Time Series and Machine Learning Approach for Forecasting the Demand for Small Equipment, Tools, and Consumables for Industrial Construction Projects

by

Elnaz Jafari

A thesis submitted in partial fulfillment of the requirements for the degree of

Master of Science

in

Construction Engineering and Management

Department of Civil and Environmental Engineering
University of Alberta

ABSTRACT

The high consumption and utilization of demand for small equipment, tools, and consumables in construction projects underscores the necessity for effective procurement strategies. Accurate estimation of these consumables is crucial for moving toward project completion in a timely manner. With recent advancements in time series analysis, artificial intelligence, and machine learning, these technologies can be employed to formulate predictive models.

This research aims to explore the advantages of using time series and machine learning—in combination with historical data from past projects—to identify key factors that impact demand for these consumables, as well as develop an efficient predictive model that analyzes and learns from historical data thereby facilitating precise estimations for future projects. The research involves collecting and analyzing historical data, analyzing current industry practices for estimating requirements for small equipment, tools, and consumables, and implementing time series analysis and machine learning algorithms to forecast demand for various types of consumables in construction projects. This study investigates crucial factors that influence these items, bridging the gap between literature review and industry practices.

Finally, this research proposes time series and machine learning models capable of predicting quantities in industrial projects using historical data. The proposed models provide an estimation of monthly requirements for various types of consumables throughout the project, which assists project managers in estimating required quantities, offering them accurate insights to help facilitate effective procurement strategies.

ACKNOWLEDGMENTS

First and foremost, my deepest gratitude is extended to my supervisor, Dr. Simaan AbouRizk, whose unwavering guidance and support over the past two years have been invaluable. Dr. AbouRizk has been an exemplary mentor, imparting wisdom that has enriched my professional and personal growth immeasurably. His role in my journey cannot be overstated, and for his mentorship, I am profoundly thankful.

I am equally grateful to Dr. Lingzi Wu, her readiness to assist and offer insightful advice has been instrumental in the success of my research. Her contributions have been pivotal, and her support is deeply appreciated.

I must extend profound thanks to my friend Mohamed ElMenshawy, whose unparalleled support was crucial throughout my master's journey. His dedication, insightful contributions, and hands-on assistance were instrumental in every phase. Mohamed's involvement was not just as a guide but as a pillar of strength and a constant source of motivation. I am profoundly grateful for his generosity of spirit, expertise, and the countless hours he devoted to my work.

My appreciation extends further to Dr. Yasser Mohamed, Maria Al-Hussein, Brenda Penner, and Stephen Hague for their invaluable dedication and support. Each has played a crucial role in my journey, offering their expertise and assistance when most needed. Their collective efforts have significantly contributed to the success of my research, and their support is wholeheartedly appreciated. Alongside Kohlbey Ozipko, whose meticulous efforts in manuscript editing have been essential. Her contributions have significantly enhanced the quality of my work.

The financial and resource support provided by the National Science and Engineering Research Council (NSERC) and Alberta Innovates has been fundamental to the accomplishment of this research. Their backing not only facilitated the research but also underscored the importance of advancing scientific knowledge.

A special acknowledgment to my friends, Mahsa, and Nastaran for their unwavering support and motivation. Their camaraderie and encouragement have been a source of strength and inspiration, propelling me forward during challenging times.

My heartfelt thanks are reserved for my family, whose endless love and belief in my capabilities have been the cornerstone of my achievements. To my sisters, Parinaz and Khatereh, for their support and for being my emotional anchors; and most importantly, to my parents. Their mentorship, faith, and unconditional love have shaped me in innumerable ways. Their

encouragement has been my guiding light, reminding me of the potential within me and the goals ahead.

In closing, this journey has been enriched by the collective support and encouragement of each individual mentioned and many unmentioned. Your contributions have left an indelible mark on my personal and professional development. Thank you for being part of my journey.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

LIST OF ABBREVIATIONS

Abbreviation	Explanation
AI	Artificial Intelligence
ANN	Artificial Neural Networks
AR	Autoregressive
ARIMA	Autoregressive Integrated Moving Average
AUC-ROC	Area Under the ROC Curve
BIM	Building Information Modelling
BP	Back Propagation
CSCM	Construction Supply Chain Management
DT	Decision Tree
EDA	Exploratory Data Analysis
EPC	Engineering, Procurement, and Construction
ERP	Enterprise Resource Planning
GDP	Gross Domestic Product
GRU	Gated Recurrent Units
KNN	K-Nearest Neighbor
LR	Linear Regression
LSTM	Long Short-Term Memory
MA	Moving Average
MAE	Mean Absolute Error
MAPE	Mean Absolute Percentage Error
ML	Machine Learning
MLP	Multilayer Perceptron
MSE	Mean Square Error
NN	Neural Network
PCA	Principal Component Analysis

PSO	Particle Swarm Optimization
RBF	Radial Basis Function
RF	Random Forest
RMSE	Root Mean Squared Error
RMSProp	Root Mean Squared Propagation
RNN	Recurrent Neural Network
SVM	Support Vector Machine
SVR	Support Vector Regression

CHAPTER 1: INTRODUCTION

1.1 Background and Problem Statement

The construction sector is one of the largest industries in the world economy. Billions of dollars are spent every year to execute and deliver construction projects to the public. In this industry, cost of materials—including small equipment, tools, and consumables among others—accounts for 30-50% of total project expenses (Heravi & Eslamdoost, 2015). Consequently, effective management of these resources is integral to the success of construction projects. Through such management, demand for resources can align with a project's requirements, allowing for a realistic and precise demand estimation. Generating accurate estimates of these resources is difficult as their utilization is contingent on various factors. Currently, estimations of these consumables are primarily conducted by project managers leveraging prevalent software tools and personal judgement. Thus, current practices rely on the expertise and experience of project professionals, which could lead to imprecise estimates. Moreover, the existing approach also often overlooks the latent value of historical data gathered from past projects.

Interestingly enough—and given recent advancements—machine learning (ML) algorithms have the potential to serve as robust tools for generating forecasting models and harnessing historical data to achieve precise resource estimation. Although ML algorithms have seen widespread adoption over the past two decades, their application for estimating small equipment, tools, and consumables resources is still emerging (Iwu, 2016; Gondia et al., 2020). The demand for these consumables varies significantly throughout the project duration, and insufficient forecasting methods contribute to inaccurate planning. This leads to delays in material delivery, directly impacting project schedules and efficiency. This study aims to develop an efficient predictive analytics model that learns from previous construction projects' data, facilitating the accurate estimation of necessary resources for diverse project types. Such a model will empower construction professionals to utilize their data effectively, enabling precise predictions for future projects' needs. The small equipment, tools, and consumables include, but are not limited to, the following: discs, gloves, helmets, kneepads, respirator filters, respirator masks, safety glasses, and welding jackets.

1.2 Research Objectives

This research seeks to forecast the demand of small equipment, tools, and consumables in industrial construction projects to improve project efficiency and reduce delays. The proposed objective will be achieved by doing the following: (1) examining current industry practices in estimating requirements for these resources; (2) identifying key factors affecting quantities of these consumables; (3) conducting Exploratory Data Analysis (EDA) for extracting insights and patterns between various features; (4) conducting various time series analysis techniques to forecast quantities throughout the project lifecycle; and (5) developing and evaluating innovative ML algorithms to measure their performance in forecasting the demand of small equipment, tools, and consumables.

1.3 Expected Contributions

The expected academic contributions of the proposed research can be summarized as follows:

- Identifying current industry practices in estimating demand requirements for small equipment, tools, and consumables.
- Investigating major factors impacting demand for consumables in construction projects.
- Investigating the use of time series analysis in building a forecasting model to predict requirements for various consumables in the projects.
- Developing a forecasting model using ML algorithms to predict requirements for these consumables in the projects.
- Implementing the proposed model on a case study.

The expected industrial contributions of the proposed research can be summarized as follows:

- Developing a framework to predict quantities for small equipment, tools, and consumables on a project level using a company's historical data.
- Proposing an accurate forecast for required consumables throughout the project lifecycle, as opposed to limiting the estimation to a single order for total quantities of consumables for the whole project.

1.4 Research Methodology

To achieve the research objectives, the proposed methodology is as follows: (1) conduct literature review and investigate current practices in the construction industry for estimating these resources; (2) identify factors that substantially influence the requirements of small equipment, tools, and consumables; (3) collect historical data of previous projects; (4) conduct feature selection to identify contributing factors and input variables of the model; (5) develop a forecasting model that utilizes time series and ML algorithms such as Artificial Neural Networks (ANNs); (6) train the model with preprocessed data and assess the algorithms' performance; and (7) generate predicted quantities for small equipment, tools, and consumables throughout the lifecycle of new projects. Figure 1 shows the summary of steps in the proposed methodology.

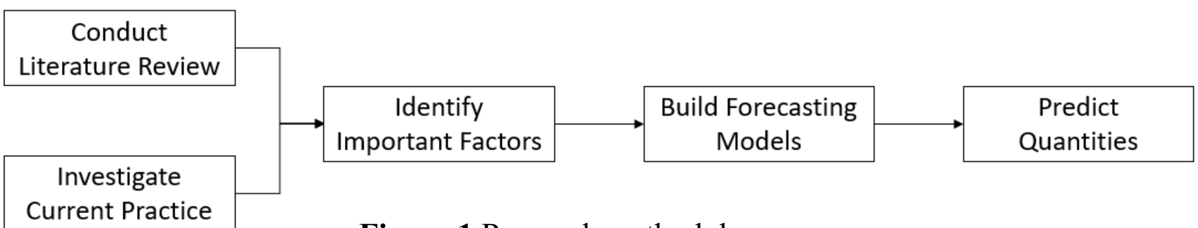


Figure 1 Research methodology

1.5 Research Questions

This research seeks to answer the following questions:

- What are the patterns of demand (such as trends, cycles, and random fluctuations) for small equipment, tools, and consumables?
- Which traditional forecasting methods are most accurate for the forecasting challenge at hand?
- Considering various performance metrics, which ML techniques are most appropriate for the problem?
- What are the primary indicators of demand, and how can we best select them?
- Which is the most effective measure for evaluating each model's forecast accuracy (for instance, Mean Square Error [MSE], Mean Absolute Error [MAE], Mean Absolute Percentage Error [MAPE])?

1.6 Thesis Organization

This thesis is organized into five main chapters to ensure a coherent structure and systematic flow of information. Chapter 1 establishes the foundation of the research, presenting the background, problem statement, objectives, expected contributions, methodology, and research questions. Chapter 2 provides a comprehensive review of existing literature, covering areas such as construction supply chain and procurement management, the impacts of material delays, procurement strategies, supply chain demand forecasting, and ML applications in construction projects, identifying research gaps in the process. Chapter 3 details the research methodology including factors affecting consumables, data collection, model development, and evaluation methods. Chapter 4 applies the developed methodologies to a real-world situation involving EDA, time series analysis, and ML forecasting models, followed by a discussion of the results. Finally, Chapter 5 summarizes the research findings, outlines the academic and industrial contributions, addresses limitations of the research, and suggests directions for future research.

CHAPTER 2: LITERATURE REVIEW

2.1 Introduction

The objective of this thesis is to develop a predictive model—fundamentally based on time series and ML algorithms—for precise estimation of demand forecasting for small equipment, tools, and consumables throughout the lifecycle of construction projects. This chapter presents the previous efforts made in several areas including construction supply chain management, impact of material delays on project performance, procurement strategies to mitigate delays, demand forecasting, factors affecting demand forecasting, time series analysis, and, finally, ML algorithms. The following sections discuss the aforementioned areas in detail.

2.2 Construction Supply Chain and Procurement Management

Construction Supply Chain Management (CSCM) is complex as it involves multiple stakeholders such as owners, contractors, consultants, regulators, and suppliers, each with distinct objectives. This complexity is increased by uncertainties arising from project time delays, market volatility, changing customer needs, fluctuating project and material costs, and governmental regulations. Such uncertainties necessitate a cooperative approach and the coordination of activities among all parties to mitigate risks and enhance owner satisfaction. Furthermore, procurement—representing a substantial portion of a construction project's total value—involves multiple steps from material requisition to final supplier selection and compliance verification with standards and codes, as outlined by the International Standards Organization (2020). The complexity of procurement processes and their role throughout a project's lifecycle are emphasized in research conducted by Mubin and Mannan (2013) who identify procurement as a high-risk activity. Indeed, challenges in procurement, such as selection of underqualified suppliers, late procurement team engagement, and misalignment of material supply with construction timelines, significantly affect project outcomes (Thomas et al., 2005; Drew et al., 2004).

The importance of procurement extends beyond acquisition of materials. It serves as a mechanism for corrective actions within construction projects, which is supported by case studies from leading Engineering, Procurement and Construction (EPC) firms (Micheli & Cagno, 2016). The choice of procurement methods—influenced by factors such as cost, project delivery methods, and contractual terms—plays a vital role in project success (Eriksson & Westerberg, 2011).

Inefficient procurement strategies can lead to schedule and budget overruns (Laedre et al., 2006) while effective material management helps in avoiding construction delays and optimizing productivity (Mawdesley & Al-Jibouri, 2010; Thomas et al., 2005). Ruparathna and Hewage (2015) investigated the conceptualization of construction procurement, making distinctions between those focused on contractual purchases and those encompassing all activities necessary for achieving project objectives. This distinction confirms the extent of procurement's impact on construction projects and highlights the need for strategic, well-informed decisions in procurement processes to prevent material shortages or surpluses and minimize deviations from project plans.

2.3 Impact of Material Delay on the Performance of Construction Projects

Performance evaluation is carried out using a range of factors typically divided into two categories depending on the perspective being examined. The first category pertains to dimensions such as time, cost, and quality (Micheli & Cagno, 2016; Ling et al., 2014; Chan & Chan, 2004) while the second category relates to entities within the CSCM including the project owner, contractor, subcontractor, and consultant (Pheng & Chuan, 2006). Additionally, the most common metrics used by stakeholders to assess outcomes of a construction project are time, cost, quality, health, and safety. According to Dissanayaka and Kumaraswamy (1998), the control of time and cost performance is influenced by factors like the procurement system, nature of the project, collaboration, and performance of stakeholders. The following paragraphs detail some of the factors that influence project performance.

Construction delays have long been a significant obstacle to successful project completion. As a result, researchers have attempted to identify and classify such delays to reduce/avoid associated costs. Assaf et al. (1995) used semi-structured interviews to identify causes of construction delay and grouped them into nine categories: funding, materials, contractual relations, alterations, extended permit approval processes, workforce, scheduling and control, equipment, and environmental factors. Within these categories, material delay was found to be a significant challenge. This finding was corroborated by the work of Thomas et al. (2005) who identified ineffective material management as a substantial contributor to project delays and financial losses. Furthermore, a survey conducted by Wang et al. (2016) pinpointed five primary risks to EPC including inflation, government inefficiency, local material shortages, fluctuating financial markets, and unstable political situations. The research of Enshassi et al. (2009) confirmed that the highest ranked factors influencing construction project performance were related to material management, particularly the unavailability of materials and resources. Based on a study of approximately 125 projects, Horman and Thomas (2005) concluded that material management related delays were indeed the most commonly documented. Thus, it is clear from previous research that material shortages are a recurring challenge for EPC contractors. Factors related to materials that impact project performance involve slow delivery, onsite material shortages, storage damage, specification changes during construction, and import challenges. The role of transportation and communication between suppliers and contractors must also be considered as these can contribute to procurement-related delays.

Construction projects rely on timely supply of necessary materials. As a result, any delay in the supply chain can significantly impact project timelines. Such material supply delays have been recognized as a primary factor in time overruns (Dey, 2000), and a number of studies have investigated the impact of material delays or mismanagement on project performance (Horman & Thomas, 2005; Thomas et al., 1999; Thomas et al., 1989). Specific construction materials such as pipeline, rebar, tiles, glass, rubber, cement, bulk filling materials like soil and rocks, ceramics, gravel, lead, paints, plastics, and plywood are consumed in large quantities and any disruption in their supply can also obstruct construction progress. Certain materials—termed as long-lead items—are required earlier in the construction process and are incorporated into project scheduling and contracting plans. With this being said, effective material planning—which includes maintaining records, determining target inventory levels, and setting the frequency of material delivery—can provide guidance for all subsequent activities and greatly influence the success of projects (Payne et al., 1996).

2.4 Procurement Strategies for the Supply Chain to Mitigate Delays

To enhance project performance, companies must develop plans to improve efficiency and effectiveness of their logistics and procurement processes (Dainty et al., 2001; Vrijhoef & Koskela, 2000). Disruptions can have a negative impact on a firm's performance. Hence, a strategic plan is needed to develop proactive actions that mitigate uncertainties and vulnerabilities. A suitable and efficient procurement route should be tailored to the unique characteristics, objectives, and performance expectations of each project, given that no two projects are identical and no single approach is universally applicable. Unfortunately, organizations often resist adopting new procurement strategies and opt to follow familiar routes even when they are unsuitable for a specific project, which contradicts recommended guidelines (Laedre et al., 2006). Some strategies that aid in mitigating delays and uncertainties are outlined in the following subsections.

2.4.1 Inventory/Buffer Management

Buffers and buffer management are employed to mitigate uncertainties in projects and procurement (Yeo & Ning, 2002). For instance, unnecessary delays in construction can be minimized by adding a buffer to the project timeline, which covers all planned dates for starting and completing project activities and milestones. In terms of material management, an inventory buffer can be introduced between the defined delivery date and the required onsite date (Yeo & Ning, 2006). An inventory buffer for construction material is crucial for improving construction performance (Horman & Thomas, 2005) and, furthermore, efficient inventory management by stakeholders in the supply network helps mitigate supply delays (Huang et al., 2012).

2.4.2 Early Sourcing and Purchase Order

An early sourcing strategy involves engaging suppliers at the initial stages of a project to reduce costs, mitigate risks, and improve quality and lead times. It is utilized to address procurement challenges in EPC firms (Azambuja et al., 2014). Jergeas (2009) suggests early purchase orders for materials as a means to mitigate delays in construction projects, while Seshadri et al. (1991) developed a model illustrating the relationship between multi-sourcing—where a company sources a particular product, service, or material from multiple suppliers—and its impact taking into account factors such as seller’s profit, buyer’s profit, and number of bids. To enhance reliability of supply, the multi-sourcing strategy is often adopted. However, it is worth noting that using a backup strategy is not prevalent in construction due to the time and cost involved in selection of suppliers.

2.4.3 Expedite

The procurement team should engage with suppliers in collaborative discussions aimed at accelerating production and shipment of materials and equipment to ensure timely delivery. This refers to purchase orders, project requirements, and schedules. Such a process demands meticulous planning as regular information sharing between contractor and supplier is crucial. In the service supply chain, backlogs are managed through capacity adjustment (Akkermans & Dellaert, 2005).

By exchanging expedite reports between the project control and procurement team, delays can be mitigated and supply chain performance enhanced through lead time reduction and sharing information on demand. Anderson et al. (2005) investigated this in their research and suggest several policies to diminish backlogs, which are often indicative of a system's responsiveness.

2.4.4 Material Decision Support Models

Experts have introduced decision support systems designed to assist managers in ensuring timely delivery of materials to construction sites. These systems consider factors like material inventory levels and storage requirements. Such models provide a deeper understanding of the procurement system's behavior over time and aid in formulating efficient guidelines and strategies for managing materials. Previous research has explored the creation of policies and their influence on various facets of construction projects. For instance, some studies have analyzed the effects of material supply decisions on construction labor productivity (Thomas et al., 1999; Thomas et al., 1989) while other studies have focused on establishing principles for onsite material management (Thomas et al., 2005). Moreover, systems intended to support material supply decisions have also been the subject of past research (Tserng et al., 2006; Polat et al., 2007). Jaśkowski et al. (2018) put forth a decision model that employs a fuzzy framework to decrease inventory costs for bulk construction materials. Conversely, other decision models focused on the process of supplier selection (Cengiz et al., 2017; Patil & Adavi, 2012; Lam et al., 2010; Ho et al., 2007).

2.5 Supply Chain Demand Forecasting in Construction Projects

Demand uncertainty is a key factor affecting supply chain performance, leading to inadequate or surplus inventories, inaccurate product forecasts, uncertain lead times, and inconsistent production planning (Lee et al., 1997; Chaharsooghi & Heydari, 2010). In the construction industry, demand uncertainty is often recognized as a core characteristic (Naim & Barlow, 2003; Green et al., 2005; Ala-Risku & Karkkainen, 2006). This uncertainty—leading to inconsistency in demand for construction materials—results from the distinctive nature of construction projects, a broad spectrum of material requirements, indeterminate construction site locations, and seasonal workload variations (Arbulu et al., 2003; Caniato et al., 2011; Vidalakis et al., 2011). These industry-specific and project-specific aspects can impede suppliers' structural decisions regarding warehouse location, material quantity and size, product variety, and mode of transportation (Silver et al., 1998). Simultaneously, these elements can impose considerable limitations on operational procedures such as inventory restocking plans, delivery scheduling, load consolidation, and backhauling, which reduces transportation efficiency (Shakantu et al., 2008). Based on the aforementioned impacts, there is a need for an efficient method for sharing project data and deliveries information throughout the project lifecycle to improve operations management.

Efficient operational management of a supply chain requires centralized synchronization of vital data (Lee & Billington, 1995). For industrial production, this usually involves data like forecasts, inventory conditions across all locations, backlogs, production strategies, delivery schedules of suppliers, and pipeline inventory. Essentially, supply chain integration requires all nodes within the network to exchange and disseminate detailed and up-to-date information. Negligence in sharing such precise information results in inaccurate demand data thus triggering what is known as the bullwhip effect (Lee et al., 1997). The bullwhip effect is a phenomenon characterized by amplified demand fluctuations up the supply chain caused by minor changes in consumer demand at the retail end (Lee et al., 1997). This effect exacerbates demand uncertainty and inventory problems, necessitating advanced planning and coordination. Additionally, Arbulu et al. (2005) concluded that the introduction of Just-in-Time (JIT) in construction is contingent upon the project team's capability to regulate supply and precisely predict demand. The introduction of JIT principles—which aim to reduce waste and increase efficiency by receiving goods only as they are needed in the production process—can help mitigate these challenges. Similar to any production system, demand and supply greatly rely on each other meaning that any form of variability will affect the successful management of the project, eventually impacting overall performance of the production system by escalating costs and duration while diminishing quality and safety (Arbulu & Ballard, 2004). Such a situation seems to be prevalent in construction supply chains where high levels of complexity and uncertainty exist. Consequently, significant deviations in plans and material delivery can occur at every phase.

To alleviate adverse effects of fluctuating demand, it is essential to be able to predict these uncertainties. Demand forecasting forms the foundation for numerous managerial decisions within the supply chain such as demand planning, order fulfillment (Narayanan et al., 2019), production planning (Donohue, 2000), and inventory management (Silver et al., 1998). Accurate and trustworthy demand forecasts offer essential insights for supply chain managers to aid their planning and decision-making processes. However, achieving high precision in demand forecasting is usually challenging due to inherent volatility and uncertainties (Syntetos et al., 2016) and the fact that the accuracy of demand forecasting can influence demand volatility (Gilliland, 2010). Due to these challenges that come along with demand forecasting, it is not a straightforward task and, as a result, many companies and planners fail to implement a scientific forecast (Armstrong & Green, 2017).

Inaccurate predictions may result in unnecessary expenses in procurement and transportation, manpower, service levels, and inventory (Torkul et al., 2016). Hence, it is critical to establish an appropriate strategy to manage volatility. Scholars and industry professionals have suggested and implemented various strategies to manage demand volatility. For instance, one method to mitigate adverse effects of demand volatility is to augment inventory levels. While this can help offset demand fluctuations, it can also result in significant costs for companies (Chopra & Meindl, 2021). Another tactic involves increasing capacity, but this approach is typically unappealing due to the high costs it imposes on the supply chain. So, even though these methods can help mitigate challenges related to demand volatility, they may not always be cost-effective.

Although there are obviously difficulties and complexities associated with implementing demand forecasting, it is an important prerequisite for strategies aiming to manage demand volatility (Hope & Fraser, 2003) as it represents the first step toward addressing uncertainty and volatility within the supply chain. Over the past several decades, a variety of models have been utilized for demand forecasting (Aye et al., 2015; Fildes et al., 2008; Huang et al., 2014; Hyndman & Athanasopoulos, 2018; Ma & Fildes, 2017; Ma et al., 2016; Pai & Lin, 2005; Syntetos et al., 2016). Yet, there is no one-size-fits-all solution that can resolve all forecasting challenges and consistently outperform all other forecasting models. Likewise, no specific methods universally outperform all others under all circumstances and conditions: certain models might perform better than others under specific conditions.

2.6 Factors Affecting Supply Chain Demand in Construction Projects

Effective demand forecasting for small equipment, tools, and consumables is imperative for successful construction project management (Vrijhoef & Koskela, 2000). The supply chain in this sector is highly sensitive to a multitude of factors that influence demand. This section explores key studies that have focused on identifying and analyzing the factors affecting supply chain demand for small equipment, tools, and consumables in construction projects.

2.6.1 Economic Conditions

The state of the economy influences the construction sector and, consequently, the demand for small equipment, tools, and consumables. Navon (2005) suggests that fluctuating economic conditions can have an immediate and prolonged impact on construction activities.

2.6.2 Labor Intensity and Tool Usage

The number of workers on a construction project directly affects demand for small equipment, tools, and consumables. The productivity of the workers might be impacted by the shortage of consumables. Accordingly, more workers may result in increased tool usage, leading to higher demand for replacements and consumables (Naoum, 2016).

2.6.3 Work Packages and Activities

The design of work packages and sequence of construction activities often dictate the types and quantities of small equipment, tools, and consumables required at various project phases. A study by Chan et al. (2004) expanded on the correlation between work packages and project success. While the study does not focus explicitly on supply chain demand, it does emphasize the role of well-planned work packages and activities in project performance, which invariably includes supply chain efficiency.

2.6.4 Project Size and Complexity

The nature and scale of a construction project directly affects the requirement for small equipment, tools, and consumables. Williams (2017) examined how project complexity and size impact demand, emphasizing that larger and more complex projects typically require a wider range and larger quantity of small equipment, tools, and consumables.

2.6.5 Project Warehouse Capacity

Project warehouse capacity plays a vital role in determining demand for small equipment, tools, and consumables. The size and storage capabilities of the warehouse influence how much inventory can be stored onsite. A larger warehouse may allow for bulk purchases and the storage of reserve supplies, reducing the frequency of replenishment orders (Kasim et al., 2005).

2.6.6 Technological Advancements

Emerging technologies can shift demand patterns for small equipment, tools, and consumables. Azhar et al. (2008) explored how Building Information Modelling (BIM) can alter demand patterns for these supplies, emphasizing the role of digitization in construction.

2.6.7 Regulatory and Environmental Factors

Legal regulations—particularly safety and environmental codes—dictate the types of small equipment, tools, and consumables that can be used. In their research, Teo and Loosemore (2001) consider how regulations affect procurement of materials and tools.

2.6.8 Supplier Relationships

Supplier collaboration can significantly influence supply chain efficiency. According to a study by Meng (2012), stable relationships with suppliers lead to more reliable logistics and delivery, affecting demand forecasting accuracy.

2.7 Time Series Forecasting in Construction Sector

As a result of the time-sensitive nature of construction projects, time series forecasting has emerged as a not only useful but necessary tool. Time series forecasting involves predicting future values based on past observations, and focuses on understanding and leveraging patterns like seasonality, trend, and cyclic behavior in historical data (Box et al., 2015). One of the most popular methods for time series forecasting is the ARIMA model. Hyndman and Athanasopoulos (2018) effectively demonstrated its application in construction supply chain demand, revealing its capacity to model various complex time-dependent models, especially when supplemented with seasonality adjustments. This section provides more detail on pivotal studies that highlight the adoption and effectiveness of time series methods in predicting supply chain demands in construction.

2.7.1 Autoregressive Integrated Moving Average (ARIMA)

Box and Jenkins (1970) established the foundation for ARIMA, stressing its adaptability in addressing non-stationary time series data. ARIMA—with its components Autoregressive (AR), differencing (I), and MA—offers flexibility to model a variety of time series structures. This is noteworthy since the volatile nature of material demand in construction projects necessitates reliable forecasting models. To this end, Hosny et al. (2023) investigated ARIMA's application in forecasting prices of construction materials. Hwang et al. (2012) also developed another automated time series model for material cost forecasting. The findings highlighted ARIMA's ability to adjust to trends and seasonality inherent in construction projects. In addition to material demand, labor demand is a critical aspect of construction. Wong et al. (2005) employed ARIMA for labor demand forecasting for construction projects, noting the model's efficiency in capturing the complex and diverse nature of the industry. For seamless construction workflows, timely procurement of equipment is also essential. Hayat and Soenandi (2018) adopted the ARIMA model coupled with non-linear Backpropagation ANN (BPNN) to predict demand for key building materials, observing the model's resilience in responding to fluctuations in demand.

