Case Fill Rate Prediction

by

Madeleine Lee
Master's in Supply Chain Management, Curtin University, Australia

and

Kamran Siddiqui

Master's in Business Administration, Institute of Business Administration

SUBMITTED TO THE PROGRAM IN SUPPLY CHAIN MANAGEMENT
IN PARTIAL FULFILLMENT OF THE REQUIREMENTS FOR THE DEGREE OF

MASTER OF APPLIED SCIENCE OR MASTER OF ENGINEERING IN SUPPLY CHAIN MANAGEMENT
AT THE
MASSACHUSETTS INSTITUTE OF TECHNOLOGY

May 2023

Signature of Author: _____
Department of Supply Chain Management
May 12, 2023

Signature of Author: _____
Department of Supply Chain Management
May 12, 2023

Certified By: _____
Dr Elenna R Dugundji
Research Scientist, Supply Chain Management Program
Capstone Advisor

Certified by: _____
Dr Thomas Koch
Postdoctoral Associate, Supply Chain Management Program
Capstone Co-Advisor

Accepted by: _____
Prof. Yossi Sheffi
Director, Center for Transportation and Logistics

Case Fill Rate Prediction
by
Madeleine Lee
and
Kamran Siddiqui
Submitted to the Program in Supply Chain Management
on May 12, 2023, in Partial Fulfillment of the
Requirements for the Degree of Master of Applied Science or Master of Engineering in Supply Chain Management

Capstone Advisor: Dr Elenna R Dugundji
Title: Research Scientist, Supply Chain Management Residential Program
Capstone Co-Advisor: Dr Thomas Koch
Title: Postdoctoral Associate

ACKNOWLEDGMENTS

We want to express our heartfelt appreciation and gratitude to our advisor, **Dr. Elenna Dugundji** for her unwavering guidance, expertise, and support throughout our capstone project. Their dedication and willingness to share their knowledge and experience have been invaluable to us. We are truly grateful for their mentorship and inspiration. Thank you for believing in us and pushing us to exceed our own expectations.

We would like to express our sincere appreciation to our sponsor company, **Mr. K and Mr. S**, for their remarkable expertise and generosity in supporting our capstone project. Their deep technical knowledge and innovative approach to the field have been instrumental in guiding us and providing us with strong guidance along the way. We are truly honored to have had the opportunity to collaborate with such accomplished professionals, and we are immensely grateful for their dedication and support.

We would like to express our sincere gratitude to our classmates, postdoctoral colleagues, and faculty members from the **Center of Transportation and Logistic**s for their unwavering support and camaraderie throughout our capstone project. Late nights in the CTL lab were made enjoyable by your presence, and your encouragement and constructive criticism were invaluable in helping us refine our ideas and improve the quality of our work. We cherish the friendships we have formed with you, and we are grateful for the positive impact you have had on our lives and our academic journey.

ABSTRACT

Stockouts present significant challenges for Fast-Moving Consumer Goods (FMCG) companies, adversely affecting profitability and customer satisfaction. This Capstone report investigates key drivers causing Case Fill Rate (CFR) to fall below target levels and identifies the best model for predicting future CFR for the sponsor company. By utilizing hypothesis testing and feature permutation techniques, we conclude that forecasting error is the most critical driver influencing CFR. Machine learning techniques including classification, regression, gradient boosted trees (XGBoost), convolutional neural network-long short term memory model (CNN-LSTM), and multi-step LSTM techniques were deployed to predict CFR. Advanced machine learning techniques demonstrated potential in predicting short term CFR. To improve longer-term forecasts, a combination of models should be incorporated, along with extended historical data, promotions data, and consideration of exogenous variables. Companies should prioritize forecasting accuracy and optimize inventory policy to improve CFR in the long run.

Table of Contents

LIST OF FIGURES

LIST OF TABLES

LIST OF EQUATIONS

1.0 INTRODUCTION

1.1 Business Motivation

Fast-Moving-Consumer-Goods (FMCG), products intended for everyday consumption, are valued at over $10 trillion USD, and projected to reach $15 trillion USD in 2025 (Shankar, 2022). The FCMG industry is often faced with challenges driven by the high complexity of supply chains and wild fluctuations of demand due to frequent promotional campaigns and changes in consumer behavior. Recent events like COVID-19, the Suez Canal blockage, US port congestion and geopolitical issues in Europe have led to massive supply chain disruptions. These disruptions include raw material and labor shortages, reduced transportation and production capacities and major hikes in freight and utility tariffs, which have further enhanced the vulnerability of the FMCG industry, causing a ripple effect of customer demand being unfulfilled. These disruptions have become increasingly frequent in the last decade, driving the urgency of companies to build a resilient and robust supply chain network to proactively mitigate the negative impact of such scenarios.

The sponsoring company, an FMCG company with multiple product lines, is also vulnerable to these disruptions. The capstone focused on one product line due to the business maturity with well-established customer demand patterns.

As shown in Figure 1, the company's products were typically produced either in their own manufacturing plant or by manufacturing partners and subsequently moved to distribution centers (DC) for storage. The company's primary customers, usually large retailers, would place

orders that would be fulfilled by the DC. The DC would then select and package the ordered volume and ship it to the customer, who would later sell it to consumers. The company used Case Fill Rate (CFR) as a performance metric to evaluate their performance against customer orders and set a predetermined CFR target level that they wanted to attain. CFR was computed by dividing total shipments by total customer orders. An example of CFR calculation can be seen in the Table 1 where we have data for different transactions and respective cut quantities are calculated. Further calculation for CFR is highlighted in Equation 1, where we can calculate CFR based on the data in Table 1. Multiple factors, including demand forecast, inventory levels, and production planning, influenced CFR performance. The sponsor company employed the term "cut" to describe ordered volumes that were unable to be fulfilled.

Figure 1

Supply Chain Network of Sponsoring Company

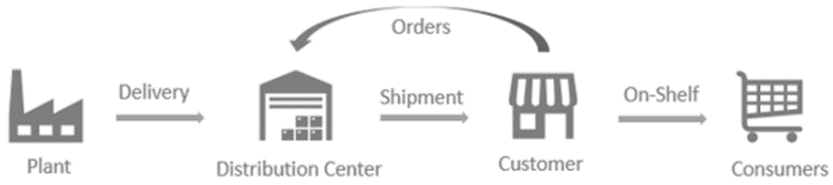


Table 1
Example of case fill rate calculation

Order ID	SKU #	Customer ID	Customer Volume	Order	Shipped Volume	Cut
0001	101	AAA	2,000		1,500	500
0001	103	AAA	2,000		2,000	0
0002	102	BBB	1,500		1,250	250
0003	106	CCC	3,000		2,000	1,000
0003	105	CCC	2,500		2,500	0
Total			11,000		9,250	1,750

Equation 1
Case Fill Rate Calculation

$$CaseFillRate = \frac{(TotalShipmentOrders)}{(TotalOrders)} = \frac{9{,}250}{11{,}000} = 84.09\%$$

A low CFR translates to lost sales. In this case, for the sponsoring company, a drop of one percent in sales contributes to over millions of losses in net profit margin. Furthermore, lost sales negatively impact customers' loyalty and brand confidence, as switching costs for FMCG products are relatively low. A recent study (NielsenIQ, 2022) reports that 70% of consumers will purchase an alternate brand if their regular choice is out of stock. With the further pressure of rising inflation, consumers are more price sensitive and therefore actively seeking alternate brands that are priced lower. Moreover, it is exceedingly difficult to regain customers if their trust in the brand is diminished. Consequently, the company will need to increase spending on advertising and marketing campaigns to regain market share and customer loyalty. Occasionally, there are

contractual penalties that are required to be paid by the sponsoring company when customer orders are unfulfilled due to CFR being below the target level.

1.2 Problem Statement and Research Questions

The company currently relies on regional supply chain managers to provide a forward-looking four-week Sales Projected Inventory (SPI) forecast on a weekly basis. The supply chain managers generally use a non-mathematical approach, combining knowledge with a heuristic approach to derive forecasts. The forecast is often overly optimistic and understates the risk that could potentially result in a low CFR. The current methodology used to derive SPI forecasts is also unable to support intelligent business decisions for Incremental Business Assessment (IBA). IBA is a framework that determines whether the company should accept volume purchases with discounts proposed by customers. IBA often requires a rapid response within a short timeframe for the company to determine whether they should commit to fulfilling these incremental orders by customers. The absence of a forward-looking CFR prediction in the past prevented the company from making timely adjustments to prevent CFR from falling below the target level.

The capstone addresses the following questions:

1. What is the major risk driver(s) that was causing the company's CFR to fall below target level for the last 3 years?
2. Which risk drivers are relevant to predict future CFR?
3. What is the best model to project future case fill rate driven by identified risk driver?

1.3 Scope: Project Goals and Expected Outcomes

The objective of the capstone project was to gain a better understanding of the major risk factor(s) that had contributed to the consistently low CFR over the past three years. Additionally, the project aimed to develop a predictive model that could estimate future CFR based on the key risk driver(s) identified through the analysis of data from the previous three years.

A data-driven approach was hypothesized to be the most appropriate approach for this capstone, which was divided into two stages. First, the major risk driver that was driving the low CFR in the past three years was identified, which also set the feature of future case fill rate projection. Second, the major risk driver(s) identified in the first stage were used as variables to be incorporated into a model that could predict CFR for the upcoming weeks so that the company could take appropriate actions to mitigate this loss. The literature reviewed in this capstone included impact of stock outs, challenges of demand forecasting and inventory forecasting in the FMCG industry, and machine learning modelling techniques.

The deliverables to the company include:

1. Analytics to identify the major risk driver(s) that were causing CFR to fall below the target level for the last three years.
2. A forecast model that provides a forward looking 13-week CFR

To build a CFR prediction model for the sponsor company, our project plan included reviewing literature related to inventory management, demand forecasting challenges, supply chain

metrics, and stock out impacts, as well as machine learning algorithms used in similar projects. We collected and examined quantitative data from the sponsoring company and conducted weekly meetings with key stakeholders to collect qualitative data such as operations and logistic network and identify key business variables related to the project. Descriptive analytics were performed on the data, and hypothesis testing was conducted for preliminary data validation. Multiple machine learning techniques were used for pattern recognition, clustering, identifying key variables, and predictive models. Based on the machine learning technique that provided the best output, model validation was performed using a test dataset. Lastly, we formulated managerial insights and recommendations for the company as well as areas for future research.

2.0 LITERATURE REVIEW

To dive deeper into the core problem of the capstone – identifying the major risk driver(s) causing a low CFR and how can we predict CFR – we reviewed literature on: (1) inventory management in FMCG and demand forecasting challenges, (2) supply chain metrics & impact of stock outs, and (3) big data and machine learning algorithms, as these are most relevant to this capstone project.

2.1 Inventory Management in FMCG industry

Fast-moving consumer goods (FMCGs), as the name suggests, refer to everyday products that sell quickly at a relatively low cost to a broad consumer base. To cater to this dynamic demand and to mitigate this quick turnover challenge, companies must maintain optimum inventories both on the shelves as well as in different echelons in their supply chains (ITC Infotech, 2020). FMCGs generally operate on a built-to-stock model, as they have shorter lead times to fulfill

customer orders so they must invest in finished goods inventory to cover forecasted demand while minimizing their financial exposure (Gundogdu et al., 2019).

Deleted: .

One major challenge in supply chain management is determining the optimal inventory level for each level of the supply chain (Inderfurth, 1991). Insufficient inventory can result in stockouts and a decrease in the customer base, while excess inventory can lead to high costs such as storage and financial expenses. Inventory, or safety stock, is typically kept in a supply chain to cover inefficiencies such as demand forecast accuracy, supply plan compliance, delays in production and upstream supply chains, transportation disruptions, logistics disruptions, and so on. Inderfurth (1991) emphasized that safety stock is a critical component of inventory management, and it acts as a buffer to compensate for demand forecast inaccuracy and demand fluctuations, which will be discussed in the next section.

2.1.1 Demand Forecasting

As the expression goes, "forecasts are never accurate": maintaining high forecast accuracy for a consumer goods company is a major challenge, as sales operations are heavily dependent on the accuracy of the demand forecast. However, research by EKN (2016) has shown that the average forecast accuracy for the consumer goods industry is approximately 60% irrespective of the forecast method used.

Operating with high accuracy in demand forecast allows a company to maintain low inventories in terms of safety stock and reap the financial benefits that come with this. Conversely, a high

forecast error requires greater investment in safety stock inventory to cover demand variation. Gruen (2002) suggests that approximately 47% of stock out events are due to poor demand forecast accuracy, which leads to either understocking or overstocking of inventory. Overstocking inventory has a negative financial impact on the organization, whereas understocking inventory leads to stock outs, loss of sales, and damage to brand confidence, which jeopardizes customer relationships (Raman and Kim, 2002).

Moreover, poor demand forecast accuracy also leads to multiple supply chain inefficiencies, such as increased logistics costs to expedite the shipments, and reallocation of inventory from sub optimal regions within the network to fulfill backorder (Martinsson & Sjoqvist,2019). Low forecast accuracy over time also causes a bullwhip effect from upstream to manufacturing, resulting in frequent production plan changes. Suppliers in the upper echelons tend to build more inventories within the network to cater to sudden demand variations (Chen et al, 2000).

