Machine Learning-Based Late Delivery Risk Prediction for Supply Chain Resilience: An Intelligent Warehouse Fulfillment Decision Support Framework

Empirical machine-learning manuscript for submission preparation

Abstract—Late delivery is a persistent operational risk in digitally connected supply chains because it can weaken service reliability, complicate fulfillment coordination, and amplify downstream disruption. This study evaluates whether pre-fulfillment supply-chain information can provide useful predictive intelligence for identifying elevated late-delivery risk before final delivery outcomes are known. Using the DataCo Smart Supply Chain dataset containing 180,519 transactions and 53 original variables, a leakage-aware binary-classification pipeline was developed. Four variables that reveal or closely encode realized delivery outcomes—Days for shipping (real), Delivery Status, Order Status, and shipping date (DateOrders)—were excluded. Three temporal features were engineered, and an Order Id-grouped 80/20 split produced 144,650 training and 35,869 test observations. Logistic Regression, Random Forest, and XGBoost were compared with a majority-class baseline using accuracy, precision, recall, F1, and ROC-AUC. Random Forest achieved the strongest ROC-AUC (0.7675), 71.60% accuracy, and 85.10% precision, exceeding the baseline accuracy by 16.76 percentage points. Feature-importance analysis highlighted scheduled shipment duration, shipping-mode indicators, order hour, and geographic variables as influential predictive signals. The findings validate a prediction layer, not operational intervention effects. Accordingly, the paper proposes a Predictive Warehouse Resilience Decision Support Framework that positions model outputs as human-governed risk-prioritization inputs for future warehouse and fulfillment exception management.

Keywords—Supply Chain Resilience; Intelligent Warehouse Management; Machine Learning; Late Delivery Prediction; Predictive Analytics; Decision Support Systems; Random Forest; Digital Transformation; Supply Chain Analytics.

1. INTRODUCTION

Supply-chain resilience has moved from a specialist risk-management concern to a central operations capability because disruptions can propagate across tightly coupled suppliers, facilities, transportation networks, and customer commitments. Resilience is commonly framed as the capacity to withstand, adapt to, and recover from disruption while maintaining required performance [1], [2]. More recent work extends this logic toward viability and survivability in intertwined supply networks, emphasizing that disruption management requires both structural preparedness and timely decision-making [3]. The COVID-19 period further demonstrated that preparedness, digitalization, recovery, information sharing, and adaptive coordination are closely connected in disruption response [4], while empirical work across manufacturing and service settings highlighted the role of real-time information and advanced digital technologies in resilience strategies [5].

At the same time, logistics and warehousing are being reshaped by digital technologies. Logistics 4.0 research connects Internet of Things technologies, cyber-physical systems, big data, cloud platforms, and human-technology interaction with increasingly data-intensive logistics tasks [6]. Digital-supply-chain research similarly argues that digitalization changes how supply-chain processes are managed rather than merely digitizing isolated transactions [7]. Within warehouses, e-commerce pressure has accelerated the adoption of automation, dynamic order processing, and new fulfillment configurations [8]. Robotized and automated systems can increase flexibility and throughput, but they also introduce new planning and control requirements [9]. Even in conventional order-picking environments, warehouse performance depends on the interaction of storage, batching, zoning, routing, and other operational decisions rather than isolated optimization [10]. These developments make the warehouse a natural decision point for predictive intelligence: operational teams increasingly have data, but the managerial challenge is deciding which signals require attention before service failures occur.

The expansion of analytics capabilities creates the technical foundation for this shift. Big-data analytics can generate business value when it is embedded in organizational capabilities rather than treated as an isolated technical tool [11]. Similarly, artificial-intelligence capability depends on coordinated data, technology, skills, and organizational resources [12], while broader AI-business-value research shows that implementation outcomes depend on the mechanisms through which AI is integrated into operational processes [13]. These insights align with digital-transformation research, which describes transformation as an organizational process in which digital technologies alter value-creation paths, structures, capabilities, and routines [14], [15]. From this perspective, the transition from retrospective delivery reporting to predictive risk intelligence is not simply an algorithmic improvement; it represents a change in when managers become aware of operational risk and how exception-handling priorities may be established.

Late-delivery prediction offers a focused context for examining this transition. A recent systematic review of delivery-delay prediction found that machine learning is increasingly used for direct delay prediction but that much of the literature remains oriented toward marginal model-performance improvement rather than practical integration. The review specifically identified prediction timing, data combination, industrial integration, and decision-support recommendations as open priorities [16]. Recent empirical studies have also applied advanced learning models to supplier efficiency and late-delivery risk [17], while DataCo-based research has investigated deep-learning and explainability combinations for shipping-time and delivery-risk prediction [18]. Importantly, a 2026 DataCo study explicitly demonstrated that data leakage can inflate apparent model performance and advocated leakage-aware baselines before operational deployment [19]. These findings reinforce a central methodological principle of the present study: a model intended to support proactive decisions should be trained primarily on information plausibly available before the outcome it seeks to predict.

Prediction, however, does not automatically become a better decision. Explainable-AI research in supply-chain cyber resilience shows that transparency can support agile decision-making [20], while broader XAI literature identifies interpretability, accountability, and audience-appropriate explanation as key requirements for responsible deployment [21]. SHAP and related methods offer a formal approach for attributing predictions to input features [22]. Human-AI research likewise cautions that effective decision support requires deliberate socio-technical design rather than simple replacement of human judgment [23]. The automation-augmentation literature similarly argues that automation and human augmentation are interdependent rather than cleanly separable organizational choices [24]. For intelligent warehouse applications, this means that a late-delivery classifier should not be interpreted as an autonomous fulfillment controller. A more defensible role is that of a risk-prioritization mechanism that helps operational personnel identify orders that deserve review.

Against this background, the central research question is: How effectively can machine-learning models predict late-delivery risk using pre-fulfillment supply-chain information to support proactive intelligent warehouse fulfillment decisions and supply-chain resilience? Four supporting questions guide the study: (RQ1) How effectively can pre-fulfillment transactional, shipment, product, geographic, and order characteristics distinguish between on-time and late-delivery outcomes? (RQ2) How do Logistic Regression, Random Forest, and XGBoost compare across accuracy, precision, recall, F1, and ROC-AUC? (RQ3) Which pre-fulfillment variables contribute most strongly to the best model's classifications, and what do they imply for intelligent fulfillment decision support? (RQ4) How can predictive risk intelligence be incorporated into proactive exception management while preserving human oversight and avoiding unsupported claims of operational improvement?

The study makes five contributions. First, it provides a leakage-aware empirical workflow for predicting late-delivery risk from a restricted set of pre-fulfillment variables. Second, it compares three widely used supervised algorithms against a naive majority-class baseline rather than interpreting model scores without context. Third, it uses Order Id as the grouping variable for train-test separation, reducing the risk that related rows from the same order appear in both partitions. Fourth, it analyzes model feature importance to identify the operational attributes carrying the strongest predictive signal without treating those associations as causal effects. Fifth, it proposes a Predictive Warehouse Resilience Decision Support Framework (PWR-DSF) that separates the empirically evaluated prediction layer from the conceptually proposed human-governed decision and intervention layer.

2. THEORETICAL AND LITERATURE BACKGROUND

2.1 Supply Chain Resilience

Supply-chain resilience is a multidimensional capability rather than a single performance outcome. Ambulkar et al. [1] conceptualized firm resilience around disruption orientation, resource reconfiguration, and risk-management infrastructure, emphasizing the organizational capacity to respond to disruption. Hosseini et al. [2] reviewed quantitative resilience methods and organized them around capacities to absorb, adapt, and recover. Ivanov and Dolgui [3] broadened resilience toward supply-network viability, recognizing that firms operate within networks whose survival may require reconfiguration under extraordinary shocks. Across these perspectives, resilience involves temporal stages: exposure to disruption, recognition of impact, response, adaptation, and recovery.