2.7.2 Linear Regression (LR)

Linear Regression (LR) establishes a linear relationship between a dependent variable and an independent predictor, making it a straightforward yet potent tool for forecasting. Draper and Smith (1981) introduced LR as a staple method for modeling and interpreting quantitative data in various domains including construction. Construction machinery is foundational for planning processes and vital in the administration of core operations in the construction industry. Therefore, precise forecasting of equipment is critical for organizations to ensure site operations can continue without disruption. A study conducted by Aktepe et al. (2021) utilized LR analysis to accurately predict future customer requests for spare parts. Data spanning from 2010 to 2018 were employed for the analysis with demand predictions made for 2018 used as a benchmark against actual figures. Another study—focused on Singapore's residential, industrial, and commercial construction demand—showcased the application of LR utilizing quarterly time series data spanning from 1975 to 1994. The primary aim of this analysis was to determine the dependability of this method in predicting sector-specific demand (GOH, 1998). Flanagan and Norman (1983) utilized LR to investigate the accuracy of quantity surveyors' predictions of project costs. This helps with determining early warning signs as a means of getting projects back within budget. Finally, a study conducted by Al-Momani (1996) was used to predict the construction cost of school buildings across years. Their model was built using records of 125 school projects gathered in Jordan from 1988 to 1994. This model would then be used for predicting cost of other projects.

2.7.3 Artificial Neural Networks (ANNs)

ANNs are computational models inspired by the human brain's structure and function. They are increasingly utilized in time series analysis for their ability to learn and model complex patterns and trends from historical data. This being said, ensuring that materials are available in the right quantities at the right time is essential for project success. In their study, Aktepe et al. (2021) focused on the production facility for construction machinery spare parts, where demand forecasting played a pivotal role in operational planning and inventory management. Their research was aimed at accurately predicting future customer demands for spare parts using historical sales data. They developed ANNs to handle complex time series data and the analysis included data spanning from 2010 to 2018. Another study—conducted by GOH (1998)—addressed the construction industry's inherent volatility by making precise demand forecasting for stakeholders such as developers, builders, and consultants. The study addressed the residential sector in Singapore and used ANNs as the forecasting method, which showed potential by generating accurate demand predictions. An additional study evaluated forecasting capabilities of the Neural Network Autoregressive (NNAR) model in predicting construction output over the medium term. The study used quarterly time series data of Hong Kong's construction output from the first quarter of 1983 to the fourth quarter of 2014. The findings revealed that the NNAR model provides reliable and accurate forecasts for various categories of construction output (Lam & Oshodi, 2016).

Perhaps most significantly, a study completed by Ihnatovich (2017) discussed predicting the construction equipment market demand using economic indicators to navigate market shifts effectively. The study ultimately provides valuable insights for construction equipment manufacturers, distributors, and suppliers, enabling a deeper understanding of market dynamics and anticipating market demand. Such foresight allows businesses to refine their strategies, optimize production capacities, allocate resources efficiently, maintain optimal inventory levels, and minimize costs while enhancing profitability and market competitiveness. Moreover, the study confirmed that demand for construction equipment is significantly impacted by economic conditions. They implemented ANNs for forecasting and successfully predicted the sales of construction equipment up to a year in advance across various countries including Germany, the United Kingdom, France, Italy, Norway, Russia, Turkey, and Saudi Arabia.

2.8 Machine Learning (ML) Algorithms in Construction Projects

Construction projects consist of diverse activities that are interdependent, impact one another, and are influenced by uncertainties like weather conditions, geological features, and human factors. Therefore, effective construction management is critical to efficiently achieve construction goals and is fundamental to success of a project (Bush, 1973). However, due to uncertainties and the dynamic nature of the construction industry, practical construction management challenges are intricate and hard to predict (Li, 1996). Various tools have been developed and successfully applied to tackle these problems in construction management. Artificial Intelligence (AI)—a tool that has been somewhat underutilized in the construction industry thus far—refers to computer system designs that manage and solve problems intelligently, mimicking processes that occur in the human brain. With AI technology continuing to advance, applying AI paradigms is becoming suitable in addressing construction management challenges (Haykin, 1998; Tommelein et al., 1992). In fact, ML—a subfield of AI—is one of the top technologies widely employed across various industries such as railways, aviation, and medicine (Sanni-Anibire et al., 2020). It is used for data modeling and developing mathematical representations of data that can then be used by computers to deliver accurate predictions. Despite the broad acceptance of ML algorithms over the last two decades, the adaptation of ML for construction management is still in the early stages (Jantan et al., 2009; Iwu, 2016; Gondia et al., 2020).

ML algorithms are categorized into three main types: supervised learning, unsupervised learning, and reinforcement learning. Each differ in their approach to learning and problem-solving. Supervised learning, the most commonly used category, involves training a model on an already existing labeled dataset (Xu et al., 2021). The objective is for the model to learn to map inputs to outputs, making it suitable for problems that involve regression and classification. Examples of supervised learning algorithms include Linear Regression (LR) for continuous output prediction, Logistic Regression (Cox, 1958), Support Vector Machines (SVMs) (Cortes & Vapnik, 1995), and Artificial Neural Networks (NNs) for classification tasks (Adeli & Yeh, 1989). Unsupervised learning, in contrast, deals with unlabeled data where the goal is to explore structure and patterns within the data. Without predefined labels, the algorithm tries to organize data, often by finding commonalities among different inputs (Xu et al., 2021). Common unsupervised learning tasks include clustering, a process by which the algorithm seeks to group data points into distinct categories based on their features, and dimensionality reduction, where the algorithm simplifies inputs by reducing the number of variables. Examples include K-Means clustering (MacQueen, 1967) and Principal Component Analysis (PCA) (Pearson, 1901). Reinforcement learning is a different paradigm where an agent learns to make decisions by performing certain actions and receiving feedback from said actions in the form of rewards or penalties (Xu et al., 2021). The algorithm learns a policy, which tells the agent what action to take according to certain conditions. This type of learning is much like a trial-and-error learning process and is often used in areas such as robotics, gaming, and navigation. It involves evaluating actions based on the reward system and updating the strategy accordingly to maximize cumulative reward (Xu et al., 2021). Together, these three types of ML algorithms form a comprehensive toolkit that can be used for addressing a wide range of data analysis and problem-solving issues, from predictive modeling and data classification to autonomous decision-making in complex environments. The following subsections provide further detail on several ML methods.

2.8.1 Artificial Neural Network (ANN) Forecasting

Networks are an effective strategy for decomposing complex systems into more understandable subsets, and are comprised of nodes as well as connections between them (Sanni-Anibire et al., 2020; Wu & Chan, 2009). Nodes serve as computational entities in networks, while connections facilitate information flow among these nodes (Haykin, 1998). In the ANN approach, nodes are referred to as artificial neurons, which are computational models inspired by biological neurons. Within artificial neurons, inputs are multiplied by corresponding weights and are then processed by a specific activation function (Patterson, 1996). ANNs consist of such artificial neurons (Gershenson, 2003). ANNs were designed to process information in a way that mirrors the human brain and are composed of a set of interconnected input/output units with each link having an associated weight. The ANN method allows for modeling of large, intricate problems involving numerous interdependent variables, and excels in the areas of prediction, pattern recognition, data compression, preliminary resource planning, and decision-making (Chukwu & Adepoju, 2012; Mourya & Gupta, 2012; Paliwal & Kumar, 2009; Wu et al., 2021).

Many variations of ANNs have recently been developed including activation functions, hybrid models, accepted values, and learning algorithms (Wu & Chan, 2009). For instance, the Recurrent Neural Network (RNN) is one of the most advanced ANN models (Hibat-Allah et al., 2020). This type of network is widely used in speech recognition, natural language processing, and for modeling sequential data (Zaremba et al., 2014). The algorithm can learn sequential features of data, utilize patterns to predict probable scenarios, and permits previous outputs to serve as inputs while maintaining hidden layers (Apaydin et al., 2020). Although ANNs are applicable in both classification and regression problems, the models typically demand extended training periods, and interpreting the meanings of the nodes' computed weights can be challenging (Han et al., 2011; Mourya & Gupta, 2012).

2.8.2 Linear Regression (LR)

An essential component of supply chain management is demand forecasting, which predicts future material, equipment, and labor needs. Materials constitute a large portion of construction costs, and forecasting their demand is therefore essential. LR—a foundational statistical tool—has been employed in various studies to address demand forecasting (Hua & Pin, 2000). Significantly, Skitmore and Ng (2003) demonstrated that LR could effectively predict prices and demands of essential construction materials by using economic indicators and historical data showing material usage rates. Persad et al. (1995) also applied LR models to forecast labor demand by using project features such as size, complexity, and type as predictors.

LR is considered a simple algorithm to apply, yet it does have some limitations. One of the major limitations is that assumptions of linearity may not always hold true in complex supply chain systems. The construction industry—influenced by external shocks such as political events, economic downturns, or unaccounted factors—may not always follow linear patterns (Wong et al., 2011). However, as computational capabilities grew, hybrid models combining LR with other techniques emerged as a solution to this issue. For instance, Tsai and Wu (2009) combined LR with ANNs to improve forecasting accuracy for cement demand, reflecting the industry's move toward more complex and integrative models. While LR provides a foundational method for forecasting supply chain demand, the inherent complexity of construction projects necessitates continuous evolution and adaptation of forecasting models. As the industry advances and data becomes more accessible, the integration of LR with other techniques promises more accurate and dynamic forecasting solutions.

2.8.3 Decision Tree (DT)

A Decision Tree (DT) operates by dividing source data into subsets based on values of input features. This process is inherently recursive, resulting in a tree-like model of decisions and their potential outcomes. Its hierarchical nature allows for easy visualization and interpretation, making it particularly valuable for stakeholders who are non-experts in ML (Quinlan, 1986).

Desai and Joshi (2010) utilized DTs to analyze labor productivity which impacts labor demand forecasting. The study revealed that the model could capture non-linear patterns in demand for labor affected by factors such as location, temperature, and age group. This was an important finding as labor constitutes a significant portion of construction expenses. Additionally, Shehadeh et al. (2021) proposed a DT model to forecast equipment demand by evaluating their residual values. This study used several features to be included in the model including, but not limited to, equipment type, manufacture, model, and age. Another subset of studies used DTs—mainly in the manufacturing industry—for predicting the price of several materials such as cotton, vehicles, and used cars (Deepa et al., 2023; Alshboul et al., 2023). The studies highlighted the model's ability to handle multiple factors such as prices in different location, car type and model, project size and complexity, and regional constraints. In previous studies, DTs have also been compared against other forecasting models. For example, in a review study conducted by Boyko and Lukash (2023), which forecasted the cost of heavy machinery based on several features, DTs were compared against several algorithms. Although each model had strengths, hybrid approaches combining DTs with other models showcased potential for enhanced accuracy.

2.8.4 Random Forest (RF)

The random forest (RF) algorithm is renowned as one of the most accurate forecasting techniques, and can be used for classifications and regressions that consist of multiple DTs (Wang et al., 2016). In the RF algorithm, each DT has a root node, which then divides into branches based on all conceivable outcomes. This division is repeated for each branch until reaching a node where all instances share the same classification (Witten et al., 2011). A randomly chosen subset of features in each tree is utilized to establish optimal threshold for data splitting. Consequently, numerous trees are trained, each making a distinct prediction (Probst et al., 2019). It is worth mentioning that trees that perform best in certain segments of the sample space may provide imprecise estimates in other sections (Pedregosa et al., 2011). Finally, the ensemble's prediction is obtained by aggregating—via majority vote or averaging—the predictions of the ensemble (Wang et al., 2016). The overall aim of the RF algorithm is to offer a solid predictive model less prone to overfitting by averaging several DTs, each of which might individually exhibit high variance (Bonaccorso, 2018). Additionally, the RF algorithm's division method also incorporates a moderate level of randomness compared to a single DT (Probst et al., 2019).This being said, the primary benefits of using RF algorithms include: simplicity of rules produced, as well as learning and classification processes with no constraints on numerical or categorical data (Gorunescu, 2011); effective operation on large datasets without being sensitive to noise or overfitting; and ability to handle a considerable number of inputs while having fewer parameters compared to other ML algorithms such as ANNs.

Due to the potential associated with DTs, there is now a trend toward leveraging the RF algorithm. As a result, there are construction industry studies wherein researchers attempt to use this algorithm to address various problems such as predicting occupational accidents, construction project delay risks, and project costs (Yaseen et al., 2020; Huang & Hsieh, 2020; Meharie & Shaik, 2020).

2.8.5 K-Nearest Neighbor (KNN)

The K-Nearest Neighbor (KNN) algorithm is a fundamental and straightforward classification technique rooted in predicting new records based on similarity metrics. It was initially designed to address reliable parametric estimations of probability density during discriminant analysis when these were unknown. In the KNN framework, k denotes the count of neighbors participating in the majority voting procedure (Guo et al., 2003). More specifically, the KNN classifier requires a metric d and a positive integer k value (Kubat & Cooperson, 2001). When a new input needs classifying, the distance between the new record and training records is computed. Based on the pre-determined threshold for the number of neighbors (which is k), the k with closest records of the smallest distances are identified and selected. The class that has the majority of samples is then chosen as the new input class. For instance, if k is 3, the prediction for a new record will be determined based on the three closest neighbors. Consequently, the only parameter to be fine-tuned is the value of k. A smaller k value might bias the model towards outliers, while a larger k value could make the modeling process computationally intensive (Guo et al., 2003). The k value that delivers optimal performance can be determined through trial and error. Some existing studies suggest different formulas for determining the optimal *k* value (Zhu et al., 2010; Lall & Sharma, 1996). Indeed, various methods can be used to compute the distance in the KNN algorithm, but the Euclidean distance metric is perhaps the most simple and straightforward method for calculating distances in a multi-dimensional input space.

Numerous studies employed the KNN algorithm across various research fields for solutions such as learning methods, mapping, and recognition (Dang et al., 2005; Franco-Lopez et al., 2001; Lee & Scholz, 2006). The KNN algorithm excels in handling classification and learning from extensive datasets (Rosa et al., 2003). Thus, it is a widely adopted method capable of producing competitive results even when compared to the most sophisticated ML methods (Song et al., 2007; Duda et al., 2001).

2.8.6 Support Vector Regression (SVR)

Support Vector Regression (SVR) is a robust supervised learning method where the outcome is numeric. It employs advanced algorithms to discover patterns within intricate datasets and has been leveraged across various fields including demand forecasting, as highlighted by Villegas et al. (2018). Kandananond (2012) conducted a comparative study between two ML algorithms, Multilayer Perceptron (MLP) and SVR, comparing them against the traditional ARIMA prediction approach. By assessing consumer demand forecasts for six products, the research utilized Ljung-Box-Q statistics to test data autocorrelation prior to model application. The evaluation metric, MAPE, confirmed SVR's superiority over ARIMA and MLP in forecasting for the majority of products.

A novel combination of SVR and Particle Swarm Optimization (PSO) algorithm was then proposed by Chen and Liu (2013). Their PSO-SVR model was employed to predict coal transportation demands, taking into account variables like railway freight turnover volume and coal consumption among others. Utilizing data from 1995 to 2011 and incorporating the Radial Basis Function (RBF) kernel into their predictive model, the combined algorithm proved its mettle by outperforming ANNs Back Propagation (BP) both in terms of forecast accuracy and error reduction. Carbonneau et al. (2008) analyzed the efficacy of cutting-edge ML algorithms including ANNs, RNNs, and SVR in forecasting supply chain demand datasets. The study compared these contemporary models against classical methods such as LR and Moving Averages (MAs). The results revealed promising potential of both RNNs and SVR in demand forecasting.

2.9 Research Gaps

Small equipment, tools, and consumables have consistently been vital for executing construction projects. The effective and efficient management of these items presents a challenge, though, one which can determine the success of any given project. In the past, extensive research in construction has been conducted regarding resource allocation. However, this research has overlooked the estimation of these consumables, leaving this research area relatively unexplored. Although forecasting models are one of the solutions that could address estimation issues in a practical manner, there is a noticeable lack of forecasting models that have been developed to focus on predicting total demand for small equipment, tools, and consumables for the entirety of the project while taking into account its general specifications. Having only a rough estimation of demand for the whole project—without an understanding of how this demand will be distributed over time—does not provide an efficient strategy for project management. Some professionals rely on their expertise to estimate quantities of these consumables but, again, this is subjective and changes from one professional to another, as well as from one project to another. Additionally, it is often the case that, during early stages of a project, there may not be sufficient detail about project activities. This can result in substantial errors and misallocated quantities when making predictions at this level. Consequently, there is a need for a forecasting model that can provide accurate estimates throughout a project.

Current research aims to improve industry practices by exploring relevant literature to develop a more robust and practical forecasting model that aids in predicting monthly demand for small equipment, tools, and consumables at the project level for construction projects. Acquiring insights into demand at project level is not only more efficient, but also more practical than making broad and rough estimations for the entire project based on project managers' experience. This approach removes the need for detailed information about every planned activity during the project. The intention of this thesis is to devise a generic model that provides project managers with reliable results. This goal will be achieved by: (1) proposing a structured approach to assist project managers with collecting project information in a way that leads to more accurate predictions and (2) developing a forecasting model that employs ML algorithms to predict consumables required over the duration of the project. While some forecasting models use LR methods, others employ a non-linear approach to forecast resources. This research—recognizing the strengths of various approaches—investigates an array of ML algorithms including both LR and ANNs, chosen specifically for their ability to handle large datasets and deliver accurate forecasts.

CHAPTER 3: METHODOLOGY

3.1 Introduction

The aim of this study is to forecast demand of small equipment, tools, and consumables for the work tasks of a project by leveraging historical project data that includes important features for forecasting demand of these consumables. This research encompasses numerous phases including: comprehending industry procedures and methodologies; studying researchers' methods in forecasting required small equipment, tools, and consumables resources at the project level; and scrutinizing the factors impacting quantities for different small equipment, tools, and consumables resources. Then, historical data is collected based on industry practice and an agreement with the industry partner involved in the study. After data collection, data cleaning and preprocessing EDA are performed to gain understanding of the data through trends analysis and visualizations. Following, time series analysis is conducted using techniques such as ARIMA, LR, and ANN. ML algorithms are employed to pinpoint significant elements and formulate a model capable of forecasting quantities for varying types of small equipment, tools, and consumables. In the final phase, various metrics are used to evaluate and compare the models' performance. Figure 2 summarizes the aforementioned phases.

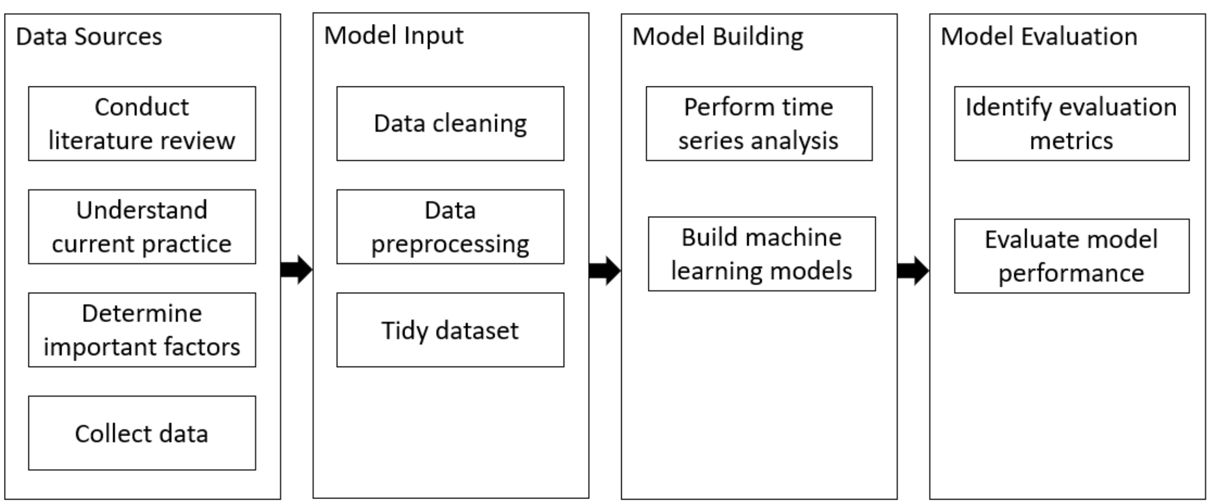


Figure 2 Proposed methodology

The purpose is to identify project characteristics that impact quantities of small equipment, tools, and consumables for construction projects. The model can define the features that have significant impacts on quantities forecasting. It is worth mentioning that the results of the model should be logical and reliable to construction professionals—particularly project managers—to provide them with an accurate demand forecasting model for small equipment, tools, and consumables resources.

This chapter focuses on: (1) exploring demand forecasting strategies; (2) pinpointing crucial project factors that impact the demand forecasting process; (3) collecting historical project information for utilization in the forecasting model; (4) performing data cleaning and preprocessing for data analysis; and (5) explaining factors in the dataset and presenting data through graphs/charts to elaborate upon the impact of said factors. The procedure for forecasting model formulation is illustrated in depth and encompasses the data preparation technique, feature selection, outlining model inputs, and the ML algorithms employed to predict required quantities for small equipment, tools, and consumables resources in construction projects.

3.2 Factors Affecting Required Consumables

Forecasting quantities of small equipment, tools, and consumables has been among the concerns of project managers for some time. Various methods have been followed by industry practitioners and researchers to estimate small equipment, tools, and consumables required for a project. One of the objectives of this research is to delve into the potential factors affecting quantities of small equipment, tools, and consumables on a project level, and to determine factors that could be utilized to forecast small equipment, tools, and consumables. This section is an exploration into prevailing industry practices, which is conducted in order to pinpoint project attributes that ought to be considered when estimating required quantities for construction projects.

3.2.1 Industry Practices in Demand Forecasting of Consumables

Exploring current industry practices is the first step in detecting key attributes required for accurate prediction. The aim is to implement an investigation approach regarding impacting factors considered by industry experts when estimating labor resources required to complete a work package. Informative discussions with experienced project managers working in the largest construction companies lead to a clear understanding of current resource allocation methodologies in the industry. Typically, each organization has its own way of identifying staffing requirements and adopts qualitative and quantitative techniques including heuristic rules based on their experienced managers, regression models from previous projects, and their own framework to rank projects based on multiple project characteristics. Demand forecasting techniques used by industry organizations include the following:

- Expert judgment: Construction organizations generally place high reliance on their experts' evaluations and engage experienced specialists who have completed similar work, seeking their opinions regarding which resources are needed.
- Project management software: Some software include features designed to assist project managers in estimating required quantities as construction schedules are resource-loaded.
- Bottom-up estimating: This involves dividing complex tasks into smaller ones and calculating quantities for each activity. It is a process of estimating individual activity resource needs and then aggregating them to a total estimate. This method is favored among construction companies due to its accuracy and simplicity. However, it demands a considerable amount of time to perform bottom-up estimating as every activity must be assessed and estimated meticulously to be incorporated in the aggregation process.
- Referenced estimating data: Many project managers in the construction industry employ such data to estimate quantities. They depend on articles, books, or journals for analysis.

The process of demand forecasting of small equipment, tools, and consumables includes various techniques: qualitative and quantitative. However, each of these frameworks have limitations that impact accuracy and lead practitioners to inaccurate estimation from employing their methodologies. The main setbacks are as follows: solely depending on project managers' knowledge and experience, and not evaluating actual project values compared to estimated values.

3.2.2 Factor Selection

Key attributes impacting small equipment, tools, and consumables demand are identified through literature review analysis, as well as determining factors applied by industry experts through understanding their methods in estimating quantities. One of the objectives of current research is to gather significant key features from historical data to forecast quantities required for a given construction project. As mentioned in the literature review chapter, there are barely any prediction models developed by researchers that can estimate quantities of small equipment, tools, and consumables required for construction projects. Yet, models exist for estimating other materials in the construction industry. These models typically account for elements such as project type and budget but often exclude other critical aspects like complexity and project delivery method (Elkholosy, 2020; Golabchi & Hammad, 2023). Furthermore, current models predominantly offer projections for overall materials demand for a project.

In contrast, this research seeks to develop a model for predicting small equipment, tools, and consumables needs for each individual time increment—be it daily or weekly—throughout the project's lifecycle. Some commonly considered determinants in existing models include: economic conditions (Navon, 2005); labor intensity and tool usage (Naoum, 2016); work packages and activities (Chan et al., 2004); project size and complexity (Williams, 2017); project warehouse capacity (Kasim et al., 2005); technological advancements (Azhar et al., 2008); regulatory and environmental factors (Teo & Loosemore, 2001); and supplier relationships (Meng, 2012). It is worth noting that some factors mentioned cannot be generalized to all types of projects.

3.3 Data Collection Process

Since quality of data significantly impacts accuracy and dependability of the developed forecasting model, historical data—encompassing various features—must be collected from contractors and then analyzed to gain insight into the different features. Data collection is the fundamental step for then being able to employ the data in the forecasting model as a means of estimating small equipment, tools, and consumables quantities.

The information required for collection—which is also based on data availability—is first discussed with subject matter experts where the aim is to collect historical data at the project level. Significantly, the data flow between manufacturers, vendors, and contractors until it reaches the workface onsite is based on consumables availability. On each site, there are multiple tool cribs for workers to use throughout the project. Each of these cribs contains different consumables/items required by workers to perform various tasks. Once quantities start running low in the cribs, the contractor's warehouse can be used to restock these items. If the contractor's warehouse cannot fulfil the ordered consumables, the request can be directed to the vendor, and the vendor can then directly ship the consumables to the project site or contractor's warehouse. The vendor may also use the distribution center to satisfy orders. In some cases, these distribution centers are located near the project site. As a result, it is more convenient to ship from the distribution center directly.

In cases where the vendor cannot satisfy the requested items, the vendor must reach out to the manufacturer. The manufacturer is the last entity in the supply chain and covers both the manufacturer's distribution center and the manufacturing facilities where consumables are produced. Figure 3 delineates the flow of consumables across the aforementioned entities.

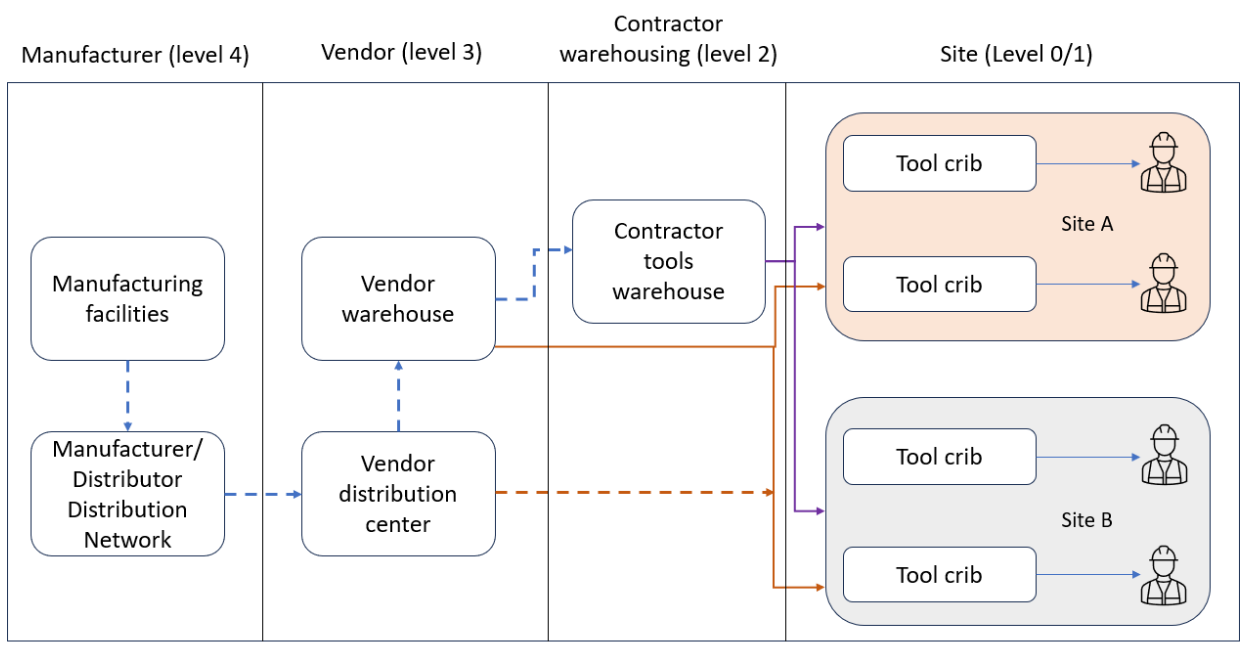


Figure 3 Consumables flow across entities

In this research, historical data regarding consumables information are collected via spreadsheets provided by the industry partner, which are primarily exported from the company's Enterprise Resource Planning (ERP) systems.

3.3.1 Essential Attributes for the Forecasting Model

The following are the imperative attributes of a project: consumable category, work package schedule, project duration, total work hours, manpower, budget for each order, and type of project. This study focuses on collecting attributes from completed historical projects. Data related to quantities of small equipment, tools, and consumables are collected and used to train the forecasting model.