2.1.2 Demand Uncertainty Caused by Discounts & Promotions

In the FMCG industry, sales promotions and discounts are key marketing strategies to boost sales. Products can be selected for promotion for multiple reasons like slow moving or dead inventory, high sales targets, and remaining shelf life. However, the main goal of all promotions is to boost sales. Sales promotions are mainly of two types: trade and consumer sales promotions as illustrated in Figure 2 (Nigam, 2016). Trade promotions are used to boost primary sales and are targeted to the trade (retailers, wholesalers, distributors) in the form of discounts, commissions, and incentives so that they stock more products, give better visibility to the product, and thus

generate more sales from the end consumers. Consumer promotions are sales promotion activities, they are targeted directly toward the end consumer and are advertised in public media to attract the attention of the masses (Nigam, 2016). Promotional activities vary from region to region and are highly influenced by local consumer buying patterns.

Figure 2
Types of Consumer Promotions

(Note: Adapted from: Dibbs S., Simkin L., Pride W.M., and Ferrell O.C., (2006).)

Promotions drive an increase in sales revenue of approximately 45% as compared to non-promotional sales. Promotions entice consumers through a financial incentive. (Ashraf et al., 2014).

Selected products are put on promotion by manufacturers and sometimes retailers, to increase the customer base, increase the retailer's margin for the product (Ashraf et al., 2014), promote brand switching, and market new products (Blattberg et al.,1995).

2.2 Supply Chain Metrics & Impact of Stock Outs

In supply chain, the main goal is "to get the right product in the right quantity in the right place and at the right time, at minimum total cost" (Carvalho et al., 2017). The KPIs or metrics are selected based on their relevance and impact on the supply chain process efficiency for the organization. To build an efficient supply chain, companies tend to follow numerous supply chain metrics or key performance indicators (KPIs) to gauge and improve their processes (Lohmann et al, 2004). These metrics vary at each echelon and function of the supply chain. All these KPIs and performance drivers contribute to the efficiency of the organization's supply chain to achieve customer satisfaction by fulfilling demand. For example, in the upper echelon of the supply chain, demand forecast accuracy is used to gauge the deviation of actual demand from forecasted demand within the demand planning function, whereas total cycle time measures the time required to convert raw materials into finished products within the production function (Drew, 2021). Conversely, OTIF (On-Time-In-Full) and Case Fill Rate (CFR) are commonly used metrics in the lower echelons to measure the performance of customers' orders fulfillment and the ability to deliver as per the promised date (Calhoun, 2021). In this capstone, the sponsoring company uses CFR as a key metric and a benchmark to measure the performance of customer delivery.

Product stock outs remain a key challenge for consumer goods industries and the entire world saw the severe effects of stock outs during COVID pandemic (EKN Research, 2016). In a detailed study, EKN Research (2016) published the below statistics on the impact of out-of-stock events in fast moving consumer goods industry.

- In North America, out-of-stock events cause an annual loss of approximately $129.5 billion while in Europe, 7-10% of annual sales are lost due to stock-outs. Accumulated losses from overstocking and stock-outs are worth $1.1 trillion every year.
- The average stock-out rate is around 8% for all categories while the out-of-stock rate exceeds 10% for promotional products.
- It is more likely that consumers switch brands rather than switch stores/retailers however, repeated out-of-stock products lead to 70% of consumers switching stores/retailers rather than searching for an alternative brand.
- Inaccurate demand forecasts make the highest contribution (47%) to out-of-stock events.

As the objective of this capstone is to predict the CFR impact caused by stock out events using the data driven approach, we will discuss the topics of big data and machine learning algorithms in the next section.

2.3 Big Data & Machine Learning Algorithms

Big data is defined as distributed computing architectures that consist of three Vs: big volumes, more velocity, and a great variety (Henry, 2019). Big data is incredibly valuable for businesses of

all industries to increase productivity and competitiveness. Manyika et al. (2011) projected that if the private sector efficiently leveraged big data to promote operational efficiency and quality, profit margins could grow by up to 60%. The data avalanche generated by a vast amount of transactional data in the FMCG business had proficiently enhanced advanced analytical to drive revenue growth. For example, the history of sales data and promotion activity provided vital insight into consumers' behavior in relation to price changes, empowering companies to develop marketing strategies to influence customers' purchases (Infosys, 2020).

Big data has empowered machine learning and predictive analytics with advanced algorithms to improve forecast accuracy within the supply chain, compared to traditional forecast models. Machine learning is mainly classified into supervised and unsupervised learning. Supervised learning requires pre-defined labels and is designed to train the algorithm in classifying data and high accuracy of predictive outcome, while the latter does not require pre-defined labels (Dickson,2020). Chase (2016) discussed that traditional forecasting models based on timer series are restricted to only a few variable factors like demand history. In contrast, forecast models based on machine learning can incorporate unlimited variable factors that are relevant to the forecasting model to improve robustness and accuracy. A study by Carbonneau et al. (2007) showed that forecast error in machine learning techniques is lower compared to traditional techniques such as naïve and moving average. Table 2 depicts a summary of literature review of studies using various machine learning techniques.

Table 2

Literature review of studies using machine learning technique.

Title	Industry	Methods	Features	Conclusion
A Comparison of Various Forecasting Methods for Autocorrelated Time Series (Kandananond, 2012)	Consumer goods packaging	Artificial neural network (multilayer perceptron and radial basis function), support vector machine, and autoregressive integrated moving	Consumer product brands and demand	Support vector machine perform better than ARIMA and artificial neural networks
Customer and Product Clustering in Retail Business(Ondrej, Vlamidir, Tomas , 2020)	Drugstore	K means clustering	Basket data, purchase record, demographic data, customer choice of online delivery or store pickup	One general cluster and six specialised clusters that can be use for promotioncal campaigns
Forecasting Seasonal Footwear Demand Using Machine Learning (Liu and Fricke, 2018)	Footwear	Regression trees, random forest, k-nearest neighbors, linear regression and neural networks	Store count, month, lifecycle month, gender, AUR, year, basic material, MSRP, color, lifecycle, cut description, product class description	Ensemble methods (median and average) and random forest gave the best predictive performance
Identifying the Root Causes of Stockout Events in e-commerce Using Machine Learning Techniques (Chao and Izaguirre, 2021)	E-commerce	Multiple lineaer regression, logistic regression, omnibus test, Nagelkerke R square, confusion matrix	Order quantity, order size, inventory avaialbility	Multiple linear regression is better than logistic regression. Accuracy of logistic regression is not high enough to make reliable predictions
Oral-Care Goods Sales Forecasting Using Artificial Neural Network Model (Vhatkar and Dias, 2016)	Oral care	Back-propagation neural network model	Sales forecast and brand category	Back propagation learning algorithm provides the best prediction and most accurate results to predict future sales
Predicting Shipping Time with Machine Learning (Jonquais, Krempl 2019)	Shipping	Linear regression, random forest ,neural network	Shipment records, unloading date, carriers, shippers, routes	Random forest performs the best out of all three models.
Predictive Demand Models in the Food and Agriculture Sectors:(Pezente, 2018)	Food industry	Linear regression, ARIMA, artificial neural network	Consumption, price, volume, GDP, population	Neural network models were significantly more accurate
Spare Parts Predictive Analytics for Telecommunications Company (Mamakos)	Telecommnication spare part	Naïve Bayes, decision tree, random forest	Forecast demand, geographic, number of sites (active, deactive)	Random Forest provides the best accuracy. Naïve Bayes gives an accurate true positive, but performs poorly in negative and it is not suitable for the model. Decision tree outperforms Naïve Bayes, but misses major clusters
Transforming eCommerce Product Segmentation with Machine Learning (Arora and Bosch, 2022)	Consumer goods packaging	Support vector machine and artificial neural network	Unit price, annual/quarterly demand, demand fluctuation, inventory turnover, profit margin, and brand priority influence	Both models fit well, no signs of overfit. Both models is effective for multi attribute inventory classification problem especially growing portfolios
Utilizing Artificial Neural Networks to Predict Demand for Weather-Sensitive Products at Retail Stores (Taghizadeh, 2017)	Retails product	The multilayer perceptron, time delay neural networks, recurrent neural networks, bagging, linear regression	Sales of potentially weather-sensitive products, weather	The multilayer perceptron with the back propagation learning algorithm is the best model

2.3.1 Supervised Machine Learning

Classification and regressions are the main algorithms in supervised learning. Regression algorithms such as linear regression and logistic regression examine the correlation between independent and dependent variables and predict numerical variable output, such as stock out rate. In contrast, classification algorithms like Support Vector Machine (SVM), Decision Tree, and Random Forest, groups test data into specific categories and predict categorical output. For example, can a DC ship a full quantity? (Delua,2021). These algorithms are frequently applied to identify root causes and prediction models on stock out events.

In a study that examines the availability of spare parts during unplanned asset failure, Mamakos (2022) uses classification algorithms including Naive Bayes, Decision Tree and Random Forest in their model to predict the potential stock out due to demand spike. The study uses multiple variables such as the actual run rate of each material and the frequency of demand spikes due to asset failure in different regions. Mamakos (2022) concluded that Random Forest provides the best outcome as it leverages multiple exogenous variables to co-relate to demand spikes, such as traffic and extreme weather conditions.

Multiple linear regression and logistics regression were used by Choa and Izaguirre (2022) to identify the major drivers of stock out events in the e-commerce industry; they found order size and order quantity are key variables that influence stock out event. They concluded that variables related to demand are more correlated to predict future stock out events as compared to supply variables in their studies. In our study, we will use multiple supervised learning algorithms including decision tree, naïve bayes to identify the key variables that have the most impact on a

low CFR, and we further develop a predictive model to forecast CFR using both regressions and classification algorithms to test the correlation between multiple variables.

2.3.2 Unsupervised Machine Learning

In contrast to supervised learning, unsupervised learning requires fewer manual data preparation as it does not require pre-classifications of labels unlike supervised learning (Alzubaidi,2022). Clustering and dimensionality reduction are algorithms within unsupervised learning commonly used to explore segmentations based on similar data points to understand the most influential dependent variables, detect anomalies using pattern recognition, and gain hidden insights. Examples of unsupervised learning include K-means, Nearest Neighbor, Principal Component Analysis (PCA). In a study conducted by Sokol et al. (2021), they examine consumers' basket data for a drugstore using a clustering algorithm, K-means to group customers' shopping behavior based on purchased items and customer demographic. The study later reveals interesting insights on special cluster groups with the correlation of social economic status and product group, for instance, perfumes are usually bought by wealthy customers. This insight could translate to the requirement for higher inventory levels for perfumes in retail stores that are positioned in wealthy neighborhoods. Furthermore, a study by Nikolopoulos et al (2016) also found evidence that the Nearest Neighbor technique can improve forecasting for sporadic demand through repetitive patterns.

3.0 DATA and METHODOLOGY

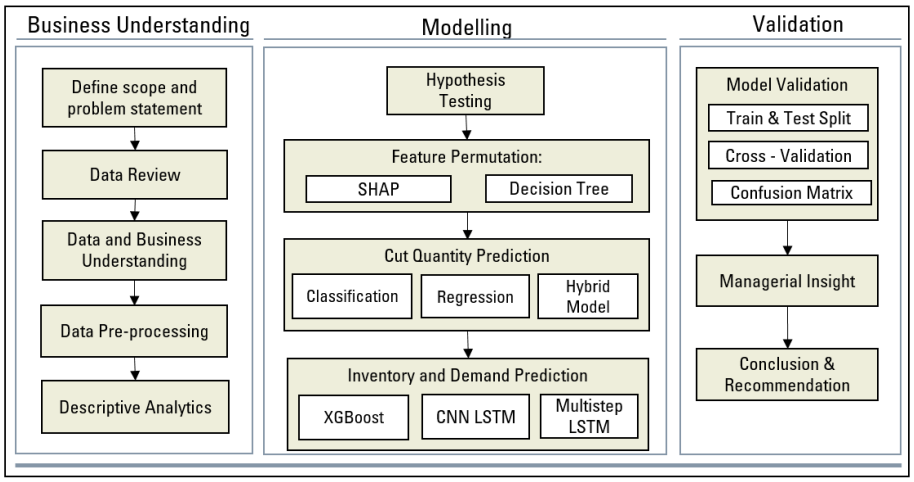
3.1 Scope

The capstone analyzes the performance of the sponsor company's one product line. The study is focused on that one product line as the first step, analyzing the orders received, forecasting accuracy, ability to fulfill this demand i.e., Case Fill Rate and inventory management for that product line. The capstone provided cut quantity prediction, insights, and recommendations to the sponsor company to address the following questions:

1. What is the major risk driver(s) that was causing CFR below target level for the last 3 years?
2. Which risk drivers are relevant to predict future CFR?
3. What is the best model to project future case fill rate driven by identified risk driver?

In this capstone, we segmented our approach into 3 major phrases. (1) Data and Business Understanding (2) Modelling (3) Validation, as shown in Figure 3.

Figure 3
Methodology for CFR Prediction



3.1 Data Understanding and Business Understanding

The dataset provided by the sponsor company contained three years of sales transactional data, including details on customer purchase history, inbound and outbound shipment of distribution centers, daily inventory levels, demand forecast, manufacturing plans, SKU master data, and labor capacity as shown in Table 3.

The dataset consists of multiple tables:

1) **Daily Forecast**: This table contains the forecasted demand for each SKU for the next 60 days. The system uses statistical forecasting by analyzing historical sales data, demand patterns, and incorporates relevant factors like promotion data to generate the forecast

60 days prior to the actual date. The demand planner reviews and validates the forecast plan on a weekly basis and adjusts before finalizing the plan for production.

2) **Planned Production Plan**: This table contains information on the planned quantity for production, actual quantity produced, and projected inventory consisting of stock on hand, inventory inflow, and outflow.

3) **Inventory**: This table contains information on the stock on hand level, safety stock levels, and demand for the current week plus the expected demand for the next 1 to 25 weeks. If the stock on hand is higher than the shipment cut quantity, the data will be dropped from the dataset for modelling.