The present study addresses only one narrow component of this broader construct: anticipation. A classifier that estimates delivery risk before the final delivery outcome may enhance visibility and earlier attention, but a predictive score is not itself resilience. Resilience would require the organization to interpret the signal, take an appropriate action, coordinate resources, and ultimately maintain or recover performance. This distinction is important because the empirical experiment measures classification performance, not resilience improvement.

2.2 Intelligent Warehouse Management

Warehousing has traditionally centered on receiving, storage, order picking, staging, and shipping. E-commerce and service-level pressure have made these activities increasingly time-sensitive [8]. Automated and robotic systems extend the physical execution capacity of warehouses, while integrated planning is required to coordinate system design, order batching, routing, storage, and resource assignments [9], [10]. Logistics 4.0 adds a digital layer in which operational objects, systems, and people exchange data through connected platforms [6].

An intelligent warehouse can therefore be understood not merely as an automated facility, but as an environment in which operational data are converted into timely decision support. In such an environment, machine learning may contribute to exception detection, workload prioritization, inventory-risk identification, labor planning, or shipment-readiness monitoring.

The DataCo dataset used in this study is not a granular warehouse-execution event log; it does not contain the full sequence of picker scans, dock events, labor assignments, or material-handling telemetry. Accordingly, this paper uses the more defensible term intelligent warehouse fulfillment decision support. The model identifies elevated delivery risk at the order/transaction level, while the proposed framework illustrates how such risk scores could later be integrated with warehouse-management and transportation-management data.

2.3 Machine Learning in Supply Chain Analytics

Supply-chain analytics combines data, analytical methods, and organizational decision processes. Wamba et al. [11] showed that analytics capabilities create value through dynamic organizational capabilities, while Mikalef and Gupta [12] argued that AI capability requires complementary technical, human, and organizational resources. Enholm et al. [13] further found that AI business value depends on adoption enablers, use mechanisms, and organizational effects. These studies help distinguish algorithmic performance from implementation capability: an accurate classifier is only one component of an operational AI system.

For delivery prediction specifically, Müller et al. [16] found a heterogeneous literature spanning classification and regression models, data drawn from enterprise and logistics systems, and varying levels of practical integration. Their review identified a need to move from isolated prediction models toward decision-support systems capable of converting forecasts into recommendations. Abushaega et al. [17] integrated supplier-efficiency evaluation with temporal convolutional networks for late-delivery prediction, while Ahmed et al. [18] combined deep-learning elements and SHAP-based interpretation using the DataCo dataset. These studies show the increasing technical sophistication of the domain. However, higher model complexity also heightens the need for transparent validation and careful assessment of whether predictors would actually exist at prediction time.

2.4 Late Delivery and Fulfillment Risk

Late delivery can arise from multiple interacting conditions, including scheduling assumptions, transport-mode choices, order timing, supplier behavior, capacity constraints, geography, and disruption. From a decision-support perspective, the most valuable prediction is not necessarily the one made closest to the final outcome; it is the one made early enough to permit a meaningful response while still achieving useful discrimination. Müller et al. [16] identify prediction timing as a significant research question because the availability of variables changes throughout the fulfillment lifecycle.

This issue creates a direct connection to leakage prevention. If a model uses realized shipping duration or a status field that encodes the final outcome, its apparent performance may be high while its operational usefulness is low. Katangoori [19] demonstrated this concern directly using DataCo by contrasting unconstrained and leakage-aware models. The present study therefore adopts a conservative feature-selection strategy and treats leakage prevention as part of model validity rather than a secondary data-cleaning step.

2.5 Predictive Analytics and Decision Support

Predictive analytics estimates future or unknown outcomes; decision support helps a human or organizational process choose what to do with those estimates. The distinction is central to this paper. A risk probability can rank orders, but it does not determine whether an order should be expedited, reprioritized, re-routed, reallocated, or simply monitored. Those actions depend on operational constraints, costs, customer commitments, resource availability, and policy.

Explainability provides one bridge between prediction and action. Barredo Arrieta et al. [21] emphasize that explanation must be designed for the audience and purpose. Lundberg and Lee [22] introduced SHAP as a unified additive feature-attribution framework for explaining predictions. Although the present experiment uses Random Forest impurity-based feature importance rather than SHAP, the explainability literature informs the interpretation: model importance identifies predictive contribution, not causal effect. Future deployment would benefit from instance-level explanations and calibrated risk scores so that operations personnel can understand why an order was flagged.

2.6 Digital Transformation and Human-AI Decision-Making

Digital transformation involves changes in organizational processes and value creation, not merely the adoption of new software [14], [15]. In fulfillment operations, retrospective dashboards answer what happened; predictive systems estimate what is likely to happen; decision-support systems identify which potential outcomes require attention. This temporal shift can alter roles, escalation paths, performance routines, and managerial judgment.

Storey et al. [23] describe human-AI decision systems as socio-technical designs in which technical capabilities and human behavior must be jointly considered. Raisch and Krakowski [24] similarly caution against treating automation and augmentation as mutually exclusive. These ideas support a human-governed interpretation of the present model: the algorithm can reduce the search space by flagging higher-risk orders, while personnel retain authority to interpret context and decide whether intervention is warranted.

3. CRITICAL LITERATURE SYNTHESIS

The literature converges around seven themes relevant to the present research. First, resilience research consistently emphasizes visibility, preparedness, adaptability, and recovery [1]–[5]. Yet resilience metrics typically operate at firm, network, or disruption-response levels, whereas many predictive models operate at transaction level. This creates a scale mismatch: order-level risk prediction may contribute to anticipatory capability, but it should not be equated with measured resilience.

Second, intelligent warehousing research shows that logistics performance increasingly depends on connected technologies and integrated operational decisions [6], [8]–[10]. Automation increases execution capability, but the associated control problem becomes more complex. Predictive analytics can complement automation by identifying exceptions, although the warehouse literature still contains a strong operations-research emphasis on routing, batching, storage, and robotics rather than predictive service-risk classification.

Third, AI and analytics research stresses that data and algorithms alone do not create business value [11]–[13]. Organizational resources, skills, and process integration determine whether insights become useful decisions. This supports evaluating a late-delivery model not only by ROC-AUC but also by its prospective fit with an operational exception-management process.

Fourth, direct delivery-delay prediction has become an identifiable research stream. Müller et al. [16] show that the field uses diverse algorithms but still faces gaps in practical applicability, prediction timing, and decision-support integration. Abushaega et al. [17] and Ahmed et al. [18] illustrate a move toward sophisticated architectures and interpretability. Katangoori [19], however, highlights a foundational validity problem: model performance can be inflated when post-outcome fields are included. The tension between complexity and validity is therefore more important than simply selecting the model with the highest published accuracy.

Fifth, explainability research suggests that trust and actionability depend on understanding model behavior [20]–[22]. Supply-chain decision-makers operate under cost, timing, and service constraints; a generic feature-importance chart may support global understanding, but local explanations are likely necessary for order-level action. This is particularly relevant when false positives can generate unnecessary expediting cost and false negatives can leave service failures undetected.

Sixth, human-AI research challenges the assumption that algorithmic recommendations automatically improve decisions [23], [24]. The design of interaction, authority, explanation, and escalation remains critical. For fulfillment, the sensible objective is not to eliminate human judgment but to focus it on a smaller set of potentially consequential exceptions.

