Estimating On-Shelf Availability of CPG Products at Nanostores in India

by

Stephanie Gabriela Gomez Prieto
Bachelor of Engineering, Industrial Engineering, Universidad Católica San Pablo (2012)

and

Matthias Eder
BSc. Business Administration / BA Economics, Webster University Vienna (2016)

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

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

May 2024

Signature of Author: _____
Department of Supply Chain Management
May 10, 2024

Signature of Author: _____
Department of Supply Chain Management
May 10, 2024

Certified by: _____
Dr. Inma Borrella
Research Scientist, Academic Lead MITx MicroMasters Program in Supply Chain Management
Capstone Advisor

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

Estimating On-Shelf Availability of CPG Products at Nanostores in India

by

Stephanie Gabriela Gomez Prieto

and

Matthias Eder

Submitted to the Program in Supply Chain Management
on May 10, 2024 in Partial Fulfillment of the
Requirements for the Degree of Master of Applied Science in Supply Chain Management

ABSTRACT

This project presents an innovative approach to estimating On-Shelf Availability (OSA) within nanostores, key components of retail channels in emerging markets like India. Utilizing sales data and field study findings in Mumbai, we developed and validated two distinct models: a probabilistic model and a Machine Learning model. The probabilistic model was constructed based on sponsors' available sales data and by the formulation of five key assumptions, which were afterwards assessed through both qualitative and quantitative components of a field study. This approach revealed that certain assumptions were not fully validated, thereby weakening the model. Moreover, the approach revealed new insights into the purchasing behavior of store owners which had not been previously considered, such as their tendency to buy from wholesalers instead of directly from the sponsor company. In contrast, the classifier Machine Learning model, notably Random Forest, yielded superior accuracy. This classifier model was trained on actual inventory data gathered during the field study and is capable of relying solely on features derived from the sponsor company's data without the need for qualitative assumptions. Our findings underscore the significance of robust modeling techniques based on relevant data to enhance OSA estimation for Consumer-Packaged Goods (CPG) companies operating in nanostore contexts. Our recommendations include the adoption of the Machine Learning model, emphasizing its scalability and robustness as well as the importance of data collection to extend OSA visibility beyond Mumbai or India to broader regions. This study offers valuable insights and actionable recommendations for tackling the lack of empirical data for OSA estimation within nanostores.

Capstone Advisor: Dr. Inma Borrella
Title: Research Scientist, Academic Lead MITx MicroMasters Program in Supply Chain Management

ACKNOWLEDGMENTS

Joint

We would like to extend a big thank you to Dr. Inma Borrella for her commitment and great insights which have been more than crucial for the development of this project. Inma, your support, time, and knowledge have been truly appreciated.

A special thank you to the sponsor company, particularly the team that was so committed with this project. Thank you for giving us your time, patience and sharing your expertise. It has been an enriching experience and we hope this project supports your growth and success.

Stephanie Gomez

I cannot express enough how thankful I am for the support of my family, especially my mom. She has always been the most empowering role model. Feeling her company throughout this journey has been such a relief.

My accomplishment in this program is also thanks to my closest friends, who always made me believe in myself. And to my partner, Arturo, who has been my solid rock and biggest cheerleader. Thank you, I do feel blessed.

Finally, thank you Matthias, you were the best capstone partner I could have asked for. Vielen Dank, lieber Freund!

Matthias Eder

Apart from Stephanie's and my joint acknowledgements, I want to thank my parents for supporting me throughout this intense academic journey. Vielen Dank.

I also want to thank my professional mentors, Robert and Alice. You played a larger role in all of this than you likely realize. Thank you & merci beaucoup.

And lastly, thank you, Stephanie. You were my better half, keeping me sane and grounded. Muchas gracias.

TABLE OF CONTENTS

1. INTRODUCTION

1.1. Motivation

Consumer Packaged Goods (CPG) companies produce and sell daily-use items on a big scale, reaching final customers through retailers. These companies often manage two different retail channels: traditional and modern. The modern channel consists of brick-and-mortar shops and E-commerce (Bisen et al., 2020), while the traditional channel is composed of little, family-owned stores.

Among these small stores, we can find nanostores, which, according to Ortega, Amador, Parada et al. (2022), are considered small neighborhood stores that provide proximity and informal credit to customers. In nanostores, customers cannot go behind the counter and have limited stock visibility. Therefore, if the customer is unable to see the product on-shelf the final purchase decision will rely on the information the shop owner shares: if he indeed has the product but it is not visible or if he has an alternative to what the customer needs.

Nanostores, in emerging countries, are the channel that sells more than half of the total volume of CPG products, performing a key role for these companies (Fransoo et al. 2017). India excels in nanostores' performance, being responsible for more than 90% of CPG sales in this country (Escamilla et al., 2021).

Our sponsoring company is a multinational CPG company with operations in India where they supply 2.4 million nanostores. Approximately 20% of our sponsor's sales in India comes through nanostores *(Director, distributor logistics and Operations (India), personal communication, October 10, 2023)*.

On-Shelf Availability (OSA) is a metric that measures the quantity of products available for the customer to purchase at a specific moment. This indicator is key because it helps companies to have visibility through their products on retailers, identify stock-out events and, most importantly, devise strategies to increase sales. Our sponsor's desire is to know this metric for the nanostores they work with. However, there is an information gap when trying to measure OSA at this level. The sponsor has the distributors replenishment orders data and manage their own deliveries information, but granular data about inventory (product availability) at nanostores is not easy to gather and there is no system that allows the company to collect this data *(Director, distributor logistics and Operations (India), personal communication, October 10, 2023).*

The goal of this project is to have better visibility over OSA in the nanostores working with the available limited data and developing a robust model from it. The expectation is that by assuring OSA of the company's products in each nanostore stockout events will decrease and product visibility will improve.

Additionally, this project can yield significant potential sales growth for the company since it can be a pilot to extend to the AMA region (Asia, Middle East and Africa) where the company has a nanostore coverage of 4.9 million from a universe of 18 million.

