Demand Forecasting of Consumer Goods for the Indian Subcontinent

by

Lanyan Feng
Bachelor of Science in Supply Chain Management and Entrepreneurship, Syracuse University, 2016

and

Kristen Foster
Bachelor of Commerce in International Management, McGill University, 2014

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

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

June 2023

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

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

Certified by: _____
Dr. Ilya Jackson
Postdoctoral Associate
Capstone Advisor

Accepted by: _____
Prof. Yossi Sheffi
Director, Center for Transportation and Logistics
Elisha Gray II Professor of Engineering Systems
Professor, Civil and Environmental Engineering

Demand Forecasting of Consumer Goods for the Indian Subcontinent

by

Lanyan Feng

and

Kristen Foster

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 in Supply Chain Management

ABSTRACT

Our sponsor company, a leading fast-moving consumer goods corporation in the Indian Subcontinent (ISC), faces ongoing challenges in forecasting the volatile demand in this region. Currently, company's demand planning relies heavily on human judgment, resulting in persistent inaccuracies and biases. This capstone develops a standardized and quantitative demand forecast methodology through three explorations. First, we build a variety of time series forecasting models. Second, we include exogenous variables relevant for Indian demand, such as rainfall and Consumer Price Index (CPI). Third, we explore ways to mitigate the impact of data outliers during the COVID-19 pandemic. These enhancements reduce the average mean absolute percent error (MAPE) of demand forecast across all product categories to 8.0%. We consolidate our analysis and model selection algorithms into an application named *The Demand Forecaster.* This powerful tool not only enables the sponsor company to retrieve our existing forecast results for all product categories, but also allows them to perform forecasts in the future with updated demand data and model specifications. By adopting our demand planning methodology, the sponsor company is expected to increase margins with better inventory management and production planning.

Capstone Advisor: Dr. Ilya Jackson
Title: Postdoctoral Associate

ACKNOWLEDGMENTS

I would like to express my deepest gratitude to Dr. Ilya Jackson, our capstone advisor. His steadfast support throughout this project was invaluable. His office was always open for us, offering a space where we could ask questions and seek guidance directly. Next, I want to extend my sincere thanks to our sponsor company for providing the opportunity to work on this remarkable capstone project. Special thanks are due to my capstone partner, Kristen, whose companionship and commitment were a constant source of inspiration. I also wish to acknowledge my family's unwavering support. Without them, I would not be where I am today. Their belief in me has been my strength, and for that, I am eternally grateful. To my boyfriend, Shushen Hou, words cannot express how grateful I am. His love, support, and encouragement have not only helped me become the person I am today but have also been my guiding light. Each one of them has played a critical role in the success of this capstone project, and for that, I am profoundly grateful. Their contributions have shaped not only this thesis but my entire academic journey. It's their continued support that has made this journey worthwhile.

Lanyan Feng

I extend my heartfelt appreciation to our advisor, Dr. Ilya Jackson, for his unwavering encouragement, guiding us in the right direction, and ensuring that our project remained focused within its intended scope. I would like to thank the sponsor company for providing us with invaluable access to high-quality data, which laid a solid foundation for our research and allowed us to hit the ground running. Special thanks go to Toby and Pamela for their meticulous attention to detail and constant encouragement throughout the writing process. I express my gratitude to my capstone partner, Lanyan, for her instrumental role in driving us to start strong and for her creative contributions to the development of our application. I would also like to acknowledge the flexibility shown by my company, One Acre Fund, in granting me a sabbatical to pursue this research endeavor. Lastly, I am deeply thankful to my partner for his patience and for sparking my initial interest in Python programming.

Kristen Foster

TABLE OF CONTENTS

LIST OF ABBREVIATIONS

AIC	Akaike Information Criterion
ARIMA	Autoregressive Integrated Moving Average
ARIMAX	Autoregressive Integrated Moving Average with eXogenous Variables
CPI	Consumer Price Index
FMCG	Fast Moving Consumer Goods
GST	Goods & Services Tax
ISC	Indian Subcontinent
MA	Moving Average
MAPE	Mean Average Percent Error
SARIMA	Seasonal Autoregressive Integrated Moving Average
SARIMAX	Seasonal Autoregressive Integrated Moving Average with eXogenous Variables
SKU	Stock Keeping Unit
XGBoost	eXtreme Gradient Boosting

LIST OF EQUATIONS

LIST OF TABLES

LIST OF FIGURES

1. INTRODUCTION

The sponsor company is a multinational corporation that specializes in fast moving consumer goods (FMCG). The focus of this capstone is the company's operations in the Indian subcontinent (ISC), which refers to India, Pakistan, Bangladesh, Bhutan, Sri Lanka, and the Maldives. With a population of over 1.7 billion (Kumar, 2019), the ISC is the largest such entity in the world, making it a critical market for the sponsor company's growth strategy. Leveraging its extensive experience and expertise, the sponsor company has emerged as one of the largest and fastest-growing FMCG companies in the region, cementing its position as a key player in the ISC market.

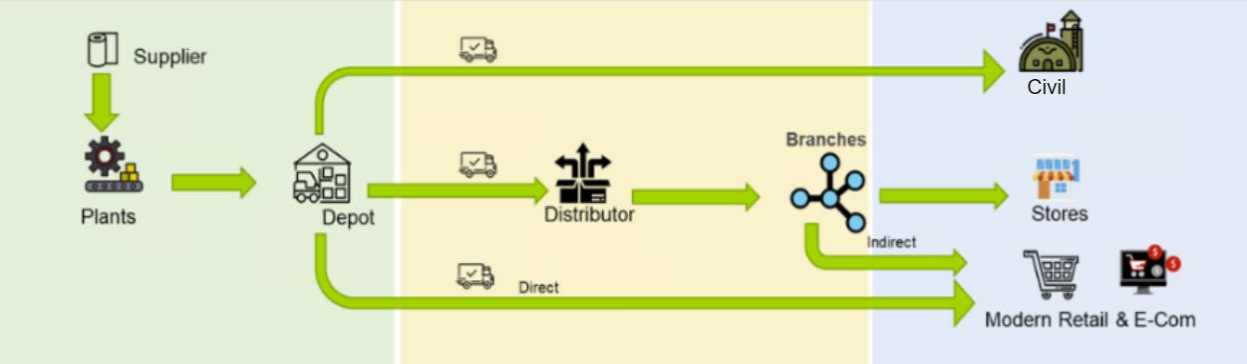
India has been a significant contributor to global industrial development for several decades, exhibiting economic growth potential comparable to that of China, and the FMCG industry is a prime example of India's consistent growth (Mahajan, 2020; Hawksworth, 2006). In fact, FMCG is the 4th largest sector in the Indian economy (India Brand Equity Foundation, 2022), mainly due to company engagement with the bottom of the pyramid – the largest, but poorest socio-economic group – and a fast growing middle class (Mishra, 2008). Various sources estimate that lower-income markets spend between 42% and 56% of their income on FMCGs (Karn, et al., 2003; Banerjee & Duflo, 2007; Mall, et al., 2012).

To serve hundreds of millions of customers in the ISC, the sponsor company operates plants and contract manufacturing sites, producing nearly 1,000 stock keeping units (SKU). As explained in Figure 1, products are delivered from the sponsor company's depot to distributors, who then ship to distributor branches throughout the region, eventually arriving at over 2 million

retail stores in the ISC. Roughly 90% of the sponsor company's demand in the ISC is through a distribution network to retailers, which is the focus of this capstone project.

Figure 1

ISC Supply Network



Note: Adapted from Sponsor Company personal communication, 2023.

The ISC is one of the most complex geographies in which the sponsor company operates. The complexity of the ISC stems from multiple sources. Consumer demographics are highly diverse across regions and quickly evolving. Weather can be volatile and extreme during monsoon season, significantly disrupting access to supply chain channels and products. As FMCG is a highly competitive market, opaque competitor trends such as pricing, new products, or marketing investments also greatly affect demand for the sponsor company's products. Finally, each individual distributor's demand for the sponsor company's products is driven by its cash flow and the demand of its own customers.

Due to the substantial business scale and the myriad of complicating factors mentioned earlier, the sponsor company experiences high volatility in sales within the ISC. Improved forecasts would significantly contribute to increased margins by enabling more efficient

management of raw material suppliers, manufacturing lines, and warehousing contracts. Given the sponsor company's extensive scale and the low-margin nature of the FMCG industry, even a modest 1% improvement in forecasts could result in substantial cost savings.

At present, the ISC demand forecasting model is acknowledged by the management to have four major deficiencies. First, efforts have been focused on shorter term demand forecasts, from one- to four-week span. While the demand planning team observes the presence of demand growth and seasonality in the longer term, they lack a standardized method for forecasting in this more strategic horizon. Second, the shorter-term model currently utilizes only internal historical variables to predict future demand, despite unanimous assertions from the demand planning team that demand is also heavily influenced by economic factors and competition trends. Third, the construction of the current forecasting model is too reliant on human judgments, with model specifications determined by planners' subjective understanding of the current business plans. Lastly, COVID-19 added an unprecedented variability in long-term demand trends that requires special treatment. As a result of these four drawbacks, the current approach consistently exhibits a bias toward overestimating demand and fails to capture crucial trends, resulting in inadequate accuracy.

Due to the significant intra-firm heterogeneities observed in the sponsor company's diverse range of products and operations, it was imperative to design forecasting models that can accurately account for these variations. Therefore, we developed product category-specific forecasting models, allowing us to tailor the models to each category's unique characteristics and test various explanatory variables and model specifications. As a result, we were able to

determine the most suitable variables and models for each category, thus increasing the forecast accuracy per category.

In light of the sponsor company's present challenges and concerns, the ultimate goal of this capstone is to find the most accurate demand forecasting model for the six-month term at the product category level. The secondary goal is to identify exogenous variables that both increase the accuracy of the forecasting model and are accessible to the company in the future.

First, we conducted a literature review to better understand the FMCG market, promising exogenous variables, and the various forecasting models available.

2. STATE OF THE ART

To achieve the ultimate goal of this capstone, which is to enhance the accuracy of category-level demand forecasting in the six-month term using exogenous and internal variables, we conduct literature review in the following areas: (1) market dynamics of the FMCG industry in the ISC and similar geographies; (2) application of exogenous variables to demand forecasting; and (3) demand forecasting models and algorithms.

2.1 Fast Moving Consumer Goods Market Dynamics

Fast moving consumer goods (FMCG) are products that are regularly purchased from local retailers, demanded by a broad variety of consumers seeking convenience, and characterized by volatile demand and low consumer loyalty (Liu, 2013). These goods exhibit relatively low costs and high turnover (Keller, 2008). Therefore, while the margin on an individual product is usually relatively small, FMCG companies can generate significant profits due to the large quantities they

sell (Eisenhardt, 2007; Sean, 2002). This high turnover can be further explained in different ways based on the type of product. Processed food products are perishable and have an expiration date that limits their time on the shelf (Ramanuj, 2007). Fast fashion and electronics quickly become obsolete (Cohen, et al., 2017). Other products have many competing alternatives and must regularly be repurchased (Tarallo, et al., 2019). This final type is where the sponsor company's product line in the ISC falls, which includes a wide range of products.

As previously discussed, India, and particularly the FMCG industry in India, has been a major source of industrial development by engaging with the bottom of the pyramid and a fast growing middle class. These new markets engage with a wide variety of distribution channels, such as street-side vendors, hawkers, and roughly 12 million unregulated kiranas (neighborhood mom-and-pop stores), that tend to be fragmented and experience a wide range of evolving regulatory challenges for manufacturers and retailers (Mishra, 2008). FMCG companies, however, have had relative success in reaching these new rural and lower-income markets by introducing smaller packages for products (Mahajan, 2020).

The struggle for the FMCG sector in the ISC began in the late 2010s, before the COVID-19 pandemic. Global events like the US-China trade war and Brexit impacted both international trade and growth in the ISC. Cash flow stress in the market, reduced job generation, and stagnant salaries led to reduced household consumption and reduced demand for FMCG products (Shetty, et al., 2020). This was exacerbated by extreme weather in rural areas in the form of floods and droughts, which led to food inflation, putting further strain on already tight household budgets (Subramanian & Feldman, 2019).

With the onset of the COVID-19 pandemic in early 2020, the FMCG market was weakened even more drastically. FMCG industry sales fell 34% in April 2020 from the previous year. Mahajan (2020)'s study of demand trends in specific product categories within FMCG reveals a more intriguing phenomenon. Products, such as those in Category 3, 5, 6, and 10, saw steep declines in demand and were slow to recover for months after lockdowns because of consumers avoiding public events for safety. Products that could be considered in Category 2, 4, and 8 saw an initial decline in demand, likely as a consequence of budget constraints, but did return to normalcy. Products one would see in Category 7 saw sustained increases in demand, even after the panic buying phase. In addition to category-level trends, geography trends were also observed. Due to migrant worker populations relocating to rural homes, rural demand increased suddenly. These positive demand trends would continue as long as companies could meet the supply (Mahajan, 2020).

FMCG companies have struggled to supply the market demand throughout the many phases of the COVID-19 pandemic. Sourcing raw materials, manufacturing, last-mile distribution, and retail were all affected by lockdowns, travel restrictions, volatile demand, and liquidity challenges. On the manufacturing side, many sectors were limited to manufacture and distribute only necessary products and are still producing at suboptimal levels (Shetty, et al., 2020). On the retail side, it is claimed that more than 600,000 kirana outlets might have shut down during lockdowns because of liquidity challenges or the owners returning to their villages. While many retail closures were more short term, these smaller kirana outlets have a harder time bouncing back (Mahajan, 2020).

Looking into the future, a key trend affecting distribution and consumer behavior for FMCGs in the ISC is the rise in e-commerce. While COVID-19 has accelerated the pace at which India adopts digital means of buying FMCG products, the trend has been upward for a while now. The skyrocketing smartphone penetration in the region, combined with some of the cheapest data rates in the world, has helped bring the majority of the Indian population online. By 2030, e-commerce is expected to contribute about 11% of FMCG sales (Mahajan, 2020).

