Identifying the Root Causes of Stockout Events in e-commerce Using Machine Learning Techniques

by

Tzu-Ning Chao
Bachelor of Business Administration

and

Federico Guillermo dos Santos Izaguirre
Bachelor of Science in Business Economics

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 2021

Signature of Author: _____
Department of Supply Chain Management
May 14, 2021

Signature of Author: _____
Department of Supply Chain Management
May 14, 2021

Certified by: _____
Josué C. Velázquez
Executive Director, Supply Chain Management Program
Capstone Advisor

Certified by: _____
Cansu Tayaksi
Postdoctoral Associate
Capstone Co-Advisor

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

Identifying the Root Causes of Stockout Events in e-commerce using Machine Learning Techniques

by

Tzu-Ning Chao

and

Federico Guillermo dos Santos Izaguirre

Submitted to the Program in Supply Chain Management
on May 14, 2021 in Partial Fulfillment of the
Requirements for the Degree of Master of Applied Science in Supply Chain Management

ABSTRACT

The year 2020 marked an unprecedented worldwide growth in e-commerce driven mainly by the COVID-19 pandemic. The lockdown restrictions created significant spikes in the demand for several products causing severe disruptions throughout the supply chains. The pandemic created significant challenges for companies to maintain efficiency in the supply chain and product availability on the digital shelves. Stockout events rose considerably in the online platforms, and companies across industries needed to find ways to address the problem.

The focus of this project was to identify the main reasons that lead to stockouts for the sponsoring company to a major online retailer and to develop a model to predict the stockouts. Using supervised machine learning models, we developed a model that predicts the missing order quantities for every specific order. The analysis shows that the variables associated to the demand such as order quantity have a higher impact than variables associated with the supply, such as inventory on hand. Additionally, the product categories and brands associated with each category play an important role in the stockout prediction. With the continued growth of e-commerce and customers changing their shopping preferences, our predictive model will help the sponsoring company analyze the orders and make informative decisions to predict the stockouts and improve the inventory allocation.

Capstone Advisor: Josué C. Velázquez
Title: Executive Director, Supply Chain Management Program

Capstone Co-Advisor: Cansu Tayaksi
Title: Postdoctoral Associate

ACKNOWLEDGMENTS

First, we want to thank Dr. Josué C. Velázquez and Dr. Cansu Tayaksi for guiding us in the right direction, providing helpful feedback, and brainstorming with us when we encountered obstacles throughout the project.

We also want to thank our sponsoring company for such interesting project, especially Neil Ackerman, Nitza Pierce, Denise Miller, Andrew Yeh, and Jeffrey Khoudary. They were very supportive, and not only shared the data required for analysis, but also answered all questions we had throughout our analysis, which was instrumental to the success of this project.

Furthermore, we need to thank Pamela Siska for helping us with the writing. She helped us restructure the paper and make it more concise. Finally, we thank our classmates who cheered and supported us throughout the process. Especially, the help and guidance on Python debugging, machine learning modeling, and engaged conversation that sparked new ideas.

We are very grateful for all the support and valuable feedback on this master's capstone project.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

1 INTRODUCTION

In 2020, US e-commerce sales increased by 44% in comparison to 2019 reaching more than USD 861 billion in sales, boosted heavily by the COVID-19 pandemic (Digital Commerce 360, 2021). The increasing demand in e-commerce during the pandemic generated important challenges for businesses across industries to maintain product availability on the digital shelves and consumer health products were not an exception. With the globalization of pharmaceutical production, the supply chains rely on manufacturing plants across the world (Socal, Apr2021). In 2020, only 28% of manufacturing facilities making active pharmaceutical ingredients to serve the US market were based in the United States (Piatek, Ning, & Touchette, Nov 2020), and the pandemic made the supply chains very vulnerable, especially with the governments of the manufacturing countries imposing different restrictions.

With a spike in demand of more than 100% for consumer health products in 2020 in comparison to 2019, manufacturers struggled to keep product availability (Ivanov, 2021). A stockout is defined as an event when the stock on hand drops to zero (Silver, Pyke, & Thomas, 2017), and the reasons for stockouts vary. Stockouts can occur when the demand for a certain product is higher than expected or when the inventory on hand (including the safety stock) is too low to fulfill the order. Stockouts can lead to lost sales, the increase the likelihood of customers switching brands, reduces customer satisfaction, and can impact brand loyalty in the long term.

The negative impact of stockout on e-commerce is even more significant than on brick-and-mortar, which are physical stores to which customer can visit (Ward, 2020), as it is easier for customers to switch to other vendors online, and the drop in sales ranking will in return further lead to a decrease in a store's online visibility.

Given the importance of stockouts in e-commerce, in this capstone we conducted an in-depth statistical analysis to better understand the root causes of stockouts for a leading pharmaceutical company that serves a key online retailer. We then built a predictive model using machine learning models to predict stockout events.

1.1 Problem Statement

The focus of this research is to find out the causes that led to stockout events in 2020 for the e-commerce business of a pharmaceutical consumer health products company in the United States. The company aims to understand the reasons of stockout in the rapidly growing e-commerce sector to improve its supply chain and better serve its customers. Therefore, this capstone has three main objectives:

1. Determine the key drivers of stockout in the Company's e-commerce through the online retailer's platform.
2. Apply machine learning techniques to understand the relationships between predictor variables and stockout.
3. Provide insights and recommendations for improving fulfillment of e-commerce.

1.2 Company background

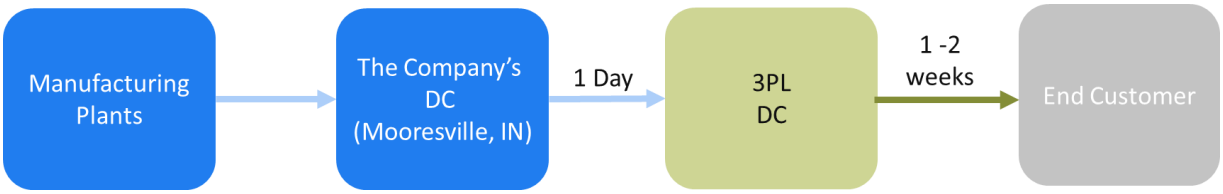
The Company is one of the leading pharmaceutical companies in the world with more than USD 3.6 billion in sales in the consumer health sector and a portfolio of over 350,000 SKUs. The Company's e-commerce business is divided into two main operation models: Direct to Customer (DTC) and online retail stores, such as Amazon, Walmart, or Target. Until 2019, the company's e-commerce business was growing at an

average rate of 35% per year. The effect of the COVID-19 global pandemic in 2020 marked an unprecedented growth of 85% in e-commerce, with spikes in the demand for several products in the Company's portfolio including well-known brands. The Company's supply chain experienced significant challenges during 2020, and with the company unable to serve its customers, buyers may either turn to other retailers or third-party vendors on online retailers to fulfill their needs, leading to losses in sales, increases in chargebacks, and decrease in overall customer satisfaction. In addition, search ranking on customers' websites drop when orders are not fulfilled and to counteract the search effect, the Company has to increase its advertisement spending. As a result, managing the stockouts and improving e-commerce performance is an imperative task for the Company, as consumer behavior evolves in a pos-COVID-19 pandemic era.

As illustrated in Figure 1, DTC serves only for two product lines and is fulfilled by one distribution center and operated by a third-party logistic (3PL) operator for picking, packing, and shipping. In general, the delivery lead time from the Company's distribution center to the 3PL's distribution center is one day and it takes one to two weeks for delivery to end customers.

Figure 1

Value stream map for DTC fulfillment



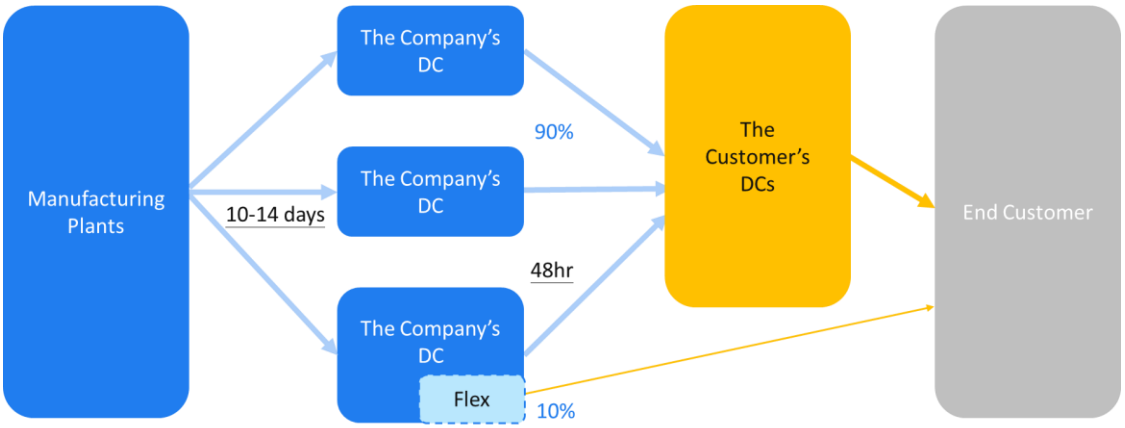
The Company's main e-commerce business is operated through large retailers, including a key online retailer. Three of the Company's distribution centers are involved to fulfill consumer product orders. The

finished goods are delivered from Company's distribution centers to customers' distribution centers. Since the scope of our project is to analyze the unfulfilled orders from a key online retailer, the information collected, and value stream map detailed in Figure 2, is specifically for serving this key online retailer's orders. The online retailer places orders on a weekly basis, and the Company must fulfill the order within 48 hours, including order reconciliation and delivery in pallets to the retailer distribution centers. As the Company does not allow backorders, the current Item Fill Rate (IFR) is around 80%, calculated by dividing the fulfilled order quantity by total order quantity for each item. The chargebacks for unfilled orders are confirmed by both companies through the order reconciliation process.

Apart from the traditional distribution network, 10% of the sales to the online retailer have been handled via the "flex" model, in which the online retailer operates independently inside one of the Company's distribution centers, handling picking and packing the orders for delivery to end customers. As Flex's space is limited, it mainly serves top-performing products with high turnover. The internal operation is set up to achieve quicker replenishment, shorter time to serve end customers, and cost-saving on transportation.

Figure 2

Value stream map for the online retailer order fulfillment



This capstone project is divided into six main sections. In the Introduction, we introduce the reader into the company, the problem, and the importance of addressing the problem. In the Literature Review we present the research that has been done in the field and how we applied it to our analysis. In the Methodology section we dive deep into our analysis, our approach to the problem, the methods we used, and how we suggest solving it. We also include hypothesis testing and both multiple linear regression and logistic regression models. We then present the Results of our analysis for the different models we developed. Finally, we include a Discussion section in which we present the key insights from the results, the limitations of our models, and suggestions for future research. Finally, our Conclusion section on which we summarizes the complete capstone project.

2 LITERATURE REVIEW

We divided the relevant literature into three categories: inventory management in e-commerce, causes of stockouts in the retail industry, and application of machine learning in stockout analysis.

2.1 Inventory Management in E-commerce

Supply chains are evolving rapidly in a fast-growing e-commerce environment and inventories play a relevant role for every retailer, both to serve the customer and to measure business performance. In the past years, inventories have been declining as a percentage of sales, shifting downstream, closer to the customer and companies are innovating to gain efficiencies in inventory management (Silver, Pyke, & Thomas, 2017). To understand the impact of e-commerce in the supply chain and how it is affecting the companies' strategic decisions, Agatz et al. (2008) introduces the differences when designing the supply chain for e-commerce, brick-and-mortar, or omni-channel sales structure.

An important aspect of e-commerce is data generation and real time processing of data. This allows the retailer to manage a more flexible pricing policy and even dynamic pricing. It is then the retailer's decision how to manage the pricing if they operate using an omni-channel strategy, and if the pricing of on-line and in-store prices are the same. Promotions are another important aspect of on-line sales, in which it is more precise and faster to do an on-line promotion than across a network of brick-and-mortar stores.

As Agatz et al. (2008) explains, another relevant subject in handling the supply chain for on-line sales is the inventory management. Inventory management is a branch of business management that studies and develops models to maintain desired stock levels for specific product or items (Toomey, 2000). Inventory's relevance in a Company's supply chain is mainly due to its main function, which is serving the final customer. As in-store space is limited for the brick-and-mortar model, it is cheaper to have a larger product selection and higher inventory levels in a centralized fulfillment center or a warehouse than

through retail stores by creating a pooling effect. Risk Pooling can help reduce the total variability of the demand and lead time, and thus helping reducing the uncertainty and risk in the system.(Oeser, 2015)

According to As Agatz et al. (2008), a key strategic decision for e-commerce operations is the design of the fulfillment network, addressing three fulfillment strategies.

1. Integrated fulfillments, which is having the e-commerce operation integrated into the traditional warehouse operation to serve a brick-and-mortar operation.
2. Dedicated fulfillment: which addresses a complete redesign of the traditional warehouse and adequate operation to have e-fulfillment capabilities only. The fulfillment centers are designed to serve on-line retail sales, small order quantities with picking and packing capabilities.
3. Store fulfillment: the traditional warehouse model to serve brick-and-mortar stores. Orders are prepared in advance and in bulk to serve the stores inventories.

Considering the sustained growth of e-commerce, it is important that companies understand how retail is evolving and adapt their strategic supply chain decisions. Inventory management is a key aspect of the supply chain to respond the customers' needs and critical to every retailer's operation.

As part of the research, an important part of our analysis focuses on understanding the Company's operation in the e-commerce business and how the inventory management could potentially affect the stockouts. We developed and tested a hypothesis defining inventory as a key variable and the reviewed literature guided us in the inventory analysis.

2.2 Causes of Stockouts in the Retail Industry

In a worldwide study on out-of-stock (OOS) of fast-moving-consumer-goods (FMCG) products in the retail industry, inaccurate forecasting and shelf-replenishment are found to be two major causes of OOS

(Corsten & Gruen, 2005). In the US, over 73% of the stockout events resulted from poor store operation, including store forecasting, store ordering, and store shelving. The authors further identified 43 reasons that lead to stockouts, which are categorized into 3 processes – planning, ordering, and replenishing. This happens over multiple supply chain levels, including stores, distribution centers, wholesalers, and suppliers. The distribution center and wholesaler level are more relevant to our capstone as we looked into the replenishment from the Company's distribution centers to the online retailer's distribution centers. However, supplier and store level are out of our capstone scope. Therefore, we used Corsten & Gruen's (2005) concept of the root causes to guide our discovery and form hypothesis with order data from those online-retail customers.

Another research used both qualitative and quantitative analysis to find out root-causes of stockouts under promotion and non-promotion period for a CPG (Consumer Packaged Goods) company and its retail stores (Nigam, 2016). Nigam (2016) concluded that forecast error, sales volume, and price have correlation with OOS frequency. The number of SKUs carrying per store and replenishment patterns, however, have no significant relationship with stockouts. Delivery systems, product sales, store characteristics are variables in another research on CPG products in retail that tested the relationship with OOS probability by applying probit regression (Milićević et al., 2018). Delivering through retailer DC results in a higher OOS probability than DSD (Direct Store Delivery) system, in which suppliers bypass retailers for product replenishment. Products with slow turnover rate and high variation in sales also increase the OOS probability. Another study on multi-echelon supply chain and its impact on inventory level and stockout ratio turned different supply chain network models and forecasting methods into binary variables, concluding that the forecasting techniques have a significant impact on the stockout rate (Wan & Evers, 2011).

Though there are not a lot of research on stockout in e-commerce, we referred to the research on causes of stockout in the retail industry to form our hypotheses, gather data, and create key variables for our

predictive model. The key difference between e-commerce and the traditional retail is that the order is placed through online platform instead of brick-and-mortar, and our research can prove whether the same variables including in previous research are also significantly impacting stockout in e-commerce.

2.3 Machine Learning in Stockout Analysis

Machine Learning (ML) is one of the most studied fields in Artificial Intelligence. It involves the study and development of computational models for learning processes on which the objective is to make the computers capable of improving their performance with practice and of acquiring knowledge on their own (Michel et al 1986).

Machine Learning comprises two main types of learning. In **supervised Learning,** the ML algorithm receives prelabeled input examples and intends to converge to the best as possible classifier $f : X \rightarrow Y$, so one can predict labels for unseen examples with high accuracy. **Unsupervised learning** is associated with the process of building up models after analyzing the similarities among input data. For example, the clustering algorithm K-Means attempts to find k representative groups according to the relative distance of points (de Mello and Ponti. 2018).

As discussed by (Sen & Srivastava, 1990)**,** regression is any method of fitting equations to data. The equations are valuable for two main purposes; making predictions about the data and judging the strength of relationships between the variables. Regression belongs to supervised machine learning algorithms and linear regression and logistic regression are the most commonly used models in research on analyzing causes of stockout events. The key difference is that linear regression requires the dependent variable to be continuous while the dependent variable of logistic regression is binary, suggesting the impact of each variable on the odds ration of the observed event (Sperandei, 2014). When the dependent variable is modeled as the occurrence of stockout, probit regression and logistic regression are great candidate.

Milićević et al (2018) further used Bayesian information criterion (BIC) and Akaike information criterion (AIC) to evaluate both probit regression and logistic regression on his study of stockout in retail and concluded that probit regression presents a better model fit. Usman (2008) used both supervised and unsupervised machine learning algorithms to test the drivers of stockout of retail stores, including multiple linear regression, logistic regression, and K-means. With the same independent variables input, the dependent variable of multiple linear regression is the number of stockout events for one store while the dependent variable of logistic regression is the occurrence of a stockout at the store. K-means, an unsupervised machine learning algorithm, was applied to cluster retail stores with similar characteristics and average stockout, then found out the common characteristics among stores with similar performance.

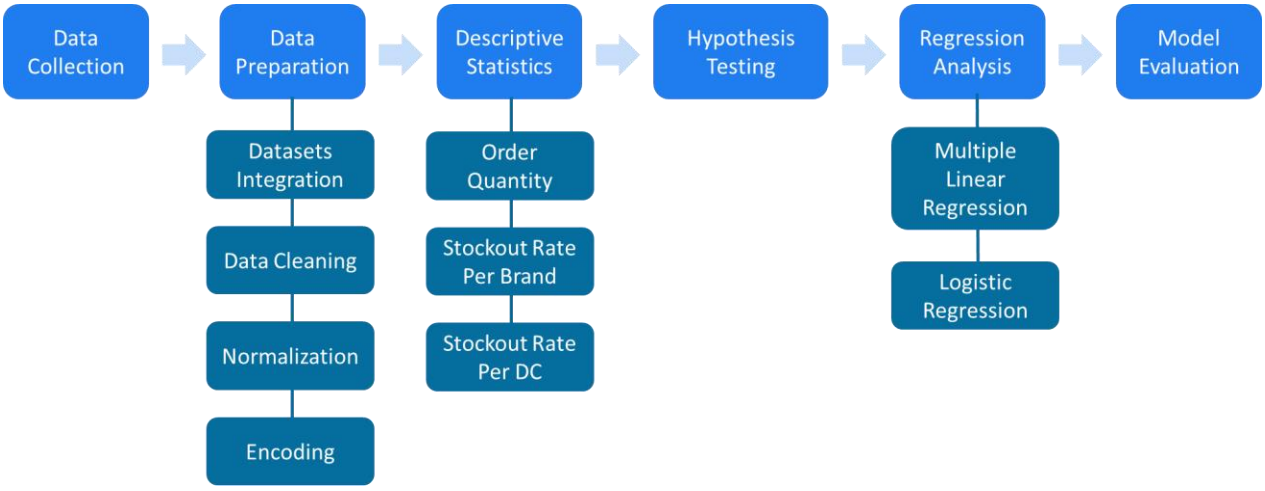
Research related to determining the causes of stockout event used both supervised and unsupervised machine learning algorithms. Different approaches were taken according to the goal of the research and the format of the datasets. Regression models presented in the previous researches served as a straight-forward tool to determine the relationship between the causes and stockout and make future predictions. Therefore, we referred to both multiple linear regression and logistic regression to build our predictive model and analyze the impact of each cause on stockout.