4) **Order delivery:** This table contains information on the orders placed by customers, including the timestamp of the order, the required delivery date, and the actual shipment creation date. If the total ordered quantity is not shipped in full, the remaining quantity is considered as a shipment cut.

5) **Master Data**: This table contains the specifications and characteristics of each product, including packaging, sizes, and minimum quantity per stocking unit.

Table 3
List of key features from the dataset

Table	Feature	Description
Daily Forecast	Forecast Date	Date for which forecast was created
Daily Forecast	Forecast Quantity	Forecast value in SU, this is forecast for one day (Forecast Date)
Daily Forecast	Forecast Generated Date	Date when forecast was generated
Daily Forecast	Location	Distribution Center code - ship from location
Daily Forecast	Product ID	Distribution center identifier, ship from location
Inventory	Calendar Date	Date, all values are reported in the end of the day of calendared day
Inventory	Available Stock on Hand	Stock available for shipment
Inventory	Days forward Coverage	Coverage (in days) of available stock, taking into consideration current forecast
inventory	Plant ID	Distribution Center code - ship from location
inventory	SKU ID	SKU Product Identifier
Order Deliveries	Product ID	SKU Product Identifier
Order Deliveries	Material Available Date	Date on which product must be ready to be shipped based on customer requested delivery time
Order Deliveries	Plant ID	Distribution center identifier, ship from location
Planned Production	Calendar Date	Planned production date generated
Planned Production	Ship From	Shipped from Location
Planned Production	Actual SOH	Actual Stock on hand available
Planned Production	Material	Finish product code

3.2 Data Preprocessing

1) **Data cleaning:** During the data cleaning phase, the dataset containing over 5 million rows was reviewed to ensure the integrity of the data for the model. Null, outlier, and inconsistent data were identified and handled. Null data were either dropped or replaced using mean imputation or the previous value. Outliers were identified using quartiles and

removed from the dataset. Inconsistent data such as negative purchase order quantity was also removed prior to modeling. Besides, inconsistent data like negative purchase order quantity is also removed from the dataset prior to modelling.

2) **Data integration:** Data integration was a significant effort, as the dataset consisted of multiple tables with different datatypes and data structures. Detailed data manipulation was necessary to ensure the correctness of the output data. For example, the date format varied in most tables, including yyyymmdd, mm-dd-yy, and ddmmyyy, requiring extra care in formatting prior to merging and joining the dataset. Additionally, data in different tables consisted of multiple types, requiring changing data types of features, scaling numerical values, and encoding numerical values to categorical values as a prerequisite to fitting them into the model. The tables in the database are linked together based on key attributes such as SKU, plant, and date, which act as the essential conditions for joining the tables.

3) **Feature selection and reduction:** Feature selection is done to help reduce the complexity of the data, improve accuracy, and increase efficiency for computation time for modelling. In feature reduction, the number of features in the dataset is reduced by selecting a subset of the most informative and relevant features for a particular modeling task. Critical features are identified to be included in the model as shown in Table 3. Besides, some features with common attributes are grouped together. For instance, features related to customer behavior, such as monthly order frequency and order volume may be

grouped together to create a composite feature that captures the overall behavior of the customers as one feature. Similarly, features related to product attributes, such as SKU type (liquid or powdered) or packaging size, may be grouped together to simplify the analysis or modeling task.

4) **Date aggregation:** This is a key step in data preprocessing before we can proceed with modelling. This was a process of combining multiple data points into a single data point to create a more manageable and informative dataset. This step helped reduce the noise in the data, improve the accuracy and reliability, simplify modelling and address data quality issues. The dataset provided by the sponsor company contained the daily transactional data for sales however orders from customers are not on a daily cycle basis due to which there is high random variation in the data, there were some dates with zero orders and some with very large orders, this variation was reduced by aggregating the data on a weekly basis.

3.3 Modelling

3.3.1 Time series analysis

Time series analysis is a widely employed technique in academic research for understanding patterns by examining a set of historical data observed over time. The primary goal is to identify trends, seasonality, and cycles, which can then be used for forecasting future values. In the

context of case fill rate prediction, two major approaches for time series analysis are univariate and multivariate time series analysis.

Univariate time series analysis focuses on identifying trends and seasonal patterns using a single variable over time, such as customer order quantity of a single SKU or inventory levels. Autocorrelation function (ACF) is a key concept in univariate analysis, which measures the correlation between a time series and its lagged values. ACF can reveal patterns such as seasonality and help in determining the appropriate model for forecasting.

Multivariate time series analysis, on the other hand, examines multiple variables over time, such as projected inventory, labor availability, and demand to project case fill rate. In this approach, partial autocorrelation function (PACF) plays a significant role as it measures the correlation between a time series and its lagged values while controlling for the effects of other lagged variables. PACF can help identify the direct influence of a specific variable on the target variable like cut quantity, projected inventory, and forecast demand.

3.3.2 Data Visualization

Data visualization is an essential tool that provides insights into trends and patterns, making data more interpretable and understandable. Data visualization also allows users to detect anomalies in historical data easily. For example, a sudden drop in inventory level may indicate issues with production that need to be addressed to avoid a low case fill rate. Moreover, data visualization

can be used to project spikes in demand during weekend or holiday periods, or regular fluctuation in customer order overtime. The capstone includes many data visualizations for users and sponsoring company which provide insights for informed decision making.

3.3.2 Hypothesis testing

One of the key questions of our project is to determine the factors that influence the Case Fill Rate (CFR) the most. Based on our findings in the literature review and discussions with the industry experts, we formulated the hypotheses below

(i) H_o : The higher the Coefficient of Variation (CV), the lower the CFR

(ii) H_o : The higher the Stock out Frequency (SOF), the lower the CFR

(iii) H_o : The higher the Coefficient of Variation (CV), the higher the SOF

(iv) H_o : The higher the Forecasting Error (FE), the lower the CFR

(v) H_o : The higher the Days inventory cover (DFC), the higher the CFR

(vi) H_o : The higher the shipments cuts (Cuts), the lower the CFR

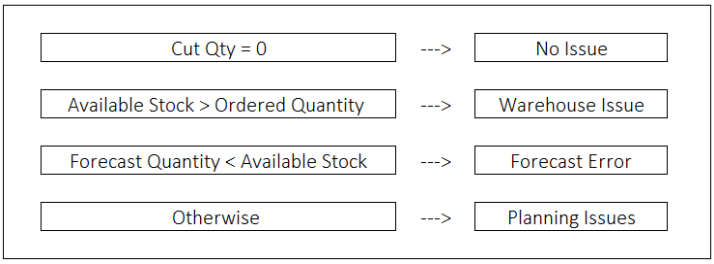
(vii) H_o : The higher the production error, the lower the inventory

(viii) H_o : The higher the demand forecast, the higher the inventory

3.3.3 Feature Permutation

To identify the major driver of a low case fill rate, multiple methods are used for this study. Firstly, a root cause analysis is conducted using a decision tree matrix to systematically investigate the problem and identify its underlying causes as mentioned in Figure 4.

Figure 4
Decision Tree Matrix



Furthermore, statistical methods such as SHAP (Shapley Addictive Explanation) values were utilized to identify the factors strongly associated with low case fill rates and cut quantity. SHAP values provided a measure of feature importance, indicating the extent to which each factor contributed to the overall performance of the system. By incorporating both decision tree analysis and SHAP values, this study aimed to comprehensively identify and understand the drivers of low case fill rates.

3.4 Modelling Technique

We approached CFR prediction by employing various machine learning time series models to identify seasonal and cyclical patterns between different variables and case fill rate. The goal is to provide insights into how these factors will impact future case fill rates. The capstone explores different techniques, including classification, regression, convolutional neural network Long Short-Term Memory (LSTM), and XGBoost, which are explained in detail below.

We used two distinct approaches to predict the case fill rate. The first model employs a hybrid approach, combining classification and regression machine learning techniques to predict cut quantity. Cut quantity is equivalent to unfulfilled demand, leading to a low case fill rate. The target variable is cut quantity.

The second approach involves a dual forecasting method, predicting inventory availability and order quantity on future dates. We utilized advanced machine learning techniques such as XGBoost, CNN LSTM, and multistep LSTM for forecasting. The target variables are inventory availability and forecasted order quantity received.

In this section, we will first discuss various machine learning techniques, followed by the two primary modeling approaches.

3.4.1 Modelling Approach – Hybrid model Cut Quantity Prediction

In our first approach, we used a hybrid model in predicting cut quantity. In our two-step approach, we first utilize classification methods to predict whether there will be a cut on a given day (binary outcome). Then, we apply regression methods to predict the magnitude of the cut quantity for the days identified as having cuts. This allows us to effectively estimate both the occurrence and the magnitude of cut quantities in our case fill rate predictions.

3.4.1.1 Data Preparation and Train-Test Split

Before applying the models, we split the dataset into training and testing sets, with the training set comprising data before January 1, 2022, and the testing set containing data from January 1, 2022, onwards. This separation ensures that the models can be evaluated on unseen data, providing a more accurate assessment of their performance.

3.4.1.2 Classification Models

We begin by employing various classification models to predict the occurrence of cuts, including Random Forest, Logistic Regression, Naïve Bayes, Decision Tree, and K-Nearest Neighbors. Our predictor variables for the classification models are forecast error, production error, projected inventory for 30 days, and forecast order for 30 days. The performance of each classification model is evaluated based on accuracy, precision, recall, specificity, and confusion matrix.

3.4.1.3 Regression Models

Once we have identified the days with cuts using the classification models, we then proceed to predict the magnitude of the cut quantity for these days. For this purpose, we tested several regression models, such as Linear Regression, Ridge Regression, Lasso Regression, Elastic Net, Random Forest Regression, Gradient Boost Regression, and Support Vector Regression. The performance of each regression model is assessed using metrics such as symmetric mean absolute percentage error (sMAPE) and root mean square error (RMSE).

3.4.1.4 Model Evaluation and Selection

Our goal is to identify the most suitable combination of classification and regression models for predicting the occurrence and magnitude of cut quantities, which ultimately determine the case fill rate. We compare the performance of our baseline model (based on forecast error, production error, projected inventory for 30 days, and forecast order for 30 days) with other advanced classification and regression models, selecting the best-performing models based on accuracy, RMSE, and confusion matrix.

3.4.2 Modelling Approach – Forecasting Inventory Availability and Order Quantity

Our second approach focuses on forecasting inventory availability and order quantity on a future date. We employ more advanced machine learning techniques that include XGBoost, CNN-LSTM, and Multistep LSTM models for this purpose. This two-section approach enables us to predict case fill rate by considering both inventory availability and expected order quantity.

3.4.2.1 Data Preparation and Train-Test Split

Similar to the first approach, we split the dataset into training and testing sets into 80:20, the dataset consists of 80 % of data for training set and remaining 20% keep for testing set. This separation ensures that the models can be evaluated on unseen data, providing a more accurate assessment of their performance.

3.4.2.2 Inventory Availability Forecasting Models

We employ advanced machine learning techniques including XGBoost, CNN-LSTM, and Multistep LSTM models to forecast inventory availability on a future date. The predictor variables for these models include historical inventory data, production data, lead and lag features, and other relevant features. The performance of each inventory availability forecasting model is evaluated based on root mean square error (RMSE) and symmetric mean absolute percentage error (sMAPE).

3.4.2.3 Order Quantity Forecasting Models

In parallel with inventory availability forecasting, we utilize the same advanced machine learning techniques (XGBoost, CNN-LSTM, and Multistep LSTM) to forecast order quantity on a future date. The predictor variables for these models include historical order data, projected forecast order and lead and lag features. The performance of each order quantity forecasting model is assessed using two main metrics, root mean square error (RMSE) and symmetric mean absolute percentage error (sMAPE).

3.4.2.4 Model Evaluation and Selection

Our goal is to identify the most suitable combination of inventory availability and order quantity forecasting models for predicting case fill rate. We compare the performance of our advanced forecasting models with baseline models based on historical data and other relevant features, selecting the best-performing models based on RMSE and SMAPE. By combining the predictions

from the selected inventory availability and order quantity forecasting models, we can estimate the case fill rate for the sponsoring company.

3.4 Metrics and Performance Measure

To evaluate the performance of our models in predicting case fill rate, we employ a range of metrics suitable for classification, regression, and other forecasting models. Here, we explain each metric and how it applies to predicting case fill rate.

Confusion Matrix: The confusion matrix is a table that compares the predicted and actual values for each category. It is particularly useful for classification problems, as it helps to identify where the model is making accurate predictions and where it may be misclassifying data. By analyzing the confusion matrix, we can better understand the model's performance in predicting whether there will be a cut on a given day.

Recall: known as sensitivity or true positive rate, measures the proportion of true positive predictions out of all actual positive cases. It assesses the model's ability to identify days with no cuts on a given day. High recall values indicate a lower rate of false negative predictions.

Specificity: known as true negative rate, measures the proportion of true negative predictions out of all actual negative cases. It evaluates the model's ability to correctly identify days with cuts. High specificity values indicate a lower rate of false positive predictions.

Root Mean Squared Error (RMSE): measure the difference between the predicted and actual values. It is a common metric used for regression and forecasting problems, as it indicates how well the model is predicting cut quantity versus the actual value. Calculation for RMSE is shown in Equation 2 below.

Equation 2
Root Mean Squared Error Calculation

$$RMSE = \sqrt{\frac{\sum_{i=1}^{N} \|y(i) - \hat{y}(i)\|^2}{N}},$$

Symmetric Mean Absolute Percentage Error (sMAPE): accounts for the scale of the data and addresses the issue of asymmetry in the percentage errors. Lower sMAPE values indicate better model performance in predicting case fill rate. Calculation for sMAPE is shown in Equation 3 below.