The FMCG sector has been extensively researched in the literature due to its wide range of products and substantial revenue. However, the existing literature on forecasting for this sector predominantly relies on historical sales data as the sole input in time series models. In Section 2.2, we explore commonly employed exogenous variables that help enhance forecasting models.

2.2 Exogenous Variables

The majority of demand forecast literature relies on historical sales data. Some authors, however, emphasize the importance of using multiple variables to forecast demand, such as the product's attributes and the economic environment (Guo, et al., 2013). There are a few frameworks that can help categorize and understand exogenous variables affecting demand in any industry. According to Kotler & Armstrong (2013), one needs to look at the marketing environment, which includes product, price, place, and promotion, as well as the macroenvironment, which includes economic, natural, demographic, technological, political and cultural forces.

Based on the literature on marketing environments, promotions and price are the most promising avenues. According to Kumar et al. (2015), 50% of FMCG consumers in India are classified as lower-income and are highly price-conscious. In addition, less-educated consumers are more likely to refer to price as an indication of quality (Shapiro, 1973). Cohen et al. (2017) assert the importance of promotions, and found that 12% to 25% of FMCGs in select European countries were sold as promotions. Tarallo et al. (2019) developed a model where the research shows that promotions are key for demand accuracy at the specific products level, but that promotions significantly increased the complexity of their model because of the different possible combinations of product types, categories, periods, and geographies. Nyaga (2014) found that prices for FMCG products, specifically relative to competition, affects demand for products the most. Yang et al. (2015) observed the most improvements in their forecasting model for FMCG distributor demand when looking at the retailer level demand data, which was forecast based on retailer price and promotion information.

The literature on the macroenvironment points to economic, natural, demographic, and cultural forces having the greatest impact on total demand of FMCGs. Shetty et al. (2020) asserts that the growth of the FMCG market in India has been mostly in the double digits and closely linked with the GDP growth of the economy, which stands to reason because better macroeconomic indicators mean consumers have more spending power. Spending power can also be linked to employment rates and stock market indexes. A potential economic alternative to GDP is inflation rates, which more accurately aligns with the growth of the Romanian FMCG market (Stanciu, et al., 2019). Dachyar et al. (2021) found that using festival days and weather led to the most accurate demand forecasting, more so than population or income rates. Abu Talib

et al., (2023) studied water demand and found that sociodemographic factors, such as population ages and household types in an area, had the strongest correlations. Secondary to sociodemographic were seasonal factors, like holidays, and natural factors, like temperature, humidity, precipitation, and air pollution. Finally, according to our interviews with the sponsor company's team in Japan that has recently done similar research, the most relevant exogenous variables contributing to demand are the number of daily COVID-19 cases, COVID-19 state of emergency, temperature, rainfall, and household trends such as the number of purchasers, units purchased, frequency of purchase, and market penetration.

It is worth noting that, while most of the reviewed literature postdate the COVID-19 pandemic, they did not address how the pandemic and its impact on the variability of demand data should be accounted for by exogenous variables.

While access to data on exogenous variables can sometimes be difficult to obtain, they can lead to more accurate forecasts by explaining underlying trends in the demand data. It is necessary to analyze each exogenous variable against the demand for each product category to understand which exogenous variables are the most important for the sponsor company's demand forecasting.

2.3 Forecasting Models

The following section introduces forecasting models that will be used in this paper. The simplest forecasting method and thus the benchmark model in this paper is the naive forecast, which assumes the predicted value as being equal to the previous value. While the naive forecast

is not expected to add any value in analyzing the demand data, it is helpful as a benchmark for comparison against more sophisticated models.

2.3.1 Moving Average

In statistics, a moving average (rolling average or running average) is a calculation to analyze data points by creating a series of averages of different subsets of the full data set. Given a series of numbers and a fixed subset size, the first element of the moving average is obtained by taking the average of the initial fixed subset of the number series. Then the subset is modified by "shifting forward"; that is, excluding the first number of the series and including the next value in the subset (Booth, et al., 2006).

2.3.2 Autoregressive Integrated Moving Average

Predicting future demand is an example of time series forecasting, which is the attempt of predicting future values based on previously observed values. Among the many models in time series analysis, we start with Autoregressive Integrated Moving Average (ARIMA), a method first created by Box et al. (2008) and has now become one of the most general classes of models for forecasting a time series (Nau, 2016). ARIMA is a combination and generalization of the autoregression (AR) model, which predicts the future value of a variable with past observations of itself, and the moving average (MA) model, which predicts the future value of a variable with past forecasting errors.

ARIMA models are generally denoted ARIMA(p,d,q), characterized by three non-negative integer parameters p, d, and q. Among these, p is the order of the autoregressive model, d is the

degree of differencing (the number of times the data have had past values subtracted), and *q* is the order of the moving-average model (Hyndman & Athanasopoulos, 2018).

To determine the appropriate values of the three parameters for a given model, the first step is setting the order of integration, denoted *d*, to the minimum number of times a series must be differenced to become stationary. After that, multiple ARIMA(*p,d,q*) models can be estimated with different combinations of AR order *p* and MA order *q*, and the (*p,q*) combination that gives rise to lowest prediction error, quantified by the Akaike Information Criterion (AIC), would be eventually selected as model parameters (Peixeiro, n.d.).

2.3.3 Seasonal Autoregressive Integrated Moving Average

The ARIMA model can be modified to account for seasonality in a time series. The resulting seasonal autoregressive integrated moving average (SARIMA) model allows us to take into consideration periodic patterns when forecasting a time series, which generally cannot be achieved by the ARIMA(*p*,*d*,*q*) model (Peixeiro, n.d.).

SARIMA models are usually denoted SARIMA(*p*,*d*,*q*)(*P*,*D*,*Q*). The three additional parameters are *P*, which is the order of the seasonal AR(*P*) process; *D*, which is the seasonal order of integration; and *Q*, which is the order of the seasonal MA(*Q*) process. All of these are estimated in ways similar to those of their ARIMA(*p*,*d*,*q*) model counterparts *p*, *d*, and *q* (Peixeiro, n.d.).

2.3.4 Seasonal Autoregressive Integrated Moving Average with Exogenous Variables

Other than seasonality effects, the ARIMA model can also be adapted to include external variables that predict the time series, which leads to the SARIMAX model. To initiate the SARIMAX

model, not only must all six parameters be decided in the SARIMA(p,d,q)(P,D,Q) model without exogenous variables, but a list of external predictors must also be examined to make sure they are correlated with the dependent variable (Peixeiro, n.d.).

There has been extensive literature on the application of ARIMA, SARIMA, and SARIMAX models in forecasting. For example, Andrews, et al. (2013) built an ARIMAX model to predict long-term disability benefit application rates by using not only autoregressive and moving average terms but also external indicators such as GDP, employment, and income.

2.3.5 Autoregressive Integrated Moving Average with Exogenous Variables

Autoregressive Integrated Moving Average with Exogenous Variables (ARIMAX) is a variant of ARIMA that incorporates additional external factors, known as exogenous variables, to improve the accuracy of the forecast. These variables can help capture the impact of external factors, such as holidays, economic indicators, or marketing campaigns, on the time series. In ARIMAX, the exogenous variables are included as additional input variables in the model, alongside the lagged values of the target variable and the forecast errors from the MA component.

The ARIMAX model is denoted as ARIMAX(p, d, q)(P, D, Q)_m, where p, d, and q represent the order of the autoregressive, integrated, and moving average components of the time series, while P, D, and Q represent the order of the autoregressive, integrated, and moving average components of the exogenous variables, and m represents the number of time periods in each seasonal cycle (Andrews, et al. 2013).

2.3.6 Machine Learning

The aforementioned time series models work particularly well when datasets are small, defined as having fewer than 10,000 data points, and coarse-grained, defined as the seasonal period being monthly, quarterly, or yearly. If the dataset is very large or highly granular, as might be the case with some of the data provided by our sponsor company, those statistical models would become very slow and their performance would quickly degrade (Peixeiro, n.d.).

In those situations, we have to turn to machine learning. Machine learning is a subset of machine learning that focuses on building models on neural network architecture. It tends to perform better as more data is available, making it an appropriate choice for forecasting high-dimensional time series.

Machine learning captures specific trends in data differently. First, it does not need to have a time-series order. For example, if predicting revenue based on ad spend, it does not matter when a certain amount was spent on ads. Instead, one must relate the amount of ad spend to the revenue. Sometimes randomly shuffling the data can make it even more robust. Machine learning techniques are also able to capture nonlinear relationships, such as more than one seasonality within a dataset. For example, hourly temperature data would have seasonality within a day, when it gets colder at night, as well as within a year, when it gets colder in the winter (Peixeiro, n.d.).

There are many machine learning models, such as logistic regression, tree-based models, and neural networks. One of the most popular tree-based models in time series studies is

eXtreme Gradient Boosting (XGBoost). It is an ensemble method that uses decision trees as base models and combines them to make predictions (Gumus & Kiran, 2017).

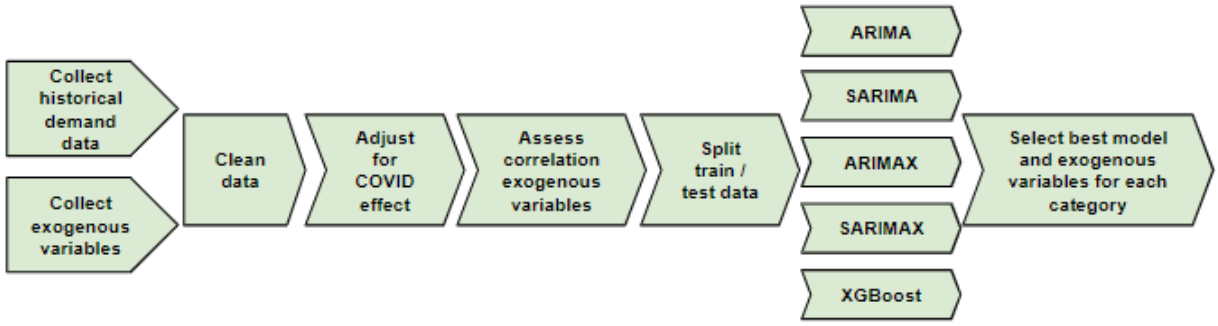
Among the ARIMA family, SARIMAX is the most advanced, but it does not necessarily produce superior forecasting results over the other models for a variety of reasons. For example, exogenous variables might not add value for products with non-elastic demand. Some product categories may be so essential that their demand is not affected by such factors as weather or even advertising budgets. Either way, SARIMA would not be better than ARIMA in the case that a product has no seasonal pattern. Additionally, classic time-series forecasting, like moving average, may appear more reliable if the product has a simple demand pattern.

3. METHODOLOGY

As standard in the demand forecast literature, the best approach to this capstone is to take a model-agnostic stance and to choose the most appropriate model for each category by conducting numeric experiments with real data and benchmarking according to metrics such as the mean absolute percentage error (MAPE). The specific steps are illustrated by Figure 2 and are elaborated on in the following sections: (1) data cleaning and processing; and (2) models selection and validation.

Figure 2

Methodology Steps



3.1 Data Cleaning and Processing

To initiate our research, we met with the sponsor company team in India to conduct qualitative interviews on their experience with the data and understand exogenous variables they believe are likely to influence demand in various categories. It is important to note that our sponsor company's Japan team has a demand planning system with exogenous variables, but we did not adopt their approach, because the Indian market is much more complex and diverse, making comparisons to the Japan market impossible. We did however arrange interviews with the Japan team to leverage their experiences in forecasting process, identified exogenous variables, forecasting models, deliverables, and realistic expected outcomes to benchmark against.

We collected the dependent variable from the sponsor company. Historical distributor sales data at month-level and product category-level was available from July 2010 to March 2023. All sales were measured in a standard unit to account for irrelevant variations across specific stock keeping units (SKUs). To clean the data, we removed discontinued product lines and used

descriptive statistics to identify missing data and outliers to replace using simple imputing. Products were ultimately grouped into 10 unique product categories that were most relevant for the business.

For exogenous variables, we acquired data the sponsor company had already purchased as well as data available in academic databases that aligned with our literature review. Variables must be easy for the sponsor company to obtain in order to update for future forecasts. The two main sources of data were NielsenIQ and CEIC Data. NielsenIQ collects a random sample of retailer sales data and extrapolates it to share consumer trends with companies (NielsenIQ, 2023). CEIC Data curates economic, industrial, and financial time-series data to gain market insight – with a particular focus on emerging economies (CEIC Data, 2023). A summary of the variables we considered is in Table 1.

Table 1

Exogenous Variables

Variable	**Type**	**Source**	**Length**	**Depth**
Total COVID-19 Cases	Natural	CEIC Data	3 years	Regional
Consumer Price Index (CPI)	Macroeconomic	CEIC Data	12 years	National
Goods and Services Tax (GST)	Macroeconomic	CEIC Data	7 years	Regional
Consumer Trends	Macroeconomic	NielsenIQ	2 years	National; product level
Rainfall	Natural	CEIC Data	15 years	Regional

We converted all exogenous data, where needed, to month-level so that they can be

matched with the dependent variable. Some variables that were prevalent in the literature review, such as holidays, weren't prioritized because they were not relevant for month-level data. We also prioritized exogenous factors that were available for as long a time period as possible.

For all variables, we considered lags of both six and 12 months when looking at correlation and assessing forecast accuracy. This method is necessary to prevent data leakage. As the priority of our research was to forecast for six months into the future, a minimum of six-month lag is needed. We also looked at a 12-month lag to see if there would be any significant difference in case the sponsor company wanted to forecast for even further into the future later.

