Integrated Predictive Pipeline for Supply Chain Risk Mitigation and Operational Optimization

Date: October, 2025 | Domain: Supply Chain Analysis

Executive Summary

This white paper details the construction, validation, and strategic application of the **SARIMAX-XGBoost Cascade Model (SXCM)**, a novel, two-stage predictive architecture developed to manage daily operational risk across our global supply chain. The project's primary outcome is the transformation of historical data into a comprehensive diagnostic and predictive tool, enabling a shift from reactive crisis management to proactive, data-driven operational control.

Key Findings and Strategic Mandates

Area	Finding	Strategic Implication
Model Performance	Achieved an exceptional historical fit with **Root Mean Squared Error (RMSE) of 0.0273** on the Daily High-Risk Proportion (P_High_Risk).	**Reliable Deployment:** The model is validated and ready to provide accurate, day-ahead risk predictions for resource allocation and operational scheduling.
Structural Bottleneck	Identified a pervasive ~**5.1-hour structural delay** that is uniform across all historical operational risk categories (Low, Moderate, High).	**Long-Term Mandate:** External volatility is secondary. The priority must be a comprehensive process re-engineering effort (Lean/Six Sigma) to eliminate this fixed, internal systemic inefficiency.
Forecast Paradox	The 365-day forecast predicts a baseline state of **chronic risk saturation** ($0.75 \leq \hat{P} \leq 0.79$).	**Systemic Warning:** The current operational structure lacks the necessary "slack" or buffer capacity to absorb normal daily fluctuations, leading to perpetual high-risk exposure.
Actionable Strategy	Developed **Dynamic Percentile Bucketing (DPB)** to classify the saturated forecast into actionable Low, Moderate, and High relative risk categories.	**Tactical Deployment:** Enables daily, data-driven resource allocation (e.g., reserving expensive backup capacity only for the 25% worst-predicted days).

Conclusion

The SXCM pipeline is both a forecaster and a diagnostic instrument. While delivering a high-accuracy prediction, it also confirms that the organization is operating at an unsustainable ~75% risk baseline, a condition driven entirely by internal process constraints rather than fluctuating external factors. The immediate strategic priority is the deployment of the DPB forecast for tactical planning, coupled with a mandated, aggressive long-term project to eliminate the ~5.1hr structural bottleneck. **Until this** ~5.1 **hour "process tax" is removed, the system will remain incapable of truly low-risk operation.**

Chapter 1: Project Overview and Methodology

1.1 Strategic Imperative and Project Goals

The core challenge addressed by this project is the historical reliance on intuition and lagged indicators for operational decision-making. This approach inherently results in a reactive stance, where resources are only deployed *after* a delay or failure has occurred. Our strategic imperative is to integrate a forward-looking, high-fidelity model that synthesizes temporal patterns and external factors (e.g., specific traffic patterns, localized weather events) to predict the **Daily Proportion of High-Risk Hours** (P_High_Risk), thereby enabling proactive intervention. This is a fundamental shift from logistics management to **predictive supply chain intelligence**.

1.1.1 Two-Phase Model Objectives

1. **Temporal Forecasting (SARIMAX):** Predict Delivery Time Deviation (hours). The SARIMAX model captures the critical seasonal, trend, and autoregressive components of delays, integrating filtered, uncorrelated exogenous variables (traffic, weather, port capacity). The output is a highly stable prediction of expected delay.

2. **Phase II (XGBoost):** This phase handles the non-linear transformation. The XGBoost regressor predicts the complex, continuous target, **Daily Operational Risk Proportion** (P_High_Risk). Crucially, the model uses the Phase I forecast as its primary feature, leveraging its non-linear decision tree capabilities to combine the temporal forecast with highly predictive lagged operational data (system memory), which dictates the network's current state of stress.

Chapter 2: Exploratory Data Analysis (EDA) and Data Preparation

2.1 Raw Data Characteristics and Granularity

The foundation of this analysis is three years of granular, hourly operational data, totalling over 32,000 observations which allowed us to capture transient, hour-by-hour changes in risk drivers such as localized congestion and micro-weather events. This depth provides the necessary resolution to decompose temporal patterns accurately. The variables being considered for our analysis can be summarised from the data description provided here.

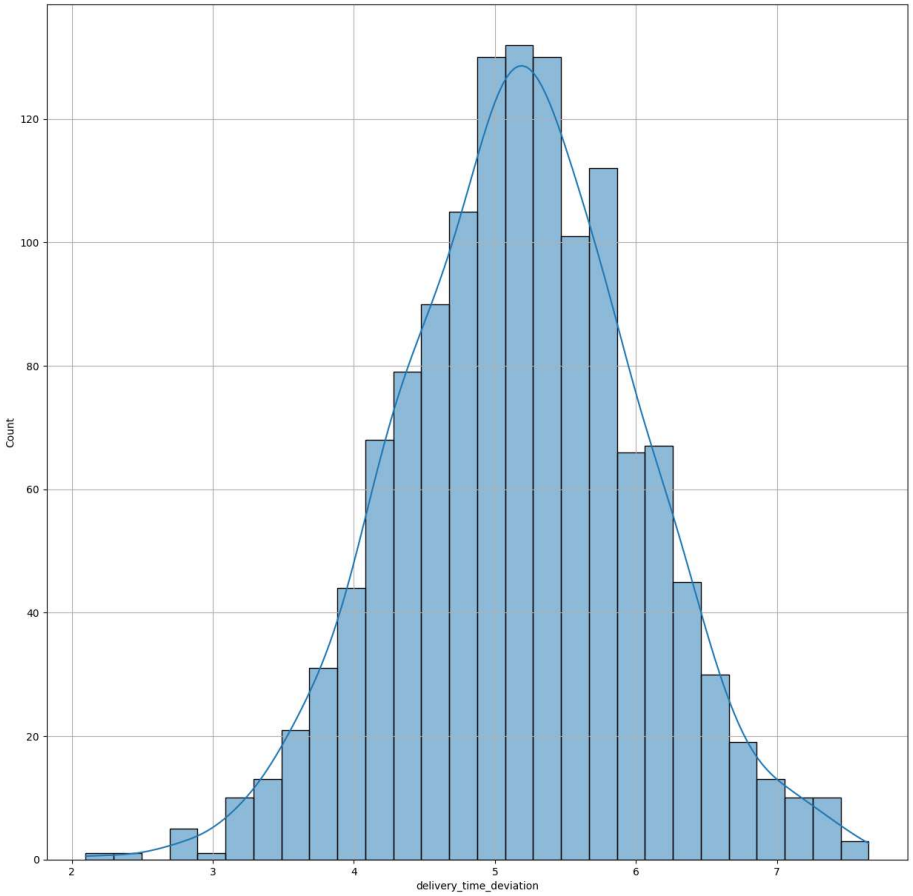
Sr. No	Feature Names	Variable type	Description	Datatype
1	**Timestamp**	Feature	The date and time when the data was recorded (hourly resolution)	*datetime*
2	**Vehicle GPS Latitude**	Feature	The latitude coordinate indicating the location of the vehicle	*float*
3	**Vehicle GPS Longitude**	Feature	The longitude coordinate indicating the location of the vehicle	*float*
4	**Fuel Consumption Rate**	Feature	The rate of fuel consumption recorded for the vehicle in litres per hour	*float*
5	**ETA Variation (hours)**	Feature	The difference between the estimated and actual arrival times	*float*
6	**Traffic Congestion Level**	Feature	The level of traffic congestion affecting the logistics route	*float*
7	**Warehouse Inventory Level**	Feature	The current inventory levels at the warehouse (units)	*float*
8	**Loading/Unloading Time**	Feature	The time taken for loading or unloading operations in hours	*float*
9	**Handling Equipment Availability**	Feature	Availability status of equipment like forklifts	*float*
10	**Order Fulfilment Status**	Feature	Status indicating whether the order was fulfilled on time	*float*
11	**Weather Condition Severity**	Feature	The severity of weather conditions affecting operations	*float*
12	**Port Congestion Level**	Feature	The level of congestion at the port	*float*
13	**Shipping Costs**	Feature	The costs associated with the shipping operations in USD	*float*
14	**Supplier Reliability Score**	Feature	A score indicating the reliability of the supplier	*float*
15	**Lead Time (days)**	Feature	The average time taken for a supplier to deliver materials	*float*
16	**Historical Demand**	Feature	The historical demand for logistics services (units)	*float*
17	**IoT Temperature**	Feature	The temperature recorded by IoT sensors in degrees Celsius	*float*
18	**Cargo Condition Status**	Feature	Condition status of the cargo based on IoT monitoring	*float*
19	**Route Risk Level**	Feature	The risk level associated with a particular logistics route	*float*
20	**Customs Clearance Time**	Feature	The time required to clear customs for shipments	*float*
21	**Driver Behaviour Score**	Feature	An indicator of the driver's behaviour based on driving patterns	*float*
22	**Fatigue Monitoring Score**	Feature	A score indicating the level of driver fatigue	*float*
23	**Disruption Likelihood Score**	Feature	A score predicting the likelihood of a disruption occurring	*float*
24	**Delay Probability**	Feature	The probability of a shipment being delayed	*float*
25	**Risk Classification**	Target	A categorical classification indicating the level of risk	*categorical*
26	**Delivery Time Deviation**	Target	The deviation in hours from the expected delivery time	*float*

2.2 Univariate Analysis and Distributional Skewness

Detailed univariate analysis revealed significant statistical properties that directly informed the necessary data preparation steps.

