

Machine Learning Models for Anticipatory Production Capacity Planning: The Case of the ASML Field Service Hub

Ruben Siemerink, Report Number: 2024.MME.8996

October 24, 2024

Abstract

Efficient production capacity planning is essential for industries facing uncertain demand, such as semiconductor manufacturing. This thesis explores machine learning models to improve demand forecasting and optimize production capacity, focusing on the ASML Field Service Hub. Models including XGBoost, Random Forest, and ARIMA are evaluated to predict demand across multiple product flows under uncertainty. These forecasts are integrated into an optimization model to minimize costs related to overcapacity, undercapacity, and labor fluctuations.

The proposed framework demonstrates substantial improvements over ASML's current benchmarks, reducing forecast errors by up to 37.4% and simulated costs by 48.3%. Additionally, robustness testing was conducted using jitter tests, ensuring the model's stability in the face of noisy data. The explainability and feature importance of the machine learning models were investigated using SHAP values. The study's results highlight the model's adaptability, offering a valuable tool for improving capacity planning in various industries dealing with demand volatility.

Contents

1	Introduction	2
2	Literature Review	3
2.1	Introduction	3
2.1.1	Background	3
2.2	Classification	4
2.3	Forecasting Techniques	5
2.3.1	Time-series Based Forecasting: Mathematical	6
2.3.2	Time-series Based Forecasting: Artificial Intelligence	6
2.3.3	Engineering: Reliability-Based Forecasting	8
2.4	Integrating the Forecast in the Supply Chain	8
2.4.1	What to Optimize?	8
2.4.2	Optimization Methods	10
2.5	Research Gaps to Pursue	10
3	Problem Statement	12
3.1	Research Questions	12
3.2	Methodology	13
4	Research Question 1: What methods are best suited to forecast demand under uncertainty?	15
4.1	Introduction	15
4.2	Classical Statistical Forecasting Methods	15
4.2.1	Time-Series Models	15
4.2.2	Intermittent Demand Forecasting	15
4.3	Machine Learning Approaches to Forecasting	16
4.3.1	Decision Tree-Based and Gradient Boosting Models	16
4.3.2	Neural Networks and Deep Learning	16
4.4	Approach for the Forecasting Model	16
4.4.1	Requirements for the Forecasting Model	17
4.4.2	Model Selection Process	17
4.4.3	Final Model Selection	17
4.5	Explainability of Machine Learning Models	18
4.5.1	Application of SHAP Values to Explain Model Behavior	18
4.6	Conclusion	19
5	Research Question 2: How should forecasting models be designed for optimal capacity planning?	20
5.1	Model Design: Selecting Features and Variables	20
5.2	Data for Forecasting Models	20
5.2.1	Aggregation Level	21
5.3	Timeframe	21
5.4	Multi-Step Forecasting: Multi-Output Regression and Chain Regression	21
5.4.1	Multi-Output Regression	22
5.4.2	Chain Regression	22
5.5	Robustness Testing and Model Evaluation	23
5.5.1	Jitter Test	23

6	Research Question 3: How can forecasts be translated into production capacity decisions?	24
6.1	Capacity Planning Based on Forecasts	24
6.2	Simulation of Production Capacity Strategies	25
6.2.1	Production Capacity Strategies Based on Forecasts	25
6.3	Decision-Making Framework: Linking Forecasts to Action	25
6.4	Validation: Cost Simulation of Production Capacity Strategies	26
6.4.1	Simulation Setup: Rolling Window Forecasting and Simulation	26
6.5	Mathematical Notation: Capacity Planning and Optimization Problem	26
7	Research Question 4: How is the ASML Field Service Hub case applied to the general model?	28
7.1	Introduction	28
7.1.1	Adapting the General Model to ASML	28
7.2	Specific Challenges of the ASML Field Service Hub	28
7.3	Data Preparation	29
7.3.1	Data Aggregation Level	29
7.4	Analysis	30
7.4.1	Standard Data Checks	30
7.4.2	Conclusion from Data Analysis	30
7.5	Data Challenges and Limitations	30
7.6	Application of Forecasting Models to ASML Data	31
7.6.1	The design of the forecasting model	31
7.6.2	What methods are best used to forecast demand for these flows?	31
7.6.3	Technical Specifications	31
7.6.4	Forecast Process	32
7.6.5	Forecasting Results	32
7.7	Application of Capacity Planning Model to ASML Data	32
7.7.1	How to Translate the Forecast into Decisions for Production Capacity?	33
7.8	Conclusion of the Use Case	33
7.9	Insights for Broader Application	33
7.9.1	Flexibility of the Model	34
7.9.2	Data-Driven Decision-Making	34
8	Research Question 5: How are the models validated and compared to benchmarks?	35
8.1	Introduction	35
8.2	Performance Metrics	36
8.2.1	Forecasting Accuracy Metrics	36
8.2.2	Validation and Simulation Metrics	36
8.2.3	Cost-Based Performance Comparison	37
8.3	Comparison Against ASML's Existing Benchmarks and Models	38
8.3.1	Benchmark	38
8.4	Results of the Implemented Forecasting Models at FSH	38
8.4.1	Error Comparison	38
8.4.2	XGBoost rolling window forecast	39
8.4.3	Random Forest, shorter timeframe	39
8.4.4	Forecasting Out of Sample	39
8.4.5	Explainability in the Use Case	39
8.4.6	Robustness in the Use Case	40
8.5	Validation: Cost Simulation of Production Capacity Strategies at ASML FSH	41
8.5.1	Realized Cost in Minutes	42
8.5.2	Realized Production Minutes Missed / Late Orders	42
8.5.3	Total Monthly FTE Used Over 8 Months	42
8.5.4	Volatility in Personnel Change	42
8.5.5	Key Insights:	42
8.6	Conclusion	43
8.6.1	Forecast Accuracy and Model Performance	43
8.6.2	Explainability and Robustness	43
8.6.3	Capacity Planning and Cost Optimization	43

8.6.4	Comparison to Benchmark Results	43
8.6.5	Final Remarks	44
9	Discussion	45
9.1	Summary of Key Findings	45
9.1.1	Performance of the Different Forecasting Models	45
9.1.2	Unexpected Outcomes	46
9.2	Interpretation of Results	46
9.2.1	Best Performing Models for Cost and Accuracy	46
9.2.2	Robustness and the Impact of Noise	47
9.3	Practical Implications and Strategic Benefits for ASML	47
9.3.1	Proactive Operations and Data-Driven Decision Making	47
9.3.2	Reduction in Forecast Errors and Cost Savings	47
9.3.3	Workforce Planning and Production Scheduling Improvements	47
9.3.4	System Integration and Future Optimization for ASML	48
9.3.5	Limitations of the Current Benchmark Model and Strategic Next Steps for ASML	48
9.3.6	Conclusion Practical Implications and Strategic Benefits for ASML	48
9.4	Insights from the Use Case for Broader Application	48
9.4.1	Flexibility of the Model	48
9.4.2	Data-Driven Decision-Making	49
9.5	Conclusion	49
9.5.1	Main Research Question: How to design an improved forecasting and production capacity optimization model that leads to an improved mid-term production capacity planning that can be applied to the ASML Field Service Hub?	49
9.5.2	Research Question 1: What methods are best suited to forecast demand under uncertainty?	49
9.5.3	Research Question 2: How should forecasting models be designed for optimal capacity planning?	50
9.5.4	Research Question 3: How can forecasts be translated into production capacity decisions?	50
9.5.5	Research Question 4: How is the ASML Field Service Hub case applied to the general model?	51
9.5.6	Research Question 5: How are the models validated and compared to benchmarks?	52
9.6	Limitations	52
9.6.1	Data Limitations	52
9.6.2	Model Limitations	52
9.6.3	External Limitations	53
9.6.4	Risk Assessment	53
9.6.5	Ethical Considerations	53
9.7	Implementability at FSH	53
9.7.1	Scalability	53
9.7.2	Technical Environment	54
9.7.3	User Interaction	54
9.7.4	Next Steps	54
9.8	Future Research Directions and Improvements	54
9.8.1	Testing and Generalizability	54
9.8.2	Investigating the Supply Side	54
9.8.3	Panel Data and Advanced Models	55
9.8.4	Bias in Forecasting Models	55
9.8.5	Preventing Overfitting	55
9.8.6	Cost of Undercapacity and Overcapacity	55
9.8.7	ARIMA Strategy Improvement	55
10	Reflection	56
10.1	Insights Created	56
10.1.1	Forecasting	56
10.1.2	Production Capacity Optimization	56
10.2	Implications for the Research Field	56
10.3	Strategic Benefits for ASML	57

10.4 Final Remarks 57

Chapter 1

Introduction

In today's rapidly evolving and highly competitive technological landscape, managing production capacity effectively is a crucial challenge for organizations. Industries characterized by volatility and complexity, such as manufacturing, logistics, and technology, face significant uncertainty in demand and operational flows. Efficient capacity planning is no longer just a competitive advantage, it is essential for maintaining service reliability, operational efficiency, and customer satisfaction. Companies across sectors are increasingly turning to advanced forecasting techniques, including Artificial Intelligence (AI) and Machine Learning (ML), to address the inherent uncertainties of demand fluctuations and optimize their production capacity.

A fundamental problem faced by many organizations is the uncertainty of product flows and the variability of demand, which can arise from multiple factors such as market volatility, seasonality, and unpredictable customer behavior. In environments where product flows are unknown or difficult to predict, production planning becomes particularly challenging.

To tackle these issues, businesses are turning to machine learning techniques, which offer more sophisticated approaches to demand forecasting compared to traditional statistical models. By learning patterns from vast amounts of historical data, machine learning models can provide more accurate predictions even in complex, uncertain environments. In turn, these forecasts can be integrated into capacity planning models, allowing organizations to optimize the allocation of resources, minimize costs, and improve service delivery.

This study addresses the general problem of demand uncertainty and capacity planning by focusing on the development of robust forecasting models that integrate machine learning approaches with production optimization. The goal is to design a solution capable of enhancing operational efficiency, reducing costs, and improving decision-making in industries where uncertainty is a critical challenge.

The remainder of this thesis will focus on applying these concepts to a specific use case, ASML's Field Service Hub (FSH). ASML, a leader in the semiconductor industry, operates a highly complex and specialized service network where managing production capacity is a critical operational need. The ASML FSH is a throughput facility for ASML and faces unique challenges due to the highly technical nature of its service offerings and the unpredictable demand patterns associated with machine servicing and spare parts management. By exploring the ASML case in detail, the study will demonstrate how advanced forecasting techniques can be applied to real-world scenarios to address capacity planning issues effectively.

In summary, this research aims to develop a machine learning-driven forecasting model that improves the alignment of demand forecasts with production capacity planning, thereby enhancing overall operational performance. The following chapters will investigate the relevant literature on forecasting and capacity optimization, present the research methodology, and provide insights into the model development and validation processes.

Chapter 2

Literature Review

2.1 Introduction

Efficient spare parts management is crucial for maintaining operational continuity and ensuring customer satisfaction across various industries. As organizations strive to optimize inventory levels, enhance forecasting accuracy, and integrate supply chain processes, the classification, forecasting, and integration of spare parts have gained significant attention in both research and practice. This literature review explores the latest advancements, challenges, and best practices in spare parts management, with an emphasis on forecasting and integration, given the thesis focus on production capacity optimization rather than spare parts classification.

2.1.1 Background

The global spare parts industry is a substantial market, valued at approximately \$200 billion, as reported by McKinsey (Gallagher et al., 2005). Effective management of spare parts is critical for minimizing machinery downtime and ensuring continuous operation, making it essential to optimize not only inventory levels but also the broader supply chain. This optimization is complicated by challenges such as long lead times and the unique characteristics of spare parts.

Spare parts management presents unique challenges due to several factors:

- **Intermittent Demand:** Many spare parts experience frequent periods of zero demand. Approximately 60% of Stock Keeping Units (SKUs) in various inventory categories, including spare parts, exhibit intermittent demand characteristics (A. A. Syntetos et al., 2016).
- **Lumpy Demand:** Non-zero demand intervals for spare parts often exhibit high variability, complicating inventory management (Hua et al., 2007).
- **High Number of Unique Spare Parts:** The large variety of spare parts contributes to management complexity (Cohen et al., 2006).
- **Varied Supply Characteristics:** Spare parts within the same organization or product can come from numerous suppliers, each with different supply characteristics (A. Syntetos and Boylan, 2008).
- **High Costs:** Even inexpensive spare parts can create substantial costs due to equipment downtime if not managed properly (Boström and Lundell, 2020).
- **Risk of Obsolescence:** Spare parts are susceptible to stock obsolescence depending on their position in the product lifecycle (Teunter et al., 2011).
- **Limited Historical Data:** Often, there is insufficient historical data to accurately forecast demand (Hua and Zhang, 2006).
- **Lifecycle Dependency:** The demand for spare parts is influenced by the equipment's usage and maintenance, with increased wear and less effective maintenance leading to higher demand (Kennedy et al., 2002).

Given these complexities, a variety of methods exist for optimizing spare parts management, each with its own advantages and limitations. This literature review categorizes spare parts management into three fundamental steps:

- **Classification:** Identifying and grouping spare parts to determine the most effective inventory and supply chain management strategies.
- **Forecasting:** Applying forecasting techniques to predict demand for each category of spare parts.
- **Integration:** Implementing the forecasts into supply chain strategies or production planning to improve overall efficiency.

Due to the specific use case of production capacity optimization, spare part classification is not required. Instead, the focus is on aggregating the demand forecasts into a total production capacity forecast. This approach simplifies the model by treating the aggregated demand as a single input, rather than differentiating between various spare parts.

While the classification, forecasting, and integration of spare parts are addressed in various studies, the focus often remains on individual steps rather than their combined application. For instance, Bacchetti and Saccani explore the gap between research and practice in spare parts management (Figure 2.1, Bacchetti and Saccani, 2012), while other studies, such as those by Boström and Lundell, concentrate on specific use cases (Boström and Lundell, 2020). This review aims to find best practices across different scenarios and contexts, highlighting how integrating these steps can enhance spare parts management strategies, with the final goal of implementing the new ideas for the use case at ASML to validate the proposed framework.

The research gap identified in this review is the need to explore how the integration of forecasting and supply chain management can improve production capacity planning. While existing literature often focuses on these steps individually, there is a need to explore how their integration can lead to enhanced production capacity optimization, bridging the gap between theoretical advancements and practical applications.

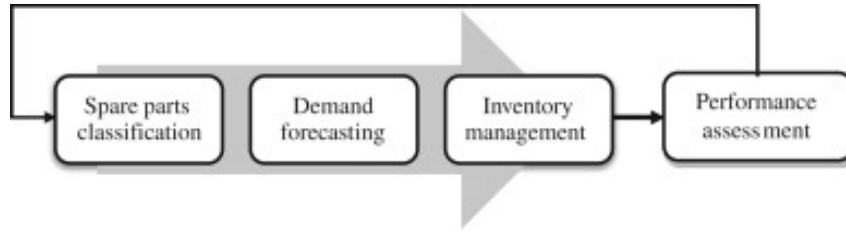


Figure 2.1: An integrated approach to spare parts management (Bacchetti and Saccani, 2012).

2.2 Classification

Effective forecasting and inventory management of spare parts depend on distinguishing parts based on their criticality and demand patterns. Classification by criticality helps prioritize key items for forecasting and stocking, while demand pattern classification allows for the selection of forecasting models that align with the specific characteristics of demand. This dual approach is vital because, in most industries, about 20% of SKUs typically generate nearly 80% of sales (Zhuang et al., 2022).

As discussed in subsection 2.1.1, spare parts exhibit diverse characteristics, making it crucial to focus on the most impactful components to optimize supply chain performance. Different spare parts display varying demand patterns, requiring a classification framework that can group parts into distinct categories, each of which may necessitate different forecasting techniques (Heinecke et al., 2013). Moreover, demand patterns change depending on the level of data aggregation. Whether data is aggregated across multiple parts or over different periods (e.g., daily vs. monthly aggregation), the patterns that emerge from the analysis may differ significantly. As later sections will explain, demand patterns observed in daily data aggregation often vary considerably from those seen at the monthly level, influencing the choice of forecasting model.

The concept of Coefficient of Variation (CV), along with Average Demand Interval (ADI) and specific cut-offs, plays a crucial role in determining the effectiveness of forecasting methods such as Croston or Syntetos-Boylan Approximation (SBA) (Syntetos et al., 2005). This method is seen as the benchmark to classify and forecast and is used extensively (e.g. Amirkolai et al., 2017). It is shown in Equation 2.1 and Figure 2.2. The only other methods of pattern-based classification that was found within the current scope is k-means clustering. Qiu et al. use this method to classify spare parts for an agricultural company

(Qiu et al., 2022). Their goal is different, however, meaning that for spare parts classification used for forecasting of spare part demand there is only one method widely used.

$$\gamma = \frac{CV_s}{CV_I} = \frac{\sigma_s/\mu_s}{\sigma_I/\mu_I} = \frac{\text{Coefficient of variation of demand size}}{\text{Coefficient of variation of demand interval}} \quad (2.1)$$

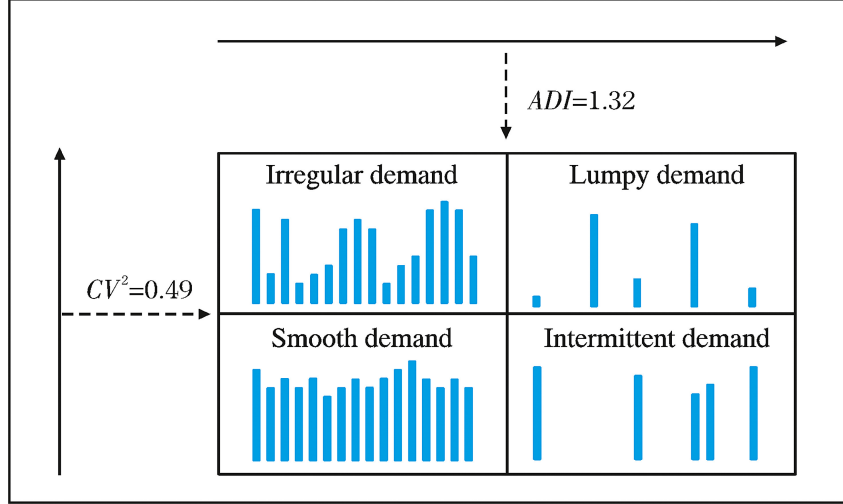


Figure 2.2: Pattern-based Demand Classification (Zhuang et al., 2022)

In summary, while pattern-based classification is key for selecting appropriate forecasting models, it is less relevant to this study's focus on production capacity optimization. Since this research aggregates demand into a single forecast for capacity planning, individual part classification is not required. Instead, attention is directed toward aggregate demand, simplifying the forecasting process and aligning to optimize total production capacity.

2.3 Forecasting Techniques

This study evaluates a variety of forecasting techniques to determine their effectiveness in predicting demand and improving supply chain operations. These techniques range from traditional time-series models, Croston's method, and bootstrapping, to more advanced machine learning and artificial intelligence approaches. Initially, forecasting was applied to spare parts individually, but the focus shifted towards predicting total spare parts demand for optimizing production capacity. An overview of the literature for forecasting techniques is shown in Table 2.1.

Forecasting spare part demand is a crucial aspect of spare parts management, as it helps organizations plan their inventory levels and optimize their supply chain operations. Accurate forecasting can lead to reduced costs, improved customer satisfaction, and enhanced overall efficiency.

Forecasting spare part demand can be divided into three main categories; judgmental forecasting, reliability-based forecasting, and time-series-based forecasting (Hu et al., 2018). Judgmental forecasting relies on expert opinions and subjective assessments to predict spare part demand. This approach involves gathering insights from industry experts to make informed decisions about future demand levels. While judgmental forecasting plays a significant role in decision-making, this research will primarily focus on the more extensively researched reliability-based and time-series-based forecasting methods to provide a comprehensive understanding of forecasting spare part demand.