3 METHODOLOGY

In this chapter we discuss our approach to addressing the problem of the Company's stockout events during the COVID-19 pandemic. As previously introduced, the analysis of this project is focused on the Company's e-commerce business through the online retailer's platform. Using and analyzing the data for several specific product sales, the main objective is to determine the reasons for the stockout events and create a methodology to avoid them. Our approach consists of six distinct steps, shown in Figure 3.

Figure 3

Step-by-Step Methodology



The first step to approach the analysis included the data collection phase. An key aspect of the data collection is to clearly understand what would be the relevant data for the research. After receiving the data, we conducted the data preparation, including datasets integration, data cleaning, normalization, and encoding, as discussed in Section 3.1. Afterwards, in Section 3.2, we reviewed descriptive statistics, including comparison of order quantity, stockout rate per brand, and stockout rate per DC in 2019 and 2020. Hypothesis testing was developed based on the observations of data and detailed in Section 3.3. Lastly, data were aggregated to a weekly level and trained with multiple linear regression and logistic

regression, the variables used, assumption testing, and model evaluation methods were detailed in Section 3.4.

3.1 Data Preparation

3.1.1 Data

To further analyze the drivers that lead to stockouts, the data we collected included all customers' orders to the Company, inventory level in the distribution centers, and product information. We used data from January 2019 to December 2020 to conduct descriptive analysis, comparing the fulfillment performance in both years, especially these two years represents the company's ecommerce sales to the major online retailer before and after COVID outbreak. Then we used data from March 2020 to November 2020 to develop regression models with the following datasets:

1. **Online retailer orders**

The Company's orders from the online retailer is the base of our analysis, which includes the SKU number, product information, the quantity ordered, and the fulfilled quantity during each week, from which DC that the order is fulfilled, and the reason this order was not fulfilled. There are six main "Cut Reasons" – which are the reasons the orders were not fulfilled completely:

1) Inventory cuts: The order is cut due to insufficient inventory.
2) Block: The order is cut due to insufficient "ATP inventory," the inventory is blocked by other orders come beforehand.
3) Warehouse cuts: The most common reason in this category is when the order is too big to fit on a truck. It is challenging to load products with different sizes and shape onto the truck, and sometimes there may be wasted space on the truck.

4) Pricing: The order is cut when the online retailer places an order with the price lower than the Company's list price.
5) Order before STT: "STT" stands for "ship to trade date." The order is cut since the product is not available yet. This happens only to new products as new products are set up in the online retailer's system before launch to ensure a smooth launch, but the product is not available yet.
6) Obsolete: The order is cut since the product is no longer produced.

To further analyze the stockout events, we calculated the missed order quantity by subtracting quantity fulfilled from order quantity. Since the inventory data we got was the inventory level for each SKU in each distribution by week, we aggregated the order data into the same format, summing up the order quantity for the same SKU in the same week and counting the total unfulfilled transactions under each cut reasons. After that, we also calculated SO% (the percentage of missed quantity by dividing the missed order quantity by order quantity) which gave us the information about the magnitude of the stock out - 0 indicated order fulfilled while 1 indicated that the entire order was unfulfilled. With the completed order data in 2019, we calculated the average order quantity of each SKU in 2019 and compared the order quantity in 2020 with the average in 2019.

2. <u>Other customers' orders</u>

Another dataset we used was the orders from all other customers, excluding the online retailer orders and not limited to e-commerce. This dataset also provided the same information about each order. Combining both datasets (orders from the online retailer and orders from other customers) gave us an idea of the total weekly demand. We further divided the total of other customers order of the same SKU in the same week by the online retailer order to represent the magnitude of the competing order.

3. **Inventory level**

This dataset includes the weekly inventory level for each product in each distribution center. Three distribution centers that serve the e-commerce orders are Tobyhanna, PA(DC13), Fontana, CA (DC16), and Mooresville, IN (DC19). The inventory level we used is Available to Promise (ATP), the inventory that an incoming order can take, excluding the inventory allocated to a previous order or being processed.

4. **Product information**

The product information data provided us information on each product, including the brand and the franchise and category it belongs to. We combined this data with customers orders' data to see the fulfillment performance in each group. We further calculated the average stockout rate of each brand in each week by dividing the brand's total unfulfilled order by total unfulfilled order in the week. Brand of each product was also transformed to binary variable for model building.

After integrating the above datasets, we got 193,260 rows and each data point consists of the following columns: Week, SKU Number, DC, Brand, Order Quantity, Cut Quantity, Percentage of Cut Quantity (SO%), Inventory, Other Customer's Order Quantity/Online Retailer Order Quantity, Order Quantity / 2019 Average Order Quantity, SO% of Brand, and Brand (binary).

3.1.2 Data Cleaning

Null data should not be included in regression analysis, so we checked the missing values for all columns. 118,660 rows were null in inventory column, mostly because the inventory data in our dataset started from week 14 in 2020 and the online retailer order dataset started from week 1 in 2019. After dropping the rows due to time mismatch, we had 74,600 rows left, with missing values in "Other Customer's Order Quantity/Online Retailer's Order Quantity", "SO_Brand", and "Order Quantity / 2019 Average Order Quantity" columns. Both ""Other Customer's Order Quantity/Online Retailer Order Quantity" and "SO_Brand" columns were replaced with 0 since they came from calculations of order data and there

were no orders of the SKU at that week. Rows with missing values in "Order Quantity / 2019 Average Order Quantity" column were dropped since these SKU did not exist in 2019 order data and thus are not comparable. After dealing with the null data, we got 68,862 rows to use proceed with machine learning models.

Following, we checked the outliers and found out that the order unfulfilled would be treated as outliers with IQR method since there are 25% of the orders were not fulfilled completely. Therefore, we decided not to remove the outliers. The final dataset which was used for the regression analysis contained 68,862 data points.

3.1.3 Normalization

Normalization is an approach to transform numeric data when dealing with parameters of different

units and scales. Among all the methods for data normalization, Min-Max normalization is best suited for the case where the maximum and minimum values of the parameters are known (Jain et al., 2005). Min-Max normalization performs a linear transformation and the minimum and maximum scores are shifted to 0 and 1 with: $x' = \frac{x - \min(x)}{\max\ (x) - \min(x)}$ (Al Shalabi & Shaaban, 2006). Since the range of our key independent variables are very different (*Table 1),* we used Min-Max scaling to normalize the values of numeric columns in the dataset to a common scale. We used normalized data to proceed with regression model to ensure correct interpretation of coefficients.

Table 1

Descriptive Statistics of Key Independent Variables

	Order Quantity	ATP Quantity	Other/Amazon	Order Qty2020/2019 Avg	SO Brand
count	68,862	68,862	68,862	68,862	68,862
mean	1,115	15,858	70.3	8.7	0.1
std	5,706	59,112	323.0	57.8	0.1
min	1.0	0.0	0.0	0.0	0.0
25%	36	1,245	0.2	0.5	0.0
50%	144	5,451	3.9	1.7	0.1
75%	684	15,325	31.0	5.4	0.2
max	809,952	2,234,601	18,146	8,026	0.4

3.1.4 Encoding

Binary Variables: Are variables converted values such as 0 and 1 which represent the presence or absence of that variable in the data. We converted some variables into binary variables so the both multiple linear and logistic regression model could work correctly. The variables converted to binary were:

- Brands
- Product Categories: the Company has three product categories for their consumer health products, Skin Health, Self-Care, and Essential Health Products.
- Distribution Centers: Three distribution centers serve the e-commerce business.

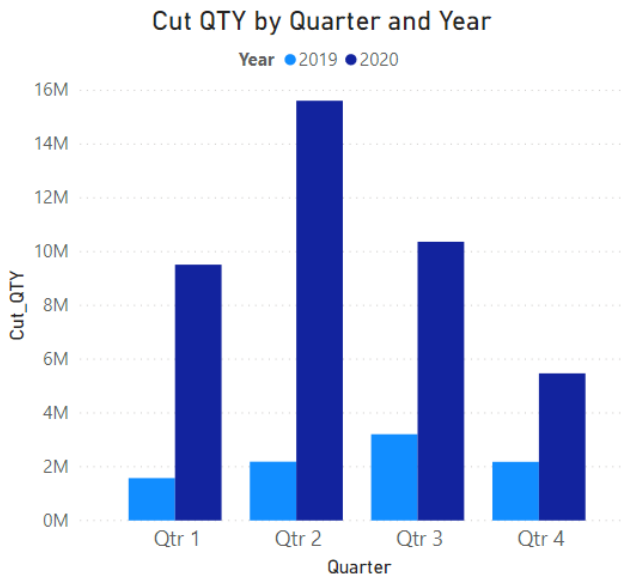
3.2 Descriptive Statistics

An important aspect prior to the analysis is to have a clear understanding of the data. For this, we conducted descriptive statistics to have a comprehensive knowledge of the dataset. An important first observation was that indeed the Company had a considerable increase of missing orders in the first

months of the COVID-19 pandemic, especially in Q2 (Figure 4). To our project, **Missing Orders** are orders that were placed to the Company but it was not able to fulfill

Figure 4

Missed Order per Quarter and Year



Through descriptive analysis, we were able to determine which brands were affected the most by missing orders, having Pepcid as the most affected product line (Figure 5). Additionally, we classified the different products using the Company's own product category classification. The three categories are: **Skin Health Products**; **Self Care,** and **Essential Health Products**. From our analysis, we concluded that the most affected product category was Self Care, followed by Essential Health Products and finally Skin Health *(Table 2)*.

Figure 5

Average of SO% by Brand and Year

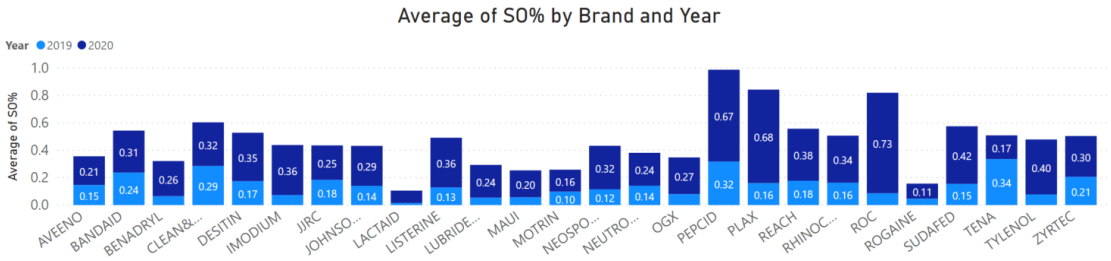


Table 2

Stockout ranking by product type

Product Category	Missed Order	Order Quantity	% Missed Order
Essential Health Products	8,326,843	39,698,601	21.0%
Self Care	7,245,666	18,875,558	38.4%
Skin Health	3,453,177	20,761,680	16.6%

3.3 Hypothesis Testing

Apart from the descriptive statistics, we applied both Multiple Linear Regression and Logistic Regression on our datasets to understand the influence of each variable on stockout rate. Hence, we developed the hypothesis below for later examination:

H1: The higher the order quantity in the same week, the higher the stockout rate.

Order quantity is aggregated by the quantity of each SKU in the same week. The measurement of the stockout rate, that is the dependent variable in our analysis, is the percentage of quantity missed.

Nigam (2016) observed correlation between sales volume and average OOS and Milićević et al. (2018) also found that products with high average sales were critical to OOS probability.

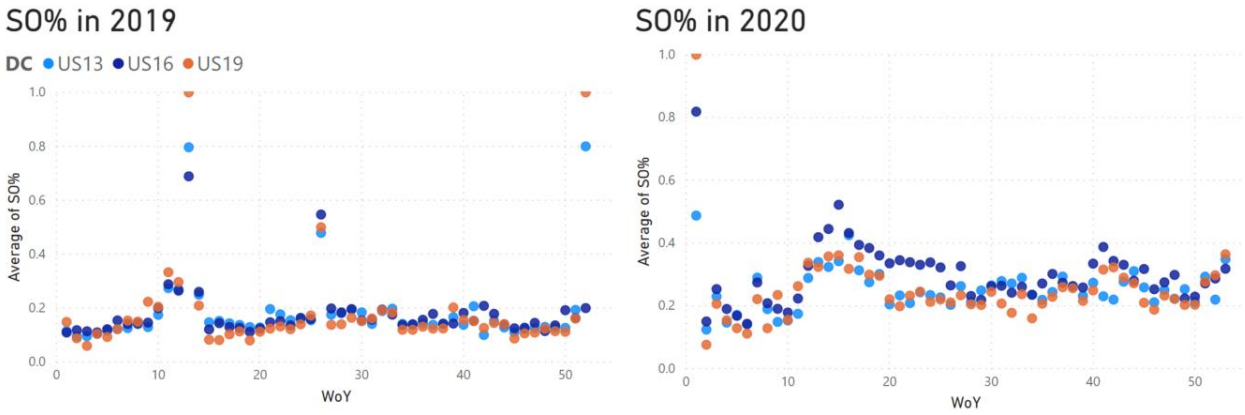
In 2020, there was an 85% increase in e-commerce sales. In Figure 6 and Figure 7, we plotted the sales volume and SO% with time in 2019 and 2020 respectively and also found a spike in demand and stockout in 2020. Moreover, the demand planning team found it difficult to forecast demand accurately as demand also fluctuated with price and promotion. Therefore, we believed that order quantity has positive impact on stockout rate.

Figure 6

Order Quantity in 2019 and 2020

Figure 7

SO% in 2019 and 2020

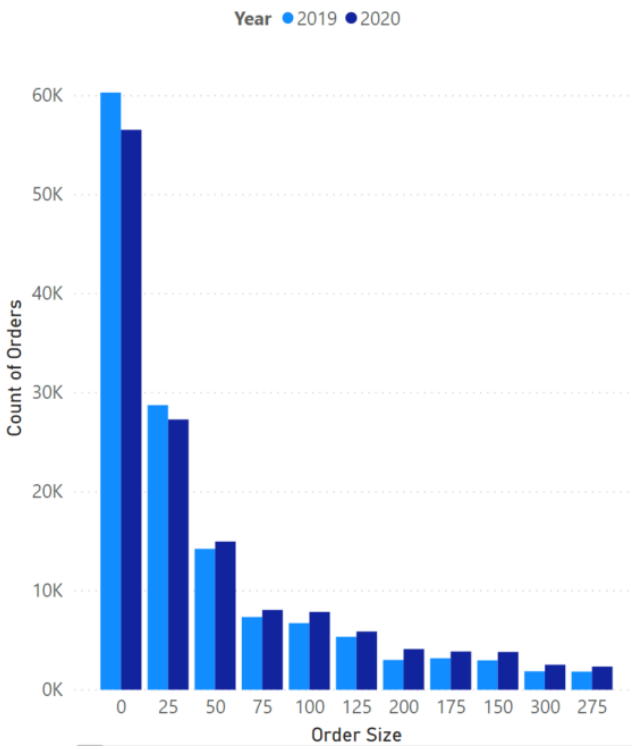


H2: The bigger the order size compared to the average order size in 2019, the higher the stockout rate.

Another observation from the online retailer's order data was that order with smaller order size decreased while large quantity orders increased compared to 2019 (Figure 8). Therefore, we believed an unexpected increase in order size has a positive relationship with stockout.

Figure 8

Order Size Comparison



H3: The higher the available inventory in the distribution center, the lower the stockout rate.

The availability of the inventory in distribution center is one of the key reasons of stock out (Corsten & Gruen, 2005). According to the order fulfillment policy between the Company and the online retailer, the available inventory plays an important role in stockout as the Company has to fulfill the online retailer's

order within 48 hrs. The available inventory is given by the company in a weekly snapshot. We then filtered the inventory stored in the distribution center that served the online retailer orders.

The company faced a shortage in raw materials, such as pumps for beauty products, due to the global shortage of ingredients for sanitizer. Besides, the capacity was also impacted since the company had to allocate down time to ensure the cleaning practices after COVID-19 outbreak. Therefore, we tested whether there was a negative correlation between available inventory and stockout.

H4: The higher the order quantity of the same SKU in the same week from other customers, the higher the stockout rate.

The company allocated its inventory to different customers' orders based on the requested fill rate and chargeback policy, so we believe that the more the competing order at the same week, the higher the chance of stockout for the online retailer's order.

H5: The higher the stockout of the same brand in the same week, the higher the stockout rate for other products under the same brand.

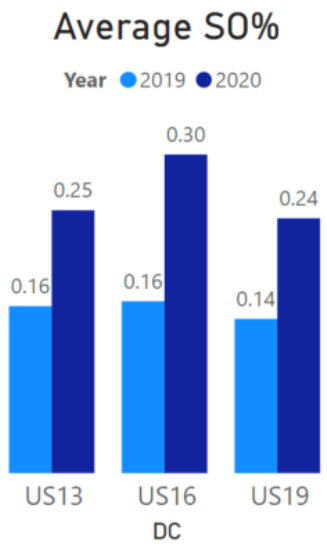
From the descriptive analysis above, we observed that some brands, such as Pepcid and Sudafed, have higher stockout rate on average. Besides, we learned from the Marketing team that the demand forecasting is made by brand, then break down to SKU level by planning team. Therefore, we believed that brand has certain impact on stockout rate. Hence, we calculated the average stockout rate of each brand in each week and used it to test the relationship with the stockout rate.

H6: The DC from which the order is fulfilled is related to stockout rate.

By comparing the stockout rate in different DCs in 2019 and 2020, we could observe that the performance of the different DCs was similar, with an average of 15% stockout rate. However, the performance of the three DCs varied in 2020, given that the overall stockout rate increased (Figure 9). Therefore, we believed that the DC from which the order is fulfilled is relevant to explain the stockout event.