Seventh, digital-transformation research frames predictive decision support as an organizational change in routines and capabilities [14], [15]. A warehouse or logistics team that moves from end-of-period lateness reporting to risk-based daily exception queues changes both the timing and structure of operational attention. The technology therefore has implications for governance, roles, data quality, and continuous learning as well as model accuracy.

TABLE IV. Critical literature synthesis and relevance to the present study.

Theme	Representative refs.	What literature establishes	Unresolved issue addressed here
Resilience and anticipation	[1]-[5]	Resilience requires absorption, adaptation, recovery, viability, information sharing, and response capabilities.	Prediction accuracy is not itself a resilience metric; the study treats prediction as an anticipatory input.
Digital logistics / warehousing	[6]-[10]	Logistics 4.0, automation, e-commerce fulfillment, and integrated warehouse planning increase data and control complexity.	How late-delivery risk scores could enter warehouse exception handling remains underdeveloped.
Analytics / AI capability	[11]-[15]	Analytics and AI create value when embedded in organizational capabilities and transformation processes.	The study separates technical prediction from organizational decision use.
Delivery-delay prediction	[16]-[19]	ML can predict delay risk, but practical integration, timing, data choice, and leakage control remain critical.	A restricted pre-fulfillment feature boundary and grouped split provide a conservative empirical baseline.
Explainability / decision support	[20]-[22]	Explainability can support transparency, trust, and interpretation of complex models.	Global feature importance is used cautiously; local explanation is reserved for future work.
Human-AI collaboration	[23]-[24]	Human-AI systems require socio-technical design; automation and augmentation are interdependent.	PWR-DSF keeps human review between risk score and operational action.

4. RESEARCH GAP

Four related gaps motivate the study. The first is a prediction-validity gap. Late-delivery datasets often contain variables recorded after shipment or variables strongly coupled with final delivery status. If those fields are used without a prediction-time audit, the model can learn the outcome rather than forecast risk. Recent DataCo research explicitly demonstrates this danger [19], yet leakage-aware design remains a practical concern across predictive logistics.

The second is a model-comparison gap. Performance values are difficult to interpret without a naive reference point. A classifier may appear useful because its accuracy exceeds 70%, but the relevant question is whether it materially outperforms the class distribution and whether it discriminates between classes. This study therefore includes a majority-class baseline and evaluates multiple metrics.

The third is a prediction-to-decision gap. Delivery-delay research increasingly identifies this issue directly: Müller et al. [16] call for integrated decision support capable of turning predictions into recommended actions. Yet a research article should not infer intervention effectiveness from classification performance. The present study addresses this gap conceptually by proposing a decision-support framework while explicitly separating tested prediction from untested operational action.

The fourth is a performance-to-resilience gap. Supply-chain resilience is a broad organizational capability, whereas late-delivery prediction is a narrow analytical task. The study therefore positions risk prediction as an anticipatory input that may contribute to resilience-related visibility and preparedness, not as a direct measurement or proof of resilience.

5. RESEARCH METHODOLOGY

5.1 Research Design and Dataset

The research uses a quantitative comparative machine-learning design. The DataCo Smart Supply Chain dataset was obtained from the publicly available dataset created by Constante, Silva, and Pereira [25]. The source contains 180,519 records and 53 original variables associated with provisioning, production, sales, and commercial distribution. The target variable, Late_delivery_risk, is binary: 1 denotes late delivery risk and 0 denotes the non-late/on-time class. The uploaded dataset contained 98,977 late observations and 81,542 on-time observations, yielding a late-delivery prevalence of approximately 54.83%.

5.2 Feature Selection and Leakage Prevention

The experiment intentionally did not use all available source variables. Twenty-two pre-fulfillment or order-context predictors were retained across fulfillment, transaction, customer, product, geographic, and temporal categories. Numeric predictors included Days for shipment (scheduled), Sales per customer, Order Item Discount, Order Item Discount Rate, Order Item Product Price, Order Item Profit Ratio, Order Item Quantity, Sales, Order Item Total, Product Price, Latitude, and Longitude. Categorical predictors included Type, Category Name, Customer Segment, Department Name, Market, Order Region, and Shipping Mode.

Three temporal variables were engineered from order date (DateOrders): order_month, order_dayofweek, and order_hour. The original dataset therefore remained a 53-variable source, while the analytical frame temporarily contained three additional engineered variables.

Four variables were deliberately excluded because they reveal or closely encode realized delivery outcomes: Days for shipping (real), Delivery Status, Order Status, and shipping date (DateOrders). The objective was to create a more credible prediction-time setting. This exclusion strategy is consistent with the broader leakage concern raised in recent DataCo research [19]. It does not prove that every retained predictor is perfectly available at every operational prediction point; rather, it removes the most obvious post-outcome fields and makes the prediction task materially more conservative.

TABLE I. Dataset variables and research roles.

Category	Variables / Attributes	Role	Type	Research relevance
Target	Late_delivery_risk	Dependent variable	Binary	Classifies late (1) versus non-late (0) delivery risk.
Fulfillment / shipping	Days for shipment (scheduled); Shipping Mode	Predictors	Numeric / categorical	Represents planned fulfillment timing and shipping-service conditions.
Order / transaction	Sales per customer; discounts; price; profit ratio; quantity; sales; item total	Predictors	Numeric	Captures transaction scale and commercial characteristics available before final delivery outcome.
Customer / business	Customer Segment; Type; Market	Predictors	Categorical	Represents business context and customer/market segmentation.

Category	Variables / Attributes	Role	Type	Research relevance
Product / organization	Category Name; Department Name	Predictors	Categorical	Captures product and organizational context.
Geography	Latitude; Longitude; Order Region	Predictors	Numeric / categorical	Adds spatial and regional predictive signal.
Engineered time	order_month; order_dayofweek; order_hour	Engineered predictors	Numeric	Extracts temporal structure from order date without using realized shipping date.
Leakage exclusions	Days for shipping (real); Delivery Status; Order Status; shipping date (DateOrders)	Excluded	Mixed	Avoids variables that reveal or closely encode realized delivery outcomes.

5.3 Train-Test Strategy

A grouped train-test split was created using GroupShuffleSplit with test_size = 0.20 and random_state = 42. Order Id served as the grouping variable. This design reduces the possibility that related transaction rows from the same order appear in both training and testing partitions. The resulting training set contained 144,650 rows and the test set 35,869 rows.

5.4 Preprocessing

Numeric variables were imputed using the median and standardized using StandardScaler configured for sparse-compatible transformation. Categorical variables were imputed using the most frequent category and transformed with one-hot encoding, with unknown test categories ignored. After preprocessing, the analytical matrix contained 115 transformed features.

5.5 Models

Four classifiers were evaluated. A DummyClassifier using the most-frequent strategy established the majority-class baseline. Logistic Regression represented a conventional linear classification approach. Random Forest represented bagged ensemble trees capable of learning nonlinear interactions. XGBoost represented gradient-boosted decision trees.

The Logistic Regression model used max_iter = 1000, solver = liblinear, and random_state = 42. Random Forest used 150 estimators, max_depth = 18, min_samples_leaf = 2, n_jobs = -1, and random_state = 42. XGBoost used 200 estimators, max_depth = 6, learning_rate = 0.08, subsample = 0.85, colsample_bytree = 0.85, tree_method = hist, eval_metric = logloss, and random_state = 42. These are specified configurations, not the outcome of an exhaustive hyperparameter search.