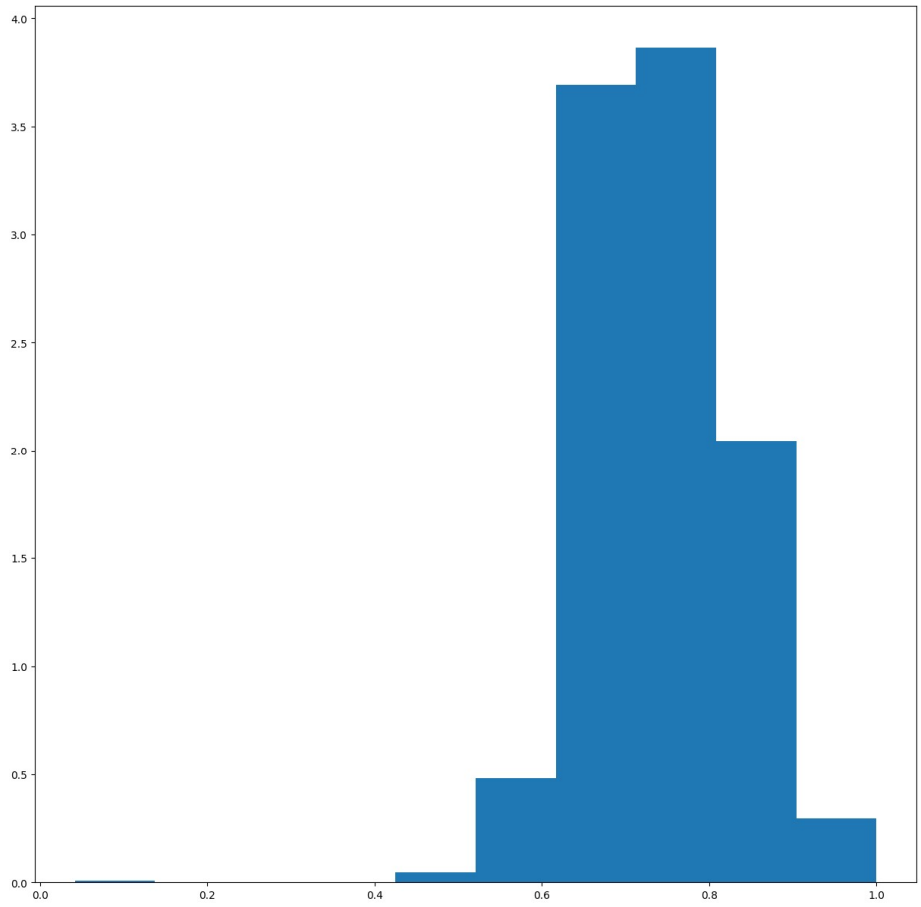
2.2.1 Target Variable Characteristics

- **Delivery Time Deviation ($Y_{Deviation}$):** This variable exhibited normality (skew ~0.5). The distribution was dominated by observations near the mean or slightly right-skewed data (representing delays in deliveries of around 5 hours), indicating that the process has an inherent delay baked into it.



(Fig. 2.2.1: Normality of 'delivery time deviation')

- **High-Risk Proportion (P_{High_Risk}):** The empirical distribution was heavily concentrated in the upper quartiles (near 0.75), confirming the prior observation of a high operational risk baseline. The objective became to predict the small, but critical, continuous movement of this P_High_Risk value around this high mean, justifying the use of a high-fidelity **XGBoost Regressor** capable of subtle boundary prediction.



(Fig. 2.2.2: Histogram for 'High Risk Proportion').

Risk Classification	Frequency	Proportion
High Risk	**23944**	**74.67%**
Moderate Risk	5011	15.63%
Low Risk	3110	9.7%

(Fig. 2.2.3: Disproportionate risk classification)

These plots visually confirm the severe right-skewness and the high concentration of the risk proportion near the 75% mark, serving as a critical diagnostic visualization that most if not all the deliveries are under the 'High risk' category throughout the time period.

2.3 Outlier Identification and Strategic Retention (Isolation Forest)

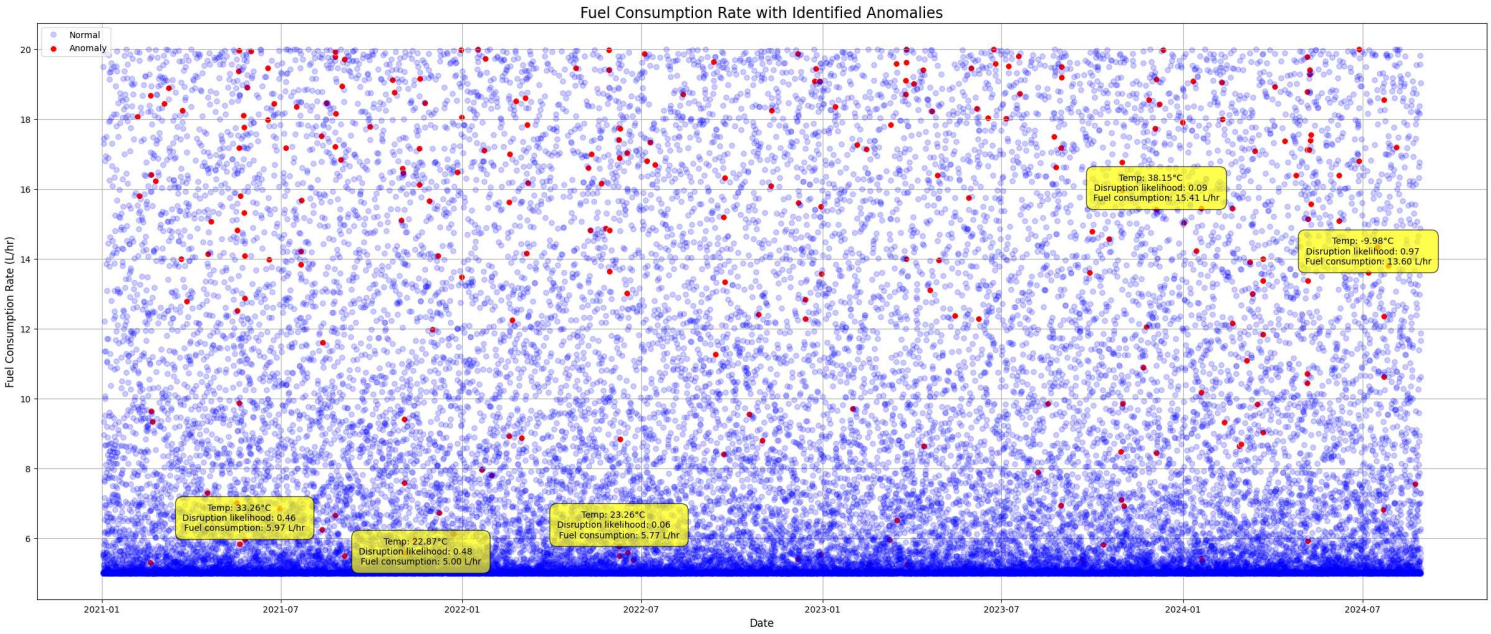
Outliers, while statistically rare, contain crucial information about non-standard, extreme logistical events. Deleting these data points would erase evidence of significant, albeit non-repeatable, historical failures. Instead of using traditional descriptive statistics, the more robust **Isolation Forest** machine learning algorithm was employed to identify anomalies across the multi-dimensional feature space.

2.3.1 Outlier Analysis and Retention Decision

The Isolation Forest analysis revealed that true, statistically extreme outliers constituted **less than** 1% **of the entire 32,000+ data points**. Given this extremely low volume and the critical intelligence these extreme events represent (genuine historical failures that the model must be aware of), the decision was made to **strategically retain these outliers without modification or capping.** The primary justification for retention was two-fold:

1. **Low Volume:** The small fraction of outliers was deemed insufficient to significantly bias the overall statistical mean or variance, especially after the subsequent application of **first-order differencing** in the SARIMAX model (expanded in Section 4.1).

2. **XGBoost Robustness:** The Phase II **XGBoost Regressor** is an inherently tree-based model that is non-parametric and highly robust against outliers. Retaining the raw, uncapped values allows the XGBoost to learn the full spectrum of high-impact events without the dampening effect of imputation or capping.



(Fig. 2.3.1: Top 5 outliers for the data identified using Isolation Forest)

2.4 Multicollinearity Screening (Variance Inflation Factor - VIF)

Multicollinearity, or high linear correlation among independent variables, must be rigorously managed. If two exogenous features are highly correlated, their coefficients in the linear SARIMAX model become unstable and statistically meaningless, reducing both accuracy and interpretability.

2.4.1 VIF Methodology and Threshold

The Variance Inflation Factor (VIF) metric quantifies how much the variance of a regression coefficient is inflated due to collinearity. Features were iteratively removed **iff** they exceeded a conservative threshold of VIF>5. This process ensured that the **remaining 21 exogenous features** (X_{VIF}) were statistically orthogonal enough to guarantee the stability and reliable inference of the Phase I model. From the VIF check, we see that the highest VIF seen was no more than 1.63, indicating that ALL the features were free from multi-collinearity and were well below the threshold.