Equation 3
Symmetric Mean Absolute Percentage Error

$$SMAPE = \frac{1}{n} \times \sum \frac{|\ forecast\ value\ - actual\ value\ |}{(|\ actual\ value\ |\ +\ |\ forecast\ value\ |)/2}$$

By employing these metrics, we comprehensively evaluate the performance of our models.

4.0 RESULTS and DISCUSSION

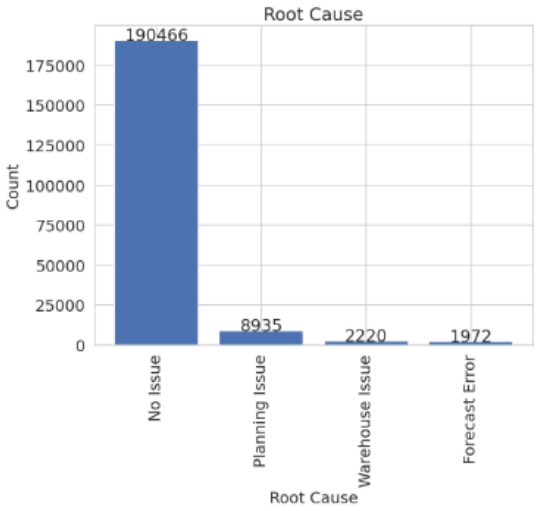
4.1 Decision Tree Matrix Result

To determine the major driver of low CFR, we ran an analysis for orders received through the decision tree matrix. Based on the results obtained from the decision tree matrix shown in Figure 5, it was found that the major driver of low CFR was planning issues, which accounted for 4% of the orders. This category indicates that even though the forecasted quantity was higher than the ordered quantity, there were issues in the production process that prevented the orders from being fulfilled. The largest portion, representing 94% (190,466) of the orders, were fulfilled with no issues.

Next, warehouse issues represented a relatively small portion of the orders and included situations where inventory was available in inventory the system, but orders were unable to be fulfilled. This category likely includes issues related to physical inventory shortages or warehouse staff shortages. To ensure data cleanliness, this category was removed from the analysis going forward to remove noise from the analysis.

Lastly, forecast issues represented the smallest portion of the orders, accounting for less than 1%. This category indicates that the forecasted order was lower compared to the actual order, leading to a shortage of inventory and unfulfilled demand.

Figure 5
Decision Tree Result Matrix



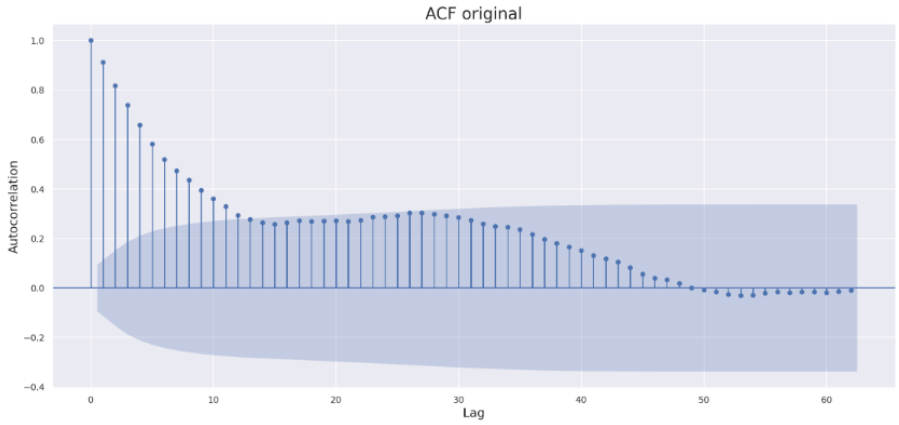
4.2 Descriptive Analytics

We conducted correlation functions of Autocorrelation (ACF) and Partial Autocorrelation (PACF), which are used to identify the presence of autocorrelation in a time series data set. Autocorrelation refers to the correlation between a time series and a lagged version of itself.

The ACF plot shows the correlation between the values of the time series at different lags. The ACF plot in Figure 6 shows that there is significant correlation at lag 1 and a gradual decrease as lag increases, this suggests that the time series may be stationary with an AR (1) structure.

Figure 6

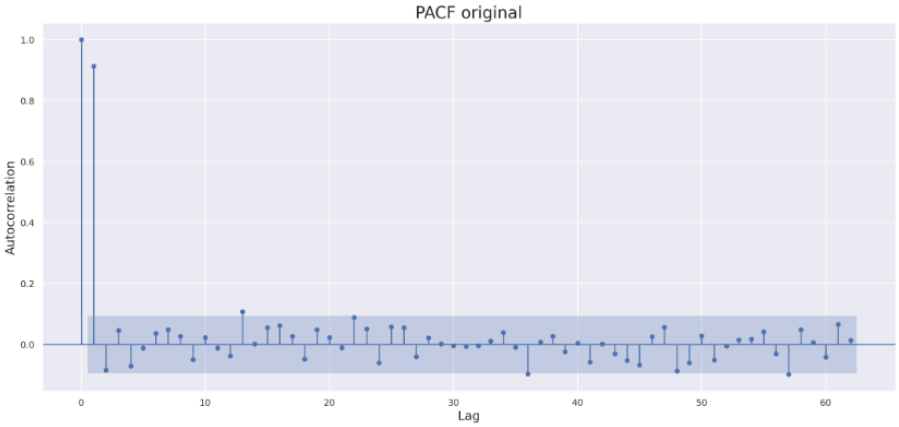
Autocorrelation Function (ACF)



The PACF plot shows the correlation between the values of the time series at different lags after removing the effects of shorter lags. The PACF plot in Figure 7, shows a significant spike at lag 1 and the remaining lags are insignificant, this suggests that the time series may have an MA (1) structure.

Figure 7

Partial Autocorrelation Function (PACF)



4.3 Hypothesis Testing

One of the key questions of our project was to determine which factors influence the Case Fill Rate (CFR) the most. Based on our findings in the literature review and discussions with the industry experts, we formulated the below hypotheses.

(i) H_o : The higher the Coefficient of Variation (CV), the lower the CFR

(ii) H_o : The higher the Stock out Frequency (SOF), the lower the CFR

(iii) H_o : The higher the Coefficient of Variation (CV), the higher the SOF

(iv) H_o : The higher the Forecasting Error (FE), the lower the CFR

(v) H_o : The higher the Days inventory cover (DFC), the higher the CFR

(vi) H_o : The higher the shipments cuts (Cuts), the lower the CFR

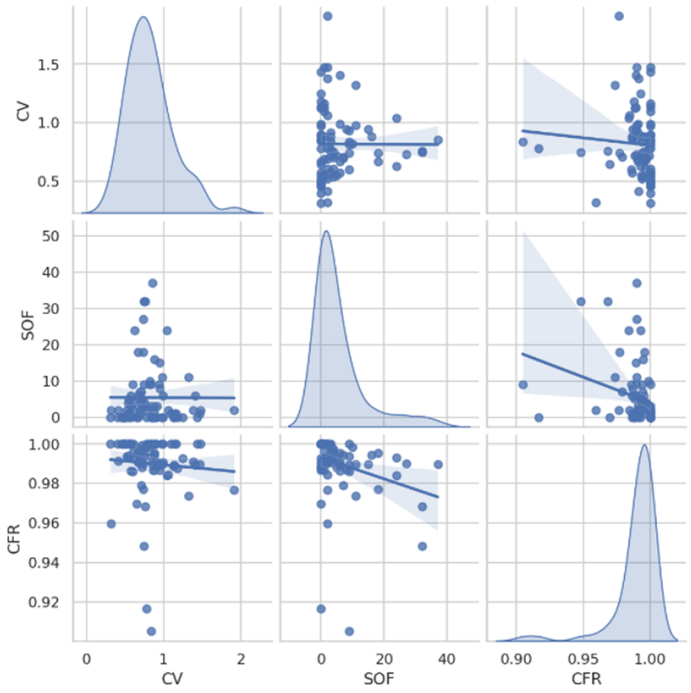
(vii) H_o : The higher the production error, the lower the inventory

(viii) H_o : The higher the demand forecast, the higher the inventory

We tested our hypotheses after organizing the data as explained above. In our first three hypotheses, data was organized by aggregation at the brand level so that we could study the linear relationships amongst our initial key variables: Coefficient of Variation of demand (CV), Case Fill Rate (CFR) and Stock Out Frequency (SOF). These relationships were visualized as seen in Figure 8 by using pair plot to analyze the above hypotheses.

Figure 8

Pairplot for CFR, SOF & CV of Demand



Our first hypothesis is about the relationship between two variables - the coefficient of variation (CV) of demand and the Case Fill Rate (CFR). The CV of demand is a measure of how much variation there is in the demand for a product over a given time period. The CFR is a measure of how well a company is able to meet the demand for its products.

Upon analysis of the data, the hypothesis was only partially supported, while a relationship between the two variables was observed, it was not as strong as expected. Through further analysis and discussions with industry experts, it was concluded that the accuracy of demand

variability forecasting could impact the relationship between CV and CFR. Accurate forecasting could help companies plan their inventory better and avoid stockouts.

Another relationship was also observed between CV and CFR, which showed that as the CV of demand increases, the CFR initially decreases but then starts to improve after reaching a certain point (at CV = 1). In our discussion with experts, we concluded that this could be due to the inventory planning methodology used by the company, which builds up excess inventory to cater to demand variability, thus improving CFR for products with highly variable demand.

This analysis highlights the importance of accurately forecasting demand variability to optimize inventory levels and improve CFR. It also emphasizes that the relationship between CV and CFR is not straightforward and can be influenced by other factors, such as demand forecasting accuracy and inventory policy. Companies can use these insights to better manage their inventory levels and improve their product availability and customer satisfaction.

Our second hypothesis focused on the relationship between two variables: stock out frequency (SOF) and Case Fill Rate (CFR). Stockout, as described earlier, occurs when a company is unable to fulfill an order due to a lack of inventory, and SOF measures how frequently these stockouts occur. The CFR, as mentioned earlier, measures how well a company is able to meet the demand for its products.

To determine the presence and strength of the relationship, we used a method that compared actual shipments against actual orders. Any unfulfilled order due to insufficient inventory was classified as a stockout. Our analysis found a significant correlation between stockouts and CFR, indicating a strong relationship between the two variables. Specifically, we observed that as the company's ability to meet demand decreased, the frequency of stockouts increased. Therefore, maintaining high inventory levels and accurate demand forecasting is critical to avoid stockouts and improve the CFR. Overall, these findings confirm the importance of proper inventory management for companies to optimize their supply chain performance.

Our third hypothesis focused on the relationship between two variables: the coefficient of variation (CV) of demand and stock out frequency (SOF). As mentioned earlier, the CV of demand measures the amount of variation in the demand for a product over a given period of time, while SOF measures how frequently a company experiences stockouts.

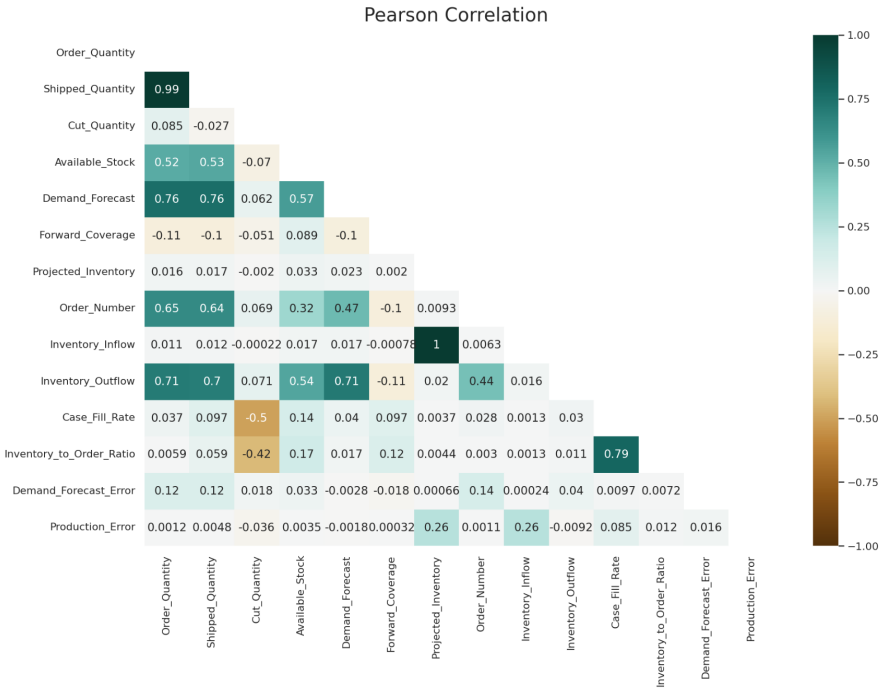
In our analysis of the data, we found this relationship is akin to the one observed in our first hypothesis, where we studied the relationship between the CFR and CV of demand. Initially, as the CV of demand increased, there was a corresponding increase in the SOF, indicating more stockouts. However, this pattern only persisted up to a certain point (at CV = 1), beyond which the SOF improved with further increases in the CV.

We concluded that the inventory levels are closely linked with the variation in demand. As the variation in demand increases, the company can build up excess inventory to cater to this

variability. Hence, when there is a high CV of demand, the company can build up excess inventory to decrease the frequency of stockouts.