5.6 Evaluation Metrics

Performance was evaluated using Accuracy, Precision, Recall, F1, and ROC-AUC. For a binary classification problem with true positives (TP), true negatives (TN), false positives (FP), and false negatives (FN):

$$Accuracy = (TP + TN) / (TP + TN + FP + FN) \quad (1)$$

$$Precision = TP / (TP + FP) \quad (2)$$

$$Recall = TP / (TP + FN) \quad (3)$$

$$F1 = 2 \times (Precision \times Recall) / (Precision + Recall) \quad (4)$$

Accuracy summarizes overall correctness but can be misleading when the class distribution is uneven. Precision measures the reliability of positive risk flags, whereas recall measures the share of eventual late deliveries identified. F1 combines precision and recall but must still be interpreted alongside the confusion matrix and ROC-AUC. ROC-AUC summarizes discrimination across classification thresholds and is especially useful for comparing ranking ability independent of one fixed threshold.

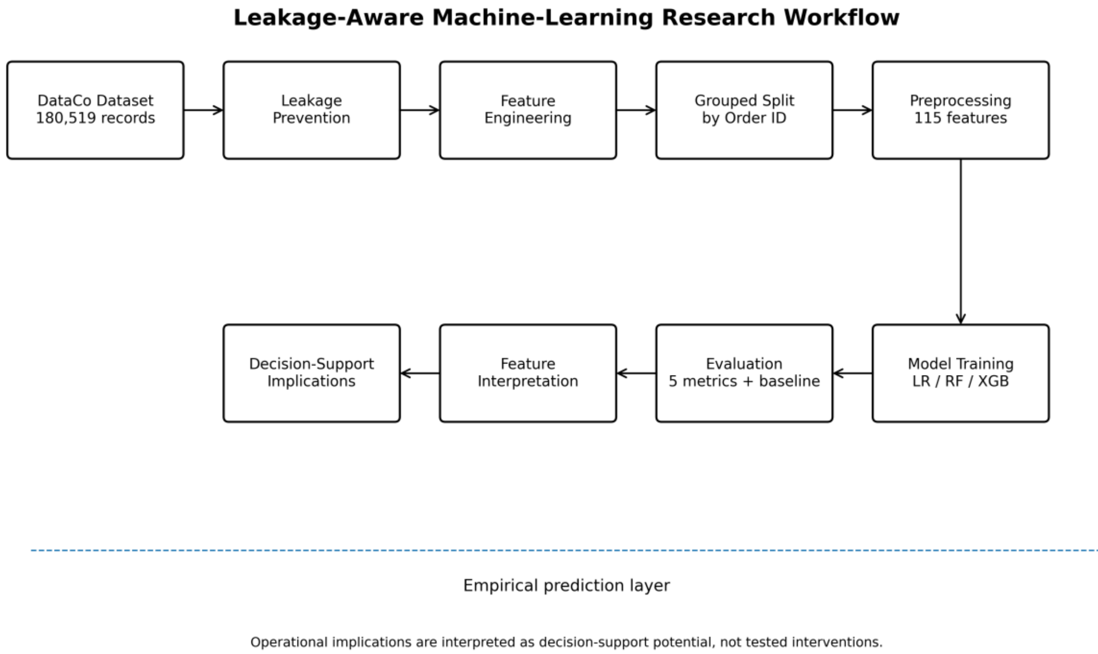

Fig. 1. Leakage-aware machine-learning research workflow used in the study.

6. EXPERIMENTAL RESULTS

6.1 Model Comparison

Table 2 reports the actual test-set results. Random Forest achieved the highest ROC-AUC at 0.7675 and the highest accuracy at 0.7160. XGBoost performed closely, with ROC-AUC 0.7652 and accuracy 0.7117. Logistic Regression achieved ROC-AUC 0.7414 and accuracy 0.7065. The majority baseline achieved accuracy 0.5484 and ROC-AUC 0.5000.

Random Forest therefore improved accuracy by approximately 16.76 percentage points over the majority baseline. The improvement is reported in percentage points because it is the absolute difference between 71.60% and 54.84%, not a relative percentage improvement.

The baseline requires careful interpretation. Because the late class is the majority, the most-frequent classifier predicts every observation as late. This produces recall = 1.0000 and an F1 value of 0.7083, which superficially exceeds the F1 scores of the trained models. However, the baseline has no discrimination capability: it never predicts the on-time class, its ROC-AUC is 0.5000, and its accuracy merely reflects the class prevalence. This result illustrates why a single metric should not drive model selection.

TABLE II. Model performance on the grouped holdout test set.

Model	Accuracy	Precision	Recall	F1	ROC-AUC
Random Forest	0.7160	0.8510	0.5845	0.6930	0.7675
XGBoost	0.7117	0.8336	0.5925	0.6926	0.7652
Logistic Regression	0.7065	0.8454	0.5689	0.6801	0.7414
Majority Baseline	0.5484	0.5484	1.0000	0.7083	0.5000

6.2 ROC-AUC Analysis

The ROC curves show Random Forest and XGBoost tracking closely across false-positive rates, with Random Forest holding a small advantage in AUC. Logistic Regression remains meaningfully above the random line but provides weaker discrimination. The narrow difference between Random Forest and XGBoost means the experiment should not be interpreted as proving universal Random Forest superiority. Instead, under the present feature set, split design, preprocessing, and model configuration, Random Forest produced the strongest overall discrimination.

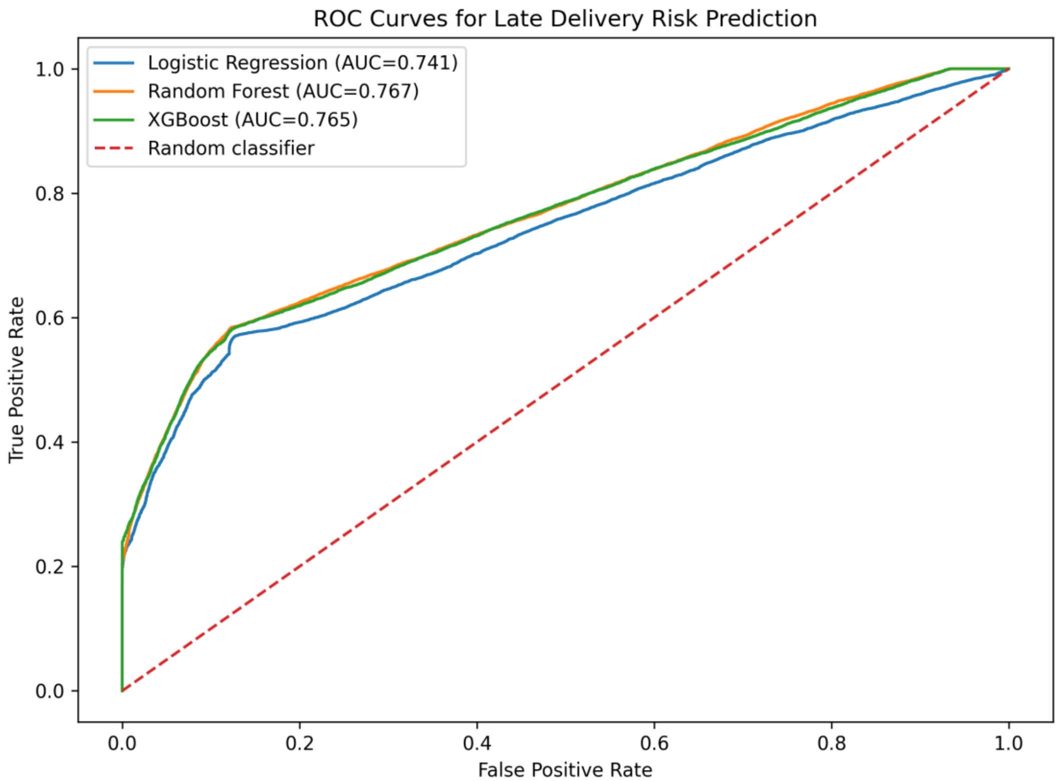

Fig. 2. ROC curves for Logistic Regression, Random Forest, and XGBoost on the grouped holdout test set.