Figure 9

Average Stockout Rate in Three Distribution Centers



H7: Product categories affect the missing order quantities and therefore, the stockout events

The company owns multiple brands and we found that stockout rate varies and order quantity among different brands. As Figure 10 shows, Pepcid and Plax had the highest average stockout rate in 2020. And

the order quantity for products under Neutrogena, Aveeno, and Tylenol were significantly higher than other brands (Figure 11). Therefore, we believed that brand plays an important role in determining stockout. The consumer health products are divided into three main categories and each brand falls under a product category. After analyzing the data, we believe that some product categories may have a higher impact on the stockout than others

Figure 10

Average Stockout Rate in 2020 by Brand

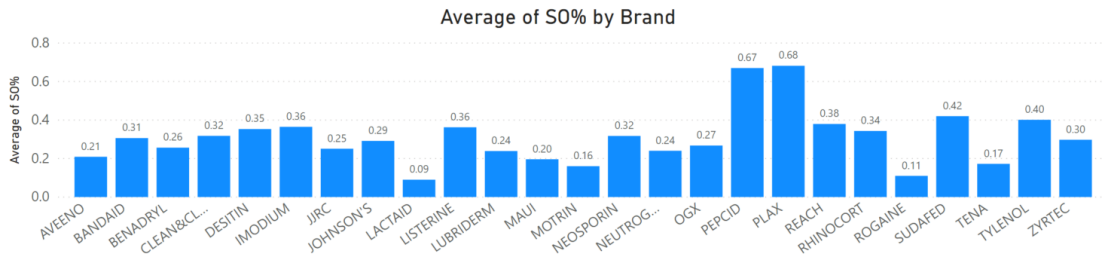
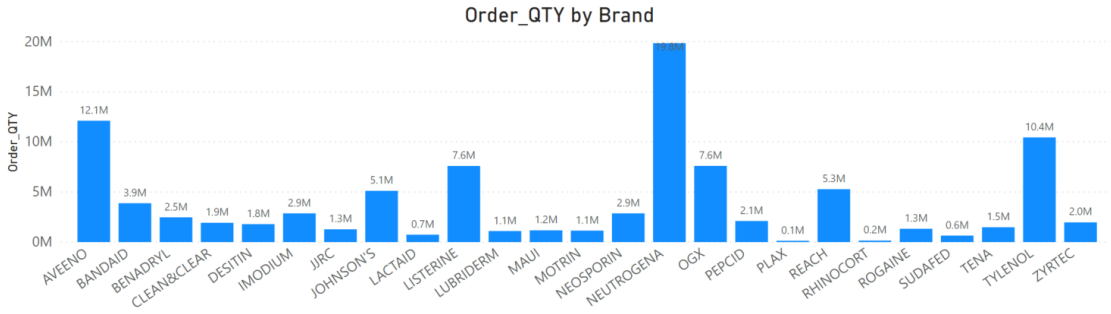


Figure 11

Total Order Quantity in 2020 by Brand



3.4 Regression Analysis

3.4.1 Multiple Linear Regression

The regression analysis focuses on finding a relevant relationship between the dependent variable *y* and the independent variables, or predictor variables. We started testing the independent variables using simple linear regression for each variable to see if they were statistically relevant or not. By conducting independent simple regression, we avoided falling into a multicollinearity situation. Multicollinearity happens when "the quality of estimates, as measured by their variances, can be seriously and adversely affected if the independent variables are closely related to each other"(Sen & Srivastava, 1990).

$$\hat{Y} = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \ldots + \beta_p X_p$$

X = Explanatory Variables or Independent Variable

Y = Dependent Variable

$\boldsymbol{\beta_0}$ = *Intercept*

$\boldsymbol{\beta_p}$ = Slope Coefficient for Each Predictor

The key assumptions for the linear regression models are:

- The true relationship is linear
- Errors are normally distributed
- Homoscedasticity of errors (or, equal variance around the line).
- Independence of the observations

We started the analysis by defining the dependent variable as the percentage of missing quantities that were not filled to the online retailer (Missing Order Quantity divided by the Total Order Quantity). After running several regressions using different independent variables to predict the percentages of missing

quantities, we concluded that the results did not show any correlation between the dependent and independents variables. The independent variables that were used for the analysis were: Normalized Inventory Levels at the Distribution Centers, Normalized Order Quantity, Which Distribution Centers served the order, Normalized magnitude of the previous online retailer order with respect to 2020, and Product Category. None of the results were statistically significant so we decided to switch the stockout analysis.

Our second approach focused on predicting the missing order quantities instead of the percentage of missing orders. The missing order quantities was defined as our dependent variable and the same independent variables were used for the analysis - Normalized Inventory Levels at the Distribution Centers, Normalized Order Quantity, Which Distribution Centers served the order, Normalized magnitude of the previous online retailer order with respect to 2020, and Product Category. We noticed that there were some variables that had strong correlations and statistically significant and others that were not. By narrowing down the scope of the analysis, were able to focus the project on the variables that are relevant and significant.

While focusing on the Missing Order Quantities, we developed our analysis on the 2020 dataset. We chose to use the 2020 dataset only is it is the most complete set, including inventory levels, and the one that could lead to greater insights.

Following, we split the data set between the training and testing sets. This way we trained the model with 70% of the datapoints to later test the model with the remaining 30%. By doing this, we can check the accuracy of the model and the predictions.

3.4.2 Logistic Regression

Similar to linear regression, logistic regression uses coefficient values for independent variables (x) to predict an output value (y). The key difference is that logistic regression is used for classification with two outcomes, so the dependent variable is a binary value instead of numeric value. In our capstone, we used logistic regression to build a model to predict the occurrence of a stockout event. The general equation of logit model is written below with k independent variables, the dependent variable is a logit, that is the natural log of the odds; while the coefficient (β) is the amount the logit (log-odds) changes with one unit change in x.

$$\ln\left[\frac{p}{1-p}\right] = \beta_0 + \beta_1 x + \beta_2 x_2 + \cdots + \beta_k x_k$$

Major assumptions of logistic are as follows:

1. The dependent variable should be binary.

   We used the same dataset for multiple linear regression and added additional column "Missed Order" as our dependent variable - 0 indicates order fulfilled and 1 indicates stockout.

2. The observations should be independent from each other.

   Each order from the online retailer to the Company is independent from each other.

3. The independent variables should not be correlated with each other.

   We checked the correlation among independent variables by Variance Inflation Factor (VIF).

4. The independent variables should be linearly related to the logit of the outcome, where the logit function is $logit(p) = \log(\frac{p}{1-p})$ and p is the probabilities of the outcome.

   We conducted Box-Tidwell Transformation Test with SPSS which included logistic model interaction terms in the model to test the linearity. Logistic model interaction terms were transformed by each independent times its natural logarithm ($\mathrm{x}(\ln(\mathrm{x}))$).

5. A large data set is required.

   688,62 data points would be used in the analysis, in which 17,321 are stockout event.

To run the logistic regression we used the SPSS software, and the model was evaluated in four ways:

1. **Omnibus Test:**

   The Omnibus test was used to check whether the new model with more independent variables is an improvement over the baseline model. In our capstone, we did not use stepwise logistic regression or blocking, so the values under chi-square are expected to be the same. However, the p-value of the chi-square test could determine whether the overall model is statistically significant.

2. **Nagelkerke R Square**

   This is one of pseudo-R-square statistics; however, it is not equivalent to the R square in OLS regression, which represents the proportion of the variance explained by independent variables. Therefore, we took it for reference but did not overly emphasize it.

3. **Confusion Matrix**

   A confusion matrix is a table that tells how accurately the model classifies the outcomes. We will focus on how many cases of stockout are correctly predicted and the overall percentage (the overall percent of cases that are correctly predicted by the model).

4. **Variables in the Equation Table**

   We looked at the coefficient of each variable to see the relationship between the independent variable and dependent variable, and the p-value tells whether the variable is significant. However, the dependent variable is on the logit scale, so we used Exp(B) (the Odds Ratio) to interpret how many times more likely it is that stockout will occur for every one unit increase in the independent variable.

4 RESULTS

In this section, we present the models that were used to determine which attributes are significant in stockout and which model is more accurate in predicting stockout. Different models were developed with different combination of attributes at the aggregate level and calibrated on the individual DC. In the end, we discussed the results and insights obtained from both analyses.

4.1 Multiple Linear Regression Results

The multiple linear regression formula that applies to our model is:

$$\hat{Y} = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \ldots + \beta_p X_p$$

where $\hat{Y}$ is the predicted or expected value of the dependent variables from 1 to p. β_0 is the intercept coefficient, meaning that it is the expected value of Y when every independent variable is equal to 0.

The model has numerical and binary variables. Binary variables included in multiple linear regression model are:

- Brands: after analyzing the brands with the highest stockouts, we selected the most representative brands;
- Product categories: Skin Health, Self-Care, and Essential Health Products.
- Distribution Centers: Three distribution centers serve the e-commerce business for the Company, including DC13, DC16, and DC19. The variables are shown as DC_US13, DC_US16, and DC_US19 respectively in the model.

All numerical variables used in the model were normalized, they are abbreviated as below:

- Online Retailer Order Quantities to the Company ➔ Order_QTY_n
- Inventory Position ➔ Inventory_n

- Magnitude of the Order Quantity in comparison to 2019 Order Quantities

  ➔ Order_Qty/2019Avg_n

Finally, the dataset to be used in the Multiple Linear Regression is the following:

Table 3

Multiple Linear Regression Dataset Sample

Order Quantity_n	Inventory_n	Order Qty2020/2019 Avg_n	ProductType_EssHealth	ProductType_Self Care	ProductType_Skin Health	DC13	DC16	DC19
0.00001	0.00575	0.00000	0	0	1	1	0	0
0.00535	0.00218	0.00099	0	0	1	0	1	0
0.00006	0.00051	0.00007	0	0	1	1	0	0
0.00058	0.00024	0.00066	0	0	1	0	1	0
0.00015	0.00187	0.00010	0	0	1	1	0	0

As the data was already normalized prior to the analysis, we did not re-scale the non-binary features. The binary variables are already values of either “0” or “1” that do not to be re-scale.

To build the model, we started by dividing the training dataset that analyzed 70% of the dataset into the dependent variables and the independent variables.

Following, we built the linear model and analyzed the results.

Table 4

Summary of the Initial Multiple Linear Regression Model of the Training Dataset

OLS Regression Results

Dep. Variable:	Cut_QTY	**R-squared:**	0.839
Model:	OLS	**Adj. R-squared:**	0.839
Method:	Least Squares	**F-statistic:**	4.492e+04
Date:	Mon, 19 Apr 2021	**Prob (F-statistic):**	0.00
Time:	17:13:20	**Log-Likelihood:**	-6.2268e+05
No. Observations:	68862	**AIC:**	1.245e+06
Df Residuals:	68853	**BIC:**	1.245e+06
Df Model:	8		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
Order_QTY_n	6.232e+05	1423.248	437.896	0.000	6.2e+05	6.26e+05
Inventory_n	-3354.1957	295.578	-11.348	0.000	-3933.528	-2774.864
Order_Qty/2019Avg_n	6.177e+04	1391.018	44.409	0.000	5.9e+04	6.45e+04
ProductType_EssHealthProducts	-109.2949	56.024	-1.951	0.051	-219.101	0.511
ProductType_SelfCare	-203.5431	59.857	-3.400	0.001	-320.863	-86.223
ProductType_SkinHealth	-93.5527	55.932	-1.673	0.094	-203.179	16.073
DC_US13	-352.2108	55.726	-6.320	0.000	-461.433	-242.988
DC_US16	-449.7228	55.561	-8.094	0.000	-558.622	-340.824
DC_US19	-182.3698	56.574	-3.224	0.001	-293.254	-71.486

Omnibus:	34833.356	**Durbin-Watson:**	1.745
Prob(Omnibus):	0.000	**Jarque-Bera (JB):**	952231908.752
Skew:	-0.139	**Prob(JB):**	0.00
Kurtosis:	579.086	**Cond. No.**	204.

First, we analyzed the "goodness of fit" of the model by analyzing the adjusted R^2. The adjusted R^2 takes into consideration the weighted proportion of all the predictors. In this first model, we got an R^2 of 0.839, which we considered it to be acceptable to continue our analysis.

Following, we tested each individual coefficient. For this, we established our hypothesis tests for each variable, setting the confidence intervals to 95%. The variables that have a p-value lower than 5% are considered statistically significant in our model.

We identified that two product categories; Essential Health Products and Skin Health Products had p-values higher that 0.05. To better fit the model we excluded the variables that were not statistically significant one by one and tested the model again after removing the variables. We started removing the

variables that had the highest p-values first, therefore the Essential Health Products and following, Skin Health.

Another important aspect to check while working on multiple linear regression is to check on the collinearity of the variables. For this, we analyzed the Variance Inflation Factor (VIF) which is a quantitative value that expresses how correlated are each feature with each other. Reasonable values of the VIF within the variables are less than 5.00. We calculated the VIF for the first model with all the included variables and determined and could see that some variables had high collinearity among each other.

Table 5

VIF Figures Before Removing the Statistically Insignificant Variables from the Multiple Linear Regression Model

Features	VIF
DC_US13	22.03
DC_US16	20.31
ProductType_SkinHealth	13.37
ProductType_EssHealthProducts	13.06
DC_US19	11.22
ProductType_SelfCare	5.70
Amazon_Order_QTY_n	1.76
Order_Qty/2019Avg_n	1.76
Inventory_n	1.01

After removing all the statistically insignificant variables (Essential Health and Skin Health Products), we analyzed the results and discovered that the model had an adjusted R^2 of 0.871 and all variables were statistically significant (p-value < 0). It is important to notice the R^2 improvement after removing the insignificant variables, from 0.839 to 0.871.

Table 6

Summary of the Resulting Multiple Linear Regression Model of the Training Dataset

OLS Regression Results

Dep. Variable:	Cut_QTY	R-squared:	0.871
Model:	OLS	Adj. R-squared:	0.871
Method:	Least Squares	F-statistic:	5.403e+04
Date:	Mon, 19 Apr 2021	Prob (F-statistic):	0.00
Time:	17:37:19	Log-Likelihood:	-4.3574e+05
No. Observations:	48203	AIC:	8.715e+05
Df Residuals:	48196	BIC:	8.716e+05
Df Model:	6		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
const	-331.9879	7.978	-41.611	0.000	-347.626	-316.350
Order_QTY_n	6.446e+05	1604.027	401.850	0.000	6.41e+05	6.48e+05
Inventory_n	-4136.9248	354.263	-11.678	0.000	-4831.285	-3442.565
Order_Qty/2019Avg_n	6.519e+04	1575.743	41.374	0.000	6.21e+04	6.83e+04
ProductType_SelfCare	-125.4394	30.595	-4.100	0.000	-185.407	-65.472
DC_US13	-138.7589	12.675	-10.947	0.000	-163.602	-113.916
DC_US16	-240.3329	13.000	-18.487	0.000	-265.814	-214.852
DC_US19	47.1039	16.299	2.890	0.004	15.158	79.050

Omnibus:	34600.218	Durbin-Watson:	2.008
Prob(Omnibus):	0.000	Jarque-Bera (JB):	397410764.521
Skew:	-1.615	Prob(JB):	0.00
Kurtosis:	447.812	Cond. No.	1.42e+16

After removing the variables that were not statistically significant, we analyzed the VIF results and found out that all values were below 5.0. Therefore, our model did not have multicollinearity.

Table 7

VIF After Removing the Statistically Insignificant Variables

Features	VIF
Amazon_Order_QTY_n	1.76
Order_Qty/2019Avg_n	1.75
DC_US13	1.10
DC_US16	1.09
DC_US19	1.03
Inventory_n	1.01
ProductType_SelfCare	1.01

Finally, we evaluated the definite model with the test data. This was the last step to finalize the analysis and find out the prediction accuracy of the model. The process consisted of using the 30% of the testing data with the training set and evaluate the performance.

The final model equation was as follows:

$$\hat{Y} = -332 + 644{,}600 \times X_1 + -4{,}136 \times X_2 + 65{,}190 \times X_3 - 125 \times X_4 - 139 \times X_5 - 240 \times X_6 + 47 \times X_7$$

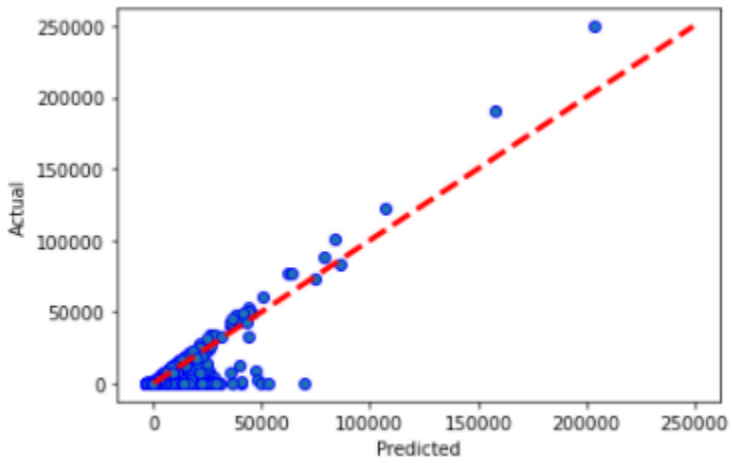
The testing model goodness of fit resulted in an R^2 of 0.63. Additionally, the Mean Absolute Error was 777 and the Mean Squared Error 4,333,758.

Model Performance Metrics	Result
R^2 Score - "Goodness of Fit"	0.63
Mean Absolute Error - MAE	777
Mean Squared Error - MSE	4,333,758

Following, an illustration of the Actual vs. Predicted values:

Figure 12

Multiple Linear Regression Model Applied to the Testing dataset



Key Insights from the Multiple Linear Regression Model:

- Essential Health Product and Skin Health product categories are not significant to predict the missing order quantities. The only product category that is statistically significant and relevant to make predictions is the Self Care product category. Some products in that category are Tylenol, Pepcid, and Sudafed. These products were also among the products with the highest missing order quantities during 2020.
- Order Quantities: The coefficient for the variable is positive, which means that the missing orders and Amazon's Order Quantity are positively related. When the online retailer increases the order quantities, it is more likely to have missing order quantities.
- Inventory: Inventory quantities, opposite from the order quantities, is negatively related to the missing order quantities. As the inventory position is higher, the probability to have missing orders is lower. The coefficient however is lower than that of the order quantity, which means that it has a lower impact on the missing order prediction than the order quantity.

- Order Quantity / 2019 Average Order Quantity: This coefficient is positively related to the missing order quantity and it explains how the magnitude of the order quantity at a specific time during 2020 in comparison to the average order quantity in 2019 affects the missing order quantity. We used this variable to see the impact of the growth in the online retailer orders during 2020 in comparison to 2019.
- Self-Care Product Type: This coefficient is negatively related to the missing order quantities and it is the only product category that is statistically significant in the model. To recall, the product category is a binary variable. The interpretation from the negative coefficient is that if the missing quantity from the product to be predicted falls in the Self-Care product category, the missing quantity would be less than if it is from other categories, such as Essential Health or Skin Care products.
- Distribution Centers: The Company has three distribution centers that serve the e-commerce business, and each performs differently with regards to predicting the missing order quantities. DC13 and DC16 are negatively related to the missing order quantities, which means that if the order is served from either of those Distribution Center, the missing order quantity would be less than if it is served from DC19. Interestingly, the magnitude of the coefficient is also different between DC13 and DC16, on which DC16 performs 73% better than DC13. On the contrary, the coefficient of DC19 binary variable suggests that orders served from DC19 will most likely increase the missing order quantity.