While COVID-19 was considered in the form of COVID-19 cases, we also considered a few other options to address this variability in the demand trend. This was relatively experimental as COVID-19 is so new and its effects have not been considered yet in most literature. We added an exogenous binary variable for COVID-19 lockdown from March to May 2020. We also experimented with replacing the dependent variable during those 3 months with demand data from the previous year during the same months.

Correlation analyses were run to confirm the validity of a variable in the forecasting model. We also looked at correlations between exogenous variables, but found none except for different lags of the same variable. In the case where an exogenous variable had a strong correlation for multiple lags, we chose just one with the strongest correlation to avoid multicollinearity.

Finally, we split our sample, including both demand data and exogenous data, into test and training datasets. As the test set must be equivalent to the forecasting horizon, the test set consists of data in the final six months of the sample data, and all of the data in the preceding periods belong to the training set. Our forecasting models are estimated from data in the training set, and then they generated demand forecasts over the test period, which were compared with the actual demand during the same period. To leverage the limited available data and ensure a more robust result, we assessed our model over more than one time period to achieve a sense of cross validation.

3.2 Models Selection and Validation

Once predicting variables were selected, we incorporated them into a broad variety of time series models to predict future demand. Since the scope of this project is relatively high level, by category instead of SKU, and the available internal and exogenous data is monthly, it was unlikely that machine learning techniques would add much value. Instead, we prioritized statistical models. Contemporary Python frameworks, such as *autoarima* and *pmdarima*, are adopted by our study to automate the experiments in identifying the optimal parameters. The python versions of these formulas are listed in Appendix A.

For naive models, a variable is forecasted as being equal to its previous value as seen in Equation 1:

$$y_t = y_{t-1} + \epsilon t \quad (1)$$

where y_t is the actual demand at time t.

For ARIMA, ARIMAX, SARIMA, and SARIMAX models, our task was to estimate coefficients ϕ, θ, and β in Equation 2:

$$y'_t = C + \phi_1 y'_{t-1} + \cdots \phi_p y'_{t-p} + \theta_1 \epsilon'_{t-1} + \cdots + \theta_q \epsilon'_{t-q} + \beta_1 X_{1t} + \cdots + \beta_n X_{nt} + \epsilon_t \quad (2)$$

where y'_t is the dependent variable to be forecasted, $\phi_p y'_{t-p}$ represents the past values of the differenced series, $\theta_q \epsilon_{t-q}$ is the past error terms, $\epsilon y'_t$ is the current error term, and X_{nt} is the external explanatory variables.

The estimation follows these procedures:

For all four models, the first step is to check whether the data is stationary. A simple plot of the dependent variable against time allows us to observe whether the series has an increasing or decreasing trend. However, a more rigorous approach to determine stationarity is to apply the augmented Dickey-Fuller (ADF) test. If the ADF statistic is a large negative number and the associated p-value is below 0.05, we can reject the null hypothesis that the series is not stationary and proceed to the next step. Otherwise we have to determine how many times the series must be differenced to become stationary, which is done by repeatedly first-differencing the series until the ADF statistic and p-value of the resulting series become small enough. The number of times the original series has been differenced during this process sets the order of integration.

If the data displays periodic patterns and we are estimating the SARIMA model that accounts for the seasonality, the seasonal order of integration must also be determined at this stage. It is set to the minimum number of times seasonal differencing is applied in order to make the time series data stationary.

If we are forecasting demand using exogenous variables as well as by estimating the SARIMAX model, variables would be selected at this time. We would perform a correlation analysis, in which the coefficient of correlation between the dependent variable and various candidate independent variables to be calculated, and those variables that have a strong correlation, i.e., a positive coefficient close to 1 or a negative coefficient close to -1, would be chosen as explanatory variables in the SARIMAX model.

Now we are left with four parameters to pin down: the autoregressive order *p*, the moving average order *q*, and their seasonal equivalents *P* and *Q* if we are fitting the SARIMA or SARIMAX model. Their values are optimally chosen by fitting every combination of SARIMA(*p,d,q*)(*P,D,Q*) and selecting the model with the lowest Akaike information criterion (AIC). AIC rewards goodness of fit (as assessed by the likelihood function), but it also includes a penalty that is an increasing function of the number of estimated parameters. The penalty discourages overfitting, which is desired because increasing the number of parameters in the model almost always improves the goodness of the fit.

Alternatively, we can also select and compare models based on their out-of-sample performance. Since we had divided our sample data into a training and a test subset, each model is able to make a demand forecast over the test period, and we can calculate the mean absolute percentage error (MAPE) for each model by computing the average difference between the forecasted demand and the actual demand. The model with the best out-of-sample performance is then the one that yields the lowest MAPE.

In our case, we made sure that each result had an AIC of four digits or less and that the AIC of the winning model was less than the AIC for the naive forecast. However, MAPE is better at comparing accuracy across data sets, such as across product categories or other products for management purposes. As such, we are prioritizing explaining the accuracy of these models with MAPE.

4. RESULTS AND ANALYSIS

To improve the accuracy of category-level demand forecasting for the next six months, this study investigates the impact of exogenous and internal variables. The following sections detail the results: (1) best results; (2) time-series models; (3) exogenous variables; (4) COVID-19 abnormalities; (5) product categories; (6) coefficient of variation; and (7) the Demand Forecast application.

4.1 Best Performance

The methodology in Section 3 is implemented for the demand data of all 10 of the sponsor company's product categories. The best model, exogenous variables, and treatment of COVID-19 abnormalities, as well as forecast errors using the mean absolute percentage error (MAPE) are shown in Table 2.

Table 2

Six-Month Forecast Best Results with MAPE

Category	Best Model	Exogenous Variables	COVID	MAPE Jan - Jun 22	MAPE Oct - Mar 23
1	SARIMA (1,1,1)(0,1,1,12)	n/a	n/a	7.2%	12.9%
2	ARIMAX (0,1,1)	CPI/Rain lag 6	n/a	11.8%	13.8%
3	SARIMAX (1,0,0) (1,0,0,12)	CPI/Rain lag 6	Replace	4.2%	8.4%
4	SARIMAX (2,1,2)(2,0,0,12)	CPI/Rain lag 6	Replace	3.6%	5.3%
5	SARIMAX (1,0,0) (2,0,0,12)	CPI/Rain lag 6	n/a	5.3%	16.1%
6	SARIMAX (1,0,1)(1,0,0,12)	CPI/Rain lag 12	Replace	6.2%	10.5%
7	SARIMA (1,0,0)(1,0,2,12)	n/a	n/a	20.2%	14.1%
8	SARIMAX (2,1,1)(1,0,0,12)	CPI/Rain lag 12	n/a	3.4%	15.7%
9	ARIMA (1,0,1)	n/a	Replace	7.4%	10.8%
10	SARIMA (3,0,0)(1,0,1,12)	n/a	Replace	10.5%	7.2%
Average				**8.0%**	**11.7%**

The average MAPE across all product categories, with test data from January 2022 to June 2022, is 8.0%. Results were also run again with updated data, making the test data from October 2022 to March 2023, for which the average MAPE across product categories was 11.7%. This method of cross validation confirms the robustness of the model. Our primary focus in assessing accuracy lies in the MAPE due to its comparability across diverse categories.

In Sections 4.2, 4.3, and 4.4, we go into more detail to compare the models, exogenous variables, and ways of dealing with COVID-19-related abnormalities.

4.2 Time-Series Models

The models used for comparison are naive, ARIMA, ARIMAX, SARIMA, SARIMAX, and XGBoost. Using the optimal exogenous variables and approach for COVID-19 in Table 2, we compared all models for the January 2022 to June 2022 test data, as shown in Table 3.

Table 3

Six-Month Forecast Model Comparison with MAPE

Category	Naive	ARIMA	ARIMAX	SARIMA	SARIMAX	XGBoost
1	55.00%	35.10%	30.80%	7.20%	10.60%	49.50%
2	24.50%	17.80%	11.80%	21.30%	14.20%	19.50%
3	7.60%	4.20%	6.40%	8.20%	4.20%	6.50%
4	9.80%	6.80%	4.50%	7.20%	3.60%	7.60%
5	7.00%	5.90%	6.40%	8.70%	5.30%	7.10%
6	10.90%	12.40%	9.80%	6.50%	6.20%	15.00%
7	107.30%	31.40%	50.60%	20.20%	31.40%	67.30%
8	17.20%	6.10%	5.30%	4.00%	3.40%	7.40%
9	12.60%	7.40%	9.50%	7.50%	9.50%	11.30%
10	21.00%	16.30%	19.10%	10.50%	17.00%	12.00%
Average	**27.30%**	**14.30%**	**15.40%**	**10.10%**	**10.50%**	**20.30%**

Our analysis reveals compelling evidence for the superiority of SARIMA and SARIMAX models in forecasting demand for the sponsor company. We conducted a rigorous evaluation of various models and found that the naive forecast, the simplest model, performed the worst. This was not surprising, but it served as a useful baseline against which to compare other models. Our findings are aligned with the literature review, which suggested that statistical models like SARIMAX tend to outperform machine learning models like XGBoost. Moreover, our project uses

month-level data, which does not benefit from evaluating more complex models that excel with multiple seasonality.

SARIMA and SARIMAX models were the clear winners in our evaluation, performing significantly better than the other models in most of the product categories. The results suggest that there is strong seasonality in 80% of the product categories, and that exogenous variables such as CPI and rainfall play a crucial role in explaining at least some of the variability in demand. We found that incorporating exogenous variables improved forecast accuracy in 60% of the categories. By identifying the most relevant exogenous variables and building models that account for their influence, we have provided the sponsor company with a powerful tool for forecasting demand at the product category level.

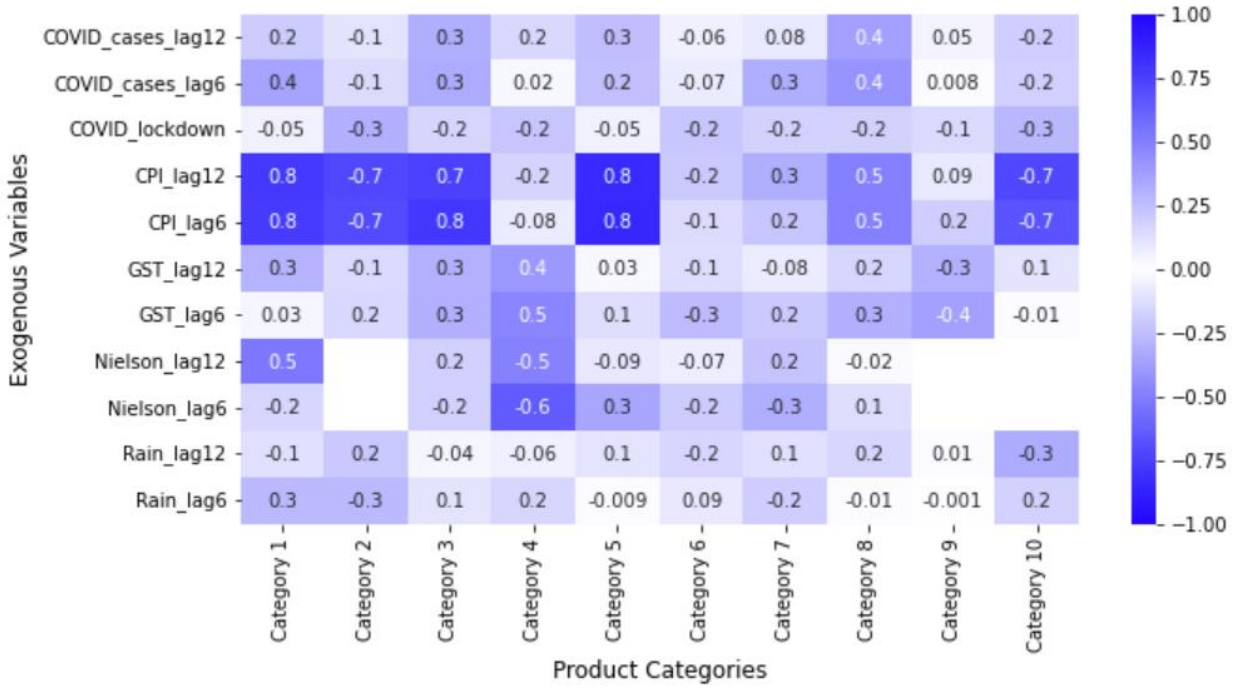
4.3 Exogenous Variables

The most helpful exogenous variables were CPI and rainfall, which improved the forecast accuracy in 60% of categories. Variables that are only available for three years (Nielsen and GST) did not help the forecast. Nielsen data correlated well in some product categories, like Category 4, as can be seen in Figure 3. However, since Nielsen's data only covers three years, it would not be as effective as using the full 12 years of demand data. The strength of statistical models, like SARIMAX, benefit from more data to show a more reliable trend.

To avoid data leakage, all variables were lagged by at least six months. We also tried creating a lag of 12 months to see if the correlation was different. Figure 3 shows the correlations between the dependent variable for each product category and each exogenous variable, lagged by both six and 12 months.

Figure 3

Correlation Heatmap



To ensure the validity of our forecasting models, we conducted thorough correlation analyses on exogenous variables. While we found no significant correlations between different exogenous variables, we did uncover strong correlations between different lags of the same variable. To prevent multicollinearity, we selected only the lag with the strongest correlation for use in our models.

To further improve our models, we experimented with forecasting exogenous variables using the ARIMA method based on their historical patterns. However, our findings showed that this approach did not result in a lower MAPE for the product categories compared to using lagged variables.

4.4 COVID-19 Abnormalities

To improve the accuracy of the forecasting model in light of the abnormalities caused by COVID-19, we explored three different approaches. The first approach involved using the total number of COVID-19 cases in a month as an exogenous factor, which was then lagged by six. For the second approach, we created a binary variable to identify periods encompassing significant lockdowns. This variable was set to 1 for the period from March 2020 to May 2020 and 0 for the rest of the months over the 12-year period. The third approach involved replacing demand data during the three months affected by COVID-19 with demand data from the same months in the previous year.