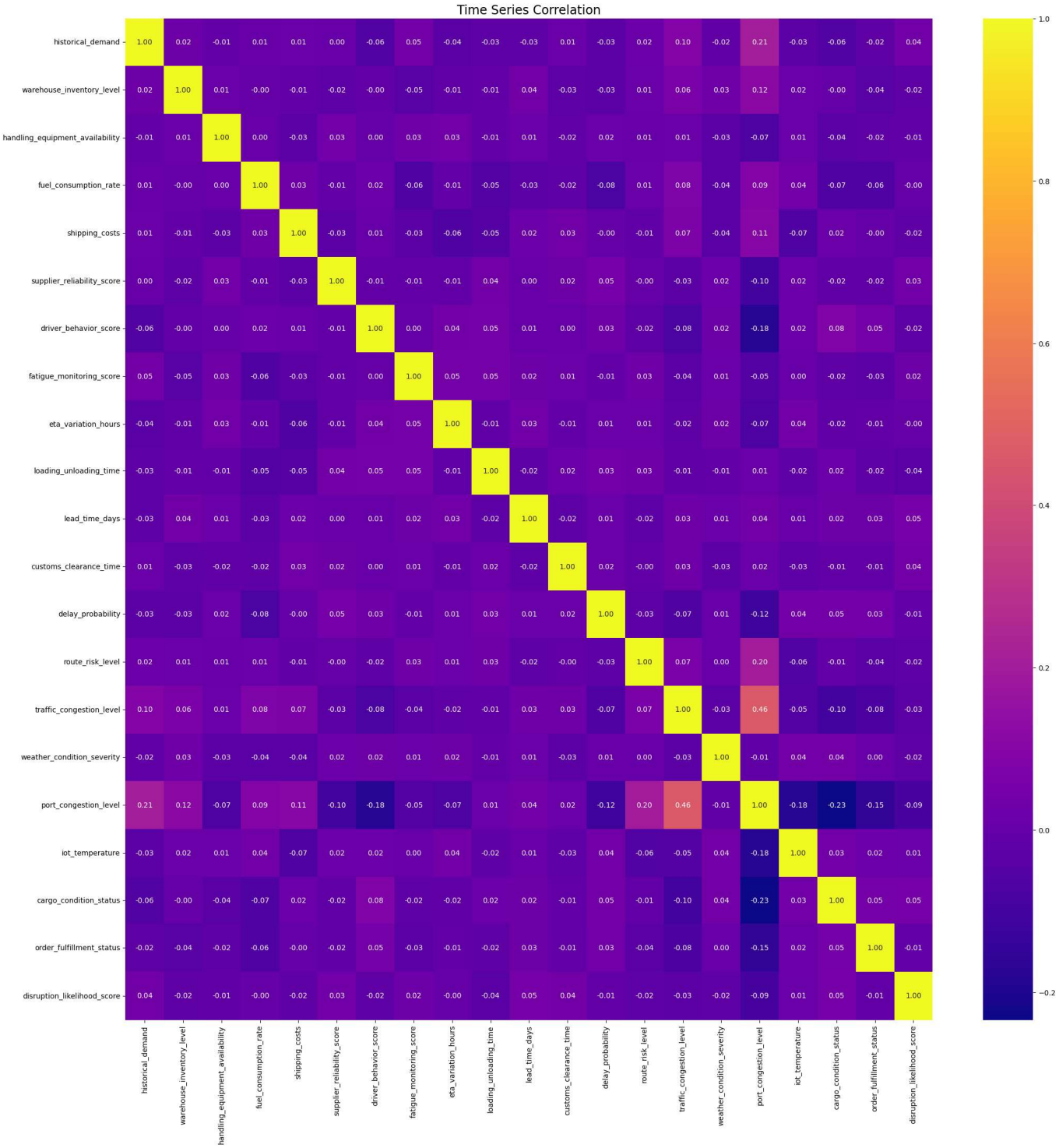
Sr. No	Feature	VIF
1	port_congestion_level	1.625268
2	traffic_congestion_level	1.284417
3	cargo_condition_status	1.080174
4	historical_demand	1.057317
5	route_risk_level	1.050522
6	iot_temperature	1.045549
7	driver_behavior_score	1.045484
8	fuel_consumption_rate	1.035413
9	order_fulfillment_status	1.032955
10	shipping_costs	1.029183
11	delay_probability	1.025552
12	disruption_likelihood_score	1.024855
13	fatigue_monitoring_score	1.024219
14	warehouse_inventory_level	1.023997
15	supplier_reliability_score	1.022137
16	loading_unloading_time	1.019434
17	handling_equipment_availability	1.018218
18	eta_variation_hours	1.017435
19	lead_time_days	1.016505
20	weather_condition_severity	1.012446
21	customs_clearance_time	1.008894

(Fig. 2.4.1: Variance Inflation Factor analysis of the features to detect multi-collinearity)

2.5 Bivariate and Temporal Analysis

2.5.1 Correlation

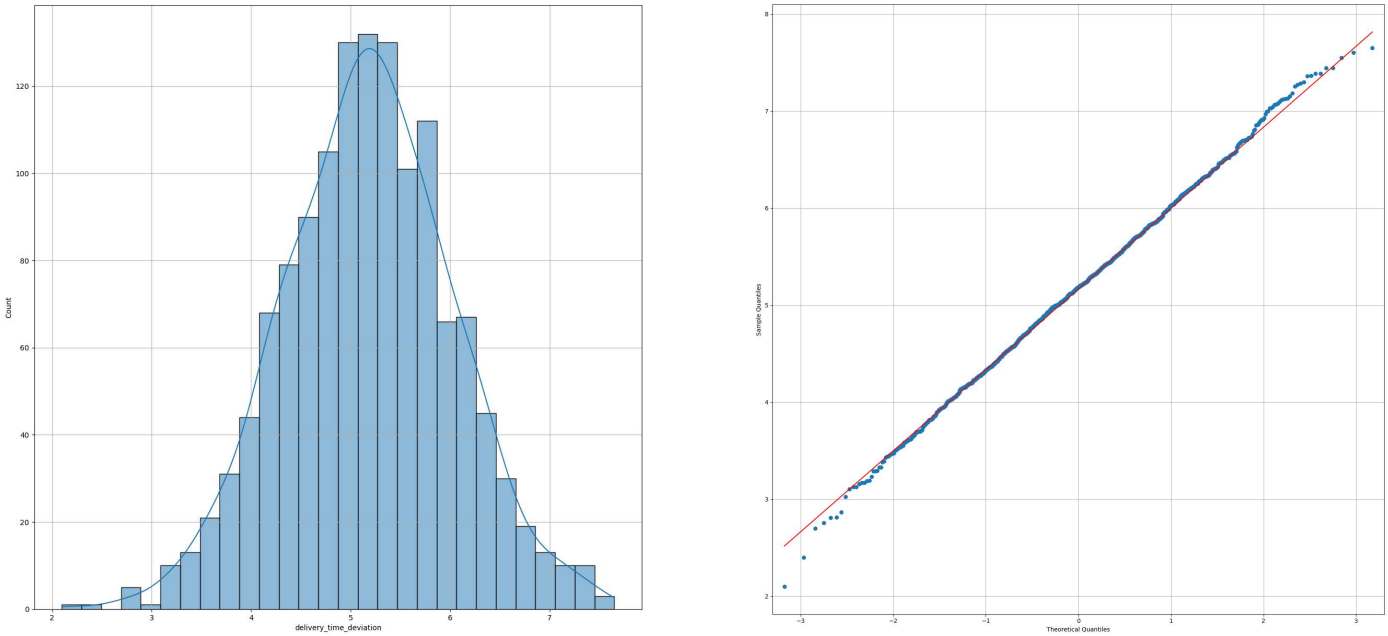
The Pearson Correlation heatmap shows that the only moderate correlation was between the traffic level and the port congestion level at any given time, indicating that the driver or rather, the delivery system got stuck in a never-ending loop where the traffic levels were being defined by the amount of congestion at the port, which in turn drove up traffic levels. However, the other features had weak correlation, indicating that the features taken for our analysis, were truly independent from each other and had minimal interactions with each other, confirming our earlier analysis from VIF scores.



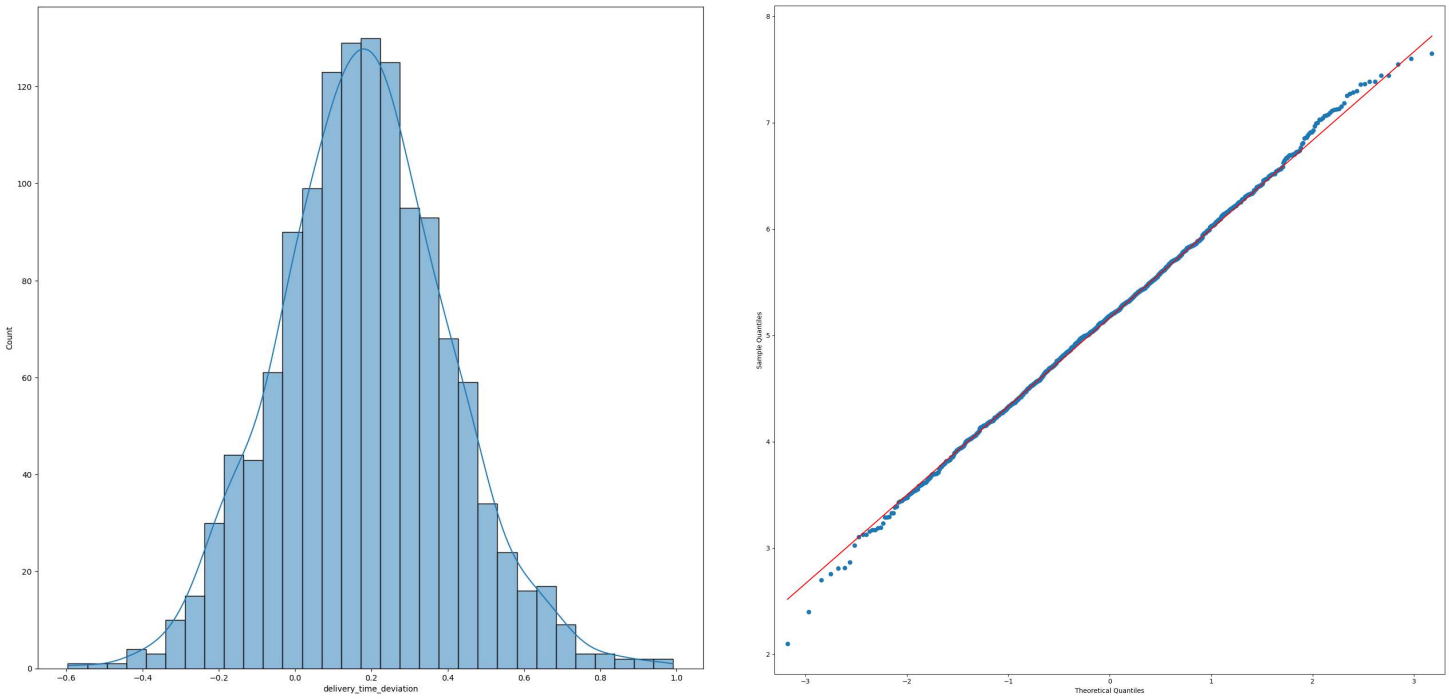
(Fig. 2.5.1: Correlation heatmap between the features)

2.5.2 Time Series Decomposition

Time Series analysis of the target data was assumed to be **Additive** due to the relative exactness of the Gaussian nature of the histogram plots, both the original as well as the log-lagged plots. Although visually showing normal distributions, their relative unchanging nature even after log-lagged transformations were statistically confirmed when checking their Q-Q plots as shown below.