Time-series-based forecasting involves the use of historical data to predict future demand patterns. This can be achieved through mathematical models or artificial intelligence techniques. Mathematical models, such as exponential smoothing and auto-regressive integrated moving average (ARIMA), are widely used in the industry for their simplicity and effectiveness. On the other hand, artificial intelligence-based forecasting methods, such as machine learning, neural networks and deep learning algorithms, provide advanced predictive capabilities by learning intricate patterns in the data. However, these techniques come at the cost of substantial computing power and the need for a large amount of data to ensure accurate predictions.

Engineering-based forecasting methods, specifically reliability-based forecasting, focuses on understanding the underlying reliability of the spare parts and their associated equipment. This approach takes into account factors like failure rates, maintenance schedules, and equipment lifetimes to estimate future demand.

2.3.1 Time-series Based Forecasting: Mathematical

Time-series forecasting is widely employed, with simple methods like Simple Exponential Smoothing (SES) being among the earliest (Little and Brown, 1956). However, SES struggles with intermittent demand, which is common for spare parts. This bias makes it unreliable for lumpy or erratic demand patterns (Gardner, 1985, Syntetos et al., 2015).

Croston’s Method

Croston’s method, introduced in 1972, is specifically designed for intermittent demand by separately estimating the demand size and frequency, then providing a ratio-based forecast (Croston, 1972). Croston’s method has been improved upon with variants like the Syntetos-Boylan Approximation (SBA, Syntetos et al., 2005), the Teunter-Syntetos-Babai (TSB, Teunter et al., 2011) method, and the Shale, Boylan & Johnston method (SBJ, Shale et al., 2006), which reduce bias and handle obsolescence more effectively.

ARIMA

Beyond Croston and its variants, there is also the ARIMA method, which is on the border between parametric mathematical methods and machine learning methods. Known as auto-regressive integrated moving average, ARIMA is a linear regression model that learns parameters from historical data and employs them for forecasting. Unlike machine learning algorithms, ARIMA can learn in real-time and does not require as much data to provide reliable insights (Fattah et al., 2018). Studies show that ARIMA models can outperform traditional methods like moving averages and exponential smoothing (Vargas and Cortés, 2017).

Bootstrapping

Bootstrapping is a resampling technique that helps estimate the distribution of future demand, particularly useful when historical data is limited (Willemain et al., 2004). It simulates entire demand distributions instead of relying on point estimates (Willemain and Smart, 2001). Although bootstrapping can slightly improve forecast accuracy compared to methods like exponential smoothing and Croston’s, the added complexity may not always justify its use (Syntetos et al., 2015, Hua and Zhang, 2006).

2.3.2 Time-series Based Forecasting: Artificial Intelligence

Supervised Machine Learning

Supervised machine learning models, such as Random Forest and XGBoost (eXtreme Gradient Boosting), have become popular for demand forecasting (Breiman, 2001). These models require large datasets (smaller than for deep learning however) but tend to offer better explainability and faster training than deep learning models. XGBoost, in particular, has been successful in providing accurate forecasts with minimal computational resources (Chen and Guestrin, 2016).

Deep Learning

Deep learning models, such as Artificial Neural Networks (ANNs, Shafi et al., 2023, Aktepe et al., 2021), Recurrent Neural Networks (RNN, Amin-Naseri and Tabar, 2008), and Long Short-Term Memory (LSTM, Bandara et al., 2020, Shafi et al., 2023) networks, can capture complex patterns in demand data. Although these models offer high accuracy, they require extensive data and computational power. Another problem is that these models are black box models, and are thus hard to use in combination with input from people. Deep learning models, which can capture sequential dependencies in time-series data, have been shown to outperform traditional models in several cases, but their implementation and maintenance can be costly and difficult.

Table 2.1: Reviewed literature for the forecasting of spare parts

Paper	Proposed Method and Benchmarks	Use Case
Mathematical Approach to Spare Part Demand Forecasting		
Little and Brown, 1956	Single Eexponential Smoothing	Theory only
Croston, 1972	Croston’s method	Theory only
Syntetos et al., 2005	Croston SBA, EWMA, Croston	Data from Automotive Industry
Shale et al., 2006	Croston SBJ	Theory only
Teunter et al., 2011	Croston TSB, SBA, Croston, SES	Simulation
Boström and Lundell, 2020	Croston SBA	Stock Control in High-Tech Industry
Vargas and Cortés, 2017	ARIMA, ANN, ARIMA-ANN hybrid, MA, ES	Automobile Spare Parts
Willemain and Smart, 2001	WSS: Patented Bootstrapping Method	Patent for Bootstrapping Method
Willemain et al., 2004	Bootstrapping, Croston, ES	Nine Large Industrial Data Sets
Artificial Intelligence Approach to Spare Part Demand Forecasting		
Zhu et al., 2019	CBR, ARIMA, SVM, Holt-Winters, Fuzzy Theory, ANN, GA, ML	Forecasting Demand of Emergency Resources
Amirkolaie et al., 2017	ANN: MLP-NN-BP, MA, SES, Croston, Croston SBJ, Croston TSB, Croston SBA	Business Aircraft Spare Parts Supply
Shafi et al., 2023	ANN, Croston, Auto-ARIMA, SES, RNN, LSTM, SVM, Holt-Winter, Random Forest, XGBoost, MLP, Auto-SVR	Historical Data Sets of Aviation Spare Parts
Zhuang et al., 2022	IDCF (combination of BTIDCF and TBIDCF), SBA, Bootstrapping	Data Set of Auto After-sales Business
Aktepe et al., 2021	Multivariate linear regression, ANN, Multivariate nonlinear regression, SVR with polynomial kernel	Construction Machinery Company
Amin-Naseri and Tabar, 2008	ANN, MLP, GRNN, Croston, SBA,	Process Industries
Bandara et al., 2020	LSTM-MSNet, other LSTM	Different Open Data Sets
Hua and Zhang, 2006	LRSVM, ES, Croston, MB, IFM, SVM	Petrochemical Enterprise
Reliability: Engineering Approach to Spare Part Demand Forecasting		
Romeijnders et al., 2012	two-step method that considers demand and type of component, Croston	Service Provider in the Aviation Industry
Van der Auweraer and Boute, 2019	SMI, SES, SBA	Simulation with Spare Part Data Set
This Research	XGBoost, Random Forest, ARIMA	ASML FSH: Demand forecast

SBA = Syntetos & Boylan Approximation, EWMA = Exponentially Weighted Moving Average, SBJ = Shale, Boylan & Johnston, TSB = Teunter-Syntetos-Babai, SES = Single Exponential Smoothing, ARIMA = auto-regressive integrated moving average, ANN = Artificial Neural Network, WSS = Weighted Subseries Smoothing, MA = Moving Average, ES = Exponential Smoothing, CBR = Case Based Reasoning, SVM = Support Vector Machine, GA = Genetic Algorithm, ML = Machine Learning, MLP = Multi-Layered Perceptron, RNN = Recurrent Neural Network, LSTM = Long Short-Term Memory, SVR = Support Vector Regression, IDCF = Intermittent Demand Combination Forecasting, BTIDCF = Best Threshold Intermittent Demand Combination Forecasting, TBIDCF = Transfer-Based Intermittent Demand Combination Forecasting, GRNN = Generalized Regression Neural Network, LRSVM = Logistic Regression Support Vector Machine, MB = Markov Bootstrapping, IFM = Integrated Forecasting Method, SMI = Service Maintenance Information

2.3.3 Engineering: Reliability-Based Forecasting

Reliability-based methods incorporate data beyond historical demand, such as failure rates, maintenance schedules, and product life cycles (PLC, see Figure 2.3). These methods can outperform time-series models because they are tailored to the operational aspects causing the demand, like equipment failure (Romeijnders et al., 2012). However, their requirement for installed base data limits their applicability to specific contexts (Van der Auweraer and Boute, 2019).

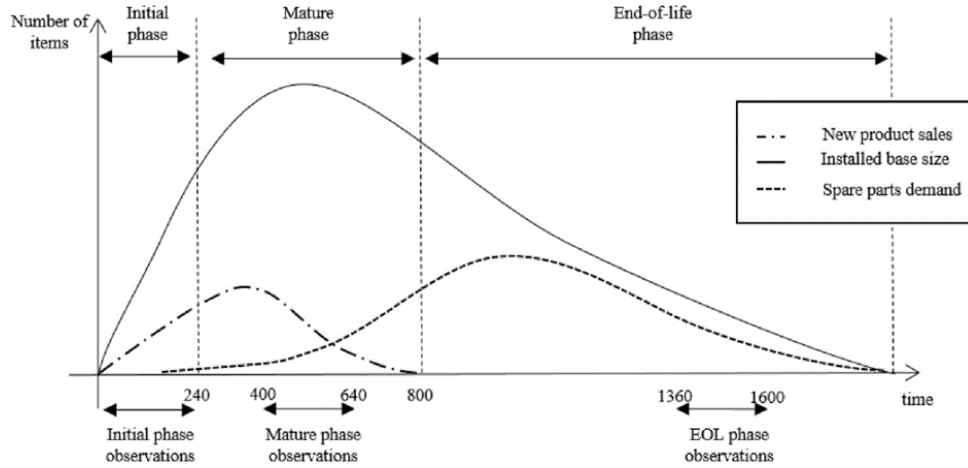


Figure 2.3: Product Life Cycle (Van der Auweraer and Boute, 2019)

2.4 Integrating the Forecast in the Supply Chain

This section examines the various approaches to integrating forecasting insights into supply chain operations, focusing on optimizing inventory strategies, production, and capacity planning based on demand forecasts. Several challenges such as managing obsolescence, dealing with long lead times, and ensuring data accuracy are also highlighted. Ultimately, the goal is to determine how the ASML Field Service Hub can utilize these insights to enhance its operational efficiency. An overview of literature for supply chain integration can be seen in Table 2.2.

The integration of a forecast into the supply chain can take multiple forms. Forecasting can optimize inventory levels, production planning, or production capacity, among other elements. Selecting the right optimization focus is crucial to maximizing benefits, particularly for spare parts management, as it directly influences operational readiness and cost management.

2.4.1 What to Optimize?

When integrating forecasting into the supply chain, several areas can be optimized, each contributing to improved efficiency in distinct ways. Below, potential optimization areas are explored based on insights from the literature.

- **Stocking Policies:** This involves optimizing reorder policies and inventory levels to ensure spare parts are available when needed, minimizing both stockouts and excess inventory. Bacchetti and Saccani stress the importance of classifying spare parts to align stock control policies with forecasted demand patterns (Boylan et al., 2008).
- **Inventory Management:** Methods such as Just-in-Time (JIT), Economic Order Quantity (EOQ), and Vendor-Managed Inventory (VMI) are frequently employed to streamline inventory operations. For instance, Deng and Liu propose a deep learning model for inventory demand prediction, achieving high prediction accuracy and significant cost reductions (Deng and Liu, 2021).
- **Production Planning and Scheduling:** Aligning production with demand forecasts can minimize lead times and improve manufacturing efficiency. Albey et al. present a model that updates production planning based on forecast evolution, reducing inventory costs and backlog (Maccarthy and Liu, 1993, Albey et al., 2015, Mula et al., 2006).

Table 2.2: Reviewed literature for the supply chain integration of spare parts

Paper	Optimization Method	Optimization Case
Elmachtoub and Grigas, 2022	SPO: Smart “predict, then optimize”, predict-then-optimize	Shortest Path and Portfolio Optimization Problems
Basten and Arts, 2017	Marginal Analysis Approach, SPO+, Random Forest	Fleet Readiness: Stocking Spare Parts and High-tech Assets
Bacchetti and Saccani, 2012	Different Re-order Policies based on Forecasts are Compared	Optimize Stock Control
Boylan et al., 2008	Simulation of Different Stock Control Methods	Optimize Stock Control
Mula et al., 2006	Artificial Intelligence Models, Simulation Models, Analytical Models, and Conceptual Models	Production Planning under Uncertainty: Aggregate Planning, Hierarchical Production Planning, Material Requirement Planning, Capacity Planning, Manufacturing Resource Planning, Inventory Management, and Supply Chain Planning
Albey et al., 2015	MMFE: Integrating Production Planning and Forecast Evolution Models	Production Planning Semiconductor Manufacturer
Yan and Wang, 2022	SPO, Predictive Prescription, Semi-SPO, Marginal Analysis Approach, PLC perspective	Applications in Transportation Management: Shortest Path Problem, Maximum Travel Bonus Problem
Tang and Khalil, 2022	Python Library for SPO and Predict-then-Optimize	Shortest Path, Multiple Knapsack, and the Traveling Salesperson Problem
Deng and Liu, 2021	DIM: LSTM for both Forecasting and Decisions	Inventory Management
Maccarthy and Liu, 1993	SPT, WSPT, EDD, Moore-Hodgson Algorithm, Johnson’s algorithm	Production Scheduling Theory Comparison
Biçer and Seifert, 2017	Dynamic Optimization Model: MMFE with Rolling Horizon	Production Scheduling under Capacity Constraint and Demand Forecast Updating
Chien et al., 2022	AHP and Multi-Objective Integer Optimization	Supplier Selection in the Electric-Mechanical Equipment Industry
Santana et al., 2017	ABC TDABC	Production Capacity Optimization in Industry 4.0
Sharma et al., 2017	Simulation	Preventive Maintenance for Equipment of the Army
do Rego and De Mesquita, 2015	Simulation	Automotive Spare Parts for a Brazilian Car Manufacturer
This research	Simulation with Exhaustive Search	Optimizing Production Capacity and Decide on FTE

MMFE = Martingale Model of Forecast Evolution, PLC = Product Life Cycle, DIM = Deep Inventory Management, LSTM = Long Short-Term Memory, SPT = Shortest Processing Time, WSPT = Weighted Shortest Processing Time, EDD = Earliest Due Date, ABC = Activity Based Costing, TDABC = Time-Driven ABC

- **Supplier Selection:** Chien et al. developed a multi-objective optimization model for supplier selection, balancing cost, quality, and reliability, demonstrating how supply chain efficiency can be improved through strategic supplier partnerships (Chien et al., 2022).
- **Fleet Readiness:** Fleet readiness depends on spare parts availability to minimize downtime and ensure asset reliability. Basten and Arts developed an optimization model for fleet readiness, focusing on minimizing the costs of holding spare parts while maintaining high asset availability (Basten and Arts, 2017).
- **Production Capacity Optimization:** Optimizing production capacity is another critical area where forecasts can guide decisions on resource allocation. Santana et al. emphasize the importance of aligning capacity planning with demand forecasts to avoid labor shortages or inefficiencies, which can lead to increased costs (Santana et al., 2017).

2.4.2 Optimization Methods

Once the area of optimization is determined, the next step is to select the appropriate method for integrating forecasting with optimization. Three primary approaches have been identified:

Predict-Then-Optimize (PTO)

Traditional supply chain models often use a two-stage framework where the forecast is generated first and then used as an input for optimization. This approach, while simple, does not always result in the most effective decisions since the optimization model does not directly influence how the forecast is generated. As a result, suboptimal decisions can be made if the forecast is inaccurate (Bertsimas and Kallus, 2020).

Smart Predict-Then-Optimize (SPO)

Recent advancements advocate for a more integrated approach, called Smart Predict-Then-Optimize, where forecasting and optimization are combined into a single framework (Elmachtoub and Grigas, 2022). This method leverages the structure of the downstream optimization problem to influence the forecast, leading to better decisions overall, even if the forecast itself has higher error margins. For instance, models using SPO can achieve improved cost-effectiveness and operational efficiency compared to traditional PTO methods. However, this is harder to implement, requires more computing power and time, and is in a lot of cases not worth the effort (Bertsimas and Kallus, 2020, Yan and Wang, 2022).

Simulation

Another option is to simulate results instead of optimizing mathematically. This can reduce time and complexity, but may not always find the optimal answer.

Simulation cannot be combined with SPO and creates the need for a split between forecast and optimization.

Sharma et al. use a combination of forecasting and simulation to optimize the timing for preventive maintenance (Sharma et al., 2017).

Do Rego and De Mesquita present results of a large-scale simulation study on spare parts demand forecasting and inventory control to select best policies within each SKU category for a Brazilian car manufacturer (do Rego and De Mesquita, 2015).

2.5 Research Gaps to Pursue

The literature review reveals several research gaps that are not just general problems, but directly applicable to improving the ASML Field Service Hub. In this section, gaps identified from existing research are discussed:

1. Feature Identification for Forecasting and Supply Chain Integration:

- **Gap:** While there is a growing body of research on supply chain forecasting, there is limited understanding of which specific features (such as failure rates, historical demand trends, and lead times) have the greatest impact on spare parts forecasting and how these forecasts should be integrated into broader supply chain decisions.

- Opportunity: A detailed analysis is needed to identify the most relevant features influencing demand forecasting and how they can be aligned with spare parts procurement, inventory management, and capacity planning.
2. Continuous Improvement of Forecasting Models:
- Gap: Forecasting methods, especially those dealing with seasonal patterns, product lifecycle variations, and obsolescence, still face challenges in maintaining accuracy over time.
 - Opportunity: Improving forecasting methods is crucial given a company's need to manage complex, long-life-cycle products and ensure parts availability. Research into hybrid models that combine traditional and AI-driven forecasting techniques could provide more reliable forecasts for these scenarios.
3. Empirical Studies and Case Studies:
- Gap: Much of the research on forecasting and supply chain integration remains theoretical, with fewer empirical studies that test models in real-world settings.
 - Opportunity: Given ASML's unique position in the semiconductor industry, a case study approach could be valuable to empirically evaluate how different forecasting models perform when integrated with inventory and capacity planning decisions.
4. Comparison of Models for Specific Product and Data Patterns:
- Gap: A key limitation in the existing literature is the lack of quantitative comparisons between forecasting models in specific use cases, especially for spare parts management. Reliability-based methods have not been extensively tested across diverse data patterns.
 - Opportunity: Research could explore which forecasting models work best for the specific data patterns associated with spare parts, such as high variability in demand, long lead times, and the criticality of certain parts.
5. Managing Uncertainty and Risk in Forecasting:
- Gap: While some research touches on uncertainty, there is a need for more robust models that integrate risk management into spare parts forecasting, especially in industries like semiconductor manufacturing where supply chain disruptions can have significant impacts.
 - Opportunity: Supply chains could be optimized by implementing models that dynamically incorporate uncertainty, such as robust models that include noise, real-time data on supplier performance or machine downtime, to adjust forecasts and capacity plans in response to real-world conditions.
6. Production Capacity Optimization in Response to Forecast Variability:
- Gap: While there is extensive research on production capacity planning, there is a lack of studies that explore how to dynamically optimize production capacity in response to fluctuating demand forecasts, especially in industries with long lead times and high product complexity.
 - Opportunity: Optimizing production capacity is particularly challenging due to the variability in spare parts demand and the criticality of maintaining machine uptime for customers. Research could focus on developing adaptive models that align production schedules and resource allocation with evolving demand forecasts. Such models would need to factor in constraints like production lead times, workforce availability, and machine capacity, helping organizations avoid over- or under-utilization of its resources.

By addressing these gaps, this thesis aims to contribute to the body of knowledge in supply chain forecasting and optimization, with an extra focus on the ASML Field Service Hub.

Chapter 3

Problem Statement

In industries characterized by high complexity and uncertainty, such as manufacturing and service delivery, effective production capacity planning is crucial for maintaining operational efficiency and meeting customer demands. One of the primary challenges organizations face is accurately forecasting demand in environments where product flows are uncertain and volatile. These uncertainties often arise from factors such as fluctuating market conditions, seasonality, and the unpredictable behavior of customers. The inability to predict demand accurately leads to inefficiencies in production capacity, resulting in either overutilization, which strains resources, or underutilization, which leads to increased operational costs.