The results of our analysis revealed that replacing the dependent variable demand with demand data from the previous year during the same months improved forecast accuracy in 50% of product categories. This approach was found to be more effective compared to the other two approaches we tried. Using the binary lockdown variable was slightly less effective than using previous year's demand data, resulting in a 1% lower forecast accuracy on average. Similarly, using the number of COVID-19 cases as an exogenous factor was less effective than using previous year's demand data, resulting in a 3% lower forecast accuracy on average.

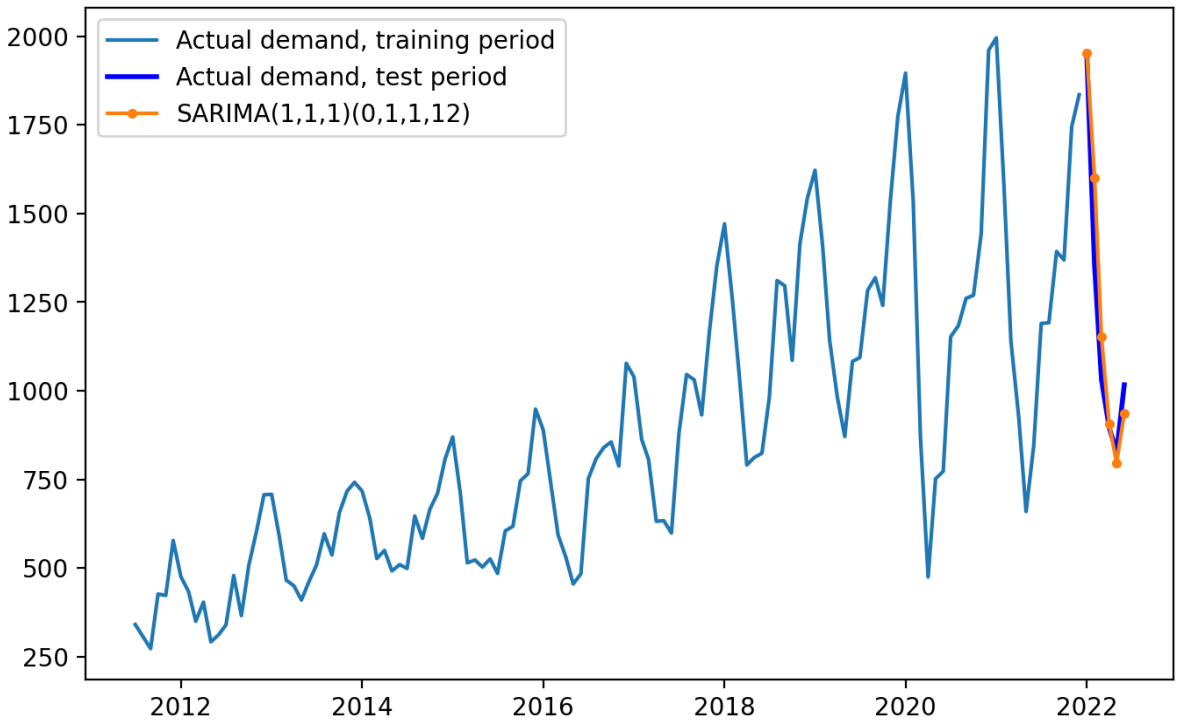
4.5 Product Categories

We apply the model selection procedures explained in Section 3.2 to the time series demand data of all 10 of the sponsor company’s categories. For the clarity of presentation, we take Category 1 as an illustrative example to lay out the full details of parameter calibration and model estimation. We then briefly discuss the trends of each specific product category.

Figure 4 shows the time series of product Category 1 demand, characterized by a noticeable growing trend and strong seasonal fluctuations. The first 131 months of the series are taken out to train the data, while the last six months are separated for data validation. At the same time, two sets of exogenous lagged time series data are prepared, i.e., CPI and rainfall with a six-month lag, and CPI and rainfall with a 12-month lag.

Figure 4

Category 1 Forecast Results



As discussed in Section 3, stationarity of the forecast target must first be guaranteed. Upon examination, the series displays both linear trend and seasonal trend. The combination of a 1st-order differencing and a 1st-order seasonal differencing proved to be both necessary and sufficient to transform the data into a stationary time series suitable for further analysis.

On the stationary demand data, ARIMA(*p,1,q*), ARIMAX(*p,1,q*), SARIMA(*p,1,q*)(*P,1,Q,12*), and SARIMAX(*p,1,q*)(*P,1,Q,12*) models are fitted, as the differentiation and seasonal differentiation parameters, *d* and *D*, have already been fixed to be 1 and the seasonality parameter, *m*, has been manually set to 12 to reflect the monthly data. A stepwise algorithm outlined in Hyndman and Khandakar (2008) is used to determine the AIC-minimizing parameters in each family of models. ARIMA(*1,0,4*), ARIMAX(*3,0,2*) lag 6, SARIMA(*1,1,1*)(*0,1,1,12*), and SARIMAX(*1,0,0*)(*2,0,0,12*) lag 6 are found to be the optimal models in each family, among which SARIMA(*1,1,1*)(*0,1,1,12*) is chosen to be the best model, with the lowest AIC of 1397 and lowest MAPE of 7.2%.

In order to gain a more comprehensive understanding of the outcome of the SARIMAX model, a detailed analysis of Category 1 was conducted and is shown in Table 4.

Table 4

Category 1 Estimated Coefficients - SARIMAX Model

	coef	se	z	p	0.25	0.75
cpi	244.74	281.06	0.87	0.38	-306.12	795.60
rain	-31.15	31.20	-1.00	0.32	-92.30	30.01
ar1	0.40	0.12	3.44	0.00	0.17	0.62
ma1	-0.84	0.09	-9.73	0.00	-1.01	-0.67
mas12	-0.33	0.10	-3.52	0.00	-0.52	-0.15

To interpret the results of the SARIMAX model in Category 1, the forecasted demand for month *t* can be derived using Equation 3:

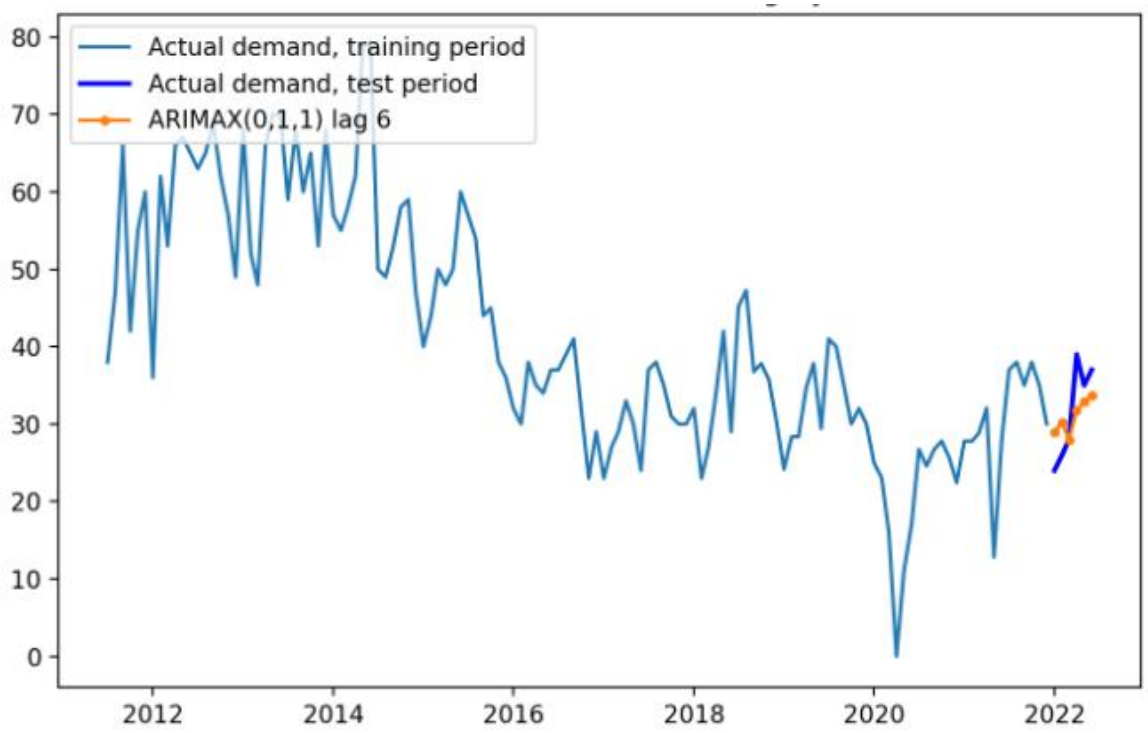
$$Demand_t=244.73CPI_{t-6}-31.15Rain_{t-6}+0.40Demand_t-1-0.84Error_t-1-0.33Error_{t-12} \quad (3)$$

This means that the demand for Category 1 products in any month is positively affected by CPI six months ago and the demand one month ago, and negatively affected by the forecast errors one month ago and 12 months ago, as well as by rainfall six months ago.

Next, we look at the remaining product categories in less detail, starting with Category 2. Figure 5 shows the history of product Category 2 demand with MAPE of 11.8% and AIC of 870.

Figure 5

Category 2 Forecast Results



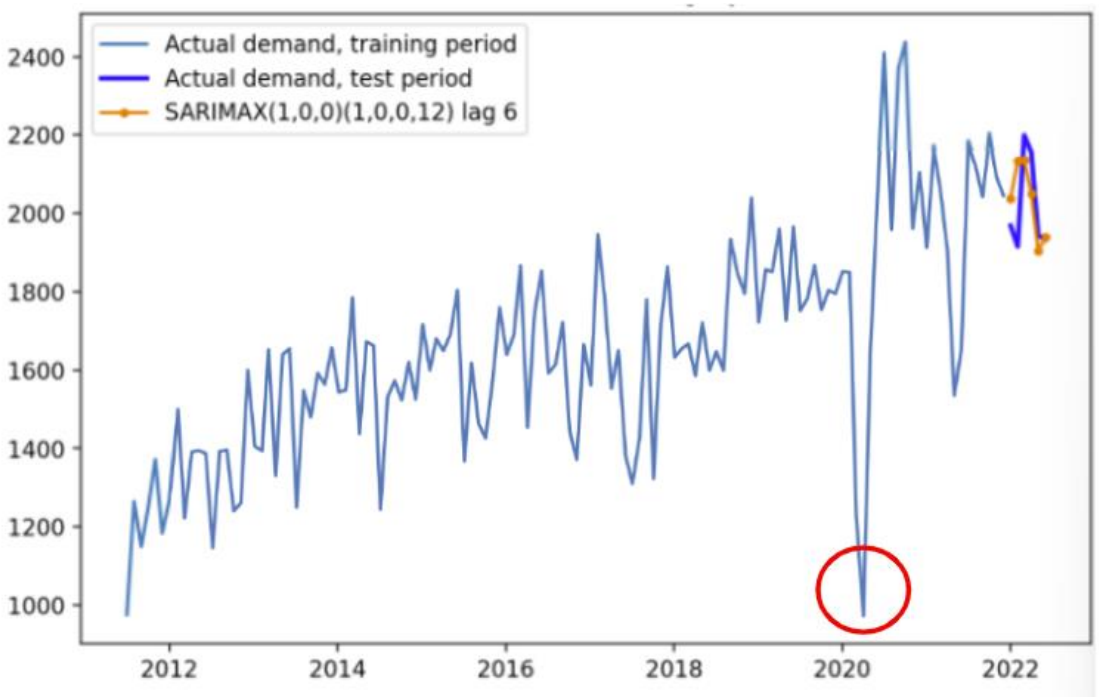
There is no clear seasonality with Category 2, confirmed by the fact that the most accurate model is ARIMAX (0,1,1). This model assumes that the current value of the time series is a function of its previous forecast error, after differencing the time series once as $d = 1$. $p = 0$. This implies that the present value of the time series is independent of its past values, signifying a

moving average model. It essentially represents a random walk process, without exhibiting a prominent trend. However, it is worth noting that exogenous variables CPI and rainfall had a strong effect on the forecast accuracy, improving results from 17.8% to 11.8%.

The time series data of product Category 3 is displayed in Figure 6, with a corresponding MAPE of 4.2% and AIC of 1649.