6.3 Confusion Matrix and Precision-Recall Trade-off

The Random Forest confusion matrix contained 14,187 true negatives, 2,013 false positives, 8,173 false negatives, and 11,496 true positives. Its precision of 0.8510 means that when the model flagged an observation as late, approximately 85% of those flags were correct in the test sample. This property is attractive for exception-management use because a high-precision queue reduces the proportion of alerts that prove unnecessary.

The recall of 0.5845 creates the main limitation. More than 8,000 late observations in the test partition were not flagged by the default decision threshold. In operational terms, the model is therefore better characterized as a prioritization aid than as a comprehensive detector of all late-delivery risk. An organization that places greater cost on missed delays could lower the probability threshold or use cost-sensitive learning, but that would likely increase false positives. Threshold optimization should therefore be tied to the economics of intervention rather than performed solely for a better summary score.

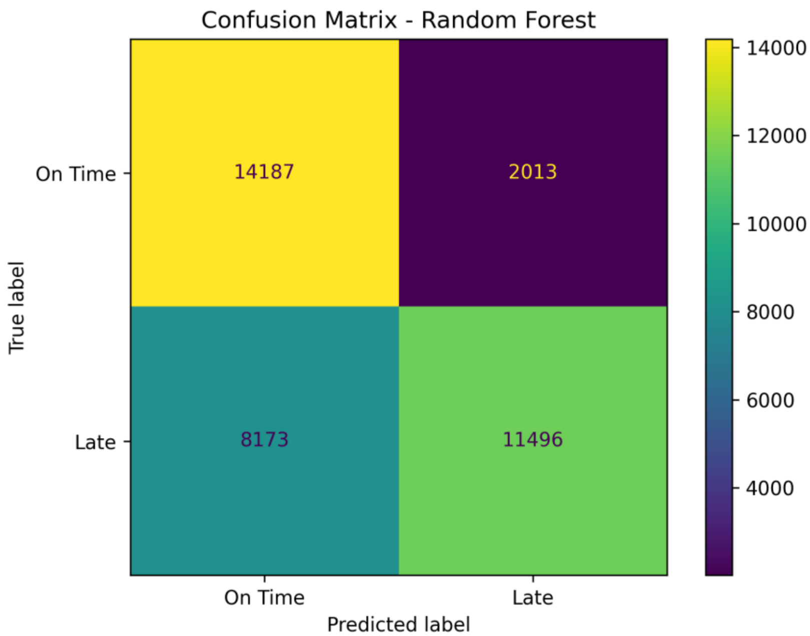

Fig. 3. Random Forest confusion matrix on the grouped holdout test set.

6.4 Feature Importance

Random Forest feature importance identified Days for shipment (scheduled) as the leading single transformed predictor with importance 0.2103. Shipping Mode levels followed: Standard Class (0.1573), First Class (0.1377), Second Class (0.0641), and Same Day (0.0393). Order hour contributed 0.0716. Latitude (0.0349) and Longitude (0.0332) also appeared among the leading predictors, followed by Order Item Profit Ratio (0.0216) and Sales per customer (0.0190).

These values describe how the trained Random Forest used features to reduce impurity across trees. They do not establish causal relationships. For example, a large importance for a shipping-mode indicator does not prove that selecting that mode causes lateness; the mode may encode service-level structures, order characteristics, geography, or other correlated conditions. Similarly, geographic importance indicates predictive signal in the model rather than a causal geographic mechanism.

TABLE III. Leading Random Forest feature-importance values.

Rank	Predictor	Importance	Operational interpretation (non-causal)
1	Days for shipment (scheduled)	0.2103	Planned shipping-time allowance carries the strongest model-level predictive signal.
2	Shipping Mode - Standard Class	0.1573	Service-mode category contributes materially to risk separation.
3	Shipping Mode - First Class	0.1377	Service-mode category contributes materially to risk separation.
4	order_hour	0.0716	Order timing contributes temporal predictive signal.
5	Shipping Mode - Second Class	0.0641	Service-mode category contributes to classification.
6	Shipping Mode - Same Day	0.0393	Service-mode category contributes to classification.
7	Latitude	0.0349	Geographic position contributes spatial signal.
8	Longitude	0.0332	Geographic position contributes spatial signal.
9	Order Item Profit Ratio	0.0216	Transaction economics provide additional predictive information.
10	Sales per customer	0.0190	Customer-level transaction value provides additional predictive information.

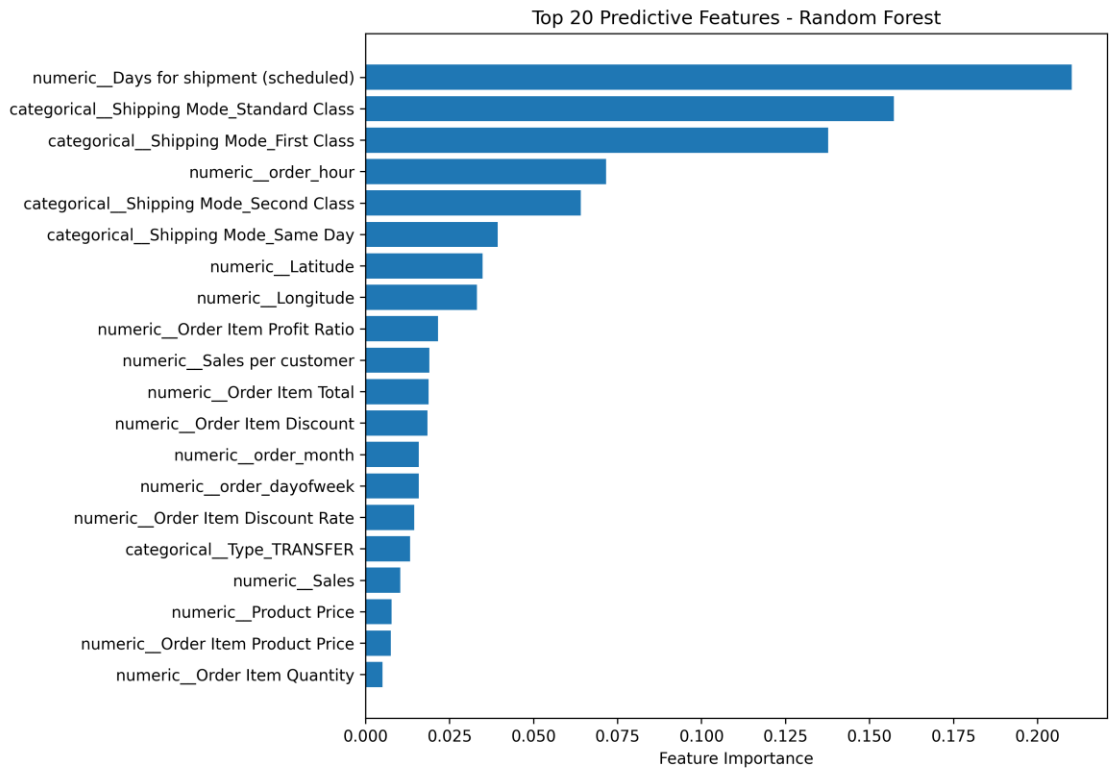

Fig. 4. Top Random Forest feature-importance values. Importance is predictive, not causal.

