# Using Machine Learning Approaches to Improve Long-Range Demand Forecasting

by

Katherine Gail Nowadly
B.S., Industrial Engineering, Pennsylvania State University, 2016

and

Sohyun Jung
B.B.A, Yonsei University, 2007

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

May 2020

Signature of Author: ____________________________________________
Department of Supply Chain Management
May 8, 2020

Signature of Author: ____________________________________________
Department of Supply Chain Management
May 8, 2020

Certified by: ____________________________________________
Tugba Efendigil, PhD
Research Scientist
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

Using Machine Learning Approaches to Improve Long-Range Demand Forecasting

by

Katherine Gail Nowadly
and
Sohyun Jung

Submitted to the Program in Supply Chain Management
on May 8, 2020 in Partial Fulfillment of the
Requirements for the Degree of Master of Applied Science in Supply Chain Management

## ABSTRACT

Achieving an accurate long-range forecast is a challenge many companies face due to the uncertainty in anticipating demand several years out. Since companies make strategic decisions based on these forecasts – such as long-term investments and supply and capacity planning – it is critical that the long-range forecast be as accurate as possible. For a large player in the pharmaceutical industry like our capstone project sponsor, an improvement in its forecasting process could have significant financial and organizational benefits. While traditional statistical methods have been extensively used in demand forecasting, due to technological developments, machine learning approaches have been widely studied and increasingly applied in forecasting. Since machine learning has shown improvements in forecasting, especially in short-term forecasting, could machine learning be applied to improve long-range forecasting? This study explores this question by testing several machine learning methodologies and approaches. First, we used support vector machine (SVM) to determine relevant features. Next, we used two types of approaches, direct and recursive, to develop one-step and multi-step long-range forecasting models. We developed forecasts using four machine learning algorithms: random forest (RF), artificial neural network (ANN), linear regression (LR), and support vector machine (SVM). We found that RF, ANN, and LR produced relatively accurate results in the one-step models. However, when extending the forecasting horizon using a multi-step forecast, the accuracy declines. By observing the results of the feature selection process and comparing the results among our forecasting models, we determined which features are critical when forecasting long-range demand for certain drugs. Additionally, we found that the machine learning model performance differed greatly based on data availability, forecasting horizon, and individual product. The biggest challenge in pursuing the application of machine learning approaches in long-range demand forecasting is data management. Given that using a machine learning approach in long-term forecasting has inconclusive performance, and creating a data management program would require a large up-front investment, a detailed cost-benefit analysis along with internal discussion is advised before pursuing further applications of machine learning in long-range demand forecasting.

Capstone Advisor: Tugba Efendigil, PhD
Title: Research Scientist

## Acknowledgements

We would like to thank several individuals that have been instrumental in this project. First, we give thanks our advisor, Dr. Tugba Efendigil, for providing guidance and support throughout the year. We also would like to thank members of the MIT Supply Chain Management program, notably Dr. Josué C. Velázquez Martínez for his mentorship and leadership, Dr. Chris Caplice for his forecasting expertise, and Dr. Sergio Caballero for teaching us the technical skills required for this project. Additionally, we would like to thank our sponsor company for giving us this opportunity and being engaged and supportive throughout this project.

- Katie & Sohyun

Thank you, Sohyun, for being a great partner for this project. I am grateful for being paired with such an intelligent and hardworking partner. I would also like to thank my classmates who have been supportive during my time at MIT. Last but not least, thank you to my friends and family, and especially my dog.

- Katie

I would like to express my gratitude to my capstone partner Katie for being a great partner and to my friends and family who have supported me throughout this journey.

- Sohyun

# Table of Contents

## List of Figures

## List of Tables

# 1 Introduction

This capstone explores the use of machine learning (ML) applications on long-range[1] demand forecasting for the sponsoring company, a large pharmaceutical manufacturer. The goal of the project is to see if and how machine learning could improve the overall long-range demand forecasting process. Throughout the report, we discuss the research gaps, key challenges, and the results of testing several machine learning methodologies on the data provided by the sponsoring company.

## 1.1 Motivation

Companies use long-range forecasting to estimate the projected demand for the next 3+ years. Since companies make strategic decisions based on these forecasts – such as long-term investments and supply and capacity planning – it is critical that the long-range forecast be as accurate as possible. If a forecast underestimates demand, the company might not have the adequate capacity and resources to meet the needs of its customers. If a forecast overestimates demand, the company incurs the cost of excess capacity and waste. Since neither of these outcomes is ideal, it is important to establish an accurate long-range forecasting process.

Achieving an accurate long-range forecast is a challenge many companies face due to the uncertainty in anticipating demand several years out. Many factors impact long-range forecasts, such as market conditions, new product launches, and supply considerations. In the case of the pharmaceutical industry, long lead times from

[1] “Long-range” is defined as 3+ years throughout this document

suppliers compound the challenge of aligning long-range forecasts to current demand. As a large player in the pharmaceutical industry, the sponsor company seeks to explore new methodologies to improve its long-range forecasting process.

## 1.2 Problem Statement

The purpose of this capstone project is to test the feasibility of applying machine learning, defined as the use of algorithms and statistics to detect patterns, to develop long-range demand forecasts in the pharmaceutical industry. Long-range forecasting is typically used for various strategic decision-making purposes. Many forecasting methods are used to develop long-range forecasts. Traditional methods include qualitative methods such as Delphi and panel consensus, as well as quantitative methods such as regression, time-series, econometric models, and diffusion index. Due to technological developments, in addition to the traditional forecasting models, machine learning approaches have been widely studied and applied in forecasting. Since machine learning has shown improvements in forecasting, especially in short-term forecasting, we find it worthwhile to study the application of machine learning in long-range forecasting (Caplice, 2019).

Accurate forecasts are needed in virtually all aspects of business planning (e.g., capital investment, product development, and supply chain design). While research has been conducted on applying machine learning to short-range forecasting (0-3 years), more research on how to apply machine learning to long-range forecasting needs to be done. Since many business units depend on long-range forecasts, this report will be meaningful to any company considering the use of machine learning in long-term demand forecasting for its business.

### 1.3 Methodology Summary

To understand the nature of long-range forecasting, we researched traditional long-range forecasting methodologies such as econometric modeling and time series models. We studied the limitations of these traditional models and whether these shortfalls can be improved with machine learning. Additionally, we reviewed the sponsor company's current long-range forecasting models, which were Excel-based econometric models. We identified the methodology of the model, data types, variables, and the key assumptions in modeling. Since a large amount of data is needed to build a machine learning model, we also identified other data sources to incorporate into our machine learning model.

Several different methodologies for machine learning are applicable to this project. To be able to find the most effective one for long-range forecasting, we reviewed the models used in other studies (Ekonomou, 2010; Hong, 2009; Liu, Ang, & Goh, 1991; Mohamed & Bodger, 2005). Long-range forecasting methodologies were actively studied in capital-intensive industries where companies need to make large investment decisions, such as the energy industry. Long-range forecasting was also actively performed in sociology, where researchers predicted social trends and change in society. To investigate the most effective machine learning forecasting methods, our study reviewed not only demand forecasting studies but also social trend forecasting and sales forecasting studies. Machine learning methodologies from these researches included random forest algorithm (RF) and artificial neural network model (ANN). We identified the data type, indicator, result, and methodology for each of the forecasting projects and studied whether the same methodology can be applied to forecast the target variable for this model. This process allowed us to identify the opportunities and challenges of applying machine learning in long-term demand forecasting and to suggest a further direction for improving the sponsor company's demand forecasting.

### 1.4 Industry Background

The sponsor company is a global pharmaceutical company supplying pharmaceuticals, vaccines, and consumer products across the world. The pharmaceutical industry faces unique challenges in forecasting and planning due to the regulatory nature of the industry. Regulations vary among countries, with constraints regarding which products can be sold in each country. The pharmaceutical supply chain needs full traceability to ensure that the final product, and the ingredients within the product, are approved in each market where they are sold. These conditions and other complexities make it particular difficult to collate market, regional and therapy area forecast and translate to long range forecast. With the help of machine learning, the sponsor company can incorporate these considerations into their long-range forecasting process.

## 2 Literature Review

In order to understand how machine learning can be used to improve long-range forecasting, we researched several forecasting techniques that have been developed to solve diverse forecasting problems. Since each forecasting problem has its own purpose, challenge, and application, finding the right methodology is crucial to have an accurate performance. In this section, we provide an overview of diverse forecasting methods, especially for long-range demand forecasting. We explore the traditional, statistical approaches and the algorithmic approaches that have been developed in recent years. Additionally, we review examples of demand forecasting and long-range forecasting methods (beyond 3-year horizon) that have been proved to be effective. Later, we further study some of the real industry application of long-range forecasting examples in diverse cases focusing on methodologies, variables, and performances.

## 2.1 Demand Forecasting Methods

Forecasting methods fall into two categories: subjective and objective. Subjective forecasting methods can be further subdivided into judgmental and experimental. These forecasting methods are used when historical data is not available, as in the case of a firm launching a new product. Objective forecasting methods are either causal or time-series approaches. In causal methods, there is a relationship between a set of variables, which is then used to formulate a forecast. Time series methods are used when there is a pattern in the data. Figure 1 summarizes the various forecasting methods (Caplice, 2019).

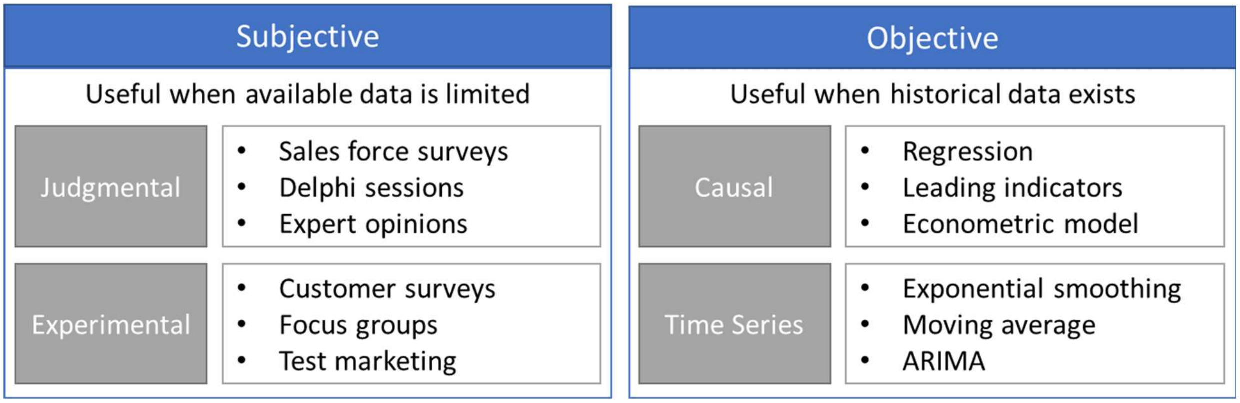
Subjective approaches are typically used for marketing and sales forecasts when there is limited data. Judgmental forecasting uses several methodologies, such as the Delphi method, where the opinions of experts are used to predict the future. Additionally, companies can use experimental methods, such as taking a local sample and extrapolating the results, to predict sales of a new product.

Contrary to a subjective approach, various objective statistical approaches such as time series analysis and causal analysis use historical data to predict the future. Time series analysis identifies seasonality, cycle, and trend in the data and assumes that in the future the prediction will show a similar pattern. Time series has been widely used in various industries and has shown a high degree of accuracy, especially in short-term forecasting. This includes moving average, exponential smoothing, and autoregressive integrated moving average (ARIMA) models.

The causal forecasting method is for when there is historical data and a causal relationship between the external factors that are considered to cause the event and the event itself. A causal approach is widely used not only in business but also in the social sciences. Regression models and econometric models can be defined in this approach. Figure 1 summarizes the forecasting methods explained in this section.

**Figure 1**

*Summary of forecasting methods and examples (Caplice, 2019)*



Subjective

Useful when available data is limited

Judgmental
- Sales force surveys
- Delphi sessions
- Expert opinions

Experimental
- Customer surveys
- Focus groups
- Test marketing

Objective

Useful when historical data exists

Causal
- Regression
- Leading indicators
- Econometric model

Time Series
- Exponential smoothing
- Moving average
- ARIMA

Over the last few decades, many of the methodologies mentioned in Figure 1 have been tested for demand forecasting. For the data showing a trend and seasonality component, traditional forecasting techniques such as moving average, Holt method, Winters exponential smoothing, Box-Jenkins ARIMA model, and multivariate regression methodologies have been proposed and widely used in various industries. Peterson (1993) showed that large retailers are more likely to use time-series methods for industry forecasts. However, in recent years, algorithmic approaches in demand forecasting have gathered a lot of attention; comparing their performance with traditional forecasting methods has become a major focus of research in demand forecasting.

Alon, Qi, and Sadowski (2001) compared artificial neural networks (ANN) and traditional methods including Winters exponential smoothing, Box–Jenkins ARIMA model, and multivariate regression for US retail sales forecasting. The results indicated that ANNs fare favorably in relation to the more traditional statistical methods on average, followed by the Box–Jenkins mode. Hribar, Potočnik, Šilc, & Papa (2019) studied ANN methodologies along with linear regression and kernel machine to forecast gas consumption demand and found recurrent neural network provides the most accurate result. Nonetheless, an algorithmic approach does not always show better performance than traditional statistical models. For example, in the study of Carbonneau, Laframboise, & Vahidov. (2008), they compared the forecasting result between machine learning technique such as neural networks, recurrent neural networks, and support vector machines and traditional ones such as naïve forecasting, trend, moving average, and linear regression with the monthly sales data. In this study, even though recurrent neural networks and support vector machines showed high accuracy in forecasting, their forecasting accuracy was not statistically significantly better than that of the regression model.