Traditional forecasting models, while useful for simpler environments, often struggle to adapt to the dynamic and complex interactions between multiple product flows, leading to reactive decision-making. Organizations are increasingly recognizing the need to shift from reactive to proactive strategies by adopting more sophisticated forecasting tools that can better handle these complexities. Machine learning-based forecasting models have emerged as a powerful solution to this problem, enabling more accurate demand predictions and offering the potential to integrate those predictions into production capacity planning.

This thesis seeks to address the general problem of demand uncertainty and production capacity planning by developing a forecasting model that leverages advanced machine learning techniques. The objective is to create a proactive approach to decision-making, where production capacity can be dynamically adjusted based on more reliable and precise demand forecasts.

The specific use case explored in this research involves ASML's Field Service Hub, which faces similar challenges due to the highly specialized nature of its operations in the semiconductor industry. The facility plays a critical role in ensuring machine uptime for customers by managing the procurement, inventory, and availability of spare parts. Due to the complexity of the semiconductor industry and the unique nature of ASML's product flows, accurate forecasting and production capacity planning are essential to meeting customer demand. However, the current forecasting and planning models used by ASML are reactive and limited in their ability to dynamically adapt to fluctuating demand, especially given the complex interactions between the various product flows through FSH.

By focusing on this use case, the thesis aims to demonstrate how the general principles of demand forecasting and capacity optimization can be applied to complex, real-world scenarios. The ultimate goal is to design a model that shifts production capacity planning from reactive to proactive, enabling data-driven strategies that improve efficiency and customer satisfaction.

3.1 Research Questions

Based on the identified research gaps, the following research questions will guide the thesis:

Main Research Question:

- How to design an improved forecasting and production capacity optimization model that leads to an improved mid-term production capacity planning that can be applied to the ASML Field Service Hub?

Research Questions:

- **Research Question 1:** What methods are best suited to forecast demand under uncertainty?

- **Research Question 2:** How should forecasting models be designed for optimal capacity planning?
- **Research Question 3:** How can forecasts be translated into production capacity decisions?
- **Research Question 4:** How is the ASML Field Service Hub case applied to the general model?
- **Research Question 5:** How are the models validated and compared to benchmarks?

3.2 Methodology

To answer the research questions, a systematic approach is followed to ensure that the methods used are generalizable and not limited solely to the ASML use case. However, the ASML Field Service Hub serves as a practical example to demonstrate how the proposed methods perform in real-world applications. Below is an outline of how each research question will be addressed.

Main Research Question: How to design an improved forecasting and production capacity optimization model that leads to an improved mid-term production capacity planning that can be applied to the ASML Field Service Hub?

The main research question will be answered through the combined responses to the sub-questions. The proposed solution involves developing and validating a Python-based model that forecasts demand and optimizes production capacity based on those forecasts. The conclusions will be drawn by comparing the performance of the new model to the current benchmark used by ASML.

Research Question 1: What methods are best suited to forecast demand under uncertainty?

To answer this question, a set of requirements will be created based on the literature review. The suitability of different forecasting methods will be assessed by building upon the existing research, focusing on methods such as machine learning algorithms and time-series forecasting models.

Research Question 2: How should forecasting models be designed for optimal capacity planning?

This question will be answered by exploring how forecasting models can be designed in tandem with optimization frameworks. Existing literature on the integration of forecasting and capacity planning will form the foundation, which will then be expanded upon by applying these principles to the ASML use case.

Research Question 3: How can forecasts be translated into production capacity decisions?

The answer to this question will be derived from the developed model, which will link demand forecasts to capacity planning decisions. The process will involve reviewing existing decision-making frameworks and refining them through simulation and optimization techniques, ultimately creating a Python model based on the derived mathematical model.

Research Question 4: How is the ASML Field Service Hub case applied to the general model?

This question focuses on the application of the general forecasting and capacity optimization model to the ASML FSH use case. The goal is to demonstrate how the general principles can be adapted to address the real-world operational challenges faced by ASML, thus providing a practical validation of the model. Using ASML experts and data, the general model will be rewritten to accommodate for the intricacies of FSH, such as the unique product flows.

Research Question 5: How are the models validated and compared to benchmarks?

To address this question, the literature on model validation and quantitative comparison methods will be reviewed. The performance of the developed model will be assessed against ASML's existing benchmarks to determine whether it offers a measurable improvement in production capacity planning and demand forecasting accuracy.

The methodology involves a combination of theoretical research, practical model development, and empirical validation, ensuring that the proposed solutions are both academically sound and practically applicable in real-world scenarios like the ASML FSH.

Chapter 4

Research Question 1: What methods are best suited to forecast demand under uncertainty?

4.1 Introduction

Forecasting demand under uncertainty is a critical challenge faced by industries with volatile product flows, fluctuating customer behavior, and varying market conditions. Traditional forecasting methods often struggle to account for the complexity and dynamic nature of demand patterns, particularly in industries like manufacturing, logistics, and service-based sectors. This chapter explores the forecasting techniques best suited to handle demand uncertainty, with a focus on both classical statistical models and advanced machine learning methods.

4.2 Classical Statistical Forecasting Methods

Statistical forecasting methods have long been used to predict demand based on historical data. These models, while effective in stable environments, often face limitations in volatile and uncertain demand contexts.

4.2.1 Time-Series Models

Time-series models are commonly used in demand forecasting and are based on the assumption that historical patterns will continue into the future. Popular time-series methods include:

Exponential Smoothing

Techniques like exponential smoothing are employed to model trends. These methods give more weight to recent observations, but they rely on historical stability, which can be a limitation in environments where demand unpredictability is prevalent.

ARIMA

ARIMA models are widely employed for their simplicity and ability to model both trend and seasonality in data. However, ARIMA assumes that relationships between past and future values remain stable over time, making it less suitable for handling highly volatile demand.

While these time-series models have been effective in industries with moderate demand variability, they often fall short in capturing complex relationships and sudden changes in demand patterns.

4.2.2 Intermittent Demand Forecasting

For industries with intermittent demand, where demand occurrences are sporadic and irregular, specialized models like Croston's method and its variants (e.g., Syntetos-Boylan Approximation) have been developed. Croston's method is particularly suited for forecasting when demand events are infrequent,

but it assumes independence between demand occurrences and quantities, which may not always be valid in highly complex supply chains.

4.3 Machine Learning Approaches to Forecasting

Recent advancements in machine learning have introduced more powerful tools for forecasting demand under uncertainty. Machine learning models offer the flexibility to capture non-linear relationships, complex interactions, and adaptive patterns, using multiple inputs or features, making them particularly well-suited to environments with significant demand volatility.

4.3.1 Decision Tree-Based and Gradient Boosting Models

Decision tree-based models such as Random Forest and gradient boosting models such as XGBoost have become increasingly popular in demand forecasting due to their ability to handle large datasets and capture complex interactions between multiple variables. These models use an ensemble of decision trees to improve prediction accuracy, and their robustness makes them effective in dynamic environments with fluctuating demand.

Random Forest

This method constructs multiple decision trees and combines their results to provide a more accurate prediction. Random Forest is capable of handling both numerical and categorical data, and its ability to model complex relationships makes it highly adaptable to uncertain demand patterns.

XGBoost

As an enhancement to Random Forest, XGBoost incorporates gradient boosting, which sequentially builds trees to correct errors from previous ones. XGBoost has been widely adopted due to its speed and accuracy, especially in environments with highly dynamic and nonlinear relationships.

4.3.2 Neural Networks and Deep Learning

Artificial Neural Networks and deep learning models, including Long Short-Term Memory networks, have shown promise in demand forecasting, particularly when handling time-series data with complex temporal dependencies.

ANNs

By mimicking the way the human brain processes information, ANNs can capture non-linear patterns and relationships in data. They can be trained on large datasets to identify underlying patterns, making them suitable for environments where demand is highly irregular or uncertain.

LSTM Networks

As a specialized type of recurrent neural network, LSTMs are particularly effective in capturing long-term dependencies in time-series data. This makes them ideal for forecasting scenarios where past demand trends influence future events, even when those patterns are irregular.

Despite their accuracy, neural networks often require large amounts of data and significant computational resources, which can be a barrier for some organizations. However, in highly complex environments, their ability to adapt to changing conditions and learn from intricate data patterns makes them a powerful forecasting tool.

4.4 Approach for the Forecasting Model

The forecasting model should be designed based on the available data and the requirements of the use case. A variety of models were explored, including simple linear models and machine learning models.

The goal is to choose from the methods found in literature review, and compare and test three of them, to finally conclude which method works best.

4.4.1 Requirements for the Forecasting Model

To determine the most suitable methods for forecasting demand, both the literature review conducted in section 2.3 and the following key requirements were considered:

1. **Literature-Based Model Selection:** The forecasting model should be well-supported in academic and industry literature. It is essential to use established models that have demonstrated success in similar contexts to ensure reliability and credibility.
2. **Time Series Approach:** Given the sequential nature of the data, the model should focus on time series analysis to capture trends, seasonality, and other temporal patterns. This is critical for maximizing the value of the available data and ensuring that the model can handle historical fluctuations effectively. Models using panel data are left out of scope.
3. **Proven Methods:** Given the constraints of this project, specifically the short timeframe of two years of historical data, testing experimental or unproven methods is not feasible. The model must be fast, reliable, and backed by a strong track record of performance in similar cases.
4. **Limited Historical Data:** With only 1.5 years of training data available to forecast 6 months ahead, the model needs to be optimized for short data windows and sub-optimal data quality. It should be capable of leveraging time series data alongside additional features to improve accuracy.
5. **Multi-Step Forecasting:** The model must support forecasting multiple steps ahead. For mid-term production capacity planning, a time horizon of up to six months is necessary, meaning the model should be capable of multi-output predictions.
6. **Explainability:** The model should not operate as a "black box." It must be explainable, allowing stakeholders to understand how predictions are made and what factors contribute to the forecast to make informed decisions.
7. **Robustness:** A key criterion is the robustness of the model. It must deliver reliable predictions even under varying data conditions, such as missing or noisy data points, and also be reliable for out of sample forecasts.
8. **Implementation Feasibility:** The model needs to be implementable within ASML's operational and technical environment. Specifically, it should be compatible with Python, a key programming language used at ASML.
9. **Balancing Complexity and Performance:** While simplicity is preferred, model performance is the priority. The goal is to balance model complexity with forecast accuracy, opting for a simpler model if it provides reliable results. However, more complex models should also be considered if they offer significantly better performance.
10. **Future Growth Integration:** The model should be capable of incorporating anticipated growth, such as increases in demand or machine capacity expansions, to produce more accurate long-term forecasts.

4.4.2 Model Selection Process

Based on these requirements, the focus shifted to machine learning techniques that can handle the limitations in data availability and quality. While deep learning models typically require larger datasets and longer time frames to perform well, machine learning models are better suited to handling the current data constraints. Simpler parametric models, while easier to implement, were not considered sufficient due to their limited ability to handle the complexity of the operations that are being forecasted, but can be used as benchmark.

Exponential smoothing was ruled out for multi-step forecasting because it tends to be unstable when forecasting further into the future without fresh inputs. This instability renders it unsuitable for this case, especially with the mid-term 6-month forecasting horizon required.

4.4.3 Final Model Selection

After evaluating various models, three were selected for implementation based on the requirements listed before:

XGBoost

A highly efficient and scalable machine learning algorithm that handles structured data well, especially with limited historical data. XGBoost has been shown to outperform many other models in time series forecasting due to its ability to model complex relationships in the data.

Random Forest

A robust and versatile model that is easy to implement and interpret. Random Forest is known for its ability to handle noisy data and its resistance to overfitting, making it a strong candidate for this case.

ARIMA

A traditional time series forecasting model that performs well when working with small datasets and can capture linear trends and seasonality. ARIMA is included to compare how a simpler statistical approach that does not use features fares against machine learning models that have multiple features as input.

By comparing the performance of these three models, the best-suited forecasting method will be identified, with the final decision balancing accuracy, complexity, robustness, and explainability.

4.5 Explainability of Machine Learning Models

While machine learning models like Random Forest, XGBoost, and neural networks can provide highly accurate forecasts, their complexity often leads to a lack of interpretability. In recent years, techniques such as SHAP (SHapley Additive exPlanations) have been developed to provide insights into how machine learning models make decisions. SHAP values explain the contribution of each feature to the prediction, offering transparency in forecasting models and enabling decision-makers to understand the factors driving demand variability.

Understanding the decision-making process of a machine learning model is crucial, even in cases where black-box models are acceptable, as in the later investigated internal use case. While black-box models can produce accurate forecasts, gaining insight into the reasoning behind their predictions allows for a deeper understanding, validation of results, and verification of model robustness. This transparency is essential for sanity checks, such as verifying whether the most important features align with domain knowledge and expectations.

Overall, an explainable model should use the right mix of input features.

4.5.1 Application of SHAP Values to Explain Model Behavior

One of the most effective ways to enhance the explainability of a machine learning model is through Shapley values, originating from cooperative game theory (Sundararajan and Najmi, 2020). Shapley values allocate the contribution of each feature to the final prediction, distributing the overall prediction (or "payout") fairly among the features based on their importance. This method is implemented through the SHAP library in Python (Lundberg and Lee, 2017), which has become a standard tool for explaining complex models.

SHAP values provide a unified measure of feature importance, where the contribution of each feature is calculated by comparing all possible combinations of feature inputs. The goal is to estimate the marginal contribution of a feature by examining how much the prediction changes when it is included or excluded from the model. The key advantage of SHAP values is their consistency and local accuracy:

- **Consistency:** If feature A contributes more to the model than feature B, SHAP will assign a higher importance score to feature A.
- **Local Accuracy:** For any given prediction, the SHAP values of all features sum up to the difference between the model output and a baseline value (typically the mean prediction). This ensures that the SHAP explanation fully accounts for all points of the prediction.

SHAP values reveal how individual features (e.g., historical minutes, and mix of product flows) influence the predicted output. By visualizing SHAP values, one can see which features push the prediction higher or lower and by how much. This is particularly useful when features interact in complex, non-linear ways, which can make direct interpretation challenging.

Using SHAP values also helps in determining the robustness of the model:

- **Verify Feature Importance:** By checking that the features deemed important by SHAP align with expectations, model developers can ensure that the model’s behavior makes logical sense. For example, if historical production minutes significantly influence the forecast, SHAP will highlight this, confirming the feature’s relevance.
- **Detect Anomalies or Biases:** SHAP values can uncover situations where the model relies too heavily on certain features or ignores others that should be important. This can be a signal to revisit the model design or training process.
- **Improve Trust in Forecasting:** By understanding which factors contribute to predictions, decision-makers can gain confidence in the model’s recommendations, even if the model itself remains complex.

4.6 Conclusion

Forecasting demand under uncertainty requires a careful balance between accuracy, adaptability, and interpretability. Traditional time-series models, while useful in certain contexts, are often limited in environments where demand patterns are volatile. In contrast, machine learning methods, particularly decision tree-based models, offer the flexibility and power needed to handle complex, uncertain demand scenarios. The use of SHAP values for model explainability provides a holistic solution for industries seeking to enhance their forecasting and capacity planning processes.

Chapter 5

Research Question 2: How should forecasting models be designed for optimal capacity planning?

5.1 Model Design: Selecting Features and Variables

Designing an effective forecasting model for optimal capacity planning requires a careful selection of features and variables. These features need to capture the underlying factors driving demand, particularly in uncertain and volatile environments. The chosen variables should provide sufficient historical information and relevant signals to enhance the model's predictive performance.

Feature and variable selection depends on the use case and available data, but in general, forecasting models rely on historical production data or demand data to predict future outcomes. However, demand patterns are rarely influenced by a single factor. To capture the full complexity of demand, additional underlying data can be included as features in the model. This may include external drivers like economic indicators, seasonality, or product-specific flows that contribute to overall demand. For example, if the total demand is influenced by multiple product lines, it is often better to include data on those individual product flows to improve the model's accuracy and flexibility.

When designing forecasting models, it is essential to use a time-series format and not use panel data which is sometimes available, particularly for models that rely on sequential data (e.g., ARIMA, XGBoost, Random Forest). Time-series models assume that the data points are temporally ordered, meaning the model can identify patterns based on historical sequences. This temporal structure is critical for forecasting future demand, as it allows the model to learn from past behaviors and project them forward. In contrast, panel data, which combines cross-sectional and time-series data, may not be suitable for models that focus exclusively on sequential predictions.

Feature selection is one of the most important steps in model design, as irrelevant or redundant features can degrade the model's performance, leading to overfitting or underfitting. Techniques such as correlation analysis, feature importance scoring (e.g., SHAP values), and domain expertise are typically employed to refine the selection of input features. By focusing on the most relevant variables, the model can generate more reliable and interpretable forecasts for capacity planning.

It is difficult to generalize the feature selection process step because every process or data set has their own characteristics.

5.2 Data for Forecasting Models

The reliability of a forecasting model heavily depends on the quality and relevance of the data used. A well-designed forecasting model requires clean, structured, and comprehensive data that accurately represents the demand drivers. The data must not only be historical but also reflect the most recent demand trends, ensuring that the model adapts to changing patterns over time.

Data preparation is a crucial step in creating a reliable forecasting model. This process involves obtaining the necessary raw data, cleaning and remediating it, and transforming it into a format that can be used effectively by the machine learning model. Cleaning the data is essential to ensure that it does not contain errors, outliers, or inconsistencies that could mislead the model and negatively impact its

performance. For example, missing values, duplicate records, or incorrect timestamps must be corrected to avoid bias in the predictions. Since this is use case specific, more will be explained in section 7.3.

Additionally, feature engineering plays a key role in preparing the data. This involves generating new variables from the raw data that can enhance the model's ability to capture the underlying demand trends. Lag features and transformations of key variables are commonly used in time-series forecasting models to enrich the dataset. The selection of features should be aligned with the forecasting model's objectives, ensuring that the input data provides a strong signal for the target variable; future demand.

Finally, a train-test split needs to be made to create a training and testing data set. The division between the two depends on the data set and use case, but the training data should be the biggest chunk of available data.

5.2.1 Aggregation Level

An important decision in designing a forecasting model for capacity planning is determining the appropriate aggregation level of the data. The aggregation level refers to the time intervals at which the data is grouped, such as daily, weekly, or monthly.

The choice of aggregation level directly impacts the granularity of the forecasts and must align with the specific goals of the forecasting model. For instance, in scenarios where operational decisions, such as production capacity adjustments, are made every month, it is more practical to aggregate the data at the monthly level. This reduces the noise inherent in daily fluctuations and allows the model to focus on broader trends that drive longer-term demand. In contrast, if the objective is to manage short-term fluctuations in demand, daily or weekly aggregation may be more appropriate.

The key consideration here is the trade-off between granularity and stability. Higher granularity (e.g., daily data) may capture short-term variability but can introduce more noise, making it harder to identify stable demand patterns. On the other hand, lower granularity (e.g., monthly data) smooths out short-term fluctuations but may miss important demand spikes or troughs that occur within a month. In the context of capacity planning, where long-term trends are more important than daily variability, a monthly aggregation level is often preferred. This is further investigated for the use case in subsection 7.3.1.

5.3 Timeframe

A critical decision to be made regards the timeframe used for modelling. Different machine learning methods require varying time frames as input. Three approaches were tested:

1. **Shortest Timeframe:** Using only the previous time step's data to forecast.
2. **Longer Timeframe:** Including one more time step back for the most important feature, the time-series of the total production.
3. **Longest Timeframe:** Adding data from multiple previous time steps, and including longer time-frames of multiple features, enabling the model to capture temporal dependencies more effectively.

A trade-off exists between adding more data points and risking data overload, which could cause overfitting or run into the problem of not having enough training data because there are too many features.

5.4 Multi-Step Forecasting: Multi-Output Regression and Chain Regression

Additional complexity was introduced by creating multi-step forecasts for up to six months ahead, which allows the model to plan production capacity more efficiently in response to demand forecasts. For the XGBoost and Random Forest model, there is a decision to be made here. For ARIMA there is not, only a sequential chain regression is possible since ARIMA does not use any other features.