7. DISCUSSION

The empirical findings support three conclusions. First, pre-fulfillment supply-chain information contains meaningful signal for late-delivery classification even after obvious outcome variables are removed. A ROC-AUC of 0.7675 is materially above random discrimination and the 71.60% accuracy exceeds the majority baseline by 16.76 percentage points. This is a more cautious result than near-perfect scores sometimes reported on DataCo, but caution is appropriate because the model is intended to approximate prediction rather than post hoc recognition.

Second, model complexity alone did not determine performance. Random Forest and XGBoost were nearly tied in ROC-AUC, and Logistic Regression remained competitive in accuracy and precision. The result is consistent with a dataset in which scheduled service, shipping mode, time, geography, and transaction features carry both nonlinear and relatively stable structural relationships. The narrow gap between ensemble models also suggests that future work should prioritize data quality, prediction timing, calibration, and external validation rather than treating small algorithmic differences as the primary research objective. This aligns with Müller et al. [16], who argue that delivery-delay research needs stronger practical integration, not merely marginal predictive gains.

Third, the high precision and moderate recall define the most plausible operational use. A warehouse or fulfillment operation could use the model to create a high-risk exception queue. Because most positive flags are correct, supervisors could review those orders for readiness, staging, inventory availability, carrier handoff, customer commitment, or other contextual constraints. Yet because recall is only 58.45%, the queue cannot replace existing monitoring. It should augment broader operational controls.

Cross-study performance comparisons require caution. Ahmed et al. [18], Abushaega et al. [17], and Katangoori [19] report higher scores under different architectures, feature treatments, and validation choices. These studies are valuable evidence that DataCo supports delivery-risk research, but their results cannot be directly ranked against the current experiment without reproducing identical prediction-time assumptions and splits. The present study's contribution is therefore methodological transparency: the feature restrictions, grouped split, baseline, and exact model settings are reported explicitly.

The feature-importance pattern is also operationally coherent without being causal. Scheduled shipment duration and shipping mode represent service-design information that exists before final delivery. Order hour may capture operational timing or cut-off effects. Geographic coordinates can act as coarse proxies for network position. These signals are potentially useful for a risk-ranking system because they are interpretable at a managerial level. However, future work should test whether these variables remain important when richer WMS, TMS, carrier, inventory, labor, weather, and facility-capacity data are introduced.

8. PREDICTIVE WAREHOUSE RESILIENCE DECISION SUPPORT FRAMEWORK

To translate prediction into a defensible operations concept, this study proposes the Predictive Warehouse Resilience Decision Support Framework (PWR-DSF). The framework contains eight stages: (1) operational data capture, (2) data quality and leakage controls, (3) pre-fulfillment feature construction, (4) machine-learning late-delivery risk prediction, (5) risk prioritization and exception queueing, (6) human operational review, (7) fulfillment decision or potential intervention, and (8) outcome capture with continuous learning.

Stages 1–4 correspond to the empirically evaluated layer. The present experiment demonstrates that structured supply-chain data can be transformed into a model that discriminates late-delivery risk with ROC-AUC 0.7675. Stages 5–8 are proposed rather than experimentally tested. Their purpose is to show how a prediction could become a socio-technical decision-support process without overstating the empirical evidence.

In a future intelligent warehouse, the risk score could be displayed in an exception-management dashboard integrated with WMS or control-tower workflows. Orders above a threshold could be prioritized for review, with supporting context such as shipping mode, planned service duration, order age, inventory readiness, staging status, carrier cut-off, and customer priority. A supervisor could then decide whether to expedite picking, adjust staging sequence, escalate a material shortage, coordinate with transportation, or take no action. The framework therefore treats prediction as attention allocation.

This structure is consistent with human-AI decision-system research [23] and the automation-augmentation paradox [24]. The model performs a task at which machines are strong—screening a large number of transactions consistently—while human personnel retain contextual authority. Explainability becomes part of the handoff between prediction and judgment. Global feature importance is sufficient for the present research analysis, but future operational deployment should provide order-level explanations using methods such as SHAP [22] and should document confidence, calibration, and data freshness.

The resilience connection is anticipatory rather than outcome-based. Earlier identification of elevated delivery risk may increase visibility and preparedness, two capabilities discussed throughout the resilience literature [1]–[5]. Whether the framework actually reduces lateness or improves recovery requires a field experiment or quasi-experimental deployment. That validation is explicitly outside the scope of the current study.

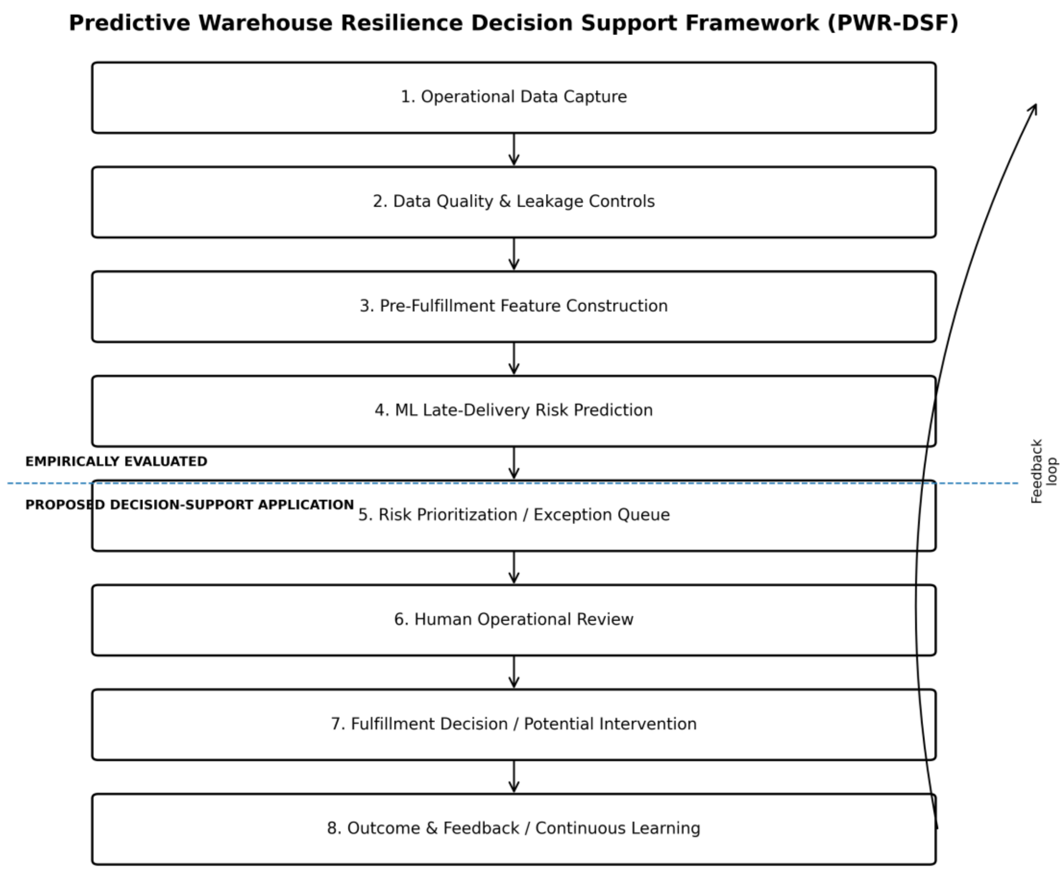

Fig. 5. Predictive Warehouse Resilience Decision Support Framework (PWR-DSF), separating the evaluated prediction layer from the proposed human-governed application layer.