## 2.2 Forecasting Methods in the Pharmaceutical Industry

The pharmaceutical supply chain is unique in that the products are highly complex and heavily regulated, and there can be catastrophic consequences if there is a shortage. Pharmaceutical drugs are formulated by combining the Active Pharmaceutical Ingredient (API) with other additives. APIs often have long lead times, sometimes a year or more, as they have to go through the chemical reaction process. This adds to the complexity of pharmaceutical long-range forecast planning. Further contributing to this complexity are factors such as patents expiring and generic drugs entering the market, limited suppliers due to the regulatory nature of the industry, and high stockout costs. These reasons contribute to the difficulty and inefficiencies in the long-range forecasting process (Merkuryeva, Valberga, & Smirnov, 2019).

As discussed in the forecasting methods section 2.1, there are two approaches to forecasting: subjective and objective. In the pharmaceutical industry, time-series models are used most often (52%) and causal models account for 24%, while judgmental forecasts account for 19% and the remaining 5% represent mixed or combined models (Merkuryeva, Valberga, & Smirnov, 2019).

Historically, the pharmaceutical industry has been hesitant to make changes regarding its forecasting process unless the monetary value is apparent and significant (Faggella, n.d.). With Artificial Intelligence (AI) and ML becoming more ubiquitous in the healthcare industry, more sophisticated forecasting methods have emerged. Companies use system dynamics modeling to study disease progression and create feedback loops that influence demand forecasts. Simulation and visualizations can provide valuable insight to decision makers. Despite these new technologies, many companies still prefer a mix of objective and subjective approaches (Merkuryeva, Valberga, & Smirnov, 2019).

## 2.3 Long-range Forecasting Methods

Compared to the development of forecasting techniques in short-term forecasting, there has not been much investigation of long-range forecasting techniques. It is difficult to achieve a high degree of accuracy in forecasting as well, since forecasting starts with the assumption that the current patterns in the forecasting model will continue in the future, but there is a high possibility that the pattern would change because of numerous unexpected events over a longer period of time. This uncertainty should be considered in the model, but the high level of uncertainty is difficult for a model to capture.

Even though it is challenging to forecast for a long-term horizon, there has been a wide range of studies in long-range forecasting in the energy industry. Having an accurate long-term forecast is important in the energy industry since it becomes the foundation that guides the decision-making process for capital investments. Since capital investments are planned years ahead to ensure a sufficient supply of energy to communities, these decisions are critical for the sustainable development of the economy and society. Many types of methodologies have been tested for long-term electricity demand forecasting. Below are a few examples of long-range forecasting studies in the energy industry (see Table 1 for a summary):

1. Mohamed and Bodger (2005) proposed multiple linear regression models to forecast electricity consumption of New Zealand. It was found that one of the models showed comparable results with other existing forecasts. However, the accuracy of this regression model relied on the accuracy of forecasts made for the dependent variables, which makes forecasting accuracy vulnerable to error. Similar to short-term forecasting, algorithmic approach has been widely tested in long-term as well.
2. Liu, Ang, & Goh (1991) made a comparison between the econometric model and neural network model, both of which are formulated as a causal model. The study shows that econometric model showed better forecasting accuracy than the neural network model because of high elasticities of the model.

3. Ekonomou (2010) also suggested using an artificial neural networks (ANN) model and a multilayer perceptron model (MLP). The models' performance was compared with the results of a linear regression method, a support vector machine method, and the actual recorded data. The new models had greater accuracy.
4. In the study by Hong (2009), support vector models were suggested to minimize the errors and when compared with ANN, it showed better long-term forecast.

Through the literature review, we found that no forecast method is always right for every circumstance. Among multiple methodologies available for demand forecasting, the selected forecasting model should fit well with the task considering data type, data availability, data horizon, number of factors that affect the result and character of the trend itself.

**Table 1**

*Long-term demand forecasting in the energy industry*

| Title | Methodology | Variables | Target Variable | Time Horizon |
|---|---|---|---|---|
| Forecasting of electricity consumption: a comparison between an econometric model and a neural network model **(**Liu, Ang, & Goh ,1991) | Econometric/ Neural Networks | Gross domestic product (GDP), real electricity prices, population (1960-1984) | Annual total electricity Consumption | 1-5 years |
| Electric load forecasting by support vector model (Hong, 2009) | Support vector regression (SVR), VR model with IA (SVRIA) | Historical annual load demand data (1981-1992) | Electric load | 1-4 years |
| Forecasting electricity consumption in New Zealand using economic and demographic variables (Zaid Mohamed, Pat Bodger, 2005) | Multiple linear regression | Gross domestic product (GDP), real electricity prices, population (1965-1999) | Annual total Electricity consumption | 1-15 years |
| Greek long-term energy consumption prediction using artificial neural networks (Ekonomou, 2010) | Multilayer perceptron model (MLP)/ linear regression method/ support vector machine method | Ambient temperature, Power capacity, Electricity consumption, GDP (1992–2004) | Energy Consumption | 1-10 years |

## 2.4 Machine Learning Applications in the Pharmaceutical Industry

While there has been research on using AI to improve long-range forecasting for the energy industry, we wanted to look at how the healthcare industry and the pharmaceutical industry are currently using AI and ML to support their business. The healthcare industry uses ML in various applications, including “for discovery and biomarker identification; within pharmacovigilance activities, including adverse event case processing or gathering regulatory intelligence; and with real-world data, which can involve the use of large data sets of claims or electronic health data" (Lamberti et al., 2019, p 1415).

Additionally, ML has been especially prevalent in epidemiology. ML models use large amounts of data from a variety of sources such as social media and satellites to monitor epidemic outbreaks. For example, artificial neural networks are used to predict malaria outbreaks by taking into consideration environmental factors such as temperature and rainfall, as well as tracking the number of patients with hospital data (Faggella, n.d.). It is evident that ML is utilized within the healthcare industry, however, using ML to improve long-range forecasting is still a relatively unexplored frontier.

# 3 Data and Methodology

This study focuses on finding applicable machine learning methodologies for long-range demand forecasting for the sponsor company. This section explains the origins, preparation, and analysis of data collected internally and externally. Additionally, we discuss the methodologies used for variable selection and forecasting.

First, we reviewed the data the sponsor company provided. Since the methodology used in a forecasting model depends on the type of variables, dataset, and forecasting purpose, we decided to test the feasibility of machine learning methodology for one specific model, the model used to forecast a subset of products within the US market. This model had the most accessible data available for our purposes.

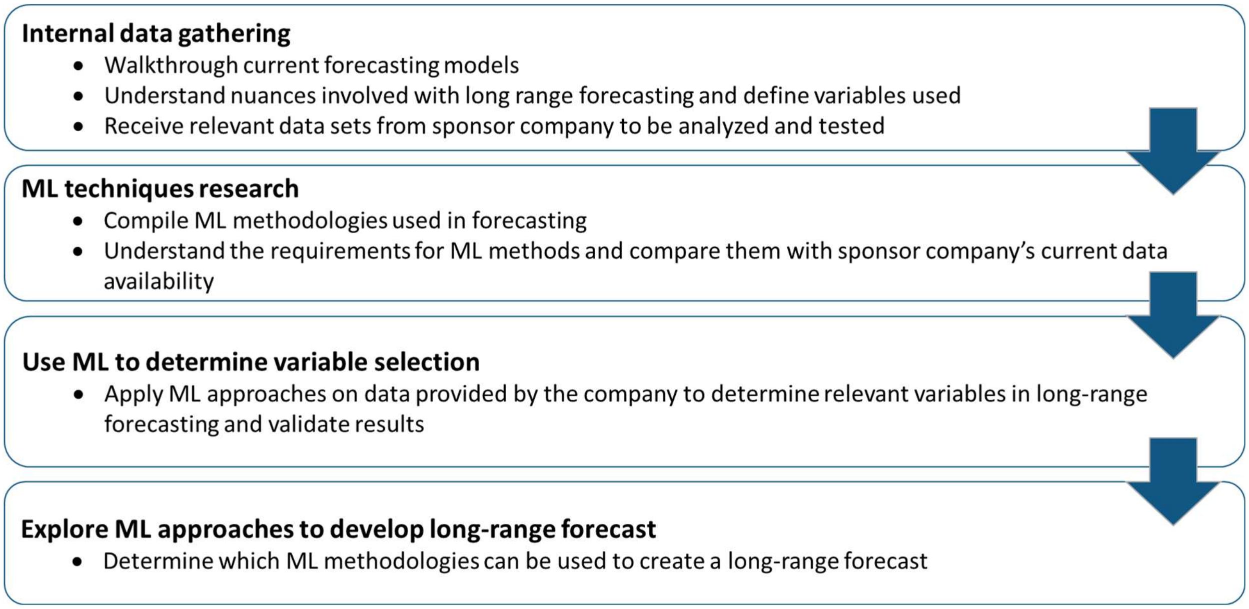
Next, we studied general machine learning methodologies for forecasting. After investigating some of the prevalent machine learning methodologies used to improve forecasting practices, such as support vector machine (SVM), decision tree, artificial neural network (ANN), and k-nearest neighbors, we identified which methodologies to test on the sponsor company's data.

After compiling the data, variables were standardized and only the most relevant features were selected through the feature selection method. Two different forecasting approaches were tested: direct modeling and recursive modeling. Direct modeling only uses historical data to forecast future time, whereas recursive modeling makes a prediction for a one-time period and feeds that prediction into the model as an input in order to predict the subsequent time steps. Direct and recursive modeling approaches were tested using four different machine learning methodologies: random forest (RAN), support vector machine (SVM), artificial neural network (ANN), and linear regression (LR).

By comparing the existing research on long-range forecasting using machine learning, the current forecasting model, and the underlying data used in the current model, we determined how machine learning could be used to improve long-range forecasting for the sponsor company. Figure 2 shows a summary of our approach.

**Figure 2**

*Summary of project approach*



**Internal data gathering**

- Walkthrough current forecasting models
- Understand nuances involved with long range forecasting and define variables used
- Receive relevant data sets from sponsor company to be analyzed and tested

**ML techniques research**

- Compile ML methodologies used in forecasting
- Understand the requirements for ML methods and compare them with sponsor company’s current data availability

**Use ML to determine variable selection**

- Apply ML approaches on data provided by the company to determine relevant variables in long-range forecasting and validate results

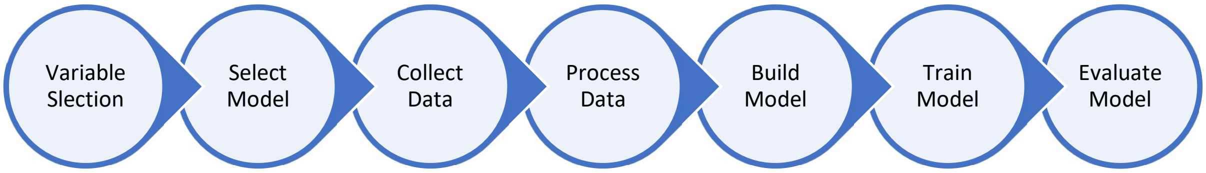
**Explore ML approaches to develop long-range forecast**

- Determine which ML methodologies can be used to create a long-range forecast

The machine learning forecasting modeling process follows the steps shown in Figure 3. The modeling process can be customized to meet the needs of specific situations and purposes.

**Figure 3**

*Machine learning forecasting modeling process*



## 3.1 Current Model

The sponsor company's current demand forecasting model is constructed with three main steps, as shown in Figure 4. It consists of baseline modeling, nonretail factor-up[2], and effects of events. Baseline modeling uses TRx[3] data as the main source for forecasting. As the first step, the model forecasts the total prescription quantity for each category according to the market projection. Using the linear regression methodology, it then forecasts the market share trend of each product. In this procedure, the forecasters may manually adjust the period of time used for linear regression according to their own discretion. By multiplying the market share by the total prescription quantity of a category, the model forecasts the prescription quantity for each product. Since many

[2] Nonretail factor-up is a method used by the sponsor company to transform historical data to better represent reality

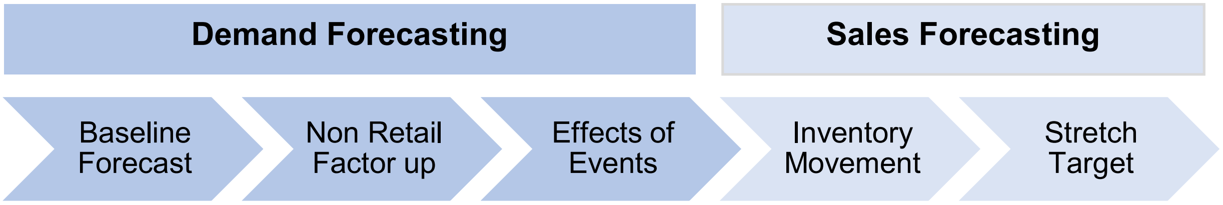
[3] TRx is an abbreviation for total prescriptions, which includes quantity of new and refilled prescriptions

events in the market, such as launch of a new product or generic entry, can influence product demand, the forecasting division then collects the opinions of experts in the sales and marketing division and applies corrections representing the expected effects of events to the model.

As shown in Figure 4, the sponsor company adds two more steps in the sales forecasting process after the three-step demand forecast process. The sponsor company adjusts the gross sales forecast according to the channel inventory analysis, or when there is a corporate sales target for a specific brand, the company stretches the sales target accordingly.

**Figure 4**

*The sponsor company's current forecasting model*



## 3.2 Data Preparation

### 3.2.1 Data Collection

The sponsor company provided their current weekly demand forecasting model which reflects weekly demand history. The company also provided product features data for their different products. This research also used external census data. Table 2 describes the data used in this research.

**Table 2**

*Data sources used in the machine learning models*

| TRx Data | Product Features Data | Census Data |
|---|---|---|
| • Product Name<br>• Product Category<br>• Weekly Prescription Quantity<br>• Date | • Product Name<br>• Product Launch Year<br>• Type of Device<br>• Type of Disease<br>• Number of Products in Category | • Number of Asthma Patients<br>• US Prescription Drug Consumption Trends |

#### 3.2.1.1 Total Prescription (TRx)

TRx represents the total quantity of prescriptions sold, including new prescriptions and refills. It measures the demand of pharmaceutical products that is generated by physicians in the process of treating patients (Cook, 2006). Pharmaceutical companies, including the sponsor company, use TRx as a base unit for sales and demand quantity. This excludes over-the-counter commercial sales, but since commercial sales quantity shows a stable relationship to the prescription quantity, the sponsor company forecasts TRx quantity and multiplies it by a fixed constant to calculate the total demand quantity. This research aims to forecast TRx quantity as a target variable. This research used 10 years of weekly prescription data (Jan 2010 – Nov 2019) of 27 products in the U.S. drug group. Table 3 shows the attributes in TRx data.