3.3.2 Data Exploration

After performing a literature review analysis and having discussions with industry experts, the following factors have been chosen for further exploration in this study:

- Consumable category: This denotes the specific type of consumable required for various tasks such as gloves, helmets, or discs.
- Work package schedule: This indicates the timeline for different tasks or activities in the project, helping determine when specific consumables might be needed.
- Project duration: This represents the total timespan for project completion. It is crucial for understanding how long a specific consumable might be in demand.
- Total work hours: This refers to cumulative hours anticipated for the project, which can indicate the intensity and therefore the potential consumption of consumables.
- Manpower: This highlights the number of workers or professionals engaged in the project. A larger workforce might indicate higher consumption rates of certain tools or materials.
- Budget for each order: This provides financial perspective, detailing how much is allocated for each order. It can be an indicator of the quantity of the consumables expected to be used.
- Type of project: This specifies the nature of the construction project such as oil and gas, power, or agricultural chemicals. Different project types have unique consumable requirements and usage patterns.

Some of the factors that might have an impact on small equipment, tools, and consumables requirements are not considered in the case study of this research due to unavailability or confidentiality. For example, project complexity—which could be a factor affecting required manhours—is difficult to collect as contractors do not include such information in their records and each contractor might have its own definition of complexity. Thus, only the seven factors listed above are used as input to the forecasting models in this research. The selected factors have been studied by experienced project managers to ensure accuracy. Additionally, the project managers have been requested to provide feedback regarding any overlooked factors that require consideration.

3.3.3 Data Sources

Each contractor possesses their unique methodology for collecting project data. Some might deploy their own developed collection systems, while others may engage ERP systems such as Oracle and Microsoft. The industry partner engaged in this research has devised its own tracking system consisting of various spreadsheets, each tailored for collecting specific data. Every record has a unique order and project code. The initial spreadsheets available include: type of consumables (Line Description), project ID (Business Unit), date order is placed (Order Date), date order is needed by (Request Date), date supplier promises to deliver (Promised Date), unit cost of consumables (Unit Cost), and quantity of ordered consumables (Quantity Ordered).

3.3.4 Missing Data

Although contractors generally track orders and quantities meticulously, certain attributes—such as complexity—are often not collected. Additionally, a uniform format for data collection across all projects may not be followed. For instance, without a well-defined format, naming various orders can result in difficulty and confusion since each project manager may employ their own naming conventions. Establishing a predefined data acquisition system can help solve these challenges thus providing a clean, well-structured dataset necessary for accurate analysis, swift challenge identification, and future actions. The industry partner involved in this research developed a robust tracking system internally and then refined it over years, resulting in limited missing values in their collected data. However, as mentioned earlier, certain attributes are often absent from the contractor's tracking systems and records. Consequently, these records are removed from the dataset. Additionally, outlier detection processes are completed to remove some records and ensure more accurate data is fed to the model.

3.3.5 Data Transformation

As noted above, the initial dataset provided by the industry partner includes the following features: Line Description, Business Unit, Order Date, Request Date, Promised Date, Unit Cost, and Quantity Ordered. These features are not adequate for building a data-driven model. As such, several modifications, transformations, and features must be added to improve model input.

Further explanation regarding these transformations and modifications is provided in the case study chapter.

3.4 Model Development

The aim of this research is to pinpoint substantial factors influencing small equipment, tools, and consumables quantities through feature selection methodologies, as well as construct time series and ML models to forecast small equipment, tools, and consumables quantities required for projects. The proposed model is developed by leveraging time series analysis and ML algorithms, and will, in turn, assist project managers in allocating budget proficiently during preliminary planning phases.

In the previous section, seven factors affecting small equipment, tools, and consumables quantities were identified. A framework is detailed for developing a forecasting model for small equipment, tools, and consumables quantities where the identified factors are used as inputs, as illustrated in Figure 4. The collected dataset first undergoes cleaning and preprocessing steps, and any outliers within the dataset are pinpointed through outlier detection techniques. Then feature selection procedures are employed to analyze the features bearing significant impact on the performance of the predictive model. Once this step is completed, a range of time series techniques and ML algorithms are examined regarding their capabilities to opt for the most suitable algorithm for this study. The selected algorithm is then employed in the formulation of the forecasting model. Finally, the performance of the forecasting model is evaluated by various metrics.

The coming subsections are organized as follows: data preprocessing and outlier detection are explained; feature selection strategy is detailed; several time series techniques and ML algorithms are developed and trained using the training dataset; and the performance evaluation of the forecasting model and validation procedure are studied in depth.

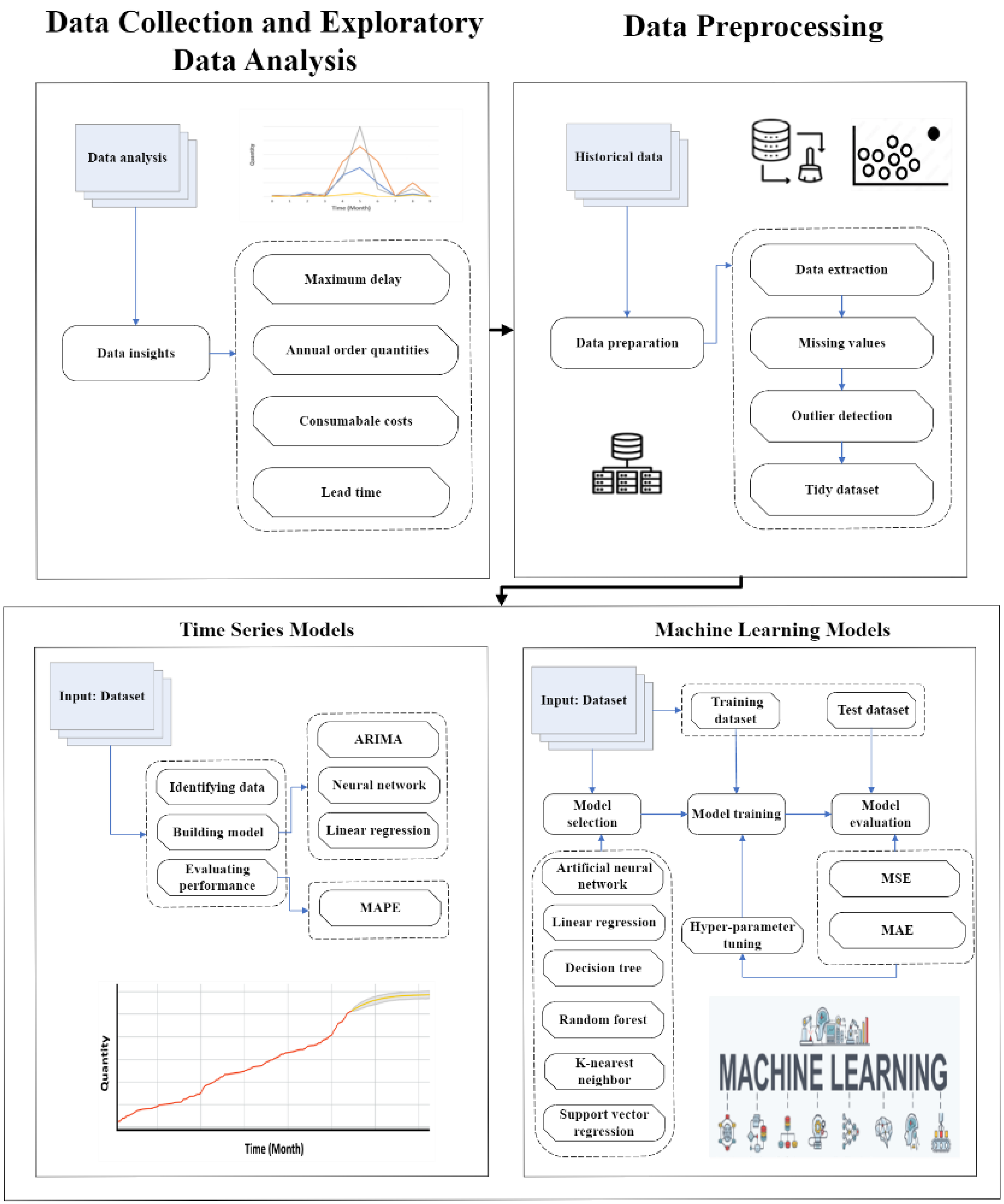


Figure 4 Methodology of developing the forecasting model

3.4.1 Data Preprocessing

Most ML algorithms require data cleaning and preprocessing before training can take place as the efficiency of these algorithms is influenced by quality of input data (Wu et al., 2022). Therefore, it is imperative that the data is prepared properly and aligns with the algorithm to avoid misleading or inaccurate outcomes. One of the stages in data preparation entails addressing missing values encountered in the dataset. The occurrence of missing information is attributable to inadequate tracking throughout project duration. Additionally, outliers may be identified in the collected datasets, which are usually a result of shortcomings in tracking and storing procedures. Subsequent to this, it is essential to scale the attributes in the dataset to prevent the forecasting model from assigning unjustified weights to attributes owing to their larger values. Moreover, setting up categorical attributes is essential. The management of categorical attributes is a prerequisite of preprocessing, especially given that certain algorithms only accept numerical inputs such as regression and ANNs. Consequently, necessary alterations should be executed on the dataset prior to advancing to the training phase.

This being said, proper formatting of the input dataset is fundamental. This process consists of several crucial steps aimed at acquiring acceptable input data for the model and includes the following: deleting records with missing values, removing outliers, managing categorical project attributes, and normalizing the training dataset.

3.4.1.1 Missing Values

In the provided dataset, certain records are missing features for specific orders related to the project. This occurs due to improper tracking and monitoring of the orders. Therefore, a number of records are removed from the input data as it is impracticable to obtain this information through alternative means. The small equipment, tools, and consumables resources forecasting must be undertaken with a profound understanding of existing data and its limitations. The planning in data handling plays a significant role in ensuring predictions for quantities are not only accurate but also reflective of actual needs in construction projects, thereby enabling a more effective and efficient quantity estimation in early stages of project planning.

3.4.1.2 Outlier Detection

Outlier detection is an important step in the data preprocessing stage of ML, aimed at identifying anomalous or atypical observations in the data. These outliers arise due to various factors such as data entry errors, measurement errors, or inherent discrepancies in the underlying process. They have the potential to significantly impact the training and performance of ML models if included. Various techniques exist for outlier detection including statistical methods, clustering algorithms like K-Means, and ML algorithms such as Isolation Forest and One-Class SVM. These methods endeavor to either model the "normal" data to distinguish outliers or directly identify outliers based on certain criteria. Handling outliers appropriately—by removing, correcting, or segregating them—is essential to build robust models capable of generalizing efficiently from the training data to test data. The choice of outlier detection method depends on the nature of the data and the problem domain, making it an area that requires careful consideration in the ML process.

3.4.1.3 Features

Handling categorical attributes may pose more challenges compared to numerical ones, necessitating certain techniques to convert them into numerical variables. This conversion is a pivotal process in data preparation. Some features in the collected dataset encompass nominal attributes such as consumable category. Nominal attributes embody discrete categorical values labeled without any inherent order. Conversely, ordinal attributes encapsulate categorical values that bear a distinct order amongst them, where the interval between these values could influence the small equipment, tools, and consumables forecasting model.

3.4.1.4 Normalizing and Data Split

Upon conversion of categorical attributes into numerical counterparts, the entirety of values within the dataset transition to a numerical form. These features exhibit variation in magnitude since each has a distinct unit. It is paramount to execute scaling on these features prior to the training of an ML model, with normalization standing as a prevalent method for this scaling step. This involves deducting the mean and then dividing by the standard deviation for each respective feature.

ML algorithms are indifferent to the units as their operation relies solely on the magnitude of the values. Hence, all project features are standardized to provide a favorable environment for the ML algorithm to function optimally. This standardization is realized by scaling each of the project features to a mean of 0 and a standard deviation of 1. The computation of the mean and standard deviation should be exclusively based on the training data, ensuring that the models have not been subjected to values within validation and test datasets. It is debatable, however, whether the model should remain uninformed of future values within the training set during the training phase, proposing that this normalization should harness MAs.

Subsequently, the training dataset is arbitrarily partitioned into three distinct subsets: the training set, the validation set, and the testing set. The training set serves as a subset of the dataset dedicated to construction of predictive models. Within the developed model, the training subset encompasses 80% of the records. The validation set is a subset of the dataset designated to evaluate the model's performance during the training phase. It provides a test bed for refining a model's parameters and electing the best performing model. The validation set embodies 20% of the dataset in the prevailing model. The testing set is a subset of the dataset deployed to evaluate the anticipated future performance of a model. In the developed model, 20% of the dataset is allocated to the test subset for small equipment, tools, and consumables demand forecasting.

3.4.2 Time Series Forecasting Methods

Time series forecasting is a quintessential approach in predictive analytics—especially when data exhibits chronological sequences—and is tailored to predict future data points based on previously observed values. Indeed, it captures patterns, seasonality, trends, and cyclic behaviors present in historical data, and is able to deconstruct data into its fundamental components: trend, seasonality, and residuals. Time series data is structured with time as its primary axis, where observations are recorded at consistent intervals. These intervals could range from milliseconds (such as in high-frequency trading data) to decades (such as in climate change studies). Recognizing and modeling these components are crucial for enhancing forecast accuracy.

Significantly, the granularity of data and its temporal structure makes time series forecasting distinct from other predictive modeling techniques. Given the temporal nature of data, model evaluation in time series forecasting requires specialized techniques including rolling-forecast origin and walk-forward validation. Various methods exist for time series forecasting, each with its unique assumptions, complexities, and applicability. From classical statistical methods, like ARIMA, to advanced ML techniques, such as ANNs and Long Short-Term Memory (LSTM) networks, the choice of methodology often hinges on the nature of the data and specific forecasting objectives. Time series forecasting is not without its challenges, though. For example, challenges such as stationarity, autocorrelation, and structural breaks can influence model efficacy. Moreover, real-world time series data can be noisy, sparse, and may have missing values.

The following subsections discuss these challenges and potential approaches to ensure that time series forecasts are both robust and reliable, as well as offer a panoramic view of time series forecasting thereby arming the reader with the knowledge and tools to navigate the complexities of predicting future data points based on temporal sequences. Through a blend of theoretical underpinnings and practical insights, the aim is to foster a comprehensive understanding of this critical predictive analytics domain.

3.4.2.1 Autoregressive Integrated Moving Average (ARIMA)

ARIMA is a cornerstone methodology for time series forecasting. Rooted in both AR and MA models, ARIMA is tailored to address data that exhibits temporal dependencies, trends, or seasonality, making it especially apt for tasks such as demand forecasting.

3.4.2.1.1 Basic Concept

- Components: ARIMA is an amalgamation of three primary components including AR, Integrated (I), and MA. AR uses the relationship between an observation and a certain number of lagged observations. Integrated (I) makes the time series stationary by differencing. MA uses the correlation between an observation and a residual error from an MA model applied to lagged observations.
- Stationarity: A prerequisite for ARIMA is a stationary time series, meaning statistical properties like mean and variance are consistent over time. The Integrated (I) component aids in achieving this by differencing the series until it becomes stationary. The model can be represented as,

$$\emptyset(B)(1-B)^d y_t = \delta + \theta(B)\epsilon_t \tag{1}$$

where y_t is the time series, B is backshift operator $B^K y_t = y_{t-k}$, d is degree of differencing, $\emptyset(B)$ is autoregressive operator of order p, $\emptyset(B) = 1 - \emptyset_1 B - \emptyset_2 B^2 - \cdots - \emptyset_p B^p$, $\theta(B)$ is MA operator of order q, $\theta(B) = 1 - \theta_1 B - \theta_2 B^2 - \cdots - \theta_q B^q$, ϵ_t is error term at time t, and δ is a constant term (if included in the model).

3.4.2.1.2 Learning Process

- Model Parameters: ARIMA is defined by three key parameters—denoted as p, d, q—where p is the order of the AR term, d is the number of differencing required to make the series stationary, and q is the order of the MA term.
- Akaike Information Criterion (AIC) and the Bayesian Information Criterion (BIC): AIC and BIC are often used to assess and choose the best combination of parameters that will result in the optimal fit for the model.

3.4.2.1.3 Training

- During the training phase, ARIMA utilizes historical data to estimate parameters that will minimize the prediction error for future points. Leveraging temporal dependencies, ARIMA captures patterns and trends to make forecasts.

3.4.2.1.4 Prediction

- Once the ARIMA model is trained, it predicts future data points based on patterns discerned from historical data. It uses lags of dependent variables and previous error terms to make these predictions.

3.4.2.1.5 Evaluation

- The effectiveness of ARIMA predictions can be assessed using standard metrics such as MAPE, MAE, and Root Mean Squared Error (RMSE).

3.4.2.1.6 Assumptions and Features

- No Explicit External Factors: Traditional ARIMA does not consider external regressors.
- Parameter Estimation: The selection of parameters p, d, q is crucial and is often determined using tools such as autocorrelation and partial autocorrelation plots. Properly estimated parameters ensure efficacy of the ARIMA model.
- In forecasting scenarios, especially demand prediction, ARIMA presents a robust technique that capitalizes on patterns in sequential data. Its ability to extrapolate temporal patterns—be they short-term fluctuations or longer-term trends—ensures that forecasts are both accurate and insightful. However, it is imperative to ensure stationarity and appropriate parameter selection to fully exploit ARIMA's potential.

3.4.2.2 Artificial Neural Networks (ANNs)

ANNs have emerged as a robust tool in the domain of time series forecasting. Owing to their ability to model and predict complex non-linear relationships, ANNs can capture intricate patterns and dependencies in time series data, offering enhanced precision in forecasts. Especially in scenarios where traditional linear methods falter, ANNs present a flexible approach to time series analysis as they can handle various temporal structures with ease.

3.4.2.2.1 Basic Concept

- Layered Structure: ANNs are structured with an input layer, one or more hidden layers, and an output layer. Each layer consists of neurons that process input data and transfer information to subsequent layers.
- Weighted Summation: Within each neuron, input data is subject to weighted summation followed by an activation function, which introduces non-linearity thereby enabling the network to learn intricate patterns. An ANN focusing on a single time series feature could be formulated as,

$$y_t = f\ (w_1 . x_{t-1} + w_2 . x_{t-2} + \cdots + w_n . x_{t-n} + b) \tag{2}$$

  where y_t is predicted value at time t, $(x_{t-1}, x_{t-2}, \ldots, x_{t-n})$ are previous n observations in the time series, $(w_1, w_2, \ldots, w_n)$ are weights the model learns during training (corresponding to each previous observation), f is an activation function, and b is the bias term.

3.4.2.2.2 Learning Process

- Backpropagation: ANNs employ the backpropagation algorithm, which involves calculating the gradient of the loss function concerning each weight by using the chain rule. This process ensures minimization of prediction errors.
- Optimization: Techniques such as Gradient Descent or its variants—including Adam and Root Mean Squared Propagation (RMSProp)—help in adjusting weights of the network to reduce error between predicted and actual time series values.

3.4.2.2.3 Training

- The model ingests sequences of historical time series data, adjusting its internal weights to best represent underlying patterns. The architecture might involve recurrent structures such as LSTM or Gated Recurrent Units (GRU), which are particularly adept at capturing long-term dependencies in time series data.

3.4.2.2.4 Prediction

- Once trained, the ANN can produce forecasts for future data points in the time series. Given the input sequence, the network generates output values based on learned patterns, ensuring that predictions consider both recent and historical trends.

3.4.2.2.5 Evaluation

- The efficacy of ANNs in time series forecasting is gauged using metrics such as MAE, MSE, and R-squared. Furthermore, validation techniques like walk-forward validation are often employed to assess the model's performance on unseen data.

3.4.2.2.6 Assumptions and Features

- Non-Linearity: Unlike some traditional models, ANNs do not assume a linear relationship in data. They are intrinsically capable of modeling complex, non-linear dependencies.
- Temporal Dependencies: With architectures like RNNs, LSTMs, or GRUs, ANNs can recognize and learn long-term and short-term dependencies in data, which is important for time series forecasting.
- In the realm of time series forecasting, ANNs represent a cutting-edge approach, particularly suited for complex datasets where other models might not yield satisfactory results. Their capability to learn from both long-term and short-term patterns in data—combined with their adaptability—positions them as a potent tool for a myriad of forecasting challenges. However, they do necessitate careful tuning and validation to ensure optimal performance and to avoid pitfalls like overfitting.

3.4.2.3 Linear Regression (LR)

LR is a cornerstone of predictive analytics and has been adapted effectively for time series forecasting. This statistical approach is tailored to understand and predict future points in a series, particularly when data demonstrates a linear trend over time. By assessing relationships between independent time variables and series' values, LR can derive actionable insights, making it an invaluable tool for forecasting demand in various scenarios.

3.4.2.3.1 Basic Concept

- Trend Identification: In the context of time series, LR focuses on deciphering linear trends within data. This involves establishing a relationship between time—or time-derived variables—and the series' values.
- The following equation can be used to predict the value of the time series at time t:

$$y_t = \beta_0 + \beta_1 t + \epsilon_t \tag{2}$$

where y_t is the value of the time series at time t, t is the time variable, β_0 is intercept of the regression line, β_1 is slope of the regression line (representing the trend over time), and ϵ_t is the error term at time t.

3.4.2.3.2 Learning Process

- Coefficients: The objective of the learning process is to determine optimal values of *m* and *c* that lead to the best fit line, which in turn provides the most accurate forecasts.
- Optimization: The Ordinary Least Squares (OLS) method is commonly employed to minimize the sum of squared residuals, ensuring the derived line fits closely to the observed data points.

3.4.2.3.3 Training

- The training phase involves feeding a set of historical data points into the LR algorithm. The model then learns the linear relationship inherent in this data, adjusting its coefficients to minimize prediction errors.

3.4.2.3.4 Prediction

- Once trained, the LR model can forecast future values by simply extrapolating the established linear trend. It is imperative to understand that, while this method is powerful, its forecasts are based on the assumption that future trends will mirror past behaviors.

3.4.2.3.5 Evaluation

- The accuracy of the LR model's predictions can be assessed using metrics like MAE, MSE, and R-squared.

3.4.2.3.6 Assumptions and Features

- Linearity: One of the fundamental assumptions of this method is that the relationship between the time variable and the series' values remains linear. This implies that changes in the independent variable correspond to proportional changes in the dependent variable.
- Independence: It is essential that the residuals—the differences between observed and predicted values—are independent of one another. Serial correlation in these residuals can lead to inaccurate predictions.
- In the domain of time series forecasting, LR offers a straightforward yet effective approach. Its strength lies in its simplicity, transparency, and ease with which its results can be interpreted. Whether predicting stock prices, sales, or demand for resources, when historical data demonstrates a clear linear trend, LR can provide dependable forecasts. However, it is crucial to acknowledge its limitations and ensure the data's behavior aligns with the model's underlying assumptions.

3.4.3 Machine Learning (ML) Forecasting Methods

The accuracy of traditional methodologies is often lacking, especially when an excess of interconnected variables is involved. Indeed, several advantages of ML techniques render them as superior choices when it comes to demand forecasting.

First, traditional methods are typically constrained to a limited set of demand determinants, while ML techniques are designed to accommodate a more extensive range of factors as independent variables. In other words, one-dimensional algorithms are usually employed by traditional models whereas multi-dimensional data spaces are navigated by ML methods. Second, ML techniques exhibit an unparalleled adaptability, making them applicable across an array of business challenges. A singular ML algorithm can be applied in areas such as sales, finance, and marketing, and traditional methods are simply not equipped to adapt and understand complex interdependencies between variables in the same flexible manner as ML. Third and last, vast data volumes can be analyzed by ML algorithms, leading to enhanced forecast accuracy, which remains the primary aim of this research. A diverse array of data types—be they numerical or categorical—can be integrated by ML thereby refining its forecast precision. Such inherent ability to continually refine forecasting is not found in conventional methodologies.

The following subsections evaluate the efficacy of a number of ML algorithms, offering insights into their suitability for this specific forecasting challenge. Based on performance metrics, the most optimal algorithm is then chosen to forecast consumables quantities during a project lifecycle.

3.4.3.1 Artificial Neural Networks (ANNs)

ANNs in ML are inspired by the structure and functional aspects of biological neural networks. They are composed of a large number of highly interconnected processing elements, known as neurons, working in harmony to solve specific problems. ANNs, like people, learn by example. They are configured for specific applications—such as pattern recognition or data classification—through a learning process. Learning in biological systems involves adjustments to synaptic connections that exist between neurons. Similarly in ANNs, it involves modification of the weights between nodes. During the training phase, ANNs learn from data to approximate a function. They have a remarkable ability to derive meaningful insights from complex and noisy datasets, making them instrumental in fields such as image and speech recognition, natural language processing, and medical diagnosis.

3.4.3.1.1 Basic Structure

- Nodes/Neurons: The fundamental units of ANNs are nodes or neurons, analogous to the neurons in a biological brain.
- Layers: Nodes are organized into layers including an input layer, one or more hidden layers, and an output layer. Each layer can have a variable number of nodes.
- Connections: Nodes from adjacent layers have connections with weights assigned to them. These weights are the heart of learning in ANNs. Figure 5 shows a simple ANN that includes inputs, one hidden layer, and an output.

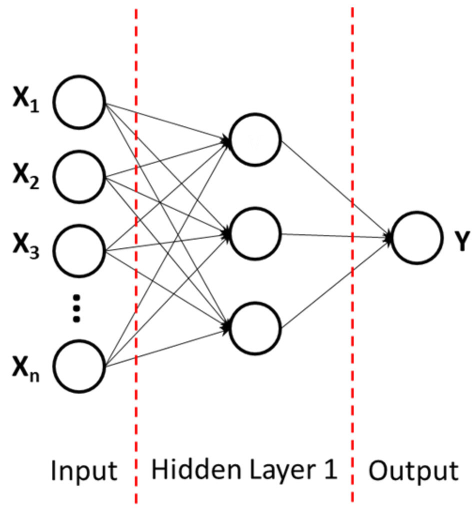


Figure 5 Artificial neural network

- Equation: The basic formula for a neuron's output in a single-layer network is,

$$y = f\left(\sum_{i=1}^{n} w_i x_i + b\right) \quad (3)$$

where y is the predicted value, f is an activation function, w_i are the weights, x_i are the inputs, and b_i is the bias term.

3.4.3.1.2 Learning Process

- Training: ANNs learn during a training phase. The network is fed a dataset, and the weights between nodes are adjusted to minimize the error between the predicted output and actual target values.
- Backpropagation: This is a common algorithm used for training ANNs. It computes the gradient of the loss function concerning the weights, which is then used to update the weights to minimize loss.
- Activation Functions: Activation functions introduce non-linear properties into the system. Common examples include the sigmoid, tanh, and ReLU functions.

3.4.3.1.3 Training

- Initialization: Starts with random weights and learns to reduce errors through training.
- Epochs and Batches: Training usually occurs over many cycles (epochs) and can involve breaking the data into batches for efficiency and effectiveness.

- Optimization Algorithms: Techniques like stochastic gradient descent or Adam are used to change the weights and reduce the loss function.

3.4.3.1.4 Prediction

- Forward Propagation: Uses the learned weights to make predictions or classifications on new data.
- Output Interpretation: The final layer's output is interpreted as the ANN's prediction, which can be a class label, a value, or a set of values.

3.4.3.1.5 Evaluation

- Validation Data: Separate data used to evaluate the performance of the model, ensuring it hasn't just memorized the training data but can generalize well.
- Metrics: Performance is measured using metrics such as accuracy, precision, recall, and F1 score for classification tasks; or mean squared error and mean absolute error for regression tasks.

3.4.3.1.6 Assumptions and Features

- Data Preparation: ANNs assume that input features are numerical and often normalized or standardized.
- Feature Scaling: Essential for networks to learn effectively, as it ensures that all inputs are treated equally.
- Overfitting: ANNs can overfit to training data, learning it too well and failing to generalize to new data. Techniques like dropout, regularization, and validation can prevent this.
- Complexity and Capacity: ANNs can model complex nonlinear relationships, but their capacity must be balanced against the risk of overfitting.

3.4.3.2 Linear Regression (LR)

LR is a foundational algorithm in ML and statistics, primarily used for predictive analysis in a variety of fields. It establishes a linear relationship between a dependent variable and one or more independent variables by fitting a line to the observed data. The aim is to find the line that minimizes the distance between the observed points and the line itself, often quantified using measures such as MSE. Multivariate LR extends this concept further to multiple independent variables. Training an LR model involves finding values of b that minimize the cost function, often achieved through optimization techniques such as Gradient Descent. Once trained, the model can predict the dependent variable for new data points, making it a powerful tool for forecasting. Despite its simplicity, LR is an effective tool, forming the foundation for understanding more complex ML algorithms, and is also easily interpretable, making it a valuable asset in the toolkit of analysts and data scientists.

3.4.3.2.1 Basic Concept

- Equation: The core idea behind LR is to fit a straight line that minimizes the distance (usually the square of the distance) between actual and predicted values. The equation for a simple LR is,

$$y = \beta_0 + \beta_1 X_1 + \epsilon \tag{4}$$

  where y is the predicted value, β_0 is the intercept, and x is the input feature.
- Multivariate LR: When there are multiple input features, the algorithm extends to multivariate LR and the equation becomes,

$$y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_n X_n + \epsilon \tag{5}$$

  where y is the predicted value, β_0 is the intercept, β_1, β_2 are the coefficients of the features, and x is the input feature.