1.2. Problem Statement & Research Questions

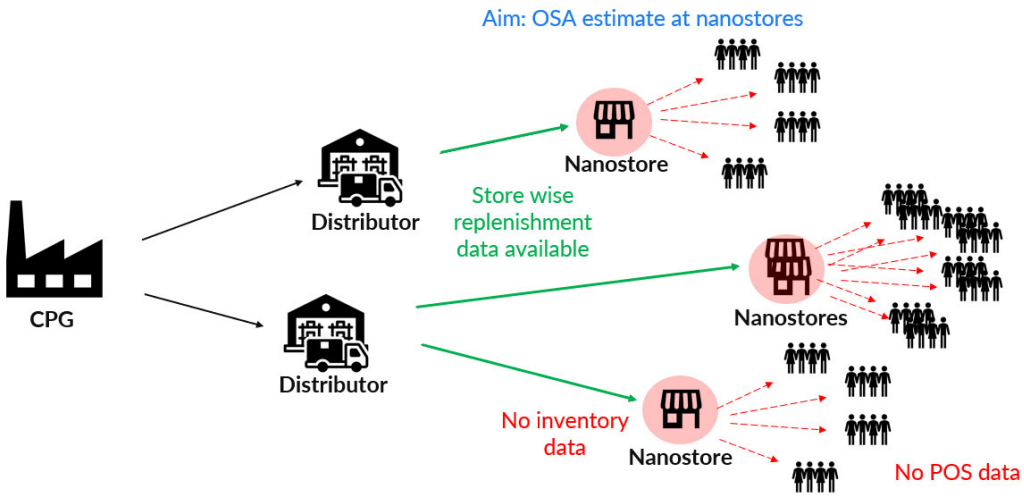
Our sponsoring company's goal is to drive increasing sales through increased On-Shelf Availability (OSA) in nanostores in India. Our project supported this goal by providing them with two different models to estimate OSA within the nanostores they supply to.

During this project's development, the company measured OSA for its modern retail channels but had no visibility into OSA at a nanostore level. As mentioned before, the sponsoring company serves approximately 2.4 million nanostores in India (*Director, distributor logistics and Operations (India), personal communication, October 10, 2023*) and due to the small size of these stores, inventory and point-of-sales data are not available (see Figure 1). Therefore, the sponsoring company was seeking to create a model to estimate OSA for key products our sponsor expects the nanostores to hold.

Nanostores are supplied through a net of distributors, each serving a certain area within India. The distributors send their salespersons to each nanostore, and during the visit the salesperson takes orders from the store owner and shares promotional information. After a store owner places an order through the salesperson, the order is fulfilled within the next day. Due to the small business volume of each individual store and the quantity of stores, the stores are typically only visited twice per month. The sponsoring company desired to communicate information about possible stock-outs at nanostores for the salesperson to drive targeted conversations during their regular check-ins with these stores.

Figure 1

Supply Flow to Nanostores



Note. This graph illustrates the flow of products from the sponsor company to the final consumer. While the information flow is continuous up to the distributor level, data from the nanostores regarding the inventory levels and purchases by the final consumers is unavailable.

In this context, the research questions to be answered include:

1. How can OSA at nanostores be estimated with no inventory or POS data available?
2. How can the proposed model to estimate OSA be validated?
3. What are key recommendations to increase visibility over OSA in nanostores that emerge from this project?

1.3. Scope: Project Goals & Expected Outcomes

The project's overall goal is to provide our sponsor with a model to estimate OSA with the limited available data. This model should enable the salespersons to target inventory gaps at the nanostores through which additional revenue is expected. The company had designated the area of Mumbai to act as a representative sample for broader India and our project is therefore focused on nanostores located there. Further, we also focused on the highest sales volume SKUs for the area in focus. In that context, the deliverables to the sponsor company include:

1. Two models capable of estimating OSA for key products at the store level: a probabilistic model and a Machine Learning model.
2. Validation process to assess the performance of the OSA models and analyze their results.
3. Set of recommendations, limitations, and conclusions aimed at enhancing OSA visibility at the nanostore level.

2. STATE OF THE ART

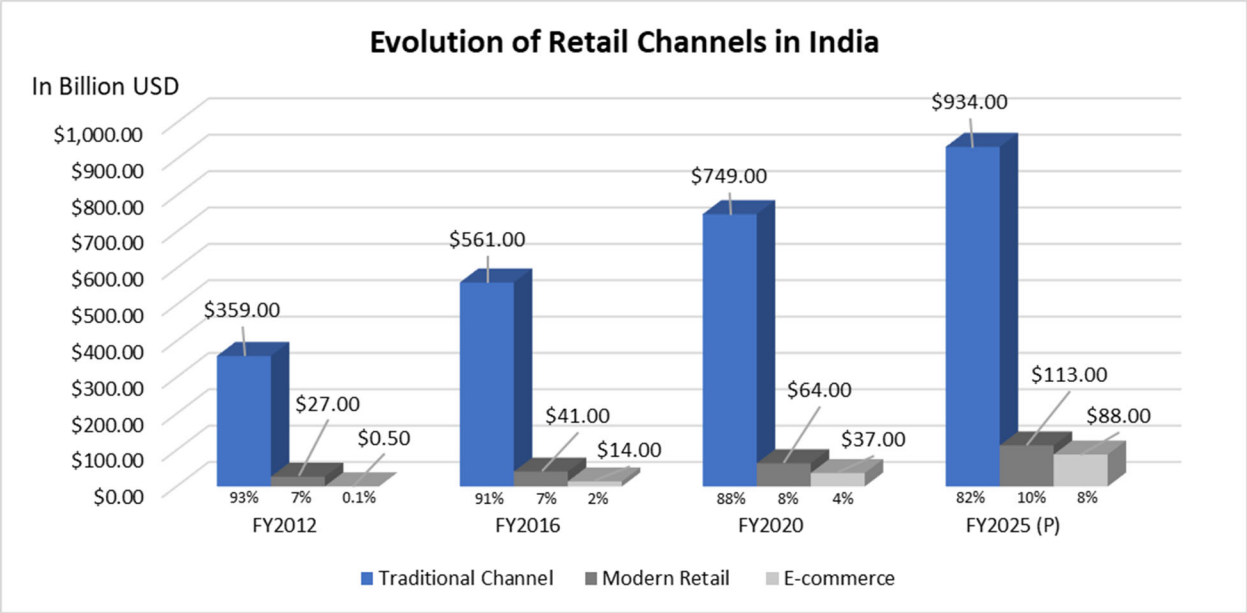
In this section we cover the literature addressing the objective of our capstone project of estimating On-Shelf-Availability with the limited existing data. The research focused on understanding Indian nanostores, inventory management in nanostores and measuring on-shelf-availability with imperfect visibility or data.