(Fig. 2.5.2: Normal distribution of raw delivery delay times, and its confirmation using the Q-Q plot)



(Fig. 2.5.3: Normal distribution of log-lagged delivery delay times, and its confirmation using the Q-Q plot)

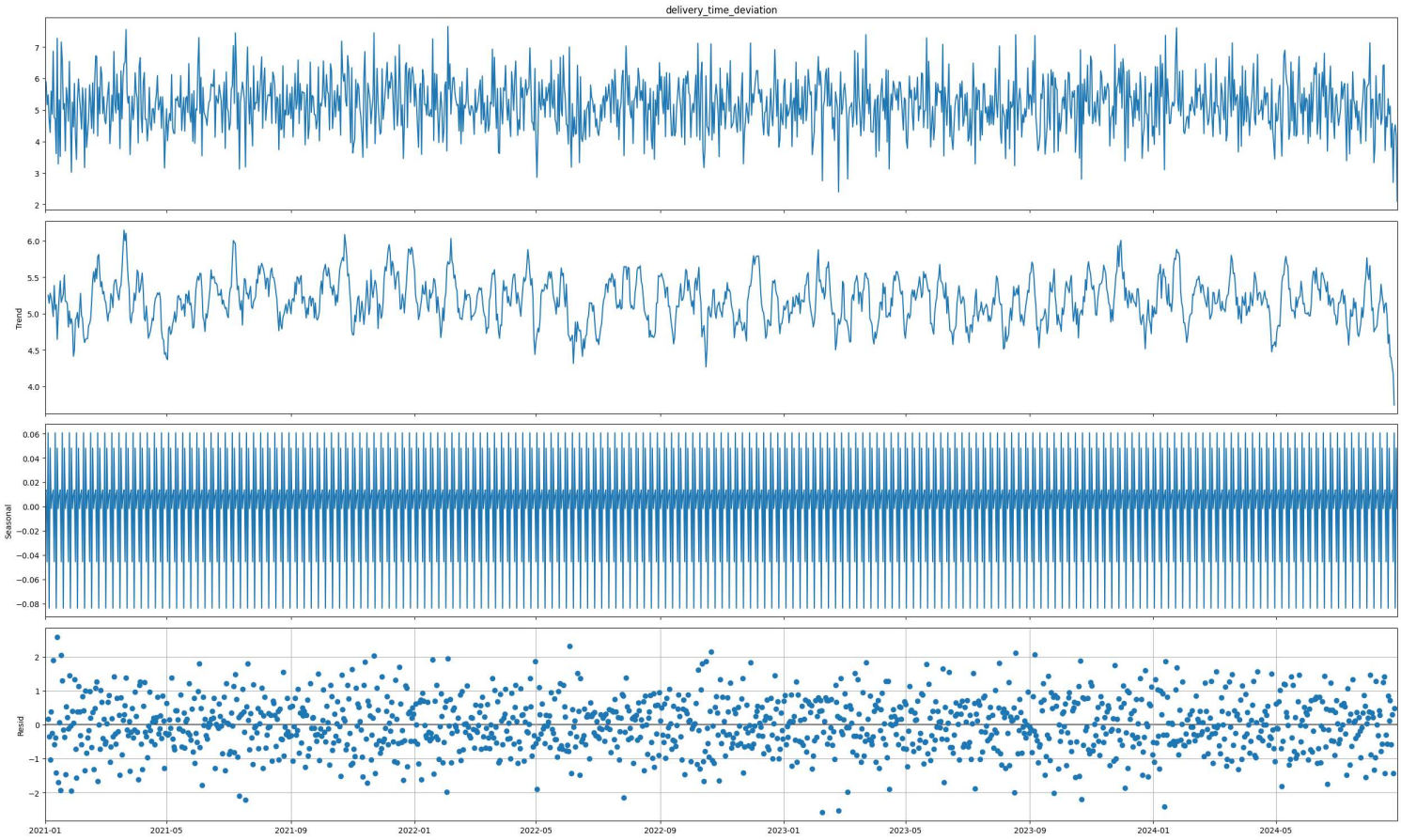
The plots show no difference in the nature of the normality, and the log-lagged transformation even shares a near-identical plot distribution with the original histogram and the Q-Q plot. This test can later be used to indicate the nature of the model when setting up the SARIMAX parameters.

The Decomposition plots of the Delivery Deviation series, showcases **minimal trend, but extremely high (weekly) seasonality,** along with the Additive nature of the time-series data:

- **Trend:** A subtle, but persistent, positive trend was noted, indicating a marginal, gradual degradation in long-term delivery performance, likely due to small operational improvements to account for logistical shocks.

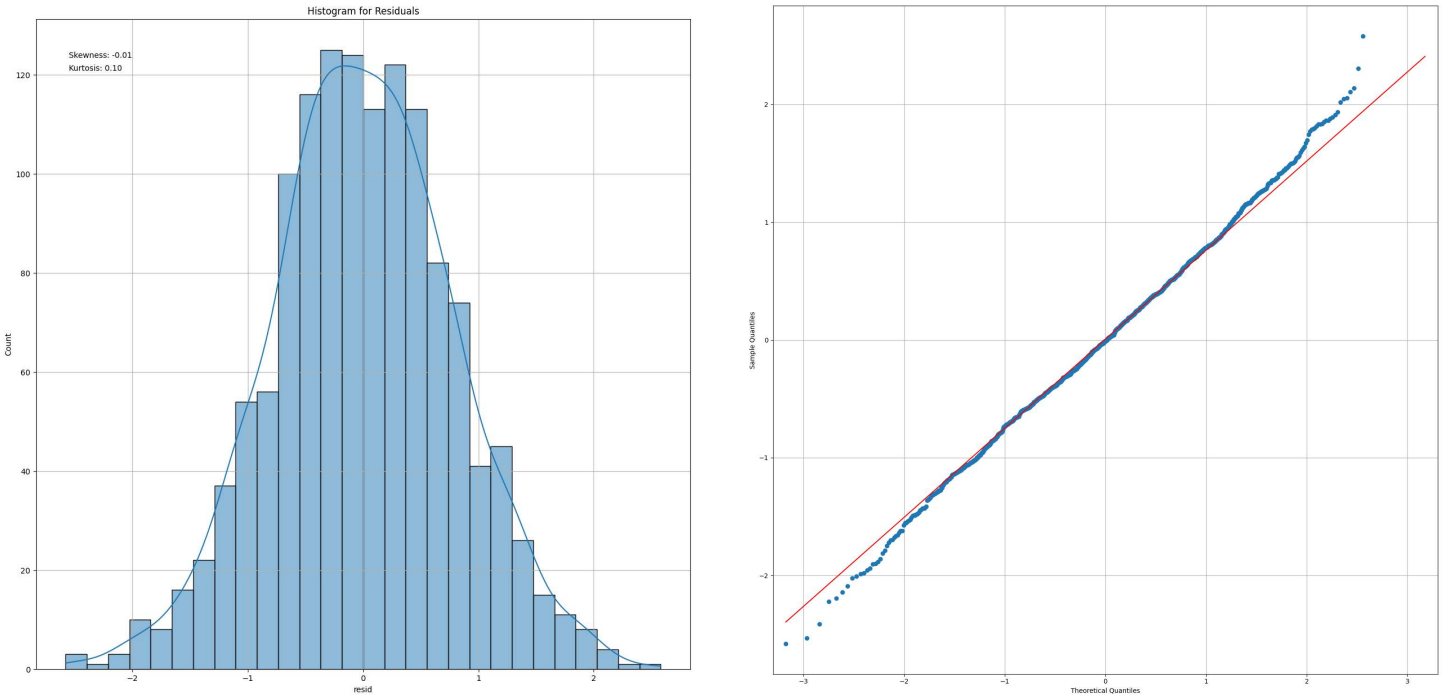
- **Seasonality:** Strong **weekly seasonality** (s=7) was confirmed, with repeatable peaks on specific weekdays corresponding to logistical cycles (e.g., Monday backlog accumulation or Saturday rush volume).

- **Residuals:** Random distribution of the residuals which showed no pattern (indicating its stochastic nature such as no slope or polynomial relations), or changing variance (such as funnels) which was later confirmed by its normality in the Q-Q plot.