Figure 6

Category 3 Forecast Results



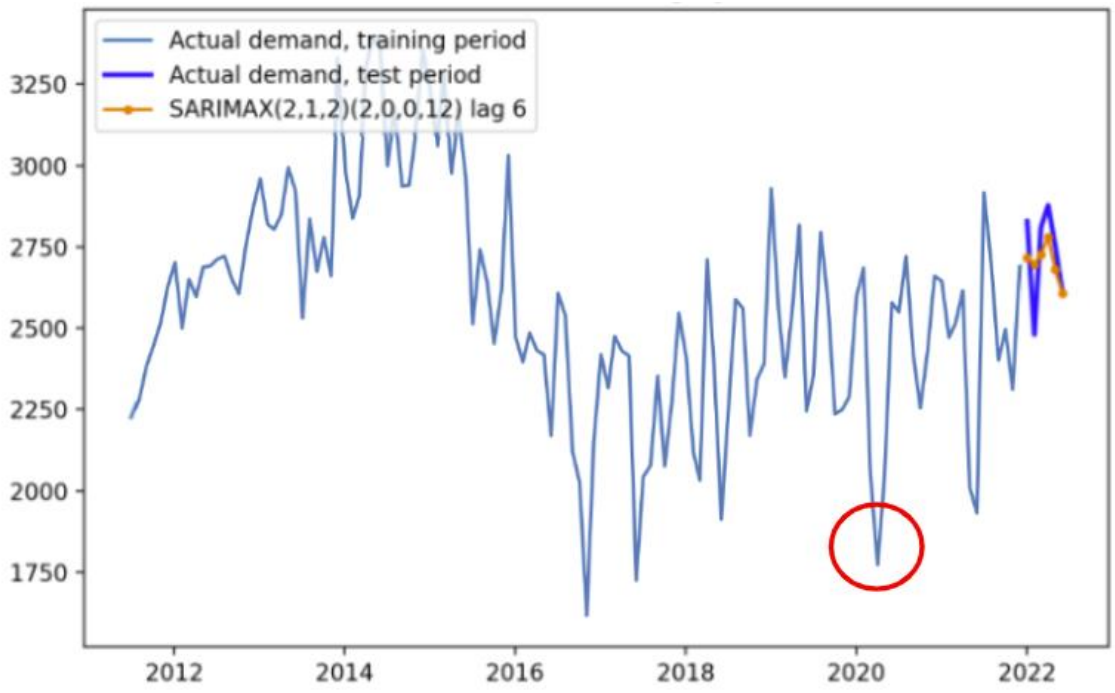
Category 3 is the first category we explore that benefitted from replacing demand data during the first three months of COVID-19 with data from the previous year. The red encircled region denotes a notable anomaly during the corresponding time frame, which may have otherwise disrupted the model. The winning model, SARIMAX (1,0,0)(1,0,0,12) lag 6, suggests that both seasonality and exogenous variables significantly contribute towards the explanation

of the model. Both the non-seasonal and seasonal components show an autoregressive term of order 1, which means that the value of the time series at the current time step is related to its value at the previous time step and at the same time step in the previous year.

The MAPE and AIC for product Category 4 are 4.5% and 1734, respectively, as shown in Figure 7's time series plot.

Figure 7

Category 4 Forecast Results



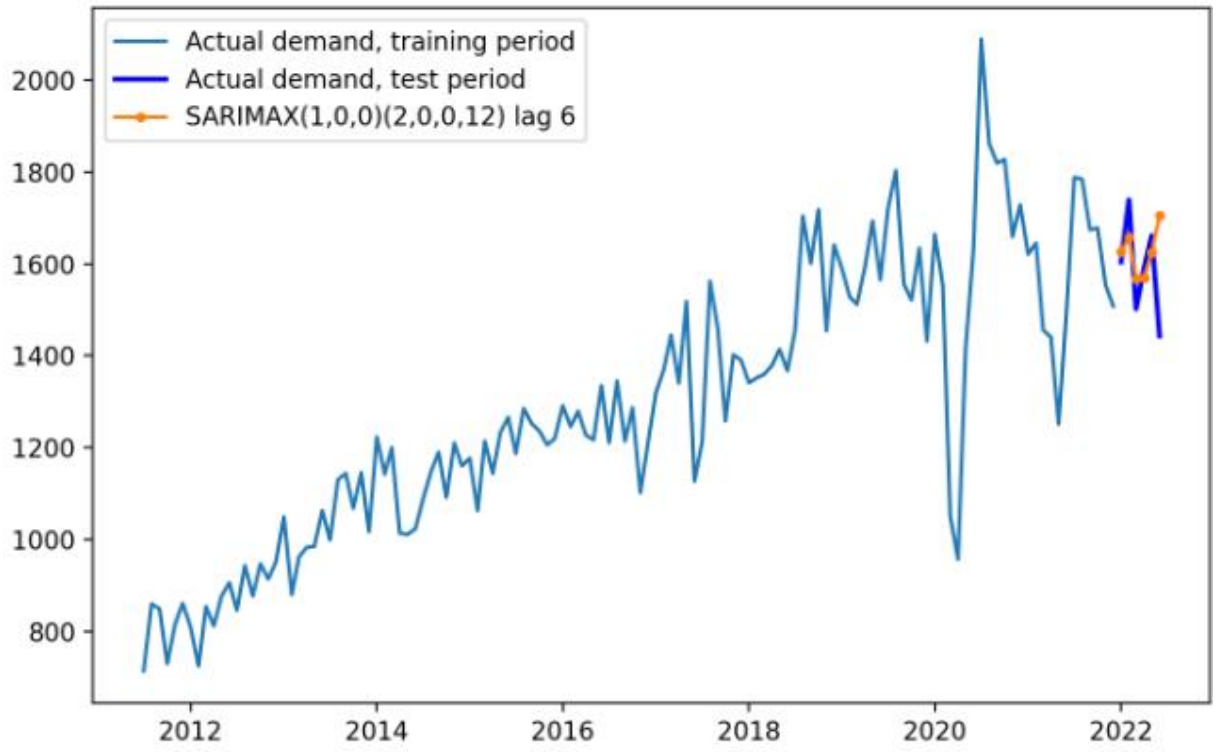
Category 4 is characterized by bi-annual seasonality and is also affected by exogenous variables CPI and rain. This category does have a strong correlation of 0.6 with exogenous variable Nielsen (consumer trends), but using that variable would mean limiting the statistical model to only three years' worth of trends and the seasonality effect would not have been as apparent. Replacing the March 2020 to May 2020 demand data with the previous year indeed helps with

the forecast accuracy, as there was an effect of COVID-19 on this product category. The low standard deviation of the demand data means that most models perform relatively well on this data set, as can be seen by even a naive forecast having 9.8% accuracy.

Moving on to the next analysis, Figure 8 displays the time series data for product Category 5, exhibiting a MAPE of 5.3% and an AIC value of 1579.

Figure 8

Category 5 Forecast Results



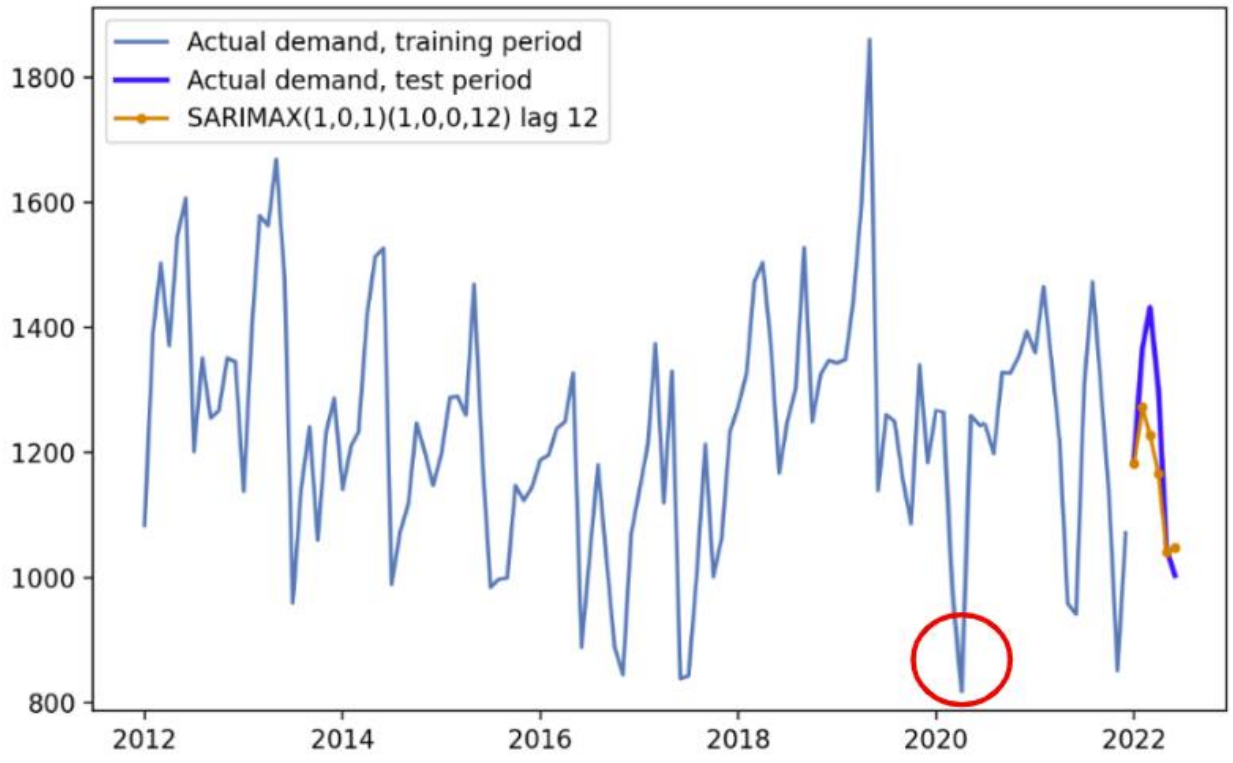
Category 5 is characterized by a clear linear trend and some seasonality. Since the seasonal autoregressive term (*P*) is equal to 2, the value of the time series in the current month depends on the values of the time series in the same month of the previous two years. Exogenous variables, particularly CPI, which had a 0.8 correlation with product Category 5 demand, also

helped interpret this model. Rainfall, with a much smaller correlation, had less impact but still helped improve the accuracy. Despite the significant impact of COVID-19 on product Category 5, the model did not derive any benefit from adjusting for this abnormality.

Figure 9 illustrates the time series data for Category 6, which had a MAPE of 6.2% and an AIC of 1551.

Figure 9

Category 6 Forecast Results



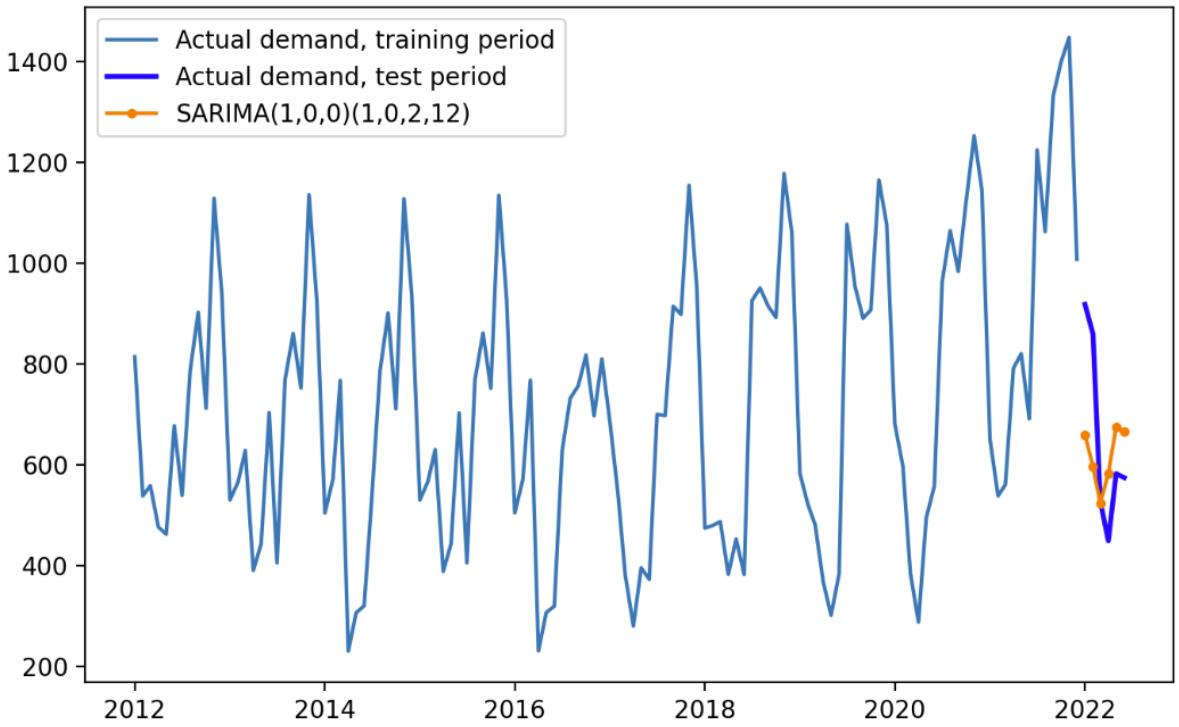
Category 6 sees a strong seasonality with increasing demand during the dry winter months and lower demand during the rainy summer months. Similar to other categories, this model benefits from replacing demand data during COVID-19, as highlighted in the red circle. This category does not see much of a linear trend over time, hence there is no need to make the

data stationary (*d* = 0). For this category, the predictive performance of lagged exogenous variables by 12 months was found to be superior to that of being lagged by six months, marking the first instance of such a trend in our analysis. This can be attributed to the fact that the correlation is twice as high for the 12-month lag than for the six-month lag. However, despite this finding, it is important to note that the correlation between the exogenous variables, namely CPI and rainfall, is relatively low. This explains why the model only slightly outperforms the SARIMA model.

Moving on to Figure 10, we can see the time series data for Category 7, which has a MAPE of 20.2% and an AIC of 1544.

