: A Real-life Case Study," 2022, doi: 10.21203/rs.3.rs-2020401/v1.

[13] S. Islam and S. H. Amin, "Prediction of probable backorder scenarios in the supply chain using Distributed Random Forest and Gradient Boosting Machine learning techniques," *Journal of Big Data 2020 7:1*, vol. 7, no. 1, pp. 65-, Aug. 2020, doi: 10.1186/S40537-020-00345-2.

[14] M. Shajalal, P. Hajek, and M. Z. Abedin, "Product backorder prediction using deep neural network on imbalanced data," *Int. J. Prod. Res.*, vol. 61, no. 1, pp. 302–319, Jan. 2023, doi: 10.1080/00207543.2021.1901153.

[15] S. M. Ali, A. U. Rahman, G. Kabir, and S. K. Paul, "Artificial Intelligence Approach to Predict Supply Chain Performance: Implications for Sustainability," *Sustainability 2024, Vol. 16, Page 2373*, vol. 16, no. 6, p. 2373, Mar. 2024, doi: 10.3390/SU16062373.

[16] I. H. Sarker, "Data Science and Analytics: An Overview from Data-Driven Smart Computing, Decision-Making and Applications Perspective," *SN Comput. Sci.*, vol. 2, no. 5, Sep. 2021, doi: 10.1007/S42979-021-00765-8.

[17] E. B. Tirkolaee, S. Sadeghi, F. M. Mooseloo, H. R. Vandchali, and S. Aeini, "Application of Machine Learning in Supply Chain Management: A Comprehensive Overview of the Main Areas," *Math. Probl. Eng.*, vol. 2021, no. 1, p. 1476043, Jan. 2021, doi: 10.1155/2021/1476043.

[18] M. Z. Babai, M. Arampatzis, M. Hasni, F. Lolli, and A. Tsadiras, "On the use of machine learning in supply chain management: a systematic review," *IMA Journal of Management Mathematics*, vol. 36, no. 1, pp. 21–49, Dec. 2024, doi: 10.1093/IMAMAN/DPAE029.

[19] R. Carbonneau, K. Laframboise, and R. Vahidov, “Application of machine learning techniques for supply chain demand forecasting,” *Eur. J. Oper. Res.*, vol. 184, no. 3, pp. 1140–1154, Feb. 2008, doi: 10.1016/J.EJOR.2006.12.004.
[20] GuyonIsabelle and ElisseeffAndré, “An introduction to variable and feature selection,” *The Journal of Machine Learning Research*, Mar. 2003, doi: 10.5555/944919.944968.
[21] T. Mart, S. Duran, and I. S. Bakal, “Tactical inventory and backorder decisions for systems with predictable production yield,” *Int. J. Prod. Econ.*, vol. 143, no. 2, pp. 294–303, Jun. 2013, doi: 10.1016/J.IJPE.2012.01.029.
[22] Y. Xu, A. Bisi, and M. Dada, “A finite-horizon inventory system with partial backorders and inventory holdback,” *Operations Research Letters*, vol. 45, no. 4, pp. 315–322, Jul. 2017, doi: 10.1016/J.ORL.2017.04.007.
[23] S. Adana, S. Cevikparmak, H. Celik, and H. Uvet, “Predicting Backorders Using Machine Learning Techniques,” in *50th Annual Conference of the Decision Science Institute*, New Orleans: Decision Sciences Institute, Nov. 2019, pp. 1158–1178. Accessed: Aug. 22, 2026. [Online]. Available: https://decisionsciences.org/wp-content/uploads/2020/05/DSI-2019-Proceedings.pdf
[24] J. Bouganim and K. Olsson, “Using Deep Learning to Predict Back Orders: A Study in the Volvo Group Aftermarket Supply Chain,” Linköping University, Linköping, Sweden, 2019.
[25] R. W. Tatko, C. R. White, G. M. Williams, and R. K. Myers, *Backorder Prediction*. 2021.
[26] M. Shajalal, A. Boden, and G. Stevens, “Explainable product backorder prediction exploiting CNN: Introducing explainable models in businesses,” *Electronic Markets 2022 32:4*, vol. 32, no. 4, pp. 2107–2122, Nov. 2022, doi: 10.1007/S12525-022-00599-Z.
[27] F. Ahmed, M. Hasan, M. S. Hossain, and K. Andersson, “Comparative Performance of Tree Based Machine Learning Classifiers in Product Backorder Prediction,” *Lecture Notes in Networks and Systems*, vol. 569 LNNS, pp. 572–584, 2023, doi: 10.1007/978-3-031-19958-5_54/SAVE-RESEARCH.
[28] M. Qi *et al.*, “A Practical End-to-End Inventory Management Model with Deep Learning,” *https://doi.org/10.1287/mnsc.2022.4564*, vol. 69, no. 2, pp. 759–773, Dec. 2022, doi: 10.1287/MNSC.2022.4564.
[29] A. Rezasoltani, A. Jafarnejad, and A. M. Khani, “A voting-based hybrid machine learning model for predicting backorders in the supply chain,” *Journal of Decisions and Operations Research*, vol. 10, no. 1, pp. 194–213, Jun. 2025, doi: 10.22105/DMOR.2025.511401.1924.
[30] W. Bouaguel, M. A. Ben Hajkacem, C. E. Ben Ncir, and A. Qaffas, “Explainable Ensemble of Deep Learning Models for Improved and Interpretable Product Backorder Prediction,” *IEEE Access*, vol. 14, pp. 82975–82990, 2026, doi: 10.1109/ACCESS.2026.3697891.