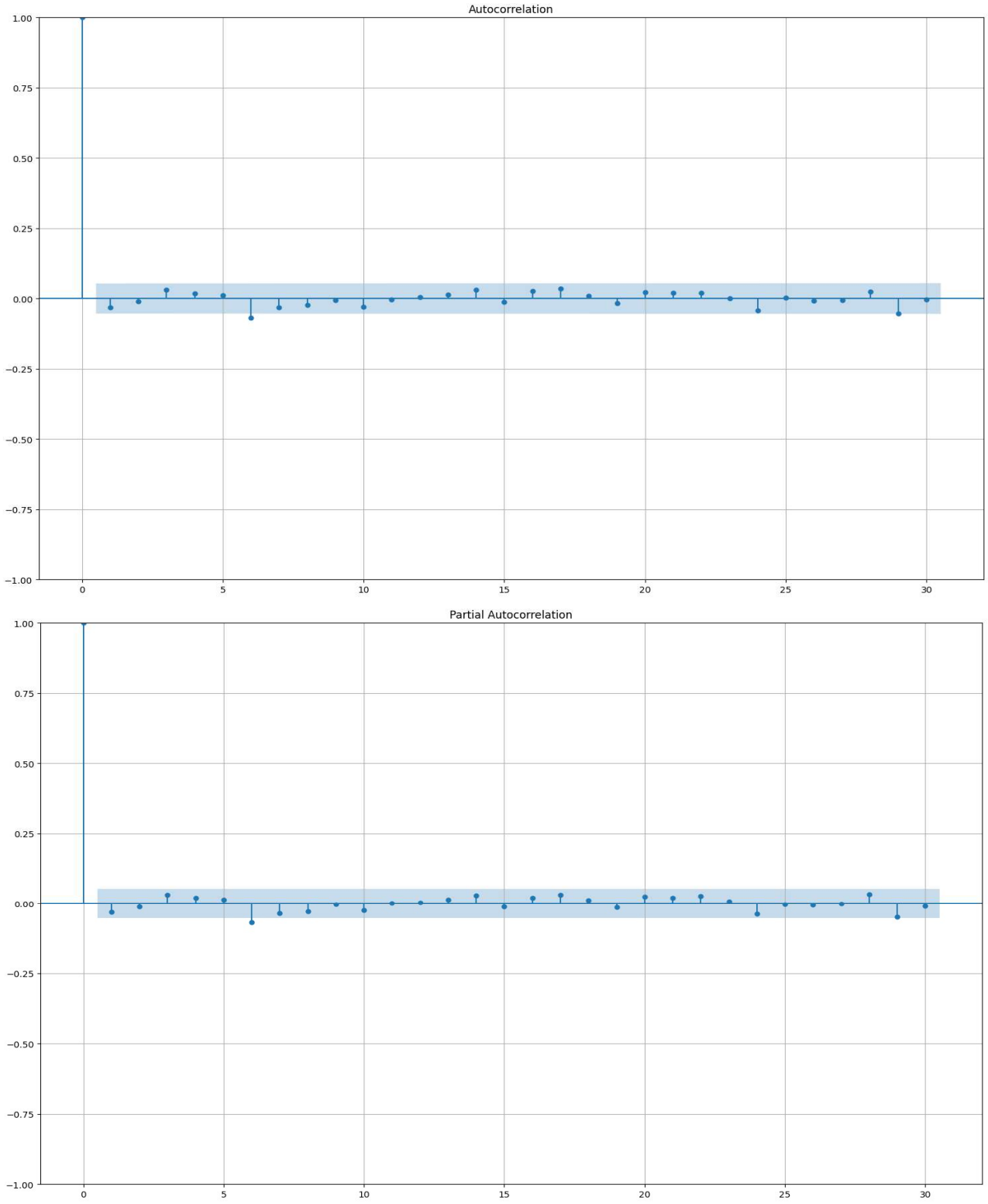
(Fig. 2.5.4: Time Series Decomposition Plot for Delivery Deviation. This plot visually confirms the clear weekly cycle and the marginal upward trend over the historical period.)

Nature: Further confirmation of the Additive nature of the plot can be seen from the Residuals of the Time Series decomposition plot, where the graphs show now change in variance due to randomised distributions, as well as no fluctuations in the amplitude or any set patterns, which further lead credence to the fact that the time-series may have an additive nature. The final confirmation can be provided by mapping the distribution and Q-Q plot of the residuals.



(Fig. 2.5.5: Normal distribution of residuals, and its confirmation using the Q-Q plot)

Autocorrelation: Detailed analysis of the Autocorrelation Function (ACF) and Partial Autocorrelation Function (PACF) plots confirmed the specific auto-regressive and moving-average dependencies required to accurately specify the SARIMAX model order, which turned out to be 0 for both the P and Q parameters of the SARIMAX model.



(Fig. 2.5.6: ACF – PACF plots of the target variable time-series, note the lack of Auto-Regressive and Moving Average features)

Stationarity: A fundamental requirement for SARIMAX modelling, stationarity in Time-series has statistical properties — (mean and variance) that remain constant over time. This constancy ensures that the patterns and relationships learned from the past (the training data) are reliable and will hold true in the future, allowing us to generate valid and accurate forecasts. As such, we will use the 2 most widely used tests to check for complete stationarity, viz., the Augmented Dicky Fuller (ADF) Test and the Kwiatkowsky Phillips Schmidt Shin (KPSS) Test. These statistical tests completely cement our assumptions that the time series for Delivery Delays is completely Stationary, by looking at the following Test characteristics of each.

- **ADF Test**: The p-value for the test is less than the Critical p-value measured at 95% Confidence Level for a left-tailed test; we can thus, safely reject the null hypothesis, meaning there is a complete absence of a Unit Root, indicating that the Time Series is Stationary. Another formal way to confirm this is the test statistic of -37.52 which is much lesser as compared to even the 1% Critical value of -3.43, which is another indication that the Null Hypothesis can be rejected.

Augmented Dickey-Fuller (ADF) Test results	
Test Statistic	-37.523795
p-value	0.000000
#Lags Used	0.000000
Number of Observations Used	1336.000000
Critical Value (1%)	-3.435254
Critical Value (5%)	-2.863706
Critical Value (10%)	-2.567923

- **KPSS Test:** The p-value for the test is greater than the Critical p-value measured at 95% Confidence Level for a right-tailed test, meaning we cannot reject the null hypothesis which indicates the presence of Trend Stationarity in the Time Series, which is what we had initially assumed. The other check that confirms our assumptions is the test statistic of 0.25 which is also lesser than the 10% Critical value of 0.35, leading further credence to the fact that we must NOT reject the Null Hypothesis.

Kwiatkowsky Phillips Schmidt Shin (KPSS) Test results	
Test Statistic	0.25258
p-value	0.10000
Lags Used	4.00000
Critical Value (10%)	0.34700
Critical Value (5%)	0.46300
Critical Value (2.5%)	0.57400
Critical Value (1%)	0.73900

From these observations, we can thus finally go ahead with the target Time Series as is for our modelling; as the Series is Stationary, we do not require any form of transformation for the same, meaning the Trend Differencing ‘d’ as well as the Seasonal Differencing ‘D’ in the SARIMAX model is 0.

Chapter 3: Feature Engineering and Data Finalization

3.1 Aggregation to Daily Cadence

The shift from hourly operational data to a daily prediction cadence was a strategic move to align the model's output with the organization's resource planning cycle:

- **Noise Reduction**: Hourly data can be very noisy and volatile. Customers place orders at different times of the day for many reasons, not all of which are meaningful for long-term trends. Aggregating the data smooths out this noise, making it easier to identify the underlying trends and seasonality; making any system too responsive will also contribute to making it volatile, and will be forced to become a 'reactive' system, rather than a 'predictive' system which is more stable and easier to control.

- **Seasonality**: While we have hourly seasonality (e.g., more orders during the day than at night), most strategic and tactical business decisions (like inventory management) are made on a daily, weekly, or monthly basis. Aggregating to these levels will reveal more significant, practical and actionable patterns, such as a spike in demand on weekends or during the end of the month; this will also help prevent excessive demand fluctuations and bull-whip effects.

3.1.1 Aggregation Rules

1) **Target (P_{High_Risk}):** Calculated as the **proportion** of the hourly **high-risk** indicator used every 24 hours, providing a continuous risk score for the day.

Numerical conversion for imbalanced classes				
timestamp	**High-risk hours**	**Low-risk hours**	**Moderate-risk hours**	**P_{High_Risk}**
2021-01-01	19	0	5	0.791667
2021-01-02	21	1	2	0.875000
2021-01-03	19	1	4	0.791667
2021-01-04	16	4	4	0.666667
2021-01-05	19	2	3	0.791667

(Fig. 3.1.1: Conversion of hourly categorical risk classification to daily proportion of high-risk hours as percentage of 24 hours)

2) **Features:** Aggregated based on the nature of the data. The decision to use different aggregation methods when moving from hourly to daily data for Phase I (SARIMAX) was based on the operational nature of each feature and its function as a risk predictor. Here is the breakdown of the feature aggregation rules:

   a) **Maximum (Max) Aggregation: Capturing Worst-Case Scenarios**
   The maximum value was used for features where the single, most intense event of the day is the primary driver of **operational stress and bottlenecks**.

   - Operational Stress/Worst-Case: Features like Traffic Congestion Level or Weather Condition Severity were aggregated using the maximum. Operational risk is often dictated by the peak hour of stress, not the average, as the system must withstand this extreme moment.

   - Capacity Constraints: Features representing infrastructure strain, such as Port Capacity Utilization or Equipment Downtime Duration, used the maximum to capture the highest point of capacity breach.

   b) **Mean (Average) Aggregation: Capturing Rate-Based and Continuous States**
   The mean (average) was used for variables that represent a **stable, continuous state**, a rate-based metric, or the desired output target.

   - Rate-Based Metrics: Features like Hourly Risk Proportion (PHigh_Risk), Average Processing Queue Time, or other continuous, ratio-based metrics were averaged. This provides a stable, representative daily figure suitable for the linear SARIMAX model.