3.4.3.2.2 Learning Process

- Cost Function: The objective of LR is to minimize the cost function, which is the MSE between the predicted and actual values over the training data.
- Gradient Descent: This is a common optimization algorithm used to find values of the coefficients that minimize the cost function.

3.4.3.2.3 Training

- The model learns the coefficients—or parameters—during the training phase using the training data. This learning is achieved by finding the error between predicted values and actual values and adjusting the coefficients to minimize this error.

3.4.3.2.4 Prediction

- Once trained, the model can make predictions on new data by simply adding the new data into the learned linear equation.

3.4.3.2.5 Evaluation

- Models are evaluated on how well they predict new data. Common evaluation metrics for LR include MAE, MSE, and R-squared.

3.4.3.2.6 Assumptions and Features

- LR makes several assumptions including linearity, independence, homoscedasticity, and normality. Violations of these assumptions can lead to poor or biased model estimates.

3.4.3.3 Decision Trees (DTs)

DTs are a type of supervised ML algorithm predominantly used for classification but can also handle regression tasks. They represent a series of decisions derived from training data that lead to predicted outcomes. The decisions are based on certain conditions or rules thereby forming a tree-like model of decisions.

3.4.3.3.1 Basic Concept

- Tree Structure: A DT consists of nodes that form a rooted tree, which means it is a directed tree with a node called the "root" that has no incoming edges. Each internal node denotes a test on an attribute, each branch represents an outcome of a test, and each leaf node (terminal node) holds a class label.
- Splitting: DTs select the best attribute to split the data in a way that is most informative. The process of selecting the best attribute is determined by measures such as Gini impurity, entropy, or the information gain. The decision process can be represented as a series of branching operations where $x_{feature} \leq threshold$, the process splits into further branches; otherwise, it moves to a different branch or results in a prediction.

3.4.3.3.2 Learning Process

- Recursive Splitting: Starting with the root, the data is split based on the best attribute. This process is recursively continued on each branch using subsets of the dataset until one of the termination conditions is met such as reaching maximum tree depth or a node contains data of a single class.

- Pruning: To ensure the tree does not overfit the training data, branches that have little weight in predicting target values are pruned.

3.4.3.3.3 Training

- Training a DT involves feeding it a dataset and allowing the algorithm to learn conditions or rules from the training data, which then shape the decisions made by the tree.

3.4.3.3.4 Prediction

- To make a prediction for a new data point, the algorithm starts at the root of the tree and evaluates conditions at each node moving down the tree. The path followed through the tree ends in a leaf node, and the value or class of that leaf node is the algorithm's prediction.

3.4.3.3.5 Evaluation

- The performance of a DT can be assessed using various metrics. For classification tasks, accuracy, precision, recall, and the F1-score are commonly employed. For regression tasks, metrics such as MAE and MSE might be used.

3.4.3.3.6 Assumptions and Features

- DTs assume that the decision boundaries are parallel to the axis of the attributes being considered. It is also assumed that the training data can provide all rules required to correctly classify or predict test data.
- DTs stand out due to their interpretability. The clear visualization of decisions, conditions, and outcomes mean that non-experts can easily understand the model's decision-making process. However, to handle challenges such as overfitting—which can make the model perform exceptionally well on training data but poorly on test data—regular pruning and parameter tuning must be completed. Ensemble methods like RF can help mitigate these challenges, as well.

3.4.3.4 Random Forest (RF)

RF is an ensemble learning method used for both classification and regression tasks. The RF algorithm combines predictions from several DTs to produce a more accurate and generalizable prediction. By leveraging a multitude of DTs, the model becomes more resilient to overfitting and can capture intricate patterns in data.

3.4.3.4.1 Basic Concept

- Ensemble of Trees: An RF model is a collection of DTs, each trained on random subsets of data. The subsets are created by both bootstrapping the data and selecting random features.
- Aggregated Predictions: The final prediction of the RF model is an average of the predictions from individual trees. The equation for an RF is,

$$y = \frac{1}{T} \sum_{t=1}^{T} f_t(x) \tag{6}$$

where $f_t(x)$ is the prediction of t^{th} tree, and T is total number of trees.

3.4.3.4.2 Learning Process

- Bootstrapping: RF trains each tree on a different bootstrap sample, a random subset of the dataset. This ensures diversity among trees.

3.4.3.4.3 Feature Randomness

- During the splitting process, RF selects splits based on a random subset of features rather than considering all features. This introduces additional randomness and diversity into the model.

3.4.3.4.4 Training

- Each tree in an RF is trained independently. Given the random subsets of data and features, each tree learns slightly different patterns. This process aims to make each tree as deep as possible—without pruning—ensuring it captures intricate patterns in its subset of the data.

3.4.3.4.5 Prediction

- For a given input, each tree in the RF provides its prediction. The RF then aggregates these predictions and takes the average.

3.4.3.4.6 Evaluation

- The performance of RF can be evaluated using metrics such as MAE, MSE, or R-squared.

3.4.3.4.7 Assumptions and Features

- RF, being an ensemble of DTs, does not rely on data distribution assumptions. Its strength arises from its ability to reduce variance by averaging out individual trees' predictions, thus offering more stability and resistance to overfitting. One of RF’s features is its capability to provide feature importance scores, which can be invaluable for understanding the drivers in forecasting models.
- In demand forecasting, RF can prove invaluable due to its ability to handle large datasets with higher dimensionality. It can manage missing values and maintain accuracy even when a large proportion of data are missing. Its ensemble nature enables it to capture non-linear relationships effectively, making it a versatile tool for predicting demand in complex scenarios.

3.4.3.5 K-Nearest Neighbor (KNN)

The KNN algorithm is a versatile method used for both classification and regression problems. It belongs to the family of instance-based, or lazy, learning algorithms where computation is deferred until prediction. For demand forecasting, KNN leverages historical data points to predict future demand based on similarity measures.

3.4.3.5.1 Basic Concept

- KNN Rule: The principle of KNN is to predict the output value of a new data point based on the k nearest data points in the training dataset. The closeness or similarity is typically determined using distance metrics such as Euclidean or Manhattan distance.
- Choosing k: The choice of k is critical. A smaller value of k can capture noise in the data, while a larger k can smooth out predictions, potentially overlooking finer data patterns. The prediction for a new point is given by,

$$y = \frac{1}{k} \sum_{i=1}^{k} y_i \tag{7}$$

where y_i are the outcomes of the k nearest neighbors.

3.4.3.5.2 Learning Process

- Instance-Based Learning: Unlike other algorithms that construct an explicit model, KNN memorizes the training dataset. New predictions are made by searching the entire dataset for the most similar instances and summarizing the output variable for those instances.
- Distance Metrics: Various metrics like Euclidean, Manhattan, Minkowski, or Hamming distance can be used to measure similarity or dissimilarity between data points.

3.4.3.5.3 Training

- Essentially, training in KNN involves storing the dataset in memory. There is no explicit learning phase. Instead, the algorithm uses the entire dataset in the prediction phase to find the nearest neighbors.

3.4.3.5.4 Prediction

- When predicting the output for a new data point, the algorithm searches for the *k* training examples that are closest to the point and returns the output value as the mean (for regression) or the mode (for classification) of the output values of its *k* nearest neighbors.

3.4.3.5.5 Evaluation

- The effectiveness of KNN can be assessed using conventional evaluation metrics. For regression tasks, metrics like MAE, MSE, or R-squared can be used. For classification tasks, metrics such as accuracy, precision, recall, and F1-score are appropriate.

3.4.3.5.6 Assumptions and Features

- KNN operates under the assumption that similar data points—based on the distance metric—will have similar output values. It is important to scale features for KNN since the algorithm is sensitive to variations in numerical values across different features.
- A notable feature of KNN is its sensitivity to the choice of *k* and the distance metric. Proper tuning using techniques like cross-validation is essential for optimal performance.

- In the context of demand forecasting, KNN can be particularly effective when there is a logical and consistent pattern of similarity in the demand data. Its non-parametric nature allows it to flexibly capture complex relationships without making strict assumptions about the data's underlying structure. However, it is crucial to ensure that the chosen features are relevant and scaled appropriately as the algorithm relies on distance metrics for prediction.

3.4.3.6 Support Vector Regression (SVR)

SVR—an extension of the popular SVM algorithm—is primarily used for regression tasks. SVR offers a unique way to predict real-valued outputs, making it an excellent choice for demand forecasting, especially in situations where data might have non-linear relationships.

3.4.3.6.1 Basic Concept

- Margin and Hyperplane: SVR aims to find a hyperplane that best fits the data. The algorithm focuses on maximizing the margin while limiting the deviations of data points from this hyperplane, especially points that are difficult to predict.
- Kernel Trick: One of the standout features of SVR is its ability to use different kernel functions, like polynomial, RBF, or sigmoid, to transform and fit non-linear data. The predictive model is represented as,

$$y = w \,.\, x + b \tag{8}$$

where the aim is to minimize $\|w\|$ subject to $|y_i - (w \,.\, x + b)| \leq \epsilon$ for each data point (x_i, y_i).

3.4.3.6.2 Learning Process

- Cost Function: SVR uses a cost function that not only penalizes errors larger than a certain threshold—termed as ε (epsilon)—but also tries to maintain a flat regression function to ensure simplicity and prevent overfitting.
- Lagrangian Multipliers: The optimization in SVR involves the use of Lagrangian multipliers, which helps in determining support vectors that are most informative for predicting outputs.

3.4.3.6.3 Training

- The training process for SVR is somewhat computationally intensive, especially for larger datasets. It revolves around finding the optimal hyperplane that can predict the continuous output with minimum error, subject to the constraints of maintaining a specified margin.

3.4.3.6.4 Prediction

- Once the SVR model is trained, predictions for new data points are made based on their relationship with support vectors. These support vectors are the data points that sit closest to the hyperplane and are most influential in defining it.

3.4.3.6.5 Evaluation

- The quality of SVR predictions is often gauged using typical regression metrics such as MAE, MSE, and R-squared.

3.4.3.6.6 Assumptions and Features

- Unlike many other regression techniques, SVR does not make strong assumptions about underlying data distribution. Instead, its strength lies in its flexibility to handle non-linearity through kernel functions.
- Proper parameter tuning, especially for the regularization parameter, kernel type, and epsilon value, is crucial for the SVR's performance. These parameters can be optimized using methods like cross-validation.
- In the realm of demand forecasting, SVR shines when capturing complex, non-linear relationships in data. Its capacity to delineate intricate patterns using kernel functions, coupled with its robustness to outliers due to the epsilon-insensitive loss, makes it a potent tool. Nevertheless, careful tuning of hyperparameters and kernel selection are imperative to harness the full power of SVR for forecasting applications.

3.4.4 Performance Evaluation

Performance evaluation in ML and time series is a pivotal step that underscores the effectiveness, accuracy, and generalizability of a model concerning the problem it aims to address. This phase involves employing various metrics and techniques to scrutinize how well the model performs, both quantitatively and qualitatively. Metrics like accuracy, precision, recall, F1 Score, MAE, MSE, RMSE, and Area Under the ROC Curve (AUC-ROC) are utilized based on whether the problem-at-hand is a classification or regression problem. These metrics provide quantitative insights into how well the model is performing. In particular, the confusion matrix is a visual tool that helps in understanding the performance of a classification model by showing true positive, true negative, false positive, and false negative values. Another performance metric or visualization often used is learning curves. These plot performance of the model on both training and validation datasets over time—or number of instances—and are instrumental in identifying challenges like overfitting or underfitting. They provide an intuitive view of how the model's performance evolves with experience.

In addition to testing how well the model functions, performance evaluation is also important for: hyperparameter tuning where different configurations are compared to find the most effective one; facilitating comparison of different models to choose the best one for the problem-at-hand; and evaluating resource efficiency—like computational time and memory usage—in real-world, largescale applications. Moreover, ROC and Precision-Recall Curves are necessary for understanding the trade-off between true positive rate and false positive rate, and between precision and recall, respectively. They provide valuable insights into the model's performance across different thresholds and are particularly useful in imbalanced dataset scenarios.

It is worth noting that performance evaluation is not a one-time task but a continuous process, especially in a production environment where models must adapt to new data over time. Through diligent performance evaluation, the model users can ensure the robustness, reliability, and effectiveness of models, which is critical for successful real-world applications. In this research, MAE, MSE, and MAPE are used for performance evaluation. Their equations are as follows:

$$MAE = \frac{1}{n}\sum_{i=1}^{n} |y_i - \hat{y}_i| \tag{9}$$

where n is number of observations, y_i is the actual value, $\hat{y}_i$ is the forecasted value, and $|y_i - \hat{y}_i|$ is the absolute error of each forecast;

$$MSE = \frac{1}{n}\sum_{i=1}^{n} (y_i - \hat{y}_i)^2 \tag{10}$$

where n is number of observations, y_i is the actual value, $\hat{y}_i$ is the forecasted value, and $(y_i - \hat{y}_i)^2$ is the squared error of each forecast;

$$MAPE = \frac{100\%}{n}\sum_{i=1}^{n} \left|\frac{y_i - \hat{y}_i}{y_i}\right| \tag{11}$$

where n is number of observations, y_i is the actual value, $\hat{y}_i$ is the forecasted value, and $\left|\frac{y_i-\hat{y}_i}{y_i}\right|$ is the absolute percentage error of each forecast.

3.5 Summary

In this chapter, the steps undertaken moved from data collection through model development to performance evaluation, and it was noted that data preparation presented challenges due to the encoding of a vast amount of information from disparate sources within the data. The acquired data exhibited a number of missing values, and methods of outlier detection were employed to remove the outliers present in the dataset. Time series techniques—ARIMA, LR, and ANN—and robust ML algorithms—ANN, LR, DT, RF, KNN, and SVR—were then assessed to ascertain the most suitable algorithm for the objectives of this study. Finally, performance evaluation was discussed to explain various metrics that could be used for model evaluation.

CHAPTER 4: CASE STUDY

4.1 Introduction

This chapter focuses on applying the proposed methodology—utilizing historical data from past projects—to develop a data-driven model for forecasting quantities and anticipated distribution of small equipment, tools, and consumables throughout a project's lifecycle. In alignment with the previously discussed methodology, a detailed development and implementation of a data-driven model utilizing various supervised ML algorithms to predict quantities through the project duration is presented. This model ensures holistic collection of historical data with emphasis on findings from both academic literature and prevalent industry practices. The model development also includes data cleaning and preprocessing, and EDA is conducted to identify patterns and trends. Subsequent to this, ML techniques are employed to pinpoint pivotal features thus laying the foundation for a model capable of forecasting demand of diverse small equipment, tools, and consumables. A comparison between different algorithms is shown to gain a better understanding of which model is most suitable for demand forecasting of small equipment, tools, and consumables. After the model is developed, subject matter experts validate said model by evaluating its accuracy and applicability within construction projects.

4.2 Exploratory Data Analysis (EDA)

EDA often includes visual analytics and aids in deriving trends and identifying insights within data. To gain better insight into trends, several preprocessing steps must be performed accordingly so illustrative diagrams and charts can be developed.

4.2.1 Data Preprocessing

For this case study, a dataset was provided by the local industrial partner and includes the following columns: type of consumables (Line Description), project ID (Business Unit), date order was placed (Order Date), date order was needed for (Request Date), date supplier promised to deliver (Promised Date), unit cost of consumables (Unit Cost), and quantity of ordered consumables (Quantity Ordered). The provided dataset is massive and includes approximately 217 Business Units (projects) that took place between 2017 and 2022. In addition, there were almost 70,000 orders placed and 20,000 types of consumables during the timespan. Due to the size of the dataset, rigorous data cleaning—which also included the addition of several columns to the dataset—was performed to ensure we had a reliable dataset that could be used in discovering trends and highlights between different features. Of course, having a reliable dataset is also important for building the model itself, and to make sure it is accurate and can be used in real-world applications.

This being said, the data cleaning process included several steps. The first step was to remove records that contained 0 quantity ordered. Following, some records in the date columns were corrected as they contained typos, and records with negative values under the costs and quantities columns were removed. Then, new columns and assumptions were identified based on the provided dataset as follows:

- Delay Column: A new column named Delay was introduced to represent the expected variance in the schedule. This was calculated by subtracting the Promised Date from the Request Date.
- Assumption of Actual Delivery Date: After consultation with subject matter experts, it was assumed that the Actual Delivery Date aligned with the Promised Date due to the absence of actual delivery dates in the dataset.
- Order Status Values: The statuses of orders were categorized into three values including Delay, No Delay, or Early.
- Request Day Column: An additional column called Request Day was added to indicate the day consumables should be onsite, converted from the Request Date to days since the start of the project.
- Lead Time Calculation: Lead Time was calculated by subtracting the Order Date from the Promised Date, aiding in the prediction of when orders should be placed for the project or specific tasks.
- Project Duration Estimation: Without an explicit Project Duration attribute, the duration was estimated using the first and last order date of all projects.
- Project Cost Assumption: Due to the absence of a Project Cost attribute, it was assumed to be the total of all order costs for each project, serving as a metric for project size.

The aforementioned columns were then analyzed thoroughly to understand how they might impact our model. From the provided dataset—and for the sake of this research—specific columns or features were chosen to be used while developing the model as is discussed in upcoming sections.

4.2.2 Data Transformation

The above subsection describes the most important insights and trends that were discovered by EDA after completing several operations on the dataset such as data cleaning and data preprocessing. After several rounds of discussions with subject matter experts—and after providing more information to the dataset—it was agreed upon that some modifications must take place, and these modifications should be taken into consideration while developing the model. The following modifications were incorporated into the dataset to provide a more reliable input for the model development phase:

- Each record in the dataset relates to a specific type of consumables. Accordingly, the consumables were aggregated into broad categories—omitting the specific type within each category—and focused into eight main categories as they represent the most crucial types of consumables throughout a project. These categories are: discs, gloves, helmets, kneepads, respirator filters, respirator masks, safety glasses, and welding jackets.
- The dataset includes several orders for each type of consumables, and many of these orders were placed in the same day. Therefore, these records were merged into one single record.
- The start of the project was decided based on total hours spent on the project. The first record that shows total hours spent on the project was considered to be the beginning of the project.
- Project Duration was recalculated as follows: for each project, the records were studied closely to identify the first record and last record with total hours spent on the project, then the difference between the dates of both records was assumed to represent Project Duration.
- Instead of using project codes—which is an identifier—the subject matter experts provided project names. This is crucial for understanding the project nature (i.e. petrochemicals, oil and gas, etc.).
- The Request Date column represents the date these consumables were needed.
- The total work hours throughout the project—Work Hours—were provided. This is an indication of the effort required to complete the project and is essential for estimating monthly manpower.

- Due to lack of data and confidentiality matters, the budget for each order was not provided in the dataset. As such, it was assumed the budget for each order could be estimated by multiplying Quantity Ordered by Unit Cost.

The dataset was reduced from approximately 4,300 records to 1,800 as a result of merging and categorizing records. The 1,800 records relate to 76 projects spanning from 2017 to 2020. The dataset was then studied thoroughly to discover insights and trends.

4.2.3 Data Insights and Visualization

Analysis was performed for annual ordered quantities and costs. The analysis revealed that approximately 91% of orders were delivered on time, 8% were delayed, and 1% were delivered early, as shown in Figure 6. Furthermore, the results revealed an important observation regarding noted delays across the timespan of 2017 to 2022, this being that 75% of recorded delays for each year were under 10 days, as represented in Table 1 and Figure 7.

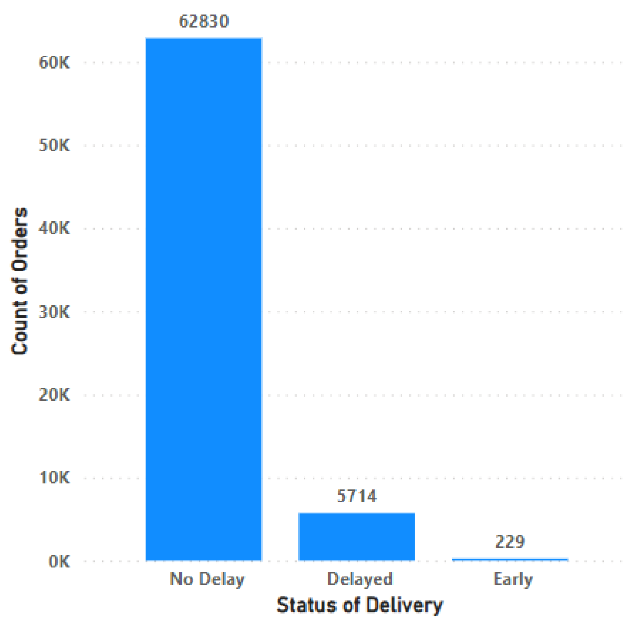


Figure 6 Orders delivery status

Table 1 Delayed orders delivery statistics in days

	count	mean	std	min	25%	50%	75%	max
Year								
2017	1413	2.86	3.96	1	1	1	3	28
2018	1105	2.11	2.34	1	1	1	3	32
2019	962	3.03	4.01	1	1	2	3	27
2020	248	6.29	6.10	1	2	4	6	21
2021	634	5.00	6.86	1	1	2	6	27
2022	1351	7.60	4.63	1	5	8	9	36

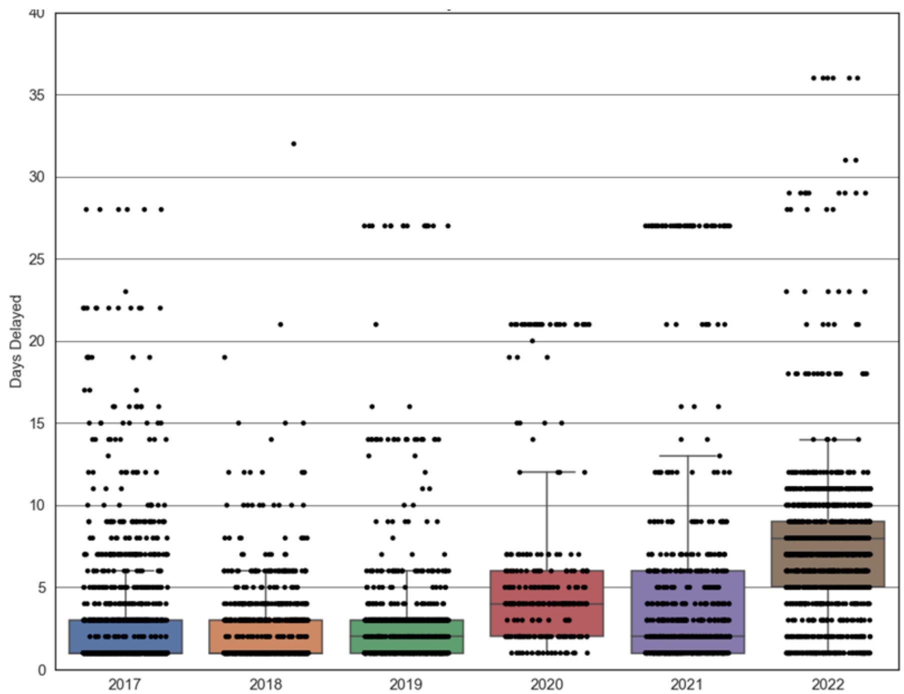


Figure 7 Boxplot of delayed delivery

Figure 8 shows another important insight. It depicts Cost of Orders as a function of Quantity Ordered, as well as Project Duration for the whole dataset. Each blue circle represents a project, and the size of the circle represents Project Duration. It can be seen that the majority of project costs and order quantities are below 1,000,000. In addition, a higher project duration does not necessarily have a higher cost or larger order quantity.

Figure 8 Expenditures based on quantity of materials and project duration

More insights can be drawn regarding total quantity ordered. Figure 9 shows total annual quantity ordered for all types of consumables for project A. This project had a high cost, long project duration, and large ordered quantity compared to other projects, and the figure reveals that most orders were placed during the beginning of the project (in its first year).

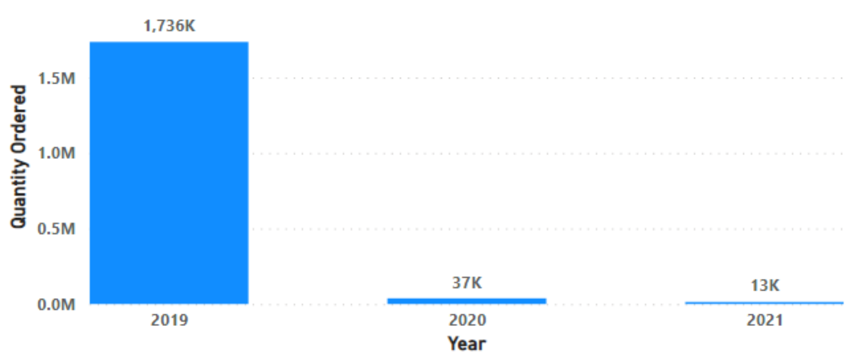


Figure 9 Annual quantity ordered for project A

There were almost 20,000 different types of consumables included in the dataset. The most frequently ordered items among these consumables were gloves. There were many different types of gloves in the dataset. Therefore, the category of gloves was analyzed separately. Figure 10 shows the order history for the glove category in project A. The first row in Figure 10—below the x-axis—indicates that there were no delays observed for this category.

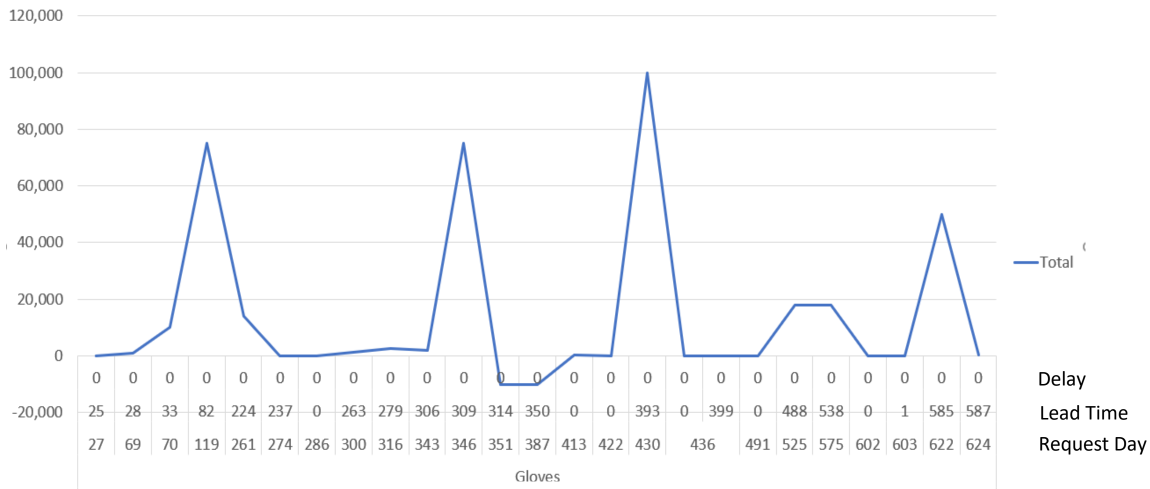


Figure 10 Order history for the glove category in project A

Figure 11 shows the percentage of each consumable category among the 76 projects where discs occupy the highest quantity and kneepads the lowest quantity. This shows that discs are essential for the majority of activities in the project, which is why they occupy the first rank.

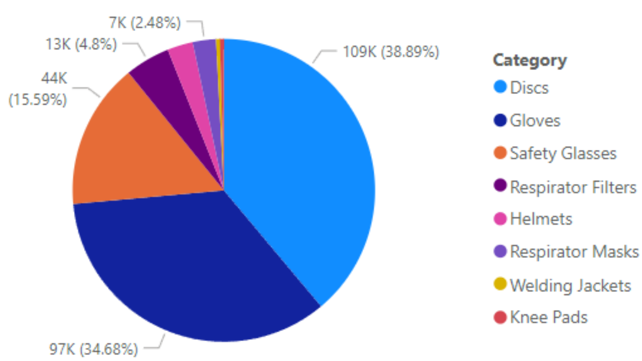


Figure 11 Consumption of each consumable category among projects

Then, the types of projects were studied to determine underlying features of the consumables types that were most used in the different project types. Figure 12 shows a comparison between two project types—oil and gas versus petrochemical—and it can be seen that the ranks of the consumables are different. The Business Unit term in the below figures refers to project.

Figure 12 Comparison of usage of consumable category among different types of projects

Figure 13 shows Project Duration in days (x-axis) and the summation of Quantity Ordered of the disc category in the projects. Each circle represents a project and the size of each circle denotes the number of total hours spent on the project. It can be gathered that longer project duration does not necessarily mean more ordered quantities or more total work hours.