2.1. Understanding Indian Nanostores

According to Blanco and Fransoo (2013) there are two different retail channels: traditional and modern. The modern channel exists in developed and emerging markets and is made up of corporate retail, which is organized, large-scale and has strong negotiation power with CPG companies. In contrast to this modern channel, we find the traditional one. This channel is composed of little family-owned stores, of different sizes; when they are as small as less than 15 square meters they are known as nanostores. In India, these nanostores dominate the retail channel and are expected to nearly triple their income by 2025 (Bisen et al., 2020). After analyzing the different retail channels, they conclude that traditional channels are in control of the country retail's share for much more than 80%, as shown in Figure 2.

Figure 2

Evolution of Retail Channels in India



Note. Evolution of three different retail channels, traditional, modern and E-commerce, since 2012 and projected to 2025. Adapted from *Impact of Indian Retail on Employment & Taxation* by Bisen et al., 2020 (https://www.technopak.com/wp-content/uploads/2021/08/Retail-Impact-Assessment-2.pdf). In the public domain.

In India, each of these small non standardized stores serves a few hundred customers within a small neighborhood (Blanco and Fransoo, 2013). Nanostores are hindered by a lack of technology, evident in the absence of Point of Sale (POS) machines and inventory systems, and operational processes (Balkrishan et al., 2019), and have cash-space constraints. They provide informal credit to customers, winning their loyalty (Mora et al., 2021). Indian customers prefer to shop for their daily needs at these small stores rather than going to modern retail stores (Khare, 2013).

Nanostores are small stores where customers cannot go inside of them, and thus the visibility of products is limited *(Director, distributor logistics and Operations (India), personal communication, October 10, 2023)*.

2.2. Inventory management in Nanostores

According to Zhang and Rajaram (2017) nanostores' inventory management is quite challenging since their main constraint is lack of space. Additionally, to this space constraint, there is also an important cash constraint. This space-cash constraint leads nanostores to sell limited types of products and not many SKUs (Escamilla et al., 2021). Das (2018) explains store space is divided into the front room and the backroom. The front room is where customers ask for products or can see the store's shelves, and the backroom is used as storage for products that cannot be on shelves due to lack of space. This excess of inventory that cannot be allocated on the shelves is called by Eroglu et al. (2013) "overflow inventory". Eroglu et al. (2013) also mentions the value and scarcity of shelf space and how a replacement order, after its arrival, might go to the overflow inventory as it is a possibility that the shelves are full. This type of inventory management has its drawbacks: SKUs can get lost or forgotten in the backroom and the shop owner must, in some way, be aware of inventory on shelves and inventory in the backroom.

When retailers find themselves limited by space, they have to make assortment and inventory decisions, since they are not able to hold everything and need to program replenishments to utilize space optimally (Zhan and Rajaram, 2017). Thus, product allocation has two strategies: each product with its own space or space sharing with many products. The first has an easier to manage and lower cost replenishment cycle by product; the second leverages space but can be more expensive (Zhan and Rajaram, 2017). This is reaffirmed by Guo et al. (2016), who propose that when sharing a storage area with large quantities of SKUs, it is recommended a classification storage policy based on item turnover. This practice can significantly improve space utilization, unlike when each item has a specific storage section. Zhan and Rajaram (2017) show that space sharing has better performance than space dedicated to small storage spaces.

How nanostores manage inventory is key for the sponsor company. This information, that nowadays is unfeasible to gather for them (Director, distributor logistics and Operations (India), personal communication, October 10, 2023). The nanostores this project is focused on are served by distributors. The sponsoring company does not have inventory information directly from these stores and considering the size of the stores (and their little or no capacity for storage) this would mean they do not have On Shelf Availability data. The data the company owns is based on distributors' replenishment orders for each nanostore (Director, distributor logistics and Operations (India), personal communication, October 10, 2023 and Senior Trade Marketing Manager (India), personal communication, October 27, 2023). In

addition, it is important to mention the sponsoring company does not have direct contact with the nanostore owners.

Going further, regarding the difficulty of gathering inventory information directly from the source, there is doubt whether nanostore owners are willing to feed information through the use of technological developments. Eustis and Sonnenberg (2023) did interesting research about Mexican nanostores, discovering that more than half of the universe studied owned a smartphone. As an inventory system, at least half kept using paper records and as few as 20% used some type of software. In addition, two-thirds of the nanostores had an internet connection. These findings, though from a different geographical area than the one we address in our capstone, show opportunity areas for inventory management through technology and maybe future possibilities of collecting reliable data directly from shop owners.

2.3. Measuring On-Shelf-Availability with imperfect visibility / data

In the realm of inventory management within the retail sector, the challenge of navigating imperfect visibility has garnered significant attention among scholars. Imperfect visibility encompasses various complexities, such as censored demand and inaccurate inventory records. One comprehensive exploration of this domain comes from Chen and Mersereau (2015), whose literature review delves into strategies for analytically addressing these issues.

