

Chapter 2

LITERATURE REVIEW

2.1 Problem Description

Accurate forecasting of incoming customer orders is essential for aligning production planning and procurement in manufacturing environments, particularly those characterized by a high variety of SKUs, seasonal demand, and fluctuating order volumes. In many industrial settings, such as heavy-duty filter manufacturing, forecasting remains a complex challenge due to irregular purchasing cycles, inconsistent data patterns, and limited visibility into downstream consumption.

Traditional forecasting approaches, such as moving averages and exponential smoothing, often fail to capture the nonlinear and volatile nature of real-world customer demand. These methods typically rely on the assumption of stable patterns and linear trends, which rarely hold in manufacturing sectors exposed to dynamic external factors, such as macroeconomic shifts, customer-specific ordering behavior, or religious calendar cycles. As a result, many manufacturers struggle to generate reliable forecasts from historical order data alone.

In addition, organizational challenges, such as fragmented planning processes in sales, production, and supply chain functions, compound the difficulty of turning forecasts into actionable plans. Although demand uncertainty has long been recognized as a planning risk, its link to data modeling and feature representation in forecasting systems remains underdeveloped in both theory and practice.

This chapter presents a comprehensive review of the literature to establish the theoretical and empirical foundation for this study. It serves the following objectives:

1. To examine prior research on the use of machine learning for forecasting customer orders in manufacturing contexts, especially where demand patterns are irregular or seasonal;
2. To identify the role of time-based, seasonal, and exogenous features in enhancing forecast accuracy;
3. To extract both methodological and practical insights that inform the development, selection, and evaluation of forecasting models using real-world customer order data.

The literature review is structured around five thematic areas relevant to the scope of the investigation.

- **Demand Forecasting in Manufacturing:** Reviewing statistical and ML-based approaches for order prediction in production systems with complex SKU dynamics;
- **Machine Learning for Forecasting:** Exploring the comparative strengths of tree-based and deep learning models in handling nonlinear demand;
- **Seasonality and Exogenous Variables:** Analyzing how prior studies have integrated time-based and external factors into forecasting models;
- **Feature Engineering and Model Evaluation:** Investigating the importance of feature design and selection in determining forecast performance;
- **Research Gaps in Forecasting Practice:** Identifying shortcomings in the existing literature, particularly the limited use of real operational data in feature-informed forecasting frameworks.

Through this synthesis of the literature, the study aims to bridge the gap between academic modeling techniques and the practical forecasting needs of manufacturers. The insights gained will guide the experimental framework used to evaluate the effectiveness of machine learning models in predicting future volumes of customer orders.

2.2 Organizational and Problem Context

The company under study operates within the manufacturing sector, specifically producing heavy-duty vehicle filter products. Its organizational structure includes functional departments such as sales, planning, production, and warehouse operations. These departments collaborate to meet customer demand while managing production capacity and fulfillment constraints.