Based on our analysis of the outcomes of the above hypotheses and in our discussion with industry experts, we concluded that we needed to explore the impact of forecasts, forecasting errors, production errors, inventory availability and coverage as well to better understand the interdependencies of these variables; this would help up identify the major drivers of stockouts for the company. We further organized the data by including these variables and conducted further hypothesis testing through a Pearson correlation matrix, Figure 9 shows the strength and direction of the linear relationship between pairs of variables in a dataset. Each element in this figure represents the Pearson correlation coefficient (r) between two respective variables, where a value of 1 represents a perfect positive linear correlation, a value of -1 represents a perfect negative linear correlation, and a value of 0 represents no linear correlation.

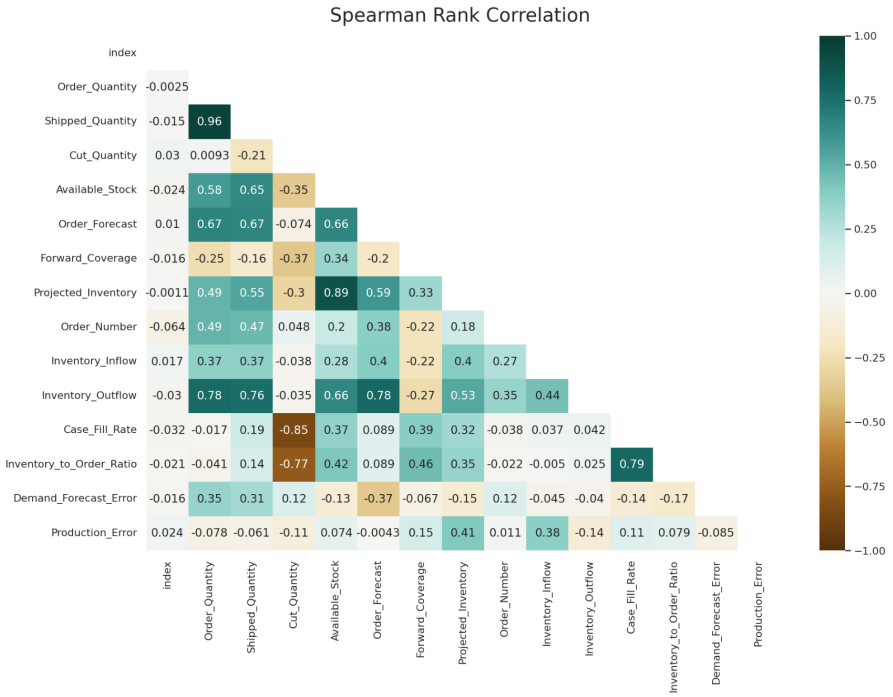
Figure 9
Pearson Correlation Matrix



We analyzed these correlations with the help of industry experts and the sponsor company to determine the relationships between these variables, during these discussions we concluded that all the variables are highly interdependent, and each SKU has different magnitudes in terms of total order quantity per month. As an example, an unfulfilled order of 100 units can be 99% overall fulfillment for one SKU; however, the same magnitude of unfulfilled order might cause the overall fulfillment to drop to 10% based on the total order quantity of that SKU per month. Such a scenario causes distortions when analyzing the linear relationships between any two

variables. To overcome this issue, we opted for the Spearman ranked correlation method as depicted in Figure 10.

Figure 10
Spearman Rank Correlation Matrix



The Spearman correlation coefficient measures the monotonic relationship between two variables. It does not assume that the relationship between the two variables is linear and is less sensitive to outliers. The Spearman correlation coefficient is calculated based on the ranked values of each variable, rather than the actual values. It ranges from -1 to +1, where a value of

+1 indicates a perfect positive correlation, a value of 0 indicates no correlation, and a value of -1 indicates a perfect negative correlation.

The fourth hypothesis focused on the relationship between forecasting error (FE) and Case Fill Rate (CFR). Forecasting error can be defined as the root mean square error (RMSE) of the forecasted demand versus the actual demand. The hypothesis was that as the forecasting error increased, the CFR would decrease.

The analysis of the data showed that the hypothesis was true, meaning that there was a negative correlation between forecasting error and CFR. However, the correlation between the two variables was not as strong as expected, and other factors appeared to be influencing the CFR as well. One possible explanation for this weaker relationship could be that other factors, such as inventory management policy and its ability to capture the demand variance in stocks have a direct relation with CFR; furthermore this can also depend on the lead times from plant to final distribution center.

Our fifth hypothesis focused on the relationship between inventory coverage (measured as Days inventory cover or Days Forward Cover (DFC)) and Case Fill Rate (CFR). Inventory coverage was defined as the number of days of inventory available to cover forecasted demand.

The hypothesis was that the higher the DFC, the higher the CFR. In other words, having more inventory on hand would lead to a higher ability to fulfill customer orders and reduce stockouts.

The analysis of the data found that the hypothesis was true, meaning that there was a positive correlation between DFC and CFR. However, the correlation was not as strong as expected, suggesting that other factors are also influencing CFR.

One possible explanation for this weaker relationship could be that having too much inventory on hand could lead to other issues, such as increased carrying costs or excess waste due to expiration or obsolescence therefore, it's not always feasible to carry high inventories just to cater to demand variability of all the products.

Overall, this hypothesis highlights the importance of managing inventory levels to ensure a balance between inventory coverage and other factors that may impact CFR. It suggests that optimizing inventory levels alone may not be sufficient to improve supply chain performance, and that other factors may need to be considered as well.

Our sixth hypothesis focused on the relationship between shipment cuts and Case Fill Rate (CFR). Shipment cuts were defined as the volume of SKUs that were not fulfilled against the demand, meaning that the company was unable to deliver the products that customers had ordered.

The hypothesis was that the higher the shipment cuts, the lower the CFR would be. This hypothesis assumes that when a company is unable to fulfill customer orders, it will result in lower customer satisfaction and a decrease in the overall CFR. The analysis of the data found that

the hypothesis was true, meaning that there was a negative correlation between shipment cuts and CFR. In other words, as the volume of unfulfilled SKUs increased, the overall CFR decreased.

This finding highlights the importance of monitoring and managing shipment cuts in order to improve supply chain performance and maintain a high CFR. If a company experiences high levels of unfulfilled demand, it may need to reevaluate its inventory levels, production capacity, or logistics processes to identify and address the root causes of the problem.

Our seventh hypothesis focused on the relationship between inventory and production error to see study the impact of production compliance on the inventory levels. Production errors for our hypothesis can be defined as the compliance of the production plan, i.e. if the plant was able to meet the production plan. Our hypothesis was that the higher the production errors, the lower the inventory.

We observed a negative correlation between inventory and production error in some cases. This can happen when the production error is systematic and occurs consistently in the same direction. For example, if a company consistently produces more than the demand forecast, it will result in excess inventory. On the other hand, if the company consistently produces less than the demand forecast, it will result in stockouts and lower inventory levels. In both cases, there will be a negative correlation between inventory and production error.

However, it is important to note that this negative correlation may not always be present and could depend on various factors such as demand variability, lead times, production processes, and supply chain dynamics. Therefore, it Is crucial to analyze and understand the underlying causes of inventory fluctuations and production errors to make informed decisions about inventory management and production planning.

Our eighth hypothesis focused on the relationship between demand forecast and available stock or inventory level. Our hypothesis was that the higher the forecast, the higher the available stock. We observed a positive correlation between forecast and inventory and found the hypothesis to be true.

When the organization forecasts a higher demand for a product, it tends to produce more units to meet that demand. This increase in production leads to a corresponding increase in inventory levels, which in turn can lead to higher available stock. On the other hand, if the organization underestimates the demand and produces fewer units, it may result in stockouts, which can lead to lost sales and dissatisfied customers.

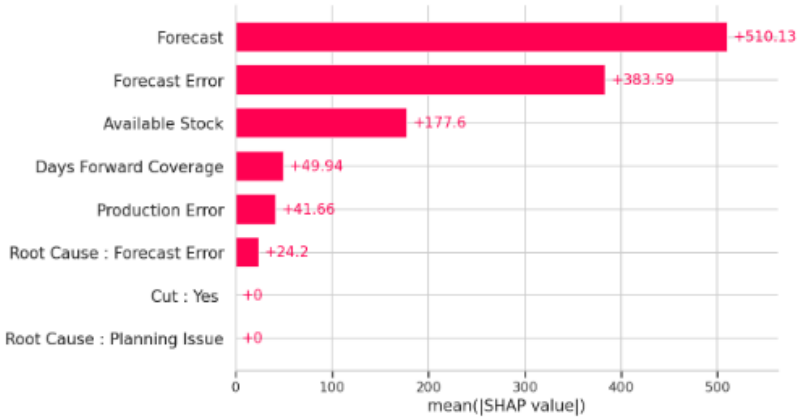
Therefore, the hypothesis that higher forecast leads to higher available stock was based on the idea that accurate forecasting of demand leads to optimized production and inventory management, which in turn leads to better availability of products for customers.

4.4 Feature Permutation

4.4.1 SHAP Value Result

To further investigate the primary causes of low case fill rate, we applied the SHAP (SHapley Additive exPlanations) value technique. This approach quantifies the contribution of each feature in driving low case fill rate, using cut quantity as the target variable. In this analysis, we used cut quantity as a target variable, cut quantity represents unfulfilled portion of an order (total order minus delivered quantity).

Figure 11
SHAP Values (Bar plot)



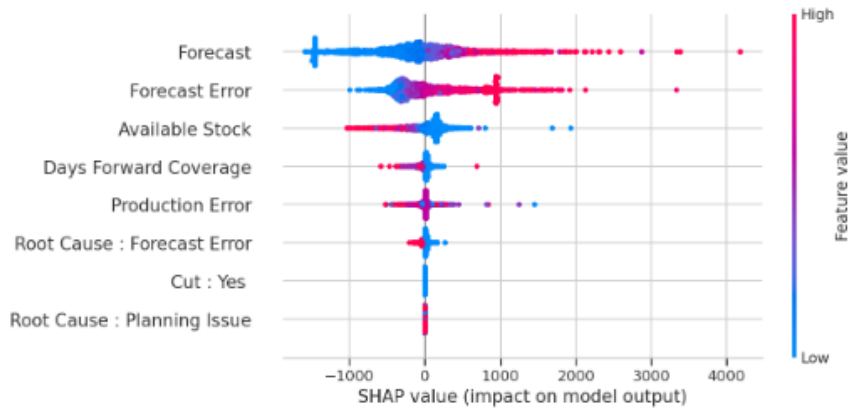
Based on the result from SHAP Value bar plot in Figure 11, features with the highest SHAP value were ‘Forecast ‘and "Forecast Error", which indicates that discrepancies between the forecasted quantity and the actual quantity ordered by customers had a significant impact on the final prediction of "Cut Quantity". A high SHAP value suggests that both forecast and forecast error are the major driver of high cut quantity that contribute to the low case fill rate. The feature

“Available Stock “also had a high SHAP Value, indicating that the available stock is a major determining factor contributing to cut quantity.

In contrast, although the feature "Available Stock" had a high feature value, it had a relatively low SHAP value for the target variable of "Cut Quantity". This suggests that while the availability of stock is important in the model, it may not have as much of an impact on the final cut quantity compared to the forecasted quantity and forecast errors. However, the results still suggest that available stock is a major driver of low case fill rate, indicating that it is critical to plan for adequate stock to meet customer demands and prevent stockouts.

We employed a beeswarm plot (Figure 12) to visualize the distribution of SHAP values for each feature in the model to make prediction using cut quantity as a target predictor. In this analysis, the results showed that forecast, forecast error and available stocks represents the highest feature predicting the target variable, indicating that these features have a high significance to the target variable’s prediction. Forecast error and forecast shows that there is a high feature values with high SHAP values, indicating that the higher the forecast error, the higher the cut quantity leading to a low case fill rate. The result suggests that the discrepancies between forecast quantity and actual quantity ordered by customers are the major drivers driving a low case fill rate.

Figure 12
SHAP Values (Beeswarm plot)



In contrast, the features of available stock have a high feature value but a relatively low SHAP values for target variable, which is cut quantity. This indicates that the lower the available stock quantity, the higher the cut quantity will be. The result suggests that the available stock is also a major driver of low case fill rate, suggesting that it is critical to plan for adequate stock to meet customer demands to prevent stock out.

4.5 Modelling Result and Validation

4.5.0 Overview of Modelling Result

Building on the insights from feature permutation, we further explore various machine learning methods to predict case fill rate. In this section, we discuss the outcomes derived from multiple machine learning methods using different target variables and features through two distinct approaches.

4.5.1 Classification and Regression Model Result

In this section, we present the performance metrics of our baseline model and the other classification and regression models. We discuss the accuracy and confusion matrix of each model, comparing them to identify the best model of predicting cut quantity.

4.5.2 Baseline Model Performance

The baseline model, which combines Logistic Regression for classification and Random Forest Regression for regression, uses the following features:

- Forecast error moving average of past 7 days.
- Forecast error moving average of past 14 days.
- Projected inventory quantity for next 7 to 30 days
- Forecasted order quantity for the next 7 to 30 days.

The target variable classification model is cut category. If cut category is equals to 1, it represents that there is a cut for the given day, in contrast, 0 represents no cut for the day with ordered quantity fulfilled.