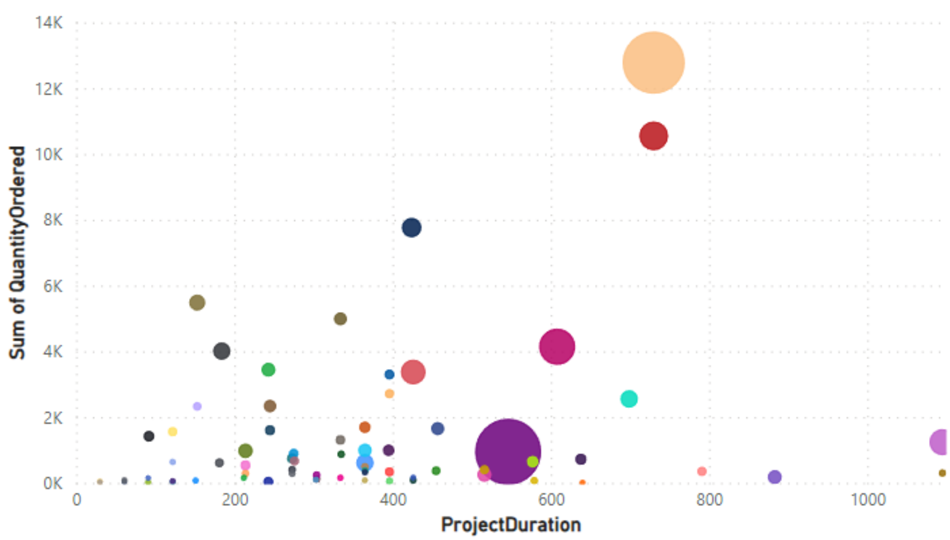


Figure 13 Scatter chart for disc category

After going through the data, Figure 14 shows that there is no trend or discovered pattern between the ordered quantity of each category and the day of order throughout the project. The figure shows fluctuations during the projects using a sample of two projects.

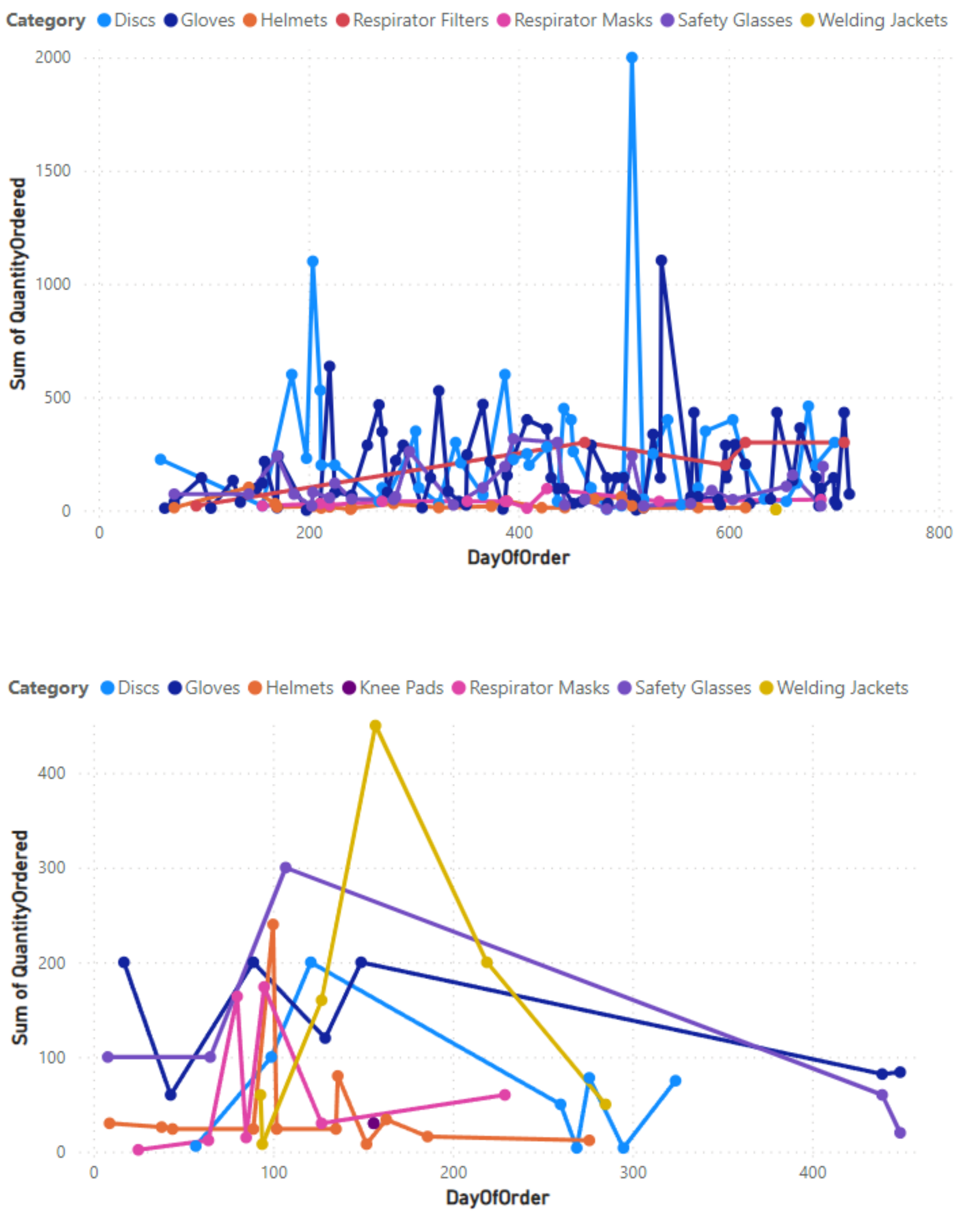


Figure 14 Order trend during two projects

4.3 Time Series Analysis

4.3.1 Autoregressive Integrated Moving Average (ARIMA)

The ARIMA results are outlined in Figure 15 and show an error of 165.99 (i.e. the average forecast error) which indicates the average difference between values predicted by the ARIMA model and actual values in the dataset. Also, a value of +/- 135.44 provides a range around that average error, indicating uncertainty or variability in the error. This range suggests that, for individual predictions, the error can be as low as (165.99 - 135.44) or as high as (165.99 + 135.44). Another measure that must be considered is the MAPE with a value of 201.58. It is a metric that expresses the average error as a percentage of actual values and is commonly used to understand the accuracy of forecasting methods in relation to the magnitude of the numbers being forecasted. A MAPE of 201.58 means that, on average, the forecasted values are off by 201.58% from the actual values. Typically, a lower MAPE value is desired as it indicates higher forecasting accuracy. A MAPE value greater than 100% can indicate substantial discrepancies between forecasted and actual values. Finally, it is worth mentioning that the confidence level is 80%.

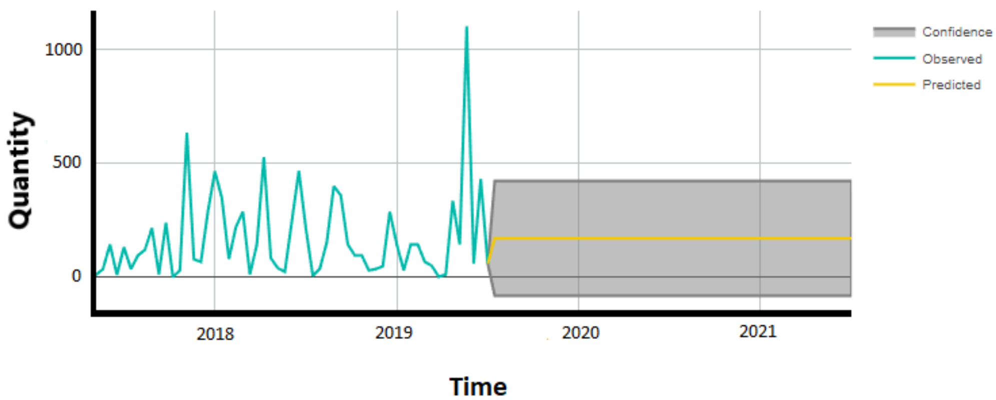


Figure 15 ARIMA time series forecasting for glove category for Project B

4.3.2 Artificial Neural Networks (ANNs)

The ANN results are displayed in Figure 16 and show an error of 165.99 (ie. the average forecast error) showcasing the mean difference between values predicted by the ANN and actual values in the dataset. Additionally, a value of +/- 429.01 indicates an interval around that average error, demonstrating uncertainty or variability in the error. This interval suggests that, for individual predictions, the error can vary and potentially be as low as (165.99 - 429.01) or as high as (165.99

+ 429.01). For the MAPE, a value of 639.37 indicates that, on average, predictions from the ANN deviate by 639.37% from actual values. Similar to the ARIMA model, the confidence level is 80%.

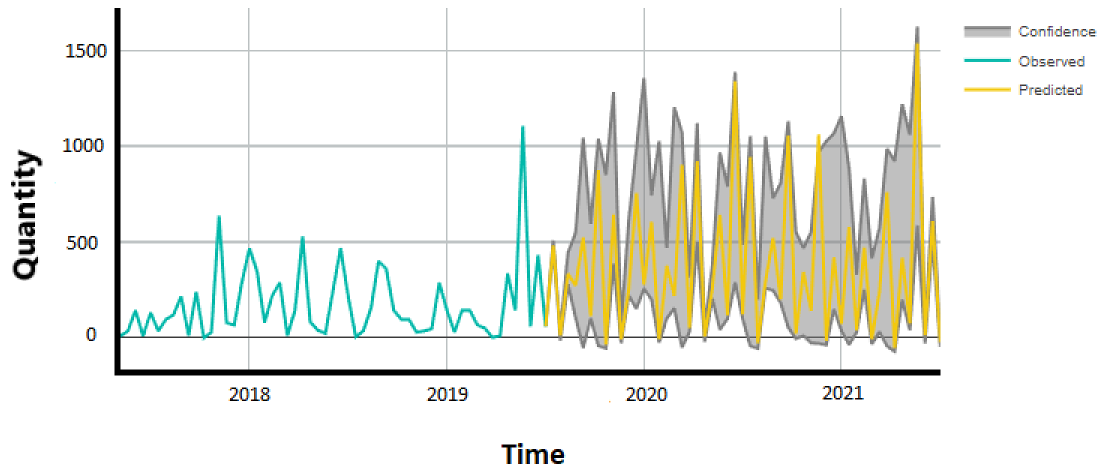


Figure 16 Artificial Neural Network time series forecasting for glove category for Project B

4.3.3 Linear Regression (LR)

LR provides a value that is an indication of the prediction error from the forecast. The LR results displayed in Figure 17 show an error of 165.99 (ie. the central forecast error), which means it is the average error between what the LR anticipated and real values in the dataset. The +/- 396.65 shows the potential variation around this central error, indicating how much the error might fluctuate. This implies that the error for specific forecasts might range from (165.99 - 396.65) to (165.99 + 396.65). For the MAPE, a value of 669.42 indicates that, on average, predictions from the LR deviate by 669.42% from actual values. Similar to the ARIMA and ANN models, the confidence level is 80%.

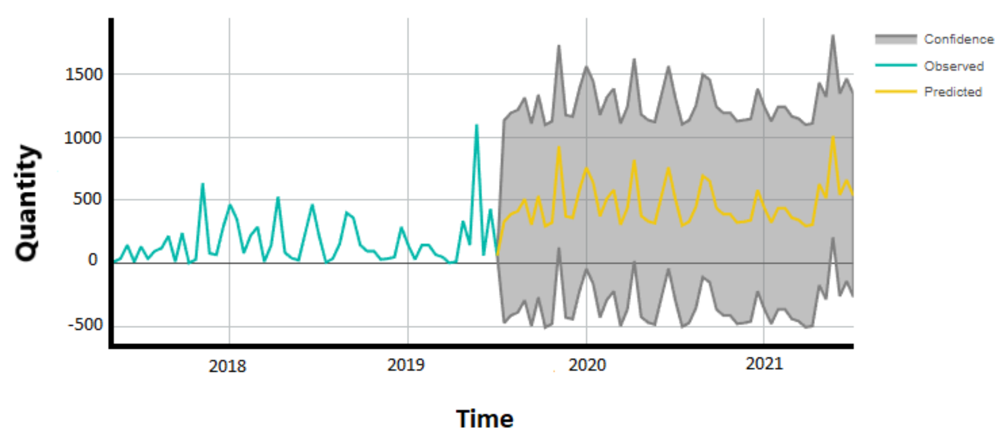


Figure 17 Linear Regression time series forecasting for glove category for Project B

4.4 Machine Learning Forecasting Models

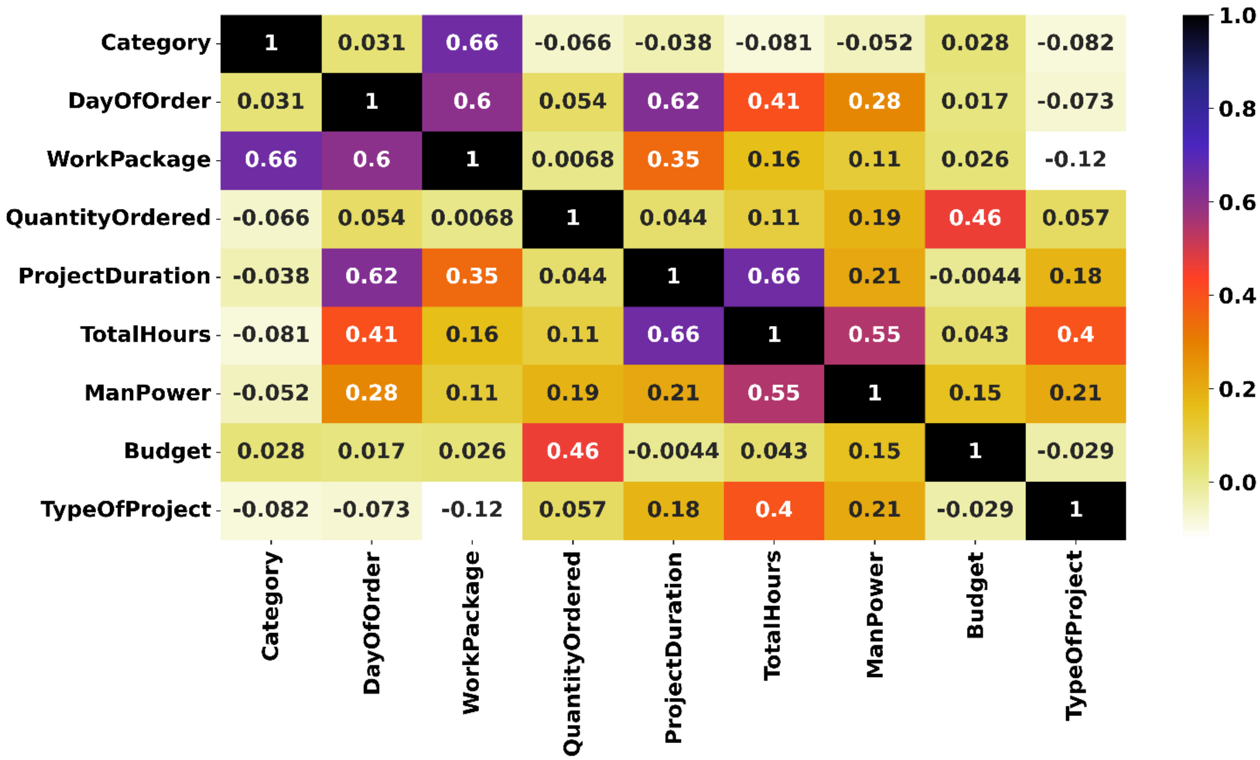
4.4.1 Data Collection and Preparation

The aforementioned modifications applied to the dataset for time series analysis were also used as the base for developing the forecasting models. To complete these modifications, several steps related to data cleaning and feature engineering were performed. First, missing values were removed to ensure we could provide a reliable dataset in the ML algorithms. Then, outliers were detected and removed from the dataset to avoid impacting its accuracy and performance. Finally, feature engineering was implemented to derive new columns or features that are useful for forecasting. After these steps were completed, the dataset includes 1,300 records, 76 projects, and eight consumables' categories.

4.4.2 Correlation and Heatmap

In data-driven, decision-making processes, understanding relationships between different variables is paramount. One of the fundamental statistical tools used to measure the linear relationship between two variables is correlation. The correlation coefficient—often represented by *r*—ranges from -1 to 1. A value of 1 implies a perfect positive correlation meaning that, as one variable increases, so does the other, proportionally. Conversely, a value of -1 indicates a perfect negative correlation suggesting that, as one variable increases, the other decreases. A value of 0 indicates no linear correlation between variables. While correlation analysis can provide a numerical understanding of relationships, visual representations often provide more immediate and intuitive insights. This is where heatmaps—graphical representations of data where individual values are represented as colors—come into play in the realm of demand forecasting. The variance in color intensity gives a quick, visual summary of information, making it easier to understand complex datasets and relationships between various factors.

In our case study, we used a heatmap to represent the correlation matrix of all variables. By visualizing our data this way, we were able to immediately spot which variables strongly related to the target variable, demand, as well as which variables might be redundant or not useful for our predictive model. This step is crucial as it can inform feature selection, ensuring that our forecasting model is both accurate and efficient. Furthermore, incorporating correlation analysis and heatmaps ensured our subsequent modeling efforts were grounded in a solid understanding of underlying relationships in the dataset. By recognizing and acting upon these insights early in the process, we could enhance the accuracy and reliability of our demand forecasting endeavors. The heatmap depicted in Figure 18 shows a high correlation between Day of Order and Project Duration, as well as between Quantity Ordered and Budget of each order. It is also observable that orders are typically placed concurrently with the start of the work package, resulting in minimal variance between the two dates. Consequently, this demonstrates a high correlation.



	Category	DayOfOrder	WorkPackage	QuantityOrdered	ProjectDuration	TotalHours	ManPower	Budget	TypeOfProject
Category	1	0.031	0.66	-0.066	-0.038	-0.081	-0.052	0.028	-0.082
DayOfOrder	0.031	1	0.6	0.054	0.62	0.41	0.28	0.017	-0.073
WorkPackage	0.66	0.6	1	0.0068	0.35	0.16	0.11	0.026	-0.12
QuantityOrdered	-0.066	0.054	0.0068	1	0.044	0.11	0.19	0.46	0.057
ProjectDuration	-0.038	0.62	0.35	0.044	1	0.66	0.21	-0.0044	0.18
TotalHours	-0.081	0.41	0.16	0.11	0.66	1	0.55	0.043	0.4
ManPower	-0.052	0.28	0.11	0.19	0.21	0.55	1	0.15	0.21
Budget	0.028	0.017	0.026	0.46	-0.0044	0.043	0.15	1	-0.029
TypeOfProject	-0.082	-0.073	-0.12	0.057	0.18	0.4	0.21	-0.029	1

Figure 18 Heatmap for small equipment, tools, and consumables

After generating the heatmap and drawing correlations between various features, we selected features as input to the forecasting model. This feature selection process relied mainly on data availability, confidentiality matters, and reasonable assumptions, and resulted in the following seven features being selected:

- Consumable category: This denotes the specific type of consumable required for various tasks such as gloves, helmets, or discs.
- Work package: This indicates the timeline for different tasks or activities in the project, helping determine when specific consumables might be needed.

- Project duration: This represents total timespan for completion of the project. It is crucial for understanding how long a specific consumable might be in demand.
- Total work hours: This refers to cumulative hours anticipated for the project, which can indicate intensity and therefore potential usage of consumables.
- Manpower: This highlights the number of workers or professionals engaged in the project on a monthly basis. A larger workforce might indicate higher consumption rates of certain tools or consumables.
- Budget for each order: This provides a financial perspective, detailing how much is allocated for each order. It can be an indicator of quantity of consumables expected to be used.
- Type of project: This specifies the nature of the construction project such as oil and gas, power, or agricultural chemicals. Different project types have unique consumable requirements and usage patterns.

As for the model output, this research aims to forecast quantities throughout the project.

4.4.3 Model Development and Evaluation

In this study, we implemented some of the most common supervised ML algorithms among construction management practices—ANNs, LR, SVR, KNN, DTs, and RF—for model development. The development process for the various algorithms included data splitting, where the dataset was split into training and testing. As a result, approximately 80% of records were used to train the model, while 20% of records were used for testing the model. Additionally, hyperparameter tuning was utilized in an attempt to achieve peak performance of each algorithm. For the developed models, we also used two metrics for evaluation. Mean Absolute Error (MAE) is a model evaluation metric that measures the average magnitude of errors in predictions, offering a simple and interpretable measure of model accuracy, while Mean Squared Error (MSE) evaluates the average of the squares of the errors, making it more sensitive to outliers and penalizing large errors more severely than MAE. Both metrics are crucial for assessing the performance of regression models, with MAE providing a straightforward average error and MSE highlighting larger discrepancies between predicted and actual values. The following subsections detail the results of each algorithm.

4.4.3.1 Artificial Neural Networks (ANNs)

For the ANN model, the following steps were taken to achieve results: input data were first fed into the network; the input data were then processed in hidden layers using weights that were adjusted during learning; and, finally, an output prediction was produced. Table 2 summarizes hyperparameter values used in the model.

Table 2 Hyperparameters used in the model

Model Parameter	Value	Model Parameter	Value
Input layer size	7	Dropout Rate	0.2
Number of hidden layers	3	Optimizer	Adam
Output layer size	2	Learning Rate	0.001
Activation Functions	relu	Epochs	300
Activation Function for output layer	linear	Batch Size	32
First hidden layer size	256	Dropout Rate	0.2
Second hidden layer size	128	Optimizer	Adam
Third hidden layer size	64		

Using Figure 19, we were able to evaluate the model's training process over 300 epochs. The training loss—shown in red—initiated at high values indicating significant initial prediction errors. However, as the epochs advanced, the training loss consistently decreased demonstrating the model's increasing proficiency in capturing patterns within the training data. The validation loss—depicted in yellow—also declined but eventually plateaued. Furthermore, the training and validation chart of the MAE depicted in Figure 20 shows the MAE of the model throughout its training. By the final epoch, the training MAE decreased substantially, but the validation MAE revealed a slight discrepancy. The metrics used to measure performance of the ANN were the MSE and MAE. For our model, the MSE for DayOfOrder was approximately 223 days, and for QuantityOrdered, it was approximately 135 units. The MAE was approximately 12 days for

DayOfOrder and 8 units for QuantityOrdered. These metrics emphasize the model's modest performance.

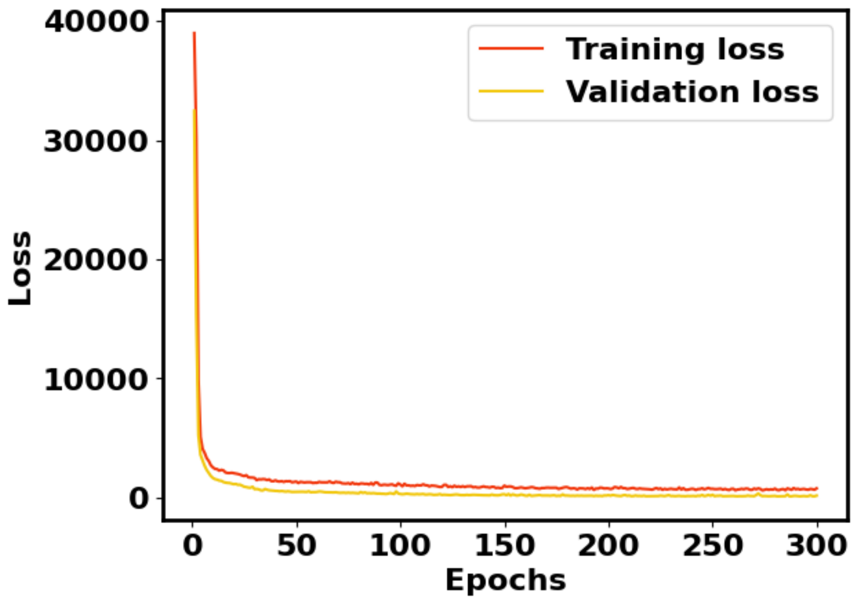


Figure 19 Training and validation loss

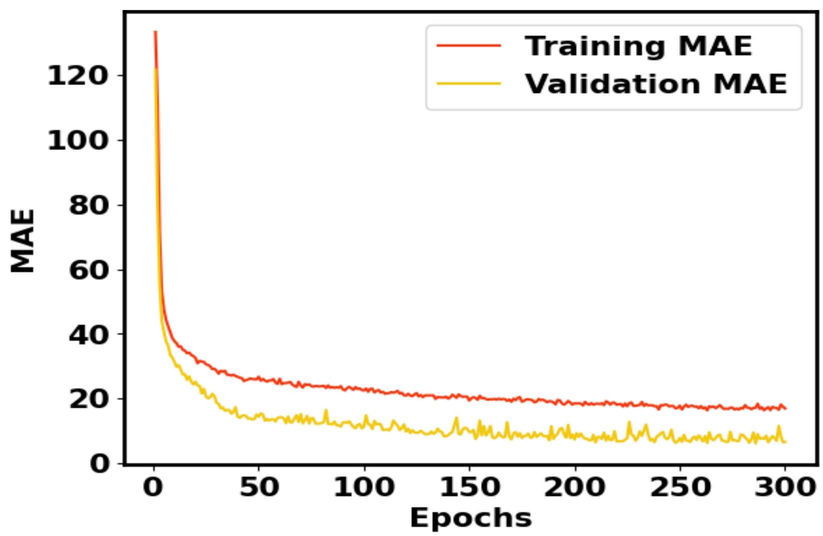


Figure 20 Training and validation mean absolute error (MAE)

Figure 21 illustrates the predictive accuracy of the ANN model by plotting actual values against model predictions for the two different variables: DayOfOrder and the QuantityOrdered. In both graphs, the x-axis shows the actual values, the y-axis represents the predicted values, and the dashed red line indicates the line of perfect prediction. Generally, the close alignment of data points to this line represent a high degree of accuracy. These visualizations serve as a quantitative assessment of the model's performance in the domain of order prediction.

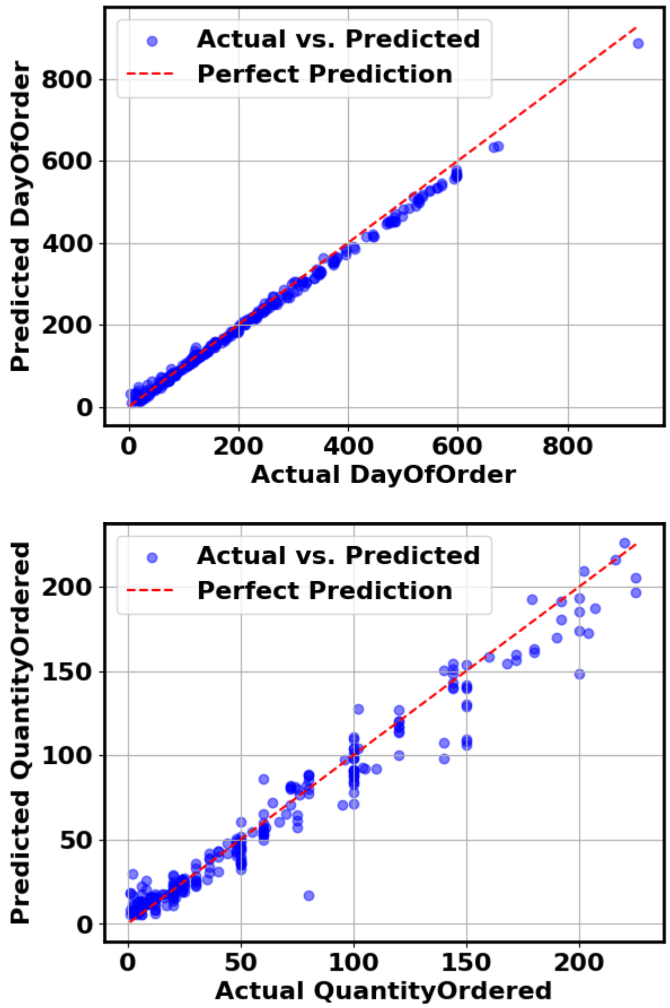


Figure 21 ANN demand forecasting

4.4.3.2 Linear Regression (LR)

In this study, the model—once trained on the scaled data—was able to predict outcomes on the test set. The metrics used to measure performance of the LR were the MSE and MAE. For our model, the MSE for DayOfOrder was approximately 24 days, and for QuantityOrdered, it was approximately 2,514 units. For DayOfOrder, the MAE stood at about 3 days, while for QuantityOrdered, it was near 39 units. This suggests that, on average, the predictions of our LR model deviate from actual values by these respective amounts. Figure 22 shows the predictive accuracy of the LR model for DayOfOrder and QuantityOrdered.