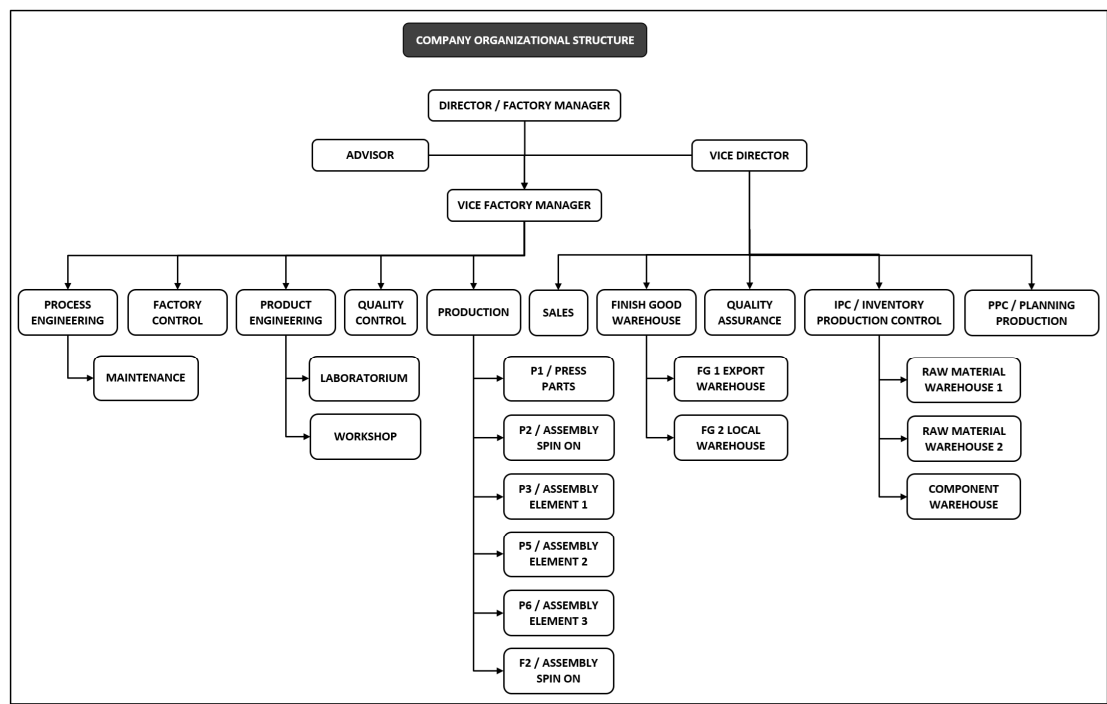


Figure 2.1: Company Organizational Structure

The organizational structure of the company illustrated in 2.1 represents a hierarchical framework designed to efficiently manage manufacturing operations by clearly defining the responsibilities and reporting lines. At the top, the Director/Factory Manager oversees overall operations, supported by an Advisor, Vice Director, and Vice Factory Manager. The structure is divided into key functional areas: Process Engineering (including maintenance), Factory Control, Product Engineering (with Laboratory and Workshop support), Quality Control, Production (covering various production lines such as Press Parts and Assembly units), Sales, Finished Goods Warehouse (separated into export and local warehouses), Quality Assurance, Inventory Production Control (managing raw material and component warehouses), and Planning/Production Control. Each function ensures smooth coordination between engineering, production, quality, sales, and logistics, while also maintaining strong control over inventory and planning. This organizational setup highlights a balance between operational execution, quality assurance, and supply chain integration, in order to enhance productivity, maintain high product standards, and ensure timely delivery to local and international customers.

2.2.1 Functional Roles

Five departments play a critical role in shaping the visibility of customer demand and the flow of order-related information.

1. **Sales Department:** Manages customer interactions, receives and processes sales orders, and communicates demand signals to planning and production teams.
2. **Planning and Production Control (PPC):** Aligns production schedules with expected customer orders. PPC relies heavily on forecasts to allocate capacity and plan procurement.
3. **Inventory and Procurement Coordination (IPC):** Supports material readiness based on projected demand, but is not the main focus of this study.
4. **Production Department:** Executes manufacturing processes based on production plans and provides downstream data on order fulfillment timelines.
5. **Warehouse Management:** Handles inbound and outbound materials. Although warehouse data are not modeled directly, it supports order tracking and historical data compilation.

These departments collectively contribute to the flow of operational and demand data used in forecasting customer orders.