   - **Target Variable**: The primary target, Delivery Time Deviation ($Y_{Deviation}$), uses the mean as the model aims to forecast the expected average delay for the day.

   c) **Sum (Sum) Aggregation: Capturing Cumulative Totals**
   The sum was used for features where the total accumulated value over the 24-hour period provides the **most relevant predictive signal** of the day's overall workload and strain.

   - Cumulative Totals/Load: Metrics representing accumulated activity, such as Total Daily Volume Processed (Count), Total Accidents Reported, or Total Non-Productive Time, were summed. The total workload dictates the full extent of the stress carried by the supply chain network.

3.2 Advanced Temporal Feature Construction

To enhance the model's ability to interpret time cyclicity, the information of the cyclicity was extracted from the timestamp as both the day of the week (ranked) and day of the year (ranked). The weekday was then converted to a numerical ordinal category which would later be encoded to provide a category-based feature map of the timestamp, this will also help the model to capture any seasonal patterns in the data such as comparison of this Monday's IoT temperature sensor compared to last Monday.

Similarly, the day of the year would be directly converted to a numerical category (ordinal) which would then be converted to a cyclical system to indicate to the model that the model is associated with a cycle and not a random collection of numbers that ranges from 1 through 366, followed by another 1. This would make no sense and would throw off the model in trying to capture non-existent patterns due to such an encoding. In order to ensure each day of the year has a unique code (sine or cosine value by themselves repeat twice in a span of 180 days) we will provide both the values to ensure each row has a unique number to identify themselves with.

$$\text{sin\_day} = \frac{sin(2\pi \times Day\_of\_Year)}{365.25}$$

$$\text{cos\_day} = \frac{cos(2\pi \times Day\_of\_Year)}{365.25}$$

This encoding prevents the **numerical discontinuity** issue inherent in treating "Day 1" as numerically far from "Day 365." By projecting the time variable onto a circle, the model smoothly interprets annual seasonality, correctly placing the predictive influence of late December close to early January.

3.3 Lagged Operational Features

Lagging is essential because the data is a time series, and even though the SARIMAX model showed no significant autocorrelation (no memory in the delay itself), the drivers of risk are highly dependent on recent history. The Regressor model we will use, the XGBoost model, by itself is time-agnostic; it treats every row (day) as an independent observation. To enable it to learn temporal patterns, we must embed the sequence of time directly into its feature matrix. Since the data had a 7-day periodic cycle from the previous AIC score SARIMAX modelling, we will do the lags for the following time-periods:

The final feature set for Phase II (X_{Full}) was purposefully built around **system memory** to quantify the impact of cumulative stress and backlog. The inclusion of lagged features is essential for capturing the autoregressive nature of supply chain stress.

- **Autoregressive Lag (Short-term memory) (*t–1*, *t-2* and *t-3*)**
  The previous day's operational risk proportion ($P_{High\_Risk,\ t-1}$) is the most powerful predictor, quantifying the immediate systemic impact of yesterday's capacity shortage on today's performance. It helps to capture the immediate carry-over effects like traffic congestion spill-over, backlog from the last few days, and momentum in high-risk events and are often the strongest short-term predictors.

- **Weekly Lag (Seasonal Memory) (t–7)**
  This lag helps capture the fixed, high-volume, and predictable cyclical stress that occurs every seven days (e.g., the cumulative effect of weekend downtime or mid-week resource clustering) that is crucial in logistics (e.g., comparing a Tuesday's risk to the previous Tuesday's risk). Since the SARIMAX model had a seasonal order of S = 7, including this lag validates that structural finding in the SARIMAX model.

The target variable is not lagged because it would introduce a severe error known as **Target Leakage** (or "look-ahead bias") into the forecasting model. A predictive model can only be trained using information that would be known at the time the prediction is made.

- **Training Failure**: If we were to use the lagged target (Y_{t-1}) as a feature to predict Y_t, the deployed model would require knowing today's actual risk proportion (Y_t) in order to predict tomorrow's risk (Y_{t+1}).

- **Deployment Failure**: In a real-time scenario, the actual risk proportion for today is unknowable until the day is over and the data is logged. Therefore, including the target's lagged value as an input makes the model unusable for making a true, proactive forecast.

Instead, we use the Lagged SARIMAX Prediction ($\hat{Y}_{SARIMAX,\ t-1}$) as a feature. This serves as the statistically sound substitute for the lagged target, providing the model with memory of the predicted outcome without violating the integrity of the forecasting process.

Chapter 4: Phase I: Delivery Deviation Forecasting (SARIMAX)

4.1 Model Specification

4.1.1 Parameter Optimization (AIC)

Optimal SARIMAX parameter (p, d, q) × (P, D, Q) $_{s=7}$ was determined via a structured grid search. The final parameters were selected based on the lowest **Akaike Information Criterion (AIC)** as well as mean RMSE score from conducting **Time-based Cross-Validations**. This methodology is statistically rigorous, ensuring the model achieves the best balance of explanatory power (fit to data) and parsimony (model simplicity), preventing an overly complex or overfit model that performs poorly on unseen data.

From the Grid Search, our initial estimates for the SARMAX parameters show that looking at the AIC scores, the parameters indicate the model is highly influenced by the Moving Average (q) past forecast error residuals over 2 days. Further investigation shows that this may be affected by the presence of high number of exogenous variables which can introduce the concept of multi-collinearity. We have already taken into account outliers in the features as SARIMAX is more robust to outliers and noisy data.

However, comparing the performance of the SARIMAX models, it was found that the RMSE for the initial SARIMAX (0,0,0) model is lower than the RMSE for the SARIMAX (0,0,2) model; this indicates that the initial model is superior for forecasting, even though its AIC was higher. The fact that the SARIMAX (0,0,2) model had a lower AIC means it was better at explaining the small, non-random noise patterns in the historical data. However, when tested on unseen data (via Cross-Validation), that complexity hurt its performance:

- **Overfitting**: The MA (2) terms were likely modelling historical noise rather than a true, persistent process.
- **Generalization Failure**: When exposed to new data, those parameters caused the model to chase non-existent patterns, resulting in larger errors (higher RMSE) compared to the simpler model.

The SARIMAX (0,0,0) model by contrast, is stable because it wisely chooses not to model the weak and unreliable correlations, relying purely on the strong signal from static exogenous variables. Thus, for our final forecasting, we will instead use the SARIMAX (0,0,0) x (0, 0, 0) $_{s=7}$ model, obtaining the final test RMSE score of 0.8097, which is lower than the Mean Cross-Validated RMSE score of 0.8678.

This suggests that the final period of data (the Test Set) was slightly easier to predict than the average historical period, or that training on the full training set immediately preceding the test set provided a marginal accuracy boost; because the Test RMSE is very close to the CV-RMSE (and even slightly better), we can be highly confident that our chosen model structure is the most stable and accurate choice for the delivery delay data.

4.2 Monte Carlo Simulation for Uncertainty Quantification

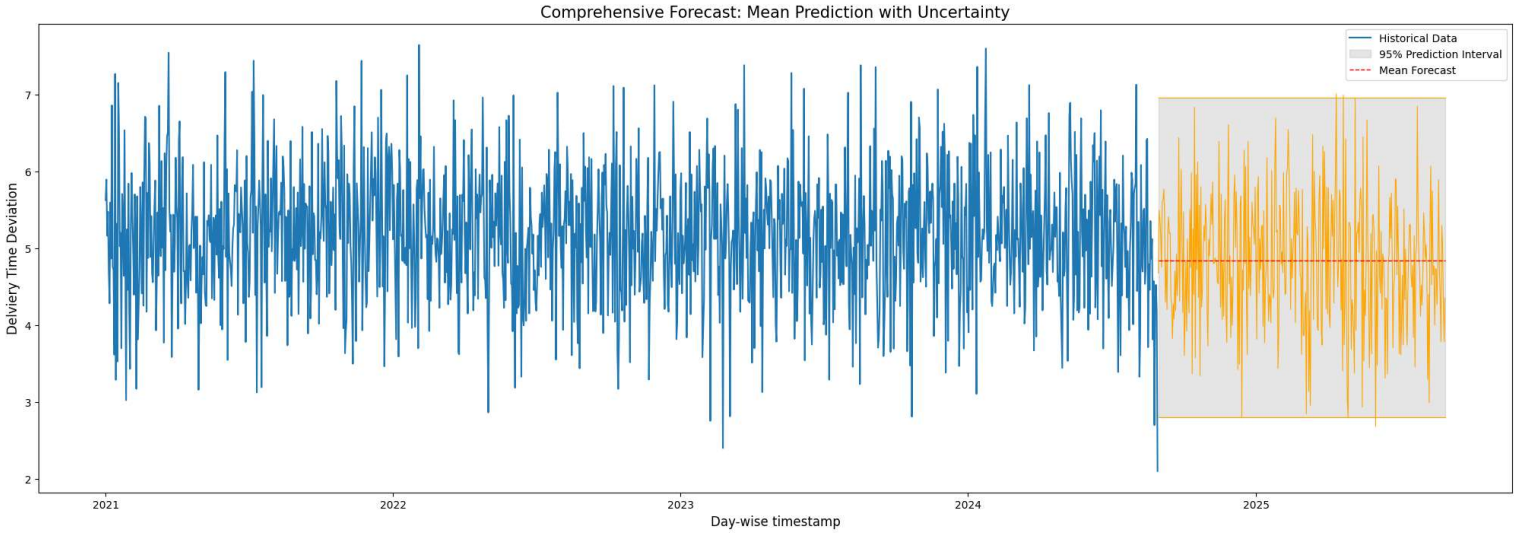
Given the inevitable complexity and noise in external inputs, a single, deterministic SARIMAX forecast line would be inadequate and misleading. We implemented a **Monte Carlo simulation** to quantify the full range of likely future outcomes.