**Table 3**

*Prescription history (TRx) data attributes*

| Attributes | Description |
|---|---|
| **Product Name** | Unique product name |
| **Quantity** | Weekly prescribed quantity |
| **Date** | Year, Month, Day (Every ending Friday) |
| **Category** | Identifier depends on the symptoms that the drug cures |

3.2.1.2 Product Features

Product feature data contained multiple features about the products in the US drug group. The data contained explanations for 27 products and 8 attributes in the dataset. Details about the attributes are explained in Table 4. We were able to match the information in this dataset to the products listed in TRx data.

**Table 4**

*Product features attributes*

| Attributes | Description |
|---|---|
| **Product Name** | Unique Product Name |
| **Manufacturer** | Name of the pharmaceutical company producing the product |
| **Type of device** | A type of Inhaler device for the product - Metered-dose inhaler (MDI), Dry powder inhaler (DPI), Soft mist inhaler (SMI) |
| **Launch date** | Product launch date (Year, Month) |
| **Indication** | Disease type to be treated |
| **Dosing** | Instruction on taking the medication |
| **Category** | A market category for the product |
| **Strength** | Amount of medication in 1 dose |

#### 3.2.1.3 Census data

Statistics for the number of asthma patients were updated with the lifetime asthma population estimates from the National Health Interview Survey (NHIS) done by the U.S. Census Bureau. The survey shows the total number of U.S. asthma patients per year (2010-2018). Regarding US drug consumption trends, this research used “Trends in use of one or more prescription drugs in the past 30 days” data from National Health and Nutrition Examination Surveys (NCHS), also by year (2007–2016).

### 3.2.2 Data Exploration

By combining product feature data with the TRx dataset, we could observe product features that may be correlated to a product’s demand such as lifecycle year[4], manufacturer, and TRx category total. TRx data included not only the sponsor company’s products but also the products from other manufacturers. We utilized all 27 products in our dataset, with the assumption that they will show similar demand patterns.

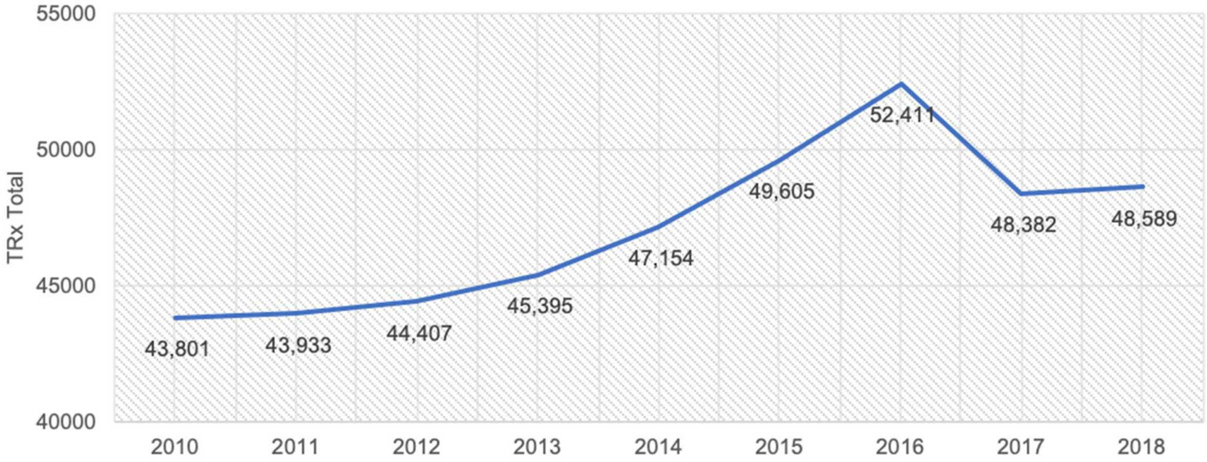
#### 3.2.2.1 Yearly Demand

From 2010 to 2016, demand for U.S. products had increased linearly, but in 2017 the demand decreased 8% from 2016, as shown in Figure 5.

[4] “Lifecycle year” represents the number of years since the product launched

**Figure 5**

*Demand for all U.S. products within a specific drug group by year*



3.2.2.2 Category Demand

The products we analyzed are divided into five categories. These products mainly contain three ingredients, namely, single-inhaler combinations of an inhaled corticosteroid (ICS), a long-acting β2-agonist (LABA), and a long-acting muscarinic antagonist (LAMA) (Vanfleteren, Fabbri, Papi, Petruzzelli, & Celli 2018). These three ingredients are combined in a way that aids patients by relieving their symptoms; the following combinations of ingredients designate the product categories in the drug market: ICS, ICS/LABA, LAMA, LAMA/LABA, and Triple Therapy. There were two products without any demand history, 13 products with 1-5 years of history, three products with 5-8 years of history, and only nine products with full 10 years of data. Table 5 summarizes the available data.

**Table 5**

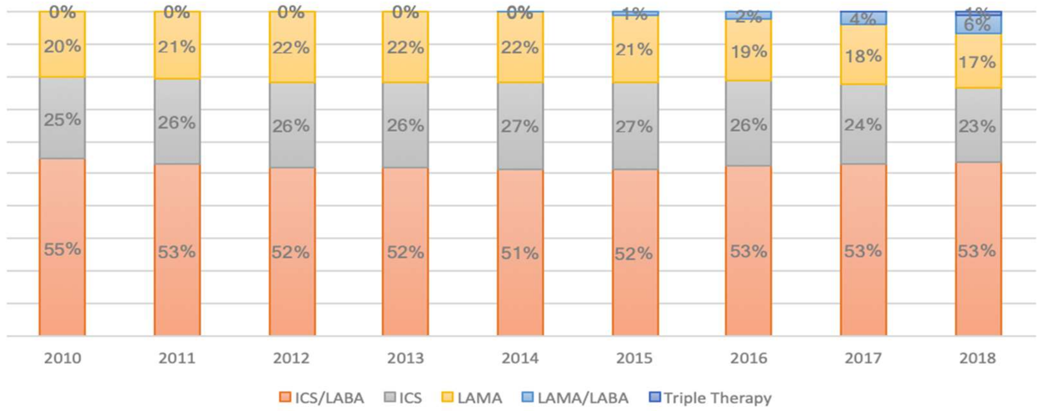
*The number of products and demand history available by category*

| Demand History (Years) | Number of Products by Category | | | | | |
|---|---|---|---|---|---|---|
| | ICS | ICS/ LABA | LAMA | LAMA/ LABA | Triple Therapy | Sum |
| 0 | 1 | 1 | - | - | - | **2** |
| 1 | 1 | - | - | - | - | **1** |
| 2 | 1 | - | - | - | - | **1** |
| 3 | 1 | 2 | 1 | 2 | 1 | **7** |
| 4 | 1 | - | - | - | - | **1** |
| 5 | 1 | - | 1 | 1 | - | **3** |
| 6 | - | - | - | 1 | - | **1** |
| 7 | - | 1 | - | - | - | **1** |
| 8 | - | - | 1 | - | - | **1** |
| 9 | - | - | - | - | - | **-** |
| 10 | 5 | 3 | 1 | - | - | **9** |
| **Total** | **11** | **7** | **4** | **4** | **1** | **27** |

When the total demand was disaggregated within the 5 categories, ICS/LABA and LAMA categories showed a decrease in market share affected by the introduction of LAMA/LABA and Triple Therapy categories, but there was little change in the ICS category market share for all 10 years. Figure 6 shows the market dynamics of demand by category in the U.S. drug group.

**Figure 6**

*Percent market share of US products (within in a specific group) for each category by year*

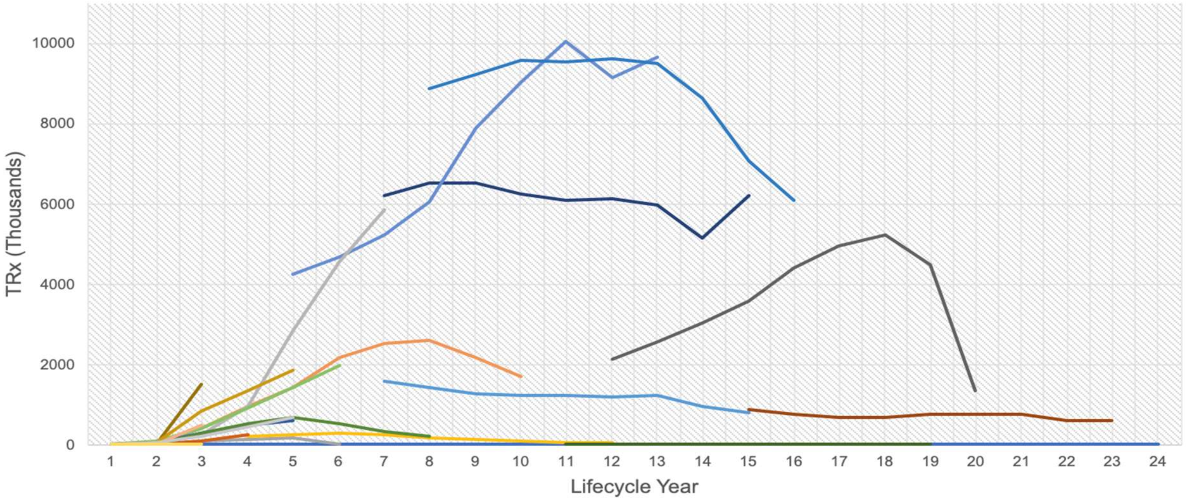


### 3.2.2.3 Lifecycle Year

For the products that have been in the market for fewer than eight years, the demand showed a general upwards trend. The demand showed stabilized patterns after nine years from their launch. Figure 7 shows product demand progression after the product launch. Since only 10 years of data were available for each product, the demand histories for each product are positioned partially in the graph.

**Figure 7**

*Demand pattern of products in U.S market by lifecycle year*

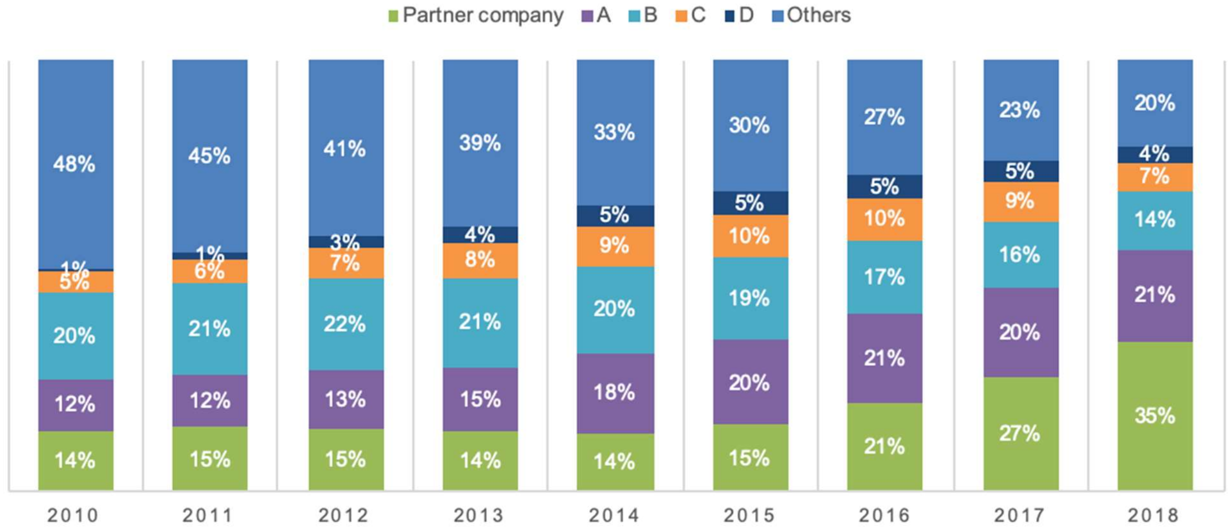


### 3.2.2.4 Manufacturer

The 27 products listed in TRx data were made by multiple manufacturers. The market share of the sponsor company's products has been growing steadily from 2010 from sales perspective (Figure 8) as well as number of products in the market (Figure 9). For products in the group we were analyzing, the sponsor company recorded a 35% market share of the entire U.S. drug market in 2018. Other manufacturers were anonymized and labeled as A, B, C, and D.