Most of the classical inventory management literature assumes that inventory values are known, though there is a significant amount of evidence that proves discrepancies between recorded inventory and actual physical available inventory (Chen and Mersereau, 2015). According to Atali et al. (2009), these inaccuracies can have three sources:

a) Misplacements where inventory is not available to a consumer but physically present in the store
b) Shrinkage where inventory has been stolen, damaged or lost
c) Transaction errors where through counting or scanning errors, physical inventory is different from recorded inventory

The literature typically assumes an “error process” (Chen and Mersereau, 2015), in which “an error random variable […] contributes to the discrepancy between available and recorded inventory each period.” In our project, this error variable would be taken to the extreme, as we do not have any information about actual inventory levels. DeHoratius et al. (2008) have suggested three ways in which

retailers can respond to inventory inaccuracies: prevention, correction and integration of inaccuracies. Our sponsor has decided to "integrate" the existing inventory inaccuracy, by exploring the option of creating a system that can account for the presence of inaccurate or missing information (DeHoratius et al., 2008).

From an integration perspective, we have identified some models that seemed promising for the context of our sponsor. Montoya and Gonzalez (2019) as well as Steeneck et al. (2022) have used a (partially) hidden Markov Chain Model to assume an out-of-stock state at an SKU level. In both papers, an OOS (Out Of Stock) state is inferred by using POS data. Montoya and Gonzalez (2019) define three unobserved demand states for their Markov Chain Model. The first one is where the demand is capped by an OOS state and no sales are recorded. The second and third states are defined as either low and high demand (sales) not limited by an OOS event. Being able to incorporate high demand states is particularly of relevance, as e.g. sales promotions can be modeled into the demand transition matrix. The probabilities of each demand state can then be estimated based on POS data (Montoya and Gonzalez, 2019).

Another model that we could see as applicable to the project is the newsvendor approach. In the classical newsvendor problem, the decision maker (in our case the nanostore owner), can only make one ordering decision prior to each sales cycle, after which the excess inventory becomes obsolete. The decision maker needs to weigh the costs of losing out on potential sales, with the cost of excess inventory (Caplice, 2016). One variant of this problem we deem relevant for our sponsor is presented by Jain et al. (2014). For their study on the effectiveness of using the timing of sales in managing censored demand, they test their model with "*checkpoint models*", where inventory levels are only observable at fixed times, not continuously. Jain et al. (2014) find that even with these checkpoint models, using timing of sales reduces the impact of information loss compared to models where demand is fully known. While we cannot directly measure POS or inventory positioning at nanostores, we can make assumptions on past sales and stocking levels due to the store owners' behavior at each bi-weekly checkpoint they have with our sponsor. While the products of our sponsor are not perishable after each sales cycle, the storage space and restricted cash availability create the cost of excess for the nanostore owner. We therefore assume that owners will only order additional units when they are either stocked out or do not have sufficient inventory to last them through the following sales cycle. An order therefore acts as a reset of inventory and the new (estimated) inventory equals the order quantity. Estimating OSA at any given point then depends on the expected demand for each product the nanostore carries. In the traditional newsvendor literature demand is assumed to be stochastic and fully known (Caplice, 2016). Typical demand

distributions used are Poisson or normal distribution (Jain et al., 2014) though in the case of our sponsor we are dealing with censored demand. Censored demand occurs, when the true demand distribution is not fully known, as demand is limited by available inventory. While considering other options besides Poisson and Normal we could explore Uniform and Triangle distributions. According to Caplice and Ponce (2021), after exploring the uniform distribution concept, it is evident it is not suitable for this project since it assumes equal likelihood for every possible value, which seems improbable in this context. The Poisson distribution is suitable for scenarios with all positive low values, which could be an option in our case depending on the sales quantities for the nanostores. The Normal distribution is recommended for large numbers and with a small coefficient of variance, which seems implausible for the expected low quantities of purchases of the nanostores, as well as the limited historical data for an accurate estimation. The triangle distribution appears to be the most promising. It aligns well with scenarios where the data distribution is unknown, requiring only the specification of a maximum value, a minimum value, and a mode. Interestingly, when dealing with anecdotal memories or non-reliable data, the suggested distribution is Triangle (Caplice & Ponce 2021).

When exploring additional alternatives we came across Machine Learning models. Following Shukla and Madhusudanan (2022), nowadays Machine Learning is a modern technology that can give companies the capacity to predict the future and manage inventory accordingly. A stock-out predictor model can be developed by employing supervised machine learning classifiers and dividing the data into training and testing datasets. Finally, after validating the model on the testing dataset it has been determined that the best classifiers are boosting algorithms, recommending XG boost, Ada boost and random forest classifiers as the primary model.

The Random Forest classifier has proven its effectiveness in a real-world scenario for a Latin American retail packaged foods manufacturing company, as highlighted by Andaur et al. (2021). Precisely on trying to solve the Out-of-Stock events situation that many retailers and manufacturers face (the inverse of On-Shelf Availability). The researchers utilized manufacturer's transactional data and data from physical audits and moved one step forward by stacking an ensemble of six classifiers. The classifiers chosen were: Random Forest, radial Support Vector Machines, Naive Bayes, Decision Tree, Logistics Regression, and Neural Networks. The ensemble classifier outperformed the individual Random Forest algorithm highlighting its predictive potential in this field.

In the above section, we have covered the crucial role that nanostores play in the retail market in general and more specifically in India. Further, we have reviewed the literature of the nanostores that typically manage their inventory given the space and cash constraints. Finally, we have then reviewed the literature on approaches for models on how to estimate OSA given the incomplete data availability our sponsor is facing.

3. DATA AND METHODOLOGY

This chapter introduces the data shared by the sponsored company, describes these datasets, and explains the techniques we considered to build two different models to estimate OSA.

As presented in the State-of-the-Art chapter, there is indeed literature on nanostores and a separate body of literature on estimating On-Shelf Availability metrics from Out Of Stock metrics based on POS data. But there is no research on how to measure On-Shelf Availability with no inventory or POS data. This chapter is divided into six subsections: data collection, data scope, data analysis, model 1 creation, model 2 creation and models' validation.