4.2.1 Simulation and Output

The Monte Carlo method involved running the 365-day forecast 1000 separate times. In each iteration, a small, random shock (noise) was introduced, sampled directly from the SARIMAX model's own historical residual distribution (the errors it made historically). This process allows the simulation to accurately model future volatility based on the system's inherent historical unpredictability.

- **Phase II Input:** The average (mean) of the 1000 simulated forecasts yielded the stable, reliable SARIMAX_Predicted_Delay feature, which represents the most probable outcome fed into XGBoost.

- **Prediction Intervals:** The 2.5th and 97.5th percentiles across the 1000 simulated forecasts established the 95% Prediction Interval, providing a statistically sound measure of forecast confidence.



(Fig. 4.2.1: SARIMAX Forecast with Monte Carlo simulations and 95% Prediction Intervals)

This visualization clearly shows the mean forecast line bounded by the shaded 95% confidence region, illustrating the stability and quantified uncertainty of the Phase I output.

4.3 Diagnostic Discovery: The Structural Bottleneck

The most critical and unexpected finding derived from the Phase I model was the diagnostic discovery of the **structural bottleneck**, revealed through a detailed cross-sectional analysis of the historical data's performance floor:

Risk Classification	Mean Delivery Deviation (Hours)
High Risk	5.184
Moderate Risk	5.184
Low Risk	5.118

The near-uniform ~5.1-hour average delay across **all** observed risk levels demonstrates that this value is a **fixed, embedded operational cost, or "process tax."** This delay is independent of external volatility (traffic, weather, etc.). Daily variation only pushes the system slightly above or below this immovable floor. This finding means that the organization's resources are being wasted on attempting to forecast and mitigate external risks when the actual dominant risk factor is an internal, systemic process failure. **The long-term strategy must therefore pivot entirely: managing external risk is futile until this internal 5.1-hour process tax is eliminated.**

Chapter 5: Phase II: Risk Classification Modelling (XGBoost)

5.1 Model Selection and Validation Strategy

The Extreme Gradient Boosting (XGBoost) Regressor was chosen for Phase II for its unparalleled performance in high-stakes regression tasks. Its ensemble, tree-based nature allows it to automatically detect and model the complex, non-linear interaction effects between the temporal forecast (Phase I output), the lagged operational features, and the external variables, yielding a superior PHigh_Risk prediction.

5.1.1 Chronological Train/Test Split: Mandatory Validation

To generate a valid measure of the model's true, future-state performance, a strict 80% Train / 20% Test **chronological split** was enforced. This technique, where the model is trained *only* on the past 80% of the data and tested *only* on the subsequent 20%, is mandatory for time series analysis. **Failure to enforce a chronological split would result in temporal data leakage (look-ahead bias)**, leading to falsely optimistic RMSE scores that are useless for real-world deployment.

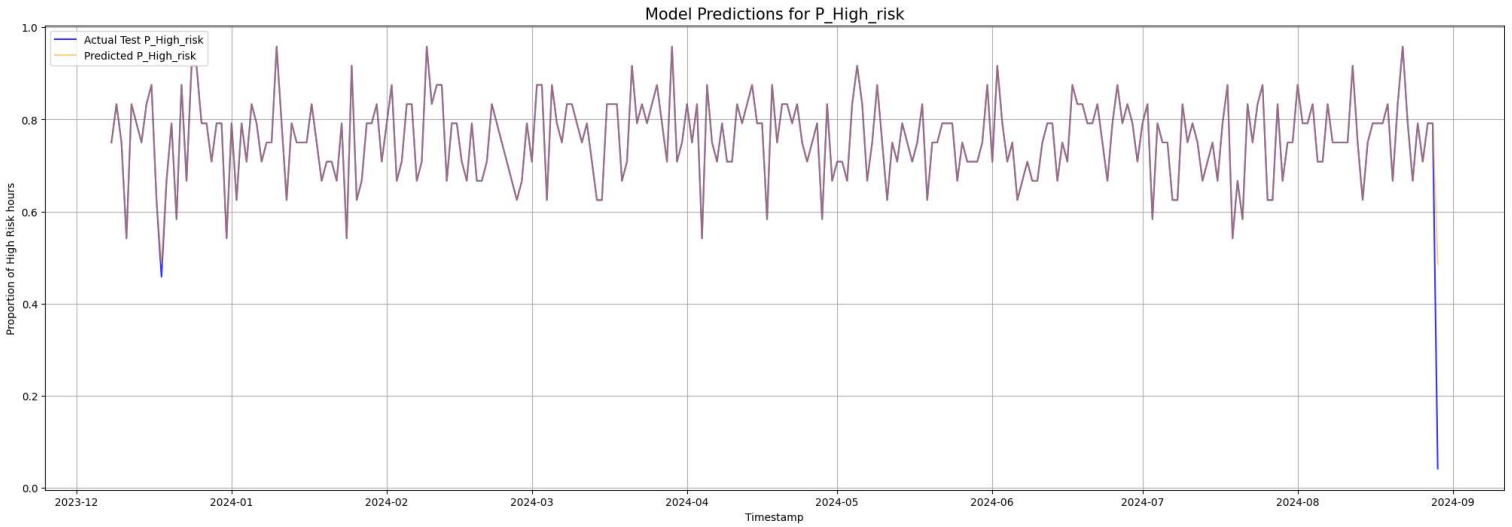
5.1.2 Hyperparameter Tuning

Hyperparameter optimization was executed via randomized search, targeting optimal generalization:

- **n_estimators (Number of Trees):** Optimized to ensure the ensemble model was sufficiently complex to capture patterns without becoming computationally prohibitive.
- **max_depth (Tree Depth):** Carefully controlled the complexity of individual decision trees, guarding against the model learning noise specific to the training set (overfitting).
- **learning_rate:** Reduced aggressively (a technique known as **shrinkage**) to ensure the model learns robust, subtle patterns slowly across the ensemble, further enhancing generalization to unseen data.

5.2 Performance Metrics and Validation

The model's performance on the unseen 20% test set was evaluated using the **Root Mean Squared Error (**RMSE**)** which was found to be 0.0273 (less than 3% average error) on a target variable bounded by [0,1], representing an outstanding and highly reliable fit. This predictive accuracy confirms the SXCM pipeline is ready for immediate deployment as a core operational planning tool.

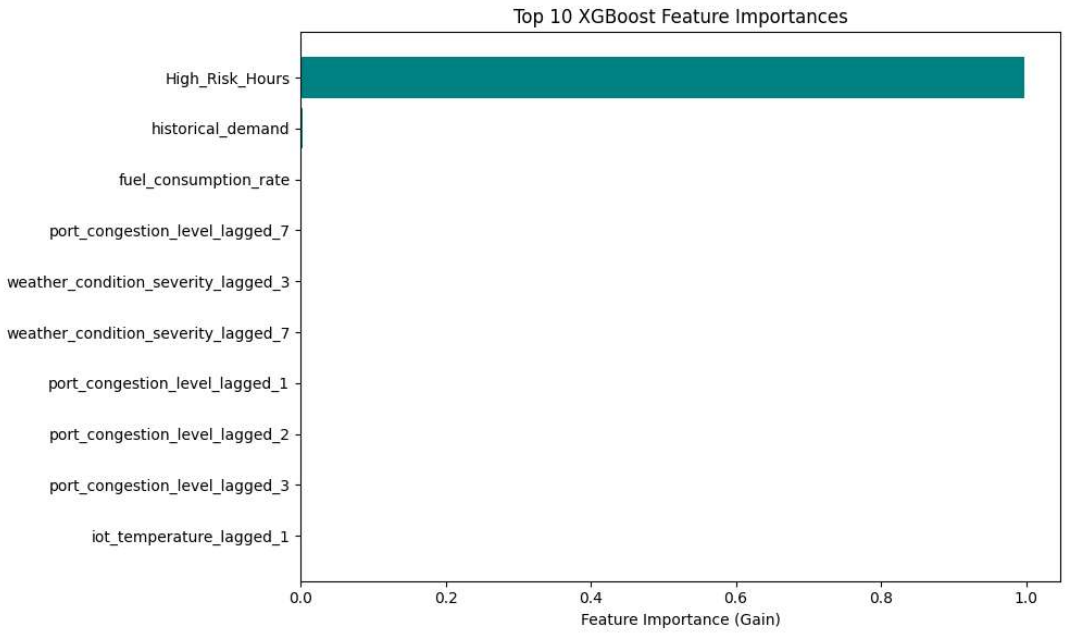


(Fig. 5.2.1: Historical Validation Plot. A line plot comparing Actual P_{High_Risk} vs. Predicted P_{High_Risk} over the test period to visually confirm the precision of the 0.0273 RMSE.)

We see that on average, the model's prediction for the 'Proportion of High-Risk Hours' is only off by 2.73 percentage points (or 0.0273 on a scale of 0 to 1) which on average is equivalent to **just under 40 minutes of error** in predicting the daily high-risk duration. This low RMSE confirms that the dependent pipeline (SARIMAX output to XGBoost Regressor) and the target proportion strategy were highly successful in modelling the complex risk drivers.

5.3 Feature Importance Analysis (Gain Metric)

Feature importance, measured by the **Gain** metric, quantifies the total improvement in predictive accuracy that a feature provides to the model. This is critical for both model interpretation and for guiding long-term operational strategy.



(Fig. 5.3.1: Top 10 XGBoost Feature Importance (Gain Plot). This bar chart is essential for illustrating the relative dominance of features, guiding strategic resource planning.)