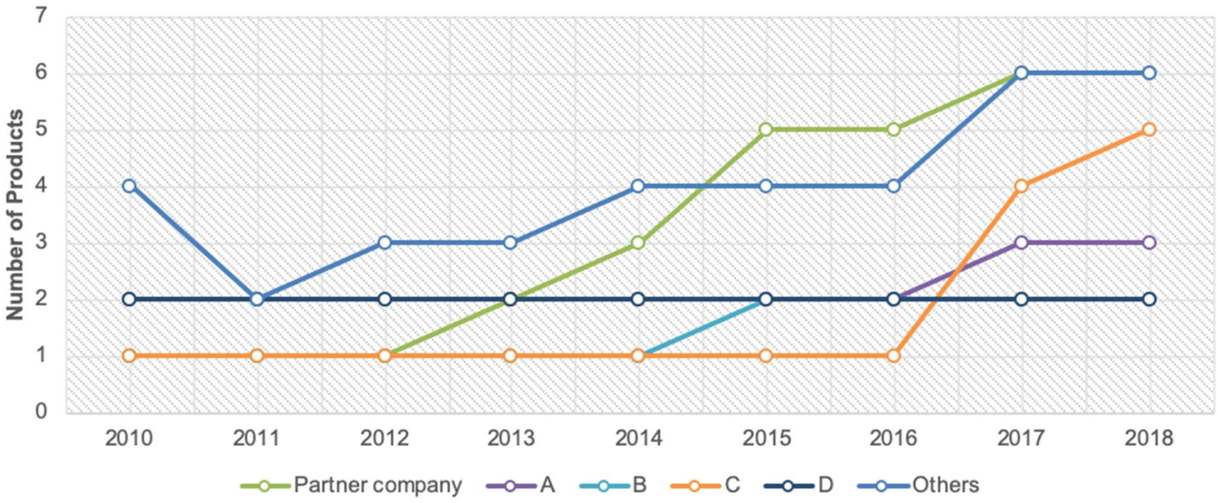
## Figure 8

*Market share of U.S. products within the drug subset by manufacturer*[5]



## Figure 9

*Number of products within the drug subset by manufacturer*[5]



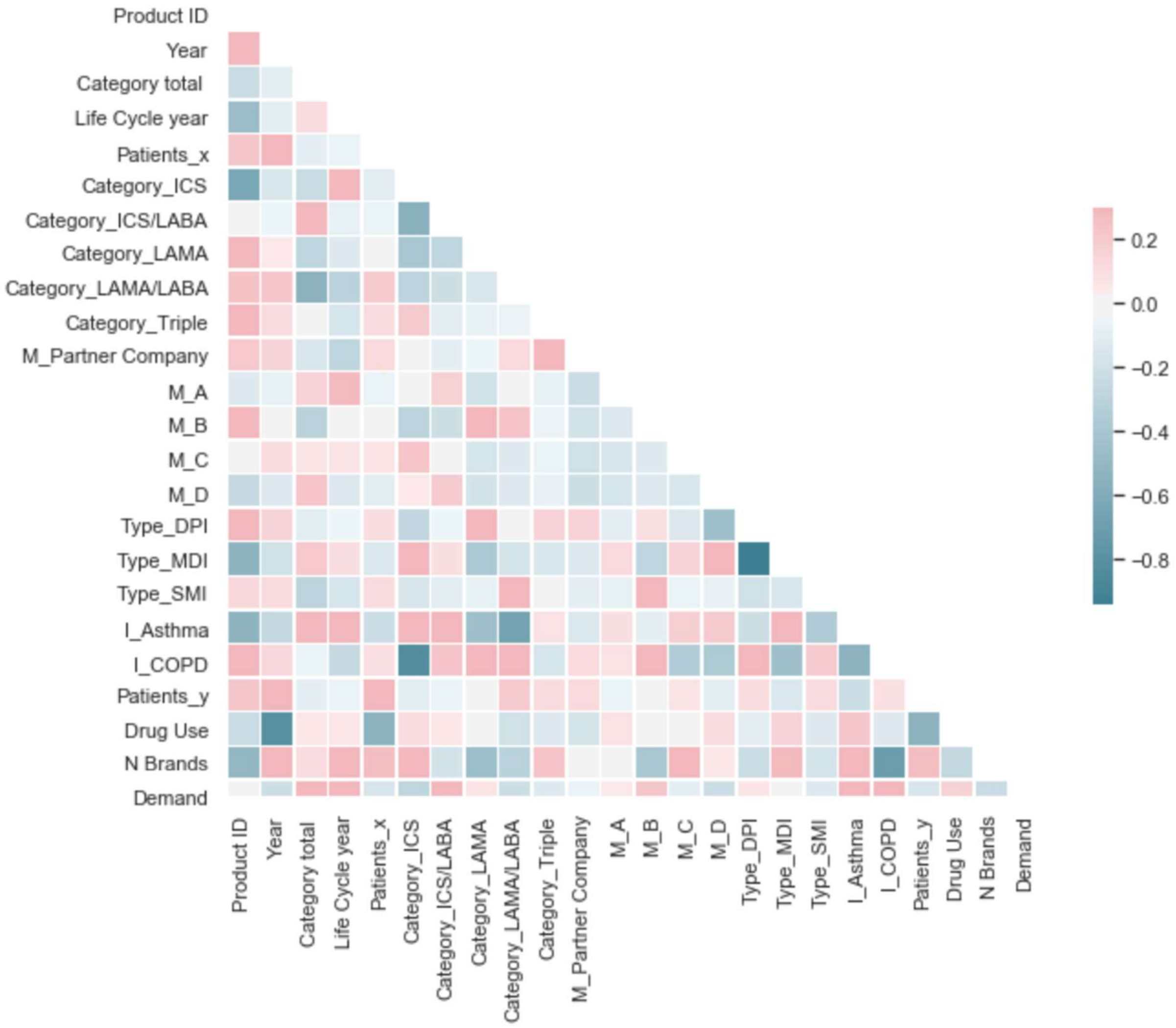
[5] Other manufacturers were anonymized and labeled as A, B, C, and D

### 3.2.2.5 Correlation Chart

After completing the dataset formation, a correlation matrix was created to identify the correlation coefficients among all variables in the dataset (Figure 10).

**Figure 10**

*Correlation coefficient chart*



### 3.2.3 Feature Engineering

This section explains the procedure of transforming the original dataset for use in the forecasting model. Census data and Product ID were merged with the demand history data in line with the Year and Product ID. Weekly time series data was aggregated on a monthly and yearly basis in order to properly predict yearly demand. Categorical features were transformed into dummy variables[6] and all the input data was standardized before model training.

#### 3.2.3.1 Date Features

Date was originally written in the form of "year-month-day." This was separated into three individual variables, "year," "month," and "day."

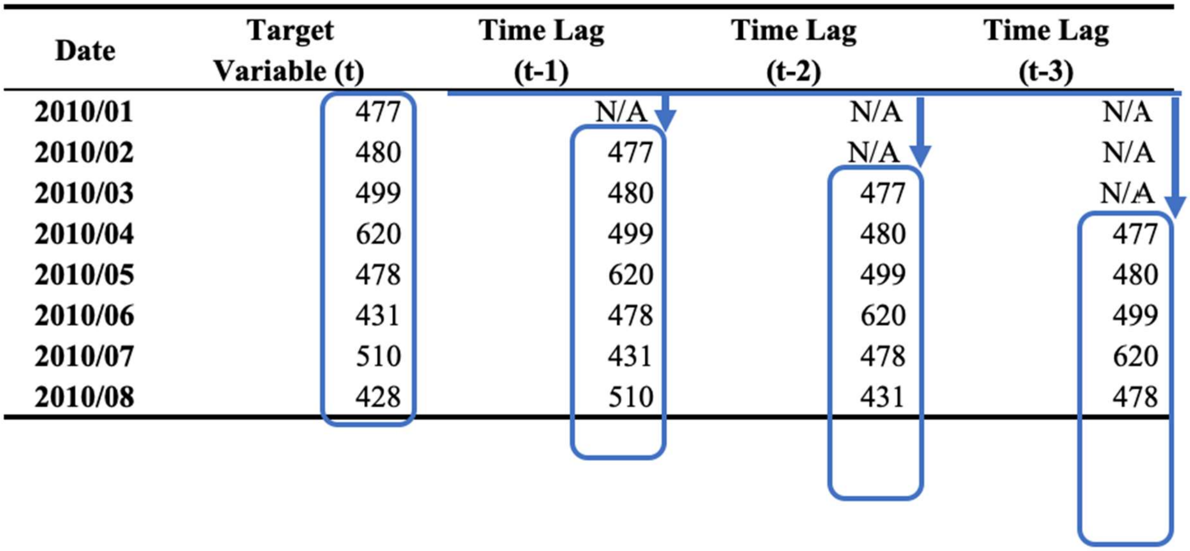
#### 3.2.3.2 Time Lag

Time series demand data (TRx data) needed proper preparation in order to re-frame the time series forecasting problem into a supervised learning[7] problem. The method of using prior time steps to predict the next time step is called the time lag method. This method captures certain movement in the demand history data. Multiple lengths of time lag were created as additional input variables for each monthly instance, as shown in Figure 11. The size of the optimal time lag is determined through empirical research.

[6] A dummy variable is a way to represent categorical data numerically. Since linear regression requires variables to be numeric, dummy variables are created to capture whether or not the categorical variable is present in the data.
[7] Supervised learning algorithms require training the model to map inputs to outputs.

**Figure 11**

*Example of time lag creation*



| Date | Target Variable (t) | Time Lag (t-1) | Time Lag (t-2) | Time Lag (t-3) |
|---|---|---|---|---|
| 2010/01 | 477 | N/A | N/A | N/A |
| 2010/02 | 480 | 477 | N/A | N/A |
| 2010/03 | 499 | 480 | 477 | N/A |
| 2010/04 | 620 | 499 | 480 | 477 |
| 2010/05 | 478 | 620 | 499 | 480 |
| 2010/06 | 431 | 478 | 620 | 499 |
| 2010/07 | 510 | 431 | 478 | 620 |
| 2010/08 | 428 | 510 | 431 | 478 |

### 3.2.3.3 Lifecycle Year

Since the demand for a drug showed correlation in line with the drug's time since launch, the lifecycle year variable was created to capture this effect. "Product launch date" from the product feature data was extracted and the lifecycle year was calculated accordingly:

$$Lifecycle\ year\ = (Demand\ year\ - Launch\ year) + 1$$

For example, the lifecycle year for a product's demand data for 2013 with a product launch in 2007 would be determined as 7, or (2013-2007 +1).

#### 3.2.3.4 Number of Brands in Category

In the pharmaceutical industry, there is a relatively small number of brands available for patients due to the requirement that they be approved by the U.S. Food and Drug Administration (FDA). Since the entry of a new brand or competitors in the same category could lead to a significant shift in demand for the existing brand(s), the "Number of Brands in a Category" feature was created to capture the competitive status of the product. The number of brands with sales history in a given year was calculated and inserted as a feature.

#### 3.2.3.5 Type of Device

The subset of products uses medical aerosol as a medium, a mixture of vaporized medication and gas, to transfer the medication. There are mainly 3 types of devices that utilize aerosol treatment: Pressurized metered-dose inhaler (MDI), dry-powder inhaler (DPI), and soft-mist inhaler (SMI). Three dummy variables were created representing the 3 types of devices (15 DPI, 11 MDI, 1 SMI).

#### 3.2.3.6 Total Prescription Quantity by Category

Total prescription quantities for each of the five categories were calculated in line with the aggregate timeframe.

#### 3.2.3.7 Census Data

Originally, the current model used by the sponsor company was not using census data to predict future demand. Since this research aims for finding the most relevant features to forecast future demand, we explored the external data available and decided to include census data as features in ML models. Lifetime asthma-population estimates, from 2010 – 2018, and U.S. drug consumption data, from 2010 – 2016, both in a yearly format, were assigned to each instance with its corresponding year. Even though asthma population data was segmented by age, gender, and race, due to the difficulty of pairing each brand with this segmented data, total asthma-population data was used for the research. Since US drug consumption was surveyed every two years, we distributed one survey result to two years each. Since the most recent survey data was for the year 2017, we used a simple linear regression model to fill in the missing data from 2017 to 2020.

### 3.2.4 Data Cleaning, Aggregation, and Transformation

This section explains the steps taken to clean, aggregate, and transform the data to prepare it for the machine learning models.

#### 3.2.4.1 Data Cleaning

The original weekly TRx data had 20,640 instances with 27 products, but this included instances of zero demand as well. To extract only the relevant data to include in training the model, the instances with zero values were deleted. After erasing the rows with zero demand, the number of instances decreased to 12,527.

#### 3.2.4.2 Data Aggregation

Originally, there were three datasets: TRx data, product feature data, and census data. This research explored two types of approaches to forecasting, direct and recursive, using two different aggregate datasets. For direct forecasting, original weekly level TRx data was aggregated into a yearly timeframe dataset, and for recursive forecasting, a monthly level of aggregation was performed. Instances within the year 2019 were dropped for the yearly timeframe dataset due to the original data lacking values for December of that year. These datasets were merged with the other two datasets with Python, using Product ID and Year as keys. The yearly dataset consisted of 134 instances and monthly dataset of 1,859 instances.

#### 3.2.4.3 Data Transformation

The dataset contained a mixture of scales for various features, such as prescription quantities, number of people, percentages, and number of brands. As machine learning methods are more effective when the attributes are of a common scale, the data was transformed with data standardization. Standardization methods were utilized to rescale variables to have a mean of 0 and a standard deviation of 1.

### 3.2.5 Feature Selection and Data Partitioning

Having many features in a dataset leads to a model becoming highly dimensional. To reduce the complexity of a model and make it easier to interpret, it is important to drop the redundant features that do not significantly improve the model’s performance. To select the important features among a group of 22 features created in the dataset, this research used the support vector machine feature selection method.

#### 3.2.5.1 Support Vector Machine (SVM) Feature Selection

SVM creates a hyperplane that uses support vectors to maximize the distance between two classes. When the SVM model is created to fit to the data, relevant coefficients are calculated; the relationship between these coefficients is then used to determine the feature importance for the data separation task.

Ten years of available data was partitioned into two datasets, a training set and a test set. As the research aimed to verify how far a machine learning model can forecast with limited data, the data was partitioned in multiple ways. When the forecasting period was extended, the training set period was reduced. The training set was used to train the model and the test set was used to evaluate the model accuracy by comparing the forecasted target variable with actual values.

## 3.3 Multi-step Forecasting Approaches

There are two types of forecasting approaches: one-step forecasting and multi-step forecasting. One-step forecasting predicts a single observation at the following time step. Multi-step forecasting predicts multiple time steps into the future. Since our research is focused on finding appropriate approaches for long-range forecasting, we tested two types of multi-step forecasting approaches: recursive and direct. This section explains our methodology for the recursive and direct approaches.

### 3.3.1 Direct Method

The direct method uses only historical data to forecast future time. As Table 6 shows, regardless of which time step the prediction is based on, the direct method only uses historical data. In other words, it doesn’t recursively feed in the predicted result from the model to predict the next step. However, in order to increase the forecasting accuracy, separate models can be built for each time step in the forecast (Taieb et al., 2014). For example, for predicting the demand of each weekday, individual models can be structured for each of the five weekdays.