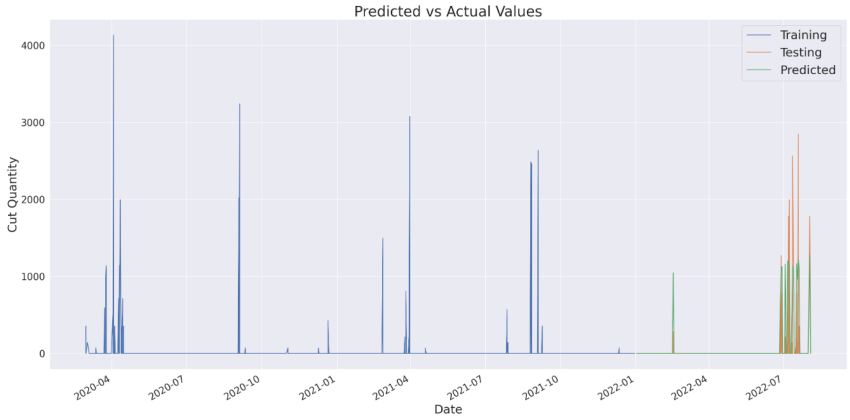
The performance of the baseline model is shown in the below diagram and in Figure 13.

			Baseline Model	
			Predicted	
			TRUE	**FALSE**
Classification	**Actual**	**TRUE**	199	1
		FALSE	12	5
	Accuracy		0.94	

Random Forest	**RMSE**	557

Figure 13

Baseline Classification-Regression Model

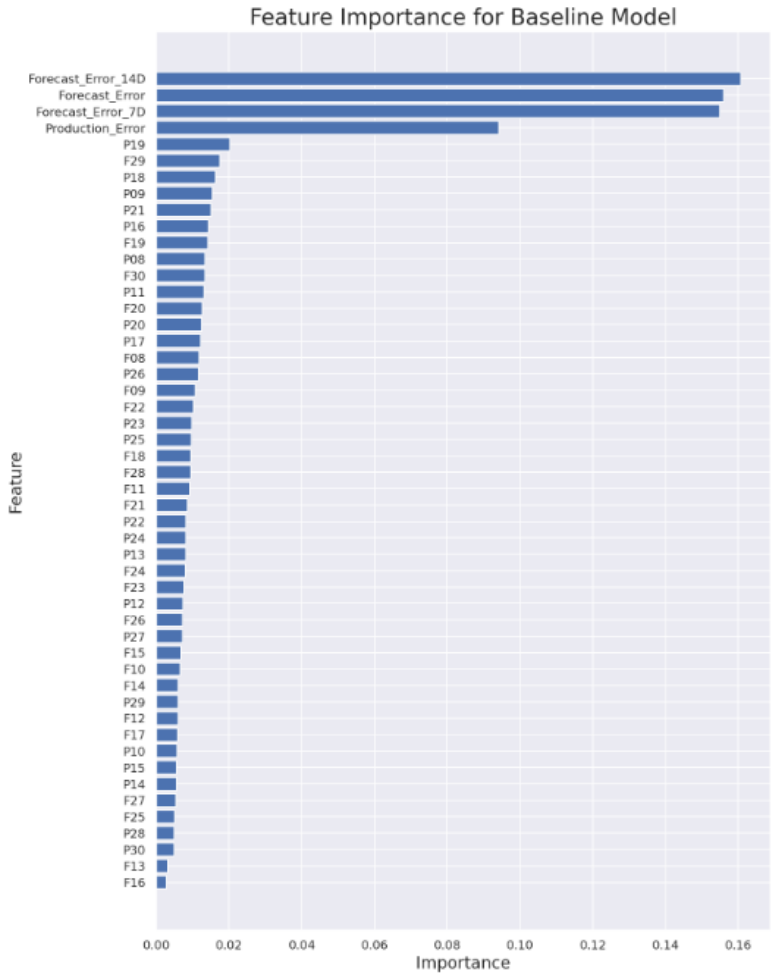


4.5.3 Feature Importance Analysis

To improve our model, we conducted a feature importance analysis to identify the most important predictors. Our results showed that forecast error of the last 7 days and forecast error of the last 14 days are the most significant predictors for the model. Consequently, we excluded the less important features and reran the models using only significant predictors. Figure 14 shows the feature importance from the baseline model based on the highest importance.

Figure 14

Feature Importance from Baseline Model



4.5.4 Discussion of Results

After removing less important features and keeping only [Forecast error 14 days] and [Forecast error 7 days], we reran the model with different classification methods with results shown in diagrams below.

Naïve Bayes	Actual	Positive	195	5
		Negative	12	5
	Accuracy		0.921	
	Precision		0.500	
	Recall		0.294	
	Specificity		0.975	

Naïve Bayes: The Naïve Bayes model showed an accuracy of 0.921 and a precision of 0.500. The model had a higher number of false negatives (12) and false positives (5) compared to the baseline Random Forest model.

Decision Tree	Actual	Positive	197	3
		Negative	15	2
	Accuracy		0.917	
	Precision		0.400	
	Recall		0.118	
	Specificity		0.985	

Decision Tree: The Decision Tree model resulted in an accuracy of 0.917 and a precision of 0.400. This model had a higher number of false negatives (15) and false positives (3) compared to the baseline model, indicating a lower performance in correctly classifying both positive and negative instances.

Support Vector Machine (SVM)	Actual	Positive	200	0
		Negative	12	5
	Accuracy		0.945	
	Precision		1.000	
	Recall		0.294	
	Specificity		1.000	

Support Vector Machine (SVM): The SVM model achieved the highest accuracy among all classifiers (0.945) and a perfect precision score of 1.000. This model showed no false positives and had a lower number of false negatives (12) compared to other models, except for KNN and Logistic Regression.

Logistic Regression	Actual	Positive	200	0
		Negative	14	3
	Accuracy		0.935	
	Precision		1.000	
	Recall		0.176	
	Specificity		1.000	

Logistic Regression: The Logistic Regression model achieved an accuracy of 0.935 and a precision of 1.000. This model showed no false positives and had a higher number of false negatives (14) compared to the Random Forest and SVM models.

Gradient Boosting Classfier	Actual	Positive	198	2
		Negative	15	2
	Accuracy		0.922	
	Precision		0.500	
	Recall		0.118	
	Specificity		0.985	

Gradient Boosting Classifier: The Gradient Boosting Classifier model showed an accuracy of 0.922 and a precision of 0.500. This model had a higher number of false negatives (15) and false positives (2) compared to the baseline model.

Our comparative analysis of different classifiers models revealed that the Support Vector Machine (SVM) model outperformed the other models in terms of accuracy and precision. This indicates that the SVM model is most effective in classifying and predicting cut quantities for case fill rate. However, as case fill rate emphasizes cut quantity (actual negative values), based on recall (sensitivity), the Random Forest and Naïve Bayes models perform best in predicting actual negatives while SVM, KNN, and Logistic Regression models have higher specificity, indicating that they perform better in identifying actual positives.

Considering both recall and specificity, the Random Forest model appears to be a good choice for predicting actual negatives, as it has a relatively high recall and the highest specificity among the top-performing models in terms of recall.

After classifying days that have a cut, we evaluated cut quantity using different regression models to predict the magnitude of cut quantity in each day that cut category is equals to 1 from predicted from the classification model. The models evaluated include Lasso Regression, Random Forest Regression, Support Vector Regression (SVR), Gradient Boosting Regression, and Multi-layer Perceptron (MLP) Regression.

Results evaluated using Root Mean Squared Error (RMSE)

Random Forest (Baseline)	RMSE	557
SVR	RMSE	1023
Gradient Boost Regressor	RMSE	854
MLP Regressor	RMSE	1195
Ridge	RMSE	786
Elastic Net	RMSE	786
Lasso	RMSE	786

The Random Forest model serves as a baseline for comparison, achieving an RMSE of 557. Comparatively, the Lasso, Ridge, and Elastic Net models show similar performance, with RMSE values around 786. These results indicate that these models are less effective than the Random Forest model for predicting cut quantity in this imbalanced dataset.

The SVR, Gradient Boosting Regressor, and MLP Regressor models have varying performance. The SVR and MLP Regressor models have higher RMSE values (1023 and 1195, respectively), indicating they are less accurate compared to other models. The Gradient Boosting Regressor, with an RMSE of 854, is more accurate than the SVR and MLP Regressor models but still less accurate than the Random Forest model.

Although the RMSE results perform well, these values may not provide an accurate representation of the model's performance given the imbalanced nature of the dataset. Most of the data consists of days with no-cut quantity, while cut quantities greater than 1 constitute

only a minor portion of the values. As a result, the imbalanced dataset may result in deceptive RMSE values, as they may not adequately reflect the model's ability to predict the less frequent cut quantities.

4.6 Advance Machine Learning Model Result

4.6.1 Seasonal Naïve Model Result

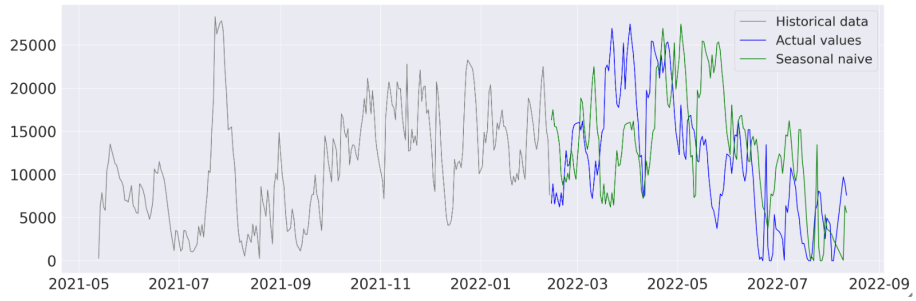
After using the classification regression models, our second approach was to explore advanced time series machine learning models. We were dealing with a dataset from fast moving consumer goods sales, which is usually very seasonal, so the Seasonal Naïve model was a suitable baseline.

In Seasonal Naïve, each forecasted value is based on the actual value in the corresponding time period in the past that has been predefined in the model based on seasonality. These seasonal periods could be a day, week, month, quarter or even a year based on the seasonality of the target variable. Since our data demonstrates multiple seasonal patterns, it was very difficult to specify a particular season. In our model, we chose 1 month (30 days) as our standard time period for seasonality after analyzing the data.

Baseline models, like Seasonal Naïve, are mainly used to check the improvement of the more comprehensive prediction models like XG Boost or neural networks. The outcome of the model can be seen in Figure 15, we split the data into an 80-20 train-test split, historical values during the training period are depicted in gray, actual values during the testing period are shown in blue

and predicted values are shown in the green. It can be seen that the predicted values mimic the actual values and are shifted by 30 days.

Figure 15
Predictions from Seasonal Naïve Model



To analyze the accuracy of models and compare them against one another, we calculated the Root Mean Squared Error (RMSE) and Symmetric Mean Absolute Percentage Error (SMAPE). Below are the accuracy parameters of the seasonal naïve model.

Seasonal Naïve	RMSE	8,108.20
	SMAPE	66.61

4.6.2 XGBoost Model Result

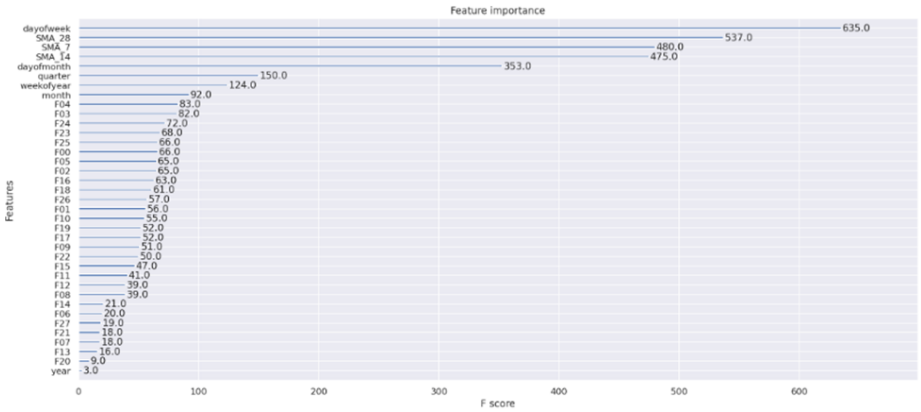
Of the various time series models that we examined, we selected XGBoost as it is particularly effective in dealing with large datasets that have many different features. This is because XGBoost can efficiently process and analyze data that is high-dimensional, resulting in more accurate predictions.

Our first target variable for the model was demand as Case Fill Rate (CFR) is highly dependent on Forecast as well as Inventory on hand. We built in the additional time series features that could help in recognizing the pattern and seasonality in the data such as Day of Week, Day of Month, Week of Year, Month, Quarter, Year, Moving average over 7, 14 and 28 days. In addition to these features, we also incorporated exogenous variables like forecast predictions done by the company's ERP system over the past 30 days that could help with the prediction of the future inventory.

Once the features were built in the model, we ran the XGBoost regressor to find out the importance of these features in the model to fine tune it as shown in Figure 16.

Figure 16

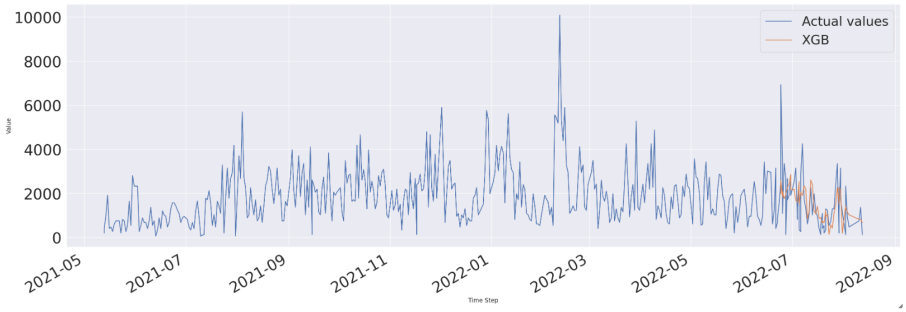
Feature Importance from XGBoost Model



We ran the XGBoost model with all the above-mentioned exogenous variables and got the prediction for demand as shown in Figure 17, we have split the data into an 80-20 train-test split, actual values during the training period as well as the testing period are depicted in blue while the predicted values from the XGBoost model are plotted in orange. We ran the XGBoost model with all the above-mentioned exogenous variables and got the prediction for demand as shown in Figure 17. We split the data into an 80-20 train-test split, actual values during the training

period as well as the testing period are depicted in blue while the predicted values from the XGBoost model are plotted in orange.