To address the need for forecasting demand up to six months ahead, two different approaches were considered for multi-step forecasting: Multi-Output Regression and Chain Regression. Both methods were evaluated and tested to determine which best fits the specific requirements of this case, offering flexibility, accuracy, and scalability.

5.4.1 Multi-Output Regression

In Multi-Output Regression, a separate model is trained for each forecasted time step. For this case, six models would be created, one for each of the next six months it tries to forecast. Each model is independent, meaning that the prediction for month one does not influence the prediction for month two, and so on.

Advantages:

- **Simplicity:** Each model can focus on a specific time horizon, potentially improving its accuracy for that particular step.
- **Parallelization:** Since the models are independent, they can be trained and run in parallel, reducing computation time.
- **Interpretability:** It is easier to understand the drivers behind each time step's forecast, as each model focuses on a distinct horizon.

Challenges:

- **Lack of interdependence:** Since the models are independent, they do not consider the relationships between consecutive time steps, which may be critical in capturing sequential trends.
- **Higher resource requirements:** Training separate models for each time step can be resource-intensive, particularly as the number of forecasted steps increases.

5.4.2 Chain Regression

Chain Regression, also known as Recursive Multi-Step Forecasting, tackles the forecasting problem by training a single model for the first time step (e.g., month one), and then using the output of that model as an input for the next time step (e.g., month two). This chaining process continues until all required time steps (up to month six) are forecasted.

Advantages:

- **Sequential learning:** The chained approach allows each prediction to take into account the results of the previous forecast, enabling the model to capture temporal dependencies more effectively.
- **Improved accuracy:** For time series with strong sequential relationships, this method often leads to more accurate long-term forecasts since each step builds on the previous one.
- **Resource efficiency:** Only one model is trained, and it can be reused across all forecasted time steps, potentially reducing computational overhead.

Challenges:

- **Error propagation:** Errors in the earlier steps can accumulate, leading to compounding inaccuracies in later steps. This is particularly problematic in long-term forecasts.
- **Complexity:** The chaining process makes the model more complex, and tuning it to prevent error amplification can be challenging.
- **Stability:** This method can create unstable outcomes that have unbound growth.

After testing and error comparison, it was found that multi-output regression performed significantly better in all cases for both XGBoost and Random Forest and was thus used for the remainder of the research.

5.5 Robustness Testing and Model Evaluation

In machine learning, the primary focus often lies on maximizing model accuracy (Tsipras et al., 2018). However, in practical applications like production capacity planning, robustness is equally important (Pranab, 2021). Robustness refers to the model’s ability to maintain consistent performance when faced with small variations in the input data, such as noise or future data that may slightly deviate from the current patterns.

5.5.1 Jitter Test

To assess the robustness of the models, a jitter test was conducted. This test involves adding varying levels of noise to the input data and evaluating the model’s error metrics, such as Mean Absolute Error (MAE) and Root Mean Square Error (RMSE). Gaussian noise was introduced at levels of 5%, 10%, and 15% to simulate possible future fluctuations in data, in line with similar robustness studies (Cai et al., 2021).

Gaussian noise was applied to the input data set to simulate the kind of variation expected in future data. The level of noise is adjusted to assess how the model behaves under small (5%), moderate (10%), and large (15%) deviations from the original data. Gaussian noise is ideal for this purpose, as it mimics random fluctuations and introduces smooth variability across all features.

The key question addressed by this test is: How much does the introduction of noise impact the model’s predictions? Models that show only minor increases in error under noisy conditions are considered more robust, indicating that they can generalize better to unseen or fluctuating data. In contrast, models that exhibit significant increases in error when noise is added may be overfitting the training data, rendering them less reliable in practical applications.

The results from the jitter tests provide insights into the robustness of the model under simulated noise. However, the extent to which these noise levels match real-world conditions remains uncertain, and additional real-world data would be required for a more accurate assessment.

Chapter 6

Research Question 3: How can forecasts be translated into production capacity decisions?

Forecasting models are not standalone tools; their value lies in how well they integrate into the broader decision-making framework of an organization. In the context of capacity planning, forecasts should guide actionable decisions that ensure the optimal allocation of resources, including labor, equipment, and materials. Proper integration of forecasts into capacity optimization models can help organizations proactively adjust their operations, avoid inefficiencies, and balance the costs of overcapacity and under-capacity.

This chapter explores how demand forecasts can be directly translated into production capacity decisions, examining different strategies for capacity planning and the role of simulation in validating the effectiveness of these strategies.

6.1 Capacity Planning Based on Forecasts

Optimizing production capacity offers the potential to significantly enhance operational efficiency. With varying demand for spare parts and complex maintenance schedules, forecasting can be used to determine the optimal number of resources (e.g., employees or machines) required to meet demand without incurring unnecessary costs.

A critical element of production capacity optimization is measuring and managing production rates. Santana et al. differentiate between three key capacities (Santana et al., 2017):

- **Design Capacity:** The theoretical maximum output that can be achieved with current resources under ideal conditions.
- **Effective Capacity:** The practical maximum output, accounting for factors like product mix, maintenance, and labor variations.
- **Actual Output:** The real production rate achieved over time, is often lower than effective capacity due to unforeseen disruptions.

These distinctions are crucial for aligning production capacity with demand forecasts and ensuring the model user can meet its service level agreements while avoiding the costs associated with overstaffing or underutilization.

There are three steps performed to determine the most cost optimal capacity planning decisions. First, a forecast is created. Then a strategy is used to set a preliminary capacity. This is done to minimize the volatility that comes from following a singular forecasted value for consecutive months. The first strategy is still to follow that singular forecasted value, but other strategies are created to decrease the volatility of the preliminary capacity planning. This will increase robustness, as the capacity will be based on multiple forecasts. The third and final step, is to do a cost optimization simulation to determine the best capacity planning based on the forecasted values. This will further smoothen the capacity planning and decrease capacity volatility.

6.2 Simulation of Production Capacity Strategies

Once a forecast is generated, the next step is to simulate different production capacity strategies to determine the most cost-effective approach. These simulations allow organizations to explore how well different capacity planning strategies perform under various demand scenarios. By testing multiple strategies, decision-makers can evaluate trade-offs between costs, resource flexibility, and service reliability.

In this context, production capacity strategies involve adjusting FTEs or machine hours based on forecasted demand. The objective is to ensure that capacity is neither over nor under the required levels, thus avoiding the costs of both underproduction (e.g., delayed orders or emergency shipping) and overproduction (e.g., idle workers or machines).

To address erratic behavior in forecasted FTE values (FTEs for each month are forecasted individually), a smoothing technique is applied to avoid excessive changes in capacity planning decisions. The optimization model performs an exhaustive search around the forecasted values to find the global optimum for FTE adjustments. During the exhaustive search, the model iterates through all possible capacity options close to the values found for the forecast to identify the configuration that minimizes cost, accounting for both under- and overcapacity and training costs. By optimizing based on forecasted demand, the model ensures that the selected capacity strategy minimizes costs while maintaining service level agreements. Given that the problem is discrete and the possible solutions are close to the non-optimized values, the exhaustive search is computationally efficient, typically taking no more than three minutes per run for the use case.

6.2.1 Production Capacity Strategies Based on Forecasts

To explore how different capacity strategies perform, three approaches to capacity planning were simulated over an eight-month period, using the forecast as input:

Strategy 1: Follow Forecast 3 Months Ahead

Adjust FTE levels based on the forecast exactly three months ahead, without smoothing fluctuations. This strategy waits until the last possible moment to change the scale of operations (upscaling or downscaling production takes three months) but will outperform the other two strategies when the forecast is perfect. While it allows for flexibility, it is sensitive to the accuracy of the forecast and may lead to abrupt shifts in capacity levels.

Strategy 2: Smoothing via Multiple Forecasts

Use the average of forecasts across different time horizons to stabilize decisions and avoid large swings in capacity planning. Here the mean of the forecasts of months three, four, five, and six are used to predict three months ahead. This approach reduces the impact of short-term fluctuations, smoothing out abrupt changes in resource allocation.

Strategy 3: Buffer + Smoothing

A buffer is added to the smoothed forecast to account for unexpected demand surges while preventing under-capacity. This strategy introduces a safety margin, preventing under-capacity and improving service reliability, albeit at a slightly higher cost.

6.3 Decision-Making Framework: Linking Forecasts to Action

Forecasting models must be embedded in a decision-making framework that translates predictions into practical actions. This involves linking the forecasted demand with operational decisions such as resource allocation, production scheduling, and workforce planning.

Key elements of this framework include:

- **Objective:** Minimize operational costs while maintaining enough capacity to meet demand.
- **Constraints:** Production schedules, labor availability, machine capacity, and costs associated with over- and under-capacity.
- **Decision Variables:** Resource levels (FTEs, machine hours) adjusted based on forecasted demand.

By structuring the forecasting model as part of a larger decision-making process, organizations can shift from reactive to proactive operations, ensuring that capacity planning aligns with future demand trends. This shift reduces last-minute adjustments, cuts costs, and enhances overall operational performance.

6.4 Validation: Cost Simulation of Production Capacity Strategies

While an improved forecast alone does not guarantee cost reductions, its value lies in providing accurate inputs for capacity planning decisions. The validation focuses on how well these decisions align with actual production needs, translating forecasts into actionable production capacity strategies.

The purpose of this validation process is to assess the effectiveness of the forecasting models by simulating production capacity strategies and comparing costs. The goal is to determine how well the forecast-driven strategies minimize costs associated with under- and over-capacity, thereby informing better capacity planning decisions.

- **Objective:** Minimize total costs while maintaining sufficient production capacity.
- **Cost Structure:**
 - Under-capacity costs are set 3x higher than over-capacity due to the disruption caused by late or emergency orders, while overcapacity only costs the salary of the operator.
 - Costs associated with hiring and releasing employees are included, with penalties for frequent changes in staffing to discourage short-term adjustments.

Using the forecast as input, the optimization model tests various capacity levels and chooses the configuration that minimizes costs while maintaining the required service level.

6.4.1 Simulation Setup: Rolling Window Forecasting and Simulation

To validate the effectiveness of these capacity strategies, a rolling window forecasting and simulation technique was employed. This method involves continuously updating the forecast as new data becomes available and adjusting the capacity decisions accordingly. The rolling window approach mimics real-world operations, where forecasts are regularly revised, and capacity adjustments are made dynamically.

The simulation tests how well each capacity strategy adapts to updated forecasts, measuring the impact on costs, capacity utilization, and service reliability. This real-time refinement process ensures that production capacity decisions remain aligned with demand trends, avoiding both over- and under-utilization of resources.

6.5 Mathematical Notation: Capacity Planning and Optimization Problem

Objective: Minimize the total cost, which consists of costs related to undercapacity, overcapacity, and employee training. Given the imbalance between undercapacity and overcapacity penalties, the model should ensure minimal disruption while balancing staffing costs.

Let:

- T be the number of months in the planning horizon (for the use case, $T = 8$).
- F_t be the forecasted production minutes for month t , obtained from the forecasting model.
- FTE_t be the Full-Time Equivalent employees needed in month t , to meet production demand.
- C_{under} and C_{over} be the cost coefficients for undercapacity and overcapacity, where $C_{\text{under}} = 3 \cdot C_{\text{over}}$.
- M_{FTE} is the production minutes an FTE can complete in a month.
- C_{train} is the cost of hiring/training an employee.
- C_{fire} is the cost of releasing an employee.

- U_t and O_t represent undercapacity and overcapacity in month t .
- S_t represents the number of staff employed in month t .

Decision Variables:

- S_t : Number of FTEs for each month t .
- $\Delta S_t = S_t - S_{t-1}$: The change in FTE from the previous month, to account for hiring and firing costs.

Objective Function:

Minimize total cost:

$$\text{Minimize } \sum_{t=1}^T (C_{\text{under}} \cdot U_t + C_{\text{over}} \cdot O_t + C_{\text{train}} \cdot \max(0, \Delta S_t) + C_{\text{fire}} \cdot \max(0, -\Delta S_t))$$

Where:

$$U_t = \max(0, F_t - S_t \cdot M_{\text{FTE}})$$

represents the undercapacity for month t , and

$$O_t = \max(0, S_t \cdot M_{\text{FTE}} - F_t)$$

represents the overcapacity for the month t .

Constraints:

1. **FTE Step Constraint:**

$$S_t \in Z^+, \quad \forall t = 1, \dots, T$$

2. **Production Capacity Constraint:** For each month, the FTEs must provide sufficient production minutes to meet forecasted demand:

$$S_t \cdot M_{\text{FTE}} \geq F_t \quad (\text{minimize undercapacity penalty}).$$

3. **Smoothing and Hiring/Firing Costs:** The optimization also includes a penalty for frequent changes in S_t :

$$|S_t - S_{t-1}| \quad (\text{penalty on hiring and firing}).$$

4. **Strategy-Specific Constraints:**

- **Strategy 1:** Uses forecast three months ahead:

$$S_{t+3} = \left\lceil \frac{F_{t+3}}{M_{\text{FTE}}} \right\rceil$$

- **Strategy 2:** Uses a smoothed forecast:

$$S_{t+3} = \left\lceil \frac{F_{t+3} + F_{t+4} + F_{t+5} + F_{t+6}}{4 \cdot M_{\text{FTE}}} \right\rceil$$

- **Strategy 3:** Adds a buffer to Strategy 2:

$$S_{t+3} = \left\lceil \frac{F_{t+3} + F_{t+4} + F_{t+5} + F_{t+6}}{4 \cdot M_{\text{FTE}}} \right\rceil + B$$

Chapter 7

Research Question 4: How is the ASML Field Service Hub case applied to the general model?

7.1 Introduction

In earlier chapters, a general forecasting and capacity planning model was developed to address the challenges posed by uncertain demand and complex production flows. While this model is designed to be applicable across various industries, its practical utility can only be fully understood when applied to a real-world scenario. The ASML Field Service Hub serves as an ideal use case for testing and validating the model's effectiveness.

ASML operates within the high-tech semiconductor industry, which is characterized by intricate service operations and highly variable demand for spare parts and machine maintenance. This complexity, coupled with the need for precise capacity planning, makes the ASML FSH an excellent test case for demonstrating how the general model can be adapted to real-world operational needs.

The aim of this chapter is to explore how the general forecasting and capacity optimization model can be applied to ASML's unique context. By focusing on ASML's specific operational challenges, this chapter will show how the general principles discussed earlier can be fine-tuned to handle the intricacies of ASML's Field Service Hub, while also providing insights into the broader applicability of the model.

7.1.1 Adapting the General Model to ASML

The next sections of this chapter outline the specific modifications made to the general forecasting and capacity planning model to suit the ASML Field Service Hub's needs. Customizations were required to handle ASML's unique data structure, product flows, and operational constraints.

The final model was developed in Python, combining both the forecasting and production capacity optimization models. The process begins by generating a forecast using one of three forecasting methods; XGBoost, Random Forest, or ARIMA, based on the prepared data. This forecast is then used in the production capacity optimization model, which performs an exhaustive search to determine the optimal number of FTE required for each month, with the goal of minimizing total operational costs.

The model was designed to improve capacity planning accuracy and reduce costs by balancing the trade-offs between overcapacity and undercapacity. The model selection and structure were guided by the requirements outlined in previous chapters (subsection 4.4.1).

7.2 Specific Challenges of the ASML Field Service Hub

ASML, founded in 1984, is a Dutch corporation and one of the world's leading manufacturers of chip-making equipment. They design and manufacture lithography machines that are essential for chip manufacturing. ASML has now become a global innovation leader, and they have expanded immensely.

The ASML Field Service Hub presents unique challenges that make it a valuable case for testing the general model. These challenges include high product complexity, unpredictable demand for spare parts, and stringent service level agreements.

7.3 Data Preparation

Data preparation was a critical step in ensuring that the forecasting model could deliver reliable predictions. The focus here was on identifying and preparing data that could be used to create a robust forecasting and optimization model.

Data preparation started with transforming raw data from the source into a structured format and remediating the data. Data remediation involved extensive cleaning and restructuring to ensure the accuracy and reliability of the dataset. First, the trend and data were analyzed to ensure data quality. Then the mistakes that were found, were rectified. This was done in coordination with production planners and data scientists from ASML. Most of the data remediation was done in Excel in CSV files, but some minor steps were undertaken in Python. Key actions taken during remediation and preparation included:

- Elimination of unrealistic peaks.
- Eliminating missing values, empty orders, and inconsistencies.
- Removal of duplicate records and non-production orders.
- Looking at the manually overwritten data and see if those changes were correct and removed lots of those manual workarounds.
- Adjustments for norm times and manual workarounds that had previously distorted the dataset.
- Disregarding the first few inputs, as they were not reliable.
- Date conversion to prepare the data for the different aggregation levels.
- Transformation of the panel data set into a usable time-series data set to prepare the data for machine learning input. This was done using Excel Pivot tables and ensured that data was ready for aggregation levels such as daily, weekly, and monthly. The decision to use time-series data instead of panel data was made because of the advantages of the models available for time-series data. For future research, it might be interesting to create a neural network that uses the panel data as input.
- An index was added for machine learning input; machine learning models cannot use the month format, so the new index gives the linearization and time dependency. Later, in subsection 8.4.5, it is shown that this feature has a big impact on the model.
- Data from multiple timeframes was shifted forwards and backward (e.g., -2, -1, +1, +2, +3, +4, +5, +6, +7, +8 months) to create lag features and test data for the machine learning models. This was also required for the rolling window approach to include historical data from previous months to capture temporal trends.

The train-test split was performed by using 17 months of data as training data and the last 8 months for testing. This rolling window allowed for an ongoing evaluation of model performance based on recent data.

Forecasting Out of Sample: Include Expected Growth

To generate valuable insights for ASML beyond the historical data, the forecasting model must demonstrate strong performance not only on past observations but also in predicting future demand. This involves incorporating anticipated growth at FSH into the model, enabling more accurate long-term planning.

7.3.1 Data Aggregation Level

Data was aggregated at various levels; daily, weekly, and monthly, to explore different patterns and trends. Daily data was highly erratic, while weekly data aligned more with ASML's planning cycles. Ultimately, monthly aggregation was selected as the most suitable for the model due to its stability, predictive capability, and the requirements for production capacity planning.

Daily Aggregation Level

The data is erratic on a daily aggregation level and this makes it hard to find the trend, Prophet (Python library created by Meta, a state-of-the-art daily forecasting model) was used to find trends and look at the data at a daily aggregation level.

Weekly Aggregation Level

The weekly aggregation level was investigated because ASML plans on a weekly level; the current benchmark is on that weekly level. Weekly is less erratic, but still not very predictable.

Monthly Aggregation Level

The monthly aggregation level was ultimately used in the forecasting and optimization model due to its stability, predictive capability, and the requirements for production capacity planning. It is smoother and thus easier to predict. It also aligns better with the goal of production capacity optimization, where it is required to look 6 months ahead.

7.4 Analysis

The analysis of the Field Service Hub data revealed several key trends and insights that were critical in shaping the forecasting model.

7.4.1 Standard Data Checks

Several standard quality checks were performed on the dataset to ensure reliability and suitability for machine learning modeling. These included:

- **Counts and duplicates:** Ensured consistency in the dataset by removing or resolving duplicate records.
- **Descriptive statistics:** Computed the mean, standard deviation, minimum, and maximum values for key metrics and features.
- **Handling missing values:** Examined the number of zero and non-zero entries to assess data completeness.
- **Autocorrelation:** Evaluated the internal dependencies within the time-series data.
- **Data smoothness:** Classified the time-series data per aggregation level to check whether it was smooth, lumpy, intermittent, or erratic. This characterization helped in selecting appropriate forecasting methods.