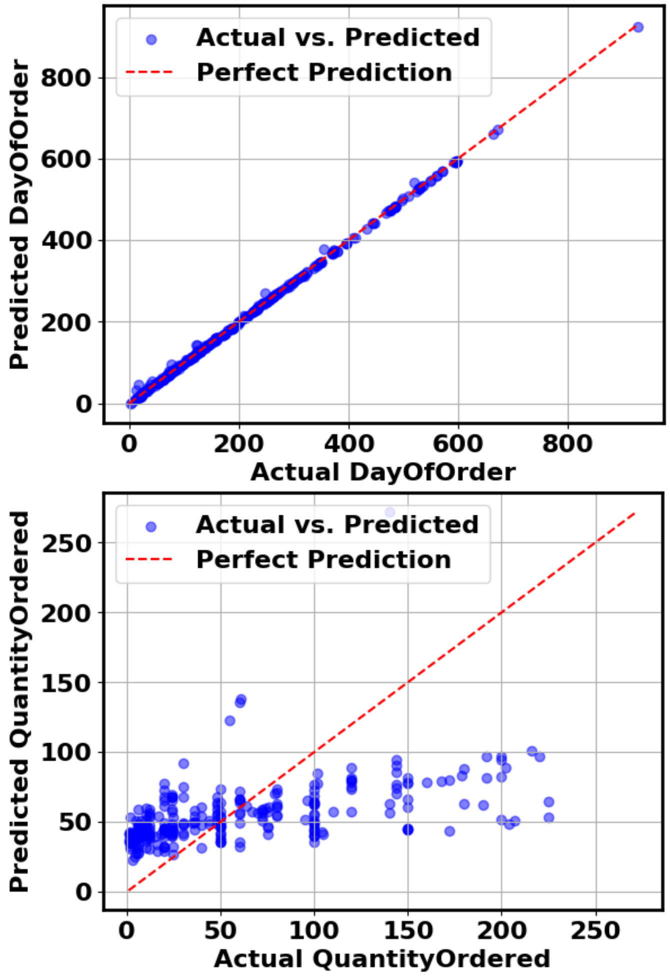


Figure 22 LR demand forecasting

4.4.3.3 Support Vector Regression (SVR)

After being trained, the SVR was employed to predict outcomes on the test set. The metrics used to evaluate the SVR's efficiency were the MSE and MAE. For our model, the MSE for DayOfOrder was approximately 17,392 days, and for QuantityOrdered, it was 3,125 units. For DayOfOrder, the MAE was approximately 91 days, and for QuantityOrdered, it reached near 40 units. This indicates that, on average, the predictions of our SVR deviate from actual values by these respective amounts. Figure 23 depicts the predictive accuracy of the SVR model for DayOfOrder and QuantityOrdered.

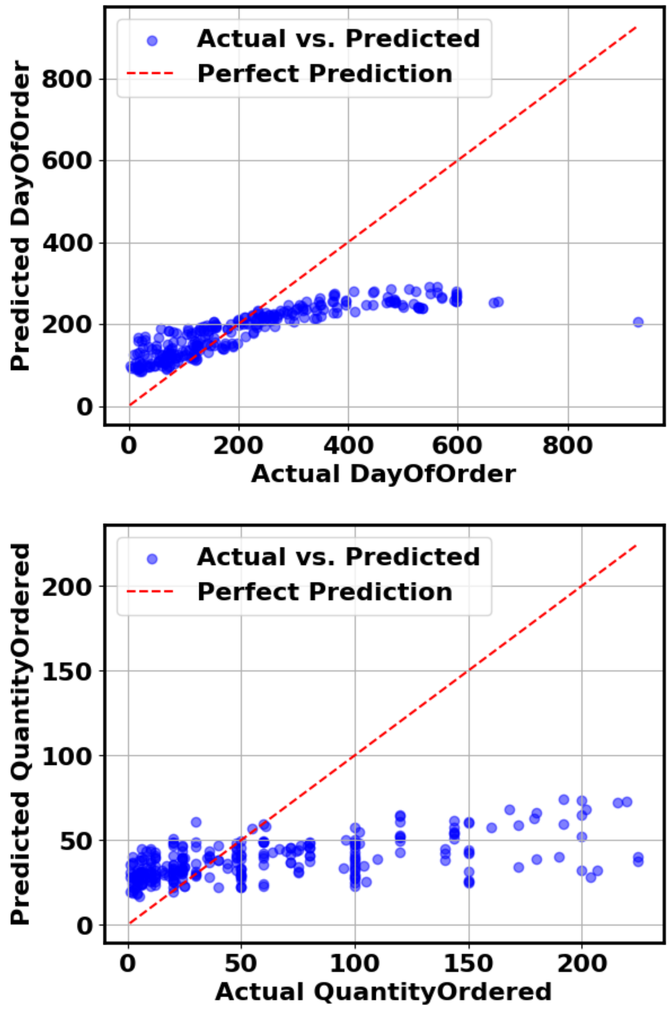


Figure 23 SVR demand forecasting

4.4.3.4 K-Nearest Neighbors (KNN) for Regression

Once the KNN algorithm was trained, it used the training data to make predictions on the test set by assessing the closeness of data points within the feature space. In our model, the MSE values were approximately 1,232 days for DayOfOrder, and 1,438 units for QuantityOrdered. These figures highlight the average squared differences between the KNN model's predictions and actual data. The MAE values were approximately 25 days for DayOfOrder, and nearly 26 units for QuantityOrdered. These values suggest that our KNN model's predictions, on average, differ from real values by these amounts. Figure 24 reveals the predictive accuracy of the KNN model for DayOfOrder and QuantityOrdered.

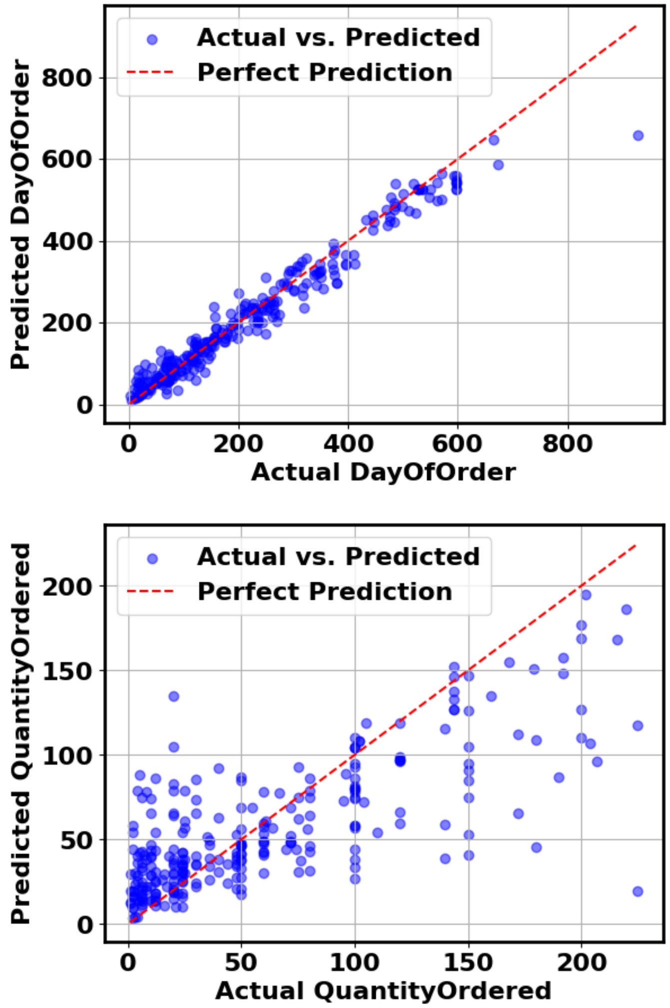


Figure 24 KNN demand forecasting

4.4.3.5 Decision Trees (DTs)

After the training phase, the DT algorithm used its hierarchical structure to navigate through decisions based on feature values, leading to predictions on the test set. For our model, the MSE values were approximately 250 days for DayOfOrder, and near 921 units for QuantityOrdered. The MAE values were close to 9 days for DayOfOrder, and nearly 18 units for QuantityOrdered. These figures suggest that the predictions from our DT model, on average, have these differences from actual observed values. Figure 25 presents the predictive accuracy of the DT model for DayOfOrder and QuantityOrdered.

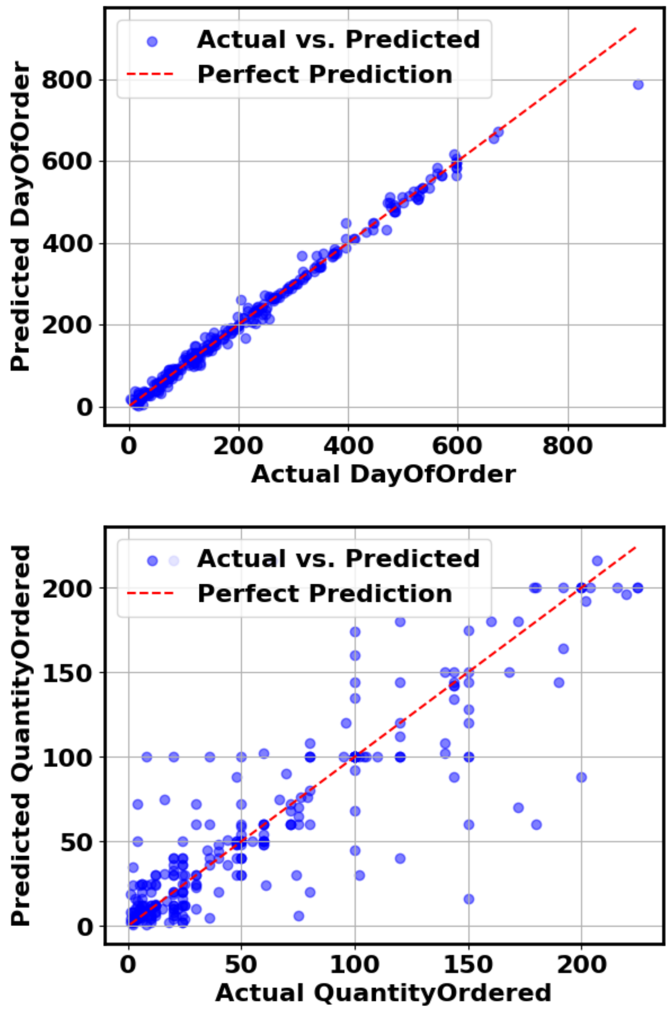


Figure 25 DT demand forecasting

4.4.3.6 Random Forest (RF)

Upon training the RF algorithm, it used the ensemble of DTs to make predictions on the test set. Each tree in the forest produced its own prediction. The final output was an average of these predictions, providing a consensus estimate. In our model, the MSE values recorded were approximately 177 days for DayOfOrder, and 653 units for QuantityOrdered. Our RF model yielded MAE values of roughly 6 days for DayOfOrder, and close to 16 units for QuantityOrdered. This indicates that the predictions made by our RF model, on an average, deviate from actual values by these margins. Figure 26 illustrates the predictive accuracy of the RF model for DayOfOrder and QuantityOrdered.

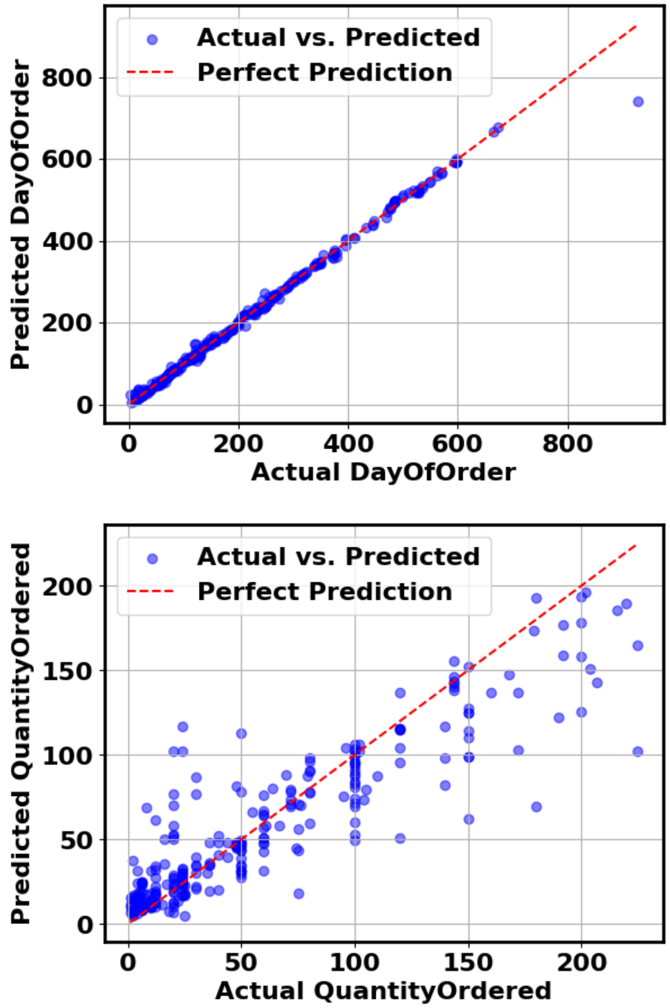


Figure 26 RF demand forecasting

4.5 Discussion

In EDA, the analysis of annual ordered quantities and costs from 2017 to 2022 indicates that, while 91% of orders were on time, a notable 75% of delays were less than 10 days. A significant insight from the data shows that most project costs and quantities were below 1,000,000 with no clear correlation between project duration, order size, or cost. In particular, Project A highlighted that substantial orders typically occurred in the initial year, reflecting early project needs. Among the 20,000 consumable types, gloves emerged as the most ordered item, prompting a detailed analysis which confirmed no delays for this category. Additionally, analysis revealed that different projects have varying essential consumables. For example, discs were dominant in the majority, indicating their critical role in project activities. Comparative studies between project types—such as oil and gas versus petrochemical—showed differing consumable priorities, highlighting the tailored nature of resource needs in diverse environments. Furthermore, data showed no consistent pattern between quantity of orders and timing within the project lifecycle, suggesting there are fluctuating requirements based on immediate project demands rather than a predictable ordering schedule. This lack of trend highlights the dynamic nature of the project and the necessity for adaptive planning and forecasting in construction projects.

After assessing the diverse range of time series and ML models, it was discovered that each offers a unique pattern of strengths and weaknesses, most notably highlighted by their MAPE for time series, and MSE and MAE metrics for ML models. From our results, several insights were revealed about the algorithms used together with their capabilities of forecasting demand of small equipment, tools, and consumables.

For time series models, the analysis confirmed that all three models—ARIMA, ANN, and LR—demonstrate a consistent average forecast error of 165.99. However, the reliability of these models varied significantly as highlighted by their error ranges and MAPE values. The ARIMA model, with its variability range of +/- 135.44 and a MAPE of 201.58%, had the lowest inaccuracies among the three. The ANN model presented a higher uncertainty interval of +/- 429.01 and a dramatically higher MAPE of 639.37%, indicating a substantial increase in prediction variability and error magnitude. Finally, the LR model showed a similar pattern with its error variability range of +/- 396.65 and the highest MAPE of 669.42%, reflecting the most significant deviation from actual values among the models. These findings underscore the challenges each model faces in forecasting, with the ANN and LR models displaying particularly severe inaccuracies and variability compared to the relatively stable ARIMA model. The observed variability and inaccuracies in the results stem from fluctuations in order days and quantities, which present limited consistent trends or patterns from which the models can learn.

For ML models—specifically in relation to DayOfOrder predictions—ANNs presented a respectable performance with an MSE of 223 days and an MAE of 12 days. These results stress the potential of ANNs in modeling intricate non-linear relationships. In contrast, LR emerged as the standout performer, achieving an impressive MSE of 24 days coupled with an MAE of a mere 3 days. Such results imply a robust ability of the LR model to decipher and represent inherent linear tendencies. However, not all models shared this level of precision. The SVR struggled considerably, registering a daunting MSE of 17,392 days and a hefty MAE of 91 days. These figures cast doubts on SVR's capability in modeling this dataset optimally. On a similar note, the KNN results—with its MSE of 1,232 days and MAE of 25 days—underline the pitfalls of instance-based learning, particularly in data where localized nuances might overshadow broader patterns. Finally, DT and RF brought a semblance of balance to the table. The DT, with its MSE of 250 days and MAE of 9 days, proved to be a balanced approach, delicately avoiding overfitting. The RF further confirmed its reputation with an MSE of 177 days and an even more impressive MAE of 6 days, revealing the strength of ensemble learning.

Regarding QuantityOrdered predictions, ANNs—with an MSE of 135 units and an MAE of 8 units—managed to navigate the complexities of the data efficiently. LR, on the other hand and despite earlier triumphs, faced hurdles as evidenced by its MSE of 2,514 units and MAE of 39 units. This stark variation accentuates the significance of understanding the nature of data when applying models. SVR continued to struggle with an MSE of 3,125 units and an MAE of 40 units. Similarly, the KNN results, with its MSE of 1,438 units and MAE of 26 units, emphasizes the need for careful calibration for optimization. In this matter, DT and RF emerged as consistent performers. The DT figures presented a balance, with its MSE of 921 units and MAE of 18 units, while RF, with its MSE of 653 units and MAE of 16 units, solidified its position as a front-runner in the ensemble learning domain.

In conclusion, this study found that—while every model offers benefits—ANN and LR distinctly shine across varied forecasting domains by demonstrating a superior performance in forecasting the quantity and the day of order, respectively. It is worth noting, however, that these insights should guide—not dictate—model choice, as the specific nature of data and overarching goals remain paramount.

CHAPTER 5: CONCLUSIONS, LIMITATIONS, AND FUTURE DIRECTIONS

5.1 Research Summary

Chapter 1 outlined the significance of the construction industry and concluded that material costs—including small equipment, tools, and consumables—account for a significant portion of total project expenses thus underscoring the necessity for precise demand estimation. Estimations currently rely on experience of project professionals using conventional software and judgment, leading to potential inaccuracies. The aim of the study and research objectives were then discussed, which included evaluating current industry practices, identifying key factors influencing demand, and developing a forecasting model utilizing time series models and ML algorithms to predict requirements of small equipment, tools, and consumables for various project types. Following, the expected contributions were detailed such as academic advancements in demand forecasting models, as well as industrial improvements in project-level consumable estimation using historical data. Finally, research questions were posed. These questions involved demand patterns, forecasting accuracy, appropriate time series models and ML algorithms, primary demand indicators, and effective evaluation metrics for model forecasting.

Chapter 2 explored the domain of CSCM, and how demand forecasting is challenged by the unpredictable nature of construction projects due to factors such as varied material requirements, fluctuating site conditions, and weather changes. Literature regarding time series forecasting models—ARIMA, LR, and ANN—was then reviewed and their suitability for different forecasting timeframes and demands assessed. It was concluded that: ARIMA is adapted for longer-term predictions due to its accommodation of non-stationary data; LR is useful for straightforward trend analysis but is limited by non-linear complexities; and ANNs excel in managing intricate and non-linear data, proving effective in various forecasting aspects of construction projects. Furthermore, a literature review was conducted upon ML models. The literature showed several efforts in the area of ML models—such as ANNs, DTs, RF, KNN, and SVR—each with specific applications and benefits in forecasting construction supply chain demands.

Chapter 3 outlined the methodology adopted to predict demand for small equipment, tools, and consumables in construction projects, beginning with an introduction to current industry methods for forecasting demand and identifying key factors influencing consumable requirements. This foundational understanding guided the development of our model aimed at enhancing precision of demand estimates for construction project consumables. The research progressed through a detailed data collection process, where historical data was gathered and analyzed to pinpoint significant demand forecasting factors. This phase combined expert judgment with various forecasting techniques to ensure a comprehensive dataset including essential attributes such as consumable category, work package schedule, project duration, and budget. Subsequent stages involved data preprocessing and transformation to refine the dataset, addressing issues like missing data and outliers, as well as standardizing inputs for model development. This ensured the dataset was ready for analysis, maintaining crucial attributes for an accurate forecasting model. It also emphasized the need for data normalization and splitting for unbiased model training and evaluation.

Chapter 3 also covered time series forecasting, focusing on unique characteristics of time-structured data. Approaches such as ARIMA were discussed for their capability to utilize temporal patterns in stationary data. Furthermore, LR was recognized for its application in linear trends, and ANN for capturing non-linear dependencies in time series data. The chapter explored various ML algorithms, as well, assessing their suitability for demand forecasting. ANNs were praised for their proficiency in complex pattern recognition and adapting biological learning processes, while LR was described as simple yet effective for identifying linear relationships. DTs were noted for their tree-like decision-making process, and RF was found to enhance DTs by combining multiple trees to improve prediction accuracy, handling large and complex datasets efficiently. KNN was valued for instance-based predictions, particularly where historical data trends were strong indicators, while SVR was found suitable for modeling non-linear relationships and its robustness against outliers recognized. Then, performance evaluation was discussed as a crucial aspect for model accuracy and generalizability, employing various metrics and methods like confusion matrices and error metrics.

Finally, chapter 4 discussed a detailed case study where we applied an EDA to understand dataset structures and relationships crucial for developing effective demand forecasting models in construction projects. After the EDA was applied, we performed data preprocessing and transformation to prepare the model input. We then applied three time series techniques—ARIMA, ANNs, and LR—to address the dynamic and complex nature of construction demand forecasting.

Following, we developed various ML algorithms to predict the demand of these consumables. Once the models were developed, we moved to the next stage, model evaluation, where the values of MAPE, MSE, and MAE were calculated. Lastly, a brief discussion about the models was presented. It is worth mentioning that ANN outperformed other algorithms in predicting the quantity. On the other hand, LR showed a superior performance when compared to the other models in predicting the day of order.

5.2 Research Conclusions

As the construction industry continues to face project delivery complexities, efficient management of its supply chain—especially material procurement—has emerged as a pivotal factor. Effective supply chain management does not only ensure timely project delivery but also maintains project quality and budgetary constraints. To this end, the role of predictive modeling in forecasting material requirements is crucial, and leveraging the potential of time series analysis and ML algorithms to streamline this process offers promising solutions.

The primary aim of this research was to understand and harness the capabilities of time series and ML to predict demand of small equipment, tools, and consumables in construction projects. The challenge was in addressing multiple factors affecting the supply chain, ranging from material supply delays to various procurement strategies designed to mitigate delays. We first investigated the development of time series models to forecast the trend of the quantity of small equipment, tools, and consumables versus the DayOfOrder. The results revealed numerous oscillations throughout Project Duration, which resulted from hidden features/factors (might be the result of lack of features in the provided dataset) that impacted DayOfOrder and QuantityOrdered. After this, we proceeded to the development and evaluation of data-driven forecasting models. Several ML models were developed to measure their performance. The results in Table 3 indicate that LR and ANNs have a higher performance in forecasting DayOfOrder and QuantityOrdered, respectively, when compared to other algorithms.

Table 3 Summary of ML models

Algorithms	Forecasting DayOfOrder (days)		Forecasting QuantityOrdered (units)	
	MSE	MAE	MSE	MAE
ANN	223	12	135	8
LR	24	3	2514	39
SVR	17392	91	3125	40
KNN	1232	25	1438	26
DT	250	9	921	18
RF	177	6	653	16

Significantly, the development of these predictive models offers construction professionals advanced tools to anticipate the need for consumables, enabling more accurate and timely procurement decisions. This capability is critical for minimizing project disruptions and maintaining schedules, directly impacting project efficiency and cost-effectiveness. By accurately predicting consumables requirements, construction managers can improve inventory management, reducing the risk of excess stock or shortages. This optimized approach to inventory management supports better budgetary control and project quality. The findings suggest that leveraging LR and ANNs in demand forecasting allows for more adaptive project planning. Professionals can adjust procurement strategies based on predictive insights, enhancing their ability to respond to unforeseen challenges.

While the models developed through this research show promising results, the dynamic nature of the construction industry necessitates ongoing adaptation and refinement of these tools. Continuous research and data updates are essential to maintain the relevance and effectiveness of forecasting models. Construction professionals and researchers are encouraged to collaborate in refining these models, integrating new data and feedback to ensure they accurately reflect current industry trends and challenges.

In conclusion, this thesis not only highlights the potential of time series analysis and ML in revolutionizing material procurement in construction but also calls for a collaborative effort towards continuous improvement. By embracing these advanced predictive tools, the construction industry can enhance supply chain efficiency, project delivery, and overall project success.

5.3 Academic Contributions

This research marks a significant advancement in the domain of construction management and demand forecasting, enriching the academic discourse with several key contributions:

- A cornerstone of this study is its exploration of current demand forecasting practices for small equipment, tools, and consumables within the construction industry. Through a systematic review, we provide a detailed overview of existing methodologies, offering a foundational understanding of industry practices and identifying areas ripe for innovation. This not only enriches academic knowledge but also sets a practical benchmark for future advancements.
- We delve into the complex dynamics influencing consumables demand including project-specific parameters, stakeholder behaviors, and market conditions. This nuanced analysis contributes a comprehensive framework for understanding demand drivers, significantly enriching the conversation around construction project management and consumable resource allocation.
- The development of a machine learning-based model for forecasting consumable requirements represents a leap forward in applying data-driven approaches to construction management. This model demonstrates the potential of ML algorithms to enhance decision-making processes, offering a novel tool for academic and practical application.
- A distinctive feature of this research is the rigorous evaluation of various ML algorithms, culminating in a detailed comparative analysis. This investigation not only identifies the most effective forecasting algorithms but also provides critical insights into their applicability in construction demand forecasting, serving as a valuable resource for both researchers and industry practitioners.

In sum, these contributions significantly advance the existing body of knowledge, providing actionable insights and innovative tools for the construction industry. By bridging the gap between theoretical research and practical application, this work fosters a more sophisticated understanding of demand forecasting and its critical role in construction project management.

5.4 Industrial Contributions

The practical contributions of this research pave the way for further advancements within the construction industry. The following are the primary areas where the research has made noteworthy industrial contributions:

- A pivotal industrial contribution is the development of a data-driven forecasting framework for small equipment, tools, and consumables at the project level. This framework distinguishes itself from traditional estimation models by utilizing a company's historical project data to generate context-specific predictions. This approach ensures forecasts are directly actionable, offering a significant leap over generalized estimates that often lead to inefficiencies.
- The research introduces a dynamic model for consumable requirements forecasting, a substantial innovation over the static models traditionally employed. By providing detailed, timeline-specific forecasts, this model enables project managers to plan resource allocation with unprecedented precision. This methodical approach aids in reducing wastage, optimizing storage costs, and mitigating resource shortage-related downtimes thereby ensuring projects proceed as scheduled without unnecessary delays.
- Equipped with precise forecasts, construction firms can enhance their procurement strategies, achieving better negotiation outcomes with suppliers and ensuring the timely availability of necessary resources. This proactive management of consumables can significantly reduce the likelihood of project delays, translating into considerable cost savings and operational efficiencies.
- Beyond project-level forecasting, the patterns and insights gleaned from the framework offer valuable strategic inputs for executive decision-making. Trends in consumable usage can inform areas for efficiency improvement, highlight investment opportunities in critical equipment, and guide the formation of strategic partnerships with suppliers.

Indeed, these contributions represent a transformative shift in how consumables are managed in construction projects, promising to drive operational optimizations, cost reductions, and improvements in project execution efficiency. By addressing the comparative limitations of existing models, providing a pathway for implementation, and suggesting solutions for potential adoption challenges, this research paves the way for widespread industry advancement.

5.5 Research Limitations

This research—while intending to construct a demand forecasting model for determining requirements of small equipment, tools, and consumables in construction projects—faced a series of limitations as highlighted below:

- Engaging industrial partners for data acquisition proved difficult, leading to a reliance on a dataset with limited attributes from several construction projects. These limited attributes resulted in coming up with assumptions so that we could include various features as model input. We narrowed down the consumables from 20,000 types to eight main categories.
- The provided dataset was limited in scope and biased towards projects with similar characteristics, and also confidentiality matters governed the dataset availability. To mitigate these limitations, modifications were made, such as the calculation of manpower using work hours. Also, the introduction of a Delay column and assumptions regarding the Actual Delivery Date and project start dates, so we could get an insight of the status of the orders.
- The envisioned comprehensive analysis was hindered by the dataset's limitations, restricting the number of input variables. This limitation was addressed by inferring some attributes from the available features, such as project duration and budget. These two features were estimated based on available records/orders of the first and last monthly work hours and total order costs.
- To address data gaps, the Lead Time was calculated by subtracting the Order Date from the Promised Date, aiding in better project planning. These steps were crucial for progressing with model development due to the limited number of project features that were available in the dataset.
- The approach to filling data gaps aimed to enhance the dataset's utility. The aforementioned assumptions together with other calculations discussed throughout the thesis were important to proceed with the model development. These assumptions were included with cautious to strike a balance between data enhancement techniques and the preservation of data integrity.

The limitations identified through this research underscore the critical role of data quality and availability in the development of reliable demand forecasting models. Despite these challenges, the study has made meaningful contributions within the confines of the available data, highlighting areas for future exploration and improvement in the field of construction project management.

5.6 Future Directions

The exploration and advancements made in this study pave the way for exciting future research opportunities in demand forecasting for construction as follows:

- Future work could explore the creation of a comprehensive framework that merges logistics, procurement, and AI-driven prediction models. This framework would aim to optimize project facets such as cost, time, and resource allocation concurrently. A potential first step could involve pilot studies with industry partners to test the framework's effectiveness in real-world scenarios.
- With algorithms like Recurrent Neural Networks (RNNs) showing promise, comparative studies could be designed to evaluate the performance of various ML algorithms across different forecasting scenarios. These studies might focus on scenarios with high variability in data or projects with limited historical data to assess the algorithms' predictive capabilities and adaptability.
- Enhancing model accuracy by integrating additional attributes, such as worker skill levels, external economic indicators, or local labor market dynamics, represents a critical avenue for research. Collaborations with industry and academic partners could facilitate access to this broader range of data, enabling the construction of more nuanced and accurate forecasting models.
- By collaborating with industry partners to deploy IoT devices and sensors at construction sites, researchers could gather real-time, high-quality data. Future research could explore the specific types of data that are most valuable for demand forecasting such as real-time usage rates of consumables or environmental conditions affecting project progress.
- The development of user-friendly forecasting tools should prioritize end-user feedback, incorporating modern UI/UX design principles. Future studies might include usability

testing with project managers and site supervisors to identify key features that enhance the intuitiveness and effectiveness of these tools.