Figure 10

Category 7 Forecast Results

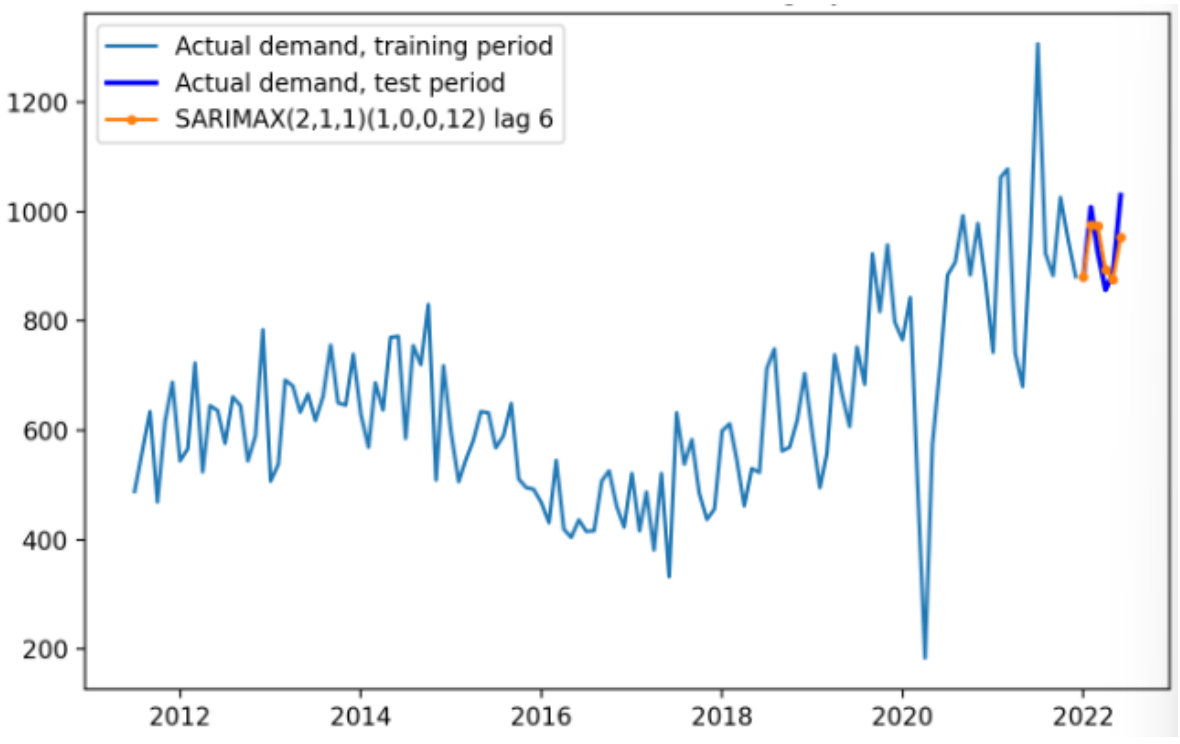


Despite the MAPE being the highest for Category 7 out of all product categories, the fact that the AIC is in line with the rest explains that this category is generally more difficult to forecast because of the variation in the demand data. Despite the high MAPE, Figure 10 shows that the model has effectively identified a trend in the data. This product follows extreme seasonality with very low demand during the dry winter months and very high demand during the rainy summer months. This best model of SARIMA (1,0,0)(1,0,2,12) tells us that the value of the time series at the current time step is related to its value at the same time step in the previous year.

Shifting our focus to Figure 11, we observe the time series data of Category 8, displaying a MAPE of 3.4% and an AIC of 1477.

Figure 11

Category 8 Forecast Results

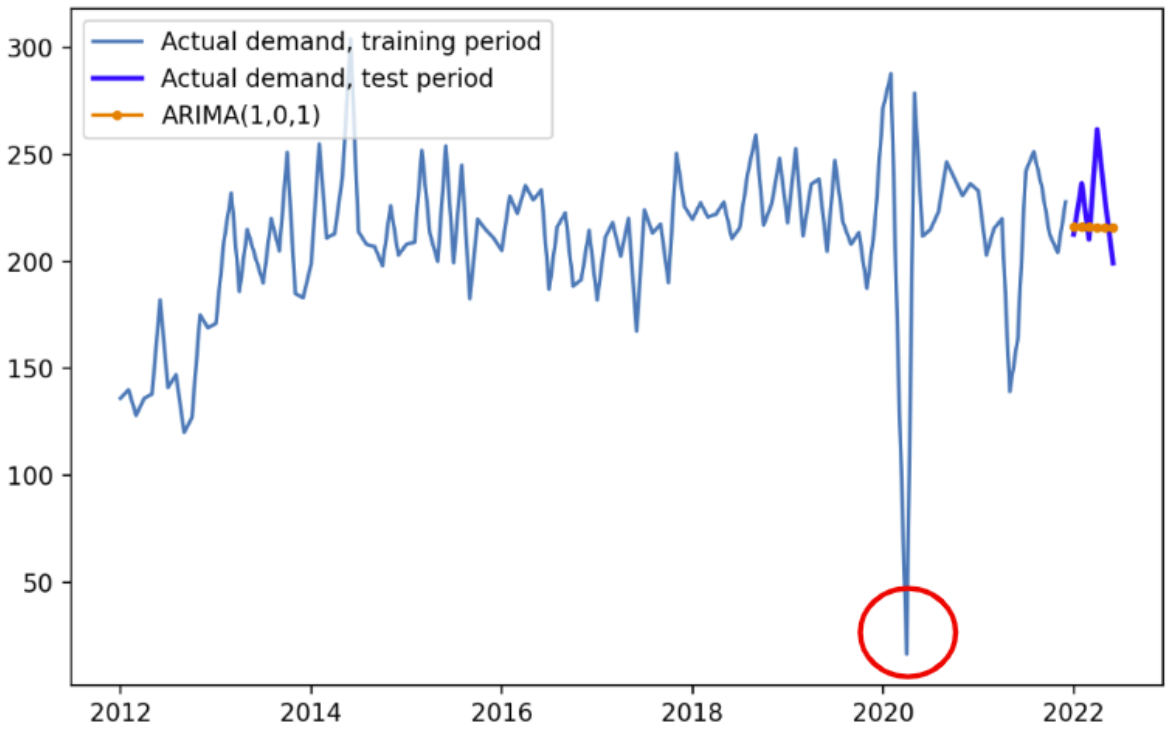


The best model for Category 8 was SARIMAX (2,1,1)(1,0,0,12). While Category 8 was clearly affected by COVID-19, this model interestingly did not benefit from correcting that abnormality. The fact that SARIMAX is the best model confirms that using exogenous variables help explain the demand trend, but this is the second category that benefitted from exogenous variables lagged 12 months instead of six months. The correlation of rainfall with demand for Category 8 is 20 times higher when lagged 12 months instead of six months. The correlation of CPI with Category 8 demand is nearly identical for both lag periods.

Turning our attention to Figure 12, we can see the time series data for Category 9, with a MAPE of 7.4% and an AIC of 1199.

Figure 12

Category 9 Forecast Results

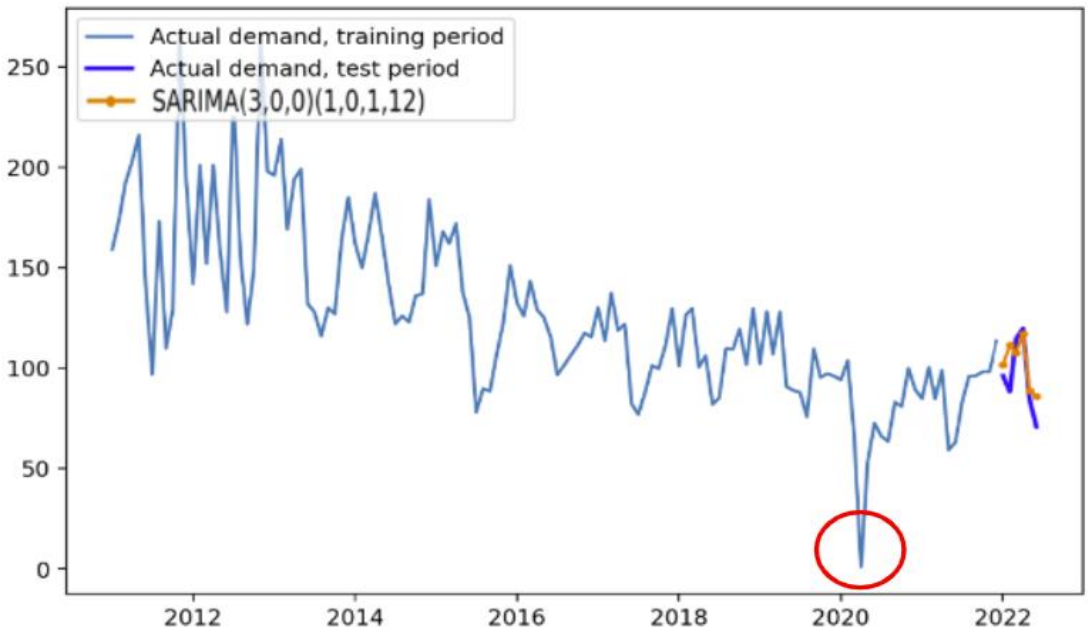


Category 9 is the only product category where ARIMA was the best model, meaning no clear seasonality was identified and the available exogenous factors didn't help explain the demand trends. In this instance, the Naive model yielded a MAPE of only 12.6%, indicating that neither model effectively captures the underlying variability. However, the ARIMA model demonstrated superior performance in capturing the moving average, particularly after accounting for the COVID-19-related demand fluctuations highlighted by the red encircled region, which were subsequently replaced. The resulting ARIMA (1,0,1) shows a degree of dependence on past values and moving average behavior, but does not show any significant trends or seasonality.

Lastly, we turn to Figure 13, which shows the time series data of Category 10 with MAPE 10.5% and AIC 1228.

Figure 13

Category 10 Forecast Results



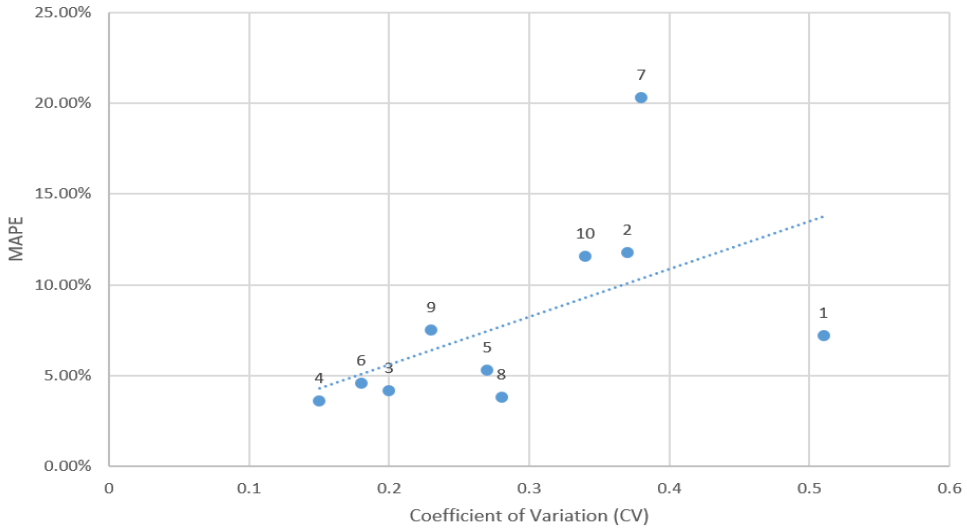
Category 10 exhibits a clear pattern of strong seasonality, with a noticeable spike in demand during dry winter months, particularly when taking into account fluctuations related to COVID-19. The most effective model for this category is SARIMA (3,0,0)(1,0,1,12), which aligns with the expected increase in demand during dry seasons for products falling under Category 10. This underscores the importance of accounting for seasonal variations and external factors such as COVID-19 in accurately forecasting demand for this product category.

4.6 Coefficient of Variation

While MAPE is a commonly used method to portray the accuracy of a forecast across different data sets, its main weakness is that it does not consider the variability in the data. Figure 4 shows the relationship between the coefficient of variation, which assesses variability in demand data by the ratio of the mean to the standard deviation, and the MAPE of the six-month forecast from January to June 2022.

Figure 14

Coefficient of Variation vs MAPE



Each dot represents one product category. The line of best fit shows an upward trend, confirming that higher variability in demand data leads to less accurate forecasts. We expect this is because less variability means that even a naive forecast would not deviate too far from the actual demand. There are some exceptions. Category 1 is well below the line of best fit, meaning the demand is highly variable but the MAPE is relatively strong. It is probable that in this instance, the model accurately predicted a consistent and stable seasonality pattern.

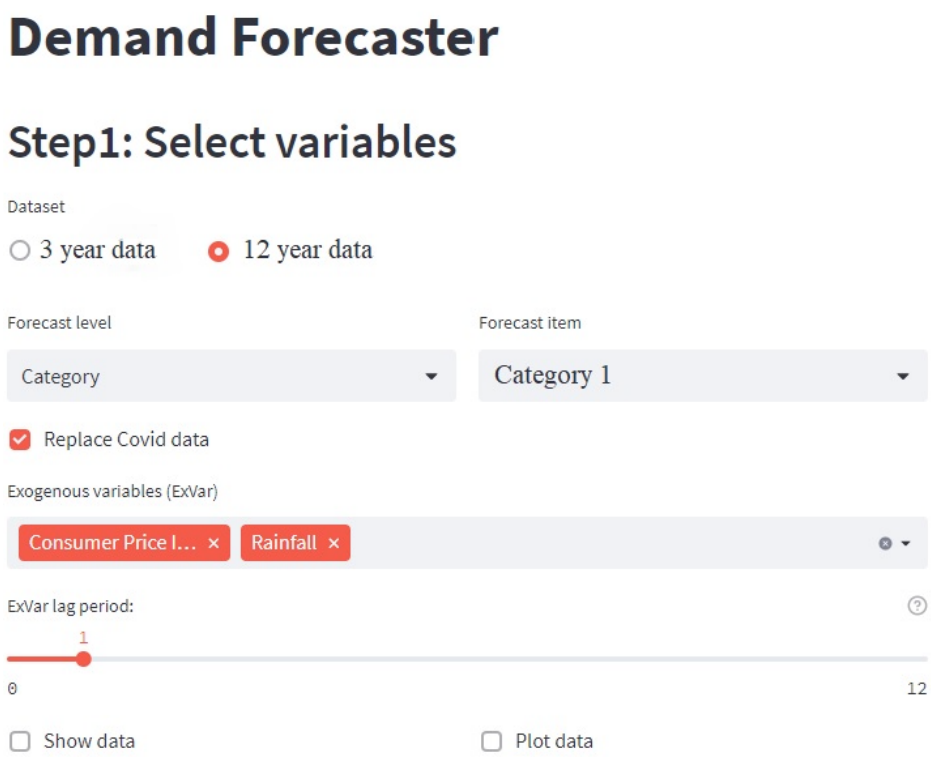
4.7 The *Demand Forecaster* Application

To best accommodate the sponsor company's diverse and fast-evolving demands, we developed a user-friendly application named *Demand Forecaster* that synthesized the findings and methods developed in Section 3. Demand Forecaster automatically executes model fitting and selecting for any category-, brand-, and sub-brand-level demand series, while allowing for the maximum degree of freedom in 1) the choices of data selection and transformation, 2) model restrictions, and 3) exogenous variables.