**Table 6**

*Direct method train and test data split*

| Training data | | | | | Predicting time | | |
|---|---|---|---|---|---|---|---|
| t-5 | t-4 | t-3 | t-2 | t-1 | t | | |
| t-6 | t-5 | t-4 | t-3 | t-2 | N/A | t | |
| t-7 | t-6 | t-5 | t-4 | t-3 | N/A | N/A | t |

### 3.3.2 Recursive Method

Recursive modeling makes a prediction for one time-step, and proceeds to feed that prediction into the model as an input in order to predict the subsequent time steps. This process is repeated until the intended steps to be forecasted are reached. Table 7 shows that predicted values are also included in the training set with other historical features to forecast the time steps ahead.

**Table 7**

*Recursive method train and test data split*

| Training data | | | | | Predicting time | | |
|---|---|---|---|---|---|---|---|
| t-5 | t-4 | t-3 | t-2 | t-1 | t | | |
| t-6 | t-5 | t-4 | t-3 | t-2 | t-1 | t | |
| t-7 | t-6 | t-5 | t-4 | t-3 | t-2 | t-3 | t |

## 3.4 Machine Learning Models

While there are many kinds of machine learning models available for long-term forecasting, this research applied methodologies that our literature review determined have been tested and proven effective in demand forecasting. Support vector machine (SVM), random forest (RF), artificial neural networks (ANN), and linear regression (LR) were chosen to be tested for long-term demand forecasting.

### 3.4.1 Support Vector Machines

The support vector machine (SVM) (Vapnik, 1998) is a supervised learning algorithm that is used for classification and regression analysis. It is based on the structural risk minimization principle, which sets a hyperplane that maximizes the margin of separation between different classes and minimizes the expected error of a learning machine. Through training the data, one unique hyperplane, called optimize hyperplane, is created to separate the data. Since most real data is not linearly distributed, a kernel function is created to classify the nonlinearly distributed data. The kernel function is applied on each data instance to map the original nonlinear instances into a high-dimensional space, where they become separable. By calculating the distances from the kernel function to the instances, SVM can be used as a regression method to predict the future values.

### 3.4.2 Random Forest

Random forest is a supervised learning algorithm for classification and regression purposes that consists of a large number of decision trees that operate as an ensemble. An ensemble method combines the predictions from multiple decision tree algorithms together to make more accurate predictions than any individual model would. Decision tree is a machine learning method that finds patterns inside of data and makes a rule to classify said data. This method classifies instances into branch-like segments, hence its nomenclature. The decision tree structure has internal nodes representing a feature and branches representing a decision rule, with each leaf node representing the outcome. The decision tree algorithm starts by selecting the best attribute using attribute selection measures (ASM) to split the data. The attribute selected becomes a decision node and breaks the dataset into smaller subsets. By repeating this process recursively for each instance, a tree is formed. At each node of the decision tree, the information entropy gain generated by the split is used to evaluate whether the variable is meaningful or not. Random forest constructs a multitude of decision trees with a training set and outputs the class for a classification problem or prediction for a regression problem.

### 3.4.3 Artificial Neural Networks (ANN)

Artificial neural network is a machine learning algorithm that resembles the workings of the human brain. ANNs discover a new pattern out of investigating a relationship between the inputs and outputs. ANN has three important elements in its structure: input layer, hidden layer, and output layer. The input layer receives the information into the model, the hidden layer computes the patterns among the input data, and the output layer obtains the result. In a neural network, there are multiple parameters and hyperparameters that affect the performance of the model. Each node in the network is assigned weights that can be interpreted as the impact that node has on the node of the next layer. Depending on the weighted sum value, an activation function defines whether a given node should be activated or not. Finally, based on the result, the model adjusts the weights of the neural networks to optimize the network following various cost minimization functions.

### 3.4.4 Linear Regression (LR)

Linear regression was developed in the statistics field, but it is used in machine learning as well. Linear regression builds a model that assumes a linear relationship between the input variables (x) and the output variable (y) along with a corresponding coefficient for each input variable. Machine learning uses gradient descent[8] to update these coefficients to reduce the error in forecasting the output variable.

## 3.5 Performance Measurement

To evaluate the model performance, the test set was used to compare the actual values against the predicted values. MAPE (Mean Absolute Percentage Error) and MPE (Mean

[8] Gradient descent is an optimization algorithm that finds the local minimum of a function by moving along points taken at the steepest descent.

Percentage Error) for the sponsor company’s products were calculated. MAPE is used to measure forecast accuracy and MPE is used to measure model bias. For the recursive model that had a monthly forecasted outcome, outcomes were aggregated on a yearly level before the model’s accuracy was evaluated.

# 4 Results

This section shows the results of applying machine learning approaches to demand forecasting for the sponsor company's products. Results are shown in the order of approaches this research has taken. Direct forecast and recursive forecast methodologies were studied to find out which approach will be a better fit for demand forecasting purposes. The recursive method showed better accuracy than the direct method and thus it was investigated further by elongating the time horizon. By extending the steps ahead from 1 month to 36 months, the forecasting accuracy fell significantly.

## 4.1 Direct Forecast

The direct forecast method was applied as the first step of modeling. Target variables were set to directly forecast the yearly demand of six the sponsor company's products (Product ID 1, 2, 3, 4, 5, 6). Ten features, which showed the strongest relevance from among 22 features chosen through SVM feature selection, were selected to train the model. Data was normalized with a standardization method. Three years of yearly demand for the sponsor company's products were forecasted and compared with actual demand values. To evaluate the effectiveness of the ML model, the model needs enough data to train and test the model. Since we had limited size of the demand data (10 years), we selected a 3-year forecasting horizon, which is the threshold the sponsor company uses to separate short-term forecasting and long-range forecasting, instead of extending the forecasting period further. Table 8 shows the structure of the direct forecast model.

**Table 8**

*Direct forecast model structure*

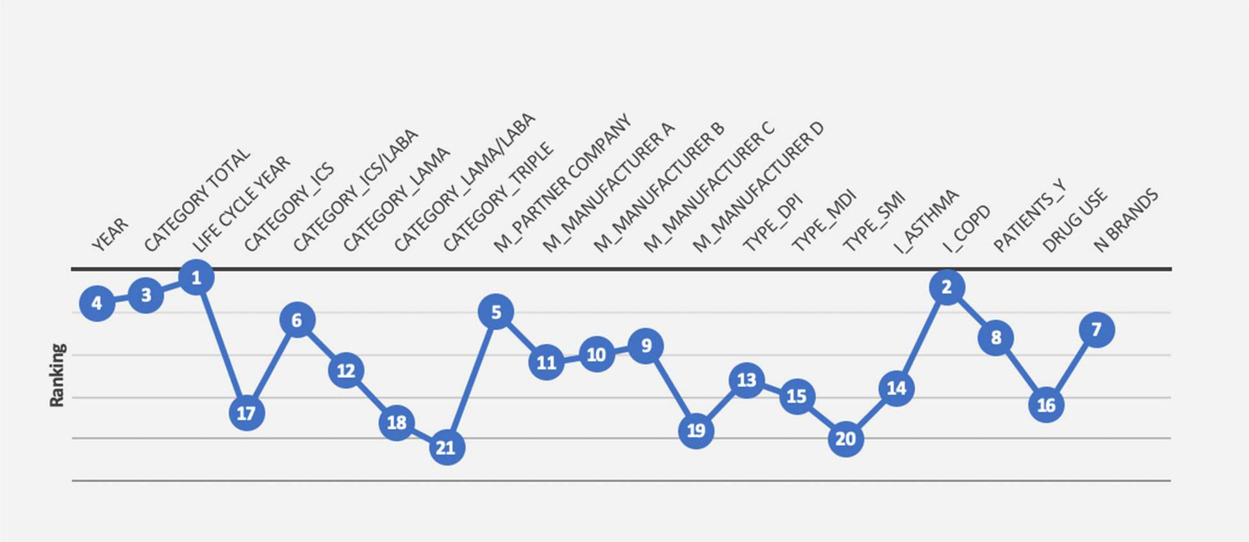
| Model Structure | Contents |
|---|---|
| **Target Variable** | Yearly Demand of The Sponsor Company's Brands |
| **Number of Features** | 10 |
| **Features** | Product Features/Census Data |
| **Data Normalization** | Standardization (mean, standard deviation) |
| **Number of Instances** | 134 |
| **Data Split** | Train (2010 - 2015) / Test (2016 - 2018) |
| **Forecasted Period** | 3 years |

### 4.1.1 Feature Selection

Support vector machine (SVM) method was applied for feature selection. Out of 22 features, 10 were selected for modeling. "Lifecycle year" ranked first and "Category Total" also showed high relevance in the feature selection ranking. The date feature "Year" was ranked third. As Figure 12 shows, most of the product features show low importance on feature selection ranking and therefore were excluded from the modeling.

**Figure 12**

*Feature importance ranking*



### 4.1.2 Model Selection

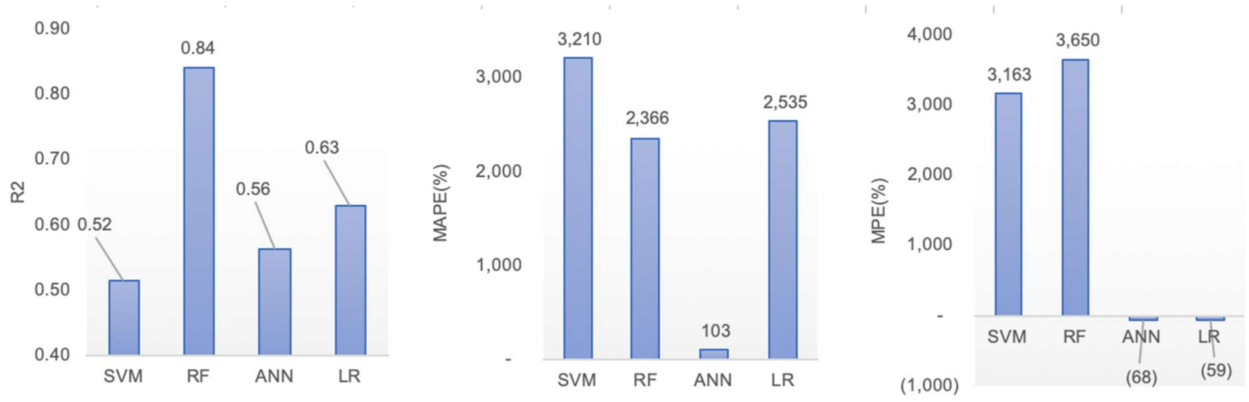
Four different machine learning methodologies, support vector machine (SVM), artificial neural network (ANN), random forest (RF) and linear regression (LR) were used to test the direct forecasting model.

### 4.1.3 Result

As Figure 13 shows, the forecasted results from the four methodologies were compared against the actual demand for five of the sponsor company’s products. Figure 13, from right to left, shows the $R^2$ value, MAPE, and MPE.

**Figure 13**

*Direct forecast model results*



## 4.2 Recursive Forecast

To include the forecasted result into next step, the recursive forecasting method was tested. Target variables were set to forecast the monthly demand of six of the sponsor company's products (Product ID 1, 2, 3, 4, 5, 6) in the same manner as direct forecasting. Creating a time-lag for each instance from its previous historical data meant that the first few instances could not be used due to the lack of history. Therefore, despite the goal of this research being to forecast demand over a long-term timeframe, monthly instances were used in order to have sufficient data to train the model. In this model, product feature and census data were not used due to the lack of positive results from direct forecasting using said variables, as well as the focus on observing the importance of time-series data on long-range forecasting. Training and test sets were split with various ratios to test the extent of machine learning's capabilities of predicting demand with limited data. Table 9 shows the structure of the recursive forecast model.

**Table 9**

*Recursive forecast model structure*

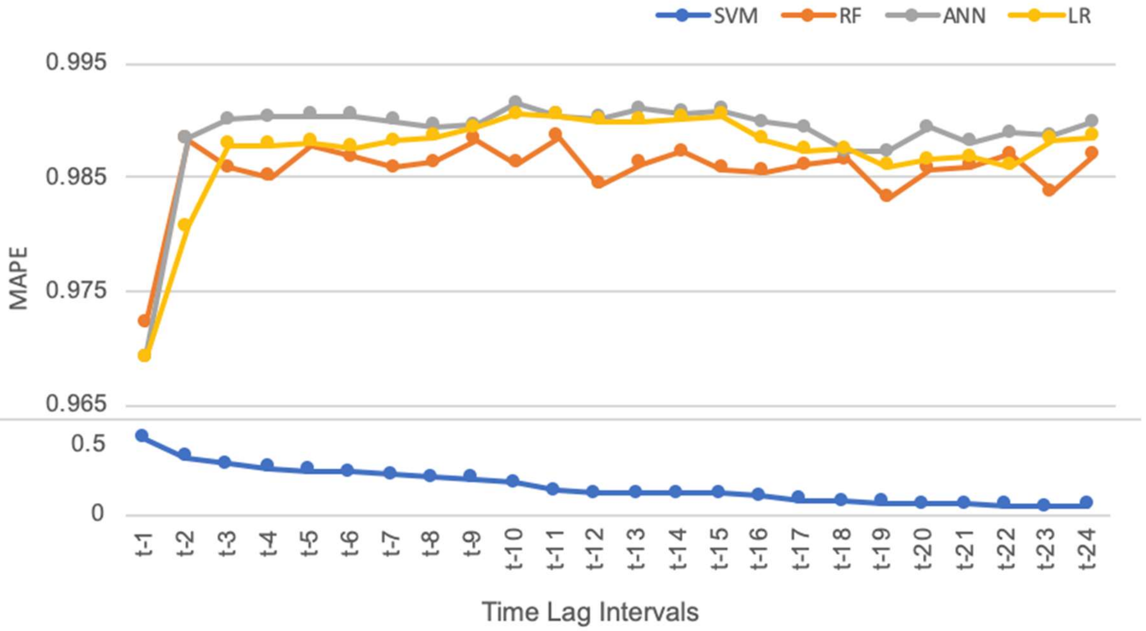
| Model Structure | Contents |
|---|---|
| **Target Variable** | Monthly Demand of The Sponsor Company’s Brands |
| **Number of Features** | 12 |
| **Features** | Date feature (Month) / Time Lags (11 Months) |
| **Number of Instances** | 1,859 |
| **Data Split** | Multiple Ratio |
| **Forecasted Period** | 1-9 years |

### 4.2.1 Time-Lag Feature Optimization

For recursive modeling, the time lag horizon must be determined to optimize the forecast. Repetitive search was applied to find the optimal time lag horizon. Time lag steps were tested from a range of 1-24 months with nine years of training set data. For a 24-month range time lag, 24 time-lag steps were included as features for one instance (one-month prior demand to 24-month prior demand). One-step leading forecast results were evaluated to determine optimal time lag, and a 1-year test set was used to evaluate the model’s performance. As the results in Figure 14 show, 11-month range time lags showed the best performance in terms of model fitness, and thus a 11-month range was used for the modeling.