7.4.2 Conclusion from Data Analysis

The data analysis indicated that the available dataset is robust enough to support a machine learning-based forecasting model. However, due to the limited amount of historical data, it was concluded that large neural network models would not be feasible. Simpler machine learning models, such as decision trees or gradient boosting, would be more appropriate for this scale of data and can give better results than parametric models or the current benchmark.

7.5 Data Challenges and Limitations

The analysis encountered several challenges and limitations, primarily around data quality and consistency:

- **Data quality:** The dataset had some issues, such as unrealistic peaks and missing values.
- **Real-time updates:** Data was not available in real-time, which limited the model's ability to respond to sudden changes in demand.

- **Limited historical data:** Only two years of historical data were available, which posed a challenge in identifying long-term trends.
- **Manual adjustments:** Many manual workarounds were found in the data, leading to aggregation errors. These need to be resolved to ensure data accuracy.
- **Order conversion:** The conversion from order lines to production minutes was based on old data, which may not reflect the current operating environment.

These challenges highlight the importance of ongoing data remediation efforts to ensure that future forecasting models can rely on higher-quality, more consistent datasets.

7.6 Application of Forecasting Models to ASML Data

7.6.1 The design of the forecasting model

The forecasting model was designed to utilize both the historical data and the expected growth forecasts provided by ASML. To incorporate future growth, a factor was applied to each month based on production capacity and expected machine installations. This approach was structured to address the need for accurate, forward-looking predictions that could support more efficient production capacity planning.

A critical decision was made regarding the timeframe used for modeling. Different machine learning methods require varying time frames as input. Three approaches were tested:

1. **Shortest Timeframe:** Using only the previous time step's data to forecast.
2. **Longer Timeframe:** Including one more time step back for the total minutes, but not all other features.
3. **Longest Timeframe:** Adding data from multiple previous time steps, including other features, enabling the model to capture temporal dependencies more effectively.

A trade-off exists between adding more data points and risking data overload, which could cause overfitting or run into the problem of not having enough training data because there are too many features. It was tested that the third option was not viable, so for the final results, only the first two options, Shortest Timeframe and Longer Timeframe, will be compared in depth.

Additionally, multiple forecast windows were created for different future months, which could be leveraged in a multi-output regression model to predict future demand for several months at once. In the final model, a 6-month window was used because larger windows did not add more information, but were less reliable.

7.6.2 What methods are best used to forecast demand for these flows?

The final forecasting model supports three methods: XGBoost, Random Forest, and ARIMA. These models were chosen based on their performance in forecasting, their ability to handle time-series data, and their fit with the available dataset based on the requirements in subsection 4.4.1.

7.6.3 Technical Specifications

Data preprocessing was conducted outside the forecasting model, as detailed in section 7.3. The final CSV file used for the forecasts contains cleaned and preprocessed data.

The key features used in the forecasting model include the data per product flow, the underlying time-series production data of FSH, the index, and the total minutes. These features were tested over different timeframes to determine the optimal configuration. After testing, a 6-month forecast horizon was chosen, as it provided a balance between accuracy and relevance to ASML's planning needs, using the three months it takes FSH to upscale or downscale production.

The model employs Multi-Output Regression, which performed better than chain regression in testing. Multi-output regression was implemented for XGBoost and Random Forest, while ARIMA utilized chain regression since it works with the time-series of total minutes without other features.

Each model follows the same training and testing process, where 1.5 years of data were used for training, and the final 8 months were reserved for testing.

XGBoost

XGBoost was optimized using grid search, which tested various parameter combinations to find the best fit for the data. The model was designed to handle the multi-output nature of the forecast, with each future time point treated as a separate regression task.

Random Forest

Random Forest followed the same setup as XGBoost, using multi-output regression and optimized through grid search. It was tested on the same features and timeframes and provided a robust, non-linear method for forecasting multiple time steps.

ARIMA

For ARIMA, the model parameters were optimized using an exhaustive grid search for combinations of p , d , and q values. Interestingly, ARIMA(1,0,0), which closely resembles a Naive Forecast, provided the best results for this dataset, especially in multi-step forecasting. While ARIMA only uses the total minutes time-series, it produces an almost straight-line forecast from the starting point of each prediction.

Other variants of ARIMA are unstable and not fit for forecasting multiple steps. Some other versions sometimes closely follow the realized value but are still worse than the (1,0,0) version for RMSE and MAE.

7.6.4 Forecast Process

The model generates an $M \times N$ matrix of forecasts, where M is the number of months forecasted and N is the number of test items. This matrix allows for a comprehensive comparison between forecasted and realized values, enabling the calculation of errors such as MAE and RMSE. The forecast results are also compared against ASML's benchmark to assess their alignment with existing capacity planning tools.

7.6.5 Forecasting Results

The results of the Forecasting Model are discussed in section 8.4 to immediately compare them to the benchmark and to also answer Research Question 5: How are the models validated and compared to benchmarks?

7.7 Application of Capacity Planning Model to ASML Data

To translate the forecast into decisions for production capacity the general model needs to be applied to the prepared ASML data. As explained before there are three strategies implemented:

Strategy 1: Follow Forecast 3 Months Ahead

Adjust FTE levels based on the forecast exactly three months ahead, without smoothing fluctuations. This strategy waits until the last possible moment to change the scale of operations but will outperform the other two strategies when the forecast is perfect.

Strategy 2: Smoothing via Multiple Forecasts

Use the average of forecasts across different time horizons to stabilize decisions and avoid large swings in capacity planning. Here the mean of the forecasts of months three, four, five, and six are used.

Strategy 3: Buffer + Smoothing

A buffer is added to the smoothed forecast to account for unexpected demand surges while preventing under-capacity.

After generating the forecast, the model calculates the required production minutes and FTE for each month based on one of the three implemented capacity strategies. Initially, the forecasted FTE values may be erratic because the model treats each month independently. This means that for each month, it chooses a starting value for the exhaustive search which is the forecasted amount of minutes divided by minutes of production per FTE. To address this erratic outcome, the capacity optimization model smooths the results to avoid excessively high training costs caused by erratic hiring and firing decisions.

The optimization model performs an exhaustive search around the forecasted FTE values. Given that the problem is discrete (per 1 FTE) and the possible solutions are close to the non-optimized values, the exhaustive search is computationally efficient, typically taking no more than three minutes per run.

Later, when all calculations were done and all model outcomes tested, the optimal starting values for the exhaustive search for each run were hardcoded in the model, so that those three minutes could be saved every time the model was run.

The model checks solutions both above and below the initial forecasted FTE and ensures the optimal solution is found within the search bounds. Since the optimization problem does not have local minima, the solution found is the global optimum if it is not at the bounds of the test set. The model also compares different strategies, such as smoothing the forecast or adding buffers, to the other models and ASML's current benchmark.

The total solution is optimized using a cost function, which includes penalties for undercapacity, overcapacity, and training costs. The model then outputs the final solution along with the corresponding cost metrics and personnel changes.

7.7.1 How to Translate the Forecast into Decisions for Production Capacity?

To convert forecasted demand into production capacity decisions, the following process was used:

1. **Order Line Conversion:** Forecasted orders were converted into production minutes using the average minutes per order.
2. **FTE Calculation:** This was translated to the number of required FTE.
3. **Capacity Adjustment:** Production capacity is adjusted in discrete steps based on the forecasted need, with optimization focused on minimizing under- and over-capacity costs.

The results of the Capacity Planning Optimization are discussed in section 8.5 to immediately compare them to the benchmark to also answer Research Question 5: How are the models validated and compared to benchmarks?

7.8 Conclusion of the Use Case

The final model integrates forecasting with production capacity optimization, providing ASML with a robust tool to plan workforce and production more effectively. The model's ability to balance forecast accuracy with cost minimization makes it a valuable addition to ASML's capacity planning processes. While more complex models such as deep learning were tested, they were not included in the final solution due to the limited data available. Ultimately, this forecasting and optimization model provides a scalable solution for future growth and fluctuating demand while keeping production costs under control that can be extended to other use cases as well.

7.9 Insights for Broader Application

The application of the general forecasting and capacity optimization model to the ASML Field Service Hub has provided valuable insights, not only for ASML but also for other industries facing similar challenges in managing uncertain demand and optimizing production capacity. While ASML operates in the semiconductor industry with its specific set of operational complexities, the lessons learned from this use case demonstrate the model's flexibility and robustness across various sectors, enabling a shift towards more proactive and data-driven operations.

7.9.1 Flexibility of the Model

One of the most significant takeaways from the ASML use case is the model's adaptability to different demand patterns and operational constraints. Importantly, because the forecasting model is based on machine learning, a new model will be trained for each specific use case. This means that the specific features used at ASML, such as data per product flow, are not limiting factors. For any other organization, as long as relevant underlying data is available, the model can be trained to perform effectively.

For ASML, this data was essential for capturing the demand flow and spare part planning. However, in other industries, different types of data might play a more critical role. The model's performance is tied not to specific features like product flow data but to the availability of quality data that reflects the underlying drivers of demand. As long as an organization has access to its operational data, whether it be sales forecasts, inventory levels, or production schedules, the machine learning model can be adapted to forecast demand and optimize capacity effectively.

For example, in the aerospace industry, the data might be replaced by aircraft maintenance schedules or parts failure rates. Similarly, in automotive manufacturing, the model could leverage production line schedules, customer orders, and parts availability to forecast demand. This flexibility ensures that the model can be adapted to different industries and data environments.

7.9.2 Data-Driven Decision-Making

One of the key outcomes of applying the model at ASML was its ability to drive data-driven decision-making. By leveraging machine learning models and advanced optimization techniques, the model provided actionable insights that led to more informed decisions about production capacity, resource allocation, and labor management. This approach can be generalized across industries where operational decisions are often based on gut feeling or limited historical data.

Chapter 8

Research Question 5: How are the models validated and compared to benchmarks?

8.1 Introduction

The validation of forecasting models is a crucial step in determining their real-world applicability and performance. In this chapter, the focus is on validating the proposed forecasting models by simulating their outputs in combination with an optimization model on realized data from the ASML Field Service Hub. The goal of this validation is to assess which forecasting model when used in combination with capacity optimization, results in the lowest total operational cost.

To evaluate the performance of the forecasting models, two key metrics are used: Mean Absolute Error and Root Mean Square Error. MAE provides a measure of the model's overall accuracy, focusing on average error across all predictions. This gives insight into how closely the forecasted demand aligns with actual demand. RMSE, on the other hand, penalizes larger errors more heavily, making it a suitable metric for evaluating the model's handling of extreme outliers. Together, MAE and RMSE provide a comprehensive view of the model's performance in both general cases and scenarios where deviations from the forecast could lead to significant operational impacts.

Additionally, the validation process includes a comparison of the proposed models against ASML's current benchmark. This benchmark is ASML's existing tool for forecasting and capacity planning and serves as the baseline for evaluating improvements. The validation begins by comparing the forecast accuracy of the new models (XGBoost, Random Forest, ARIMA) to the benchmark's predictions.

Next, the explainability of the models is assessed using SHAP values, which provide insights into how the models make predictions by showing the contribution of each feature. This analysis is important for understanding the drivers behind the forecasts and ensuring that the models are interpretable and actionable for decision-makers.

Finally, robustness testing is conducted to determine how well the models perform under uncertain conditions, by introducing noise into the input data. This step ensures that the forecasting models can handle fluctuations in demand and still provide reliable predictions.

The results of the validation process are presented through cost simulations, which integrate the forecasted demand into a capacity optimization model. This allows for a direct comparison between the models based on their total operational cost, factoring in overcapacity, undercapacity, and workforce allocation. The outcomes of these simulations are then compared to the benchmark's performance, providing a benchmark for assessing improvements in both forecast accuracy and production capacity optimization.

In summary, this chapter will evaluate the forecasting models based on accuracy (MAE and RMSE), explainability (SHAP values), robustness, and their ability to minimize total costs when integrated with a capacity optimization framework. By comparing these results with ASML's existing system, the aim is to demonstrate the advantages and limitations of each model in real-world operations.

8.2 Performance Metrics

The final model's performance was evaluated using a range of metrics that assess both the accuracy of the forecasting models and their impact on capacity planning decisions, with a focus on minimizing total operational costs. These metrics are divided into two main categories: forecasting accuracy metrics and validation/simulation metrics related to cost and capacity optimization.

8.2.1 Forecasting Accuracy Metrics

To measure how well the forecasting models predicted future demand, the following performance metrics were employed:

Mean Absolute Error

MAE calculates the average magnitude of errors between the forecasted and actual values, without considering direction. Lower MAE values indicate more accurate forecasts. This metric is particularly useful for understanding overall forecast performance, as it provides insight into the general alignment between forecasted demand and realized demand.

Root Mean Square Error

RMSE gives more weight to larger errors by squaring the residuals, making it sensitive to extreme outliers in the forecast. This metric is crucial for assessing the reliability of forecasts in extreme conditions, where large deviations from actual demand could significantly impact operational decisions.

Together, MAE and RMSE provide a comprehensive view of forecast accuracy, helping to assess how well the models handle both average errors (via MAE) and more extreme outliers (via RMSE).

8.2.2 Validation and Simulation Metrics

After generating the forecasts, they were integrated with the capacity optimization model to determine the overall efficiency and cost-effectiveness of production decisions based on those forecasts. The following metrics were used to evaluate the outcomes of the simulation:

Expected Cost in Minutes Based on Forecast

This metric represents the cost that would have been incurred if the forecast had been perfectly accurate. While it does not directly impact the outcome, the optimization model tries to minimize this before knowing what the realized demand is.

Realized Cost in Minutes Based on Forecast

This metric calculates the actual cost incurred based on the realized data and the chosen forecasting strategy. Lower realized costs indicate better performance, as this reflects the total minutes of production, including any penalties for under- or overcapacity. ASML measures cost in production minutes, as order times are tracked in minutes. Although future projects may convert these values into euros, this study uses minutes as the cost metric.

Expected Production Minutes Missed / Late Orders with this Strategy

Based on the forecasted demand, this metric shows how many minutes of undercapacity (i.e., missed production) were anticipated. Though not a direct indicator of final performance, this metric provides a reference point for comparison with the realized data.

Realized Production Minutes Missed / Late Orders with this Strategy

This metric tracks how many minutes of production were missed due to undercapacity, based on the actual, realized demand. Lower values indicate better performance, as this reflects the ability of the model to meet demand and avoid late orders or production delays.

Total FTE Used Over 8 Months

This metric measures the total number of FTE employed across the entire simulation period. Lower values indicate more efficient staffing, as fewer workers are needed to meet the forecasted demand. However, if this is too low, the costs will increase due to undercapacity.

Average FTE Used

The average number of FTEs employed over the 8 months. This is a measure of how consistently the model allocates labor resources. Lower average FTE values indicate greater efficiency, as fewer workers are required to meet production needs. However, if this is too low, the costs will increase due to undercapacity.

Total People Trained

This metric counts the number of employees hired over the simulation period, starting from a predetermined level. Lower values are generally preferred, as they indicate less reliance on hiring new employees, which incurs additional costs and operational inefficiencies.

Total People Released

The number of employees let go over the simulation period. While releases can be beneficial, this metric needs to be interpreted in combination with undercapacity performance. More releases should not come at the expense of meeting production demand.

Volatility in Personnel Change

This metric aggregates the number of hires and releases to assess the overall volatility in staffing decisions. Lower volatility is better, as it indicates more stable workforce management, avoiding the costs associated with frequent hiring and firing.

8.2.3 Cost-Based Performance Comparison

The primary focus of the validation was to determine which combination of the forecasting model and capacity planning strategy minimized total operational costs. The total cost is broken down into the following categories:

Undercapacity Costs

These represent the penalties incurred from failing to meet production demand, resulting in missed or delayed orders. In this context, undercapacity costs are significantly higher than overcapacity costs (three times as expensive), due to the critical importance of on-time service delivery at ASML.

Overcapacity Costs

These reflect the cost of excess production capacity that is not utilized, including idle labor (or machinery). While overcapacity is less expensive than undercapacity, it still represents inefficiency that should be minimized.

Training and Release Costs

These include the costs associated with onboarding new employees (training) and letting workers go (release costs). For this use case, release costs were set at zero because of the large number of operators with flex contracts, but the model does have the option to include this for other use cases. Three months for new hires is included, during which no production is achieved but the employee is paid. High turnover (frequent hiring and firing) drives these costs up, so stable staffing is incentivized.

8.3 Comparison Against ASML’s Existing Benchmarks and Models

8.3.1 Benchmark

ASML currently uses a simpler model based on future outlooks, that is generated once a year, not based on sophisticated models or historical data. This forecast is based on predicted future sales per product. This model outlines expected production (capacity), and serves as a key benchmark for the model. The model is based on a weekly aggregation level, but is for this research calculated towards minutes per month, based on the number of order lines realized multiplied by the average minutes per line and the number of (working) weeks in the month.

To validate the forecasting models, simulations were run to compare the outcomes of the model-driven decisions against ASML’s existing benchmark.

ASML’s Current Model vs. Forecasting Models

ASML’s current model was used as a benchmark to compare the performance of the forecasting-driven capacity planning strategies. By running simulations over the same eight-month period, the following aspects were analyzed:

- **Error Comparisons:** Which model resulted in the lowest MAE and RMSE?
- **Cost Comparison:** Which model resulted in the lowest total cost (including overcapacity, undercapacity, and staffing)?
- **Explainability:** How easy is it to interpret and explain the decisions made by each model?
- **Robustness:** How sensitive are the models to changes in input data?

8.4 Results of the Implemented Forecasting Models at FSH

This section presents the results of the forecasting models applied to improve production capacity planning at the ASML Field Service Hub. The three primary models, XGBoost, Random Forest, and ARIMA, were evaluated across various strategies and timeframes, focusing on their accuracy and the impact of these forecasts on cost efficiency and personnel management.

These models were tested using different strategies and across two timeframes. The total error comparison is calculated over 30 points (5 forecast times 6 steps ahead, until the last point the realized demand is known for 6 months ahead), and the strategies are compared for the 8 months of test data. The strategies were applied to assess how each model performs in both short-term and extended forecasts. In total, five forecasts were made, each predicting six months ahead, for a total of 30 forecasted points for the forecasting models. Additionally, the benchmark was included for comparison. Later, the results of the production capacity optimization model will be shown.

8.4.1 Error Comparison

The error comparison is just an intermediate result. The most important outcome to compare is cost, which will be investigated in section 8.5.

Error Comparison between Different Models

Comparing the forecasting accuracy of each model provides insight into which approach is most reliable:

- **XGBoost:** XGBoost demonstrated the best performance in terms of error reduction, particularly when using the Longer Timeframe. Strategy 1 performed notably well, achieving the lowest MAE among all models with a total reduction of 37.4% against the benchmark. XGBoost’s accuracy three months ahead (a critical period for scaling) was particularly strong, although this may be partly due to overfitting to the data, making it difficult to make assumptions about its future robustness.
- **Random Forest:** The Random Forest model also performed well but did not outperform XGBoost for MAE. When using the Shorter Timeframe, it achieved respectable accuracy, but including additional data inputs did not result in a significant performance boost. Random Forest exhibited

a slightly higher MAE compared to XGBoost, though it was still significantly more accurate than the benchmark. Random Forest achieved the best RMSE (when using Strategy 3), with a reduction of up to 28.3% compared to the benchmark.

- **ARIMA:** ARIMA had the highest MAE and RMSE among the three models, underperforming compared to both XGBoost and Random Forest. ARIMA struggled to handle the complexity of the data, especially when forecasting more than three months, and consistently performed worse than the machine learning models. However, it still outperforms the benchmark considerably.