9. DIGITAL TRANSFORMATION AND ORGANIZATIONAL IMPLICATIONS

The broader digital-transformation implication is a shift in the timing of information. Traditional operational reporting is retrospective: it identifies orders that were late and summarizes service-level performance after the fact. Predictive analytics is prospective: it estimates which orders are more likely to become late. Decision support adds a third layer: it determines which predictions deserve human attention and what contextual information should accompany them.

This progression can be represented as Data → Analytics → Prediction → Risk Intelligence → Human Review → Decision → Action → Feedback. The progression parallels the digital-transformation literature's emphasis on changes in organizational routines and value creation rather than isolated technology adoption [14], [15]. It also aligns with Logistics 4.0, where connected information and human interaction are integral to the logistics system [6].

For managers, implementation would require more than deploying a Python model. Data ownership must be defined; prediction timing must align with fulfillment milestones; alert thresholds must reflect intervention cost; false-positive and false-negative consequences must be measured; and users must understand both model capability and model limitation. Teams would need technical fluency sufficient to interpret risk scores and explanations without assuming that model outputs are certainties.

Organizationally, an exception-based operating model can redistribute attention. Instead of reviewing all open orders equally, supervisors may focus on the subset with elevated risk. This can create value only if the alert is timely, explainable, and linked to actions that personnel are authorized and resourced to take. In that sense, digital transformation is not the presence of machine learning; it is the redesign of the operational decision process around reliable digital intelligence.

10. MANAGERIAL IMPLICATIONS

The study has implications for warehouse managers, fulfillment planners, transportation teams, supply-chain analysts, and data-science teams. For warehouse and fulfillment leaders, the main insight is that risk classification can potentially become a triage mechanism. The Random Forest's 85.10% precision suggests that a positive-risk queue could be relatively focused, reducing some of the burden associated with indiscriminate alerting.

Transportation and planning teams could use the risk score as one input among carrier capacity, service level, route, shipment cut-off, and customer priority. Data-science teams should avoid optimizing only aggregate accuracy and instead define the operational cost of missed delays and unnecessary interventions. Analysts should also monitor drift because shipping policies, carrier performance, customer mix, and network configuration can change over time.

Managers should be especially cautious about interpreting feature importance as causality. Scheduled shipment duration and shipping mode are useful predictors in this experiment, but changing them may not improve delivery performance unless the underlying operational mechanism is understood. The model is best used to generate questions and priorities for review rather than to prescribe policy changes directly.

11. RESEARCH IMPLICATIONS

The research reinforces the need for prediction-time validity in supply-chain machine learning. Future studies should explicitly state which variables are available at the moment a prediction is generated. Leakage audits should become a normal part of logistics analytics research, particularly for datasets that contain both planned and realized timestamps or outcome-coded statuses.

A second opportunity is explainability. Global Random Forest importance identifies broad drivers but does not explain individual predictions. SHAP [22] or other local methods could reveal why a specific order was flagged. Future research should evaluate whether explanations improve the quality, speed, or consistency of warehouse decisions rather than assuming that explainability automatically creates trust.

A third opportunity is data integration. Müller et al. [16] identify richer combinations of enterprise, logistics, location, and environmental data as a priority. A stronger intelligent-warehouse model could integrate WMS events, TMS milestones, inventory shortages, dock schedules, labor availability, equipment state, weather, traffic, carrier reliability, and promised customer dates. Such integration would also enable more precise prediction timing.

Finally, the PWR-DSF suggests a bridge between machine learning and supply-chain resilience research. The prediction layer can be evaluated with conventional classification metrics, while the decision-support layer should be evaluated with operational and resilience outcomes such as intervention lead time, service-level adherence, recovery time, expediting cost, alert burden, and decision quality. Separating these layers can improve theoretical clarity.

12. LIMITATIONS

Several limitations constrain interpretation. First, the experiment uses one public dataset. DataCo is valuable for reproducible research but may not represent the process, data quality, network structure, or service commitments of a specific contemporary warehouse. Second, the study predicts late delivery; it does not measure supply-chain resilience as an organizational outcome. Third, no operational intervention was tested. Claims about order prioritization, staging, carrier coordination, or other warehouse actions are conceptual implications rather than experimental findings.

Fourth, Random Forest impurity-based feature importance can favor variables with particular structures and does not establish causality. Fifth, the model settings were specified but not subjected to exhaustive hyperparameter tuning. Sixth, no independent external dataset was used for validation. Seventh, the dataset is historical and structured; live operations may experience concept drift, data latency, and missing-event problems. Eighth, the leakage controls remove the most obvious outcome variables but do not constitute a formal data-lineage audit of every retained feature at every possible prediction timestamp.

Ninth, the DataCo dataset is not a complete WMS event log. It lacks detailed picker activity, workstation congestion, labor allocation, dock-door events, equipment telemetry, inventory-location changes, and many other variables relevant to intelligent warehouse management. The article therefore interprets the model as an intelligent fulfillment decision-support component rather than a complete warehouse-control model. Finally, the default classification threshold was used; threshold optimization based on operational cost was not performed.

13. FUTURE RESEARCH

Future work should validate the pipeline on external and industry-specific datasets and compare performance across warehouses, regions, product categories, and service models. Time-aware validation should be added so that models train on earlier periods and test on later periods, better approximating deployment. Cross-validation grouped by Order Id and facility should be considered where identifiers permit.

Model development should explore calibration, cost-sensitive learning, threshold optimization, and hyperparameter tuning. Because the current model has high precision but moderate recall, a particularly useful research question is how to increase recall without creating an operationally unmanageable false-positive rate. The answer should be evaluated in economic terms, not only statistical metrics.

Explainability should move from global feature importance to local explanation and human-subject evaluation. SHAP-based explanations [22] could be paired with risk scores in prototype dashboards, followed by experiments measuring decision time, intervention quality, trust calibration, and overreliance. Human-AI research indicates that interaction design strongly affects outcomes [23], [24], making this an important extension.

A richer operational model should combine WMS, TMS, ERP, inventory, carrier, IoT, and external data. Digital-twin research could eventually simulate intervention policies, but simulation should occur only after the prediction model is validated on operational data. Field experiments could then compare a control process with a prediction-supported exception-management process to measure whether intervention reduces late deliveries, cycle time, expediting cost, or recovery time.

Finally, research should operationalize the resilience contribution directly. Instead of inferring resilience from prediction accuracy, future studies could define anticipation lead time, disruption-detection rate, time-to-response, time-to-recovery, and service-continuity metrics. This would enable empirical testing of whether predictive intelligence actually strengthens resilience capability.

14. CONCLUSION

This study examined whether pre-fulfillment supply-chain information can support meaningful prediction of late-delivery risk. Using 180,519 DataCo records, a leakage-aware workflow excluded obvious post-outcome fields, engineered temporal features, separated orders through a grouped train-test design, and compared Logistic Regression, Random Forest, and XGBoost with a majority-class baseline. Random Forest produced the strongest ROC-AUC (0.7675), 71.60% accuracy, and 85.10% precision. Its performance exceeded the naive baseline by 16.76 percentage points in accuracy, demonstrating useful discrimination without relying on the most obvious outcome-leakage variables.

The results should nevertheless be interpreted with restraint. Moderate recall means the model misses a substantial share of eventual late deliveries, and feature importance does not establish causality. More importantly, predictive performance does not prove that warehouse interventions improve delivery or resilience. The empirical contribution is the prediction layer.