2.2.2 Operational Challenges Across Functional Areas

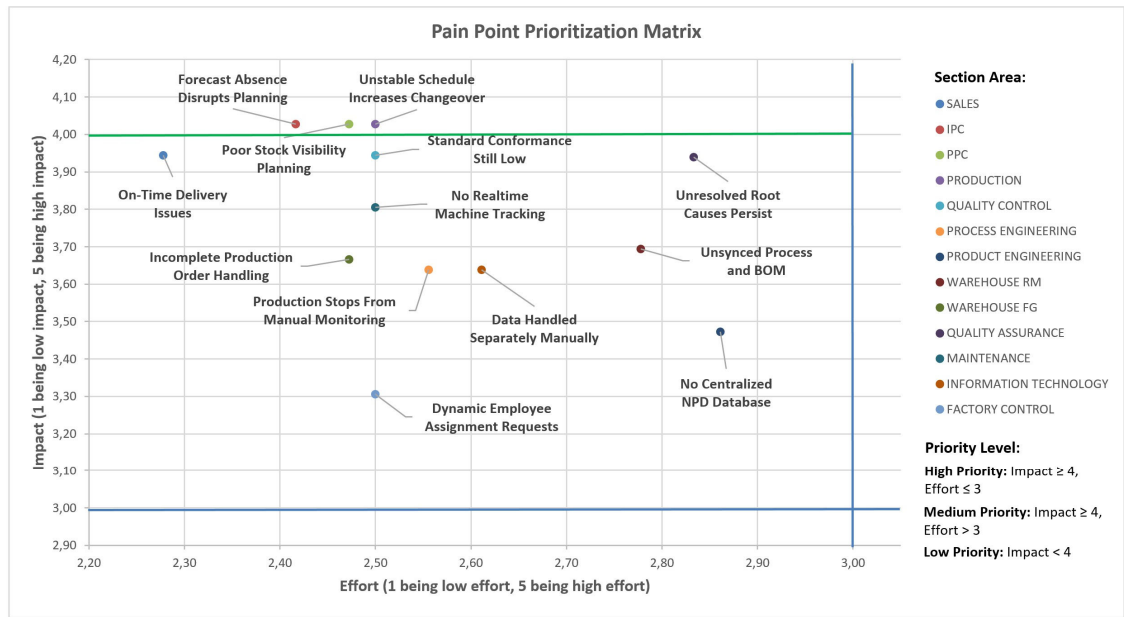


Figure 2.2: Pain-Point Prioritization Matrix

Figure 2.2 illustrates the prioritization of operational pain points identified in various functional departments in PT. Heavy Duty Vehicle Filter Manufacturer. The matrix reflects the relative severity and frequency of recurring problems that hinder the organization’s ability to anticipate and respond to customer demand effectively. These pain points are not isolated to one department but are distributed throughout the operational spectrum, from sales to warehouse operations, highlighting the cross-functional nature of the forecast challenge.

Table 2.1: Key Operational Challenges Across Functional Areas

Functional Area	Key Operational Challenges
Sales	Volatile and irregular customer order patterns; manual data entry errors
PPC	Forecast dependency for production scheduling; frequent rescheduling due to demand uncertainty
IPC	Limited upstream visibility into customer order trends
Production	Reactive to changes in short-term forecasts; difficulty aligning output with demand cycles
Warehouse	Historical shipment data not fully integrated into forecasting logic

Table 2.1 complements the matrix by providing a textual summary of the key challenges observed within each functional area. The sales department, for example, faces unpredictable customer ordering behavior and frequent manual errors, which directly affect the quality of input data for forecasting. In the planning and production control (PPC) unit, scheduling remains heavily reliant on forecasts but suffers from frequent adjustments due to fluctuating demand signals. Similar constraints are seen in IPC and production, where data fragmentation and reactive planning cycles further degrade forecast reliability. The warehouse function, while not central to forecasting itself, holds valuable historical shipment records that are not yet leveraged in the modeling process.

The combined message of the figure and the table is that forecast failures are not only technical in nature, but are also rooted in organizational fragmentation, inconsistent data practices, and lack of integrated planning systems. Together, these issues create a forecasting environment that is reactive, error prone, and poorly aligned with the actual demand cycle.

This context reinforces the need for a forecasting approach that can function independently of real-time integration, while still extracting meaningful patterns from historical customer order data. The machine learning models proposed in this study are designed with that constraint in mind: they operate on a structured order history and engineered features, rather than relying on ideal data integration or perfect system coordination.

These challenges illustrate how demand variability, manual data handling, and poor integration between departments lead to unreliable visibility into future customer orders. Inconsistent forecasting practices and limited use of analytical tools result in reactive decision-making and misalignment between expected and actual demand.

2.2.3 Relationship to Forecasting Challenges

Forecast errors in this environment are primarily caused by:

- **Irregular Order Timing:** Customers place orders unpredictably, often with short notice or variable quantities.
- **Limited Historical Feature Utilization:** Forecasts often neglect seasonality (e.g. spikes of Idul Fitri) or external signals (e.g. exchange rate fluctuations).
- **Fragmented Data Sources:** Sales, planning, and warehouse data are not integrated into a unified forecasting pipeline.

- **Manual Forecast Generation:** Forecasts are generated based on heuristics rather than predictive modeling, reducing reliability.
- **Lack of Interpretability in Planning Decisions:** Forecast drivers (features) are not explicitly analyzed or tracked.

These factors highlight the need for a machine learning-based forecasting approach that leverages structured feature inputs to improve the accuracy and consistency of customer order predictions. The focus of this study is not on inventory optimization or OCR implementation, but on evaluating the predictive performance of different ML models in capturing historical customer order behavior.