3.1. Data Collection

The data provided by the CPG company is divided into 4 types of datasets: coverage all India data, project sales reports (PSR) data, sales app outputs data and field study data.

a) **Coverage all India data** contains information about the approximately 2.4 million stores from different channels the sponsor company serves throughout India. Each row represents an individual store with a unique store code and includes information about the store's location, such as the city and geographical data.

b) **Project Sales Reports (PSR) data** includes sales or return transactions per SKU from the distributor to the different stores served every month (one file per month from January 2022 to January 2024). Each row represents a sale or return per SKU, with details about transaction date, store code, channel, product information, sales value and quantities sold.

c) **Sales App Outputs data** contains reports from August 2023 to November 2023. The sales app portrays the product basket the company expects nanostores to hold. The sales app is an algorithm that recommends which products the distributor should sell to each store based on sales patterns at proximity stores, promotions and customer behavior. This algorithm was built by our sponsor and is

leveraged by sellers on their twice-a-month visit to nanostores. Within the reports are the product information and the reason why this product is recommended (e.g. product is in trend, nearby stores have purchased it, the product is similar to previously bought products). Since the sales app algorithm has been introduced at the sponsor company only in August 2023, earlier data is not available.

d) **Field study data:** From February 15 to April 10, 2024, the sponsor company conducted a field study covering 40 nanostores in the Mumbai region, focusing on the top 20 selling SKUs, which collectively account for 52% of sales volume in Mumbai. The study was co-designed by the capstone authors in collaboration with the sponsor company. The primary objective of the field study was to improve the data conditions for determining and modeling OSA values by providing firsthand information on previously unavailable data. Additionally, it served to validate key assumptions made by the capstone team during the model development process. To achieve these objectives, we designed both qualitative and quantitative surveys to be administered by pollsters and responded to by store owners. Each nanostore was visited approximately every 4 days, sometimes 3, throughout the study period, with the OSA of the top 20 selling SKUs being recorded during each visit. The questionnaire for the qualitative survey can be found in Appendix A.

3.2. Data Scope

The sponsor company decided to focus the project's scope on the Mumbai region, which includes 80,000 nanostores. Python was chosen as the tool for data exploration and manipulation for this project.

As a first step, we narrowed down the “Coverage all India data” to Mumbai data only. We then merged this dataset with PSR Reports or sales data. In agreement with the sponsoring company, we exclusively focused on four nanostores sales channels, hereafter referred to as Channel 1, Channel 2, Channel 3 and Channel 4. These channels are categorized by size and are internally used by our sponsor to group nanostores.

Upon further review of the provided data, we determined that the “Sales App Outputs data” was not particularly useful for our project's purposes. Our focus was on nanostores, and it is important to note that these establishments typically make very small purchases. Consequently, the Sales App Outputs data mainly consisted of forecasts of one unit every time for every SKU, which limited our ability to understand purchasing behaviors. However, the PSR data contained richer information, encompassing every

transaction from every store over 25 consecutive months, making it the ideal starting point for identifying key SKUs for this project.

Additionally, it is worth noting the granularity of data available in the PSR dataset for each product. We were able to access data categorized by product (SKU), Brand Name, and Category. Following discussions with the sponsoring company, it was determined that the SKU level would be the focal point of our analysis for this project.

3.3. Data Analysis

To gain insight into the data and the current situation of OSA (On-Shelf Availability) at nanostores in India supplied by our sponsoring company, we conducted a Pareto analysis to identify the top-selling SKUs across each of the four sales channels under investigation. We selected the top 20 selling SKUs from each channel. We discovered that these top-selling SKUs exhibited similarities and recur across all four channels. While the SKU ranking may vary, certain SKUs appear repeatedly. For instance, the highest-selling SKU is first in three channels and ranks third in the fourth one, showcasing consistent popularity across multiple channels.

As a next step, we decided to find reference parameters for these nanostores. These parameters represent the typical characteristics and behaviors or average performance within the nanostores in this dataset. Afterwards, we examined potential patterns among neighboring nanostores. We chose to focus on high sales volume regions within Mumbai identified in the PSR data. Data was categorized by channel, region, and Top SKU. The central question was: are there discernible trends or seasonal variations within these clusters? Our analysis revealed no significant seasonality, which aligns with the nature of products sold in these nanostores—they are daily necessities, unaffected by seasonal fluctuations.

3.4. Model 1 Creation – Probabilistic model

After completing the data analysis, our next step involved developing and coding a model to estimate OSA. Python was our language of choice for this task, due to its scalability and extensive library support. Given the potential deployment of our model across multiple regions by our sponsor, scalability was crucial.

Considering the absence of direct demand data and point-of-sale (POS) information, we relied on PSR data (replenishment) as a proxy for the demand. To estimate demand, we analyzed historical order

patterns by channel for each key SKU and fitted them to various probability distributions. However, after conversations with the sponsoring company it was determined that they needed information for each specific store on each specific SKU. This was because they recognized that it was possible for nanostores to behave differently even within the same sales channel.

As outlined in the problem statement in Chapter 1, each store was visited twice per month by a seller. For instance, if a nanostore purchased 10 units of a particular SKU during the first monthly visit and then acquired 9 units of the same SKU during the second visit, we could infer that the initial 10 units were sold between the two visits. However, to understand the demand dynamics, considering we had 25 monthly files and the sponsor company typically examines demand at a monthly level, we chose to maintain a monthly demand occurrence with 25 checkpoints to assess demand.

Once more, in the absence of inventory and POS data, we were compelled to rely on several assumptions to construct a model for estimating OSA with the available information. At this stage, we had pinpointed five key assumptions to be tested and validated:

A. Store owner only order when their stock for a product is close to or equal to 0
B. Distributor can always fulfill store owner's demand (no backlog)
C. Lead Time is negligible (fulfillment of sponsor company within 1 day)
D. Store owner will wait for salesperson visit before replenishing
E. End customers and shop owners place significant value on purchasing sponsor company's brand (they do not purchase by product category)

Based on these assumptions we formulated the following propositions:

1. Assumption A (order at stock 0) + Assumption D (wait for salesperson) gives us the following:

$$\sum \boldsymbol{Q} = \sum \boldsymbol{D}$$

Replenishment Q is being used a proxy for Demand D

2. Assumption A (order at stock 0) + Assumption B (no backlogs) + Assumption C (no lead time) + Assumption D (wait for salesperson) gives us the following:

$$\boldsymbol{Q^*_{(t)} = I_{(t)}}$$

Inventory I is known at time (t) of Replenishment Q

The model formulated from these propositions is probabilistic in nature, enabling us to integrate uncertainty, which is a critical aspect given the absence of precise data points:

OSA likelihood = P [Demand Qty < $Q^*_{(t)}$]

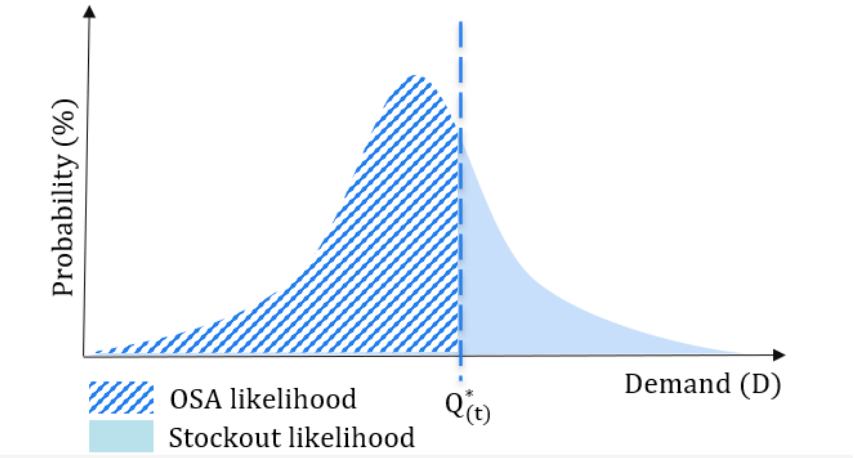
$Q^*_{(t)}$ = Last known inventory position $I_{(t)}$

Demand Qty = Probabilistic demand since time$_{(t)}$

This model is designed to provide estimations of OSA to some extent, leveraging the provided parameters. Even with limited data, the model can predict an OSA likelihood of any SKU at any given time of revision. See Figure 3 for a graphical representation.

Figure 3

OSA Likelihood Illustrated



Note. The graph illustrates the CDF of the probability of the OSA likelihood and its contrary, the stockout likelihood, based on the last known inventory position and the shaped demand.

In shaping the demand, we explored two distributions suitable for modeling it. The Poisson distribution initially seemed fitting due to the typically small purchase quantities at each store (typically less than 5 units). However, during the validation phase, we found that the Poisson distribution was causing too many false positives and therefore we did not use it further. Considering this and the need for a generalizable model coupled with the constraints posed by small quantities of data, we opted for the Triangle distribution. This choice was supported by the limited monthly data available (25 months of data and 2 replenishment orders each month for each store).

Our approach involved first identifying combinations of Store and SKU, as requested by the sponsor company for information at a store and SKU level. From each combination, we derived the parameters required for the Triangle distribution: Minimum ordered quantity, maximum ordered quantity, and most common ordered quantity (mode). Next, we determined the last ordered quantity for each store and SKU combination based on the available dataset, which became our last known inventory position.

Finally, using the Triangle distribution parameters and historical data, we estimated the shape of the demand. By comparing this estimation to the last known inventory position, we calculated the probability of the demand exceeding our inventory position. If the likelihood of the demand exceeding our inventory position is low, then the likelihood of On-Shelf Availability (OSA) is higher than the likelihood of a stockout event, which is the inverse. For instance, if the likelihood of On-Shelf Availability (OSA) is 25%, then the likelihood of demand greater than the last inventory position (resulting in a stockout event) is 75%.

3.5. Model 2 Creation – Machine Learning Model

Model 2 was created after obtaining the data from the field study. With this new data, the possibility of implementing a Machine Learning model emerged. We had access to validation data that enabled us to introduce a binary classification machine learning model. This type of model predicts the probability of an event occurring based on features provided to the model, in our case predicting whether or not a SKU is available at a store. We explored three different algorithms: Logistic Regression, XGBoost Classifier and Random Forest Classifier.

To start building this model, we began by crafting features specifically designed to offer vital insights about the SKUs. From the PSR data, we extracted the following features:

- Sales channel
- SKU
- Brand name
- Category name
- Rupees per unit (average): the average sales price over the 25 months of data
- Days since last purchase (SKU – store combination level)
- Last purchased quantity
- Days since last purchase (Store overall)
- Average daily purchases from the last 30, 60, 90, 120 and 180 days

Additionally, the inventory status feature was chosen from the inventory data collected during the field study, indicating whether the product was present each time inventory was taken.

Another consideration was the possibility of different city districts exhibiting distinct behaviors. To explore this, we utilized the latitude and longitude of each store visited during the field study and researched demographic features. We used Mumbai census data from 2011 – although dated, still potentially useful – to derive statistics from the district a store is located in by using the PointInPolygon technique. Unfortunately, 2011 is the most recent census data available, as a planned 2021 census was canceled due to the COVID-19 pandemic.

The selected features, chosen for their potential relevance to the model, included:

- Persons per household
- Male percentage
- Literacy rate
- Male worker percentage
- Worker percentage
- Population under 6

By merging all relevant data, we created the feature dataset. Each feature was assigned a column, and its values, if strings, will be then treated as columns. Consequently, we omitted the Store Code and ward name at this stage.

For the train-test split, we allocated the standard test size of 0.20 and specified a random state. We then identified categorical and numerical columns and constructed the preprocessor for our pipelines. Categorical columns were one-hot encoded, while numerical ones were scaled using Min Max Scaler. This preprocessor was applied to three separate pipelines, each corresponding to a different classifier: Logistic Regression, XGB Classifier, and Random Forest Classifier.

We defined two scenarios for the model: one incorporating demographic features and another without them, relying solely on PSR data. The latter option was designed to be simpler for the sponsor company to replicate.

3.6. Models' Validation

In this section we describe how we determined the validity of our models by testing them against the real OSA and stockout values taken in the field study.