The proposed PWR-DSF therefore places machine learning within a broader human-governed process: operational data are transformed into risk intelligence, high-risk transactions are prioritized, personnel review contextual evidence, and outcomes become feedback for future learning. The central implication is that the value of machine learning in supply-chain operations does not come simply from generating predictions. Its practical value depends on whether reliable predictions are timely, interpretable, and integrated into operational decisions that people can execute and evaluate. In this sense, predictive late-delivery intelligence offers a concrete pathway from reactive reporting toward intelligent warehouse fulfillment decision support and, potentially, a stronger anticipatory foundation for supply-chain resilience.

REFERENCES

[1] S. Ambulkar, J. Blackhurst, and S. Grawe, “Firm's resilience to supply chain disruptions: Scale development and empirical examination,” Journal of Operations Management, vols. 33–34, pp. 111–122, 2015, doi: 10.1016/j.jom.2014.11.002.

[2] S. Hosseini, D. Ivanov, and A. Dolgui, “Review of quantitative methods for supply chain resilience analysis,” Transportation Research Part E: Logistics and Transportation Review, vol. 125, pp. 285–307, 2019, doi: 10.1016/j.tre.2019.03.001.

[3] D. Ivanov and A. Dolgui, “Viability of intertwined supply networks: extending the supply chain resilience angles towards survivability. A position paper motivated by COVID-19 outbreak,” International Journal of Production Research, vol. 58, no. 10, pp. 2904–2915, 2020, doi: 10.1080/00207543.2020.1750727.

[4] M. M. Queiroz, D. Ivanov, A. Dolgui, and S. F. Wamba, “Impacts of epidemic outbreaks on supply chains: mapping a research agenda amid the COVID-19 pandemic through a structured literature review,” Annals of Operations Research, vol. 319, no. 1, pp. 1159–1196, 2022, doi: 10.1007/s10479-020-03685-7.

[5] A. Belhadi, S. Kamble, C. J. C. Jabbour, A. Gunasekaran, N. O. Ndubisi, and M. Venkatesh, “Manufacturing and service supply chain resilience to the COVID-19 outbreak: Lessons learned from the automobile and airline industries,” Technological Forecasting and Social Change, vol. 163, Art. no. 120447, 2021, doi: 10.1016/j.techfore.2020.120447.

[6] S. Winkelhaus and E. H. Grosse, “Logistics 4.0: a systematic review towards a new logistics system,” International Journal of Production Research, vol. 58, no. 1, pp. 18–43, 2020, doi: 10.1080/00207543.2019.1612964.

[7] G. Büyüközkan and F. Göçer, “Digital Supply Chain: Literature review and a proposed framework for future research,” Computers in Industry, vol. 97, pp. 157–177, 2018, doi: 10.1016/j.compind.2018.02.010.

[8] N. Boysen, R. de Koster, and F. Weidinger, “Warehousing in the e-commerce era: A survey,” European Journal of Operational Research, vol. 277, no. 2, pp. 396–411, 2019, doi: 10.1016/j.ejor.2018.08.023.

[9] K. Azadeh, R. de Koster, and D. Roy, “Robotized and Automated Warehouse Systems: Review and Recent Developments,” Transportation Science, vol. 53, no. 4, pp. 917–945, 2019, doi: 10.1287/trsc.2018.0873.

[10] T. van Gils, K. Ramaekers, A. Caris, and R. B. M. de Koster, “Designing efficient order picking systems by combining planning problems: State-of-the-art classification and review,” European Journal of Operational Research, vol. 267, no. 1, pp. 1–15, 2018, doi: 10.1016/j.ejor.2017.09.002.

[11] S. F. Wamba, A. Gunasekaran, S. Akter, S. J.-f. Ren, R. Dubey, and S. J. Childe, “Big data analytics and firm performance: Effects of dynamic capabilities,” Journal of Business Research, vol. 70, pp. 356–365, 2017, doi: 10.1016/j.jbusres.2016.08.009.

[12] P. Mikalef and M. Gupta, “Artificial intelligence capability: Conceptualization, measurement calibration, and empirical study on its impact on organizational creativity and firm performance,” Information & Management, vol. 58, no. 3, Art. no. 103434, 2021, doi: 10.1016/j.im.2021.103434.

[13] I. M. Enholm, E. Papagiannidis, P. Mikalef, and J. Krogstie, “Artificial Intelligence and Business Value: a Literature Review,” Information Systems Frontiers, vol. 24, pp. 1709–1734, 2022, doi: 10.1007/s10796-021-10186-w.

[14] G. Vial, “Understanding digital transformation: A review and a research agenda,” The Journal of Strategic Information Systems, vol. 28, no. 2, pp. 118–144, 2019, doi: 10.1016/j.jsis.2019.01.003.

[15] P. C. Verhoef et al., “Digital transformation: A multidisciplinary reflection and research agenda,” Journal of Business Research, vol. 122, pp. 889–901, 2021, doi: 10.1016/j.jbusres.2019.09.022.

[16] N. Müller, P. Burggräf, F. Steinberg, C. R. Sauer, and M. Schütz, “An analytical review of predictive methods for delivery delays in supply chains,” Supply Chain Analytics, vol. 11, Art. no. 100130, 2025, doi: 10.1016/j.sca.2025.100130.

[17] M. M. Abushaega, O. Y. Moshebah, A. Hamzi, and S. Y. Alghamdi, “Enhancing supply chain resilience with data envelopment analysis and temporal convolutional networks for supplier efficiency and late delivery risk prediction,” Alexandria Engineering Journal, vol. 128, pp. 231–246, 2025, doi: 10.1016/j.aej.2025.05.062.

[18] K. R. Ahmed, M. E. Ansari, M. N. Ahsan, A. Rohan, M. B. Uddin, and M. A. H. Rivin, “Deep learning framework for interpretable supply chain forecasting using SOM ANN and SHAP,” Scientific Reports, vol. 15, Art. no. 26355, 2025, doi: 10.1038/s41598-025-11510-z.

[19] A. Katangoori, “An empirical analysis of data-driven supply chain optimization in retail and logistics,” Supply Chain Analytics, vol. 14, Art. no. 100206, 2026, doi: 10.1016/j.sca.2026.100206.

[20] K. Sadeghi R., D. Ojha, P. Kaur, R. V. Mahto, and A. Dhir, “Explainable artificial intelligence and agile decision-making in supply chain cyber resilience,” Decision Support Systems, vol. 180, Art. no. 114194, 2024, doi: 10.1016/j.dss.2024.114194.

[21] A. Barredo Arrieta et al., “Explainable Artificial Intelligence (XAI): Concepts, taxonomies, opportunities and challenges toward responsible AI,” Information Fusion, vol. 58, pp. 82–115, 2020, doi: 10.1016/j.inffus.2019.12.012.

[22] S. M. Lundberg and S.-I. Lee, “A Unified Approach to Interpreting Model Predictions,” in Advances in Neural Information Processing Systems 30, 2017, pp. 4765–4774.

[23] V. C. Storey, A. R. Hevner, and V. Y. Yoon, “The design of human-artificial intelligence systems in decision sciences: A look back and directions forward,” Decision Support Systems, vol. 182, Art. no. 114230, 2024, doi: 10.1016/j.dss.2024.114230.

[24] S. Raisch and S. Krakowski, “Artificial Intelligence and Management: The Automation–Augmentation Paradox,” Academy of Management Review, vol. 46, no. 1, pp. 192–210, 2021, doi: 10.5465/amr.2018.0072.

[25] F. Constante, F. Silva, and A. Pereira, “DataCo SMART SUPPLY CHAIN FOR BIG DATA ANALYSIS,” Mendeley Data, Version 5, 2019, doi: 10.17632/8gx2fvg2k6.5.

Reference verification note. Bibliographic metadata for References [1]-[25] was checked against publisher or official scholarly records before manuscript preparation. A separate verification log accompanies this manuscript.