2.3 Characteristics of Inventory and Forecasting in Manufacturing Industries: Case of PT. Heavy Duty Vehicle Filter Manufacture

Demand forecasting in manufacturing environments such as PT. The heavy duty vehicle filter manufacturer is characterized by high product variety, irregular ordering patterns, and the presence of strong seasonal effects. Although an ERP system is implemented, planning and forecasting processes remain challenged by fragmented information flows and limited analytical tools. The following characteristics describe the demand-side complexities that motivate the need for a more data-driven forecasting approach:

1. **Demand Seasonality and Event-Based Spikes:** Demand for certain SKUs surges around key calendar events such as Idul Fitri, Chinese New Year, and fiscal year-end. These patterns are recurring, but not always consistent in magnitude, requiring forecasting models that incorporate seasonal and event-based features.
2. **High SKU Diversity:** The company manages a wide range of filter products, including both high-frequency standard items and low-frequency custom products. Each SKU type exhibits distinct demand behaviors, increasing the complexity of forecasting at the item level.
3. **Customer Order Irregularity:** Orders are often placed in uneven intervals and in varying quantities, especially for special order items. This results in volatile historical demand data that is poorly modeled by traditional time series techniques.

4. **Limited Use of Historical Features:** While historical sales order data exist, prior forecasting efforts have relied on manual trend estimation, with limited integration of temporal features (e.g., month, week, holiday flag) or exogenous variables (e.g., exchange rates).
5. **Siloed Functional Operations:** Sales, planning, and production departments operate with low integration, leading to delayed or incomplete transmission of demand signals. As a result, forecasts are often reactive and manually generated.
6. **Absence of Automated Forecast Models:** Forecasts are not generated using predictive modeling techniques. There is no standardized use of statistical or machine learning models to generate forward-looking order predictions.

These operational characteristics highlight the need for machine learning-based forecasting tools that can learn from historical customer order behavior, model non-linear trends, and incorporate multiple contextual features. The goal is to generate more reliable forecasts of future customer orders, enabling more effective upstream planning without requiring downstream consumption data.

Table 2.2: Demand Forecasting Characteristics at PT. Heavy Duty Vehicle Filter Manufacturer

No.	Characteristic	Description	Relevance to ML Forecasting
1	Demand Seasonality	Recurring spikes during Idul Fitri, Chinese New Year, and fiscal year-end.	Requires seasonal flags, holiday indicators, and event-based calendar features.
2	SKU Diversity	Standard high-volume and low-frequency custom SKUs coexist.	Justifies item-level forecasting with SKU-specific patterns.
3	Irregular Ordering Behavior	Customer orders vary in frequency and volume.	Requires nonlinear models to handle noisy or sparse data.
4	Underutilized Historical Features	Forecasting lacks temporal or external input features.	Highlights the benefit of feature engineering for accuracy.
5	Functional Fragmentation	Sales and planning are poorly integrated.	Forecasting must rely on structured historical order data only.
6	No Predictive Forecasting Tools	Forecasts are generated manually.	Supports adoption of ML models for automated prediction.

Table 2.2 summarizes the key characteristics of the demand environment faced by PT. Heavy Duty Vehicle Filter Manufacturer. Unlike consumer-oriented sectors with relatively stable and aggregated demand, this industrial manufacturer deals with fragmented item-level customer orders that are subject to irregular timing and strong seasonal influence. Each characteristic represents a forecasting challenge that cannot be effectively addressed by manual estimation or simple statistical models.

The message conveyed by this table is two-fold. First, it illustrates the complexity of forecasting incoming customer orders in a real-world manufacturing setting where data

availability is constrained and demand patterns are volatile. Second, it emphasizes the relevance and necessity of applying machine learning models capable of learning from both structured historical data and engineered features, such as calendar variables and product-level segmentation.

These characteristics form the operational rationale for the methodological focus of this study: building forecasting models that go beyond univariate time series methods and integrate multiple features that reflect the temporal, behavioral, and contextual drivers of demand.

2.4 Feature Engineering and Feature Selection in Demand Forecasting

In machine learning-based demand forecasting, the structure, relevance, and quality of input features play a central role in determining model accuracy and generalizability. This section outlines two key data preparation steps **feature engineering** and **feature selection** that are essential to build effective forecasting models, particularly in manufacturing environments with volatile and seasonal order patterns.