3.6.1. Model 1 Validation – Probabilistic model

To be able to validate our key assumptions, we designed a questionnaire for the qualitative part of the field study. Store owners were asked these questions by the pollster during the first visit in the initial phase of the field study, and answers were recorded. The questions had the objective of revealing the store owners' behavior, e.g. whether they actually reached a zero level before a replenishment order, whether they maintained stock of the sponsors' products or they gave importance to the sponsors' brand. For a detailed list of questions and the corresponding assumptions they were intended to validate, see Appendix A.

To identify the predictive accuracy of the probabilistic model, we validated the OSA estimations against the inventory information of the field study. As we were estimating whether a unit is on shelf, we focused only on whether inventory at a survey date was available or not. This created a binary flag (0 for not available, 1 for available) to test the model estimate against. Our intuitive assumption was that for low OSA model estimates, we should observe a lower OSA percentage within the field study data. To achieve the validation, we took 720 individual datapoints from the field study, each being a combination of date, SKU, and store, and compared those 720 datapoints to the model's OSA estimate for each datapoint.

3.6.2. Model 2 Validation – Machine Learning model

To validate Model 2, we conducted a series of tests. We analyzed a classification report and created a confusion matrix, crucial steps in evaluating how well the model performs and understanding its predictive abilities. Given the complexity of our approach, using three different classifiers for two distinct scenarios (with and without demographic data), we repeated these analyses six times.

Additionally, we recognized the importance of selecting the right features to interpret the model and improve its performance. To address this, we conducted a feature importance analysis specifically for the top classifier identified in the prior tests. By sorting and ranking the features, we gained insights into which ones had the greatest impact on the model's accuracy. This process was vital for comprehending the primary factors influencing On-Shelf Availability (OSA) in a situation with limited data.

4. RESULTS

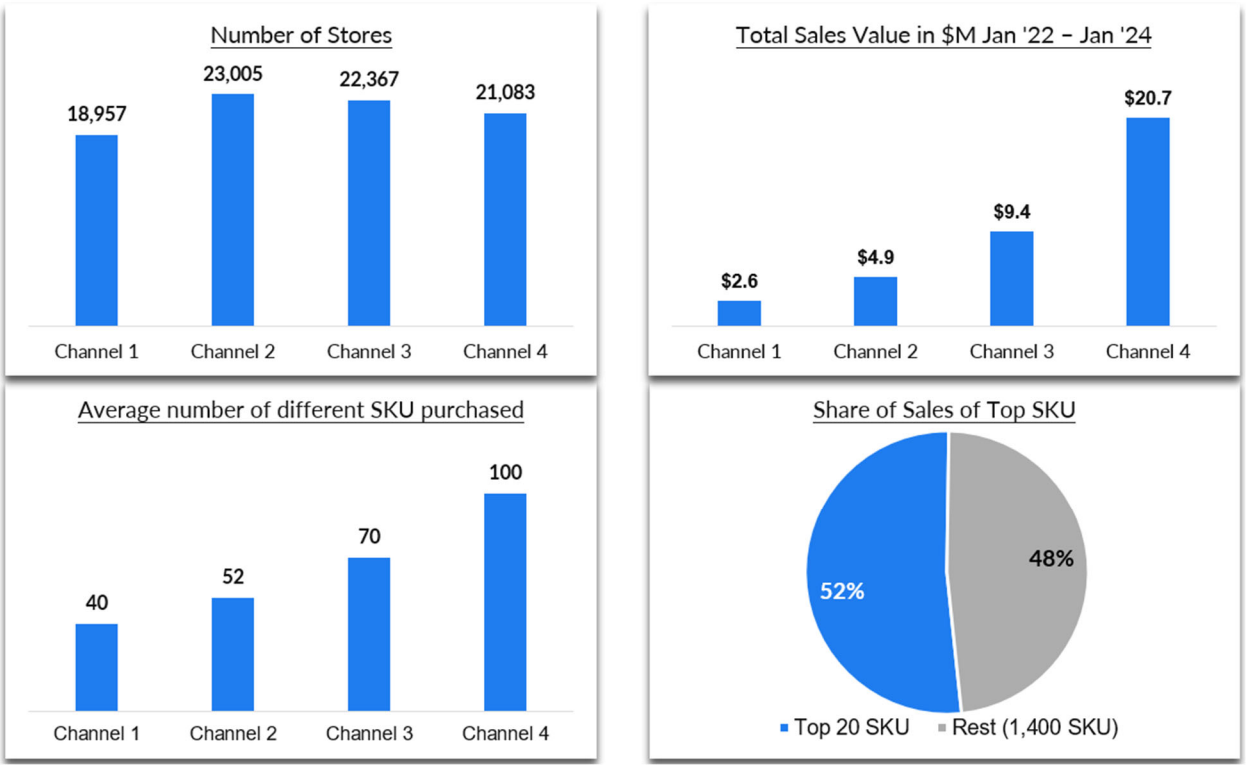
In this chapter we present the results of the models following the validation process explained in the previous chapter. It is divided into four subsections: insights from data analysis, Probabilistic model results, Machine Learning model Scenario 1 results and Machine Learning model Scenario 2 results.

4.1. Insights from Data Analysis

The initial findings of this project emerge from the data analysis. Upon examination, across the 25 months examined, Mumbai's nanostores have sales totaling $37.6 million. Their average monthly sales figures are relatively modest, with Channel 1 recording the lowest average sales at USD 5.00, while Channel 4 has the highest average sales at USD 39.00. While these figures may appear small at first glance, it is important to contextualize them within the larger landscape of nanostores across India, considering that this dataset represents just a sample of a region of the total. In addition, Purchasing Power Parity (PPP) highlights that the value of USD 1.00 differs based on where it's spent. For example, in India, one dollar can purchase more goods compared to some other locations. Figure 4 provides more detailed information on the exploratory data analysis of Mumbai.

Figure 4

Exploratory Data Analysis of Mumbai



Note. The graphs present an exploratory analysis of Mumbai, detailing the quantity of stores by channel, sales volume in Million USD, average number of SKUs purchased by channel over the dataset's time period and the sales share of the Top 20 SKUs.