In this section, we walk you through the functionality of the application, with each step illustrated by screenshots. Figure 15 presents the initial selection of targets, categorized by different forecast levels.

Figure 15

Demand Forecaster Application

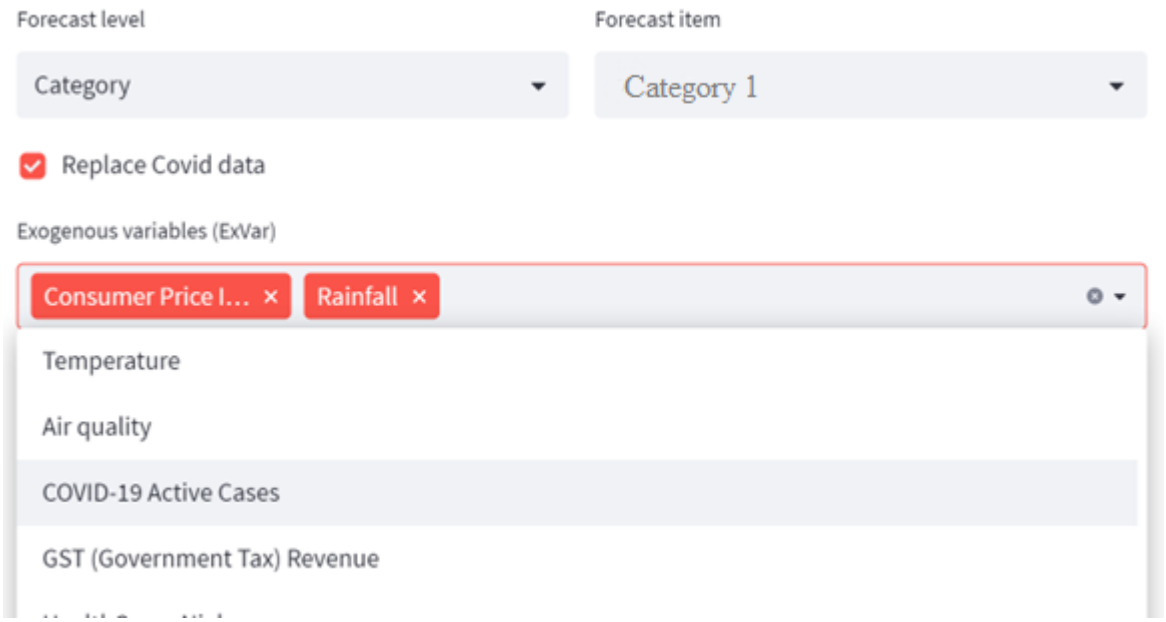


Forecast target: As shown in Figure 15, the application has a built-in data filter, which allows the user to make forecasts using either the 3-year or the 12-year dataset, and within each dataset the user also has the choice of selecting any item at category-, brand-, or sub-brand-level as the target forecast variable. A geographical filter is also included, offering the user the option of forecasting demand of the selected item in the entire country or a given state.

As illustrated in Figure 16, the application also incorporates a filter for exogenous variables.

Figure 16

Demand Forecaster Application - Exogenous Variables



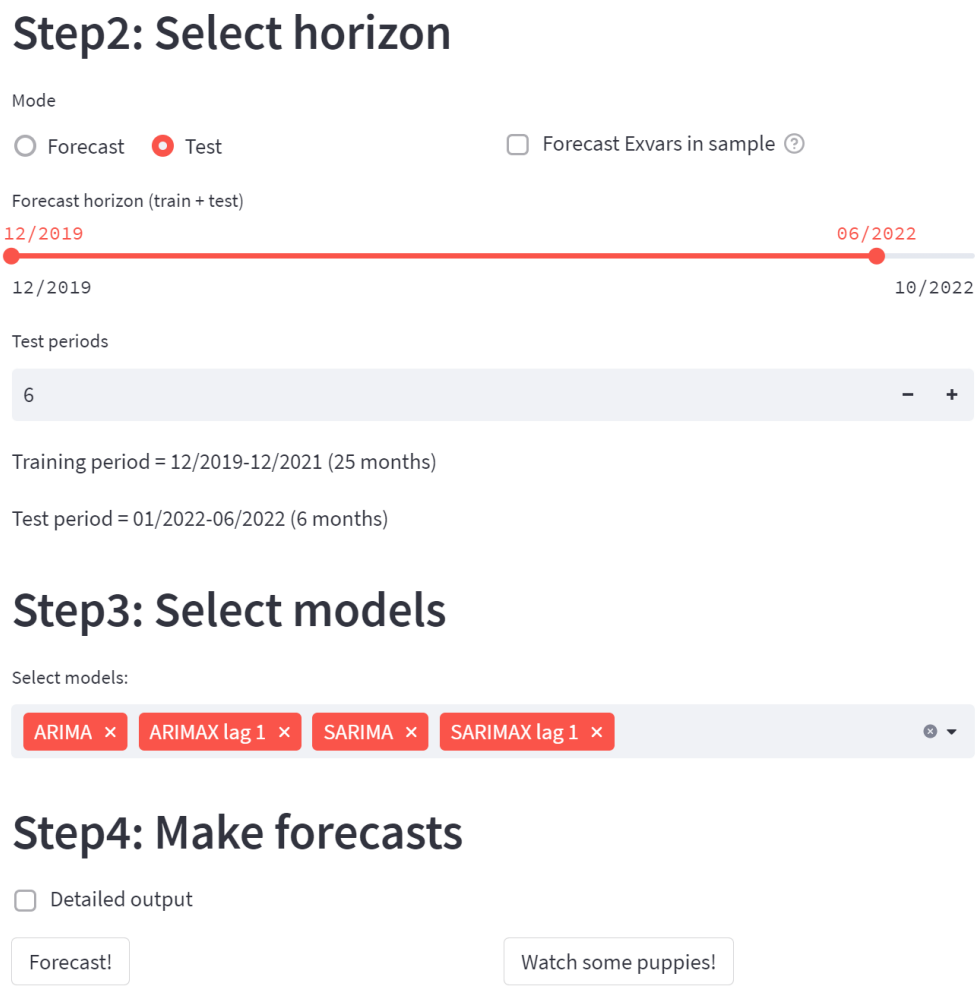
Exogenous variables: The user is allowed to pre-select the set of exogenous variables to be included into the forecast model. This selection is not necessary, as the model automatically assigns near-zero coefficients to variables not helpful for forecasting the target variable, but could be helpful if the user has and would like to use existing knowledge about exogenous variables.

Data preprocessing: Once forecast target and exogenous variables have been selected, the user is given the freedom to make changes to the data and command the application to perform different tasks with the data. This includes: indicate the lag length of exogenous variables, discard certain fractions of the data, choose between an in-sample model validation mode and an out-of-sample forecast mode, and important features. Graph and table representations of target and exogenous variables are also available with a single click.

In the final phase of using the application, users have the ability to choose both the forecast horizon and the model selection for generating the results.

Figure 17

Demand Forecaster Application - Model Selection

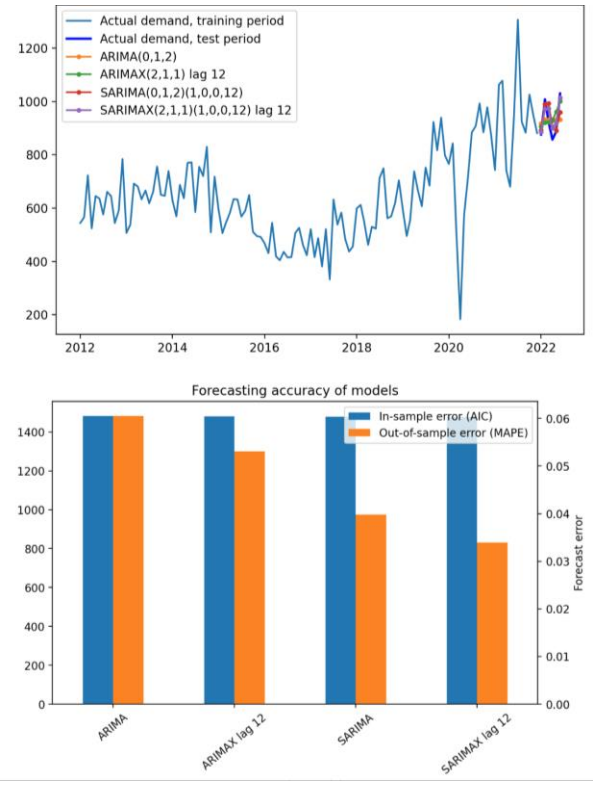


Model selection and forecast result: Users may pre-select, if desired, the set of candidate models. When all parameters are specified, the user can click the "Forecast" button which instructs the application to fit all candidate models, find the best model, and report results. Normally taking about 1 minute, the application returns the results in Figure 18: best model,

forecasted demand and its plot, forecast errors including AIC and MAPE (in-sample validation mode only), a comparison between the best model and best models of other families, and finally coefficients, standard errors, and p-values of the best model.

Figure 18

Demand Forecaster Application - Category 2 Forecast Results Summary



Summary of models

	AIC	MAPE	Accuracy
ARIMA(0,1,2)	1482.107307	6.05%	93.95%
ARIMAX(2,1,1) lag 12	1479.396225	5.31%	94.69%
SARIMA(0,1,2)(1,0,0,12)	1477.603619	3.98%	96.02%
SARIMAX(2,1,1)(1,0,0,12) lag 12	1476.663437	3.39%	96.61%

Summary of Key Forecasting Variables

	coef	std err	z	P>|z|	(.025	.975)
Consumer Price Index: 20	107.738700	97.385000	1.106000	0.269000	-83.133000	298.611000
Rainfall: All India: Actual	30.765800	13.114000	2.346000	0.019000	5.062000	56.470000
ar.L1	0.197100	0.102000	1.934000	0.053000	-0.003000	0.397000
ar.L2	-0.185200	0.106000	-1.745000	0.081000	-0.393000	0.023000
ma.L1	-0.781500	0.082000	-9.548000	0.000000	-0.942000	-0.621000
ar.S.L12	0.220800	0.100000	2.200000	0.028000	0.024000	0.418000
sigma2	12740.000000	1316.878000	9.671000	0.000000	10200.000000	15300.000000

	Impact	Significance
Consumer Price Index: 2012=100: India	5.429	73.10%
Rainfall: All India: Actual	0.323	98.10%
Demand of last month	0.1971	94.70%
Demand of m-2	-0.1852	91.90%
Forecast error of last month	-0.7815	100.00%
Demand of same month 1 year ago	0.2208	97.20%

The Demand Forecaster application is designed to be compatible with the programming language currently used by the sponsor company. This ensures a seamless integration with their existing systems and infrastructure, minimizing any potential disruptions or inefficiencies in the implementation process. By leveraging this technology, the application has the potential to enhance the sponsor company's demand forecasting capabilities, leading to improved productivity and profitability.

Furthermore, the application's modular code structure allows for easy updates and expansion of datasets as new data becomes available. This flexibility enables the sponsor company to adapt to changes in market conditions and consumer behavior, ensuring that their demand forecasting models remain accurate and effective over time. With the potential to serve as a cornerstone for a tailored time series analysis framework, the Demand Forecaster application has the capacity to drive future business research and strategic decision-making for the sponsor company.

5. DISCUSSION

The following sections interpret and explain the significance of the research findings: (1) managerial implications; and (2) limitations and future research.

5.1 Managerial Implications

The implications of forecasting research carry considerable weight for managerial decisions within organizations. It provides managers with invaluable insights into future industry or market trends, enabling them to make more informed decisions. Enhanced forecasting accuracy can elevate service levels, bolster margins, and optimize management of raw material suppliers, plant scheduling, and warehousing contracts. It can also help identify potential opportunities and threats, allocate resources more judiciously, and shape strategies to meet their objectives.

We strongly recommend the sponsoring company to invest in access to the Consumer Price Index (CPI) and rainfall exogenous variables from CEIC Data. Employing exogenous variables

like rainfall and CPI enhances forecasting accuracy in 60% of categories. In contrast, variables available for only three years, such as Nielsen Consumer Trends and Goods and Services Tax (GST), contribute less. Not all exogenous variables improve a model—particularly for products with inelastic demand. However, incorporating exogenous variables has been observed to enhance accuracy in 60% of categories, while replacing COVID-19 data from March to May 2020 with data from the previous year improved accuracy in 50% of categories. This approach was found to be more effective than considering the number of COVID-19 cases or lockdowns.

Models within the ARIMA family, despite their rigor, continue to dominate academic literature and applications for forecasting. In our study, these statistical models consistently outperformed machine learning models in terms of accuracy for this dataset. While machine learning techniques can capture more complexity, like multiple seasonality, they may not necessarily increase accuracy for data lacking such complexity. Moreover, machine learning models lack the transparency of ARIMA family models, which confirmed that seasonality enhances accuracy in 80% of categories.

Managers need to appreciate the limitations and uncertainties inherent in forecasting research. Forecasts are only as reliable as the assumptions and data they are based on, and unforeseen events can drastically affect their accuracy. For instance, complications with discontinued products or SKU reconfiguration can result in time series data inconsistencies. Therefore, data quality improvements could potentially enhance forecasts. Ultimately, higher variability in demand data often leads to less accurate forecasts.