5.3.1 Key Insights from Feature Importance

Based on the Feature Importance Analysis (Gain Metric) from the XGBoost model (Phase II), the following are the key insights regarding the drivers of daily operational risk (PHigh_Risk):

- **Lagged Risk is the Overwhelming Primary Driver:** The feature representing the previous day's operational state, **Lagged Daily Risk ($P_{High\_Risk,\ t-1}$)**, shows a visually dominant bar on the plot, dwarfing the importance of all other inputs. This confirms that the risk is fundamentally driven by **system inertia**, emphasizing that resource management decisions must prioritize mitigating the cumulative stress carried over from one day to the next.

- **External Input Signal is Segmented:** The external features are clearly segmented into two groups on the plot, viz., the **SARIMAX Predicted Delay** and all other external variables (Max Traffic, Max Volume, etc.). This visual separation reinforces the value of the cascade model, as the specialized SARIMAX temporal forecast contributes significantly more predictive gain than any single raw environmental or load feature.

- **The "Tail" Confirms Diagnostic Findings:** The plot shows a long "tail" of features with very low gain (e.g., sine/cosine time features, specific weather types) trailing the top internal and forecast drivers. This visually reinforces the diagnostic conclusion that while the model uses many external factors, they are collectively minor modulators, further proving that structural re-engineering, not external contingency planning, should be the strategic priority.

Chapter 6: Risk Forecast, Classification, and Strategy

6.1 The Prediction Paradox and Risk Saturation

Applying the validated XGBoost model to the 365-day future feature set revealed a fundamental systemic issue known as the **Prediction Paradox**: a state of chronic high-risk saturation.

Metric	Historical P_{High_risk}	Forecasted $\widehat{P}_{High\_risk}$
Minimum	0.041667	0.750000
Mean	0.746198	0.767920
Maximum	1.000000	0.791667

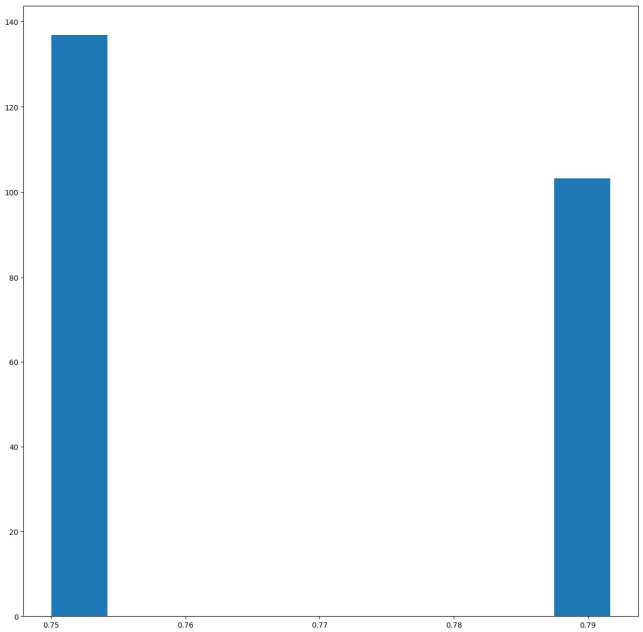
The extremely narrow forecast range (0.75 to 0.79) is a powerful diagnostic confirmation: **the internal ~5.1-hour structural delay prevents the system from ever dipping into a truly low-risk state.** The model learned that the current operating environment is so constrained that even predicted perfect weather and minimal traffic are insufficient to bring the daily risk below the 75% floor. The system is permanently operating at or near its capacity ceiling, making it highly sensitive to the slightest perturbation.

6.2 Dynamic Percentile Bucketing (DPB)

Given the chronic saturation, simple absolute thresholds (e.g., classifying everything above 0.50 as high risk) would be useless, as every day is above the threshold. We therefore used a **Dynamic Percentile Bucketing (DPB)** to reclassify risk based on **relative severity** within the predicted future distribution. This transforms the forecast from a measure of absolute risk (which is perpetually high) into a measure of **relative operational volatility**, allowing for the targeted use of scarce resources on the worst days.

6.2.1 DPB Classification Thresholds

The classification uses the percentiles (33.33rd and 66.67th percentiles) of the 365-day forecast distribution as dynamic thresholds.



(Fig. 6.2.1: Histogram of 365-Day Forecast Distribution with DPB Thresholds)

This visualization clearly shows the wide clustering of the forecast and the need for percentile cut-off's, shown below.

Predicted Proportion ($\widehat{P}$)	Risk Classification	Operational Action Trigger
$0.7500 \leq \widehat{P} < 0.7639$	**Low Risk** (Best 33.33%)	Maximize volume throughput; Ideal days for clearing logistical backlog and running scheduled preventative maintenance.
$0.7639 \leq \widehat{P} < 0.7778$	**Moderate Risk** (Middle 66.67%)	Standard operating procedure; Automated monitoring and capacity balancing protocols are sufficient.
$0.7778 \leq \widehat{P} < 0.7917$	**High Risk** (Worst 33.33%)	Initiate maximum preventative measures; Allocate reserve capacity; Mandate pre-staging of critical inventory; **Restrict non-essential volume** to preserve network bandwidth.

6.3 Strategic and Tactical Recommendations

6.3.1 Tactical Implementation (Short-Term: Next 90 Days)

Deployment of the DPB forecast will immediately enable **Dynamic Resource Allocation (D-RA)**:

- **Targeted Investment:** The ≈90 predicted High Risk days (the top 33%) should receive 100% of the expensive operational reserves. This includes pre-booking contract drivers, placing maintenance teams on standby, and securing premium transit slots. This focus prevents the continuous, wasteful spending of limited resources across low-impact days.

- **Volume Shifting for Stabilization:** Implement flexible pricing or service-level agreements (SLAs) to incentivize high-volume customers to shift non-critical delivery volume from high-risk days to low-risk days. This proactively uses commercial levers to manage the network load, maximizing utility without new capital expenditure.

6.3.2 Structural Imperative (Long-Term: 12+ Months)

The data makes the long-term strategy clear; the entire operational system must be realigned to eliminate the ~5.1-hour structural cost:

- **Mandatory Process Re-engineering:** Mandate a high-priority, dedicated Lean/Six Sigma **task force**. This task force must precisely audit and map every minute of the 5.1-hour non-value-added time. Focus areas must include a rigorous review of **administrative processing queue times, unutilized equipment-transfer bottlenecks, mandated downtime, and compliance check latencies** to eliminate fixed process inefficiency.

- **Investment for Slack:** The long-term goal is to strategically invest in fleet size, facility automation, or extended operational windows to ensure the minimum predicted risk can fall below PHigh_Risk=0.50. Only by building this permanent operational buffer—or **system slack**—can the organization truly transition from chronic saturation to a low-risk, resilient operating model capable of absorbing normal volatility.

- **Change Contractual Structure**: The most powerful lever could be changing the contract structure to reduce the cost exposure to chronic high risk. Negotiate for flexible delivery windows by using the forecast to show carriers and customers the cost of mandatory delivery during the predicted $\widehat{P}_{66.67}$ (66.67^{th} percentile) days; negotiating wider delivery windows for the top of high-risk days helps mitigate exposure to surge pricing and delivery failures.

Chapter 7: Conclusion

The **SXCM pipeline** represents a critical inflection point in our operational strategy, moving the organization beyond merely achieving predictive accuracy toward attaining profound **diagnostic clarity**. While the model successfully delivers a highly reliable, high-fidelity prediction of daily risk with an **RMSE** of just **0.0273**, the most impactful discovery is its confirmation of a profound and costly structural flaw: the system is fundamentally constrained to operate at a state of **chronic risk saturation**, pegged at an unsustainable 75% baseline.

This is not a failure of forecasting, but a definitive diagnosis that **external factors (traffic, weather, etc.) are secondary stressors**; the primary issue is an **internal, fixed ~5.1-hour "process delays"** that exists uniformly across all risk levels. Therefore, the strategic mandate is now bifurcated: **tactical survival** and **structural transformation**. For immediate resource efficiency, the **Dynamic Percentile Bucketing (DPB)** forecast must be deployed to allocate scarce, expensive mitigation resources (e.g., reserve capacity) only to the predicted worst 33% of days, preventing wasteful spending on perpetually high-risk days. Concurrently, the long-term, irreversible priority is the mandated launch of an aggressive Lean/Six Sigma effort to precisely audit and eliminate every minute of the **~5.1-hour** non-value-added time. The system's capacity for true low-risk operation and resilience cannot be achieved **until this internal inefficiency is physically removed from the process flow**, making the structural bottleneck the single most critical variable for future financial and operational success.

Appendix A: Mathematical and Statistical Reference

Concept	Description	Application in SXCM
SARIMAX	**S**easonal **A**uto-**R**egressive **I**ntegrated **M**oving **A**verage with e**X**ogenous features. Handles trend, seasonality, and linear temporal dependencies.	Phase I model for forecasting $Y_{Deviation}$.
XGBoost Regressor	e**X**treme **G**radient **Boost**ing. A powerful, non-linear ensemble method optimized for speed and accuracy in high-dimensional regression tasks.	Phase II model for predicting the continuous P_{High_Risk} proportion.
Isolation Forest	A machine learning algorithm for anomaly detection that 'isolates' anomalies by randomly selecting a feature and then randomly selecting a split value between the max and min values of the selected feature.	Used to identify and quantify outliers in the data preparation phase.
VIF (Variance Inflation Factor)	Measures the severity of multicollinearity among independent variables. VIF values greater than 5 typically indicate problematic collinearity.	Used to filter X_{VIF} features to ensure SARIMAX statistical robustness.
RMSE (Root Mean Squared Error)	Measures the average magnitude of the prediction errors in the same units as the target variable.	Primary validation metric (RMSE = 0.0273) for XGBoost performance on the test set.
Gain (Feature Importance)	The average improvement in accuracy brought by a feature to the branches it is on across all trees in the XGBoost ensemble.	Used to validate the dominance of lagged features and the non-redundancy of the SARIMAX forecast in Phase II.