Feature engineering refers to the process of transforming raw historical data into structured variables that expose meaningful temporal, categorical, or contextual signals. In the context of forecasting incoming customer orders, the most relevant engineered features include:

- *Temporal variables*: such as month, quarter and year, which help the models capture time-based fluctuations.
- *Seasonal indicators*: flags for events such as Idul Fitri or fiscal year-end that regularly influence demand patterns.
- *Lagged variables*: recent order volumes (for example, last 1 to 3 months) that capture momentum or trend effects.
- *Rolling statistics*: moving averages or demand variability over past periods to smooth noise and expose patterns.
- *Categorical encodings*: such as sales channel, region, or customer group.
- *Exogenous features*: including macroeconomic indicators such as exchange rates or commodity prices.

These features allow forecasting models to recognize both short-term fluctuations and long-term cycles, enabling better anticipation of demand dynamics in a complex SKU environment.

Feature selection involves identifying which of the engineered features provide the most predictive value while eliminating redundant or noisy input. This improves both the efficiency and interpretability of machine learning models. Common methods include:

- *Filter-based techniques:* e.g., correlation analysis, mutual information ranking.
- *Wrapper-based methods:* such as recursive feature elimination (RFE).
- *Embedded techniques:* which use built-in importance metrics from algorithms like Random Forest or LightGBM.

These methods ensure that only relevant variables are retained, reducing dimensionality and minimizing the risk of overfitting, especially when working with high-cardinality categorical encodings or multi-feature time series data.

Table 2.3: Summary of Feature Engineering and Selection in Forecasting Customer Orders

Step	Description	Examples in Order Forecasting
Feature Engineering	Creating structured variables to reveal demand patterns.	Month, quarter, seasonal flags (e.g., Idul Fitri), lagged order volume, moving average, sales region, channel, exchange rate.
Feature Selection	Choosing predictive variables and removing irrelevant ones.	Correlation filtering, permutation importance, RFE, model-based feature scoring (e.g., Random Forest).
Expected Outcomes	Improved accuracy, less overfitting, better model interpretability.	Higher R^2 , lower MAE/MAPE, more robust generalization.

Table 2.3 summarizes the key components and benefits of feature engineering and selection in the context of forecasting incoming customer orders in a manufacturing setting. These steps form the methodological foundation for the modeling approach adopted in this study.

2.5 Conceptual Description

This section outlines the core theoretical concepts that underpin this research. The goal of this study is to forecast incoming customer orders using machine learning techniques, particularly in the context of irregular and seasonal demand patterns in a manufacturing environment. Five interrelated concepts form the conceptual foundation: (1) machine learning in time series forecasting, (2) demand seasonality, (3) feature engineering, (4) feature selection, and (5) model performance evaluation.

2.5.1 Machine Learning for Demand Forecasting

Machine learning (ML) provides a data-driven and flexible approach to demand forecasting, capable of learning from complex, non-linear relationships in historical data. Unlike traditional statistical models that rely on fixed rules or linear assumptions, ML models can adapt to changing patterns and incorporate diverse features.

Tree-based models such as Random Forest and LightGBM have been widely used in time series forecasting tasks due to their ability to handle high-dimensional data, rank feature importance, and model interactions. In this study, these models are applied to forecast monthly order quantities from customers at the SKU level.

2.5.2 Demand Seasonality

Manufacturing demand is often influenced by recurring patterns tied to the calendar. In the case of heavy-duty filters, demand frequently spikes around major religious or fiscal events such as Idul Fitri, Chinese New Year, and year-end procurement periods.

Capturing this seasonality through engineered features, such as holiday flags, month, quarter, or fiscal cycle indicators, allows ML models to differentiate between cyclical fluctuations and random noise, thus improving forecast accuracy.

2.5.3 Feature Engineering

Feature engineering is the process of transforming raw transactional data into structured variables that better reflect the underlying drivers of demand. It enables the model to interpret patterns over time and across product or customer dimensions.

Examples of engineered features used in this study include:

- *Temporal features*: month, quarter, and holiday season flags
- *Lag variables*: sales from prior months (e.g., $t - 1$, $t - 2$, $t - 3$)
- *Rolling averages*: moving average of past demand to smooth out noise
- *Categorical encodings*: region, channel, item type
- *Exogenous indicators*: exchange rates, market indices

These features help the model recognize seasonality, customer behavior, and market volatility, all of which influence order patterns.

2.5.4 Feature Selection

Feature selection involves identifying the most relevant features for model training while excluding redundant or irrelevant variables. Improve model efficiency, reduce overfitting, and enhance interpretability.