Forecasts should be seen by managers as a tool for decision-making, not as an infallible guide to the future. It is prudent to consider multiple forecasts and scenarios and utilize them to inform decisions rather than relying solely on a single forecast. Managers should also periodically evaluate and adjust their forecasts to mirror changing market conditions and modify their strategies accordingly. We anticipate that the application we developed will enable managers to review these results biannually without significant computational burden. The sponsoring company might also integrate the results of the exogenous forecast at the category level as a factor in their SKU level forecast. Ultimately, the managerial implications of forecasting research hinge on how adeptly managers utilize it to inform their decisions and adapt to evolving circumstances.

5.2 Limitations and Future Research

Our exploration of exogenous factors was constrained by the project's scope and the readily accessible data. Other economic indicators or promotional information might potentially correlate more strongly with the demand data for certain product categories. Given the myriad permutations of exogenous variables that could be considered for each model, an automated process for their selection could yield a more optimal solution, alleviating the need for manual testing.

While our research primarily targeted product categories and national-level forecasting for the ensuing six months, the application can handle much finer levels of granularity. We managed to acquire demand data and exogenous variables at the product and regional levels for a three-year span. Rigorous analysis across these granular levels could potentially yield more

robust insights about the app's capabilities, the models utilized, and the involved variables. Heightened granularity could also introduce multiple seasonalities, thereby enhancing the value of machine learning models like XGBoost. In the future, the sponsor company's focus could expand beyond XGBoost to encompass other models, such as deep learning and neural networks.

The sponsoring company has also expressed interest in forecasting beyond a six-month horizon, extending up to 12 or 18 months into the future. Our preliminary tests on 12-month and regional-level forecasting returned slightly less accurate results, a finding that aligns with existing literature. Regardless, these results retain their validity and provide a solid foundation for future investigations.

Lastly, even though our demand forecasting model was designed with India in mind, it is conceivable to test it in other countries within the ISC or in other regions globally for the sponsor company. By doing so, we can assess the model's generalizability and pinpoint necessary adjustments for varying regions or product categories.

6. CONCLUSION

The Fast Moving Consumer Goods (FMCG) sector, the 4th largest in the Indian economy, operates on narrow margins, making even minute improvements in forecasting accuracy financially impactful. The long-term forecasts of the sponsoring company, historically guided by human judgment, were bereft of quantitative and exogenous variables. The COVID-19 pandemic introduced an unprecedented variance in demand trends, a factor mostly overlooked in extant literature. This capstone project explored ways to leverage both exogenous and internal variables for enhancing the accuracy of category-level demand forecasts over a six-month period.

Adopting a model-agnostic stance, we selected the most suitable model, exogenous variables, and approach to COVID-19 for each product category, based on the Mean Absolute Percentage Error (MAPE). Our experiments centered on the Autoregressive Integrated Moving Average (ARIMA) model, incorporating seasonality and exogenous variables, while also comparing its performance with machine learning methods.

The results of our research revealed a significant improvement in forecasting accuracy, rising from 72% in the naive model to 92% in our optimal model. On average, the Seasonal Autoregressive Integrated Moving Average model with exogenous variables (SARIMAX) outperformed others. Rainfall and Consumer Price Index emerged as the most influential exogenous factors. In addressing the COVID-19 impact, we found substituting dependent variable data from the initial three months of the pandemic with the equivalent data from the previous year more effective than using case counts or binary lockdown variables.

A key deliverable of this project is a fully operational application that allows the sponsoring company to easily compare forecast results across different lengths and product-level granularity, and to refine optimal models with updated data. Given the inherent uncertainty of longer-term predictions, the application ensures that near-future forecasts are more accurate. As the product catalog evolves over time, changes in seasonality and trends may necessitate model adjustments, an aspect the application seamlessly manages, requiring minimal computational resources.

Improved forecasting accuracy will enable the sponsor company to elevate service levels, enhance margins, and manage raw material suppliers, manufacturing plant scheduling, and

warehousing contracts more efficiently. This application, by reducing the effort to revise forecasts and promoting staff engagement with forecasting decisions, holds the potential to be adopted by other FMCG companies seeking to optimize their forecasting accuracy.

In conclusion, this capstone project offers a tangible and significant value proposition not only for the sponsor company but also for the wider FMCG sector. The enhanced forecasting model, supported by our intuitive application, represents a leap forward in operational efficiency. The sponsor company stands to benefit from improved margins, better resource management, and overall increased service levels - all achieved through the improved accuracy of demand forecasting. Moreover, our model's ability to adapt to new data and changing trends ensures its ongoing relevance and utility, making it a dynamic tool for strategic planning.

Beyond our sponsor, other FMCG companies could also leverage this application to improve their forecasting accuracy, thereby making more informed decisions and optimizing their operations. The potential to reduce the time and computational resources needed to revise forecasts, as well as to increase staff engagement in the decision-making process, could have far-reaching implications for the sector. In essence, the developments from this project provide a blueprint for FMCG companies looking to harness the power of data analytics and machine learning for superior forecasting, thus enhancing their competitive positioning in a low-margin, high-competition market.

REFERENCES

Abu Talib, M., Abdallah, M., Abdeljaber, A., & Abu Waraga, O. (2023). Influence of exogenous factors on water demand forecasting models during the COVID-19 period. *Engineering Applications of Artificial Intelligence, 117*, 105617. https://doi.org/10.1016/j.engappai.2022.105617

Andrews, B. H., Dean, M. D., Swain, R., & Cole, C. (2013). Building ARIMA and ARIMAX models for predicting long-term disability benefit application rates in the public/private sectors. *Society of Actuaries*, 1-54.

Banerjee, A. V. and Duflo, E. (2007). The Economic Lives of the Poor. *The Journal of Economic Perspectives, 21* (1), pp. 141-167

Booth, E., Mount, J., & Viers, J.H. (2006). Hydrologic Variability of the Cosumnes River Floodplain. *San Francisco Estuary and Watershed Science*, Volume 4, Issue 2.

Box, G.E.P, Jenkins, G.M., & Reinsel, G.C. (2008). *Time Series Analysis: Forecasting and Control*. John Wiley & Sons, Inc., 4th edition.

CEIC Data. (2023). *Global Economic Data, indicators, charts & forecasts.* https://www.ceicdata.com/en

Croston, J. D. (1972). Forecasting and stock control for intermittent demands. *Journal of the Operational Research Society*, 23(3), 289-303.

Dachyar, F.F, M., Taurina, Z., & Qaradhawi, Y. (2021). Disclosing fast moving consumer goods demand forecasting predictor using multi linear regression. *Engineering and Applied Science Research, 48*(5), 627–636. Retrieved from https://ph01.tci-thaijo.org/index.php/easr/article/view/242407

Gumus, M. and Kiran, M.S. (2017). Crude oil price forecasting using XGBoost. *International Conference on Computer Science and Engineering (UBMK), Antalya, Turkey,* pp. 1100-1103, doi: 10.1109/UBMK.2017.8093500.

Guo, Z.X., Wong, W.K., Li, M. (2013). A multivariate intelligent decision-making model for retail sales forecasting. *Decision Support Systems*, v. 55, n. 1, 247-255.

Holt, C. C. (1957). Forecasting seasonals and trends by exponentially weighted moving averages. *International Journal of Forecasting*, 20(1), 5-10.

Hyndman, R.J., & Athanasopoulos, G. (2018). *Forecasting: Principles and Practice, 2nd edition*. OTexts: Melbourne, Australia.

Indian Brand Equity Foundation. (2022). *Indian FMCG Industry Analysis.* https://www.ibef.org/industry/Fmcg-presentation

Karn, S. K., Shikura, S. and Harada, H. (2003). Living Environment and Health of Urban Poor: A Study in Mumbai. *Economic and Political Weekly, 38* (34), pp. 3575-3577, 3579-3586

Kotler, P., & Armstrong, G. (2013). *Principles of Marketing*. Pearson Education Limited, Harlow, England.

Kumar, A., Dangi, H., & Vohra, A. (2015). The Purchase Pattern of Poor for Fast Moving Consumer Goods: An Empirical Study of Poor in India. *International Journal of Management and Business Research, 5*(2), 79-94.

Kumar, R. (2019). IJU for the Indian subcontinent. *Indian Journal of Urology : IJU : journal of the Urological Society of India, 35*(1), 1. https://doi.org/10.4103/iju.IJU_372_18

Liu, Y., Li, M., & Zhu, Z. (2013). Simulated Annealing Sales Combining Forecast in FMCG. *2013 IEEE 10th International Conference on E-Business Engineering*. https://doi.org/10.1109/icebe.2013.35

Mahajan, Y. (2020). Impact of Coronavirus Pandemic on Fast Moving Consumer Goods (FMCG) Sector in India (September 05, 2020). *Journal of Xi'an University of Architecture & Technology, Volume XII*, Issue IX, 2020 http://xajzkjdx.cn/gallery/5-sep2020.pdf, Available at SSRN: https://ssrn.com/abstract=3710329

Mall, A., Ramesh, R., Sanghi, K., Singhi, A., Subramanian, A. (2012). The Tiger Roars: An In-depth Analysis of How a Billion Plus People Consume. *International Business.* Boston Consulting Group.

Mishra, D. P. (2008). FMCG distribution channels in India: Challenges and opportunities for manufacturers and retailers. *Journal of Global Business Issues*, *2*(2), 175.

Nau, R. (2016). *Statistical Forecasting: Notes on Regression and Time Series Analysis.* Fuqua School of Business, Duke University, Durham.

NielsenIQ. (2023). *NielsenIQ Insights*. https://nielseniq.com/global/en/

Nyaga, J. (2014). Factors affecting distribution of fast moving consumer goods in Kenya: A case of Eveready East Africa. *International Journal of Social Sciences and Entrepreneurship, 1* (12), 290-302.

Peixeiro, M. (2022). *Time Series Forecasting in Python*. Manning Publications Co.

Ramanuj, M. (2007). *Product Management in India, Third Edition*. PHI Learning.

Sean, B. (2002). *The Advertising Handbook, 2nd Edition.* Routledge.

Shapiro, B. P. (1973). Price Reliance: Existence and Sources. *Journal of Marketing Research, 10* (3), pp. 286-294

Shetty, G., Nougarahiya, S., Mandloi, D., & Sarsodia, T. (2020). COVID-19 and Global Commerce:

An Analysis of FMCG, and Retail Industries of Tomorrow. *SSRN Electronic Journal*. https://doi.org/10.2139/ssrn.3603028

Stanciu, S., Vîrlănuţă, F. O., Vochin, O. A., Ionescu, R. V., & Antohi, V. M. (2019). Fast Moving Consumer Goods (FMCG) Market In Romania Features And Trends. *Amfiteatru Economic, 21*(13), 778-794.

Subramanian, A. & Felman, J. (2019, December). India's Great Slowdown: What Happened? What's the Way Out?. *Harvard University.* https://www.hks.harvard.edu/sites/default/files/centers/cid/files/publications/facultyworking-papers/2019-12-cid-wp-369-indian-growth-diagnosis-remedies-final.pdf

Tarallo, E., Akabane, G. K., Shimabukuro, C. I., Mello, J., & Amancio, D. (2019). Machine Learning in Predicting Demand for Fast-Moving Consumer Goods: An Exploratory Research. *IFAC-PapersOnLine*, *52*(13), 737–742. https://doi.org/10.1016/j.ifacol.2019.11.203

Triebe, O., Hewamalage, H., Pilyugina, P., Laptev, N., Bergmeir, C., & Rajagopal, R. (2021). Neuralprophet: Explainable forecasting at scale. *arXiv preprint arXiv:2111.15397*.

Winters, P. R. (1960). Forecasting sales by exponentially weighted moving averages. *Management Science*, 6(3), 324-342.

Yang, D., Goh, G. S. W., Jiang, S., Zhang, A. N., & Akcan, O. (2015). Forecast UPC-level FMCG demand, Part II: Hierarchical reconciliation. *2015 IEEE International Conference on Big Data (Big Data)*. https://doi.org/10.1109/bigdata.2015.7363994

APPENDIX A

Python Code for Model Comparison

```
def naive(df, category, months):
    y_test = df[category][-months:]
    shift = df[category].shift(6)
    preds = shift[-months:]
    mape = round(mean_absolute_percentage_error(y_test, preds), 3)
    return mape

def arima(y_train, y_test):
    model = pm.auto_arima(y_train , X=None, test='adf', seasonal=False)
    preds = model.predict(n_periods=len(y_test))
    mape = round(mean_absolute_percentage_error(y_test, preds), 3)
    return mape, model.order

def arimax(y_train, x_train, y_test, x_test):
    model = pm.auto_arima(y_train, X=x_train, test='adf', seasonal=False)
    preds = model.predict(n_periods=len(y_test), X=x_test)
    mape = round(mean_absolute_percentage_error(y_test, preds), 3)
    return mape, model.order

def sarima(y_train, y_test):
    model = pm.auto_arima(y_train, X=None, test='adf',seasonal=True, m=12)
    preds = model.predict(n_periods=len(y_test))
    mape = round(mean_absolute_percentage_error(y_test, preds), 3)
    return mape, model.order, model.seasonal_order

def sarimax(y_train, x_train, y_test, x_test):
    model = pm.auto_arima(y_train, X=x_train, test = 'adf', seasonal=True, m=12)
    preds = model.predict(n_periods=len(y_test), X=x_test)
    mape = round(mean_absolute_percentage_error(y_test, preds), 3)
    return mape, model.order, model.seasonal_order, preds

def xgb(y_train, x_train, y_test, x_test):
    model = xgboost.XGBRegressor(max_depth=3, n_estimators=50)
    model.fit(x_train, y_train)
    preds = model.predict(x_test)
    mape = round(mean_absolute_percentage_error(y_test, preds), 3)
    return mape
```