**Figure 14**

*R-squared by time lag intervals*



### 4.2.2 Model Selection

The methodologies that were previously tested in the direct method, such as SVM, RF, ANN and LR, were chosen for recursive modeling as well with direct comparisons between each model's results being drawn to conclusion.

### 4.2.3 Forecasting Horizon

To test the feasibility of using time-lag features in forecasting, one month-lead forecasting without the recursive function was tested first. After confirming time-lag steps were useful in forecasting the target variable, a model using prior forecasted results as input features was built.

### 4.2.4 One-step Forecasting Result

To determine the relevance of time lag as an input variable on demand forecasting, a one-step forecasting model was created with time-lag variables. T-11 time lags were used for the modeling, which showed the best performance from time lag feature optimization. As shown in Table 10, ANN performed best in one-step forecasting with MAPE of 8.86%, followed by LR 10.23% and RF with 10.31%. MAPE for SVM was the highest among the four different methodologies.

**Table 10**

*MAPE of one-step forecasting for the sponsor company's products*

| Product ID | MAPE | | | |
|---|---|---|---|---|
| | SVM | RF | ANN | LR |
| 1 | 82.92% | 11.61% | 9.42% | 9.81% |
| 2 | 81.77% | 8.35% | 7.30% | 8.25% |
| 3 | 49.24% | 10.03% | 7.19% | 8.70% |
| 4 | 49.24% | 10.03% | 7.19% | 8.70% |
| 5 | 30.22% | 13.39% | 12.48% | 15.15% |
| 6 | 51.57% | 8.45% | 9.58% | 10.75% |
| Avg. | 57.50% | 10.31% | 8.86% | 10.23% |

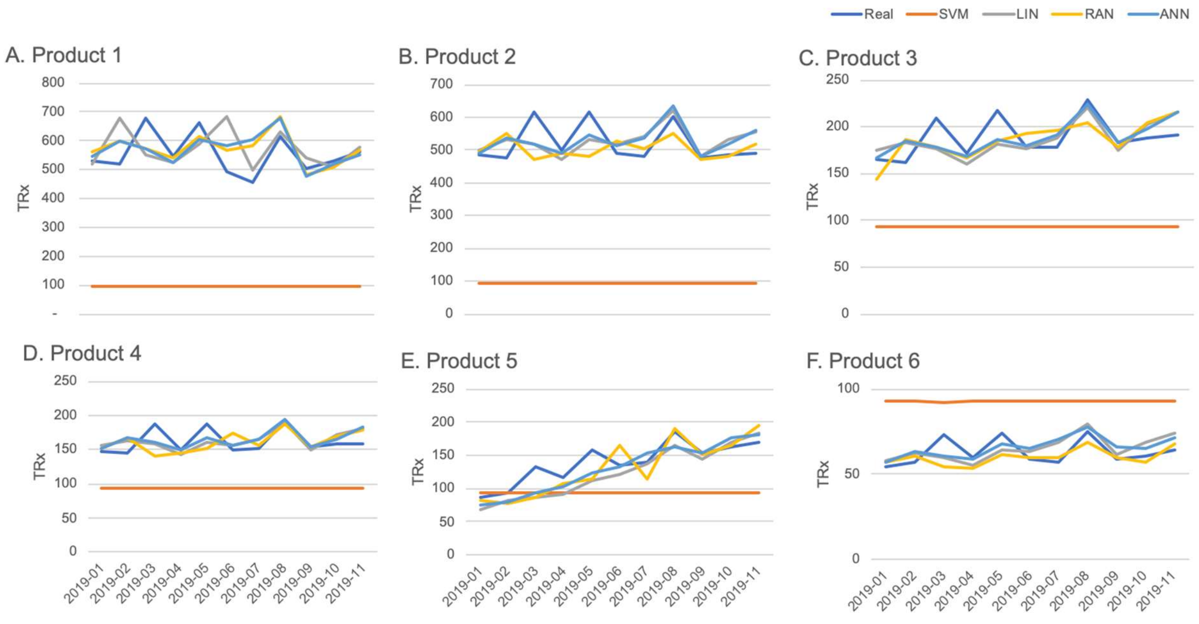
**Table 11**

*MPE of one-step forecasting for the sponsor company's products*

| Product ID | MPE | | | |
|---|---|---|---|---|
| | SVM | RF | ANN | LR |
| 1 | -82.92% | 3.47% | 3.40% | 3.44% |
| 2 | -81.77% | 2.62% | 2.45% | -1.10% |
| 3 | -49.24% | 1.35% | 0.65% | 0.11% |
| 4 | -41.28% | 3.00% | 2.29% | 0.58% |
| 5 | -24.86% | -8.15% | -13.25% | -6.58% |
| 6 | 51.57% | 5.09% | 4.24% | -2.93% |
| Avg. | -38.08% | 1.23% | -0.04% | -1.08% |

**Figure 15**

*One-step forecast results for the sponsor company's products*



As shown in Figure 15, ANN performed best in one-step forecasting for five products (Product 1, 2, 3, 4, 5) except for one product (Product 6), whose MAPE was the lowest with RF.

### 4.2.5 Multi-step Forecasting Result

For forecasting multi-step leads, a new model that transformed the forecasted monthly demand into a feature represented as a time lag was necessary. For example, in the recursive model, forecasted demand of January 2018 becomes the t-1 time-lag feature for forecasting February 2018 demand as well as the t-2 time-lag feature for forecasting March 2018 demand.

Forecasting horizons were set in six different scenarios, 1-year, 2-years, 3-years, 4-years, 5-years and 9-years and results were compared. 1-year scenario used nine years of data as a training set (2010-2018) and one year of data (2019) as a test set. The 2-year forecasting scenario used eight years of data as a training set (2010-2017) and two years of data (2018-2019) as a test set. For 3-year scenario, one year of data for the training set was removed from the 2-year forecasting training set and the same period was added to the test set accordingly. And for analysis for different time horizons, data have been split in the same manner. The R-Squared score decreased as the forecasting horizon was extended as Table 12 shows. The R-squared score was highest with ANN methodology in the 1-year scenario and RF performed the best in the 3-year scenario.

**Table 12**

*R squared scores with different forecasting horizons*

| Data | Methodology | 1 Year | 2 Years | 3 Years |
|---|---|---|---|---|
| Train Set | SVM | 0.46 | 0.64 | 0.65 |
| | RF | 1.00 | 1.00 | 1.00 |
| | ANN | 0.99 | 0.99 | 0.99 |
| | LR | 0.99 | 0.99 | 0.99 |
| Test Set | SVM | (0.01) | (0.02) | (0.03) |
| | RF | 0.97 | 0.95 | 0.96 |
| | ANN | 0.98 | 0.95 | 0.90 |
| | LR | 0.98 | 0.96 | 0.90 |

**Table 13**

*MAPE for one-year forecasting horizon*

| Product ID | SVM | RAN | ANN | LR |
|---|---|---|---|---|
| 1 | 83.31% | 4.01% | 0.05% | 6.00% |
| 2 | 82.22% | 3.71% | 6.50% | 0.90% |
| 3 | 50.94% | 1.05% | 11.22% | 4.89% |
| 4 | 42.90% | 4.01% | 6.82% | 1.28% |
| 5 | 33.46% | 29.74% | 19.43% | 33.70% |
| 6 | 46.70% | 7.30% | 5.50% | 1.60% |
| Avg. | 56.59% | 8.30% | 8.25% | 8.06% |

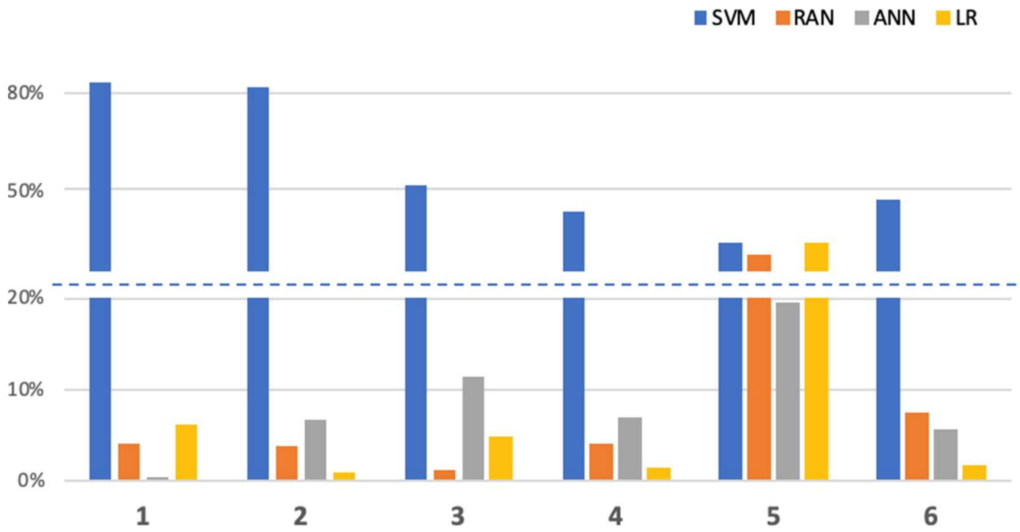
**Table 14**

*MPE for one-year forecasting horizon*

| Product ID | SVM | RAN | ANN | LIN |
|---|---|---|---|---|
| 1 | -83.3% | -4.0% | 0.0% | 6.0% |
| 2 | -82.2% | -3.7% | -6.5% | -0.9% |
| 3 | -50.9% | -1.0% | -11.2% | -4.9% |
| 4 | -42.9% | -4.0% | -6.8% | -1.3% |
| 5 | -33.5% | -29.7% | -19.4% | -33.7% |
| 6 | 46.7% | -7.3% | -5.5% | 1.6% |
| **Avg.** | **41.02%** | **8.30%** | **8.25%** | **5.53%** |

**Figure 16**

*MAPE for one-year forecasting horizon by products*



The results show that methodologies performed differently across products. For instance, Figure 16 shows LR performed the best for three products (Products 2, 4, 6), while RAN for Product 3, and ANN for Products 1 and 5 performed the best. SVM performed the worst for all six products. The dashed line in Figure 16 was used to scale the visual to improve readability.

**Table 15**

*MAPE for two-year forecasting horizon*

| Product ID | SVM | | RAN | | ANN | | LR | |
|---|---|---|---|---|---|---|---|---|
| | 2018 | 2019 | 2018 | 2019 | 2018 | 2019 | 2018 | 2019 |
| 1 | 81% | 83% | 13% | 14% | 20% | 25% | 14% | 17% |
| 2 | 80% | 81% | 10% | 13% | 12% | 15% | 8% | 10% |
| 3 | 42% | 49% | 5% | 18% | 10% | 14% | 8% | 5% |
| 4 | 37% | 40% | 21% | 30% | 14% | 10% | 10% | 3% |
| 6 | 73% | 53% | 26% | 2% | 17% | 81% | 4% | 40% |
| Avg. | 63% | 61% | 15% | 15% | 15% | 29% | 9% | 15% |

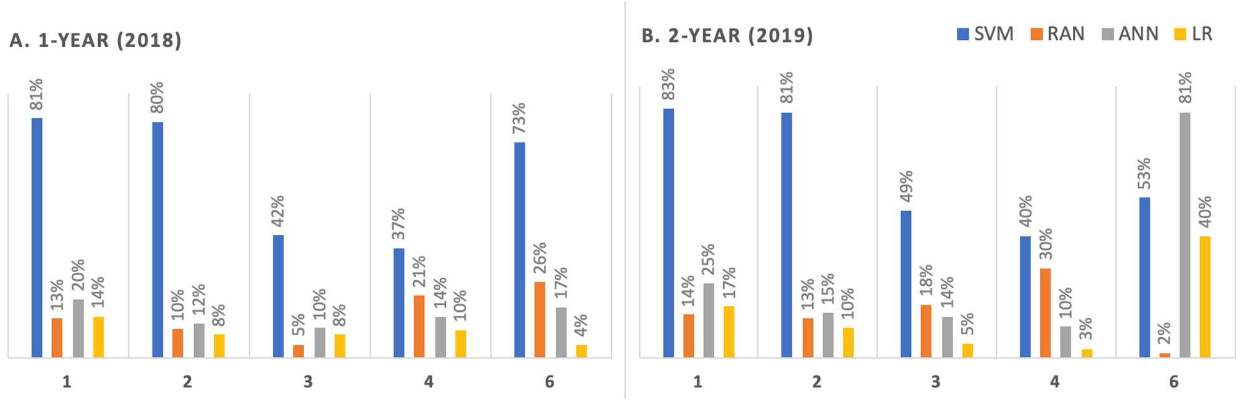
**Table 16**

*MPE for two-year forecasting horizon*

| Product ID | SVM | | RAN | | ANN | | LR | |
|---|---|---|---|---|---|---|---|---|
| | 2018 | 2019 | 2018 | 2019 | 2018 | 2019 | 2018 | 2019 |
| 1 | -81% | -83% | -13% | -14% | -20% | -25% | -14% | -17% |
| 2 | -80% | -81% | -10% | -13% | -12% | -15% | -8% | -10% |
| 3 | -42% | -49% | -5% | -18% | -10% | -14% | -8% | -5% |
| 4 | -37% | -40% | -21% | -30% | -14% | -10% | -10% | 3% |
| 6 | 73% | 53% | 26% | 2% | 17% | 81% | 4% | 40% |
| **Avg.** | **-34%** | **-40%** | **-5%** | **-15%** | **-8%** | **4%** | **-7%** | **2%** |

**Figure 17**

*MAPE for two-year forecasting horizon by Product*



As the average MAPEs shown in Table 15, each methodology performed better in the first year than the second year in the two-year forecasting horizon. When we compared the first-year forecasting performance in the two-year forecasting horizon (Table 15) to the one-year forecasting horizon result (Table 13), the performance of the first-year in the two-year forecasting horizon, which had a smaller training set, performed worse.