- As the construction industry evolves, forecasting models must also adapt. Research could focus on designing models with embedded feedback mechanisms such as ML algorithms that update their parameters based on new data or outcomes. This adaptability ensures models remain relevant and efficient over time.
- Developing platforms that enable real-time collaboration and integrate seamlessly with industry-standard tools could significantly streamline decision-making processes. Future initiatives could identify the essential features of such platforms, as well as the integration challenges and opportunities with existing project management software.

The aforementioned directions may represent a great opportunity to advance the demand forecasting for small tools, equipment, and consumables in construction. Emphasizing algorithm exploration and user interface development will be particularly important for creating tools that meet the needs of industry professionals.

REFERENCES

Adeli, H., & Yeh, C. (1989). Perceptron learning in engineering design. *Computer-Aided Civil and Infrastructure Engineering*, *4*(4), 247-256. https://doi.org/10.1111/j.1467-8667.1989.tb00026.x

Akkermans, H., & Dellaert, N. (2005). The rediscovery of industrial dynamics: The contribution of system dynamics to supply chain management in a dynamic and fragmented world. *System Dynamics Review*, *21*(3), 173-186. https://doi.org/10.1002/sdr.317

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. https://doi.org/10.1007/s10845-021-01737-8

Al-Momani, A. H. (1996). Construction cost prediction for public school buildings in Jordan. *Construction Management & Economics*, *14*(4), 311-317. https://doi.org/10.1080/014461996373386

Ala-Risku, T., & Ka¨rkka¨inen, M. (2006). Material delivery problems in construction projects: A possible solution. *International Journal of Production Economics*, *104*(1), 19-29. https://doi.org/10.1016/j.ijpe.2004.12.027

Alshboul, O., Shehadeh, A., & Hamedat, O. (2023). Development of integrated asset management model for highway facilities based on risk evaluation. *International Journal of Construction Management*, *23*(8), 1355-1364. https://doi.org/10.1080/15623599.2021.1972204

Anderson, E. G., Morrice, D. J., & Lundeen, G. (2005). The "physics" of capacity and backlog management in service and custom manufacturing supply chains. *System Dynamics Review*, *21*(3), 217-247. https://doi.org/10.1002/sdr.319

Apaydin, H., Feizi, H., Sattari, M. T., Colak, M. S., Shamshirband, S., & Chau, K. W. (2020). Comparative analysis of recurrent neural network architectures for reservoir inflow forecasting. *Water*, *12*(5), 1500. https://doi.org/10.3390/w12051500

Arbulu, R., & Ballard, G. (2004). Lean supply systems in construction. *International Group for Lean Construction,* Conpenhague, Dinamarcapp, 1-13. IGLC. http://www.iglc.net/papers/details/291

Arbulu, R., Ballard, G., & Harper, N. (2003). Kanban in construction. *Proceedings of the 11th Annual Conference of the International Group for Lean Construction*, Blacksburgh, Virginia, 16-17. IGLC. https://www.iglc.net/Papers/Details/225

Arbulu, R., Koerckel, A., & Espana, F. (2005). Linking production-level workflow with materials supply. *In 13th International Group for Lean Construction Conference: Proceedings*, Sydney, International Group on Lean Construction, 199-206. IGLC. https://www.iglc.net/Papers/Details/365

Armstrong, J. S., & Green, K. C. (2017). Demand forecasting II: Evidence-based methods and checklists. Wharton.upenn, 1-36. https://faculty.wharton.upenn.edu/wp-content/uploads/2017/05/JSA-Demand-Forecasting-89-clean.pdf

Assaf, S. A., Al-Khalil, M., & Al-Hizami, M. (1995). Causes of delay in large building construction projects. *Journal of Management Engineering*, *11*(2), 45-50. https://doi.org/10.1061/(ASCE)0742-597X(1995)11:2(45)

Aye, G. C., Balcilar, M., Gupta, R., & Majumdar, A. (2015). Forecasting aggregate retail sales: The case of South Africa. *International Journal of Production Economics*, *160*, 66-79. https://doi.org/10.1016/j.ijpe.2014.09.033

Azambuja, M. M., Ponticelli, S., & O'Brien, W. J. (2014). Strategic procurement practices for the industrial supply chain. *Journal of Construction Engineering and Management*, *140*(7), 06014005. https://doi.org/10.1061/(ASCE)CO.1943-7862.000085

Azhar, S., Nadeem, A., Mok, J. Y., & Leung, B. H. (2008). Building information modeling (BIM): A new paradigm for visual interactive modeling and simulation for construction projects. *First International Conference on Construction in Developing Countries*, *1*, 435-446. https://d1wqtxts1xzle7.cloudfront.net/35698780/045-libre.pdf?1416785984=&response-content-disposition=inline%3B+filename%3DBuilding_Information_Modeling_BIM_A_New.pdf&Expires=1709323997&Signature=fmBiqgKRR0JqK1Es29aJXYUiAy5OSRZUyWhUm4j-QpBZdby8Jkznxr9VX14Y6d1-bKvgoQk08nqHJMgC2bWyyZIwZjkbdRwmrGD1R3LXNP9-zOu3EHQ9a0UizoRlscD3slQEWCucoaVMsuJvEpzysQ62JEV1lu~c2TzBhOZJhowDtor-jdR1Nb4rYOStDS6iWn1uzL5fNiLNm9QSDMGwQIoMpRbaIHRgqrUHMt3hDNemTY8op

Fl9c2JyplaPHDMS2SdS5mMZITmOYztoUfrlmCVnFlrikBo9YPJLyUOzKVzKJTYsg91yu mMScrnSuLI9aSNhZyS8CIuIxzn8a25Wjw__&Key-Pair-Id=APKAJLOHF5GGSLRBV4ZA

Bonaccorso, G. (2018). *Machine learning algorithms: Popular algorithms for data science and machine learning*. Packt Publishing. ISBN: 9781789347999

Box, G. E., Jenkins, G. M. (1970). Time series analysis forecasting and control: WISCONSIN UNIV MADISON DEPT OF STATISTICS.

Box, G. E., Jenkins, G. M., Reinsel, G. C., & Ljung, G. M. (2015). *Time series analysis: Forecasting and control*. John Wiley & Sons. https://doi.org/10.1111/jtsa.12194

Boyko, N., & Lukash, O. (2023). Methodology for estimating the cost of construction equipment based on the analysis of important characteristics using machine learning methods. *Journal of Engineering*, 2023. https://doi.org/10.1155/2023/8833753

Bush, V. G. (1973). *Construction management: A handbook for contractors, architects, and students*. Reston Publishing. https://works.swarthmore.edu/alum-books/1042

Caniato, F., Kalchschmidt, M., & Ronchi, S. (2011). Integrating quantitative and qualitative forecasting approaches: Organizational learning in an action research case. *Journal of the Operational Research Society*, *62*, 413-424. https://doi.org/10.1057/jors.2010.142

Carbonneau, R., Laframboise, K., & Vahidov, R. (2008). Application of machine learning techniques for supply chain demand forecasting. *European Journal of Operational Research*, *184*(3), 1140-1154. https://doi.org/10.1016/j.ejor.2006.12.004

Cengiz, A. E., Aytekin, O., Ozdemir, I., Kusan, H., & Cabuk, A. (2017). A multi-criteria decision model for construction material supplier selection. *Procedia Engineering*, *196,* 294-301. https://doi.org/10.1016/j.proeng.2017.07.202

Chaharsooghi, S. K., & Heydari, J. (2010). LT variance or LT mean reduction in supply chain management: Which one has a higher impact on SC performance?. *International Journal of Production Economics*, *124*(2), 475-481. https://doi.org/10.1016/j.ijpe.2009.12.010

Chan, A. P. C., & Chan, A. P. L. (2004). Key performance indicators for measuring construction success. *Benchmarking*, *11*(2), 203-221. https://doi.org/10.1108/14635770410532624

Chan, A. P. C., Scott, D., & Chan, A. P. L. (2004). Factors affecting the success of a construction project. *Journal of Construction Engineering and Management*, *130*(1), 153-155. https://doi.org/10.1061/(ASCE)0733-9364(2004)130:1(153

Chen, P. Y., & Liu, L. (2013). Study on coal logistics demand forecast based on pso-svr. *In 2013 10th International Conference on Service Systems and Service Management*, 130-133. IEEE. http://doi.org/10.1109/ICSSSM.2013.6602656

Chopra, S., & Meindl, P. (2021). Supply chain management: Strategy, planning, and operation. Pearson. ISBN: 9780137502844

Chukwu, A. U., & Adepoju, K. A. (2012). On the power efficiency of artificial neural network (ANN) and the classical regression model. *Progress in Applied Mathematics*, *3*(2), 28-34. http://doi.org/10.3968/j.pam.1925252820120302.1255

Cortes, C., & Vapnik, V. (1995). Support-vector networks. *Machine Learning*, *20*, 273-297. https://doi.org/10.1007/BF00994018

Cox, D. R. (1958). The regression analysis of binary sequences. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, *20*(2), 215-232. https://doi.org/10.1111/j.2517-6161.1958.tb00292.x

Dainty, A. R. J., Millett, S. J., & Briscoe, G. H. (2001). New perspectives on construction supply chain integration. *Supply Chain Management*, *6*(4), 163-173. https://doi.org/10.1108/13598540110402700

Dang, Y., Zhang, Y., Zhang, D., & Zhao, L. (2005). A KNN-based learning method for biology species categorization. *In International Conference on Natural Computation*, 956-964. Springer, Berlin. https://doi.org/10.1007/11539087_127

Deepa, S., Alli, A., & Gokila, S. (2023). Machine learning regression model for material synthesis prices prediction in agriculture. *Materials Today: Proceedings*, *81*, 989-993. https://doi.org/10.1016/j.matpr.2021.04.327

Desai, V. S., & Joshi, S. (2010). Application of decision tree technique to analyze construction project data. *In Information Systems, Technology and Management: 4th International Conference*, Bangkok, Thailand, 304-313. Springer, Berlin. ISBN : 978-3-642-12034-3

Dey, P. K. (2000). Managing projects in fast track – A case of public sector organization in India. *International Journal of Public Sector Management*, *13*(7), 588-609. https://doi.org/10.1108/09513550010362677

Dissanayaka, S. M., & Kumaraswamy, M. M. (1998). Comparing contributors to time and cost performance in building projects. *Building and Environment*, *34*(1), 31-42. https://doi.org/10.1016/S0360-1323(97)00068-1

Donohue, K. L. (2000). Efficient supply contracts for fashion goods with forecast updating and two production modes. *Management Science*, *46*(11), 1397-1411. https://doi.org/10.1287/mnsc.46.11.1397.12088

Draper, N. R., & Smith, H. (1981). *Applied regression analysis*. John Wiley & Sons. ISBN:0471029955, 9780471029953

Duda, R. O., Hart, P. E., & Stork, D. G. (2001). *Pattern classification* (2nd ed.). John Wiley & Sons. ISBN: 0471056693, 9780471056690

Elkholosy, H. (2020). *A model for forecasting owner's project management resources* [Master's thesis, University of Alberta]. ERA. https://doi.org/10.7939/r3-netf-gz33

Enshassi, A., Mohamed, S., & Abushaban, S. (2009). Factors affecting the performance of construction projects in the Gaza Strip. *Journal of Civil Engineering and Management*, *15*(3), 269-280. https://doi.org/10.3846/1392-3730.2009.15.269-280

Eriksson, P. E., & Westerberg, M. (2011). Effects of cooperative procurement procedures on construction project performance: A conceptual framework. *International Journal of Project Management*, *29*(2), 197-208. https://doi.org/10.1016/j.ijproman.2010.01.003

Fildes, R., Nikolopoulos, K., Crone, S. F., & Syntetos, A. A. (2008). Forecasting and operational research: A review. *Journal of the Operational Research Society*, *59*(9), 1150-1172. https://doi.org/10.1057/palgrave.jors.2602597

Flanagan, R., & Norman, G. (1983). The accuracy and monitoring of quantity surveyors' price forecasting for building work. *Construction Management and Economics*, *1*(2), 157-180. https://doi.org/10.1080/01446198300000012

Franco-Lopez, H., Ek, A. R., & Bauer, M. E. (2001). Estimation and mapping of forest stand density, volume, and cover type using the k-nearest neighbors method. *Remote Sensing of Environment*, *77*(3), 251-274. https://doi.org/10.1016/S0034-4257(01)00209-7

Gershenson, C. (2003). Artificial neural networks for beginners. arXiv preprint cs/0308031.
https://doi.org/10.48550/arXiv.cs/0308031

Gilliland, M. (2010). *The business forecasting deal: Exposing myths, eliminating bad practices, providing practical solutions*. John Wiley & Sons. ISBN: 978-0-470-57443-0

GOH, B. H. (1998). Forecasting residential construction demand in Singapore: A comparative study of the accuracy of time series, regression and artificial neural network techniques.

Engineering, Construction and Architectural Management, *5*(3), 261-275. https://doi.org/10.1108/eb021080

Golabchi, H., & Hammad, A. (2023). Estimating labor resource requirements in construction projects using machine learning. Construction Innovation, 1-18. https://doi.org/10.1108/CI-11-2021-0211

Gondia, A., Siam, A., El-Dakhakhni, W., & Nassar, A. H. (2020). Machine learning algorithms for construction projects delay risk prediction. *Journal of Construction Engineering and Management*, *146*(1), 04019085. https://doi.org/10.1061/(ASCE)CO.1943-7862.0001736

Gorunescu, F. (2011). *Data mining: concepts, models and techniques* (Vol. 12). Springer Science & Business Media. https://doi.org/10.1007/978-3-642-19721-5

Green, S. D., Fernie, S., & Weller, S. (2005). Making sense of supply chain management: A comparative study of aerospace and construction. *Construction Management and Economics*, *23*(6), 579-593. https://doi.org/10.1080/01446190500126882

Guo, G., Wang, H., Bell, D., Bi, Y., & Greer, K. (2003). KNN model-based approach in classification. *In OTM Confederated International Conferences*, 986-996. Springer, Berlin. https://doi.org/10.1007/978-3-540-39964-3_62

Han, J., Pei, J., & Kamber, M. (2011). *Data mining: Concepts and techniques*. Elsevier. ISBN: 9780123814807

Hayat, C., & Soenandi, I. A. (2018). The hybrid-model architectural modelling based on ARIMA-BPNN methods for building materials demands forecasting. *In MATEC Web of Conferences*, *204*, 02003. EDP Sciences. https://doi.org/10.1051/matecconf/201820402003

Haykin, S. (1998). *Neural Networks A Comprehensive Foundation*. Prentice Hall PTR. ISBN: 978-0-13-273350-2

Heravi, G., & Eslamdoost, E. (2015). Applying artificial neural networks for measuring and predicting construction-labor productivity. *Journal of Construction Engineering and Management*, *141*(10), 04015032. https://doi.org/10.1061/(ASCE)CO.1943-7862.0001006

Hibat-Allah, M., Ganahl, M., Hayward, L. E., Melko, R. G., & Carrasquilla, J. (2020). Recurrent neural network wave functions. *Physical Review Research*, *2*(2), 023358. https://doi.org/10.1103/PhysRevResearch.2.023358

Ho, C., Nguyen, P. M., & Shu, M. (2007). Supplier evaluation and selection criteria in the construction industry of Taiwan and Vietnam. *International Journal of Information and Management Sciences*, *18*(4), 403-426. https://api.semanticscholar.org/CorpusID:59061371

Hope, J., & Fraser, R. (2003). *Beyond budgeting: How managers can break free from the annual performance trap*. Harvard Business Press. ISBN: 1578518660

Horman, M. J., & Thomas, H. R. (2005). Role of inventory buffers in construction labor performance. *Journal of Construction Engineering and Management*, *131*(7), 834-843. https://doi.org/10.1061/(ASCE)0733-9364(2005)131:7(834)

Hosny, S., Elsaid, E., & Hosny, H. (2023). Prediction of construction material prices using ARIMA and multiple regression models. *Asian Journal of Civil Engineering*, *24*, 1697-1710. https://doi.org/10.1007/s42107-023-00597-2

Hua, G. B., & Pin, T. H. (2000). Forecasting construction industry demand, price and productivity in Singapore: The BoxJenkins approach. *Construction Management and Economics*, *18*(5), 607-618. https://doi.org/10.1080/014461900407419

Huang, C. H., & Hsieh, S. H. (2020). Predicting BIM labor cost with random forest and simple linear regression. *Automation in Construction*, *118*, 103280. https://doi.org/10.1016/j.autcon.2020.103280

Huang, M., Yang, M., Zhang, Y., & Liu, B. (2012). System dynamics modeling-based study of contingent sourcing under supply disruptions. *Systems Engineering Procedia*, *4*, 290- 297. https://doi.org/10.1016/j.sepro.2011.11.078

Huang, T., Fildes, R., & Soopramanien, D. (2014). The value of competitive information in forecasting fmcg retail product sales and the variable selection problem. *European Journal of Operational Research*, *237*(2), 738-748. https://doi.org/10.1016/j.ejor.2014.02.022

Hwang, S., Park, M., Lee, H. S., & Kim, H. (2012). Automated time-series cost forecasting system for construction materials. *Journal of Construction Engineering and Management*, *138*(11), 1259-1269. https://doi.org/10.1061/(ASCE)CO.1943-7862.0000536

Hyndman, R. J., & Athanasopoulos, G. (2018). *Forecasting: Principles and practice*. OTexts. https://otexts.com/fpp2/

Ihnatovich, H. (2017). *Predicting the development of the construction equipment market demand using economic indicators: Artificial neural networks approach* [Master's thesis, KTH Royal

Institute of Technology]. Diva-Portal. https://kth.diva-portal.org/smash/get/diva2:1193430/FULLTEXT01.pdf

International Standards Organisation. (2020). *Construction procurement - Part 1: Processes, methods and procedures*. OBP. ISO 10845-1:2020

Iwu, C. G. (2016). Effects of the use of electronic human resource management (E-HRM) within human resource management (HRM) functions at universities. *Acta Universitatis Danubius Administratio*, *8*(1), 5-20. https://www.proquest.com/scholarly-journals/effects-use-electronic-human-resource-management/docview/2118357596/se-2

Jantan, H., Hamdan, A. R., & Othman, Z. A. (2009). Knowledge discovery techniques for talent forecasting in human resource application. *World Academy of Science, Engineering and Technology*, *50*, 775-783. https://doi.org/10.5281/zenodo.1077930

Jaśkowski, P., Sobotka, A., & Czarnigowska, A. (2018). Decision model for planning material supply channels in construction. *Automation in Construction*, *90*, 235-242. https://doi.org/10.1016/j.autcon.2018.02.026

Jergeas, G. (2009). *Improving construction productivity on Alberta oil and gas capital projects: A report submitted to Alberta finance and enterprise*. https://open.alberta.ca/publications/improving-construction-productivity-on-alberta-oil-and-gas-capital-projects

Kandananond, K. (2012). A comparison of various forecasting methods for autocorrelated time series. *International Journal of Engineering Business Management*. *4*, 4. https://doi.org/10.5772/51088

Kasim, N. B., Anumba, C. J., & Dainty, A. R. J. (2005). Improving materials management practices on fast-track construction projects. *In 21st Annual ARCOM Conference*, SOAS, University of London, *2*, 793-802. http://www.arcom.ac.uk/-docs/proceedings/ar2005-0793-0802_Kasim_Anumba_and_Dainty.pdf

Kubat, M., & Cooperson Jr., M. (2001). A reduction technique for nearest-neighbor classification: Small groups of examples. *Intelligent Data Analysis*, *5*(6), 463-476. https://doi.org/10.3233/IDA-2001-5603

Laedre, O., Austeng, K., Haugen, T. I., & Klakegg, O. J. (2006). Procurement routes in public building and construction projects. *Journal of Construction Engineering and Management*, *132*(7), 689-696. https://doi.org/10.1061/(ASCE)0733-9364(2006)132:7(689)

Lall, U., & Sharma, A. (1996). A nearest neighbor bootstrap for resampling hydrologic time series. *Water Resources Research*, *32*(3), 679-693. https://doi.org/10.1029/95WR02966

Lam, K. C., & Oshodi, O. S. (2016). Forecasting construction output: A comparison of artificial neural network and Box-Jenkins model. *Engineering, Construction and Architectural Management*, *23*(3), 302-322. https://doi.org/10.1108/ECAM-05-2015-0080

Lam, K. C., Tao, R., & Lam, M. C. K. (2010). A material supplier selection model for property developers using fuzzy principal component analysis. *Automation in Construction*, *19*(5), 608-618. https://doi.org/10.1016/j.autcon.2010.02.007

Lee, B. H., & Scholz, M. (2006). A comparative study: Prediction of constructed treatment wetland performance with k-nearest neighbors and neural networks. *Water, Air, and Soil Pollution*, *174*, 279-301. https://doi.org/10.1007/s11270-006-9113-2

Lee, H. L., & Billington, C. (1995). The evolution of supply-chain-management models and practice at Hewlett-Packard. *Interfaces*, *25*(5), 42-63. https://doi.org/10.1287/inte.25.5.42

Lee, H. L., Padmanabhan, P., & Whang, S. (1997). Bullwhip effect in a supply chain. *Sloan Management Review*, *38*(4), 93-102. http://mitsmr.com/1phEOiM

Lee, H. L., Padmanabhan, V., & Whang, S. (1997). Information distortion in a supply chain: The bullwhip effect. *Management Science*, *43*(4), 546-558. https://doi.org/10.1287/mnsc.43.4.546

Li, H. (1996). Case-based reasoning for intelligent support of construction negotiation. *Information & Management*, *30*(5), 231-238. https://doi.org/10.1016/S0378-7206(96)01058-0

Ling, F. Y. Y., Ong, S. Y., Ke, Y., Wang, S., & Zou, P. (2014). Drivers and barriers to adopting relational contracting practices in public projects: Comparative study of Beijing and Sydney. *International Journal of Project Management*, *32*(2), 275-285. https://doi.org/10.1016/j.ijproman.2013.04.008

Ma, S., & Fildes, R. (2017). A retail store sku promotions optimization model for category multi-period profit maximization. *European Journal of Operational Research*, *260*(2), 680-692. https://doi.org/10.1016/j.ejor.2016.12.032

Ma, S., Fildes, R., & Huang, T. (2016). Demand forecasting with high dimensional data: The case of sku retail sales forecasting with intra-and inter-category promotional information. *European Journal of Operational Research*, *249*(1), 245-257. https://doi.org/10.1016/j.ejor.2015.08.029

MacQueen, J. (1967). Some methods for classification and analysis of multivariate observations. *In Proceedings of the 5th Berkeley Symposium on Mathematical Statistics and Probability*, *1*(14), 281-297. https://digitalassets.lib.berkeley.edu/math/ucb/text/math_s5_v5_frontmatter.pdf

Mawdesley, M. J., & Al-Jibouri, S. (2010). Modelling construction project productivity using systems dynamics approach. *International Journal of Productivity and Performance Management*, *59*(1), 18-36. https://doi.org/10.1108/17410401011006095

Meharie, M. G., & Shaik, N. (2020). Predicting highway construction costs: Comparison of the performance of random forest, neural network and support vector machine models. *Journal of Soft Computing in Civil Engineering*, *4*(2), 103-112. 10. https://doi.org/22115/SCCE.2020.226883.1205

Meng, X. (2012). The effect of relationship management on project performance in construction. *International Journal of Project Management*, *30*(2), 188-198. https://doi.org/10.1016/j.ijproman.2011.04.002

Micheli, G. J. L., & Cagno, E. (2016). The role of procurement in performance deviation recovery in large EPC projects. *International Journal of Engineering Business Management*, *8*, 1-17. https://doi.org/10.1177/1847979016675302

Mourya, S. K., & Gupta, S. (2012). *Data mining and data warehousing*. Alpha Science International, Ltd. ISBN: 9781842657577

Mubin, S., & Mannan, A. (2013). Innovative approach to risk analysis and management of oil and gas sector EPC contracts from a contractor's perspective. *Journal of Business and Economics*, *5*(2),149-170. https://journals.au.edu.pk/ojs/index.php/jbe/article/view/58

Naim, M., & Barlow, J. (2003). An innovative supply chain strategy for customized housing. *Construction Management and Economics*, *21*(6), 593-602. https://doi.org/10.1080/0144619032000134129

Naoum, S. G. (2016). Factors influencing labor productivity on construction sites: A state-of-the-art literature review and a survey. *International Journal of Productivity and Performance Management*, *65*(3), 401-421. https://doi.org/10.1108/IJPPM-03-2015-0045

Narayanan, A., Sahin, F., & Robinson, E. P. (2019). Demand and order-fulfillment planning: The impact of point-of-sale data, retailer orders and distribution center orders on forecast accuracy. *Journal of Operations Management*, *65*(5), 468-486. https://doi.org/10.1002/joom.1026

Navon, R. (2005). Automated project performance control of construction projects. *Automation in Construction*, *14*(4), 467-476. https://doi.org/10.1016/j.autcon.2004.09.006

Pai, P. F., & Lin, C. S. (2005). A hybrid ARIMA and support vector machines model in stock price forecasting. *Omega*, *33*(6), 497-505. https://doi.org/10.1016/j.omega.2004.07.024

Paliwal, M., & Kumar, U. A. (2009). Neural networks and statistical techniques: A review of applications. *Expert Systems with Applications*, *36*(1), 2-17. https://doi.org/10.1016/j.eswa.2007.10.005

Patil, S., & Adavi, P. (2012). A survey study of supplier selection issues in construction supply chain. *International Journal of Engineering Research and Applications*, *2*(5), 1806-1809. ISSN: 2248-9622

Patterson, D. (1996). *Artificial neural networks*. Prentice Hall, Singapore. ISBN: 0132953536

Payne, A. C., Chelsom, J. V., & Reavill, L. R. P. (1996). *Management for engineers*. John Wiley & Sons, England. ISBN: 0471956031

Pearson, K. (1901). On lines and planes of closest fit to systems of points in space. *The London, Edinburgh, and Dublin Philosophical Magazine and Journal of Science*, *2*(11), 559-572. https://doi.org/10.1080/14786440109462720

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M., & Duchesnay, E. (2011). Scikit-learn: Machine learning in Python. *The Journal of Machine Learning Research*, *12*, 2825-2830. https://doi.org/10.5555/1953048.2078195

Persad, K. R., O'Connor, J. T., & Varghese, K. (1995). Forecasting engineering manpower requirements for highway preconstruction activities. *Journal of Management in Engineering*, *11*(3), 41-47. https://doi.org/10.1061/(ASCE)0742-597X(1995)11:3(41)

Pheng, L. S., Chuan, Q. T. (2006). Environmental factors and work performance of project managers in the construction industry. *International Journal of Project Management*, *24*(1), 24-37. https://doi.org/10.1016/j.ijproman.2005.06.001

Polat, G., Arditi, D., & Mungen, U. (2007). Simulation-based decision support system for economical supply chain management of rebar. *Journal of Construction Engineering and Management*, *133*(1), 29-39. https://doi.org/10.1061/(ASCE)0733-9364(2007)133:1(29)

Probst, P., Wright, M. N., & Boulesteix, A. L. (2019). Hyperparameters and tuning strategies for random forest. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, *9*(3), e1301. https://doi.org/10.1002/widm.1301

Quinlan, J. R. (1986). Induction of decision trees. *Machine Learning*, *1*(1), 81-106. https://doi.org/10.1007/BF00116251

Rosa, J. L. A., Ebecken, N. F., & Costa, M. C. A. (2003). Towards an optimized parallel KNN-fuzzy classification approach. *WIT Transactions on Information and Communication Technologies*, *29*. DOI: 10.2495/DATA030081

Ruparathna, R., & Hewage, K. (2015). Review of contemporary construction procurement practices. *Journal of Management in Engineering*, *31*(3), 04014038. https://doi.org/10.1061/(ASCE)ME.1943-5479.0000279

Sanni-Anibire, M. O., Zin, R. M., & Olatunji, S. O. (2020). Machine learning model for delay risk assessment in tall building projects. *International Journal of Construction Management*, *22*(11), 2134-2143. https://doi.org/10.1080/15623599.2020.1768326

Seshadri, S., Chatterjee, K., Lilien, G. L. (1991). Multiple source procurement competitions. *Marketing Science*, *10*(3), 246-263. https://doi.org/10.1287/mksc.10.3.246

Shakantu, W. M., Muya, M., Tookey, J. E., & Bowen, P. A. (2008). Flow modelling of construction site materials and waste logistics: A case study from Cape Town, South Africa. *Engineering, Construction and Architectural Management*, *15*(5), 423-439. https://doi.org/10.1108/09699980810902721

Shehadeh, A., Alshboul, O., Al Mamlook, R. E., & Hamedat, O. (2021). Machine learning models for predicting the residual value of heavy construction equipment: An evaluation of modified decision tree, LightGBM, and XGBoost regression. *Automation in Construction*, *129*, 103827. https://doi.org/10.1016/j.autcon.2021.103827