Error Comparison between Different Strategies

The forecasting strategies used to interpret the model predictions significantly impacted the final results:

- **XGBoost:** For XGBoost, Strategy 1 outperformed both Strategies 2 and 3, consistently resulting in lower MAE and RMSE. Strategy 3, which adds a buffer to the forecast, surprisingly performed better than Strategy 2 (no buffer), highlighting a potential bias in the forecasting model.
- **Random Forest:** In contrast to XGBoost, Strategy 1 was the least effective for Random Forest, while Strategy 3 performed the best. This suggests that different strategies should be applied depending on the forecasting model in use.
- **ARIMA:** All strategies performed worse than the total error comparison for ARIMA. This is likely due to the simplistic nature of ARIMA, which struggled to capture the necessary complexity in the forecasted demand, particularly at critical points for scaling capacity. Since the ARIMA setup is almost a Naive Forecast, it will not perform well in predicting the necessary three months.

8.4.2 XGBoost rolling window forecast

The first two months this works almost perfectly. Unfortunately, the third month is out of sync, but the fourth and fifth months are perfect again. The sixth is a bit off. The seventh is perfect, and the eighth is not perfect, but still a lot better than the benchmark.

In total, a 37.4% reduction in MAE forecast error, as well as 20.1 % for RMSE, against the benchmark was achieved.

8.4.3 Random Forest, shorter timeframe

The best performing Random Forest model, based on MAE, is with the Shorter Timeframe and Strategy 3.

8.4.4 Forecasting Out of Sample

In addition to in-sample forecasts, the models were adjusted to predict demand beyond the available data to create new insights for ASML. The models incorporated an expected growth factor to account for future demand increases, as explained before in section 7.3.

8.4.5 Explainability in the Use Case

In this thesis, SHAP values were applied to the XGBoost and Random Forest models. These models are typically considered black-box models due to their complexity and non-linear interactions. SHAP provides a valuable tool to open this black box, offering interpretability without sacrificing accuracy.

It is important to note that SHAP cannot be applied to models like ARIMA, which operate purely on time-series data without feature inputs. ARIMA forecasts are based solely on historical values of one time-series input, making SHAP's feature-based explanations irrelevant.

SHAP XGBoost

The SHAP values for the final model, using XGBoost and the shortest timeframe, were investigated because that model had the lowest error and cost and thus the best result (when used in combination with Strategy 1).

For the XGBoost model, the SHAP values indicated that the most important feature was the index of the time series. However, it was surprising that the total production, rather than individual features,

played a smaller role in the model’s predictions. It can be seen that only a few features are used extensively in the forecasting model. This inequality between features makes the model hard to explain. For example, if the features with lower SHAP values decrease or increase drastically, the model would not forecast it differently. This leads to this version of the model only working if the mix of product flows stays at the current level.

Another finding is that the most important features used in this model are not necessarily the largest product flows.

SHAP Random Forest

The SHAP values were calculated for the Random Forest model that had the lowest MAE and RMSE; the version with the shortest timeframe input and Strategy 3.

In this Random Forest model, the SHAP values revealed a more balanced use of features compared to XGBoost. Random Forest used almost all input features to some extent, which makes sense since the output should be a combination of all product flows and the total, with an increase over time using the index. This makes the Random Forest model more explainable than the XGBoost model.

It can also be seen that the Random Forest model has fewer outliers than the XGBoost model. This extra stability (better RMSE, but not MAE) that the Random Forest model has, may originate from the better balance between features.

Feature Importance and Future Robustness

Another advantage of investigating the feature importance of machine learning models is that it is another robustness check (Mishra et al., 2021). A more balanced feature use creates more robust models.

For example, for the XGBoost model, if the features with lower SHAP values decrease or increase drastically, the model would not forecast it differently. This leads to this version of the model only working if the mix of product flows stays at the current level. This does not only impact the explainability of the model but can also mean that the robustness for future forecasts is not great if the demand for the product mix changes. This unexpected result warrants further investigation in future research.

The Random Forest model’s use of all input features could also indicate greater future robustness compared to XGBoost due to the possibility of the mix of product flows changing. Even if the mix changes, the Random Forest model still uses all features and can thus probably still forecast accurately.

In conclusion, incorporating SHAP values into this analysis provided essential insights into model behavior, ensuring that the forecast-driven strategies are not only accurate but also interpretable and robust. It can be concluded that the Random Forest model is more explainable than the XGBoost model because of the balanced use of features.

8.4.6 Robustness in the Use Case

To assess the robustness of the models, a jitter test was conducted. This test involves adding varying levels of noise to the input data and evaluating the model’s error metrics, such as MAE and RMSE. Gaussian noise was introduced at levels of 5%, 10%, and 15% to simulate possible future fluctuations in data, in line with similar robustness studies (Cai et al., 2021).

The test was performed on all three models: XGBoost, Random Forest, and ARIMA, using both the Shorter and Longer timeframes. To thoroughly assess the robustness, the jitter test was applied over 30 different points, encompassing five forecasts across six months for each model. This ensured that the models were evaluated across a wide range of scenarios and forecast horizons.

Gaussian noise was applied to the input data set to simulate the kind of variation expected in future data. The level of noise is adjusted to assess how the model behaves under small (5%), moderate (10%), and large (15%) deviations from the original data. Gaussian noise is ideal for this purpose, as it mimics random fluctuations and introduces smooth variability across all features.

The key question addressed by this test is: How much does the introduction of noise impact the model’s predictions? Models that show only minor increases in error under noisy conditions are considered more robust, indicating that they can generalize better to unseen or fluctuating data. In contrast, models that exhibit significant increases in error when noise is added may be overfitting the training data, rendering them less reliable in practical applications.

The robustness tests confirmed that the models can handle noisy data. For example, after introducing 15% Gaussian noise, the Random Forest model's MAE only increased by 1.5%, maintaining stable predictive accuracy under fluctuating conditions.

Key findings included:

XGBoost Sensitivity

XGBoost was the most sensitive to noise compared to Random Forest and ARIMA, showing larger increases in both MAE and RMSE as the noise level increased. This suggests that XGBoost, while effective in clean datasets, may struggle to maintain accuracy under uncertain conditions.

Random Forest Resilience

Random Forest demonstrated better robustness, maintaining a relatively stable performance even as noise levels increased. Its performance was more balanced, possibly due to the model's ability to aggregate predictions from multiple trees, making it less susceptible to fluctuations in the input data.

ARIMA Stability

ARIMA, though generally less accurate compared to the machine learning models, also showed good resilience to noise. Its simplicity might contribute to its stability, as it relies solely on the time-series structure, without additional input features.

Unexpected Performance Boost

Interestingly, in some cases, the models performed better with noise. This could be due to overfitting in the original dataset, where slight noise helps generalize the model better. However, more testing and future research are required to investigate this in-depth.

Impact of Increased Noise

As expected, higher noise levels generally resulted in lower accuracy across all models. The errors increased with the level of noise, confirming that extreme deviations in the input data can lead to significant forecasting challenges.

The jitter test only evaluated MAE and RMSE, focusing on forecast accuracy rather than the resulting costs. While cost impacts are crucial for capacity planning, the primary concern in robustness testing is the reliability of the forecast itself.

8.5 Validation: Cost Simulation of Production Capacity Strategies at ASML FSH

In this section, the performance of different forecasting models is analyzed based on key metrics for production capacity planning, including realized costs, production delays, workforce utilization, and personnel management. The goal of the cost simulation is to validate that the improved forecasting models also lead to lower cost, and are not just a forecast error reduction.

For each forecasting model, the best strategy, based on total cost, is further investigated in this report. The four strategies that are evaluated: XGBoost with a Shorter timeframe (Strategy 1), Random Forest with a Shorter timeframe (Strategy 3), ARIMA (Strategy 3), and the benchmark.

The first conclusion here is that these strategies are not just the strategies with the lowest cost, but also the ones with the lowest errors.

These four models are compared against each other based on:

1. Realized Cost in Minutes
2. Undercapacity: Realized Production Minutes Missed / Late Orders
3. Overcapacity: FTE Used
4. Volatility in Personnel Change

As explained in chapter 6, the model starts by using the forecasted values of the forecasting model as the input for the simulation. The eight months of input are first translated from minutes to FTE, meaning the forecasted values have to be rounded towards the nearest integer of FTE, because that is the number of minutes that one FTE produces each month, which makes the optimization model discrete. Then this solution of eight FTE inputs is translated back to total minutes and used for the simulation, where it is smoothed to optimize for total expected cost. Finally, the model provides the optimized number of FTE and minutes.

8.5.1 Realized Cost in Minutes

The realized cost represents the actual production time consumed over the evaluation period. The XGBoost model (Strategy 1) outperformed all other models, making it the most efficient strategy in terms of minimizing production time. It is only slightly worse (2.89%) than the optimal cost. The optimal cost is the cost that could be achieved with perfect knowledge of the demand beforehand, which is a perfect forecast. Both Random Forest and ARIMA strategies exhibited similar realized costs. The benchmark, however, demonstrated significantly higher realized costs, indicating inefficiencies in the current system when compared to the proposed machine learning-based models. This means that using the proposed methods, simulated total costs can be decreased by up to 48.3%.

8.5.2 Realized Production Minutes Missed / Late Orders

In terms of production delays, ARIMA (Strategy 3) demonstrated exceptional performance. By aligning workforce levels with forecasted demand, this model reduced the number of missed production minutes due to undercapacity by 91% compared to the benchmark. This does however come at the cost of a much higher FTE usage. XGBoost, despite its lower realized cost, missed more production, showing higher vulnerability to production delays. The benchmark performed the worst in this category.

8.5.3 Total Monthly FTE Used Over 8 Months

The Full-Time Equivalent metric measures the number of employees required per month to implement each strategy. According to its forecast, the benchmark required the fewest employees. However, despite its efficiency in workforce utilization, its high realized costs and production delays diminish the advantage in personnel usage. The XGBoost and Random Forest strategies used more FTEs, making them more resource-intensive than the benchmark. ARIMA required the most FTEs, although this did not negatively impact its overall performance given its minimal production delays.

8.5.4 Volatility in Personnel Change

Personnel volatility, representing fluctuations in workforce size due to hiring or releasing employees, was highest in the benchmark. This high volatility can create instability in operations, potentially leading to further inefficiencies. XGBoost also exhibited a notable degree of volatility. ARIMA and Random Forest had lower volatility, reflecting more stable workforce management.

In terms of personnel training, ARIMA (Strategy 3) required the most training effort, with the most FTE trained, compared to a little less for XGBoost, Random Forest, and the benchmark. ARIMA and Random Forest did not release any personnel during the eight-month period, which contributed to their stability and low personnel volatility. XGBoost, however, released some employees, leading to a higher level of personnel change. The benchmark released the most employees, further contributing to its high volatility.

The optimized models helped reduce workforce volatility by smoothing the hiring and release processes. The Smoothing via Multiple Forecasts strategy (Strategy 2) provided the most stable workforce levels, reducing the frequency of rapid changes in personnel, which was the exact goal of the strategy.

8.5.5 Key Insights:

- **Lowest Realized Cost:** XGBoost Strategy 1 (only 2.89% above optimum).
- **Best Production Time Missed:** ARIMA Strategy 3
- **Lowest FTE Usage:** The benchmark, though this comes with the highest cost and missed production.

- **Personnel Management:** ARIMA Strategy 3 involved the highest amount of training but did not release any employees, while the benchmark released the most and had the highest volatility in personnel change.

The comparison between these forecasting strategies reveals that while XGBoost (Strategy 1) minimized production costs, it exhibited a higher degree of production delays and personnel volatility. ARIMA (Strategy 3) proved to be the most robust in preventing production delays while maintaining a relatively stable workforce, though it required the highest number of FTEs. The benchmark strategy was outperformed in nearly all categories, showing significant inefficiencies in both production time and workforce stability.

For organizations prioritizing cost reduction, as in this study, XGBoost (Strategy 1) offers the best balance between cost efficiency and operational stability.

8.6 Conclusion

In this chapter, the forecasting models were validated and compared their performance to ASML's existing benchmark. The validation process involved evaluating forecast accuracy, explainability, robustness, and total operational costs through cost simulations that integrated forecasting outputs into production capacity planning. The following key conclusions can be drawn from the validation process:

8.6.1 Forecast Accuracy and Model Performance

The forecasting models, XGBoost, Random Forest, and ARIMA, were compared in terms of their Mean Absolute Error and Root Mean Square Error. While all models demonstrated improvements over the benchmark in terms of general accuracy, XGBoost, and Random Forest outperformed the other models. XGBoost consistently achieved the lowest MAE and RMSE, indicating its strength in managing both overall accuracy and extreme outliers. These models showed clear advantages in predicting demand fluctuations more accurately than the benchmark.

8.6.2 Explainability and Robustness

The use of SHAP values added significant value to the analysis by providing transparency in the decision-making process of the models. SHAP values helped identify the most critical features driving the forecasted demand, which in turn allows (ASML's) stakeholders to understand the factors influencing demand volatility. This level of explainability is crucial for decision-makers to trust and adopt machine learning-driven forecasts.

In addition, the models were tested for robustness by introducing noise into the input data. ARIMA and Random Forest demonstrated higher resilience to noise compared to XGBoost, maintaining their predictive power even when faced with uncertain or fluctuating demand. This highlights the robustness of these models for real-world applications where demand uncertainty is a constant factor.

8.6.3 Capacity Planning and Cost Optimization

The ultimate test of the forecasting models was their ability to drive effective production capacity decisions and minimize total costs. Through cost simulations, which integrated the forecasts with capacity planning strategies, it was clear that the machine learning models (particularly XGBoost) led to substantial improvements in cost performance. These models optimized capacity planning by reducing undercapacity (missed production minutes) and controlling overcapacity (idle resources).

The simulations revealed that the realized costs were significantly lower for the machine learning models compared to the benchmark, with XGBoost providing the most cost-effective solution. By minimizing the mismatch between forecasted and actual production needs, XGBoost and Random Forest reduced the number of late orders, improved resource utilization, and stabilized workforce changes (reduced hiring and firing cycles), thereby controlling personnel volatility.

8.6.4 Comparison to Benchmark Results

The comparison against ASML's system showed that the proposed machine learning models offered clear advantages across multiple dimensions:

- **Improved forecast accuracy:** XGBoost and Random Forest outperformed the benchmark in both MAE and RMSE.
- **Better cost efficiency:** The machine learning models led to lower total operational costs by optimizing capacity decisions and reducing the penalties for undercapacity and overcapacity.
- **Increased workforce stability:** The models reduced the need for frequent hiring and firing, minimizing the associated costs of personnel volatility.

While ASML’s current benchmark has been an effective tool for ASML in the past, these results suggest that integrating advanced forecasting techniques with capacity optimization offers a more robust and cost-effective solution for the company’s future operational needs.

8.6.5 Final Remarks

Beyond the ASML use case, the results of this chapter suggest broader applicability of the proposed models to other industries facing similar challenges of uncertain demand and complex production environments.

Industries such as aerospace, automotive manufacturing, healthcare, and logistics, which also contend with volatile demand patterns and the need for optimized capacity planning, could benefit from adopting these advanced forecasting and optimization techniques. The flexibility of machine learning models to adapt to different data inputs, such as maintenance schedules in aerospace or patient flow in healthcare, means that the core principles demonstrated in this study are transferable across various operational contexts.

The ability of the models to balance forecast accuracy, cost efficiency, and workforce stability makes them well-suited to any industry where demand uncertainty leads to significant capacity planning challenges. Future research could explore adapting these models to other sectors, further validating their robustness and expanding their application.

Chapter 9

Discussion

9.1 Summary of Key Findings

This section discusses the key findings from the research, evaluating the performance of the forecasting models used for demand prediction and their impact on production capacity planning. The robustness and explainability of the models are also investigated in the context of the ASML Field Service Hub, drawing on the insights from the results. Overall, this research aimed to provide a comprehensive approach to mid-term production capacity planning by enhancing forecasting models and optimizing decision-making processes.

9.1.1 Performance of the Different Forecasting Models

The performance of the forecasting models, XGBoost, Random Forest, and ARIMA, demonstrates their applicability across industries facing challenges of demand uncertainty and capacity optimization. XGBoost and Random Forest emerged as the most accurate models, handling both general demand variability and more extreme fluctuations effectively, as evidenced by their superior performance in reducing forecast error (measured by MAE and RMSE). ARIMA, while useful for simpler time-series forecasting, struggled in more complex, multi-feature environments. These findings suggest that machine learning models are particularly well-suited to industries characterized by high volatility and complex data patterns, such as manufacturing and logistics.

XGBoost

XGBoost consistently showed high performance in terms of predictive accuracy, especially for complex, non-linear relationships in the data. It outperformed both Random Forest and ARIMA in certain cases, particularly in short-term predictions with higher fluctuations in demand. However, it also exhibited higher sensitivity to noise when subjected to robustness tests. Despite its relatively higher accuracy, the model's black-box nature poses challenges in explainability. Feature importance metrics using SHAP values did help to explain that.

Random Forest

Random Forest provided a balance between accuracy and explainability. Its predictive power was slightly lower than XGBoost, but it showed greater robustness during jitter tests, maintaining stable predictions even with the addition of noise. Random Forest's performance in mid-term forecasting scenarios was steady, and its relative interpretability made it a more practical option for capacity planners who need to understand the factors driving demand.

ARIMA

ARIMA proved to be the most stable and explainable of the models, making it highly suitable for time series data with linear trends and seasonality. Though ARIMA offers good stability under noise, its simplicity could limit its effectiveness in handling more complex, multi-featured datasets. ARIMA's simplicity and interpretability offer significant value, especially for users who prefer a transparent model.

However, its relatively high errors and limitations in capturing non-linear patterns suggest that this may not be the optimal model.

In summary, each model demonstrated specific advantages depending on the forecast horizon and data complexity. XGBoost delivered the best overall accuracy but struggled with robustness and explainability, while Random Forest provided a balance of robustness and accuracy. ARIMA, though simpler, offered good robustness and explainability, making it a reliable choice for a more stable time series. The choice for the forecasting model depends on the use case, but the optimal forecasting solution for the ASML Field Service Hub according to the used metrics and requirements is the XGBoost model due to the superior results in forecast error and total cost.

9.1.2 Unexpected Outcomes

One of the most surprising results was that introducing noise into the data through jitter tests improved the performance of some models. Normally, adding noise is expected to degrade model accuracy, as it disturbs the underlying data patterns that the models rely on. However, in this case, certain (low) levels of Gaussian noise led to more accurate forecasts in some instances, especially for the Random Forest and ARIMA models. This phenomenon could be attributed to the noise acting as a form of regularization, forcing the models to generalize better and avoid overfitting on the training data.

For example, when the data was slightly perturbed, Random Forest and ARIMA both displayed more stable performance compared to XGBoost. ARIMA, already known for its simplicity and reliance on linear patterns, seemed to benefit from noise, suggesting that its simpler structure might allow it to focus more on the primary trends in the data. On the other hand, XGBoost was more susceptible to noise, particularly with larger perturbations, indicating that while it excels in clean, complex datasets, it struggles with the variance introduced by noise.

These findings challenge conventional assumptions about the impact of noise on predictive accuracy. Rather than uniformly degrading model performance, noise can occasionally help certain models, especially those with strong regularization features like Random Forest, to avoid overfitting and provide more generalized predictions. However, the higher the noise levels, the lower the accuracy in all models, especially in XGBoost, which became increasingly sensitive as noise levels rose.

9.2 Interpretation of Results

The results of this study provide valuable insights into the relationship between forecasting model performance and production capacity planning. The key findings, particularly regarding the models' cost efficiency, accuracy, and robustness, offer practical implications for mid-term production capacity planning. This section interprets the significance of the results concerning production forecasting and the broader context of demand management in complex operational environments both for the ASML FSH and in the more general sense.

9.2.1 Best Performing Models for Cost and Accuracy

The first and most notable conclusion is that the strategies with the lowest realized costs also achieved the lowest forecast errors. For instance, XGBoost Strategy 1, which was trained on a Shorter timeframe, resulted in the lowest realized cost.

Similarly, Random Forest also performed well in terms of both cost and forecast accuracy. Despite having slightly higher realized costs compared to XGBoost, its stability in the presence of noise and its relatively lower errors across multiple scenarios highlight its reliability as a robust forecasting solution. These models' ability to strike a balance between minimizing costs and maintaining accuracy suggests that they are well-suited to the dynamic and complex demand environment faced in some industries, such as at ASML.