**Table 17**

*MAPE for three-year forecasting horizon*

| Product ID | SVM | | | RAN | | | ANN | | | LR | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | 2017 | 2018 | 2019 | 2017 | 2018 | 2019 | 2017 | 2018 | 2019 | 2017 | 2018 | 2019 |
| 1 | 77% | 81% | 82% | 20% | 1% | 6% | 9% | 13% | 22% | 14% | 3% | 5% |
| 2 | 74% | 80% | 81% | 13% | 2% | 3% | 24% | 43% | 48% | 19% | 32% | 31% |
| 3 | 18% | 41% | 48% | 1% | 29% | 36% | 23% | 43% | 49% | 12% | 21% | 11% |
| 4 | 14% | 37% | 40% | 1% | 21% | 27% | 24% | 42% | 44% | 13% | 18% | 0% |
| 6 | 162% | 75% | 54% | 34% | 58% | 63% | 6% | 30% | 39% | 1% | 19% | 67% |
| Avg. | 69% | 63% | 61% | 14% | 22% | 27% | 17% | 34% | 40% | 12% | 19% | 23% |

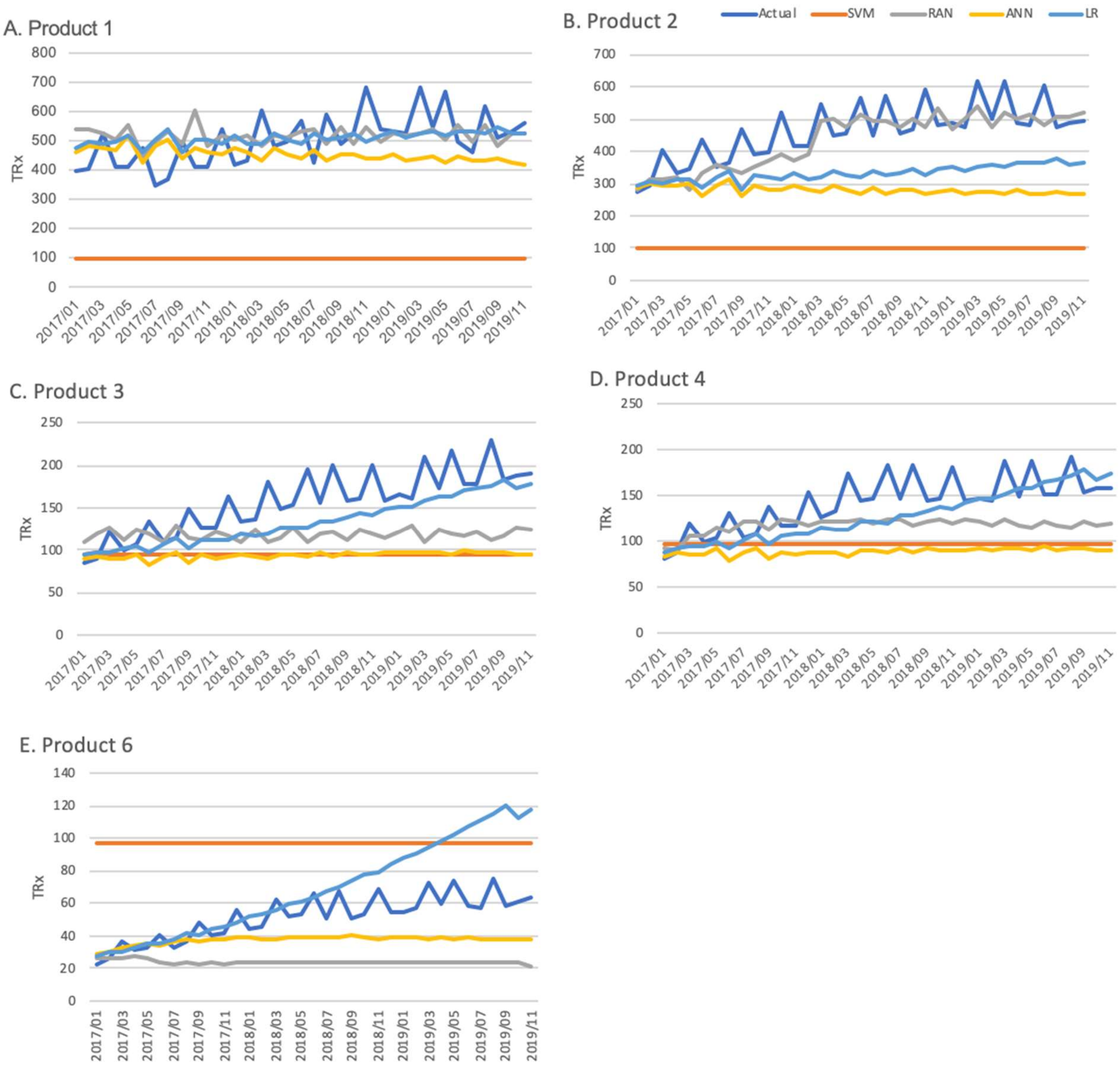
**Table 18**

*MPE for three-year forecasting horizon*

| Product ID | SVM | | | RAN | | | ANN | | | LR | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | 2017 | 2018 | 2019 | 2017 | 2018 | 2019 | 2017 | 2018 | 2019 | 2017 | 2018 | 2019 |
| 1 | -77% | -81% | -82% | 20% | -1% | -6% | 9% | -13% | -22% | 14% | -3% | -5% |
| 2 | -74% | -80% | -81% | -13% | -2% | -3% | -24% | -43% | -48% | -19% | -32% | -31% |
| 3 | -18% | -41% | -48% | -1% | -29% | -36% | -23% | -43% | -49% | -12% | -21% | -11% |
| 4 | -14% | -37% | -40% | -1% | -21% | -27% | -24% | -42% | -44% | -13% | -18% | 0% |
| 6 | 162% | 75% | 54% | -34% | -58% | -63% | -6% | -30% | -39% | 1% | 19% | 67% |
| Avg. | -4% | -33% | -40% | -6% | -22% | -27% | -14% | -34% | -40% | -6% | -11% | 4% |

**Figure 18**

*Multi-step forecasting results (3 year) for the sponsor company's products*



When forecasting performances were compared among all methodologies, LR showed the best performance out of the four in all three types of time horizon scenarios as shown in Table 13, Table 15, and Table 17. When the forecasting results were filtered by products, Product 1 showed the lowest MAPE score, notably having stable demand for all years with 10 full years of data. Table 19 shows the historical data available for each product. In contrast, the MAPE of Product 5, which had the shortest data count, scored the highest.

**Table 19**

*Data count for the sponsor company's products*

| Product ID | Demand History | Monthly Data Count |
|---|---|---|
| 1 | Jan. 2010 – Nov. 2019 | 119 |
| 2 | Nov. 2013 – Nov. 2019 | 73 |
| 3 | May 2014 – Nov. 2019 | 67 |
| 4 | Apr. 2015 – Nov. 2019 | 56 |
| 5 | Nov. 2017 – Nov. 2019 | 25 |
| 6 | Jan. 2015 – Nov. 2019 | 59 |

Figure 18 shows the forecasts becoming flatter as the forecasting period gets longer, e.g., Product 4 shows highly accurate prediction of trend and volatility in the first year of forecast but the forecasted demand does not show volatility after the 2-year timeline.

**Table 20**

*MAPE for four-year forecasting horizon*

| Product ID | SVM | | | | RAN | | | |
|---|---|---|---|---|---|---|---|---|
| | 2016 | 2017 | 2018 | 2019 | 2016 | 2017 | 2018 | 2019 |
| 1 | 80% | 77% | 81% | 82% | 6% | 24% | 2% | 5% |
| 2 | 58% | 74% | 80% | 81% | 52% | 61% | 77% | 80% |
| 3 | 29% | 17% | 40% | 48% | 18% | 29% | 64% | 69% |
| 6 | 437% | 167% | 78% | 57% | 40% | 30% | 77% | 80% |
| Avg. | 151% | 84% | 70% | 67% | 29% | 36% | 55% | 58% |

| Product ID | ANN | | | | LR | | | |
|---|---|---|---|---|---|---|---|---|
| | 2016 | 2017 | 2018 | 2019 | 2016 | 2017 | 2018 | 2019 |
| 1 | 4% | 23% | 7% | 6% | 3% | 24% | 12% | 15% |
| 2 | 30% | 71% | 54% | 49% | 50% | 49% | 59% | 49% |
| 3 | 9% | 48% | 37% | 78% | 30% | 6% | 17% | 8% |
| 6 | 78% | 66% | 376% | 531% | 8% | 220% | 83% | 168% |
| Avg. | 30% | 52% | 119% | 166% | 23% | 75% | 43% | 60% |

**Table 21**

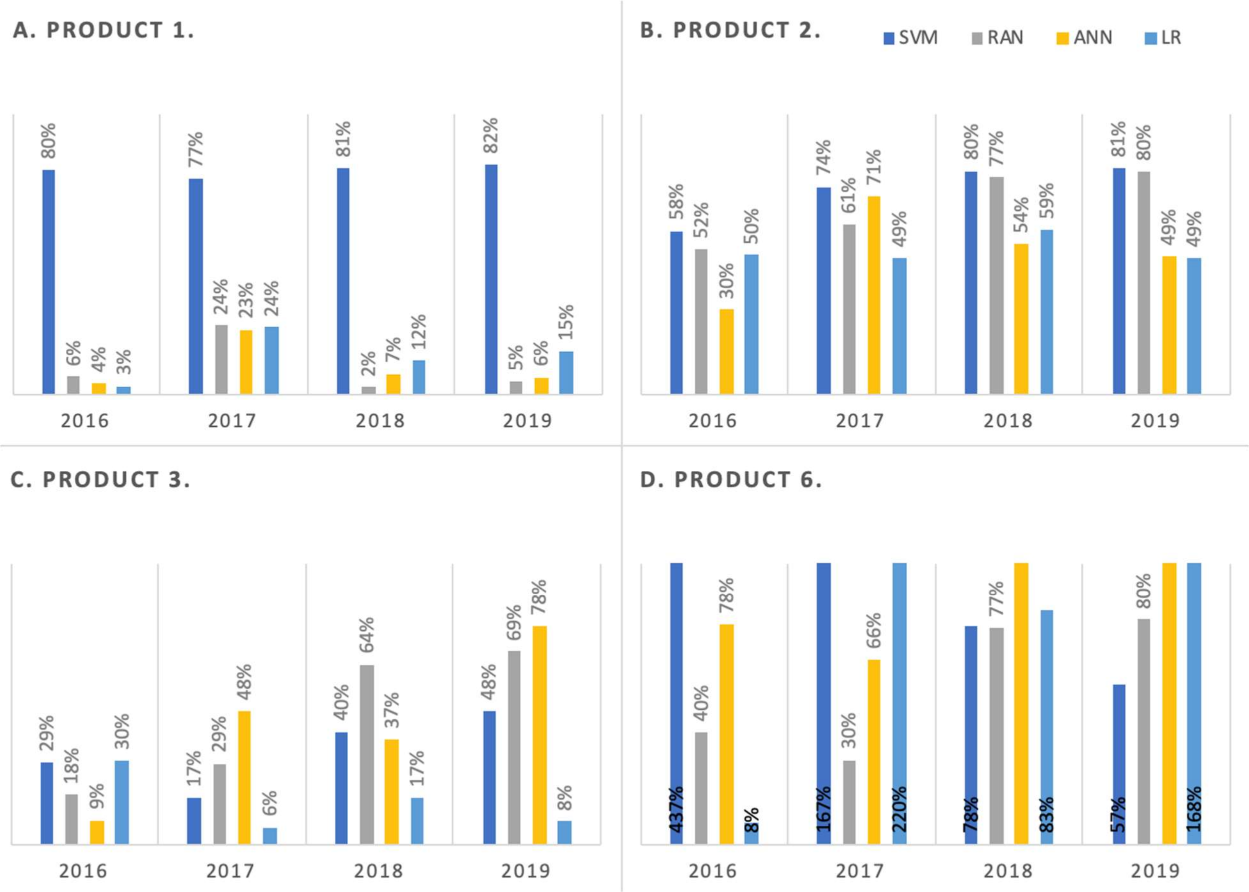
*MPE for four-year forecasting horizon*

| Product ID | SVM | | | | RAN | | | |
|---|---|---|---|---|---|---|---|---|
| | 2016 | 2017 | 2018 | 2019 | 2016 | 2017 | 2018 | 2019 |
| 1 | -80% | -77% | -81% | -82% | 6% | 23% | -2% | -5% |
| 2 | -58% | -74% | -80% | -81% | -52% | -71% | -77% | -80% |
| 3 | 29% | -17% | -40% | -48% | -18% | -48% | -64% | -69% |
| 6 | 437% | 167% | -78% | -57% | -40% | -66% | -77% | -80% |
| Avg. | 82% | 0% | -70% | -67% | -26% | -40% | -55% | -58% |

| Product ID | ANN | | | | LIN | | | |
|---|---|---|---|---|---|---|---|---|
| | 2016 | 2017 | 2018 | 2019 | 2016 | 2017 | 2018 | 2019 |
| 1 | 4% | 24% | -7% | -6% | 3% | 24% | -12% | -15% |
| 2 | -30% | -49% | -54% | -49% | -50% | -61% | -59% | -49% |
| 3 | -9% | 6% | -37% | -78% | -30% | -29% | -17% | -8% |
| 6 | 78% | 220% | -376% | -531% | -8% | 30% | -83% | -168% |
| Avg. | 11% | 50% | -119% | -166% | -21% | -9% | -43% | -60% |

**Figure 19**

*MAPE for four-year forecasting horizon by product*



As shown in Table 21, the models mostly underpredicted the demand for the products on the third year and the fourth year in the four-year forecasting horizon.