Silver, E. A., Pyke, D. F., & Peterson, R. (1998). *Inventory management and production planning and scheduling* (3rd ed.). Wiley, New York. ISBN: 0471119474

Skitmore, R. M., & Ng, S. T. (2003). Forecast models for actual construction time and cost. *Building and Environment*, *38*(8), 1075-1083. https://doi.org/10.1016/S0360-1323(03)00067-2

Song, Y., Huang, J., Zhou, D., Zha, H., & Giles, C. L. (2007). Iknn: Informative k- nearest neighbor pattern classification. *In European Conference on Principles of Data Mining and Knowledge Discovery*, 248-264. Springer, Berlin. https://doi.org/10.1007/978-3-540-74976-9_25

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. https://doi.org/10.1016/j.ejor.2015.11.010

Teo, E. A., & Loosemore, M. (2001). A theory of waste behavior in the construction industry. *Construction Management and Economics*, *19*(7), 741-751. https://doi.org/10.1080/01446190110067037

Thomas, B. H. R., Riley, D. R., & Sanvido, V. E. (1999). Loss of labor productivity due to delivery methods and weather. *Journal of construction engineering and management*, *125*(1), 39–46. https://doi.org/10.1061/(ASCE)0733-9364(1999)125:1(39)

Thomas, B. H. R., Sanvido, V. E., & Sanders, S. R. (1989). Impact of material management on productivity—A case study. *Journal of Construction Engineering and Management*, *115*(3), 370-384. https://doi.org/10.1061/(ASCE)0733-9364(1989)115:3(370)

Thomas, H. R., Riley, D. R., & Messner, J. I. (2005). Fundamental principles of site material management. *Journal of Construction Engineering and Management*, *131*(7), 808-815. https://doi.org/10.1061/(ASCE)0733-9364(2005)131:7(808)

Tommelein, I. D., Levitt, R. E., & Hayes-Roth, B. (1992). Site-layout modeling: how can artificial intelligence help?. *Journal of Construction Engineering and Management*, *118*(3), 594-611. https://doi.org/10.1061/(ASCE)0733-9364(1992)118:3(594)

Torkul, O., Yılmaz, R., Selvi, I., & Cesur, M. R. (2016). A real-time inventory model to manage variance of demand for decreasing inventory holding cost. *Computers & Industrial Engineering*, *102*, 435-439. https://doi.org/10.1016/j.cie.2016.04.020

Tsai, C. F., & Wu, J. W. (2008). Using neural network ensembles for bankruptcy prediction and credit scoring. *Expert Systems with Applications*, *34*(4), 2639-2649. https://doi.org/10.1016/j.eswa.2007.05.019

Tserng, H. P., Yin, S. Y. L., & Li, S. (2006). Developing a resource supply chain planning system for construction projects. *Journal of Construction Engineering and Management*, *132*(4), 393-407. https://doi.org/10.1061/(ASCE)0733-9364(2006)132:4(393)

Vidalakis, C., Tookey, J. E., & Sommerville, J. (2011a). The logistics of construction supply chains: The builders' merchant perspective. *Engineering, Construction and Architectural Management*, *18*(1), 66-80. https://doi.org/10.1108/09699981111098694

Vidalakis, C., Tookey, J. E., & Sommerville, J. (2011b). Logistics simulation modelling across construction supply chains. *Construction Innovation: Information, Process, Management*, *11*(2), 212-228. https://doi.org/10.1108/14714171111124176

Villegas, M. A., Pedregal, D. J., & Trapero, J. R. (2018). A support vector machine for model selection in demand forecasting applications. *Computers & Industrial Engineering*, *121*, 1-7. https://doi.org/10.1016/j.cie.2018.04.042

Vrijhoef, R., & Koskela, L. (2000). The four roles of supply chain management in construction. *European Journal of Purchasing and Supply Management*, *6*(3-4), 169-178. https://doi.org/10.1016/S0969-7012(00)00013-7

Wang, T., Tang, W., Du, L., Duffield, C. F., & Wei, Y. (2016). Relationships among risk management, partnering, and contractor capability in international EPC project delivery. *Journal of Management in Engineering*, *32*(6), 04016017. https://doi.org/10.1061/(ASCE)ME.1943-5479.000045

Williams, T. (2017). The nature of risk in complex projects. *Project Management Journal*, *48*(4), 55-66. https://doi.org/10.1177/875697281704800405

Witten, I. H., Frank, E., & Hall, M. A. (2011). *Data mining: Practical machine learning tools and techniques*. Morgan Kaufmann Publishers. https://doi.org/10.1016/C2009-0-19715-5

Wong, J. M., Chan, A. P., & Chiang, Y. H. (2011). Construction manpower demand forecasting: A comparative study of univariate time series, multiple regression and econometric modelling techniques. *Engineering, Construction and Architectural Management*, *18*(1), 7-29. https://doi.org/10.1108/09699981111098667

Wong, J. M., Chan, A. P., & Chiang, Y. H. (2005). Time series forecasts of the construction labour market in Hong Kong: the Box‐Jenkins approach. *Construction Management and Economics*, *23*(9), 979-991. https://doi.org/10.1080/01446190500204911

Wu, J. D., & Chan, J. J. (2009). Faulted gear identification of a rotating machinery based on wavelet transform and artificial neural network. *Expert Systems with Applications*, *36*(5), 8862-8875. https://doi.org/10.1016/j.eswa.2008.11.020

Wu, L., Ji, W., Feng, B., Hermann, U., & AbouRizk, S. (2021). Intelligent data-driven approach for enhancing preliminary resource planning in industrial construction. *Automation in Construction*, *130*, 103846. https://doi.org/10.1016/j.autcon.2021.103846

Wu, L., Li, Z., & AbouRizk, S. (2022). Automating common data integration for improved data-driven decision-support system in industrial construction. *Journal of Computing in Civil Engineering*, *36*(2), 04021037. https://doi.org/10.1061/(ASCE)CP.1943-5487.0001001

Xu, Y., Zhou, Y., Sekula, P., & Ding, L. (2021). Machine learning in construction: From shallow to deep learning. *Developments in the Built Environment*, *6*, 100045. https://doi.org/10.1016/j.dibe.2021.100045

Yaseen, Z. M., Ali, Z. H., Salih, S. Q., & Al-Ansari, N. (2020). Prediction of risk delay in construction projects using a hybrid artificial intelligence model. *Sustainability*, *12*(4), 1514. https://doi.org/10.3390/su12041514

Yeo, K. T., & Ning, J. H. (2002). Integrating supply chain and critical chain concepts in engineer - procure - construct (EPC) projects. *International Journal of Project Management*, *20*(4), 253-262. https://doi.org/10.1016/S0263-7863(01)00021-7

Yeo, K. T., & Ning, J. H. (2006). Managing uncertainty in major equipment procurement in engineering projects. *European Journal of Operational Research*, *171*(1), 123-134. https://doi.org/10.1016/j.ejor.2004.06.036

Zaremba, W., Sutskever, I., & Vinyals, O. (2014). Recurrent neural network regularization. *arXiv preprint arXiv:1409.2329*. https://doi.org/10.48550/arXiv.1409.2329

Zhu, X., Zhang, S., Jin, Z., Zhang, Z., & Xu, Z. (2010). Missing value estimation for mixed-attribute data sets. *IEEE Transactions on Knowledge and Data Engineering*, *23*(1), 110-121. https://doi.org/10.1109/TKDE.2010.99

APPENDIX

The following snapshots show some parts of the code used in developing the models.

Neural Network - Multiple Output

In [ ]:

```
import numpy as np
from pandas import read_csv
from keras.models import Sequential
from keras.layers import Dense, Dropout
from sklearn.model_selection import train_test_split
import pandas as pd
import matplotlib.pyplot as plt
from sklearn.preprocessing import StandardScaler

# load data and arrange into Pandas dataframe
df = pd.read_excel('Data_wo_outlier.xlsx')

feature_names = ['Category', 'DayOfOrder', 'Workpackage', 'QuantityOrdered', 'P
df.columns = feature_names

# Drop rows with NaN values
df = df.dropna()

# Split into features and target (Price)
X = df.drop(['DayOfOrder', 'QuantityOrdered'], axis = 1)
y = df[['DayOfOrder', 'QuantityOrdered']]

# Split data into training and test sets
X_train, X_test, y_train, y_test = train_test_split(X, y, test_size = 0.2, rand

# Scale data, otherwise model will fail.
scaler=StandardScaler()
scaler.fit(X_train)
X_train_scaled = scaler.transform(X_train)
X_test_scaled = scaler.transform(X_test)


# define the Neural Network model
model = Sequential()
model.add(Dense(256, input_dim= 7, activation='relu'))
model.add(Dropout(0.2))
model.add(Dense(128, activation='relu'))
model.add(Dropout(0.2))
model.add(Dense(64, activation='relu'))
model.add(Dropout(0.2))
# Output layer
model.add(Dense(2, activation='linear'))

# Compile the model
model.compile(loss='mean_squared_error', optimizer='adam', metrics=['mae'])
model.summary()

# Train the model
history = model.fit(X_train_scaled, y_train, validation_split=0.2, epochs =300,


# plot the training and validation loss
loss = history.history['loss']
val_loss = history.history['val_loss']
epochs = range(1, len(loss) + 1)
plt.plot(epochs, loss, '#F2380E', label='Training loss')
plt.plot(epochs, val_loss, '#F2C80F', label='Validation loss')
plt.title('Training and validation loss')
plt.xlabel('Epochs')
plt.ylabel('Loss')
```

```
plt.legend()
plt.show()

# Plot the training and validation MAE
acc = history.history['mae']
val_acc = history.history['val_mae']
plt.plot(epochs, acc, '#F2380E', label='Training MAE')
plt.plot(epochs, val_acc, '#F2C80F', label='Validation MAE')
plt.title('Training and validation MAE')
plt.xlabel('Epochs')
plt.ylabel('MAE')
plt.legend()
plt.show()


# Predict on test data
predictions = model.predict(X_test_scaled[:5])
print("Predicted values are: ", predictions)
print("Real values are: ", y_test[:5])


# Predict on the test data
predictions = model.predict(X_test_scaled)

# Calculate MSE and MAE for each feature separately
mse_errors = ((predictions - y_test.values) ** 2).mean(axis=0)
mae_errors = np.abs(predictions - y_test.values).mean(axis=0)

print('Mean Squared Error (MSE) for DayOfOrder:', mse_errors[0])
print('Mean Squared Error (MSE) for QuantityOrdered:', mse_errors[1])
print('Mean Absolute Error (MAE) for DayOfOrder:', mae_errors[0])
print('Mean Absolute Error (MAE) for QuantityOrdered:', mae_errors[1])
```

Heatmap

In [ ]:

```
import seaborn as sns

# Using Pearson Correlation between featurs and output
plt.figure(figsize=(15,7), dpi=500)
plt.rcParams.update({'font.size': 16, "font.weight": "bold", "axes.labelweight"

cor = df.corr()
sns.heatmap(cor, annot=True, cmap=plt.cm.CMRmap_r)
plt.show()
```

In [ ]:

```
# Predict on test data
predictions = model.predict(X_test_scaled)

# Actual vs Predicted scatter plot
plt.scatter(y_test['QuantityOrdered'].values, predictions[:, 1], color='blue',

# Add a line of perfect predictions for reference
max_value = max(max(predictions[:, 1]), max(y_test['QuantityOrdered'].values))
min_value = min(min(predictions[:, 1]), min(y_test['QuantityOrdered'].values))
plt.plot([min_value, max_value], [min_value, max_value], color='red', linestyle

# Customize the plot
plt.title('Actual vs Predicted QuantityOrdered')
plt.xlabel('Actual QuantityOrdered')
plt.ylabel('Predicted QuantityOrdered')
plt.legend()
plt.grid(True)
plt.show()
```

In [ ]:
```
# Predict on test data
predictions = model.predict(X_test_scaled)

# Actual vs Predicted scatter plot for DayOfOrder
plt.scatter(y_test['DayOfOrder'].values, predictions[:, 0], color='blue', alpha

# Add a line of perfect predictions for reference
max_value = max(max(predictions[:, 0]), max(y_test['DayOfOrder'].values))
min_value = min(min(predictions[:, 0]), min(y_test['DayOfOrder'].values))
plt.plot([min_value, max_value], [min_value, max_value], color='red', linestyle

# Customize the plot
plt.title('Actual vs Predicted DayOfOrder')
plt.xlabel('Actual DayOfOrder')
plt.ylabel('Predicted DayOfOrder')
plt.legend()
plt.grid(True)
plt.show()
```

Decision Tree - Multiple Output

In [ ]:
```
import numpy as np
import pandas as pd
from sklearn.model_selection import train_test_split
from sklearn.preprocessing import StandardScaler
from sklearn.tree import DecisionTreeRegressor
from sklearn.metrics import mean_squared_error, mean_absolute_error

# Load data and arrange into Pandas dataframe
df = pd.read_excel('Data_wo_outlier.xlsx')

feature_names = ['Category', 'DayOfOrder', 'Workpackage', 'QuantityOrdered', 'P
df.columns = feature_names

# Drop rows with NaN values
df = df.dropna()

# Split into features and target
X = df.drop(['DayOfOrder', 'QuantityOrdered'], axis=1)
y = df[['DayOfOrder', 'QuantityOrdered']]

# Split data into training and test sets
X_train, X_test, y_train, y_test = train_test_split(X, y, test_size=0.2, random

# Scale the data
scaler = StandardScaler()
scaler.fit(X_train)
X_train_scaled = scaler.transform(X_train)
X_test_scaled = scaler.transform(X_test)

# Train the Decision Tree Regressor
model = DecisionTreeRegressor()
model.fit(X_train_scaled, y_train)

# Predict on the test data
predictions = model.predict(X_test_scaled)

# Calculate MSE and MAE for each feature separately
mse_errors = mean_squared_error(y_test, predictions, multioutput='raw_values')
mae_errors = mean_absolute_error(y_test, predictions, multioutput='raw_values')

print('Mean Squared Error (MSE) for DayOfOrder:', mse_errors[0])
print('Mean Squared Error (MSE) for QuantityOrdered:', mse_errors[1])
```

```
print('Mean Absolute Error (MAE) for DayOfOrder:', mae_errors[0])
print('Mean Absolute Error (MAE) for QuantityOrdered:', mae_errors[1])
```

In [ ]:
```
# Predict on test data
predictions = model.predict(X_test_scaled)

# Actual vs Predicted scatter plot
plt.scatter(y_test['QuantityOrdered'].values, predictions[:, 1], color='blue',

# Add a line of perfect predictions for reference
max_value = max(max(predictions[:, 1]), max(y_test['QuantityOrdered'].values))
min_value = min(min(predictions[:, 1]), min(y_test['QuantityOrdered'].values))
plt.plot([min_value, max_value], [min_value, max_value], color='red', linestyle

# Customize the plot
plt.title('Actual vs Predicted QuantityOrdered')
plt.xlabel('Actual QuantityOrdered')
plt.ylabel('Predicted QuantityOrdered')
plt.legend()
plt.grid(True)
plt.show()
```

In [ ]:
```
# Predict on test data
predictions = model.predict(X_test_scaled)

# Actual vs Predicted scatter plot for DayOfOrder
plt.scatter(y_test['DayOfOrder'].values, predictions[:, 0], color='blue', alpha

# Add a line of perfect predictions for reference
max_value = max(max(predictions[:, 0]), max(y_test['DayOfOrder'].values))
min_value = min(min(predictions[:, 0]), min(y_test['DayOfOrder'].values))
plt.plot([min_value, max_value], [min_value, max_value], color='red', linestyle

# Customize the plot
plt.title('Actual vs Predicted DayOfOrder')
plt.xlabel('Actual DayOfOrder')
plt.ylabel('Predicted DayOfOrder')
plt.legend()
plt.grid(True)
plt.show()
```

Random Forest - Multiple Output

In [ ]:
```
import numpy as np
import pandas as pd
from sklearn.model_selection import train_test_split
from sklearn.preprocessing import StandardScaler
from sklearn.ensemble import RandomForestRegressor
from sklearn.metrics import mean_squared_error, mean_absolute_error

# Load data and arrange into Pandas dataframe
df = pd.read_excel('Data_wo_outlier.xlsx')

feature_names = ['Category', 'DayOfOrder', 'Workpackage', 'QuantityOrdered', 'P
df.columns = feature_names

# Drop rows with NaN values
df = df.dropna()

# Split into features and target
X = df.drop(['DayOfOrder', 'QuantityOrdered'], axis=1)
y = df[['DayOfOrder', 'QuantityOrdered']]
```

```
# Split data into training and test sets
X_train, X_test, y_train, y_test = train_test_split(X, y, test_size=0.2, random

# Scale the data
scaler = StandardScaler()
scaler.fit(X_train)
X_train_scaled = scaler.transform(X_train)
X_test_scaled = scaler.transform(X_test)

# Train the Random Forest Regressor
model = RandomForestRegressor(n_estimators=100)
model.fit(X_train_scaled, y_train)

# Predict on the test data
predictions = model.predict(X_test_scaled)

# Calculate MSE and MAE for each feature separately
mse_errors = mean_squared_error(y_test, predictions, multioutput='raw_values')
mae_errors = mean_absolute_error(y_test, predictions, multioutput='raw_values')

print('Mean Squared Error (MSE) for DayOfOrder:', mse_errors[0])
print('Mean Squared Error (MSE) for QuantityOrdered:', mse_errors[1])
print('Mean Absolute Error (MAE) for DayOfOrder:', mae_errors[0])
print('Mean Absolute Error (MAE) for QuantityOrdered:', mae_errors[1])
```

In [ ]:
```
# Predict on test data
predictions = model.predict(X_test_scaled)

# Actual vs Predicted scatter plot
plt.scatter(y_test['QuantityOrdered'].values, predictions[:, 1], color='blue',

# Add a line of perfect predictions for reference
max_value = max(max(predictions[:, 1]), max(y_test['QuantityOrdered'].values))
min_value = min(min(predictions[:, 1]), min(y_test['QuantityOrdered'].values))
plt.plot([min_value, max_value], [min_value, max_value], color='red', linestyle

# Customize the plot
plt.title('Actual vs Predicted QuantityOrdered')
plt.xlabel('Actual QuantityOrdered')
plt.ylabel('Predicted QuantityOrdered')
plt.legend()
plt.grid(True)
plt.show()
```

In [ ]:
```
# Predict on test data
predictions = model.predict(X_test_scaled)

# Actual vs Predicted scatter plot for DayOfOrder
plt.scatter(y_test['DayOfOrder'].values, predictions[:, 0], color='blue', alpha

# Add a line of perfect predictions for reference
max_value = max(max(predictions[:, 0]), max(y_test['DayOfOrder'].values))
min_value = min(min(predictions[:, 0]), min(y_test['DayOfOrder'].values))
plt.plot([min_value, max_value], [min_value, max_value], color='red', linestyle

# Customize the plot
plt.title('Actual vs Predicted DayOfOrder')
plt.xlabel('Actual DayOfOrder')
plt.ylabel('Predicted DayOfOrder')
plt.legend()
plt.grid(True)
plt.show()
```

Linear Regression - Multiple Output

```
In [ ]: import numpy as np
        import pandas as pd
        import matplotlib.pyplot as plt
        from sklearn.model_selection import train_test_split
        from sklearn.preprocessing import StandardScaler
        from sklearn.linear_model import LinearRegression
        from sklearn.metrics import mean_squared_error, mean_absolute_error

        # Load data and arrange into Pandas dataframe
        df = pd.read_excel('Data_wo_outlier.xlsx')

        feature_names = ['Category', 'DayOfOrder', 'Workpackage', 'QuantityOrdered', 'P
        df.columns = feature_names

        # Drop rows with NaN values
        df = df.dropna()

        # Split into features and target
        X = df.drop(['DayOfOrder', 'QuantityOrdered'], axis=1)
        y = df[['DayOfOrder', 'QuantityOrdered']]

        # Split data into training and test sets
        X_train, X_test, y_train, y_test = train_test_split(X, y, test_size=0.2, random

        # Scale the data
        scaler = StandardScaler()
        scaler.fit(X_train)
        X_train_scaled = scaler.transform(X_train)
        X_test_scaled = scaler.transform(X_test)

        # Define and train the linear regression model
        model = LinearRegression()
        model.fit(X_train_scaled, y_train)

        # Predict on the test data
        predictions = model.predict(X_test_scaled)

        # Calculate MSE and MAE for each feature separately
        mse_errors = mean_squared_error(y_test, predictions, multioutput='raw_values')
        mae_errors = mean_absolute_error(y_test, predictions, multioutput='raw_values')

        print('Mean Squared Error (MSE) for DayOfOrder:', mse_errors[0])
        print('Mean Squared Error (MSE) for QuantityOrdered:', mse_errors[1])
        print('Mean Absolute Error (MAE) for DayOfOrder:', mae_errors[0])
        print('Mean Absolute Error (MAE) for QuantityOrdered:', mae_errors[1])
```

```
In [ ]: # Predict on test data
        predictions = model.predict(X_test_scaled)

        # Actual vs Predicted scatter plot
        plt.scatter(y_test['QuantityOrdered'].values, predictions[:, 1], color='blue',

        # Add a line of perfect predictions for reference
        max_value = max(max(predictions[:, 1]), max(y_test['QuantityOrdered'].values))
        min_value = min(min(predictions[:, 1]), min(y_test['QuantityOrdered'].values))
        plt.plot([min_value, max_value], [min_value, max_value], color='red', linestyle

        # Customize the plot
        plt.title('Actual vs Predicted QuantityOrdered')
        plt.xlabel('Actual QuantityOrdered')
        plt.ylabel('Predicted QuantityOrdered')
        plt.legend()
        plt.grid(True)
        plt.show()
```

In [ ]:
```python
# Predict on test data
predictions = model.predict(X_test_scaled)

# Actual vs Predicted scatter plot for DayOfOrder
plt.scatter(y_test['DayOfOrder'].values, predictions[:, 0], color='blue', alpha

# Add a line of perfect predictions for reference
max_value = max(max(predictions[:, 0]), max(y_test['DayOfOrder'].values))
min_value = min(min(predictions[:, 0]), min(y_test['DayOfOrder'].values))
plt.plot([min_value, max_value], [min_value, max_value], color='red', linestyle

# Customize the plot
plt.title('Actual vs Predicted DayOfOrder')
plt.xlabel('Actual DayOfOrder')
plt.ylabel('Predicted DayOfOrder')
plt.legend()
plt.grid(True)
plt.show()
```

K-Nearest Neighbors - Multiple Output

In [ ]:
```python
import numpy as np
from pandas import read_csv
from sklearn.model_selection import train_test_split
import pandas as pd
from sklearn.preprocessing import StandardScaler
from sklearn.neighbors import KNeighborsRegressor
from sklearn.metrics import mean_squared_error, mean_absolute_error

# Load data and arrange into Pandas dataframe
df = pd.read_excel('Data_wo_outlier.xlsx')

feature_names = ['Category', 'DayOfOrder', 'Workpackage', 'QuantityOrdered', 'P
df.columns = feature_names

# Drop rows with NaN values
df = df.dropna()

# Split into features and target
X = df.drop(['DayOfOrder', 'QuantityOrdered'], axis=1)
y = df[['DayOfOrder', 'QuantityOrdered']]

# Split data into training and test sets
X_train, X_test, y_train, y_test = train_test_split(X, y, test_size=0.2, random

# Scale the data
scaler = StandardScaler()
scaler.fit(X_train)
X_train_scaled = scaler.transform(X_train)
X_test_scaled = scaler.transform(X_test)

# Define and train the KNeighborsRegressor model
model = KNeighborsRegressor(n_neighbors=5)
model.fit(X_train_scaled, y_train)

# Predict on the test data
predictions = model.predict(X_test_scaled)

# Calculate MSE and MAE for each feature separately
mse_errors = mean_squared_error(y_test, predictions, multioutput='raw_values')
mae_errors = mean_absolute_error(y_test, predictions, multioutput='raw_values')

print('Mean Squared Error (MSE) for DayOfOrder:', mse_errors[0])
print('Mean Squared Error (MSE) for QuantityOrdered:', mse_errors[1])
```

```
print('Mean Absolute Error (MAE) for DayOfOrder:', mae_errors[0])
print('Mean Absolute Error (MAE) for QuantityOrdered:', mae_errors[1])
```

In [ ]:
```
# Predict on test data
predictions = model.predict(X_test_scaled)

# Actual vs Predicted scatter plot
plt.scatter(y_test['QuantityOrdered'].values, predictions[:, 1], color='blue',

# Add a line of perfect predictions for reference
max_value = max(max(predictions[:, 1]), max(y_test['QuantityOrdered'].values))
min_value = min(min(predictions[:, 1]), min(y_test['QuantityOrdered'].values))
plt.plot([min_value, max_value], [min_value, max_value], color='red', linestyle

# Customize the plot
plt.title('Actual vs Predicted QuantityOrdered')
plt.xlabel('Actual QuantityOrdered')
plt.ylabel('Predicted QuantityOrdered')
plt.legend()
plt.grid(True)
plt.show()
```

In [ ]:
```
# Predict on test data
predictions = model.predict(X_test_scaled)

# Actual vs Predicted scatter plot for DayOfOrder
plt.scatter(y_test['DayOfOrder'].values, predictions[:, 0], color='blue', alpha

# Add a line of perfect predictions for reference
max_value = max(max(predictions[:, 0]), max(y_test['DayOfOrder'].values))
min_value = min(min(predictions[:, 0]), min(y_test['DayOfOrder'].values))
plt.plot([min_value, max_value], [min_value, max_value], color='red', linestyle

# Customize the plot
plt.title('Actual vs Predicted DayOfOrder')
plt.xlabel('Actual DayOfOrder')
plt.ylabel('Predicted DayOfOrder')
plt.legend()
plt.grid(True)
plt.show()
```

Support Vector Regression - Multiple Output

In [ ]:
```
import numpy as np
import pandas as pd
import matplotlib.pyplot as plt
from sklearn.model_selection import train_test_split
from sklearn.preprocessing import StandardScaler
from sklearn.svm import SVR
from sklearn.multioutput import MultiOutputRegressor
from sklearn.metrics import mean_squared_error, mean_absolute_error

# Load data and arrange into Pandas dataframe
df = pd.read_excel('Data_wo_outlier.xlsx')

feature_names = ['Category', 'DayOfOrder', 'Workpackage', 'QuantityOrdered', 'P
df.columns = feature_names

# Drop rows with NaN values
df = df.dropna()

# Split into features and target
X = df.drop(['DayOfOrder', 'QuantityOrdered'], axis=1)
```

```
y = df[['DayOfOrder', 'QuantityOrdered']]

# Split data into training and test sets
X_train, X_test, y_train, y_test = train_test_split(X, y, test_size=0.2, random

# Scale the data
scaler = StandardScaler()
scaler.fit(X_train)
X_train_scaled = scaler.transform(X_train)
X_test_scaled = scaler.transform(X_test)

# Define and train the SVR wrapped in MultiOutputRegressor
model = MultiOutputRegressor(SVR())
model.fit(X_train_scaled, y_train)

# Predict on the test data
predictions = model.predict(X_test_scaled)

# Calculate MSE and MAE for each feature separately
mse_errors = mean_squared_error(y_test, predictions, multioutput='raw_values')
mae_errors = mean_absolute_error(y_test, predictions, multioutput='raw_values')

print('Mean Squared Error (MSE) for DayOfOrder:', mse_errors[0])
print('Mean Squared Error (MSE) for QuantityOrdered:', mse_errors[1])
print('Mean Absolute Error (MAE) for DayOfOrder:', mae_errors[0])
print('Mean Absolute Error (MAE) for QuantityOrdered:', mae_errors[1])
```

In [ ]:

```
# Predict on test data
predictions = model.predict(X_test_scaled)

# Actual vs Predicted scatter plot
plt.scatter(y_test['QuantityOrdered'].values, predictions[:, 1], color='blue',

# Add a line of perfect predictions for reference
max_value = max(max(predictions[:, 1]), max(y_test['QuantityOrdered'].values))
min_value = min(min(predictions[:, 1]), min(y_test['QuantityOrdered'].values))
plt.plot([min_value, max_value], [min_value, max_value], color='red', linestyle

# Customize the plot
plt.title('Actual vs Predicted QuantityOrdered')
plt.xlabel('Actual QuantityOrdered')
plt.ylabel('Predicted QuantityOrdered')
plt.legend()
plt.grid(True)
plt.show()
```

In [ ]:

```
# Predict on test data
predictions = model.predict(X_test_scaled)

# Actual vs Predicted scatter plot for DayOfOrder
plt.scatter(y_test['DayOfOrder'].values, predictions[:, 0], color='blue', alpha

# Add a line of perfect predictions for reference
max_value = max(max(predictions[:, 0]), max(y_test['DayOfOrder'].values))
min_value = min(min(predictions[:, 0]), min(y_test['DayOfOrder'].values))
plt.plot([min_value, max_value], [min_value, max_value], color='red', linestyle

# Customize the plot
plt.title('Actual vs Predicted DayOfOrder')
plt.xlabel('Actual DayOfOrder')
plt.ylabel('Predicted DayOfOrder')
plt.legend()
plt.grid(True)
plt.show()
```