A notable finding is the strong performance of ARIMA in minimizing production delays, as evidenced by the lowest realized missed production minutes among all models. ARIMA's strength lies in its focus on linear trends, which aligns well with a relatively stable product flow demand. The model's simplicity allows it to remain resilient in scenarios with limited or moderate noise, making it an excellent choice for minimizing operational disruptions. In comparison to machine learning models like XGBoost, which showed sensitivity to noise, ARIMA performed consistently well under both clean and perturbed data

conditions. This highlights ARIMA’s capacity to provide reliable forecasts, particularly for product flows where demand is less volatile, and a robust model is needed to ensure uninterrupted production schedules.

All three of these models can help reduce the mismatch between supply and demand, thereby minimizing excess costs while maintaining service levels. Their ability to handle large datasets and multiple features (XGBoost and Random Forest) further increases their suitability for complex operations.

9.2.2 Robustness and the Impact of Noise

The robustness tests, which introduced jitter (Gaussian noise) into the data, revealed surprising outcomes regarding how models handle variability in the input data. It is important to note that these robustness results may be more stable for short-term predictions than long-term forecasting scenarios, where error accumulation can lead to significant challenges in accuracy.

The Random Forest model demonstrated resilience to added noise, which helped it maintain stable performance across various noise levels. This makes Random Forest a suitable model for real-world applications where data imperfections are common.

Conversely, XGBoost, while excelling in clean data conditions, proved more susceptible to noise. The model’s reliance on capturing intricate relationships in the data makes it vulnerable when those relationships are distorted by noise. This suggests that XGBoost may require higher-quality data or more sophisticated noise-handling techniques to maintain its edge in forecasting accuracy.

The unexpected improvement in forecast accuracy, when moderate noise was added, emphasizes the potential benefit of data preparation by perturbing the input data. In some cases, this jittering effect may help models avoid overfitting to historical patterns and encourage them to generalize better to future, unseen data. However, as the noise levels increased, all models, especially XGBoost, experienced a decline in performance, reinforcing the importance of ensuring data quality for optimal forecasting results.

9.3 Practical Implications and Strategic Benefits for ASML

The practical implications of the research findings are significant for ASML’s Field Service Hub and broader production capacity planning. The proposed forecasting models, XGBoost, Random Forest, and ARIMA, demonstrate substantial improvements over the current benchmark. All three models outperform the benchmark, particularly in terms of accuracy and cost efficiency, which highlights their potential to improve ASML’s capacity planning operations.

9.3.1 Proactive Operations and Data-Driven Decision Making

The implementation of one of the proposed models will enable the FSH to shift from a reactive to a proactive approach in its operations. The predictive power of these models provides FSH with the ability to forecast demand more accurately, leading to improved production schedules and optimized workforce allocation. Instead of responding to production bottlenecks as they arise, ASML can now anticipate demand fluctuations, adjust capacity accordingly, and prevent issues such as late or emergency orders. By making decisions based on data-driven models, FSH operations will become more efficient, responsive, and resilient to demand volatility.

9.3.2 Reduction in Forecast Errors and Cost Savings

While precise cost savings are challenging to quantify without further research, the data suggest a substantial reduction in forecast error compared to the benchmark of up to 37.4% in MAE. This reduction in error directly contributes to better capacity planning. With fewer forecast errors, ASML can reduce instances of overstaffing or understaffing, thus optimizing labor costs and ensuring more consistent workforce utilization.

Moreover, a decrease in forecast errors means fewer late or emergency orders, which typically incur additional costs and disrupt production flow. By minimizing these disruptions, ASML can reduce both direct labor costs and the more significant expenses associated with rush orders and production delays.

9.3.3 Workforce Planning and Production Scheduling Improvements

Improved forecasting models have direct implications for workforce planning. For instance, with more accurate forecasts, companies can allocate the appropriate number of FTEs to production without the risk

of over- or under-utilizing resources. This will likely lead to a reduction in labor costs while simultaneously improving operational efficiency.

However, capacity challenges remain for ASML, particularly during periods of high demand or in the summer months when the total workforce is at a lower capacity. Despite this, recent data suggests that the reduced capacity had minimal impact on production efficiency. This may be due to increased focus on production, leading to a higher production per FTE. Still, the current model of reduced capacity, while effective in the short term, could decrease workforce morale over time, meaning it's not sustainable in the long run. With better forecasting, ASML can avoid such extreme fluctuations in workforce utilization and provide a more stable, long-term approach to staffing.

9.3.4 System Integration and Future Optimization for ASML

For ASML to realize the full benefits of these forecasting models, integration with existing systems is crucial. Implementing the models within their systems can significantly enhance production planning. The combination of these tools will enable real-time insights into production minutes, allowing for ongoing adjustments to the workforce and scheduling based on current forecasts.

However, improvements in data quality are essential for the long-term success of any forecasting model. Ensuring reliable data access and an effective remediation process will help maintain accuracy in the models' predictions. In cases where data has been manually overwritten to solve one problem, new problems may arise, and this cycle of inconsistency can hinder long-term planning. ASML must establish protocols to ensure data integrity and avoid such pitfalls.

9.3.5 Limitations of the Current Benchmark Model and Strategic Next Steps for ASML

The current forecast by the benchmark, while functional, is no longer reliable in comparison to the proposed forecasting models. The benchmark struggles with forecast accuracy and operational reliability, which has led to inefficiencies in capacity planning. Replacing it with a more advanced, robust forecasting model will improve decision-making and ultimately lead to cost savings and operational efficiency. By adopting one of the proposed models, ASML can better plan for future growth, optimize production schedules, and create a more agile and responsive operational framework.

9.3.6 Conclusion Practical Implications and Strategic Benefits for ASML

In conclusion, the strategic benefits of implementing one of the proposed forecasting models are clear. From reducing costs and improving workforce planning to enhancing overall operational efficiency, these models provide ASML with a reliable and future-proof solution. As the Field Service Hub continues to grow and adapt to shifting market demands, a data-driven approach to production capacity planning will be critical to maintaining ASML's competitive edge in the semiconductor industry.

9.4 Insights from the Use Case for Broader Application

The application of the general forecasting and capacity optimization model to the ASML Field Service Hub has provided valuable insights, not only for ASML but also for other industries facing similar challenges in managing uncertain demand and optimizing production capacity. While ASML operates in the semiconductor industry with its specific set of operational complexities, the lessons learned from this use case demonstrate the model's flexibility and robustness across various sectors, enabling a shift towards more proactive and data-driven operations.

9.4.1 Flexibility of the Model

One of the most significant takeaways from the ASML use case is the model's adaptability to different demand patterns and operational constraints. Importantly, because the forecasting model is based on machine learning, a new model will be trained for each specific use case. This means that the specific features used in ASML, such as data for each product flow, are not limiting factors. For any other organization, as long as relevant underlying data is available, the model can be trained to perform effectively.

For ASML, these features were essential for capturing the demand flow and spare part planning. However, in other industries, different types of data might play a more critical role. The model's performance is tied not to specific features but to the availability of quality data that reflects the underlying drivers of demand. As long as an organization has access to its operational data, whether it be sales forecasts, inventory levels, or production schedules, the machine learning model can be adapted to forecast demand and optimize capacity effectively.

This flexibility ensures that the model can be adapted to different industries and data environments.

9.4.2 Data-Driven Decision-Making

One of the key goals of applying the model at ASML was its ability to drive data-driven decision-making. By leveraging machine learning models and optimization techniques, the model provided actionable insights that led to more informed decisions about production capacity, resource allocation, and labor management. This approach can be generalized across industries where operational decisions are often based on gut feeling or limited historical data.

9.5 Conclusion

In this section, the research questions are answered based on the results and insights gained from applying advanced forecasting and capacity optimization models to the general problem of demand uncertainty and production capacity planning. While the ASML Field Service Hub use case was central to this study, the findings are broadly applicable to various industries facing similar challenges.

9.5.1 Main Research Question: How to design an improved forecasting and production capacity optimization model that leads to an improved mid-term production capacity planning that can be applied to the ASML Field Service Hub?

The improved forecasting and production capacity optimization model developed in this research combines advanced machine learning techniques such as XGBoost, Random Forest, and ARIMA to predict future demand with higher accuracy and enable more efficient capacity planning.

The results demonstrate a significant reduction in forecasting errors (up to 37.4% for MAE with XGBoost) and simulated costs (48.3% reduction with XGBoost). Additionally, the model performed within 2.89% of the theoretical cost optimum, suggesting that the proposed approach offers substantial cost-saving potential. Overall, the general model provides a framework that can be adapted to different industries, facilitating better resource allocation, reducing missed orders, and improving efficiency in capacity planning.

The final model also incorporates robustness testing through jitter tests to ensure stability in predictions, even under noisy data scenarios, allowing for reliable mid-term capacity planning. Incorporating future growth scenarios and optimizing resource allocation based on the demand forecast enables improved production capacity decisions, reducing missed orders and increasing efficiency compared to ASML's current models.

9.5.2 Research Question 1: What methods are best suited to forecast demand under uncertainty?

The forecasting methods were chosen based on requirements given by the literature and the use case. The forecasting model must be adaptable, scalable, and capable of handling both short-term fluctuations and long-term trends. It must account for different levels of noise and potential demand changes.

The research identified that machine learning models such as XGBoost and Random Forest are particularly well-suited for forecasting demand under uncertainty. These models excel in handling both short-term fluctuations and long-term trends. They can adapt to various demand patterns and noise levels, making them versatile for industries with complex, volatile demand. In comparison, ARIMA performed well with simpler time-series data but struggled with more intricate relationships.

The models were evaluated using MAE and RMSE to gauge their accuracy, with XGBoost and Random Forest consistently outperforming traditional models. These metrics, combined with cost-based assessments, confirm that machine learning models are more effective at handling unpredictable demand compared to conventional time-series methods.

9.5.3 Research Question 2: How should forecasting models be designed for optimal capacity planning?

Designing forecasting models for optimal capacity planning involves a balance between accuracy, flexibility, and robustness. The following considerations are essential for building models that effectively forecast demand and translate those forecasts into actionable production capacity decisions:

- **Feature and Data Selection:** The selection of relevant features is critical for accurate forecasting. This includes identifying key variables from historical data that influence demand, such as the mix of product flows, sales trends, seasonality, product lifecycle data, and external factors like economic conditions. Proper feature selection enhances the model's ability to capture the underlying demand drivers, improving forecast accuracy.
- **Data Aggregation and Timeframe:** Determining the appropriate data aggregation level (e.g., daily, weekly, or monthly) and selecting the right timeframe (lag periods) are vital for ensuring the model aligns with the operational needs of capacity planning.
- **Forecasting Multiple Time Steps:** Multi-step forecasting can be implemented using either multi-output regression or chain regression. Multi-output regression trains independent models for each time step, providing more flexibility but lacking interdependence between time horizons. In contrast, chain regression allows each time step to influence the next, capturing sequential relationships in demand but potentially propagating errors through the chain. The decision between these approaches depends on the specific demand patterns and the desired forecast horizon. For this research and use case, multi-output regression performed better.
- **Adaptability:** The models are designed to account for future growth scenarios by incorporating flexibility in feature engineering (e.g., adding lag features and growth projections). Adaptable models are better equipped to remain accurate even as the underlying business environment evolves.
- **Incorporating future scenarios:** Future scenarios, such as expected growth in production or diversification of product lines, are integrated into the model to make it more future-proof. By including these scenarios in the forecasting process, the model can anticipate changes in demand and adjust capacity planning strategies accordingly.
- **Robustness:** Robustness is critical for ensuring the model's stability in real-world applications where data can be noisy or inconsistent. Robustness is evaluated through jitter tests, which introduce noise into the input data to simulate unpredictable demand fluctuations. While XGBoost showed some sensitivity to noise, it performed well overall, with Random Forest and ARIMA demonstrating strong stability under noisy conditions.
- **Explainability:** Explainability is crucial for model adoption, especially when the output is used for critical decision-making. XGBoost and Random Forest, though powerful, are often considered black-box models. To address this, feature importance metrics, SHAP values, were used to provide insights into how predictions are made. The ARIMA model is more interpretable by design, helping balance explainability with accuracy, which can be valuable when transparency is required for business stakeholders.

This approach ensures that the models can be customized to meet industry-specific requirements, remain interpretable for decision-makers, and provide accurate forecasts that optimize production capacity.

9.5.4 Research Question 3: How can forecasts be translated into production capacity decisions?

Forecasts are translated into production capacity decisions through an optimization process that aligns predicted demand with resource allocation. This process ensures that the organization can meet demand efficiently while minimizing costs associated with labor, undercapacity, and overcapacity. The translation from forecast to capacity decision involves the use of three key strategies for production planning: Follow Forecast 3 Months Ahead, Smoothing via Multiple Forecasts, and Buffer + Smoothing.

These strategies are incorporated into an optimization simulation, which seeks to minimize total operational costs by aligning the production capacity with forecasted demand in a way that smooths capacity volatility over time. The optimization process uses an exhaustive search around the forecasted

values, simulating different capacity scenarios for the upcoming months to determine the most cost-effective production plan.

The simulation is run in a rolling window format, where forecasts and capacity decisions are continuously updated based on new data. This method allows the model to dynamically adjust to changes in demand while ensuring that the production capacity remains flexible and responsive. The cost simulation takes labor costs and costs for under- and overcapacity into account.

To further clarify the mechanics of the optimization process, the mathematical formulation of the model was documented, providing transparency and explaining how the Python code implements the optimization logic. This mathematical model highlights the decision variables, cost functions, and constraints that guide the capacity planning process.

Ultimately, the model outputs the optimized production capacity for the upcoming months, providing recommendations on how to allocate labor and resources to meet forecasted demand while minimizing operational costs.

9.5.5 Research Question 4: How is the ASML Field Service Hub case applied to the general model?

While this model is designed to be applicable across various industries, its practical utility can only be fully understood when applied to a real-world scenario, such as the ASML Field Service Hub.

The ASML Field Service Hub case provided a real-world application of the general model, illustrating how demand forecasting and capacity planning can be optimized for highly specialized, complex production environments. Four distinct product flows were considered, each with varying demand frequency, seasonality, and volatility.

Historical demand and production data were acquired, cleaned, remediated, and prepared, and features were selected from the data set. Used features included the data per product flow, the index and total production.

Demand data was aggregated monthly to align with mid-term production capacity planning needs. Multiple timeframes were tested to capture the relationships between past and future demand, ensuring the model had enough historical context to make accurate predictions without becoming overly complex or prone to overfitting.

Before model training, standard data checks were performed, including trend analysis to ensure the integrity of the data. This step helped identify any underlying trends or seasonal patterns that could improve the model's ability to predict future demand accurately.

A Python-based forecasting model was created using various forecasting techniques, including XGBoost, Random Forest, and ARIMA. These models were evaluated based on performance metrics such as MAE and RMSE, with XGBoost emerging as the most effective model for handling the complex demand patterns in the ASML FSH environment.

Building on the forecasting model, a capacity planning model was developed using the mathematical formulation outlined earlier in the thesis. This model used demand forecasts to optimize production capacity and aimed to minimize costs while ensuring that the ASML FSH could meet its service level agreements without under- or over-allocating resources.

The results showed a tangible improvement compared to the currently employed model at ASML FSH. The new forecasting and capacity planning model reduced forecasting errors by up to 37.4% (for MAE with XGBoost) and optimized production capacity, leading to simulated cost reductions of 48.3%. The model's ability to smooth out capacity fluctuations and make more accurate predictions resulted in more efficient resource allocation, fewer missed orders, and better alignment between demand and capacity.

The ASML Field Service Hub case provides a valuable real-world application of the general forecasting and capacity optimization framework developed in this thesis. By demonstrating how the model can be applied in a high-demand, complex production environment, the FSH case proves that the combination of advanced forecasting techniques and capacity planning optimization can significantly reduce costs and improve operational efficiency. This use case serves as a blueprint for other industries facing similar challenges, validating the model's ability to drive data-driven decision-making and improve resource allocation in uncertain demand environments.

9.5.6 Research Question 5: How are the models validated and compared to benchmarks?

The proposed models were validated through rigorous comparisons against ASML’s current benchmark model using metrics such as MAE, RMSE, and cost-based simulations. The proposed models (XGBoost, Random Forest, and ARIMA) consistently outperformed the benchmark system, significantly reducing forecast errors and lowering production costs.

Key findings include:

- **Error comparison:** The machine learning models reduced forecasting errors (MAE and RMSE) compared to the benchmark, with XGBoost showing the best balance between accuracy and cost performance.
- **Cost comparison:** Simulated cost reductions were significant, with XGBoost achieving the lowest total cost, 48.3% lower than the benchmark. The cost differential between the theoretical optimum and the actual realized cost was minimal.
- **Explainability and robustness:** ARIMA was the most explainable model and handled noisy data well. XGBoost showed the best overall performance but was slightly more sensitive to noise. Random Forest was a more explainable model than XGBoost due to the more balanced use of features and was more robust under noisy conditions.

The validation process confirmed that advanced machine learning models can significantly improve demand forecasting and production capacity planning across various sectors when compared to traditional models.

9.6 Limitations

In this section, several limitations encountered during the testing and implementation of the forecasting models are addressed. These limitations highlight areas where the data, models, or external factors may have impacted the results or their generalizability. Most limitations found apply to the FSH use case but also serve as more general warnings for implementation in other cases.

9.6.1 Data Limitations

The primary limitation stems from the data quality and preparation process. The current data setup is not suitable for real-time model integration. Currently, whenever a new month of data is available, it must be manually prepared, as outlined in the data preparation steps in section 7.3. This introduces inefficiencies and delays, making it difficult to adapt the model to a dynamic production environment. Automating this process would be essential for deploying the model at scale.

9.6.2 Model Limitations

The forecasting models themselves are based on numerous assumptions, which may limit their applicability in real-world scenarios. For instance, Strategy 3 outperforms Strategy 2 in most cases, suggesting a possible bias in the models. This could indicate overfitting, as many parameters were tuned during model development. The dataset is relatively small, with few timesteps, which further increases the risk of overfitting. Future research should explore robust optimization to mitigate this issue and improve the generalizability of the models.

Another model limitation regards being future-proof. While the proposed model shows a significant improvement over ASML’s benchmark, the comparison is based on historical data that may not fully reflect future conditions. It remains speculative whether these improvements will consistently hold in the face of changing demand patterns.

Furthermore, although this model has been applied successfully to the ASML use case, its applicability to other industries will depend on the availability of high-quality data and similar demand variability characteristics. Further testing in diverse operational environments would be required to substantiate these findings.

Robustness Limitations

The robustness tests, such as jitter tests (adding noise to the data), revealed some limitations. While adding noise improved model performance in certain cases, it may not fully simulate real-world uncertainty. Real-world variability can involve more complex factors, such as demand shifts due to economic or supply chain disruptions, which the jitter tests may not capture.

9.6.3 External Limitations

External factors also limit the model's effectiveness. For example, the model was designed for the demand side, meaning it does not account for issues with supply or backlog, which are common in real-world production environments. Unexpected market fluctuations or supply chain disruptions can significantly affect production capacity, but these factors were excluded from the model, simplifying its scope.

9.6.4 Risk Assessment

There are inherent risks in relying too heavily on the forecasting and capacity planning model. Unexpected demand fluctuations, technological constraints, or changes in market conditions could limit the accuracy and usefulness of the model. It is important to emphasize that the model should not be viewed as a standalone solution but as a tool to provide insights. Human supervision and decision-making are crucial in interpreting and acting on the model's forecasts, ensuring that strategic decisions account for factors the model may miss.

9.6.5 Ethical Considerations

Several ethical considerations arise in the implementation of forecasting models, particularly around data use and privacy, black-box modeling, and the potential for automation in decision-making.