The methods used in this study include:

- Correlation analysis to eliminate highly collinear variables
- Permutation importance from Random Forest
- Recursive feature elimination (RFE)

These methods ensure that the forecasting model is driven by inputs that truly contribute to accuracy, without excessive complexity.

2.5.5 Forecast Evaluation

To ensure that machine learning models produce useful and reliable forecasts, performance evaluation is a key component of the framework. Standard metrics such as Mean Absolute Error (MAE), Mean Absolute Percentage Error (MAPE), Root Mean Square Error (RMSE), and R^2 Score are used to quantify the prediction quality.

These metrics allow for comparative analysis between different models and feature configurations, ensuring that the final model is accurate and generalizable.

2.6 Forecasting Methods and Frameworks Used in Previous Studies

In addressing the challenges of forecasting customer demand in manufacturing supply chains, recent studies have increasingly adopted machine learning approaches, particularly tree-based models such as Random Forest and LightGBM. These models have shown strong performance in capturing non-linear demand behavior, incorporating complex feature interactions, and handling structured historical data.

Unlike traditional time series models that assume stationarity and linear trends, machine learning-based forecasting enables greater flexibility in dealing with seasonal variability, irregular customer order patterns, and high SKU complexity. Studies such as Miguel Rodrigues (2024) and Feizabadi (2022) demonstrate that models incorporating engineered time-based characteristics (e.g., month, quarter, holiday indicators) and categorical variables (e.g., product type, sales region) can significantly improve forecast accuracy compared to statistical baselines.

In addition, feature selection techniques have been used, such as the importance of permutation or recursive feature elimination, to streamline input dimensionality and enhance model interpretability. These approaches reduce overfitting and increase model efficiency without sacrificing predictive performance. Several studies report measurable gains in forecast precision by applying these methods in B2B and production-oriented contexts.

Together, the reviewed literature highlights a methodological shift from traditional statistical forecasting to feature-enriched machine learning-based models. This shift supports the rationale for the focus of the current study on evaluating the effectiveness of Random Forest and LightGBM, as well as the role of feature engineering and selection in improving the accuracy of SKU-level demand forecasting in a manufacturing environment.

Table 2.4: Comparison of Machine Learning-Based Forecasting Studies, Miguel Rodrigues (2024) and Feizabadi (2022).

Title	Model	Features Used	Evaluation Metrics	Key Findings
ML-Based SKU Forecasting in Retail	Random Forest, XGBoost	Lag, calendar, holiday, sales channel	MAE, RMSE, R ²	Tree-based models outperform linear regression in volatile demand.
Random Forest Forecasting for Automotive Parts	Random Forest	Lag, moving average, month, SKU type	MAPE, MAE	Random Forest improved accuracy over moving average by 30%.
LightGBM for Demand Prediction in Manufacturing	LightGBM	Calendar variables, customer segment, economic indicators	MAE, MAPE, R ²	Feature-rich LightGBM achieved highest accuracy across months.
Feature Engineering Impact on Forecast Accuracy	Random Forest, LightGBM	Lag features, rolling stats, categorical encodings	MAE, MAPE, RMSE, R ²	FE and FS significantly improve performance; RFE reduces overfitting.

As shown in Table 2.4, Random Forest and LightGBM consistently demonstrate strong predictive performance in various operational settings, especially when enriched with well-designed temporal, categorical, and lag-based features. The reviewed studies highlight that forecast accuracy improves significantly when proper feature engineering and selection are applied, even when deep learning models are not used.

In addition, the inclusion of multiple evaluation metrics, such as MAE, MAPE, RMSE, and R2, allows a robust assessment of model performance. These findings support the methodological choices made in this thesis, which adopts both Random Forest and LightGBM as primary forecasting models. The next chapter builds upon these insights by applying these models to real-world manufacturing order data, comparing their performance before and after the application of feature engineering and selection.

2.6.1 Research Gap

Although existing studies have successfully demonstrated the value of tree-based forecasting models, few have explicitly compared model performance across different modeling stages (e.g., baseline, post-feature engineering, and post-feature selection). In addition, limited research focuses on evaluating structured order data in a B2B manufacturing setting where seasonal factors and SKU-level demand variability are prominent. This gap

motivates the present study to systematically investigate the incremental contributions of engineered and selected features to forecast performance in a real-world operational context.