4.2 Logistic Regression Results

Four major models were developed by logistic regression algorithm. The first model included all 68,862 data points and the other three models were developed by looking at only one DC at a time, that is 28,454, 26,300, 14,108 datapoints for DC13, DC16, and DC19 respectively. An average of 25% of stockout was

observed in the dataset, a summary of SO% is shown in Table 8. DC13, DC16, and DC19 are shown in the model that represent DC13, DC16, and DC19 respectively.

Table 8

Average stockout rate and datapoints per DC

DC	SO%	SO	Total Orders
US13	0.25	7005	28454
US16	0.28	7468	26300
US19	0.20	2848	14108
Total	**0.25**	**17321**	**68862**

4.2.1 Model 1- All data

Six models were developed using the entire datasets, comparison of prediction accuracy is summarized in Table 10 and the detailed output results are given in the Appendix.

(1) Model 1-1:

$$\text{Stockout} = f\left(Amazon\ Order\ Quantity, Inventory, \frac{Other\ Order\ Quantity}{Amazon\ Order\ Quantity}, \frac{Amazon\ Order\ Quantity}{2019\ Average\ Order\ Quantity}, Brand\ SO\%\right)$$

The model is statistically significant because the p-value is less than .000 and all five coefficients (β) are significant. However, the model did a better job in predicting fulfilled order (99.1% accuracy) and fail to predict stockout (5.0% accuracy). Therefore, we added other variables to see if the accuracy of predicting stockout will increase.

(2) Model 1-2:

$$\text{Stockout} = f\left(Amazon\ Order\ Quantity, Inventory, \frac{Other\ Order\ Quantity}{Amazon\ Order\ Quantity}, \frac{Amazon\ Order\ Quantity}{2019\ Average\ Order\ Quantity}, Brand\ SO\%, DC(Binary)\right)$$

By adding DC as a binary variable, the accuracy in predicting stockout slightly increase. Moreover, coefficients of both DC binary variables are significant with p-value <.001.

Multicollinearity was also checked for logistic regression model. As Table 9 shows, VIF values of all variables in the model are below 5, suggesting that no strong correlation was found among given variables. Therefore, we believed that coefficients of our logistic regression models are reliable.

Table 9

VIF of Model 1-2

Coefficients[a]

Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.	Collinearity Statistics	
		B	Std. Error	Beta			Tolerance	VIF
1	(Constant)	.214	.004		49.664	.000		
	Amazon_Order_QTY_n	6.983	.298	.113	23.406	<.001	.606	1.649
	Inventory_n	-.637	.062	-.039	-10.232	<.001	.987	1.013
	OtherAmazon_n	.678	.093	.028	7.307	<.001	.984	1.016
	Order_Qty2019Avg_n	-.087	.291	-.001	-.297	.766	.609	1.642
	SO_Brand	-.131	.018	-.028	-7.256	<.001	.983	1.017
	DC1	.044	.004	.050	9.972	<.001	.561	1.783
	DC2	.073	.005	.082	16.206	<.001	.560	1.786

a. Dependent Variable: SO

(3) Model 1-3:

$$\text{Stockout} = f\left(Amazon\ Order\ Quantity, Inventory, \frac{Other\ Order\ Quantity}{Amazon\ Order\ Quantity}, \frac{Amazon\ Order\ Quantity}{2019\ Average\ Order\ Quantity}, Brand\ SO\%, DC(Binary), Top\ 5\ Stockout\ Brands(Binary)\right)$$

Since the stockout rate varies by brand, we added five brands with the highest stockout rate as binary variables: Pepcid, Plax, Tylenol, Sudafed, and Reach. Coefficients of all five brand variables are all significant and the prediction accuracy of stockout increased by 2.9% with 75.9% overall model prediction accuracy.

(4) Model 1-4:

$$\text{Stockout} = f\left(Amazon\ Order\ Quantity, Inventory, \frac{Other\ Order\ Quantity}{Amazon\ Order\ Quantity}, \frac{Amazon\ Order\ Quantity}{2019\ Average\ Order\ Quantity}, Brand\ SO\%, DC(Binary), Top\ 5\ Order\ Brands(Binary)\right)$$

After aggregating order quantity by brand, we found several dominating brands. However, the top brands with the highest demand were not the same as brands with the highest stockout rate. In this model, we changed the brand variable to top five brands with the highest demand: Neutrogena, Aveeno, Tylenol, OGX, and Listerine. Coefficients of all five brand variables are all significant, and showed that Neutrogena, Aveeno, and OGX have negative relationship with stockout. Though the Chi-square increases suggesting that this model is explaining more of the variance in the outcome, the prediction accuracy of stockout decrease by 2.1%.

(5) Model 1-5:

$$\text{Stockout} = f\left(Amazon\ Order\ Quantity, Inventory, \frac{Other\ Order\ Quantity}{Amazon\ Order\ Quantity}, \frac{Amazon\ Order\ Quantity}{2019\ Average\ Order\ Quantity}, Brand\ SO\%, DC(Binary), Top\ 5\ Stockout\ Brands(Binary), Top\ 5\ Order\ Brands(Binary)\right)$$

In this model, we included the top five brands with the highest stockout rate and the top five brands with the highest demand. The prediction accuracy of stockout increase by 2.4% compared to simply included top stockout brands.

(6) Model 1-6:

$$\text{Stockout} = f\left(Amazon\ Order\ Quantity, Inventory, \frac{Other\ Order\ Quantity}{Amazon\ Order\ Quantity}, \frac{Amazon\ Order\ Quantity}{2019\ Average\ Order\ Quantity}, Brand\ SO\%, DC(Binary), Brands\ with\ SO\% > 0.3\ (Binary)\right)$$

Apart from Pepcid and Plax, the average stockout rate of other brands did not vary significantly. Therefore, we included all the brands with average stockout rate over 0.3 in this model: Pepcid, Plax, Sudafed, Tylenol, Reach, Imodium, Listerine, Desitin, Rhinocort, Clean & Clear, Neosporin, and Bandaid. Though the prediction accuracy is the same as Model 1-4, the Chi-square is the highest among these six models.

Table 10

Prediction Accuracy of Model 1

Model 1	True Positive (Stockout)	True Negative (Fulfilled)	Accuracy
1-1 All data	5.0	99.1	75.5
1-2 All data + DC	5.1	99.1	75.5
1-3 All data + DC+ Top 5 Brand (SO%)	7.9	98.7	75.9
1-4 All data + DC+ Top 5 Brand (Demand)	5.8	99.0	75.5
1-5 All data + DC+ Top 5 Brand (SO%)+ Top 5 Brand (Demand)	8.2	98.6	75.9
1-6 All data + DC+ All Brand (SO% > 0.3)	8.2	98.6	75.9

We picked Model 1-6 (Table 11) as the best one to predict stockout using all the data, several insights are drawn based on this model:

- Coefficients of all variables are significant at the 95 percent confidence level and the coefficient of each variable support our hypothesis that the order quantity, the competing order quantity, order size, and brand SO% have positive relationship with stockout while inventory has negative relationship with stockout. Moreover, from which DC the order was fulfilled and brand are related to stockout.
- The odds ratio of the competing orders and order size significantly increases the probability of stockout. However, the odds ratio (Exp (B)) of inventory is .000, suggesting that the inventory in the model did not impact the probability of stockout.
- The odds ratio of the brand also tells us that products under Pepcid, Plax, Tylenol, Listerine, Desitin are more than 2 times more likely to have stockout.

Table 11

Variables in the Equation Table of Model 1-6

Variables in the Equation

		B	S.E.	Wald	df	Sig.	Exp(B)	95% C.I.for EXP(B) Lower	Upper
Step 1[a]	Amazon_Order_QTY_n	73.816	2.959	622.526	1	<.001	1.143E+32	3.465E+29	3.770E+34
	Inventory_n	-12.809	.938	186.334	1	<.001	.000	.000	.000
	OtherAmazon_n	2.621	.489	28.743	1	<.001	13.748	5.274	35.838
	Order_Qty2019Avg_n	16.203	2.511	41.641	1	<.001	10888538.59	79367.056	1493822229
	SO_Brand	1.416	.118	143.507	1	<.001	4.120	3.268	5.194
	DC1	.209	.026	65.506	1	<.001	1.232	1.171	1.296
	DC2	.323	.026	155.284	1	<.001	1.381	1.313	1.453
	PEPCID	2.171	.090	582.681	1	<.001	8.771	7.353	10.462
	PLAX	3.002	.354	71.901	1	<.001	20.121	10.054	40.270
	SUDAFED	.551	.119	21.358	1	<.001	1.735	1.374	2.192
	TYLENOL	.975	.047	433.519	1	<.001	2.651	2.419	2.906
	REACH	.656	.084	60.304	1	<.001	1.926	1.633	2.273
	IMODIUM	.504	.099	25.821	1	<.001	1.656	1.363	2.011
	LISTERINE	.923	.040	531.008	1	<.001	2.517	2.327	2.723
	DESITIN	1.059	.093	131.026	1	<.001	2.885	2.406	3.458
	RHINOCORT	.721	.208	12.045	1	<.001	2.057	1.369	3.092
	CLEANCLEAR	.653	.052	158.374	1	<.001	1.922	1.736	2.128
	NEOSPORIN	.437	.080	29.908	1	<.001	1.548	1.323	1.810
	BANDAID	.731	.046	255.234	1	<.001	2.078	1.900	2.273
	Constant	-1.733	.029	3646.017	1	.000	.177		

a. Variable(s) entered on step 1: Amazon_Order_QTY_n, Inventory_n, OtherAmazon_n, Order_Qty2019Avg_n, SO_Brand, DC1, DC2, PEPCID, PLAX, SUDAFED, TYLENOL, REACH, IMODIUM, LISTERINE, DESITIN, RHINOCORT, CLEANCLEAR, NEOSPORIN, BANDAID.

Multicollinearity was checked again to ensure no strong correlation among variables when we added brands to the model. VIF values of Model 1-6 are below 5 (Table 12), suggesting that coefficient estimates and the p-values are not impacted by multicollinearity.

Table 12

VIF of Model 1-6

Coefficients[a]

Model		Unstandardized Coefficients		Standardized Coefficients			Collinearity Statistics	
		B	Std. Error	Beta	t	Sig.	Tolerance	VIF
1	(Constant)	.184	.003		52.831	.000		
	Amazon_Order_QTY_n	6.158	.296	.100	20.772	<.001	.598	1.672
	Inventory_n	-.710	.061	-.043	-11.547	<.001	.986	1.014
	OtherAmazon_n	.184	.093	.008	1.981	.048	.955	1.047
	Order_Qty2019Avg_n	.002	.288	.000	.006	.995	.607	1.648
	SO_Brand	.220	.020	.046	10.847	<.001	.757	1.321
	PEPCID	.501	.017	.111	29.591	<.001	.978	1.023
	PLAX	.620	.059	.039	10.434	<.001	.997	1.003
	SUDAFED	.112	.023	.018	4.909	<.001	.984	1.016
	TYLENOL	.216	.009	.089	23.131	<.001	.944	1.059
	REACH	.160	.016	.037	9.846	<.001	.960	1.041
	IMODIUM	.104	.019	.021	5.516	<.001	.978	1.023
	LISTERINE	.182	.008	.092	23.486	<.001	.908	1.101
	DESITIN	.253	.019	.051	13.442	<.001	.977	1.024
	RHINOCORT	.144	.041	.013	3.492	<.001	.995	1.005
	CLEANCLEAR	.122	.010	.047	12.205	<.001	.934	1.071
	NEOSPORIN	.100	.015	.025	6.653	<.001	.965	1.036
	BANDAID	.141	.009	.062	16.028	<.001	.913	1.095

a. Dependent Variable: SO

4.2.2 Model 2 - Only orders fulfilled by DC13

Five models were developed with the same iteration as Model 1, but using only the order data that were fulfilled by DC13, that is 28,454 data points. Comparison of prediction accuracy of these five models is summarized in Table 13 and the detailed output results are given in Appendix.

(1) Model 2-1:

Stockout

$$= f\left(Amazon\ Order\ Quantity, Inventory, \frac{Other\ Order\ Quantity}{Amazon\ Order\ Quantity}, \frac{Amazon\ Order\ Quantity}{2019\ Average\ Order\ Quantity}, Brand\ SO\%\right)$$

Comparing to Model 1-1 with the same independent variables, the overall accuracy and prediction accuracy of True Negative (Fulfilled) is better in Model 2-1 with 75.7% and 99.5% respectively. However, Model 1-1 has higher prediction accuracy of True Positive (Stockout).

(2) Model 2-2:

$$\text{Stockout} = f\left(Amazon\ Order\ Quantity, Inventory, \frac{Other\ Order\ Quantity}{Amazon\ Order\ Quantity}, \frac{Amazon\ Order\ Quantity}{2019\ Average\ Order\ Quantity}, Brand\ SO\%, Top\ 5\ Stockout\ Brands(Binary)\right)$$

Top five brands with the highest average stockout rate are included in this model, including Plax, Pepcid, Sudafed, Rhinocort, and Tylenol. By adding brands as binary variables, the prediction accuracy of stockout increases by 3% and the overall model accuracy increases by 0.4%. However, the coefficient of Sudafed is only 0.169 and the p-value tells us that it is not significant.

(3) Model 2-3:

$$\text{Stockout} = f\left(Amazon\ Order\ Quantity, Inventory, \frac{Other\ Order\ Quantity}{Amazon\ Order\ Quantity}, \frac{Amazon\ Order\ Quantity}{2019\ Average\ Order\ Quantity}, Brand\ SO\%, Top\ 5\ Order\ Brands(Binary)\right)$$

Neutrogena, Aveeno, Tylenol, Listerine, and Reach are the brands with the highest demand fulfilled by DC13 and are included in this model. Similar to Model 1-3, adding brands with the highest demand increases the prediction accuracy of Ture Negative (Fulfilled) and decreases the prediction accuracy of True Positive (Stockout), mainly due to less stockout of Neutrogena and Aveeno products as suggested by the negative coefficient of these two variables.

(4) Model 2-4:

$$\text{Stockout} = f\left(Amazon\ Order\ Quantity, Inventory, \frac{Other\ Order\ Quantity}{Amazon\ Order\ Quantity}, \frac{Amazon\ Order\ Quantity}{2019\ Average\ Order\ Quantity}, Brand\ SO\%, Top\ 5\ Stockout\ Brands(Binary), Top\ 5\ Order\ Brands(Binary)\right)$$

Nine brands are included in this model to represent brands with the highest stockout rate and the highest order quantity: Pepcid, Plax, Rhinocort, Sudafed, Neutrogena, Aveeno, Tylenol, Listerine and Reach. The prediction accuracy of True Positive (Stockout) slightly increases but the prediction accuracy of True Negative (Fulfilled) decreases comparing to model with only top stock out brands (Model 2-2), and the coefficient of Sudafed is also not significant in this model.

(5) Model 2-5:

$$\text{Stockout} = f\left(Amazon\ Order\ Quantity, Inventory, \frac{Other\ Order\ Quantity}{Amazon\ Order\ Quantity}, \frac{Amazon\ Order\ Quantity}{2019\ Average\ Order\ Quantity}, Brand\ SO\%, Brands\ with\ SO\% > 0.3\ (Binary)\right)$$

Again, brands with average stockout rate above 0.3 are included in the model, that is Pepcid, Plax, Rhinocort, Sudafed, Tylenol, Reach, Imodium, Listerine, Desitin, Clean & Clear, and Bandaid. This combination of independent variables (Table 14) is again the best for predicting stockout among models using only DC13 data.

Table 13

Prediction Accuracy of Model 2

Model 2	True Positive (Stockout)	True Negative (Fulfilled)	Accuracy
2-1 DC1 data	3.1	99.5	75.7
2-2 DC1 data + Top 5 Brand (SO%)	6.1	98.9	76.1
2-3 DC1 data + Top 5 Brand (Demand)	3.5	99.2	75.7
2-4 DC1 data + Top 5 Brand (SO%)+ Top 5 Brand (Demand)	6.2	98.8	76.0
2-5 DC1 data + All Brand (SO% > 0.3)	8.2	98.6	75.9

Table 14

Variables in the Equation Table of Model 2-5

Variables in the Equation

		B	S.E.	Wald	df	Sig.	Exp(B)	95% C.I.for EXP(B) Lower	Upper
Step 1[a]	Amazon_Order_QTY_n	56.758	4.632	150.128	1	<.001	4.462E+24	5.088E+20	3.914E+28
	Inventory_n	-3.261	.696	21.920	1	<.001	.038	.010	.150
	OtherAmazon_n	3.115	.883	12.451	1	<.001	22.529	3.994	127.090
	Order_Qty2019Avg_n	14.623	4.924	8.819	1	.003	2243174.260	144.360	3.486E+10
	SO_Brand	.405	.182	4.947	1	.026	1.499	1.049	2.142
	PLAX	2.950	.637	21.436	1	<.001	19.102	5.480	66.590
	PEPCID	2.012	.131	237.434	1	<.001	7.481	5.792	9.664
	RHINOCORT	1.302	.268	23.575	1	<.001	3.676	2.174	6.218
	SUDAFED	.448	.186	5.834	1	.016	1.566	1.088	2.252
	TYLENOL	.943	.073	167.730	1	<.001	2.568	2.226	2.962
	REACH	.884	.125	50.002	1	<.001	2.421	1.895	3.094
	IMODIUM	.546	.157	12.076	1	<.001	1.726	1.269	2.348
	DESITIN	1.028	.141	53.331	1	<.001	2.797	2.122	3.685
	LISTERINE	.763	.066	134.916	1	<.001	2.146	1.886	2.440
	CLEANCLEAR	.695	.081	74.079	1	<.001	2.004	1.711	2.348
	BANDAID	.662	.073	82.384	1	<.001	1.938	1.680	2.236
	Constant	-1.423	.033	1861.641	1	.000	.241		

a. Variable(s) entered on step 1: Amazon_Order_QTY_n, Inventory_n, OtherAmazon_n, Order_Qty2019Avg_n, SO_Brand, PLAX, PEPCID, RHINOCORT, SUDAFED, TYLENOL, REACH, IMODIUM, DESITIN, LISTERINE, CLEANCLEAR, BANDAID.

By comparing with Model 1-6, several insights are drawn based on this model:

- Coefficients of all variables are significant at the 95 percent confidence level and the coefficient of each variable support our hypothesis. However, the coefficients and the odds ratio are generally smaller than Model 1-6.

- The odds ratio of the inventory is .038, slightly higher than Model 1-6, meaning more impact on the probability of stockout.
- The odds ratio of the brand also tells us that products under Pepcid, Plax, Rhinocort, Tylenol, Reach, Desitin, Listerine, and Clean & Clear are more than 2 times more likely to have stockout in DC13.

4.2.3 Model 3 - Only orders fulfilled by DC16

Five models were developed with the same iteration as Model 2, but using only the order data that were fulfilled by DC16, that is 26,300 data points. Comparison of prediction accuracy of these five models is summarized in Table 15 and the detailed output results are given in Appendix.