Figure 17
Predictions from XGBoost Model



To evaluate the accuracy of this model, we used RMSE and SMAPE which were then compared with the accuracy parameters of the Seasonal Naïve model to see whether the XGBoost would give any improvement.

XGBoost	RMSE	1,022.13
	SMAPE	58.12

Overall, the XGBoost model with both time series and exogenous variables showed better performance in terms of RMSE and SMAPE compared to the Seasonal Naive model. However, the RMSE for the XGBoost model was still higher than it should be, indicating that the model may not be suitable for business needs.

4.6.3 Long-Short-Term-Memory (LSTM) Model Result

A Recurrent neural network (RNN) is a type of neural network designed to have cycles within it. These cycles allow the network to pass information between groups of neurons, known as modules, and in the case of LSTM models, they are called memory blocks. As a result, the network

can retain information, allowing it to be more effective in processing data with temporal dependencies. In contrast to other types of neural networks, where inputs are independent of each other, RNN inputs are related to each other. This makes RNNs particularly useful for analyzing data that has patterns or relationships over time.

Recurrent neural networks (RNNs) are useful for applications that require short-term memory. However, they struggle to recall long-term dependencies because there is a widening gap between relevant information and the necessary application. This is called the long-term dependency problem. Long-Short-Term Memory (LSTM) is an adapted type of RNN designed to address this problem. The standard LSTM has four layers that add and remove information to the recurrent vector between the modules of the layer, making it less likely to suffer from the vanishing gradient problem that can occur when using RNNs. Additionally, the LSTM can learn large weights to prevent the gradient from vanishing.

To implement the LSTM model, the data needs to be in the correct shape. The data can be reshaped into a three-dimensional vector that includes samples, time-steps, and the number of features. One time-step could be equivalent to one week of data instead of daily data. All time-steps make up the sequence length, which is equivalent to several weeks of data.

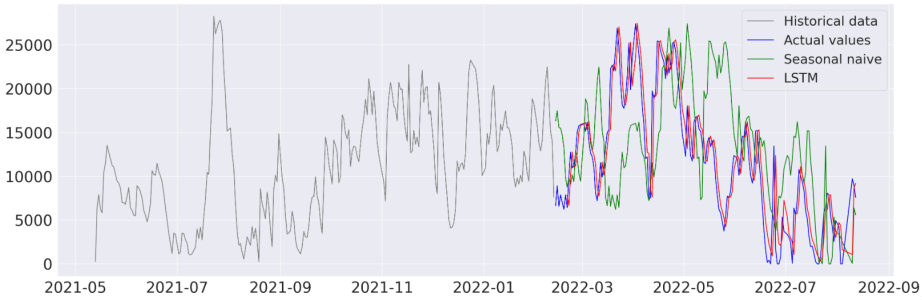
In order to use the data to train and test the LSTM model, the data was prepared according to the guidelines outlined in the Section 3.1 Data Preparation. Additionally, an 80-20 split was defined to create separate training and testing datasets for the model.

Before implementing the LSTM model, hyperparameters were defined. These hyperparameters are values that determine the behavior and performance of the model. By defining these parameters beforehand, they can be easily adjusted to improve the model's accuracy.

To optimize the hyperparameters for the LSTM model, Gridsearch was utilized. Gridsearch is a method for systematically searching through a range of parameter values to find the combination that yields the best results. By using Gridsearch, the optimum parameter values were identified that would allow the model to accurately predict the target variable.

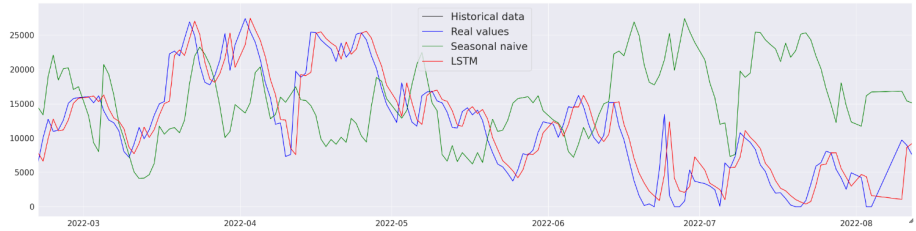
We ran the LSTM model with all the above-mentioned parameters and got the prediction for demand as shown in Figure 18, actual values (historical values) during the training period as are depicted in grey, actual values during testing period are depicted in blue while the predicted values from the LSTM model are plotted in red. We ran the LSTM model with all the above-mentioned parameters and got the prediction for demand as shown in Figure 18. Actual values (historical values) during the training period as are depicted in grey, actual values during testing period are depicted in blue while the predicted values from the LSTM model are plotted in red. The figure also depicts the predictions from Seasonal Naïve for comparison.

Figure 18
Predictions from LSTM Model



The focus of Figure 19 is to showcase the predictions made by the LSTM model during the testing period.

Figure 19
Predictions from LSTM Model with focused test split



To evaluate the accuracy of this model, we used RMSE and SMAPE which were then compared with the accuracy parameters of the Seasonal Naïve & XGBoost model to see if the LSTM would give us any improvement.

LSTM	RMSE	2,807.21
	SMAPE	35.21

From the graph, the model has made highly accurate predictions of the actual values. However, upon a closer examination of the predictions, we noticed that the model is only capable of predicting the values up to one day in advance. This is a problem because we need the model to predict the values for a period of 13 weeks into the future. The analysis shows that the LSTM model was not able to provide accurate predictions for such a long period, which is a significant limitation.

4.6.3 Multi LSTM Model Result

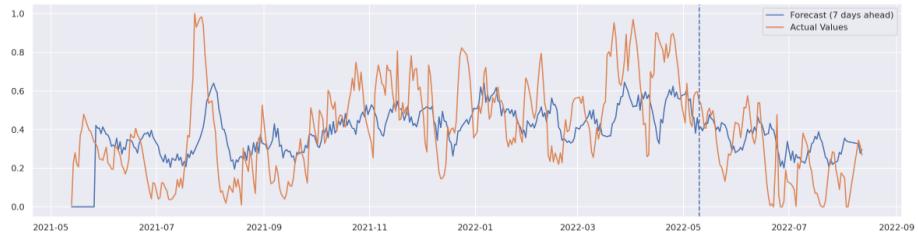
Multi-LSTM (Multiple Layer LSTM) is an extension of the LSTM model that incorporates multiple LSTM layers in its architecture. Similar to LSTM, Multi-LSTM is designed to process data with temporal dependencies by allowing the network to retain information. The key difference is that Multi-LSTM has multiple memory blocks, which enable it to learn more complex patterns and relationships between data points over time.

The data used for Multi-LSTM was organized similarly as in the case of LSTM, however, we built some new features of Multi-LSTM model like “signal”, month of the year, day of the month. "Signal" is a time-series signal generated using different components such as trend, seasonality, noise, and covariates. It is a combination of all these components and represents the overall pattern in the data. In summary, "signal" represents the true underlying pattern in the data that we want to extract and analyze.

As with the LSTM model, hyperparameters had to be defined before implementing Multi-LSTM. These values affect the behavior and accuracy of the model and can be optimized through hyperparameter tuning techniques such as Gridsearch.

Once the hyperparameters were set, Multi-LSTM was trained and tested on the data. The actual values during training and testing periods are depicted in orange, respectively, while the predicted values from Multi-LSTM are plotted in blue in Figure 20. Using Multi-LSTM, we attempted to forecast sales data for a single SKU, but the data proved to be very unpredictable and volatile. We focused on predicting values for 7 days in advance. While the model was able to correctly capture the overall pattern in the data, it struggled to accurately estimate the magnitude of the peaks and troughs in the sales data.

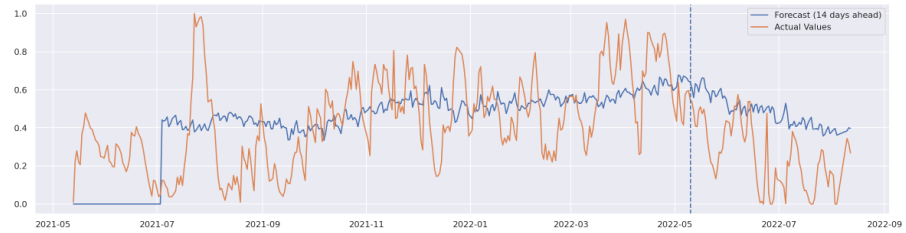
Figure 20
Predictions from Multi-LSTM Model with 7 day forward looking forecast



We then used the same model to forecast the values on a 14-day horizon and the result can be seen in Figure 21. The model further smoothened the predictions and was not able to get the magnitude accurately.

Figure 21

Predictions from Multi-LSTM Model with 14 day forward looking forecast



As the prediction horizon is increased in the Multi-LSTM model, the prediction pattern becomes more normalized, but the model struggles to accurately capture the peaks and troughs in the data. This suggests that the model is better suited for short-term predictions where the data is less erratic and easier to predict.

5.0 CONCLUSION

5.1 Managerial Insights

This capstone project identified the key drivers that influence the Case Fill Rate (CFR) for a sponsor company and developed a model that can predict inventory for 13 weeks into the future. We found several factors that impact CFR and inventory prediction, and we provided recommendations for companies to help improve their forecasting accuracy and inventory availability, which would lead to a better case fill rate.

Forecast error and demand variability are critical factors that impact the Case Fill Rate (CFR) for companies, especially in industries with high demand variability such as Consumer Packaged

Goods (CPG) companies. Therefore, improving forecast accuracy should be a priority to optimize inventory policy and increase CFR. It is crucial to incorporate the impact of promotions and exogenous factors, such as trade and retail promotions, market index and economic indices during forecasting, as these factors can significantly impact sales volumes and inventory levels. By addressing these factors and incorporating forecasting error and demand variation into their inventory policy, companies can optimize inventory levels and meet customer demand effectively.

The second question of our capstone aimed to develop a model that can predict inventory for a period of 13 weeks in advance. However, our analysis revealed that this can be a challenging task due to various factors, as discussed previously. These include the high variability in inventory availability and order demand caused by factors such as promotions and the irregular order patterns typical in B2B businesses. Nonetheless, we suggest that utilizing advanced forecasting techniques, such as machine learning models that incorporate time series and exogenous variables, could help improve prediction accuracy. It is essential to consider the strengths and limitations of such models to ensure they are suitable for specific business needs.

We found machine learning models like LSTM and Multi-LSTM show potential in predicting inventory availability and order demand, particularly for short-term forecasting. However, they struggle to capture the magnitude of peaks and troughs for longer-term forecasts, like the 13-week window required in this study. Therefore, we recommend that companies use these models in combination with other models for more accurate and reliable long-term predictions.

Furthermore, aggregation of data from SKU level to brand level which reduces the variability in the data, and accounting for exogenous variables, such as holidays, that affect customer demand, could improve forecasting accuracy.

Aggregating data from SKU level to the brand level can potentially improve forecast accuracy as a whole and be helpful in certain situations, such as ordering raw materials and production resource planning. However, this may not meet the specific business requirement of the company sponsoring the study since forecasting order demand by brand may not account for the unique demand patterns of individual SKUs. This is particularly crucial as the delivery of a specific SKU may not be substituted by another SKU, making it necessary to forecast by SKU level to ensure optimal inventory levels and meet customer demand.

5.2 Limitations

This capstone presents a few limitations. First, the dataset used in this study covers a three-year period, which may impact the model's ability to effectively learn patterns and generalize to new data points, particularly in the context of machine learning. Furthermore, the absence of certain data points such as promotional data and market indices may hinder the model's ability to learn from the available training data and create accurate predictions. A larger and more comprehensive dataset could lead to improved model performance and more accurate predictions. Besides, the dataset used is also imbalanced in nature, which may have led to biased model performance and less accurate prediction for underrepresented classes. Future research

could explore techniques specifically designed for handling imbalanced data, such as oversampling the minority class, under sampling the majority class, like cut category.

Second, certain products in the market have short life cycles, which means that they are available for a limited period. As a result, there may be insufficient training data for these specific products, which could have hindered the model's ability to accurately predict their demand and associated cut quantities. Future research could explore strategies for handling short life cycle products, such as developing specialized models or leveraging transfer learning techniques.

5.3 Future Research

We recommend that future research explore the use of more advanced machine learning techniques, such as Reinforcement Learning or Deep Reinforcement Learning, to better understand the intricate relationships between features and impact of case fill rate. Additionally, future research could consider incorporating more external data points, including a larger training dataset, promotional activities, market indices, and competitor pricing, to enhance the predictability of cut quantity and its impact on case fill rates.

Market indices offer valuable insights into overall economic conditions and consumer trends, which may directly affect product demand. Including this information in the models can help account for wider market influences when forecasting cut quantities and case fill rates. Similarly, promotional activities can have a significant impact on demand patterns. Incorporating these factors into the models can better capture the effects of sales promotions, discounts, and other

marketing efforts on inventory levels. Competitor pricing data can also be a beneficial input, providing a deeper understanding of the competitive landscape and its influence on customer preferences and purchasing behaviors. By incorporating additional data points, the model can more effectively capture the comprehensive market dynamics that influence order quantity and projected inventory, leading to improved case fill rate projections.