**Table 22**

*MAPE for five-year forecasting horizon*

| Product | SVM | | | | | RAN | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| ID | 2015 | 2016 | 2017 | 2018 | 2019 | 2015 | 2016 | 2017 | 2018 | 2019 |
| 1 | 80% | 80% | 77% | 81% | 81% | 2% | 4% | 20% | 0% | 6% |
| 2 | 26% | 58% | 74% | 79% | 81% | 40% | 73% | 84% | 88% | 88% |
| Avg. | 53% | 69% | 75% | 80% | 81% | 21% | 39% | 52% | 44% | 47% |

| Product | ANN | | | | | LR | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| ID | 2015 | 2016 | 2017 | 2018 | 2019 | 2015 | 2016 | 2017 | 2018 | 2019 |
| 1 | 8% | 23% | 56% | 43% | 47% | 4% | 12% | 37% | 23% | 25% |
| 2 | 66% | 88% | 92% | 94% | 94% | 54% | 76% | 76% | 71% | 63% |
| Avg. | 37% | 55% | 74% | 68% | 71% | 29% | 44% | 56% | 47% | 44% |

**Table 23**

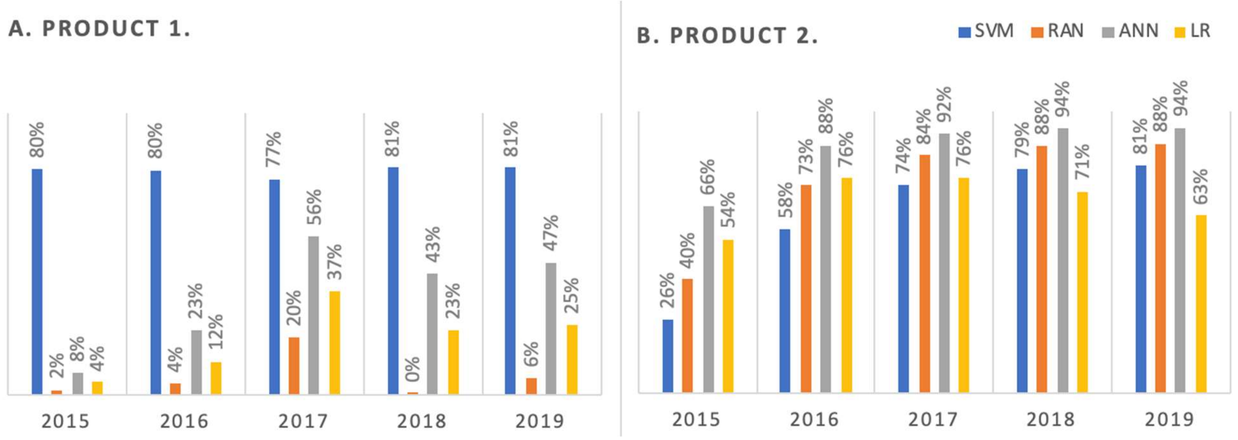
*MPE for five-year forecasting horizon*

| Product | SVM | | | | | RAN | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| ID | 2015 | 2016 | 2017 | 2018 | 2019 | 2015 | 2016 | 2017 | 2018 | 2019 |
| 1 | -80% | -80% | -77% | -81% | -82% | 2% | 4% | 20% | 0% | -6% |
| 2 | 26% | -58% | -74% | -79% | -81% | -40% | -73% | -84% | -88% | -88% |
| Avg. | -27% | -69% | -75% | -80% | -81% | -19% | -35% | -32% | -44% | -47% |

| Product | ANN | | | | | LR | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| ID | 2015 | 2016 | 2017 | 2018 | 2019 | 2015 | 2016 | 2017 | 2018 | 2019 |
| 1 | 8% | 23% | 56% | 43% | 47% | 4% | 12% | 37% | 23% | 25% |
| 2 | -66% | -88% | -92% | -94% | -94% | -54% | -76% | -76% | -71% | -63% |
| Avg. | -29% | -32% | -18% | -25% | -23% | -25% | -32% | -20% | -24% | -19% |

**Figure 20**

*MAPE for five-year forecasting horizon by product*



We continued to extend the forecasting time horizon up to nine years for Product 1, the only product that had enough historical data to train the model.

**Table 24**

*MAPE for nine-year forecasting horizon for Product 1*

| Year | SVM | LR | RAN | ANN |
|---|---|---|---|---|
| 2011 | 76% | 64% | 40% | 52% |
| 2012 | 76% | 618% | 47% | 147% |
| 2013 | 75% | 4217% | 53% | 319% |
| 2014 | 74% | 29998% | 57% | 602% |
| 2015 | 74% | 225795% | 56% | 1037% |
| 2016 | 74% | 1680408% | 60% | 1803% |
| 2017 | 70% | 13506661% | 85% | 3486% |
| 2018 | 75% | 79308067% | 54% | 4758% |
| 2019 | 76% | 478565200% | 44% | 7154% |
| Avg. | 74% | 63702336% | 55% | 2151% |

**Table 25**

*MPE for nine-year forecasting horizon for Product 1*

| Year | SVM | LR | RAN | ANN |
|---|---|---|---|---|
| 2011 | -76% | -64% | 40% | 52% |
| 2012 | -76% | -618% | 47% | 147% |
| 2013 | -75% | -4217% | 53% | 319% |
| 2014 | -74% | -29998% | 57% | 602% |
| 2015 | -74% | -225795% | 56% | 1037% |
| 2016 | -74% | -1680408% | 60% | 1803% |
| 2017 | -70% | -13506661% | 85% | 3486% |
| 2018 | -75% | -79308067% | 54% | 4758% |
| 2019 | -76% | -478565200% | 44% | 7154% |
| Avg. | -74% | -63702337% | 55% | 2151% |

Tables 24 and 25 display the results of the nine-year forecasting horizon.

# 5 Conclusion

The key research question for this project was to understand how machine learning can improve long-range forecasts. After researching current methods in machine learning and long-range forecasting, we applied four machine learning methodologies: support vector machine (SVM), random forest (RF), artificial neural network (ANN), and linear regression (LR) on real data provided by the sponsoring company. Key features affecting the demand were determined by SVM. These features were used to build long-range forecasting models using the four methodologies.

We used two types of approaches, direct and recursive, to develop multi-step long-range forecasting models. RF, ANN, and LR produced relatively accurate results in the one-step models. However, when extending the forecasting horizon using a multi-step forecast, the accuracy declines.

By observing the results of the feature selection process and comparing the results among our forecasting models, we conclude that historical demand of the individual product and product maturity level (lifecycle year) can be important features to consider when forecasting long-range demand for certain drugs. Additionally, we found that the machine learning model performance differed greatly based on data availability, forecasting horizon, and individual product.

## 5.1 Implications

With the direct forecast approach, we identified specific features relevant to product demand forecasting using machine learning. Contrary to the initial assumption, census data and product features failed to show high correlation with demand for individual pharmaceutical products. However, through the feature selection method, we discovered that “Lifecycle Year” was the most relevant feature out of 22 features used in the model. This shows that product maturity may be an important feature to consider in pharmaceutical demand forecasting.

The recursive forecast approach reaffirmed the importance of having time features in demand forecasting. When the 3-year recursive forecast result was compared to the direct method result, the MAPE score was significantly lower. Based on the recursive modeling’s forecasting results, we concluded that model performance is contingent on both the available training data sample size and the forecasting horizon. The rationale for this comes from the learning mechanisms of machine learning itself: as a model uses more data, the model can capture more granular patterns; accordingly, a bigger dataset will naturally produce better results. The forecasting model performance also differed based on the forecasting time horizon. As the horizon was set farther with no additional data being used, the model accuracy went down. Because recursive modeling uses forecasted results as an input for future forecasts, any forecasting error will continuously accumulate, thus reducing model accuracy.

One additional finding was that each forecasting methodology performed differently on each target product. For the products showing linear demand increase (e.g., Product 2, 3, 4), LR outperformed other methodologies. However, for forecasting products showing stable demand, random forest outperformed LR.

## 5.2 Limitations

Limitations of this research can be explained in two parts: data scarcity in modeling and model evaluation.

### 5.2.1 Data Scarcity

To utilize the features that are relevant, such as lifecycle year, abundant historical demand data for the entire product lifecycle is needed. However, our dataset consisted of only partial lifecycles. This is because we focused on products that had long lifecycles in a market with a limited number of products. The features that we used in the models were limited to the data that was available. Additional features, such as exogenous factors, that could affect demand were not incorporated into the model due to the data not being available. This is something that could be an area of exploration for future research.

### 5.2.2 Model Evaluation

Due to the data challenges in 5.2.1 we were not able to compare our model results with the company’s historical forecasting model output from that time, which makes it difficult to measure the performance and benefits of using this type of model over the existing forecasting models.

## 5.3 Future Research

Given the data requirements for using machine learning to forecast demand, only products that are more mature in their lifecycle stage should be forecasted using machine learning. For products that are newly introduced, future research should be focused on identifying the features that may relate to demand. Additionally, as every machine learning methodology has different advantages, a possible solution to certain limitations may be to develop multiple models for each demand category based on product features of the predicting timeline.

Lifecycle year may prove a fruitful starting point for future research. By categorizing the lifecycles into several similar groups (i.e., linear curve with a positive slope, negative slope, flat demand) the most appropriate methodology could be selected for each lifecycle category and built into models.

In order for a company to apply machine learning approaches to long-range demand forecasting, data management will be a real challenge. Applying these approaches would require long-term commitment and resources to accumulate the necessary data—in other words, the company must track all analogs related to changes in product demand. It would need cooperation between various stakeholders to manage the collected data. An added benefit of the new data is that it could be utilized not only for demand forecasting, but also for scenario analysis in internal decision making.

However, given that using a machine learning approach in long-term forecasting has inconclusive performance, and requires a commitment to establishing a data management program, a detailed cost-benefit analysis along with internal discussion is advised before pursuing further applications of machine learning in long-range demand forecasting.

## References

Alon, I., Qi, M., & Sadowski, R. J. (2001). Forecasting aggregate retail sales: A comparison of artificial neural networks and traditional methods. *Journal of Retailing and Consumer Services*, *8*(3), 147–156. https://doi.org/10.1016/S0969-6989(00)00011-4

Caplice, C. (2019). *SCM260_KeyConceptDocument_v5_PageNumberwDistTab.pdf*.

Cook, A. G. (2006). *Forecasting for the Pharmaceutical Industry: Models for New Product and In-market Forecasting and how to Use Them*. Gower Publishing, Ltd.

Ekonomou, L. (2010). Greek long-term energy consumption prediction using artificial neural networks. *Energy*, *35*(2), 512–517. https://doi.org/10.1016/j.energy.2009.10.018

Faggella, D. (n.d.). *7 Applications of Machine Learning in Pharma and Medicine*. Emerj. Retrieved November 12, 2019, from https://emerj.com/ai-sector-overviews/machine-learning-in-pharma-medicine/

Hong, W.-C. (2009). Electric load forecasting by support vector model. *Applied Mathematical Modelling*, *33*(5), 2444–2454. https://doi.org/10.1016/j.apm.2008.07.010

Hribar, R., Potočnik, P., Šilc, J., & Papa, G. (2019). A comparison of models for forecasting the residential natural gas demand of an urban area. *Energy*, *167*, 511–522. https://doi.org/10.1016/j.energy.2018.10.175

Lamberti, M. J., Wilkinson, M., Donzanti, B. A., Wohlhieter, G. E., Parikh, S., Wilkins, R. G., & Getz, K. (2019). A Study on the Application and Use of Artificial Intelligence to Support Drug Development. *Clinical Therapeutics*, *41*(8), 1414–1426. https://doi.org/10.1016/j.clinthera.2019.05.018

Liu, X. Q., Ang, B. W., & Goh, T. N. (1991). Forecasting of electricity consumption: A comparison between an econometric model and a neural network model. *[Proceedings] 1991 IEEE International Joint Conference on Neural Networks*, 1254–1259 vol.2. https://doi.org/10.1109/IJCNN.1991.170569

Merkuryeva, G., Valberga, A., & Smirnov, A. (2019). Demand forecasting in pharmaceutical supply chains: A case study. *Procedia Computer Science*, *149*, 3–10. https://doi.org/10.1016/j.procs.2019.01.100

Mohamed, Z., & Bodger, P. (2005). Forecasting electricity consumption in New Zealand using economic and demographic variables. *Energy*, *30*(10), 1833–1843. https://doi.org/10.1016/j.energy.2004.08.012

Peterson, R. T. (1993). Forecasting practices in retail industry. *The Journal of Business Forecasting Methods & Systems; Flushing*, *12*(1), 11.

Taieb, S. B., Bontempi, G., & Hyndman, R. J. (2014). *Machine learning strategies for multi-step-ahead time series forecasting*.

Vanfleteren, L., Fabbri, L. M., Papi, A., Petruzzelli, S., & Celli, B. (2018). Triple therapy (ICS/LABA/LAMA) in COPD: Time for a reappraisal. *International Journal of Chronic Obstructive Pulmonary Disease*, *13*, 3971–3981. https://doi.org/10.2147/COPD.S185975

Vapnik, V. (1998). The Support Vector Method of Function Estimation. In J. A. K. Suykens & J. Vandewalle (Eds.), *Nonlinear Modeling: Advanced Black-Box*

*Techniques* (pp. 55–85). Springer US. https://doi.org/10.1007/978-1-4615-5703-6_3