- **Data privacy** is critical, as the model relies on large datasets that may contain sensitive information. Ensuring that data is anonymized and securely stored is essential.
- **Black-box modeling** presents challenges, as complex machine learning models can be difficult to interpret. Ensuring that the model is explainable and transparent to decision-makers will help avoid blind reliance on its outputs.
- **Automation in decision-making** raises ethical concerns, particularly regarding its impact on the workforce. A model that optimizes efficiency might suggest reducing headcount, which could lead to job losses. It is essential to ensure that the model is used as an additional tool rather than an all-encompassing solution. Human oversight is necessary to consider the broader implications of decisions, especially those affecting employment.

In conclusion, while the forecasting models offer clear strategic benefits for ASML, they come with limitations that need to be addressed in future iterations. Improved data processes, refined model assumptions, and careful consideration of the ethical impacts will help ensure that these models are both effective and responsible.

9.7 Implementability at FSH

The practical implementation of the forecasting model at ASML will require careful planning and integration into the existing production capacity planning processes. Given its potential to improve mid-term forecasting, production scheduling, and workforce management, ensuring that the model is operational and easily usable by staff is essential for maximizing its impact.

9.7.1 Scalability

The model was designed with a specific focus on ASML's FSH, but its scalability potential is significant. The methods used can be adapted for other departments or product flows within ASML that require demand forecasting and production capacity optimization or even completely different organizations and sectors. By modifying the model inputs, such as product characteristics or production constraints, the forecasting system could extend to other critical areas. Additionally, given its design in Python, the

model is flexible enough to handle larger datasets or different timeframes, making it adaptable to broader operations.

9.7.2 Technical Environment

The model is implemented using Python, which is a versatile and widely used programming language in data science and machine learning. At ASML, the model can be integrated into the existing IT infrastructure by leveraging existing data pipelines and dashboards. The output from the Python model can be fed into ASML's current systems. By using the already available systems, the model can seamlessly integrate into dashboards, providing decision-makers with real-time or near-real-time insights.

For data pipelines, it is crucial to automate the flow of data from ASML's existing systems into the model, ensuring that the forecasting model is always using the most current and accurate data. This may require collaboration with ASML's IT teams to ensure that the appropriate APIs and data connectors are in place.

9.7.3 User Interaction

ASML staff will interact with the forecasting model primarily through reports, visualizations, and potentially automated alerts. The outputs, such as forecasted demand, capacity utilization, and workforce requirements, can be visualized in intuitive dashboards, helping decision-makers quickly grasp the situation and take action. Additionally, the model can be designed to trigger alerts when certain thresholds are reached, such as a sudden increase in forecasted demand or workforce shortages, providing early warning and allowing for more proactive planning.

Training the staff on how to interpret the results and how to make adjustments based on their expertise will be important. While the model can provide data-driven insights, there should be a mechanism for manual inputs, allowing users to adjust the model's outputs based on expert knowledge or external factors not captured by the model.

9.7.4 Next Steps

The next phase of the project involves working on a detailed roadmap for implementation within the Field Service Hub. This roadmap will outline the timeline and tasks needed to operationalize the model, including the necessary technical integrations, staff training, and any adjustments to the existing infrastructure. Ensuring the model is real-time capable and easy to update with new data will be crucial for its long-term success.

9.8 Future Research Directions and Improvements

While the current forecasting model provides valuable insights for production capacity planning at ASML, there are several areas where future research could lead to further improvements in accuracy, robustness, and operational efficiency.

9.8.1 Testing and Generalizability

To test the generalizability of the model, it should be applied in different use cases and industries. This would help assess whether the forecasting techniques developed for ASML's FSH can be applied to other sectors with similar capacity planning challenges.

9.8.2 Investigating the Supply Side

One important future research direction is the incorporation of supply-side factors into the model. Currently, the focus is primarily on the demand side, but understanding and incorporating the supply chain dynamics, including supplier reliability and inventory levels, could further optimize production planning. For instance, integrating known future orders as a feature could improve the model's accuracy, but would require historical data on past future orders, which is not readily available.

9.8.3 Panel Data and Advanced Models

The decision to use time-series data instead of panel data was made due to the advantages of time-series models. However, future research could explore using panel data with more advanced models, such as neural networks, to capture more complex relationships between variables. Transforming the used panel data into a usable time-series format was necessary for this project, but in future studies, using more complex models could enhance the model's predictive power.

9.8.4 Bias in Forecasting Models

The current model shows unexpected results, with Strategy 3 performing better than Strategy 2, even though Strategy 3 adds a buffer that theoretically should lead to higher errors. This suggests a possible bias in the model, likely due to the limited historical data available. Future research could focus on understanding and mitigating this bias by training the model on more representative data that better captures fluctuations in demand.

9.8.5 Preventing Overfitting

Given the small number of timesteps in the dataset, the model is at risk of overfitting. Many parameters were changed to optimize for cost, which might have led the model to fit too closely to the available data. Future research should explore robust optimization techniques to reduce the risk of overfitting and improve the model's performance on unseen data.

9.8.6 Cost of Undercapacity and Overcapacity

Investigate the factor 3 in cost for undercapacity versus overcapacity. How expensive are late orders exactly?

9.8.7 ARIMA Strategy Improvement

For ARIMA, all strategies are strictly worse than the total error comparison. This is mostly because the most important month to predict, 3 months ahead (the time to scale) is different than the starting point. Since the ARIMA setup is almost a Naive Forecast, it will not perform well. In future research, it might be interesting to look into different strategies or to choose the ARIMA setup (p,d,q) based on the strategy error instead of the total error.

In conclusion, while the current model offers substantial improvements over ASML's benchmark, future research directions outlined above could make the model more robust, scalable, and capable of handling a wider range of variables, scenarios, and use cases.

Chapter 10

Reflection

This thesis has focused on developing and testing a forecasting model to improve production capacity planning at ASML's Field Service Hub. The key findings provide valuable insights for optimizing production capacity, improving operational efficiency, and reducing costs for both ASML and more general cases. The model offers a data-driven approach that surpasses the current benchmark in terms of accuracy and cost, with practical implications for both ASML's workforce management and operational model.

10.1 Insights Created

10.1.1 Forecasting

The development and evaluation of the forecasting models revealed that machine learning techniques such as XGBoost and Random Forest outperformed the benchmark in terms of forecast accuracy and overall production capacity planning. XGBoost was found to be the most cost-effective, but it introduced higher production delays and more volatility in workforce planning. On the other hand, ARIMA demonstrated strong robustness in minimizing delays and maintaining workforce stability, although it required more FTEs to meet demand.

This research also highlighted the importance of balancing the trade-offs between forecast accuracy, cost minimization, and production stability. While XGBoost achieved the lowest cost, ARIMA was most effective in preventing production delays and stabilizing staffing levels. Therefore, the choice of model should align with the organization's priorities, whether to minimize costs or to maintain a steady production flow.

10.1.2 Production Capacity Optimization

By optimizing capacity planning based on more accurate demand forecasts, organizations can make more informed decisions regarding staffing levels and production scheduling. For organizations prioritizing production time and stability, ARIMA (Strategy 3) appears to be the most effective solution. However, if cost reduction is the primary goal, as in this study, XGBoost (Strategy 1) may offer a more attractive balance between cost efficiency and operational stability. The model simulations demonstrated that, compared to the benchmark, the proposed strategies offer superior performance not only in reducing forecast error but also in aligning workforce capacity with demand fluctuations. This alignment minimizes the risk of both overcapacity and undercapacity, leading to more efficient resource allocation and cost savings. Ultimately, the improved model offers ASML the ability to respond more proactively to changes in demand, leading to a smoother, more flexible production environment.

10.2 Implications for the Research Field

This research contributes to the broader field of production capacity planning and forecasting by demonstrating the value of integrating predictive and prescriptive models. While simpler time-series models such as ARIMA performed well, this study shows the potential of more advanced machine learning models, like XGBoost, for cost minimization in production planning.

Additionally, the use of jitter tests for robustness analysis adds to the literature on model stability under real-world conditions. Future research could further explore integrating supply-side variables,

macroeconomic factors, and more complex forecasting models such as neural networks to enhance the predictive power of production planning tools.

10.3 Strategic Benefits for ASML

The insights gained from this research offer significant strategic benefits for ASML, particularly in its production planning and capacity management. Implementing the forecasting models proposed in this study can result in:

- **Improved Capacity Planning:** The adoption of a more accurate forecasting model can reduce errors by up to 37% compared to the benchmark, leading to better resource allocation.
- **Proactive Decision-Making:** By relying on data-driven insights rather than subjective assessments, ASML can transition towards more proactive operations at FSH. The model enables better alignment between workforce capacity and demand, allowing the organization to mitigate risks of production delays or underutilization.
- **Operational Flexibility:** The model's ability to anticipate demand shifts and adjust staffing levels accordingly contributes to greater operational flexibility. For example, it can help ASML navigate the seasonal variation in workforce capacity, ensuring higher productivity during periods of undercapacity and avoiding strain during high-demand periods.

10.4 Final Remarks

In conclusion, this forecasting and optimization model offers ASML and other organizations an opportunity to significantly improve their production capacity planning processes. By leveraging machine learning to enhance forecast accuracy and operational efficiency, the model can help to better align workforce capacity with fluctuating demand.

Although further refinement of the model is possible, especially by incorporating more data and supply-side factors, the current model represents a strong foundation for more effective capacity planning at ASML's FSH. Ultimately, this research highlights the importance of data-driven forecasting in enhancing production efficiency and minimizing costs.

Acknowledgements

To improve readability, Generative AI was used to improve the writing style. It was not used to provide any code, new information or answers, but only to make coherent text out of shorter pieces of text.

This thesis was adjusted to not publish any confidential information from ASML. Thus, all confidential results, figures, tables, and text were left out of this document so it could be published to the TU Delft repository.

Bibliography

- Aktepe, A., Yanık, E., & Ersöz, S. (2021). Demand forecasting application with regression and artificial intelligence methods in a construction machinery company. *Journal of Intelligent Manufacturing*, 32, 1587–1604.
- Albey, E., Norouzi, A., Kempf, K. G., & Uzsoy, R. (2015). Demand modeling with forecast evolution: An application to production planning. *IEEE Transactions on Semiconductor Manufacturing*, 28(3), 374–384.
- Amin-Naseri, M. R., & Tabar, B. R. (2008). Neural network approach to lumpy demand forecasting for spare parts in process industries. *2008 International Conference on Computer and Communication Engineering*, 1378–1382.
- Amirkolaii, K. N., Baboli, A., Shahzad, M., & Tonadre, R. (2017). Demand forecasting for irregular demands in business aircraft spare parts supply chains by using artificial intelligence (ai). *IFAC-PapersOnLine*, 50(1), 15221–15226.
- Bacchetti, A., & Saccani, N. (2012). Spare parts classification and demand forecasting for stock control: Investigating the gap between research and practice. *Omega*, 40(6), 722–737.
- Bandara, K., Bergmeir, C., & Hewamalage, H. (2020). Lstm-msnet: Leveraging forecasts on sets of related time series with multiple seasonal patterns. *IEEE transactions on neural networks and learning systems*, 32(4), 1586–1599.
- Basten, R. J., & Arts, J. J. (2017). Fleet readiness: Stocking spare parts and high-tech assets. *IIEE Transactions*, 49(4), 429–441.
- Bertsimas, D., & Kallus, N. (2020). From predictive to prescriptive analytics. *Management Science*, 66(3), 1025–1044.
- Biçer, I., & Seifert, R. W. (2017). Optimal dynamic order scheduling under capacity constraints given demand-forecast evolution. *Production and Operations Management*, 26(12), 2266–2286.
- Boström, E., & Lundell, J. (2020). Availability vs. cost efficiency: A case study taking on an integrated approach to spare part distribution in the high-tech industry.
- Boylan, J. E., Syntetos, A. A., & Karakostas, G. C. (2008). Classification for forecasting and stock control: A case study. *Journal of the operational research society*, 59, 473–481.
- Breiman, L. (2001). Random forests. *Machine learning*, 45, 5–32.
- Cai, Z., Peng, C., & Du, S. (2021). Jitter: Random jittering loss function. *2021 International Joint Conference on Neural Networks (IJCNN)*, 1–8.
- Chen, T., & Guestrin, C. (2016). Xgboost: A scalable tree boosting system. *Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining*, 785–794.
- Chien, C.-H., Chen, P.-Y., Trappey, A. J., Trappey, C. V., et al. (2022). Intelligent supply chain management modules enabling advanced manufacturing for the electric-mechanical equipment industry. *Complexity*, 2022.
- Cohen, M. A., Agrawal, N., & Agrawal, V. (2006). Winning in the aftermarket. *Harvard business review*, 84(5), 129.
- Croston, J. D. (1972). Forecasting and stock control for intermittent demands. *Journal of the Operational Research Society*, 23(3), 289–303.
- Deng, C., & Liu, Y. (2021). A deep learning-based inventory management and demand prediction optimization method for anomaly detection. *Wireless Communications and Mobile Computing*, 2021, 1–14.
- do Rego, J. R., & De Mesquita, M. A. (2015). Demand forecasting and inventory control: A simulation study on automotive spare parts. *International Journal of Production Economics*, 161, 1–16.
- Elmachtoub, A. N., & Grigas, P. (2022). Smart “predict, then optimize”. *Management Science*, 68(1), 9–26.

- Fattah, J., Ezzine, L., Aman, Z., El Moussami, H., & Lachhab, A. (2018). Forecasting of demand using arima model. *International Journal of Engineering Business Management*, 10, 1847979018808673.
- Gallagher, T., Mitchke, M., & Rogers, M. (2005). Profiting from spare parts. *McKinsey Quarterly*, 2.
- Gardner, E. S. (1985). Exponential smoothing: The state of the art. *Journal of forecasting*, 4(1), 1–28.
- Heinecke, G., Syntetos, A., & Wang, W. (2013). Forecasting-based sku classification. *International Journal of Production Economics*, 143(2), 455–462.
- Hu, Q., Boylan, J. E., Chen, H., & Labib, A. (2018). Or in spare parts management: A review. *European Journal of Operational Research*, 266(2), 395–414.
- Hua, Z., & Zhang, B. (2006). A hybrid support vector machines and logistic regression approach for forecasting intermittent demand of spare parts. *Applied Mathematics and Computation*, 181(2), 1035–1048.
- Hua, Z., Zhang, B., Yang, J., & Tan, D. (2007). A new approach of forecasting intermittent demand for spare parts inventories in the process industries. *Journal of the Operational Research Society*, 58(1), 52–61.
- Kennedy, W., Patterson, J. W., & Fredendall, L. D. (2002). An overview of recent literature on spare parts inventories. *International Journal of production economics*, 76(2), 201–215.
- Little, A., & Brown, R. (1956). Exponential smoothing for predicting demand.
- Lundberg, S. M., & Lee, S.-I. (2017). A unified approach to interpreting model predictions. In I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, & R. Garnett (Eds.), *Advances in neural information processing systems 30* (pp. 4765–4774). Curran Associates, Inc. <http://papers.nips.cc/paper/7062-a-unified-approach-to-interpreting-model-predictions.pdf>
- Maccarthy, B. L., & Liu, J. (1993). Addressing the gap in scheduling research: A review of optimization and heuristic methods in production scheduling. *The International Journal of Production Research*, 31(1), 59–79.
- Mishra, S., Dutta, S., Long, J., & Magazzeni, D. (2021). A survey on the robustness of feature importance and counterfactual explanations. *arXiv preprint arXiv:2111.00358*.
- Mula, J., Poler, R., García-Sabater, J. P., & Lario, F. C. (2006). Models for production planning under uncertainty: A review. *International journal of production economics*, 103(1), 271–285.
- Pranab. (2021). Robustness measurement of machine learning models with examples in python. <https://pkgghosh.wordpress.com/2021/03/25/robustness-measurement-of-machine-learning-models-with-examples-in-python/>
- Qiu, C., Zhao, B., Liu, S., Zhang, W., Zhou, L., Li, Y., & Guo, R. (2022). Data classification and demand prediction methods based on semi-supervised agricultural machinery spare parts data. *Agriculture*, 13(1), 49.
- Romeijnders, W., Teunter, R., & Van Jaarsveld, W. (2012). A two-step method for forecasting spare parts demand using information on component repairs. *European Journal of Operational Research*, 220(2), 386–393.
- Santana, A., Afonso, P., Zanin, A., & Wernke, R. (2017). Costing models for capacity optimization in industry 4.0: Trade-off between used capacity and operational efficiency. *Procedia Manufacturing*, 13, 1183–1190.
- Shafi, I., Sohail, A., Ahmad, J., Espinosa, J. C. M., López, L. A. D., Thompson, E. B., & Ashraf, I. (2023). Spare parts forecasting and lumpiness classification using neural network model and its impact on aviation safety. *Applied Sciences*, 13(9), 5475.
- Shale, E. A., Boylan, J. E., & Johnston, F. (2006). Forecasting for intermittent demand: The estimation of an unbiased average. *Journal of the Operational Research Society*, 57(5), 588–592.
- Sharma, P., Kulkarni, M. S., & Yadav, V. (2017). A simulation based optimization approach for spare parts forecasting and selective maintenance. *Reliability Engineering & System Safety*, 168, 274–289.
- Sundararajan, M., & Najmi, A. (2020). The many shapley values for model explanation. *International conference on machine learning*, 9269–9278.
- Syntetos, A., & Boylan, J. (2008). Forecasting for inventory management of service parts (chap. 20). *Complex system maintenance handbook*. Springer, World Academy of Science, Engineering and Technology (To appear in 2007).
- Syntetos, A. A., Babai, Z., Boylan, J. E., Kolassa, S., & Nikolopoulos, K. (2016). Supply chain forecasting: Theory, practice, their gap and the future. *European Journal of Operational Research*, 252(1), 1–26.
- Syntetos, Babai, M. Z., & Gardner Jr, E. S. (2015). Forecasting intermittent inventory demands: Simple parametric methods vs. bootstrapping. *Journal of Business Research*, 68(8), 1746–1752.

- Syntetos, Boylan, J. E., & Croston, J. (2005). On the categorization of demand patterns. *Journal of the operational research society*, 56, 495–503.
- Tang, B., & Khalil, E. B. (2022). Pyepo: A pytorch-based end-to-end predict-then-optimize library for linear and integer programming. *arXiv preprint arXiv:2206.14234*.
- Teunter, R. H., Syntetos, A. A., & Babai, M. Z. (2011). Intermittent demand: Linking forecasting to inventory obsolescence. *European Journal of Operational Research*, 214(3), 606–615.
- Tsipras, D., Santurkar, S., Engstrom, L., Turner, A., & Madry, A. (2018). Robustness may be at odds with accuracy. *arXiv preprint arXiv:1805.12152*.
- Van der Auweraer, S., & Boute, R. (2019). Forecasting spare part demand using service maintenance information. *International Journal of Production Economics*, 213, 138–149.
- Vargas & Cortés, M. E. (2017). Automobile spare-parts forecasting: A comparative study of time series methods. *International Journal of Automotive and Mechanical Engineering*, 14(1), 3898–3912.
- Willemain, T. R., Smart, C. N., & Schwarz, H. F. (2004). A new approach to forecasting intermittent demand for service parts inventories. *International Journal of forecasting*, 20(3), 375–387.
- Willemain, T. R., & Smart, C. N. (2001, March). System and method for forecasting intermittent demand [US Patent 6,205,431].
- Yan, R., & Wang, S. (2022). Integrating prediction with optimization: Models and applications in transportation management.
- Zhu, X., Zhang, G., & Sun, B. (2019). A comprehensive literature review of the demand forecasting methods of emergency resources from the perspective of artificial intelligence. *Natural Hazards*, 97, 65–82.
- Zhuang, X., Yu, Y., & Chen, A. (2022). A combined forecasting method for intermittent demand using the automotive aftermarket data. *Data Science and Management*, 5(2), 43–56.