(1) Model 3-1:

$$\text{Stockout} = f\left(Amazon\ Order\ Quantity, Inventory, \frac{Other\ Order\ Quantity}{Amazon\ Order\ Quantity}, \frac{Amazon\ Order\ Quantity}{2019\ Average\ Order\ Quantity}, Brand\ SO\%\right)$$

Without adding any brand binary variable, the prediction accuracy of True Positive (Stockout) is already higher than the best model in Model 1 and Model 2. However, the prediction accuracy of True Negative (Fulfilled) slightly decrease to 98.4%. Moreover, the coefficient of SO% by brand is negative and is not significant, which does not support our hypothesis.

(2) Model 3-2:

$$\text{Stockout} = f\left(Amazon\ Order\ Quantity, Inventory, \frac{Other\ Order\ Quantity}{Amazon\ Order\ Quantity}, \frac{Amazon\ Order\ Quantity}{2019\ Average\ Order\ Quantity}, Brand\ SO\%, Top\ 5\ Stockout\ Brands(Binary)\right)$$

The top five brands with the highest average stockout rate were included in this model: Pepcid, Plax, Sudafed, Tylenol, and Reach. After adding these brands as binary variables, the prediction accuracy of stockout increases by 2.7%. However, Sudafed is this model appears to be insignificant in predicting stockout.

(3) Model 3-3:

$$\text{Stockout} = f\left(Amazon\ Order\ Quantity, Inventory, \frac{Other\ Order\ Quantity}{Amazon\ Order\ Quantity}, \frac{Amazon\ Order\ Quantity}{2019\ Average\ Order\ Quantity}, Brand\ SO\%, Top\ 5\ Order\ Brands(Binary)\right)$$

Neutrogena, Aveeno, Tylenol, Listerine, and Reach are the brands with the highest demand in DC16 and improves the prediction accuracy of stockout compared to Model 3-1 without any brand variable.

(4) Model 3-4:

$$\text{Stockout} = f\left(Amazon\ Order\ Quantity, Inventory, \frac{Other\ Order\ Quantity}{Amazon\ Order\ Quantity}, \frac{Amazon\ Order\ Quantity}{2019\ Average\ Order\ Quantity}, Brand\ SO\%, Top\ 5\ Stockout\ Brands(Binary), Top\ 5\ Order\ Brands(Binary)\right)$$

Eight brand variables in total are added in this model: Neutrogena, Aveeno, Tylenol, Listerine, Reach, Pepcid, Plax, and Sudafed. The combination of brands with highest demand and highest stockout rate performs better in predicting stockout compared to Model 3-2, which includes only top stockout brands. However, the coefficients of the competing order and Sudafed are not significant.

(5) Model 3-5:

$$\text{Stockout} = f\left(Amazon\ Order\ Quantity, Inventory, \frac{Other\ Order\ Quantity}{Amazon\ Order\ Quantity}, \frac{Amazon\ Order\ Quantity}{2019\ Average\ Order\ Quantity}, Brand\ SO\%, Brands\ with\ SO\% > 0.3\ (Binary)\right)$$

Fourteen brands with average stockout rate above 0.3 in DC16 are added to this model: Pepcid, Plax, Sudafed, Tylenol, Reach, Desitin, Imodium, Listerine, Clean & Clear, ZYRTEC, Rhinocort, Bandaid, Johnson's, and OGX. The prediction accuracy of stockout reach the highest given the same overall accuracy as previous models. However, coefficients of three variables – the competing order, Rhinocort, and Johnson's.

Table 15

Prediction Accuracy of Model 3

Model 3	True Positive (Stockout)	True Negative (Fulfilled)	Accuracy
3-1 DC2 data	9.0	98.4	73.0
3-2 DC2 data + Top 5 Brand (SO%)	11.7	98.1	73.6
3-3 DC2 data + Top 5 Brand (Demand)	9.6	98.3	73.1
3-4 DC2 data + Top 5 Brand (SO%)+ Top 5 Brand (Demand)	11.9	98.0	73.6
3-5 DC2 data + All Brand (SO% > 0.3)	12.3	97.9	73.6

Model 3-5 (Table 16) is the best among Model 3 in predicting stockout though not all coefficients are significant. In comparison with Model 1-6 and Model 2-5, several insights are drawn based on this model:

- Coefficients of order quantity, inventory, and SO% by brand are significantly bigger than Model 1-6

and Model 2-5. However, the odds ratio (Exp (B)) of inventory is .000, so we concluded that order quantity and SO% by brand play an important role in increasing the probability of stockout.

- The coefficient of the competing order quantity is smaller than Model 1-6 and Model 2-5 and it is insignificant. However, the positive coefficient with 3.5 odds ratio still support our hypothesis that the larger the order quantity of the same SKU at the same week from other customers, the higher chance of stockout.
- The odds ratio of the brand tells us that products under Pepcid, Plax, Tylenol, Desitin, Listerine, and Bandaid are more than 2 times more likely to have stockout in DC16.

Table 16

Variables in the Equation Table of Model 3-5

Variables in the Equation

		B	S.E.	Wald	df	Sig.	Exp(B)	95% C.I.for EXP(B) Lower	Upper
Step 1[a]	Amazon_Order_QTY_n	82.407	4.282	370.310	1	<.001	6.148E+35	1.392E+32	2.716E+39
	Inventory_n	-40.059	2.565	243.982	1	<.001	.000	.000	.000
	OtherAmazon_n	1.263	1.004	1.581	1	.209	3.535	.494	25.301
	Order_Qty2019Avg_n	22.510	3.738	36.265	1	<.001	5971987108	3929571.835	9.076E+12
	SO_Brand	1.898	.203	87.749	1	<.001	6.670	4.484	9.920
	PEPCID	2.269	.149	231.520	1	<.001	9.674	7.222	12.958
	PLAX	3.303	.761	18.853	1	<.001	27.191	6.123	120.761
	SUDAFED	.440	.176	6.248	1	.012	1.553	1.100	2.193
	TYLENOL	.842	.073	131.951	1	<.001	2.320	2.010	2.678
	REACH	.581	.130	20.035	1	<.001	1.788	1.386	2.306
	DESITIN	1.110	.139	63.658	1	<.001	3.033	2.310	3.984
	IMODIUM	.312	.143	4.736	1	.030	1.366	1.031	1.809
	LISTERINE	.859	.063	184.864	1	<.001	2.361	2.086	2.673
	CLEANCLEAR	.661	.078	71.934	1	<.001	1.936	1.662	2.255
	ZYRTEC	.618	.185	11.185	1	<.001	1.856	1.292	2.666
	RHINOCORT	-.004	.410	.000	1	.993	.996	.446	2.224
	BANDAID	.713	.070	104.763	1	<.001	2.040	1.780	2.339
	JOHNSONS	-.055	.066	.698	1	.403	.946	.832	1.077
	OGX	-.102	.045	5.081	1	.024	.903	.826	.987
	Constant	-1.338	.034	1514.317	1	.000	.262		

a. Variable(s) entered on step 1: Amazon_Order_QTY_n, Inventory_n, OtherAmazon_n, Order_Qty2019Avg_n, SO_Brand, PEPCID, PLAX, SUDAFED, TYLENOL, REACH, DESITIN, IMODIUM, LISTERINE, CLEANCLEAR, ZYRTEC, RHINOCORT, BANDAID, JOHNSONS, OGX.

4.2.4 Model 4 - Only orders fulfilled by DC19

Five models were developed with the same iteration as Model 2 and Model 3, but using only the order data that were fulfilled by DC19, that is 14,108 data points. Comparison of prediction accuracy of these five models is summarized in Table 17 and the detailed output results are given in Appendix.

(1) Model 4-1:

$$\text{Stockout} = f\left(Amazon\ Order\ Quantity, Inventory, \frac{Other\ Order\ Quantity}{Amazon\ Order\ Quantity}, \frac{Amazon\ Order\ Quantity}{2019\ Average\ Order\ Quantity}, Brand\ SO\%\right)$$

With only data from DC19, this model has the worst performance in predicting stockout compared to all the models discussed above. However, the overall accuracy is 79.9%, mainly impacted by the 99.6% prediction accuracy of True Negative (Fulfilled). Besides, the coefficient of order size compared to 2019 average is not significant.

(2) Model 4-2:

$$\text{Stockout} = f\left(Amazon\ Order\ Quantity, Inventory, \frac{Other\ Order\ Quantity}{Amazon\ Order\ Quantity}, \frac{Amazon\ Order\ Quantity}{2019\ Average\ Order\ Quantity}, Brand\ SO\%, Top\ 5\ Stockout\ Brands(Binary)\right)$$

The top five brands with the highest average stockout rate included in this model are Pepcid, Plax, Listerine, Tylenol, and Neosporin. After adding these brands as binary variables, the prediction accuracy of stockout increases by 2.8% and the overall accuracy reaches 80.1%. However, order size comparing to 2019 average remains insignificant in predicting stockout.

(3) Model 4-3:

$$\text{Stockout} = f\left(Amazon\ Order\ Quantity, Inventory, \frac{Other\ Order\ Quantity}{Amazon\ Order\ Quantity}, \frac{Amazon\ Order\ Quantity}{2019\ Average\ Order\ Quantity}, Brand\ SO\%, Top\ 5\ Order\ Brands(Binary)\right)$$

Comparing DC13 and DC16, brands with highest demand are different, including OGX, Neutrogena, Listerine, Aveeno, and Imodium. Though the prediction accuracy of stockout is lower than the model with top stockout brands, the performance of this model is still better than Model 4-1 which does not include any brand variable. Moreover, the odds ratio of the competing order quantity and SO% by brand increases significantly to 262.741 and 116.385 respectively, suggesting that products with higher order quantity from other customers and higher brand SO% are 262.741 and 116.385 times more likely to stockout.

(4) Model 4-4:

$$\text{Stockout} = f\left(Amazon\ Order\ Quantity, Inventory, \frac{Other\ Order\ Quantity}{Amazon\ Order\ Quantity}, \frac{Amazon\ Order\ Quantity}{2019\ Average\ Order\ Quantity}, Brand\ SO\%, Top\ 5\ Stockout\ Brands(Binary), Top\ 5\ Order\ Brands(Binary)\right)$$

Nine brand variables in total are added in this model, including Listerine, Imodium, OGX, Neutrogena, Aveeno, Pepcid, Plax, Tylenol, and Neosporin. This prediction accuracy of stockout is 0.8% higher than the model includes only top stockout brands.

(5) Model 4-5:

Stockout

$$= f\left(Amazon\ Order\ Quantity, Inventory, \frac{Other\ Order\ Quantity}{Amazon\ Order\ Quantity}, \frac{Amazon\ Order\ Quantity}{2019\ Average\ Order\ Quantity},\right.$$

$$\left.Brand\ SO\%, Brands\ with\ SO\% > 0.3\ (Binary)\right)$$

Six brands with average stockout rate above 0.3 in DC19 are added to this model: Pepcid, Plax, Listerine, Tylenol, Neosporin, Imodium. This model only includes one more brand comparing to Model 4-3, so the performance is also similar.

Table 17

Prediction Accuracy of Model 4

Model 4	True Positive (Stockout)	True Negative (Fulfilled)	Accuracy
4-1 DC3 data	1.7	99.6	79.9
4-2 DC3 data + Top 5 Brand (SO%)	4.5	99.2	80.1
4-3 DC3 data + Top 5 Brand (Demand)	3.5	99.2	79.9
4-4 DC3 data + Top 5 Brand (SO%)+ Top 5 Brand (Demand)	5.3	99.0	80.1
4-5 DC3 data + All Brand (SO% > 0.3)	4.5	99.2	80.1

We picked Model 4-4 (Table 18) as the best one to predict stockout with only DC19 data due to the highest prediction accuracy of True Positive (Stockout). several insights are drawn based on this model:

- Coefficients of the variables are different compared to Model 1, 2, and 3. Coefficients of order quantity from other customers and SO% by brand are significantly bigger than Model 1-6, Model 2-5, and Model 3-5. The high odds ratios (Exp (B)) suggest that the probability of stockout in this model is dominated by these two variables.
- Another key difference of this model is that the coefficient of order size compared to the 2019 average is consistently not significant, suggesting that the increase in order size didn't impact the

stockout rate in DC19.

- Though the coefficient of inventory is the biggest negative among all the models, the odds ratio (Exp (B)) of inventory is .000, which is not conclusive in explaining the relationship between inventory and stockout rate.
- The odds ratio of the brand tells us that products under Pepcid, Plax, Tylenol, Neosporin, and Listerine are more than 2 times more likely to have stockout in DC19. In contrast, products under OGX, Neutrogena, and Aveeno have lower probability of stockout.

Table 18

Variables in the Equation Table of Model 4-4

Variables in the Equation

								95% C.I.for EXP(B)	
		B	S.E.	Wald	df	Sig.	Exp(B)	Lower	Upper
Step 1[a]	Amazon_Order_QTY_n	70.124	9.132	58.968	1	<.001	2.848E+30	4.803E+22	1.689E+38
	Inventory_n	-46.977	3.193	216.418	1	<.001	.000	.000	.000
	OtherAmazon_n	4.609	.868	28.197	1	<.001	100.422	18.321	550.426
	Order_Qty2019Avg_n	2.170	4.232	.263	1	.608	8.756	.002	35068.165
	SO_Brand	4.273	.506	71.251	1	<.001	71.709	26.590	193.387
	LISTERINE	1.103	.090	151.144	1	<.001	3.013	2.527	3.592
	IMODIUM	.681	.286	5.650	1	.017	1.975	1.127	3.462
	OGX	-.286	.107	7.170	1	.007	.751	.610	.926
	NEUTROGENA	-.616	.121	25.745	1	<.001	.540	.426	.685
	AVEENO	-.502	.088	32.264	1	<.001	.606	.509	.720
	PEPCID	1.926	.234	67.930	1	<.001	6.859	4.339	10.843
	PLAX	2.781	.524	28.129	1	<.001	16.133	5.773	45.082
	TYLENOL	1.037	.117	77.879	1	<.001	2.820	2.240	3.550
	NEOSPORIN	.814	.190	18.439	1	<.001	2.257	1.556	3.272
	Constant	-1.620	.045	1273.681	1	<.001	.198		

a. Variable(s) entered on step 1: Amazon_Order_QTY_n, Inventory_n, OtherAmazon_n, Order_Qty2019Avg_n, SO_Brand, LISTERINE, IMODIUM, OGX, NEUTROGENA, AVEENO, PEPCID, PLAX, TYLENOL, NEOSPORIN.

5 DISCUSSION

5.1 Main Insights

1) The first insight of the capstone project is the impact that some of the analyzed variables have over the missing order quantities and its magnitude. The most significant variables that affect the Company's missing order quantities for the online retailer are:

   - Order Quantity: This is the amount of product order in a specific order.
   - Order Size in 2020 in comparison to 2019: The magnitude of the order size in comparison to the 2019 order size
   - Inventory availability: Inventory on hand at the moment of the order

   From the results of the multiple linear regression, the magnitude of the impact that Order Quantity and Order Size in 2020 vs. 2019 have on the missing order quantities is interestingly higher than the inventory on hand. This means that the demand has a higher impact on the missing orders than the supply. This is a powerful insight for the Company and it will be important to monitor the order sizes closely as it would be more relevant than the inventory on hand.

2) The second relevant insight is that the Product Categories play an important role in the missing order quantities, and therefore the brand portfolio brands in each category. Not all product categories impact the missing order quantities in the same way and it is 'Self Care' the category that has the highest impact. Under Self Care, brands like Tylenol, Pepcid, Motrin were highly demanded during 2020 and the missing order quantities increased substantially. Other product categories such as Skin Health and Essential Health Products had a lower impact and were even statistically insignificant under the multiple linear regression analysis

3) The third insight is that for the problem we analyzed and the dataset we were able to work with, the multiple linear regression worked better than the logistic regression. Our approaches for both analyses were different:

   - Multiple Linear Regression: The prediction (independent variable) was based on the missing order quantity
   - Logistic Regression: The prediction was based on the probability of stocking out. The classification was that "1" corresponded to a stockout event and "0" to order fulfillment.

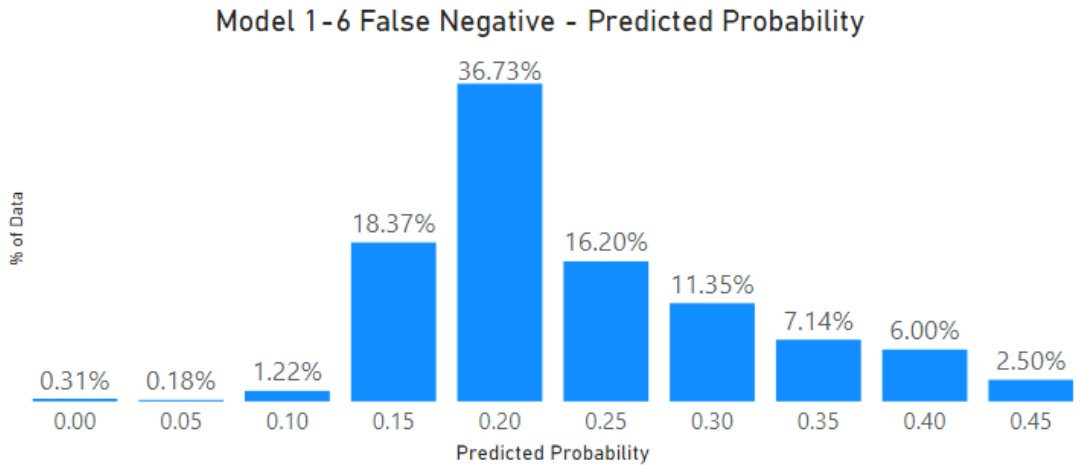
   The challenge we encountered with the logistic regression was the determination of the cut-off point and the classification of the orders as stockouts/fulfilled. The accuracy of the logistic regression we developed was not high enough to make reliable predictions and therefore we chose to use the multiple linear regression model to make the missing order quantity predictions.

5.2 Limitations

One of the limitations of our logistic regression model was that the prediction accuracy is based using the general cutoff point - 0.5. The overall model accuracy might be bias due to the imbalanced dataset, where an average of 25% of the datapoints are in class 1 while class 0 contains the remaining datapoints. Take Model 1-6 for example (Figure 13), the predicted probability of 43.2% of False Negative (classified as order fulfilled but actually stockout) are above 0.25. Therefore, the overall accuracy is expected to increase significantly when the cutoff value is set to 0.25. Hence, we suggest to include comparison of outcomes for different cutoff values and select an optimal cutoff with the best overall accuracy.

Figure 13

Predicted Probability of Model 1-6 False Negative



Another limitation of our logistic regression models was that the assumption of linear relationship between logit of the outcome and each predictor variables is violated. Box-Tidwell Transformation Test was conducted with all continuous variables included, the logistic model interaction terms are transformed by $(x(\ln(x))$ where x is the independent variable. The interaction terms are abbreviated as below:

- Amazon's Order Quantity ➔ Qty_t
- Inventory Position ➔ Inv_t
- Magnitude of the Order Quantity from Other Customers in comparison to Amazon's Order Quantity ➔ Other_t
- Magnitude of the Order Quantity in comparison to 2019 Average Order Quantity ➔ OrderSize_t
- SO% of Brand ➔ SOBrand_t

The test result (Table 19) shows that these logistic model interaction terms are significant, suggesting nonlinearity in the logit. Therefore, we suggest transforming variables, such as polynomial regression, in the future study to increase the robustness of the model.