By employing advanced machine learning techniques and integrating more external data points in future research, we believe there is potential to develop more accurate and reliable models for predicting cut quantities and case fill rates. This would enable the creation of more robust and adaptive models that can effectively forecast cut quantities and case fill rates while considering additional data points.

REFERENCES

Alzubaidi, Z. Y. (2020). A Comparative Study on Statistical and Machine Learning Forecasting Methods for an FMCG Company. *Rochester Institute of Technology Scholar Works*, 96.

Bhandalkar, S. (2022). *FMCG Market Expected to Reach $15,361.8 Billion by 2025 | AMR*. https://www.alliedmarketresearch.com/press-release/fmcg-market.html

Calhoun, S. (n.d.). *On-Time, In-Full (OTIF): A Key Supply Chain Metric*. Retrieved December 1, 2022, from https://www.veryableops.com/blog/on-time-in-full-otif

Carvalho, H., Naghshineh, B., Govindan, K., & Cruz-Machado, V. (2022). The resilience of on-time delivery to capacity and material shortages: An empirical investigation in the automotive supply chain. Computers & Industrial Engineering, 171, 108375. https://doi.org/10.1016/j.cie.2022.108375

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

Chao, T.-N., & Izaguirre, F. (2022). Identifying the Root Causes of Stockout Events in e-commerce Using Machine Learning Techniques. *Supply Chain Management Capstone Projects*, 90.

Chase, C. W. (2016). Machine Learning Is Changing Demand Forecasting. *The Journal of Business Forecasting*, *35*(4), 43–45.

Chen, F., Drezner, Z., Ryan, J. K., & Simchi-Levi, D. (2000). Quantifying the Bullwhip Effect in a Simple Supply Chain: The Impact of Forecasting, Lead Times, and Information. *Management Science*, *46*(3), 436–443. https://doi.org/10.1287/mnsc.46.3.436.12069

Delua, J. (2022, November 15). *Supervised vs. Unsupervised Learning: What's the Difference?* https://www.ibm.com/cloud/blog/supervised-vs-unsupervised-learning

Dickson, B. (2020). Machine learning: What's the difference between supervised and unsupervised? *TheNextWeb.Com [BLOG]*. Advanced Technologies & Aerospace Collection. https://www.proquest.com/blogs-podcasts-websites/machine-learning-what-s-difference-between/docview/2407960774/se-2?accountid=12492

Drew Editorial Team. (n.d.). *10 Key Performance Indicators for production management*. Retrieved December 1, 2022, from http://blog.wearedrew.co/en/10-key-performance-indicators-for-production-management

EKN Research. (2016). *EKN Research: Plugging Out-of-Stock Gaps in Consumer Goods*. RIS News. https://risnews.com/ekn-research-plugging-out-stock-gaps-consumer-goods

Gruen, T. W. (2008). *A Comprehensive Guide To Retail Out-of-Stock Reduction In the Fast-Moving Consumer Goods Industry*. https://www.nacds.org/pdfs/membership/out_of_stock.pdf

Gundogdu, B. (n.d.). Comparison and Financial Assessment of Demand Forecasting Methodologies for Seasonal CPGs. *Supply Chain Management Capstone Projects*, 66.

Henry, J. (Director). (2019). *Data Analytics and Machine Learning Fundamentals LiveLessons Video Training* (1st edition). Addison-Wesley Professional.

Invent Analytics*, How to Measure Demand Forecast Accuracy*. (n.d.). Retrieved December 1, 2022, from https://www.inventanalytics.com/blog/how-to-measure-demand-forecast-accuracy/

Inderfurth, K. (1991). Safety stock optimization in multi-stage inventory systems. *International Journal of Production Economics*, *24*(1), Article 1. https://doi.org/10.1016/0925-5273(91)90157-O

Infosys BPM, I. B. (2022). *Big Data Analytics in CPG: Insights Into Its Benefits| Infosys BPM*. https://www.infosysbpm.com/blogs/retail-cpg-logistics/why-big-data-and-analytics-is-a-must-for-profitable-growth-in-cpg.html

ITC Infotech. (2022, November 25). *Inventory Management and Optimization for an FMCG Manufacturing Company*. https://www.anylogic.com/resources/case-studies/inventory-management-and-optimization-for-an-fmcg-manufacturing-company/

Kiwop. (2015, September 22). *How Safety Stock Absorbs Demand Volatility*. ToolsGroup. https://www.toolsgroup.com/blog/how-safety-stock-absorbs-demand-volatility/

Kamath, & Liu, J. (2021). Explainable artificial intelligence: an introduction to interpretable machine learning. Springer International Publishing.

Lohman, C., Fortuin, L., & Wouters, M. (2004). Designing a performance measurement system: A case study. *European Journal of Operational Research*, *156*(2), Article 2. https://doi.org/10.1016/S0377-2217(02)00918-9

Mamakos, A. (2022). Spare Parts Predictive Analytics for Telecommunications Company. *Supply Chain Management Capstone Projects*, 66.

Manyika, J., Chui, M., & Brown, B. (2011). *Big data: The next frontier for innovation, competition, and productivity | McKinsey*. https://www.mckinsey.com/capabilities/mckinsey-digital/our-insights/big-data-the-next-frontier-for-innovation

NielsenIQ, N. (2022). Can the FMCG industry afford to lose billions from empty shelves? *NielsenIQ*. https://nielseniq.com/global/en/insights/education/2022/can-the-fmcg-industry-afford-to-lose-billions-from-empty-shelves/

Nigam, A. (2016). Product Promotion Effectiveness: Root Causes of Stock-Outs By. *Supply Chain Management Capstone Projects*, 59.

Nikolopoulos, K. I., Babai, M. Z., & Bozos, K. (2016). Forecasting supply chain sporadic demand with nearest neighbor approaches. *International Journal of Production Economics*, *177*, 139–148. https://doi.org/10.1016/j.ijpe.2016.04.013

PII*: S0927-0507(05)80035-0 | Elsevier Enhanced Reader*. (n.d.). https://doi.org/10.1016/S0927-0507(05)80035-0

Raman, A., & Kim, B. (2002). Quantifying the impact of inventory holding cost and reactive capacity on an apparel manufacturer's profitability. *Production and Operations Management*, *11*(3), 358–373. https://doi.org/10.1111/j.1937-5956.2002.tb00191.x

ASCM*, SCOR Model*. (2022, December 1). https://scor.ascm.org/practices/best%20practices

Sokol, O., Holý, V., & Cipra, T. (2021). Customer and Product Clustering in Retail Business. In S. N. Shahbazova, J. Kacprzyk, V. E. Balas, & V. Kreinovich (Eds.), *Recent Developments and the New Direction in Soft-Computing Foundations and Applications: Selected Papers from the 7th World Conference on Soft Computing, May 29–31, 2018, Baku, Azerbaijan* (pp. 529–537). Springer International Publishing. https://doi.org/10.1007/978-3-030-47124-8_43

6.0 APPENDIX

6.1 Statistical Time Series Model

Statistical models use a set of statistical assumptions to capture temporal dependencies between datapoints. They are generally simpler and more interpretable.

Seasonal Naïve

A seasonal naive forecasting model is a simple time series forecasting method that assumes the future value of a variable will be equal to its value from the same season in the previous year. In other words, the model assumes that there will be no growth or decline in sales or demand patterns over time.

In the FMCG (Fast Moving Consumer Goods) industry, where sales tend to be highly seasonal and cyclical, the seasonal naive forecasting model can be a useful tool for short-term forecasting. It is often used as a benchmark against which more sophisticated forecasting models are compared.

Moving Average

The Moving Average model is a simple statistical technique that calculates the average of a time series over a given period. It smoothens out the random variations and highlights the underlying trend in the data. The moving average model is useful when there is a steady pattern in the data, and it is easy to compute and interpret.

For the FMCG industry, moving average models can be used to forecast demand for fast-moving products that have a predictable pattern. For example, if a product has a seasonal pattern with sales increasing during the holiday season, a moving average model can be used to predict future sales based on the past seasonal data.

ARIMA

Auto Regressive Integrated Moving Average (ARIMA) combines three main components: Autoregression, Differencing, and Moving Average. Autoregression explains the correlation between current value and previous value, differencing represents the differencing that is needed to be applied to the time series to make it stationary, and moving average captures the correlation between error terms and residual. ARIMA model is also used for non-seasonal series of numbers that exhibits patterns and not a series of random event.

SARIMAX

Seasonal Autoregressive Integrated Moving Average with Exogenous (SARIMAX) incorporates external variables and seasonal trends, including temperature or rain volume as an extension to ARIMA model to improve forecasting. SARIMAX captures both seasonal patterns from historical data and relationship between timer series and external variables. SARIMAX determines the presence of seasonality and trends in time series data, and later uses the logic of ARIMA's three main components: autoregression, differencing, and moving average while capturing external factors to improve forecasting CFR.

Vector Autoregression (VAR)

VAR uses multiple time series variables for forecasting by analyzing the interrelationship between multiple variables, such as demand, inventory, and lead time to predict the impact. VAR incorporates lagged variables to capture the dynamic interaction between multiple variables to predict future CFR, as well as causal relationship between multiple variables.

6.2 Machine Learning Time Series Model

XGBoost

Extreme Gradient Boosting (XGBoost) is a powerful gradient boosting algorithm used for regression, classification, and ranking tasks. It is a highly accurate and efficient algorithm that can handle large datasets with high-dimensional features and is known for its speed, accuracy, and scalability.

The main idea behind XGBoost is to use a gradient descent algorithm to iteratively add decision trees to a model. At each iteration, the algorithm calculates the gradient of the loss function and uses this gradient to update the model's parameters. This approach ensures that subsequent trees are built on the errors of the previous ones, gradually improving the overall model's accuracy.

In the context of forecasting demand and stock outs for the Fast-Moving Consumer Goods industry, XGBoost can be used to predict demand patterns and identify factors that contribute to stock outs. By training a model on historical sales data, XGBoost can identify key drivers of

demand and predict future demand trends with a high degree of accuracy. This information can be used to optimize inventory levels and minimize stock outs.

Moreover, the algorithm can be used for feature selection, which is useful in identifying the most relevant factors that influence demand and stock outs. For example, XGBoost can be used to identify which products are most sensitive to demand variability, forecasting errors, changes in pricing, seasonality, or promotional campaigns. This information can be used to design more effective marketing strategies, adjust pricing and inventory build up to optimize revenue and minimize stock outs.

Neural Networks

Neural networks are a type of machine learning algorithm modeled on the structure and function of the human brain. They consist of interconnected nodes, or neurons, arranged in layers. Each neuron receives input from multiple neurons in the previous layer, processes that input, and passes output to neurons in the next layer. The output of the final layer is the network's prediction or classification. Neural networks are highly adaptable and can learn complex patterns in the data, making them suitable for a wide range of applications.

In the context of forecasting for FMCG products, neural networks can be trained on historical sales data to learn patterns and trends in demand. The network can then be used to make predictions about future demand, considering factors such as seasonality, promotions, and other external factors that may influence sales. Neural networks can be particularly useful for FMCG

forecasting because they are able to capture complex, nonlinear relationships between variables that traditional statistical methods may not be able to identify.

There are several types of neural networks, including feedforward networks, recurrent networks, and convolutional networks, each with its own strengths and weaknesses. Recurrent neural networks (RNNs) are particularly well-suited to time series forecasting tasks, as they can process sequences of input data and maintain a memory of past inputs. Long short-term memory (LSTM) networks, a type of RNN, are commonly used for time series forecasting tasks in the FMCG industry.

Long Short-Term Memory (LSTM)

LSTM stands for Long Short-Term Memory and is a type of recurrent neural network that is designed to remember and process long-term dependencies. LSTM networks are commonly used in time series forecasting because they can learn patterns in sequential data and can make accurate predictions based on those patterns.

In the context of forecasting demand for FMCG products, LSTM can be used to analyze historical sales data and identify patterns and trends in the data. The model is trained on a dataset of historical sales data and is then able to predict future sales based on those patterns. LSTM models can be used to predict demand and stockouts for products based on various factors such as demand variability, forecasts, forecasting errors, production errors, seasonal patterns, marketing campaigns, and external factors such as weather and holidays.

LSTM models have several advantages over traditional time series forecasting models like ARIMA and exponential smoothing as they can capture long-term dependencies in the data, which makes them more accurate at predicting future sales. They are also able to handle non-linear relationships in the data, which makes them more flexible than traditional models. Moreover, they can be used to forecast demand for multiple products simultaneously. This is important in the FMCG industry, where companies may have hundreds or thousands of products that need to be forecasted.

However, one of the challenges of using LSTM models is that they can be difficult to train and optimize. They require a large amount of training data and can be computationally expensive to train. Additionally, the model architecture needs to be carefully designed and tuned to achieve the best performance.

Neural Prophet

Neural Prophet is a machine learning library that is commonly used in time series forecasting. The library is developed by Facebook which incorporates the flexibility and power of neural networks (Kamath & Liu, 2021), using feedforward neural network with multiple layers to learn pattern in dataset to increase accuracy of forecast predictions. One of the advantages of Neural Prophet is its ability to handle multiple seasonality patterns, which is often observed in our sponsoring company's data. For instance, CFR may exhibit weekly, monthly, and yearly seasonality patterns. Neural Prophet can capture and incorporate these patterns into its model

to improve its forecasting accuracy. Additionally, Neural Prophet can handle missing values, making it suitable for CFR data that may contain missing or incomplete observations.