Table 19

Box-Tidwell Transformation Test of Logistic Regression Model

Variables in the Equation

		B	S.E.	Wald	df	Sig.	Exp(B)
Step 1[a]	Amazon_Order_QTY_n	318.938	15.295	434.817	1	<.001	3.257E+138
	Inventory_n	-193.680	5.917	1071.436	1	<.001	.000
	OtherAmazon_n	33.649	3.829	77.213	1	<.001	4.107E+14
	Order_Qty2019Avg_n	92.488	9.910	87.100	1	<.001	1.469E+40
	SO_Brand	-1.212	.344	12.378	1	<.001	.298
	Qty_t	.000	.000	287.225	1	<.001	1.000
	Inv_t	.000	.000	1022.954	1	<.001	1.000
	Other_t	.000	.000	51.157	1	<.001	1.000
	OrderSize_t	-.001	.000	66.746	1	<.001	.999
	SOBrand_t	-.711	.275	6.676	1	.010	.491
	Constant	-1.093	.027	1620.586	1	.000	.335

a. Variable(s) entered on step 1: Amazon_Order_QTY_n, Inventory_n, OtherAmazon_n, Order_Qty2019Avg_n, SO_Brand, Qty_t, Inv_t, Other_t, OrderSize_t, SOBrand_t.

5.3 Future Research

The main machine learning techniques used for the analysis of this capstone project were Multiple Linear Regression and Logistic Regression. There are however other machine learning techniques that could be applied to analyze the problem from a different perspective. One of the alternative methods that was not used in this capstone is the Random Forest model, which is works very well as a general-purpose classification and regression model. Random Forest is a supervised algorithm on which it operates diving the dataset into a collection of predictors of randomized regression trees (Biau &

Scornet, 2016). We consider that Random Forest could be a valid approach to analyze similar cases to this project as it is versatile and widely applicable.

6 CONCLUSION

The main focus of this capstone was to find out the reasons of stockout in e-commerce business. Hypotheses were developed based on the observations of order data from e-commerce customers in 2019 and 2020. Both multiple linear regression and logistic regression were applied to test the hypotheses.

In multiple linear regression, the dependent variable is missing order quantity and our model got 63% adjusted accuracy in testing datasets. The coefficients of our key variables suggested that order quantity, order size, product category that contains brands with higher stockout rate have positive impact on missing order quantity. On the other hand, inventory has negative impact on the quantity of stockout. Moreover, DCs were also tested significant in predicting stockout quantity.

In logistic regression, the dependent variable is the occurrence of stockout event. Coefficients of all key variables are also tested significant at the 95 percent confidence level. Four major models were developed with logistic regression, the result showed similar relationship found in multiple linear regression model, including order quantity, order size, and inventory. In addition, by adding brands with high stockout rate and DC as binary variables increase the prediction accuracy of all models.

Models trained with multiple linear regression and logistic regression both supported our hypotheses and suggested that key variables related to demand are more impactful in predicting stockout than variables related to supply. Brands and DCs with higher stockout rate also proved to impact the probability of stockout.

Lastly, we would like to highlight a few potential future research directions. A key area could be trying different cutoff point or other machine learning techniques to deal with imbalanced classification data. In addition, the prediction accuracy could be further improved by exploring other classification models like random forest.

References

Al Shalabi, L., & Shaaban, Z. (2006). Normalization as a Preprocessing Engine for Data Mining and the Approach of Preference Matrix. 2006 International Conference on Dependability of Computer Systems, 207–214. https://doi.org/10.1109/DEPCOS-RELCOMEX.2006.38

Agatz, N. A., Fleishcmann, M., & van Nunen, J. A. (2007). E-fulfillment and multi-channel distribution - A review. *European Journal of Operational Research*, 339-356.

Biau, G., & Scornet, E. (2016). A random forest guided tour. *TEST*, *25*(2), 197–227. https://doi.org/10.1007/s11749-016-0481-7

Corsten, D., & Gruen, T. (2005). *On Shelf Availability: An Examination of the Extent, the Causes, and the Efforts to Address Retail Out-of-Stocks* (pp. 131–149). https://doi.org/10.1007/3-540-27059-0_9

Digital Commerce 360. (2021, Jan 29) US ecommerce grows 44.0% in 2020. https://www.digitalcommerce360.com/article/us-ecommerce-sales/

Fernandes de Mello, R., & Antonelli Ponti, M. (2018). *Machine Learning: A Practical Approach on the Statistical Learning Theory.* Cham: Springer.

Forger, G. (2021, January/February). The Supply Chain's Pivot to e-Commerce. Supply Chain Management Review, pp. 12-22.

Ivanov, D. (2021). Supply Chain Viability and the COVID-19 pandemic: a conceptual and formal generalisation of four major adaptation strategies. *International Journal of Production Research*.

Jain, A., Nandakumar, K., & Ross, A. (2005). Score normalization in multimodal biometric systems. Pattern Recognition, 38(12), 2270–2285. https://doi.org/10.1016/j.patcog.2005.01.012

Milićević, N., Grubor, A., Đokić, N., & Avlijaš, G. (2018). Retail Out-of-stocks in the Context of Centralized and Direct Delivery. *Promet - Traffic&Transportation*, *30*(1), 105–114. https://doi.org/10.7307/ptt.v30i1.2466

Mitchell, T. M., Carbonell, J. G., & Michalski, R. S. (1986). *Machine Learning: A Guide to Current Research.* Norwell: Kluwer Academic Publishers.

Nigam, A. (2016). *Product promotion effectiveness: Root causes of stockouts* [Thesis, Massachusetts Institute of Technology]. https://dspace.mit.edu/handle/1721.1/107513

Oeser, G. (2015). *Risk-Pooling Essentials*. Springer International Publishing. https://doi.org/10.1007/978-3-319-14157-2

Patil, H., & Divekar, R. (2014). Inventory Management Challenges for B2C E-Commerce Retailers. *Procedia Economics and Finance*, 561-571.

Piatek, O., Ning, J. C.-m., & Touchette, D. R. (Nov 2020). National drug shortages worsen during COVID-19 crisis: Proposal for a comprehensive model to monitor and address critical drug shortages. *American Journal of Health-System Pharmacy*, 1778–1785.

Sperandei, S. (2014). Understanding logistic regression analysis. *Biochemia Medica*, *24*(1), 12–18. https://doi.org/10.11613/BM.2014.003

Sen, A., & Srivastava, M. (1990). *Regression Analysis: Theory, Methods, and Applications.* New York: Springer-Verlag.

Sheffi, Y. (2020). *The New (Ab)Normal: Reshaping Business and Supply Chain Strategy Beyond Covid-19.* Cambridge: MIT CTL Media.

Silver, E. A., Pyke, D. F., & Thomas, D. J. (2017). *Inventory Production Management in Supply Chains.* Boca Raton: CRC Press.

Socal, M. (Apr2021). The Pandemic and the Supply Chain: Gaps in Pharmaceutical Production and Distribution. *American Journal of Public Health*, 635-639.

Usman, K. (2008). *Determination of drivers of stockout performance of retail stores using data mining techniques* [Thesis, Massachusetts Institute of Technology]. https://dspace.mit.edu/handle/1721.1/45246

Wan, X., & Evers, P. T. (2011). Supply Chain Networks With Multiple Retailers: A Test of the Emerging Theory on Inventories, Stockouts, and Bullwhips. *Journal of Business Logistics*, *32*(1), 27–39. https://doi.org/10.1111/j.2158-1592.2011.01003.x

Ward, S. (2020, July 26). *The Balance Small Business*. Retrieved from https://www.thebalancesmb.com/a-definition-of-bricks-and-mortar-2947959

Appendix

Logistic Regression Output for Model 1-1

Omnibus Tests of Model Coefficients

		Chi-square	df	Sig.
Step 1	Step	2117.281	5	.000
	Block	2117.281	5	.000
	Model	2117.281	5	.000

Model Summary

Step	-2 Log likelihood	Cox & Snell R Square	Nagelkerke R Square
1	75560.710[a]	.030	.045

a. Estimation terminated at iteration number 6 because parameter estimates changed by less than .001.

Classification Table[a]

			Predicted		
			SO		
	Observed		0	1	Percentage Correct
Step 1	SO	0	51085	456	99.1
		1	16449	872	5.0
	Overall Percentage				75.5

a. The cut value is .500

Variables in the Equation

								95% C.I.for EXP(B)	
		B	S.E.	Wald	df	Sig.	Exp(B)	Lower	Upper
Step 1[a]	Amazon_Order_QTY_n	88.088	2.987	869.615	1	<.001	1.803E+38	5.168E+35	6.289E+40
	Inventory_n	-12.231	.926	174.452	1	<.001	.000	.000	.000
	OtherAmazon_n	4.543	.506	80.481	1	<.001	93.930	34.818	253.404
	Order_Qty2019Avg_n	16.625	2.500	44.225	1	<.001	16607520.46	123684.586	2229944287
	SO_Brand	-.546	.099	30.447	1	<.001	.579	.477	.703
	Constant	-1.110	.016	4675.482	1	.000	.330		

a. Variable(s) entered on step 1: Amazon_Order_QTY_n, Inventory_n, OtherAmazon_n, Order_Qty2019Avg_n, SO_Brand.

Logistic Regression Output for Model 1-2

Omnibus Tests of Model Coefficients

		Chi-square	df	Sig.
Step 1	Step	2295.517	7	.000
	Block	2295.517	7	.000
	Model	2295.517	7	.000

Model Summary

Step	-2 Log likelihood	Cox & Snell R Square	Nagelkerke R Square
1	75382.475[a]	.033	.048

a. Estimation terminated at iteration number 6 because parameter estimates changed by less than .001.

Classification Table[a]

			Predicted		
			SO		
	Observed		0	1	Percentage Correct
Step 1	SO	0	51086	455	99.1
		1	16438	883	5.1
	Overall Percentage				75.5

a. The cut value is .500

Variables in the Equation

								95% C.I.for EXP(B)	
		B	S.E.	Wald	df	Sig.	Exp(B)	Lower	Upper
Step 1[a]	Amazon_Order_QTY_n	84.216	2.968	805.357	1	<.001	3.756E+36	1.119E+34	1.261E+39
	Inventory_n	-11.378	.914	154.894	1	<.001	.000	.000	.000
	OtherAmazon_n	4.826	.510	89.628	1	<.001	124.691	45.914	338.633
	Order_Qty2019Avg_n	16.610	2.490	44.485	1	<.001	16356456.27	124127.100	2155320325
	SO_Brand	-.520	.099	27.311	1	<.001	.595	.489	.723
	DC1	.233	.025	83.864	1	<.001	1.263	1.201	1.327
	DC2	.337	.026	173.988	1	<.001	1.401	1.333	1.473
	Constant	-1.344	.025	2846.501	1	.000	.261		

a. Variable(s) entered on step 1: Amazon_Order_QTY_n, Inventory_n, OtherAmazon_n, Order_Qty2019Avg_n, SO_Brand, DC1, DC2.

Logistic Regression Output for Model 1-3

Omnibus Tests of Model Coefficients

		Chi-square	df	Sig.
Step 1	Step	3099.431	12	.000
	Block	3099.431	12	.000
	Model	3099.431	12	.000

Model Summary

Step	-2 Log likelihood	Cox & Snell R Square	Nagelkerke R Square
1	74578.561[a]	.044	.065

a. Estimation terminated at iteration number 6 because parameter estimates changed by less than .001.

Classification Table[a]

			Predicted		
			SO		
	Observed		0	1	Percentage Correct
Step 1	SO	0	50880	661	98.7
		1	15957	1364	7.9
	Overall Percentage				75.9

a. The cut value is .500

Variables in the Equation

		B	S.E.	Wald	df	Sig.	Exp(B)	95% C.I.for EXP(B) Lower	Upper
Step 1[a]	Amazon_Order_QTY_n	79.070	2.959	714.163	1	<.001	2.187E+34	6.626E+31	7.216E+36
	Inventory_n	-10.563	.901	137.498	1	<.001	.000	.000	.000
	OtherAmazon_n	4.416	.500	77.915	1	<.001	82.768	31.047	220.653
	Order_Qty2019Avg_n	15.788	2.461	41.142	1	<.001	7189375.382	57745.975	895077423.1
	SO_Brand	.020	.103	.036	1	.849	1.020	.833	1.248
	DC1	.219	.026	72.860	1	<.001	1.245	1.184	1.309
	DC2	.335	.026	169.060	1	<.001	1.397	1.329	1.470
	PEPCID	1.901	.089	453.059	1	<.001	6.691	5.616	7.970
	PLAX	2.701	.354	58.302	1	<.001	14.901	7.448	29.809
	TYLENOL	.728	.046	252.666	1	<.001	2.072	1.894	2.267
	SUDAFED	.260	.119	4.789	1	.029	1.297	1.027	1.636
	REACH	.344	.084	16.912	1	<.001	1.411	1.198	1.663
	Constant	-1.455	.026	3157.650	1	.000	.233		

a. Variable(s) entered on step 1: Amazon_Order_QTY_n, Inventory_n, OtherAmazon_n, Order_Qty2019Avg_n, SO_Brand, DC1, DC2, PEPCID, PLAX, TYLENOL, SUDAFED, REACH.

Logistic Regression Output for Model 1-4

Omnibus Tests of Model Coefficients

		Chi-square	df	Sig.
Step 1	Step	3226.657	12	.000
	Block	3226.657	12	.000
	Model	3226.657	12	.000

Model Summary

Step	-2 Log likelihood	Cox & Snell R Square	Nagelkerke R Square
1	74451.335[a]	.046	.068

a. Estimation terminated at iteration number 6 because parameter estimates changed by less than .001.

Classification Table[a]

			Predicted		
			SO		
	Observed		0	1	Percentage Correct
Step 1	SO	0	51009	532	99.0
		1	16321	1000	5.8
	Overall Percentage				75.5

a. The cut value is .500

Variables in the Equation

								95% C.I.for EXP(B)	
		B	S.E.	Wald	df	Sig.	Exp(B)	Lower	Upper
Step 1[a]	Amazon_Order_QTY_n	82.214	2.978	762.081	1	<.001	5.069E+35	1.479E+33	1.737E+38
	Inventory_n	-13.305	.945	198.038	1	<.001	.000	.000	.000
	OtherAmazon_n	2.915	.492	35.119	1	<.001	18.456	7.037	48.406
	Order_Qty2019Avg_n	10.562	2.384	19.625	1	<.001	38641.069	361.094	4135019.106
	SO_Brand	4.037	.243	276.479	1	<.001	56.678	35.216	91.222
	DC1	.274	.026	111.521	1	<.001	1.316	1.250	1.384
	DC2	.359	.026	192.566	1	<.001	1.432	1.361	1.507
	NEUTROGENA	-1.021	.054	362.640	1	<.001	.360	.324	.400
	AVEENO	-.665	.037	322.800	1	<.001	.514	.478	.553
	TYLENOL	.515	.047	118.468	1	<.001	1.673	1.525	1.836
	OGX	-.754	.051	220.675	1	<.001	.470	.426	.520
	LISTERINE	.457	.041	124.156	1	<.001	1.579	1.457	1.711
	Constant	-1.473	.026	3107.768	1	.000	.229		

a. Variable(s) entered on step 1: Amazon_Order_QTY_n, Inventory_n, OtherAmazon_n, Order_Qty2019Avg_n, SO_Brand, DC1, DC2, NEUTROGENA, AVEENO, TYLENOL, OGX, LISTERINE.

Logistic Regression Output for Model 1-5

Omnibus Tests of Model Coefficients

		Chi-square	df	Sig.
Step 1	Step	3812.261	16	.000
	Block	3812.261	16	.000
	Model	3812.261	16	.000

Model Summary

Step	-2 Log likelihood	Cox & Snell R Square	Nagelkerke R Square
1	73865.731[a]	.054	.080

a. Estimation terminated at iteration number 6 because parameter estimates changed by less than .001.

Classification Table[a]

			Predicted		
			SO		
	Observed		0	1	Percentage Correct
Step 1	SO	0	50835	706	98.6
		1	15899	1422	8.2
	Overall Percentage				75.9

a. The cut value is .500

Variables in the Equation

								95% C.I.for EXP(B)	
		B	S.E.	Wald	df	Sig.	Exp(B)	Lower	Upper
Step 1[a]	Amazon_Order_QTY_n	78.910	2.983	699.660	1	<.001	1.864E+34	5.382E+31	6.452E+36
	Inventory_n	-12.752	.937	185.148	1	<.001	.000	.000	.000
	OtherAmazon_n	2.760	.490	31.746	1	<.001	15.807	6.051	41.294
	Order_Qty2019Avg_n	10.900	2.391	20.779	1	<.001	54200.752	499.490	5881437.596
	SO_Brand	4.154	.244	289.910	1	<.001	63.687	39.481	102.735
	DC1	.266	.026	103.492	1	<.001	1.304	1.239	1.373
	DC2	.359	.026	190.285	1	<.001	1.432	1.360	1.507
	NEUTROGENA	-.960	.054	317.720	1	<.001	.383	.345	.426
	AVEENO	-.593	.037	252.757	1	<.001	.553	.514	.595
	TYLENOL	.598	.048	158.553	1	<.001	1.819	1.657	1.996
	OGX	-.695	.051	185.686	1	<.001	.499	.451	.551
	LISTERINE	.535	.041	168.384	1	<.001	1.707	1.575	1.851
	PEPCID	1.868	.090	434.148	1	<.001	6.477	5.433	7.722
	PLAX	2.793	.354	62.251	1	<.001	16.330	8.160	32.682
	SUDAFED	.307	.119	6.654	1	.010	1.359	1.076	1.715
	REACH	.393	.084	21.871	1	<.001	1.481	1.256	1.747
	Constant	-1.554	.027	3311.538	1	.000	.211		

a. Variable(s) entered on step 1: Amazon_Order_QTY_n, Inventory_n, OtherAmazon_n, Order_Qty2019Avg_n, SO_Brand, DC1, DC2, NEUTROGENA, AVEENO, TYLENOL, OGX, LISTERINE, PEPCID, PLAX, SUDAFED, REACH.

Logistic Regression Output for Model 1-6

Omnibus Tests of Model Coefficients

		Chi-square	df	Sig.
Step 1	Step	3920.193	19	.000
	Block	3920.193	19	.000
	Model	3920.193	19	.000

Model Summary

Step	-2 Log likelihood	Cox & Snell R Square	Nagelkerke R Square
1	73757.799[a]	.055	.082

a. Estimation terminated at iteration number 6 because parameter estimates changed by less than .001.

Classification Table[a]

			Predicted		
			SO		Percentage Correct
	Observed		0	1	
Step 1	SO	0	50821	720	98.6
		1	15893	1428	8.2
	Overall Percentage				75.9

a. The cut value is .500

Variables in the Equation

								95% C.I.for EXP(B)	
		B	S.E.	Wald	df	Sig.	Exp(B)	Lower	Upper
Step 1[a]	Amazon_Order_QTY_n	73.816	2.959	622.526	1	<.001	1.143E+32	3.465E+29	3.770E+34
	Inventory_n	-12.809	.938	186.334	1	<.001	.000	.000	.000
	OtherAmazon_n	2.621	.489	28.743	1	<.001	13.748	5.274	35.838
	Order_Qty2019Avg_n	16.203	2.511	41.641	1	<.001	10888538.59	79367.056	1493822229
	SO_Brand	1.416	.118	143.507	1	<.001	4.120	3.268	5.194
	DC1	.209	.026	65.506	1	<.001	1.232	1.171	1.296
	DC2	.323	.026	155.284	1	<.001	1.381	1.313	1.453
	PEPCID	2.171	.090	582.681	1	<.001	8.771	7.353	10.462
	PLAX	3.002	.354	71.901	1	<.001	20.121	10.054	40.270
	SUDAFED	.551	.119	21.358	1	<.001	1.735	1.374	2.192
	TYLENOL	.975	.047	433.519	1	<.001	2.651	2.419	2.906
	REACH	.656	.084	60.304	1	<.001	1.926	1.633	2.273
	IMODIUM	.504	.099	25.821	1	<.001	1.656	1.363	2.011
	LISTERINE	.923	.040	531.008	1	<.001	2.517	2.327	2.723
	DESITIN	1.059	.093	131.026	1	<.001	2.885	2.406	3.458
	RHINOCORT	.721	.208	12.045	1	<.001	2.057	1.369	3.092
	CLEANCLEAR	.653	.052	158.374	1	<.001	1.922	1.736	2.128
	NEOSPORIN	.437	.080	29.908	1	<.001	1.548	1.323	1.810
	BANDAID	.731	.046	255.234	1	<.001	2.078	1.900	2.273
	Constant	-1.733	.029	3646.017	1	.000	.177		

a. Variable(s) entered on step 1: Amazon_Order_QTY_n, Inventory_n, OtherAmazon_n, Order_Qty2019Avg_n, SO_Brand, DC1, DC2, PEPCID, PLAX, SUDAFED, TYLENOL, REACH, IMODIUM, LISTERINE, DESITIN, RHINOCORT, CLEANCLEAR, NEOSPORIN, BANDAID.

Logistic Regression Output for Model 2-1

Omnibus Tests of Model Coefficients

		Chi-square	df	Sig.
Step 1	Step	625.196	5	<.001
	Block	625.196	5	<.001
	Model	625.196	5	<.001

Model Summary

Step	-2 Log likelihood	Cox & Snell R Square	Nagelkerke R Square
1	31135.560[a]	.022	.032

a. Estimation terminated at iteration number 6 because parameter estimates changed by less than .001.

Classification Table[a]

			Predicted					
			Selected Cases[b]			Unselected Cases[c]		
			SO			SO		
	Observed		0	1	Percentage Correct	0	1	Percentage Correct
Step 1	SO	0	21334	115	99.5	29876	216	99.3
		1	6790	215	3.1	9850	466	4.5
	Overall Percentage				75.7			75.1

a. The cut value is .500

b. Selected cases DC1 EQ 1

c. Unselected cases DC1 NE 1

Variables in the Equation

		B	S.E.	Wald	df	Sig.	Exp(B)	95% C.I.for EXP(B) Lower	Upper
Step 1[a]	Amazon_Order_QTY_n	67.777	4.688	209.000	1	<.001	2.723E+29	2.782E+25	2.664E+33
	Inventory_n	-2.961	.652	20.640	1	<.001	.052	.014	.186
	OtherAmazon_n	4.975	.902	30.431	1	<.001	144.820	24.723	848.325
	Order_Qty2019Avg_n	16.412	4.859	11.407	1	<.001	13419444.98	980.247	1.837E+11
	SO_Brand	-1.447	.154	88.024	1	<.001	.235	.174	.318
	Constant	-1.031	.025	1645.156	1	.000	.357		

a. Variable(s) entered on step 1: Amazon_Order_QTY_n, Inventory_n, OtherAmazon_n, Order_Qty2019Avg_n, SO_Brand.

Logistic Regression Output for Model 2-2

Omnibus Tests of Model Coefficients

		Chi-square	df	Sig.
Step 1	Step	942.370	10	<.001
	Block	942.370	10	<.001
	Model	942.370	10	<.001

Model Summary

Step	-2 Log likelihood	Cox & Snell R Square	Nagelkerke R Square
1	30818.385[a]	.033	.048

a. Estimation terminated at iteration number 6 because parameter estimates changed by less than .001.

Classification Table[a]

			Predicted					
			Selected Cases[b]			Unselected Cases[c]		
			SO			SO		
	Observed		0	1	Percentage Correct	0	1	Percentage Correct
Step 1	SO	0	21215	234	98.9	29780	312	99.0
		1	6575	430	6.1	9589	727	7.0
	Overall Percentage				76.1			75.5

a. The cut value is .500

b. Selected cases DC1 EQ 1

c. Unselected cases DC1 NE 1

Variables in the Equation

								95% C.I.for EXP(B)	
		B	S.E.	Wald	df	Sig.	Exp(B)	Lower	Upper
Step 1[a]	Amazon_Order_QTY_n	63.022	4.641	184.402	1	<.001	2.345E+27	2.628E+23	2.091E+31
	Inventory_n	-2.613	.621	17.697	1	<.001	.073	.022	.248
	OtherAmazon_n	4.714	.891	27.973	1	<.001	111.531	19.439	639.899
	Order_Qty2019Avg_n	14.147	4.730	8.946	1	.003	1393125.666	131.193	1.479E+10
	SO_Brand	-.926	.159	33.953	1	<.001	.396	.290	.541
	PLAX	2.662	.637	17.470	1	<.001	14.322	4.111	49.897
	PEPCID	1.753	.130	183.062	1	<.001	5.773	4.478	7.442
	RHINOCORT	1.021	.268	14.539	1	<.001	2.775	1.642	4.690
	SUDAFED	.169	.185	.834	1	.361	1.184	.824	1.701
	TYLENOL	.710	.071	98.980	1	<.001	2.034	1.768	2.339
	Constant	-1.145	.027	1822.973	1	.000	.318		

a. Variable(s) entered on step 1: Amazon_Order_QTY_n, Inventory_n, OtherAmazon_n, Order_Qty2019Avg_n, SO_Brand, PLAX, PEPCID, RHINOCORT, SUDAFED, TYLENOL.

Logistic Regression Output for Model 2-3

Omnibus Tests of Model Coefficients

		Chi-square	df	Sig.
Step 1	Step	910.668	10	<.001
	Block	910.668	10	<.001
	Model	910.668	10	<.001

Model Summary

Step	-2 Log likelihood	Cox & Snell R Square	Nagelkerke R Square
1	30850.088[a]	.031	.047

a. Estimation terminated at iteration number 6 because parameter estimates changed by less than .001.

Classification Table[a]

			Predicted					
			Selected Cases[b]			Unselected Cases[c]		
			SO			SO		
	Observed		0	1	Percentage Correct	0	1	Percentage Correct
Step 1	SO	0	21281	168	99.2	29836	256	99.1
		1	6760	245	3.5	9820	496	4.8
	Overall Percentage				75.7			75.1

a. The cut value is .500

b. Selected cases DC1 EQ 1

c. Unselected cases DC1 NE 1

Variables in the Equation

		B	S.E.	Wald	df	Sig.	Exp(B)	95% C.I.for EXP(B) Lower	Upper
Step 1[a]	Amazon_Order_QTY_n	64.193	4.737	183.622	1	<.001	7.566E+27	7.022E+23	8.151E+31
	Inventory_n	-3.765	.752	25.054	1	<.001	.023	.005	.101
	OtherAmazon_n	3.131	.878	12.706	1	<.001	22.905	4.094	128.139
	Order_Qty2019Avg_n	10.341	4.698	4.846	1	.028	30982.594	3.108	308805684.7
	SO_Brand	.519	.227	5.226	1	.022	1.681	1.077	2.623
	NEUTROGENA	-.434	.047	84.439	1	<.001	.648	.591	.711
	AVEENO	-.401	.046	75.211	1	<.001	.669	.611	.733
	TYLENOL	.627	.072	75.789	1	<.001	1.872	1.626	2.156
	LISTERINE	.457	.065	49.325	1	<.001	1.580	1.390	1.794
	REACH	.557	.124	20.095	1	<.001	1.746	1.368	2.228
	Constant	-1.121	.028	1603.142	1	.000	.326		

a. Variable(s) entered on step 1: Amazon_Order_QTY_n, Inventory_n, OtherAmazon_n, Order_Qty2019Avg_n, SO_Brand, NEUTROGENA, AVEENO, TYLENOL, LISTERINE, REACH.

Logistic Regression Output for Model 2-4

Omnibus Tests of Model Coefficients

		Chi-square	df	Sig.
Step 1	Step	1155.486	14	<.001
	Block	1155.486	14	<.001
	Model	1155.486	14	<.001

Model Summary

Step	-2 Log likelihood	Cox & Snell R Square	Nagelkerke R Square
1	30605.269[a]	.040	.059

a. Estimation terminated at iteration number 6 because parameter estimates changed by less than .001.

Classification Table[a]

			Predicted					
			Selected Cases[b]			Unselected Cases[c]		
			SO			SO		
	Observed		0	1	Percentage Correct	0	1	Percentage Correct
Step 1	SO	0	21199	250	98.8	29749	343	98.9
		1	6571	434	6.2	9578	738	7.2
	Overall Percentage				76.0			75.4

a. The cut value is .500

b. Selected cases DC1 EQ 1

c. Unselected cases DC1 NE 1

Variables in the Equation

		B	S.E.	Wald	df	Sig.	Exp(B)	95% C.I.for EXP(B) Lower	Upper
Step 1[a]	Amazon_Order_QTY_n	61.317	4.714	169.202	1	<.001	4.260E+26	4.140E+22	4.384E+30
	Inventory_n	-3.505	.730	23.044	1	<.001	.030	.007	.126
	OtherAmazon_n	3.054	.877	12.125	1	<.001	21.205	3.800	118.317
	Order_Qty2019Avg_n	10.207	4.677	4.763	1	.029	27101.276	2.829	259586839.1
	SO_Brand	.899	.230	15.297	1	<.001	2.458	1.566	3.858
	NEUTROGENA	-.429	.047	82.108	1	<.001	.651	.594	.715
	AVEENO	-.355	.046	58.488	1	<.001	.701	.640	.768
	TYLENOL	.702	.072	94.432	1	<.001	2.018	1.752	2.325
	LISTERINE	.525	.065	64.706	1	<.001	1.690	1.487	1.921
	REACH	.654	.124	27.621	1	<.001	1.923	1.507	2.454
	PLAX	2.740	.637	18.503	1	<.001	15.485	4.444	53.961
	PEPCID	1.783	.130	188.193	1	<.001	5.950	4.612	7.677
	RHINOCORT	1.084	.268	16.386	1	<.001	2.957	1.749	4.998
	SUDAFED	.231	.185	1.560	1	.212	1.260	.877	1.811
	Constant	-1.211	.029	1722.115	1	.000	.298		

a. Variable(s) entered on step 1: Amazon_Order_QTY_n, Inventory_n, OtherAmazon_n, Order_Qty2019Avg_n, SO_Brand, NEUTROGENA, AVEENO, TYLENOL, LISTERINE, REACH, PLAX, PEPCID, RHINOCORT, SUDAFED.

Logistic Regression Output for Model 2-5

Omnibus Tests of Model Coefficients

		Chi-square	df	Sig.
Step 1	Step	1227.192	16	<.001
	Block	1227.192	16	<.001
	Model	1227.192	16	<.001

Model Summary

Step	-2 Log likelihood	Cox & Snell R Square	Nagelkerke R Square
1	30533.563[a]	.042	.063

a. Estimation terminated at iteration number 6 because parameter estimates changed by less than .001.

Classification Table[a]

			Predicted					
			Selected Cases[b]			Unselected Cases[c]		
			SO			SO		
	Observed		0	1	Percentage Correct	0	1	Percentage Correct
Step 1	SO	0	21185	264	98.8	29738	354	98.8
		1	6551	454	6.5	9564	752	7.3
	Overall Percentage				76.0			75.5

a. The cut value is .500

b. Selected cases DC1 EQ 1

c. Unselected cases DC1 NE 1

Variables in the Equation

								95% C.I.for EXP(B)	
		B	S.E.	Wald	df	Sig.	Exp(B)	Lower	Upper
Step 1[a]	Amazon_Order_QTY_n	56.758	4.632	150.128	1	<.001	4.462E+24	5.088E+20	3.914E+28
	Inventory_n	-3.261	.696	21.920	1	<.001	.038	.010	.150
	OtherAmazon_n	3.115	.883	12.451	1	<.001	22.529	3.994	127.090
	Order_Qty2019Avg_n	14.623	4.924	8.819	1	.003	2243174.260	144.360	3.486E+10
	SO_Brand	.405	.182	4.947	1	.026	1.499	1.049	2.142
	PLAX	2.950	.637	21.436	1	<.001	19.102	5.480	66.590
	PEPCID	2.012	.131	237.434	1	<.001	7.481	5.792	9.664
	RHINOCORT	1.302	.268	23.575	1	<.001	3.676	2.174	6.218
	SUDAFED	.448	.186	5.834	1	.016	1.566	1.088	2.252
	TYLENOL	.943	.073	167.730	1	<.001	2.568	2.226	2.962
	REACH	.884	.125	50.002	1	<.001	2.421	1.895	3.094
	IMODIUM	.546	.157	12.076	1	<.001	1.726	1.269	2.348
	DESITIN	1.028	.141	53.331	1	<.001	2.797	2.122	3.685
	LISTERINE	.763	.066	134.916	1	<.001	2.146	1.886	2.440
	CLEANCLEAR	.695	.081	74.079	1	<.001	2.004	1.711	2.348
	BANDAID	.662	.073	82.384	1	<.001	1.938	1.680	2.236
	Constant	-1.423	.033	1861.641	1	.000	.241		

a. Variable(s) entered on step 1: Amazon_Order_QTY_n, Inventory_n, OtherAmazon_n, Order_Qty2019Avg_n, SO_Brand, PLAX, PEPCID, RHINOCORT, SUDAFED, TYLENOL, REACH, IMODIUM, DESITIN, LISTERINE, CLEANCLEAR, BANDAID.

Logistic Regression Output for Model 3-1

Omnibus Tests of Model Coefficients

		Chi-square	df	Sig.
Step 1	Step	1365.993	5	<.001
	Block	1365.993	5	<.001
	Model	1365.993	5	<.001

Model Summary

Step	-2 Log likelihood	Cox & Snell R Square	Nagelkerke R Square
1	30017.766[a]	.051	.073

a. Estimation terminated at iteration number 6 because parameter estimates changed by less than .001.

Classification Table[a]

			Predicted					
			Selected Cases[b]			Unselected Cases[c]		
			SO			SO		
	Observed		0	1	Percentage Correct	0	1	Percentage Correct
Step 1	SO	0	18538	294	98.4	32456	253	99.2
		1	6797	671	9.0	9454	399	4.0
	Overall Percentage				73.0			77.2

a. The cut value is .500

b. Selected cases DC2 EQ 1

c. Unselected cases DC2 NE 1

Variables in the Equation

								95% C.I.for EXP(B)	
		B	S.E.	Wald	df	Sig.	Exp(B)	Lower	Upper
Step 1[a]	Amazon_Order_QTY_n	91.853	4.250	467.076	1	<.001	7.788E+39	1.878E+36	3.229E+43
	Inventory_n	-38.801	2.549	231.718	1	<.001	.000	.000	.000
	OtherAmazon_n	3.727	.975	14.622	1	<.001	41.556	6.152	280.722
	Order_Qty2019Avg_n	21.798	3.635	35.963	1	<.001	2930092063	2359491.883	3.639E+12
	SO_Brand	-.097	.159	.372	1	.542	.907	.664	1.240
	Constant	-.960	.025	1435.322	1	.000	.383		

a. Variable(s) entered on step 1: Amazon_Order_QTY_n, Inventory_n, OtherAmazon_n, Order_Qty2019Avg_n, SO_Brand.

Logistic Regression Output for Model 3-2

Omnibus Tests of Model Coefficients

		Chi-square	df	Sig.
Step 1	Step	1684.324	10	.000
	Block	1684.324	10	.000
	Model	1684.324	10	.000

Model Summary

Step	-2 Log likelihood	Cox & Snell R Square	Nagelkerke R Square
1	29699.434[a]	.062	.089

a. Estimation terminated at iteration number 6 because parameter estimates changed by less than .001.

Classification Table[a]

			Predicted					
			Selected Cases[b]			Unselected Cases[c]		
			SO			SO		
	Observed		0	1	Percentage Correct	0	1	Percentage Correct
Step 1	SO	0	18479	353	98.1	32328	381	98.8
		1	6591	877	11.7	9209	644	6.5
	Overall Percentage				73.6			77.5

a. The cut value is .500

b. Selected cases DC2 EQ 1

c. Unselected cases DC2 NE 1

Variables in the Equation

		B	S.E.	Wald	df	Sig.	Exp(B)	95% C.I.for EXP(B) Lower	Upper
Step 1[a]	Amazon_Order_QTY_n	87.488	4.248	424.184	1	<.001	9.899E+37	2.398E+34	4.087E+41
	Inventory_n	-36.960	2.531	213.232	1	<.001	.000	.000	.000
	OtherAmazon_n	3.267	.983	11.046	1	<.001	26.225	3.820	180.038
	Order_Qty2019Avg_n	20.914	3.604	33.665	1	<.001	1209682802	1034084.897	1.415E+12
	SO_Brand	.400	.165	5.908	1	.015	1.492	1.081	2.060
	PLAX	3.017	.760	15.748	1	<.001	20.435	4.604	90.690
	PEPCID	2.023	.148	187.118	1	<.001	7.563	5.660	10.107
	SUDAFED	.167	.175	.914	1	.339	1.182	.839	1.665
	TYLENOL	.616	.071	75.060	1	<.001	1.851	1.610	2.128
	REACH	.286	.128	4.973	1	.026	1.331	1.035	1.712
	Constant	-1.068	.027	1588.481	1	.000	.344		

a. Variable(s) entered on step 1: Amazon_Order_QTY_n, Inventory_n, OtherAmazon_n, Order_Qty2019Avg_n, SO_Brand, PLAX, PEPCID, SUDAFED, TYLENOL, REACH.

Logistic Regression Output for Model 3-3

Omnibus Tests of Model Coefficients

		Chi-square	df	Sig.
Step 1	Step	1560.455	10	.000
	Block	1560.455	10	.000
	Model	1560.455	10	.000

Model Summary

Step	-2 Log likelihood	Cox & Snell R Square	Nagelkerke R Square
1	29823.304[a]	.058	.083

a. Estimation terminated at iteration number 6 because parameter estimates changed by less than .001.

Classification Table[a]

			Predicted					
			Selected Cases[b]			Unselected Cases[c]		
			SO			SO		
	Observed		0	1	Percentage Correct	0	1	Percentage Correct
Step 1	SO	0	18503	329	98.3	32426	283	99.1
		1	6748	720	9.6	9406	447	4.5
	Overall Percentage				73.1			77.2

a. The cut value is .500

b. Selected cases DC2 EQ 1

c. Unselected cases DC2 NE 1

Variables in the Equation

								95% C.I.for EXP(B)	
		B	S.E.	Wald	df	Sig.	Exp(B)	Lower	Upper
Step 1[a]	Amazon_Order_QTY_n	89.679	4.328	429.397	1	<.001	8.850E+38	1.833E+35	4.273E+42
	Inventory_n	-41.081	2.573	254.913	1	<.001	.000	.000	.000
	OtherAmazon_n	1.704	.991	2.956	1	.086	5.494	.788	38.310
	Order_Qty2019Avg_n	18.889	3.661	26.618	1	<.001	159726293.1	122174.719	2.088E+11
	SO_Brand	.764	.213	12.893	1	<.001	2.148	1.415	3.260
	NEUTROGENA	-.138	.048	8.263	1	.004	.871	.792	.957
	AVEENO	-.166	.042	15.652	1	<.001	.847	.780	.920
	TYLENOL	.594	.072	68.277	1	<.001	1.811	1.573	2.084
	LISTERINE	.625	.062	102.141	1	<.001	1.868	1.655	2.109
	REACH	.264	.129	4.211	1	.040	1.303	1.012	1.677
	Constant	-1.041	.028	1411.329	1	.000	.353		

a. Variable(s) entered on step 1: Amazon_Order_QTY_n, Inventory_n, OtherAmazon_n, Order_Qty2019Avg_n, SO_Brand, NEUTROGENA, AVEENO, TYLENOL, LISTERINE, REACH.

Logistic Regression Output for Model 3-4

Omnibus Tests of Model Coefficients

		Chi-square	df	Sig.
Step 1	Step	1826.681	13	.000
	Block	1826.681	13	.000
	Model	1826.681	13	.000

Model Summary

Step	-2 Log likelihood	Cox & Snell R Square	Nagelkerke R Square
1	29557.077[a]	.067	.096

a. Estimation terminated at iteration number 6 because parameter estimates changed by less than .001.

Classification Table[a]

			Predicted					
			Selected Cases[b]			Unselected Cases[c]		
			SO			SO		
	Observed		0	1	Percentage Correct	0	1	Percentage Correct
Step 1	SO	0	18455	377	98.0	32314	395	98.8
		1	6578	890	11.9	9179	674	6.8
	Overall Percentage				73.6			77.5

a. The cut value is .500

b. Selected cases DC2 EQ 1

c. Unselected cases DC2 NE 1

Variables in the Equation

		B	S.E.	Wald	df	Sig.	Exp(B)	95% C.I.for EXP(B) Lower	Upper
Step 1[a]	Amazon_Order_QTY_n	87.680	4.311	413.624	1	<.001	1.199E+38	2.566E+34	5.605E+41
	Inventory_n	-39.382	2.555	237.558	1	<.001	.000	.000	.000
	OtherAmazon_n	1.508	.996	2.293	1	.130	4.516	.642	31.778
	Order_Qty2019Avg_n	18.625	3.646	26.101	1	<.001	122711437.7	96762.323	1.556E+11
	SO_Brand	1.079	.215	25.188	1	<.001	2.941	1.930	4.482
	NEUTROGENA	-.131	.048	7.395	1	.007	.877	.798	.964
	AVEENO	-.127	.042	9.031	1	.003	.881	.811	.957
	TYLENOL	.659	.072	83.726	1	<.001	1.932	1.678	2.225
	LISTERINE	.682	.062	121.142	1	<.001	1.978	1.752	2.234
	REACH	.346	.129	7.213	1	.007	1.414	1.098	1.820
	PLAX	3.083	.760	16.438	1	<.001	21.825	4.917	96.880
	PEPCID	2.070	.148	195.027	1	<.001	7.928	5.929	10.601
	SUDAFED	.223	.175	1.627	1	.202	1.250	.887	1.763
	Constant	-1.121	.029	1532.154	1	.000	.326		

a. Variable(s) entered on step 1: Amazon_Order_QTY_n, Inventory_n, OtherAmazon_n, Order_Qty2019Avg_n, SO_Brand, NEUTROGENA, AVEENO, TYLENOL, LISTERINE, REACH, PLAX, PEPCID, SUDAFED.

Logistic Regression Output for Model 3-5

Omnibus Tests of Model Coefficients

		Chi-square	df	Sig.
Step 1	Step	2028.439	19	.000
	Block	2028.439	19	.000
	Model	2028.439	19	.000

Model Summary

Step	-2 Log likelihood	Cox & Snell R Square	Nagelkerke R Square
1	29355.320[a]	.074	.107

a. Estimation terminated at iteration number 6 because parameter estimates changed by less than .001.

Classification Table[a]

			Predicted					
			Selected Cases[b]			Unselected Cases[c]		
			SO		Percentage Correct	SO		Percentage Correct
	Observed		0	1		0	1	
Step 1	SO	0	18440	392	97.9	32281	428	98.7
		1	6546	922	12.3	9157	696	7.1
	Overall Percentage				73.6			77.5

a. The cut value is .500

b. Selected cases DC2 EQ 1

c. Unselected cases DC2 NE 1

Variables in the Equation

								95% C.I.for EXP(B)	
		B	S.E.	Wald	df	Sig.	Exp(B)	Lower	Upper
Step 1[a]	Amazon_Order_QTY_n	82.407	4.282	370.310	1	<.001	6.148E+35	1.392E+32	2.716E+39
	Inventory_n	-40.059	2.565	243.982	1	<.001	.000	.000	.000
	OtherAmazon_n	1.263	1.004	1.581	1	.209	3.535	.494	25.301
	Order_Qty2019Avg_n	22.510	3.738	36.265	1	<.001	5971987108	3929571.835	9.076E+12
	SO_Brand	1.898	.203	87.749	1	<.001	6.670	4.484	9.920
	PEPCID	2.269	.149	231.520	1	<.001	9.674	7.222	12.958
	PLAX	3.303	.761	18.853	1	<.001	27.191	6.123	120.761
	SUDAFED	.440	.176	6.248	1	.012	1.553	1.100	2.193
	TYLENOL	.842	.073	131.951	1	<.001	2.320	2.010	2.678
	REACH	.581	.130	20.035	1	<.001	1.788	1.386	2.306
	DESITIN	1.110	.139	63.658	1	<.001	3.033	2.310	3.984
	IMODIUM	.312	.143	4.736	1	.030	1.366	1.031	1.809
	LISTERINE	.859	.063	184.864	1	<.001	2.361	2.086	2.673
	CLEANCLEAR	.661	.078	71.934	1	<.001	1.936	1.662	2.255
	ZYRTEC	.618	.185	11.185	1	<.001	1.856	1.292	2.666
	RHINOCORT	-.004	.410	.000	1	.993	.996	.446	2.224
	BANDAID	.713	.070	104.763	1	<.001	2.040	1.780	2.339
	JOHNSONS	-.055	.066	.698	1	.403	.946	.832	1.077
	OGX	-.102	.045	5.081	1	.024	.903	.826	.987
	Constant	-1.338	.034	1514.317	1	.000	.262		

a. Variable(s) entered on step 1: Amazon_Order_QTY_n, Inventory_n, OtherAmazon_n, Order_Qty2019Avg_n, SO_Brand, PEPCID, PLAX, SUDAFED, TYLENOL, REACH, DESITIN, IMODIUM, LISTERINE, CLEANCLEAR, ZYRTEC, RHINOCORT, BANDAID, JOHNSONS, OGX.

Logistic Regression Output for Model 4-1

Omnibus Tests of Model Coefficients

		Chi-square	df	Sig.
Step 1	Step	435.793	5	<.001
	Block	435.793	5	<.001
	Model	435.793	5	<.001

Model Summary

Step	-2 Log likelihood	Cox & Snell R Square	Nagelkerke R Square
1	13756.450[a]	.030	.048

a. Estimation terminated at iteration number 6 because parameter estimates changed by less than .001.

Classification Table[a]

			Predicted					
			Selected Cases[b]			Unselected Cases[c]		
			SO			SO		
	Observed		0	1	Percentage Correct	0	1	Percentage Correct
Step 1	SO	0	11219	41	99.6	40097	184	99.5
		1	2800	48	1.7	13961	512	3.5
	Overall Percentage				79.9			74.2

a. The cut value is .500

b. Selected cases DC3 EQ 1

c. Unselected cases DC3 NE 1

Variables in the Equation

								95% C.I.for EXP(B)	
		B	S.E.	Wald	df	Sig.	Exp(B)	Lower	Upper
Step 1[a]	Amazon_Order_QTY_n	76.902	9.118	71.128	1	<.001	2.501E+33	4.331E+25	1.445E+41
	Inventory_n	-43.331	3.190	184.538	1	<.001	.000	.000	.000
	OtherAmazon_n	7.599	.957	63.003	1	<.001	1996.736	305.761	13039.462
	Order_Qty2019Avg_n	5.688	4.206	1.829	1	.176	295.305	.078	1123491.237
	SO_Brand	1.264	.231	29.895	1	<.001	3.540	2.250	5.569
	Constant	-1.388	.040	1235.315	1	<.001	.249		

a. Variable(s) entered on step 1: Amazon_Order_QTY_n, Inventory_n, OtherAmazon_n, Order_Qty2019Avg_n, SO_Brand.

Logistic Regression Output for Model 4-2

Omnibus Tests of Model Coefficients

		Chi-square	df	Sig.
Step 1	Step	815.168	10	<.001
	Block	815.168	10	<.001
	Model	815.168	10	<.001

Model Summary

Step	-2 Log likelihood	Cox & Snell R Square	Nagelkerke R Square
1	13377.075[a]	.056	.089

a. Estimation terminated at iteration number 6 because parameter estimates changed by less than .001.

Classification Table[a]

			Predicted					
			Selected Cases[b]			Unselected Cases[c]		
			SO			SO		
	Observed		0	1	Percentage Correct	0	1	Percentage Correct
Step 1	SO	0	11167	93	99.2	39819	462	98.9
		1	2719	129	4.5	13390	1083	7.5
	Overall Percentage				80.1			74.7

a. The cut value is .500

b. Selected cases DC3 EQ 1

c. Unselected cases DC3 NE 1

Variables in the Equation

		B	S.E.	Wald	df	Sig.	Exp(B)	95% C.I.for EXP(B) Lower	Upper
Step 1[a]	Amazon_Order_QTY_n	68.653	9.085	57.108	1	<.001	6.541E+29	1.210E+22	3.536E+37
	Inventory_n	-46.447	3.201	210.518	1	<.001	.000	.000	.000
	OtherAmazon_n	4.445	.861	26.648	1	<.001	85.166	15.754	460.417
	Order_Qty2019Avg_n	5.254	4.246	1.531	1	.216	191.334	.047	787040.161
	SO_Brand	2.473	.247	100.460	1	<.001	11.862	7.313	19.241
	PEPCID	2.020	.233	75.276	1	<.001	7.537	4.776	11.895
	PLAX	2.817	.524	28.891	1	<.001	16.732	5.989	46.742
	LISTERINE	1.254	.084	223.618	1	<.001	3.504	2.973	4.130
	TYLENOL	1.196	.112	113.216	1	<.001	3.308	2.653	4.123
	NEOSPORIN	.879	.189	21.621	1	<.001	2.408	1.663	3.488
	Constant	-1.657	.044	1421.172	1	.000	.191		

a. Variable(s) entered on step 1: Amazon_Order_QTY_n, Inventory_n, OtherAmazon_n, Order_Qty2019Avg_n, SO_Brand, PEPCID, PLAX, LISTERINE, TYLENOL, NEOSPORIN.

Logistic Regression Output for Model 4-3

Omnibus Tests of Model Coefficients

		Chi-square	df	Sig.
Step 1	Step	698.935	10	<.001
	Block	698.935	10	<.001
	Model	698.935	10	<.001

Model Summary

Step	-2 Log likelihood	Cox & Snell R Square	Nagelkerke R Square
1	13493.308[a]	.048	.076

a. Estimation terminated at iteration number 6 because parameter estimates changed by less than .001.

Classification Table[a]

			Predicted					
			Selected Cases[b]			Unselected Cases[c]		
			SO			SO		
	Observed		0	1	Percentage Correct	0	1	Percentage Correct
Step 1	SO	0	11175	85	99.2	39984	297	99.3
		1	2749	99	3.5	13887	586	4.0
	Overall Percentage				79.9			74.1

a. The cut value is .500

b. Selected cases DC3 EQ 1

c. Unselected cases DC3 NE 1

Variables in the Equation

		B	S.E.	Wald	df	Sig.	Exp(B)	95% C.I.for EXP(B) Lower	Upper
Step 1[a]	Amazon_Order_QTY_n	69.650	9.095	58.643	1	<.001	1.773E+30	3.212E+22	9.785E+37
	Inventory_n	-47.656	3.206	221.014	1	<.001	.000	.000	.000
	OtherAmazon_n	5.571	.928	36.080	1	<.001	262.741	42.662	1618.150
	Order_Qty2019Avg_n	1.360	4.236	.103	1	.748	3.897	.001	15716.171
	SO_Brand	4.757	.498	91.310	1	<.001	116.385	43.869	308.772
	OGX	-.548	.103	28.264	1	<.001	.578	.472	.708
	NEUTROGENA	-.892	.118	57.540	1	<.001	.410	.325	.516
	LISTERINE	.893	.087	104.369	1	<.001	2.443	2.058	2.899
	AVEENO	-.729	.085	72.970	1	<.001	.483	.408	.570
	IMODIUM	.498	.286	3.040	1	.081	1.646	.940	2.883
	Constant	-1.456	.042	1211.276	1	<.001	.233		

a. Variable(s) entered on step 1: Amazon_Order_QTY_n, Inventory_n, OtherAmazon_n, Order_Qty2019Avg_n, SO_Brand, OGX, NEUTROGENA, LISTERINE, AVEENO, IMODIUM.

Logistic Regression Output for Model 4-4

Omnibus Tests of Model Coefficients

		Chi-square	df	Sig.
Step 1	Step	872.620	14	<.001
	Block	872.620	14	<.001
	Model	872.620	14	<.001

Model Summary

Step	-2 Log likelihood	Cox & Snell R Square	Nagelkerke R Square
1	13319.623[a]	.060	.095

a. Estimation terminated at iteration number 6 because parameter estimates changed by less than .001.

Classification Table[a]

			Predicted					
			Selected Cases[b]			Unselected Cases[c]		
			SO			SO		
	Observed		0	1	Percentage Correct	0	1	Percentage Correct
Step 1	SO	0	11152	108	99.0	39808	473	98.8
		1	2697	151	5.3	13345	1128	7.8
	Overall Percentage				80.1			74.8

a. The cut value is .500

b. Selected cases DC3 EQ 1

c. Unselected cases DC3 NE 1

Variables in the Equation

								95% C.I.for EXP(B)	
		B	S.E.	Wald	df	Sig.	Exp(B)	Lower	Upper
Step 1[a]	Amazon_Order_QTY_n	70.124	9.132	58.968	1	<.001	2.848E+30	4.803E+22	1.689E+38
	Inventory_n	-46.977	3.193	216.418	1	<.001	.000	.000	.000
	OtherAmazon_n	4.609	.868	28.197	1	<.001	100.422	18.321	550.426
	Order_Qty2019Avg_n	2.170	4.232	.263	1	.608	8.756	.002	35068.165
	SO_Brand	4.273	.506	71.251	1	<.001	71.709	26.590	193.387
	LISTERINE	1.103	.090	151.144	1	<.001	3.013	2.527	3.592
	IMODIUM	.681	.286	5.650	1	.017	1.975	1.127	3.462
	OGX	-.286	.107	7.170	1	.007	.751	.610	.926
	NEUTROGENA	-.616	.121	25.745	1	<.001	.540	.426	.685
	AVEENO	-.502	.088	32.264	1	<.001	.606	.509	.720
	PEPCID	1.926	.234	67.930	1	<.001	6.859	4.339	10.843
	PLAX	2.781	.524	28.129	1	<.001	16.133	5.773	45.082
	TYLENOL	1.037	.117	77.879	1	<.001	2.820	2.240	3.550
	NEOSPORIN	.814	.190	18.439	1	<.001	2.257	1.556	3.272
	Constant	-1.620	.045	1273.681	1	<.001	.198		

a. Variable(s) entered on step 1: Amazon_Order_QTY_n, Inventory_n, OtherAmazon_n, Order_Qty2019Avg_n, SO_Brand, LISTERINE, IMODIUM, OGX, NEUTROGENA, AVEENO, PEPCID, PLAX, TYLENOL, NEOSPORIN.

Logistic Regression Output for Model 4-5

Omnibus Tests of Model Coefficients

		Chi-square	df	Sig.
Step 1	Step	821.305	11	<.001
	Block	821.305	11	<.001
	Model	821.305	11	<.001

Model Summary

Step	-2 Log likelihood	Cox & Snell R Square	Nagelkerke R Square
1	13370.938[a]	.057	.089

a. Estimation terminated at iteration number 6 because parameter estimates changed by less than .001.

Classification Table[a]

			Predicted					
			Selected Cases[b]			Unselected Cases[c]		
			SO			SO		
	Observed		0	1	Percentage Correct	0	1	Percentage Correct
Step 1	SO	0	11167	93	99.2	39814	467	98.8
		1	2721	127	4.5	13383	1090	7.5
	Overall Percentage				80.1			74.7

a. The cut value is .500

b. Selected cases DC3 EQ 1

c. Unselected cases DC3 NE 1

Variables in the Equation

		B	S.E.	Wald	df	Sig.	Exp(B)	95% C.I.for EXP(B) Lower	Upper
Step 1[a]	Amazon_Order_QTY_n	68.885	9.094	57.381	1	<.001	8.248E+29	1.499E+22	4.539E+37
	Inventory_n	-46.500	3.203	210.752	1	<.001	.000	.000	.000
	OtherAmazon_n	4.392	.858	26.213	1	<.001	80.823	15.042	434.287
	Order_Qty2019Avg_n	5.261	4.248	1.534	1	.216	192.617	.047	795405.688
	SO_Brand	2.531	.248	104.156	1	<.001	12.572	7.731	20.442
	PEPCID	2.031	.233	76.068	1	<.001	7.621	4.829	12.030
	PLAX	2.830	.524	29.137	1	<.001	16.937	6.062	47.318
	LISTERINE	1.264	.084	226.513	1	<.001	3.538	3.002	4.171
	TYLENOL	1.205	.113	114.803	1	<.001	3.338	2.678	4.162
	NEOSPORIN	.891	.189	22.187	1	<.001	2.437	1.682	3.530
	IMODIUM	.754	.286	6.958	1	.008	2.126	1.214	3.724
	Constant	-1.669	.044	1419.512	1	.000	.188		

a. Variable(s) entered on step 1: Amazon_Order_QTY_n, Inventory_n, OtherAmazon_n, Order_Qty2019Avg_n, SO_Brand, PEPCID, PLAX, LISTERINE, TYLENOL, NEOSPORIN, IMODIUM.