4.2. Model 1 Results – Probabilistic model

As an initial step, we validated our key assumptions with the qualitative aspect of the field study which as explained was a self-made questionnaire to store owners. The outcomes are illustrated in Figure 5.

Figure 5

Key Assumption Validation with Qualitative Field Study

	Key Assumption	(1) Confirmed from Questionnaire	(1) Percentage of questioned nanostores confirming assumption	(2) Confirmed from Inventory Data	(2) Percentage of SBF/Store combinations confirming assumption
A	Store owners only order when their stock a product is close to or equal to 0	Not confirmed	15%	Confirmed	93%
B	Distributor can always fulfill store owner's demand (no backlog)	Confirmed	100%	N/A	-
C	Lead Time is negligible (fulfillment from sponsor company within 1 day)	Confirmed	100%	N/A	-
D	Store owner will wait for Salesperson visit before replenishing	Partially Confirmed	69%	Not confirmed	39%
E	End customers and shop owners place significant value on purchasing sponsor company's brand (they do not purchase by product category)	Confirmed	88%	N/A	-

Three out of the five key assumptions tested were confirmed. However, assumption A, namely, 'store owners only order when their stock for a product is close to or equal to 0,' was not confirmed solely with the qualitative field study part. After discussions with the sponsor company and our advisor, we recognized that store owners might perceive the field study not only as a resource to gather information and understand their behaviors but also as an evaluation of their practices, potentially biasing their responses. There might be a tendency to always indicate having stock of the sponsor's products. The sponsor company's team and existing literature suggest that nanostores typically do not maintain significant stock levels due to storage constraints and limited cash flow. With this idea in mind, we examined this key assumption with the quantitative data from the study. We analyzed 28 different store and SKU combinations across the four different sales channels. We found that in 93% of the cases, store owners made purchases only when inventory was at zero. This finding sufficiently validated assumption A.

Assumption D was also only partially confirmed after the qualitative part of the field study. Initially, 69% of store owners confirmed that they wait for the salesperson before making a replenishment. However, when tested against the quantitative field study data, only 39% of the time did store owners wait for the salesperson before replenishing, based on the analysis of the same 28 store and SKU combinations across the four different sales channels. It was observed that stock appeared without any formal purchase recorded in the PSR data, indicating that replenishment occurs through alternative means, such as visiting a wholesaler.

An intriguing aspect of our second proposition (Inventory I is known at time (t) of Replenishment Q) involves the inventory status at the time of replenishment. From the previously mentioned

combinations, we found that in 36% of cases, inventory did not increase despite a purchase being made. As previously noted, purchases typically involve very small quantities, with the most common quantity being 1. It is rational to believe that stores can immediately go out of stock again after purchasing, complicating the assessment of On-Shelf Availability (OSA).

Upon final revision of the field study results, it became evident that each SKU exhibits different behaviors, even within the same store. For instance, during the 2-month field study period, a fabric product was observed across 34 stores that purchased it. In 73.53% of cases, inventory increased without any formal purchase recorded. However, when analyzing a diaper product across the 23 stores that purchased it exclusively, this phenomenon occurred only 39.13% of the time. These findings have provided the sponsor company with valuable insights into the performance and inventory levels of its top 20 SKUs, as well as the behavior of store owners for each of them.

While we already knew that some of our key assumptions were only partially validated, we tested whether the model could overcome the inherent limitations. For this first model, we considered our sponsor's initial intention of providing salespeople with a targeted list of SKUs that are likely out of stock at a store and need to be reviewed with the store owner. Our aim was to determine whether the model could deliver a result such as: "*When the model's OSA estimate is below X%, there is a high probability of stockout*". We therefore started at our model estimate of 0% and calculated a true cumulative stockout % from 0% – 100%. The true OSA level for the whole dataset was 31% and we wished to identify the area of our estimates, where our model gives a significantly better than average prediction of a stockout event occurring. We found that models' OSA estimates <= 10% yield a relatively good result of a true OSA level of 21%, i.e. a 79% likelihood of a stockout (See Table 1). Out of the 720 datapoints, 151 fall into the category. Based on this finding, we can provide the sponsor with a tangible action list as "when the model estimate is <= 10%, there is a 79% likelihood of stockout."

Table 1

Model 1 Results

	Survey Results		Cummulative %	
OSA Estimate	Available	Stockout	Available	Stockout
0%	15	86	15%	85%
1%	1	2	15%	85%
2%	1	4	16%	84%
3%	2	4	17%	83%
4%	1	1	17%	83%
5%	2	8	17%	83%
6%	1	10	17%	83%
7%	1	6	17%	83%
8%	4	11	18%	83%
9%	7	9	20%	80%
10%	**6**	**10**	**21%**	**79%**
11%	3	7	22%	78%
12%	4	7	23%	77%
13%	6	6	24%	76%
14%	1	4	24%	76%
15%	5	8	25%	75%

4.3. Model 2 Results – Machine Learning Model – Scenario 1 with Demographic Features

In the evaluation of three algorithms–Logistic Regression, XGBoost Classifier and Random Forest Classifier–and using 1298 datapoints, the XGB Classifier and Random Forest Classifier demonstrated superior accuracy compared to Logistic Regression. With the XGB Classifier achieving 86% accuracy and the Random Forest Classifier achieving 87%, both models significantly outperformed Logistic Regression's 71% accuracy (See Table 2). The Random Forest Classifier exhibited a slight edge of 1% over the XGB Classifier, indicating its potential as the best choice for this dataset. However, it is crucial to consider the context of the inventory dataset, which was gathered from only 40 stores, raising concerns about overfitting. Given the risk associated with overfitting, particularly with models like XGB classifiers known for their susceptibility to it, the Random Forest Classifier emerges as the more robust option. Its ability to handle complex datasets and its lower tendency to overfit makes it better suited for this scenario. Also, since we are predicting OSA, there is no preference on the false positive or false negative results; the random forest classifier gives us a false negatives and false positives balanced ratio. Therefore,

considering accuracy, the risk of overfitting and the false negatives and false positives balanced ratio, the Random Forest Classifier is the recommended choice for this dataset.

Table 2

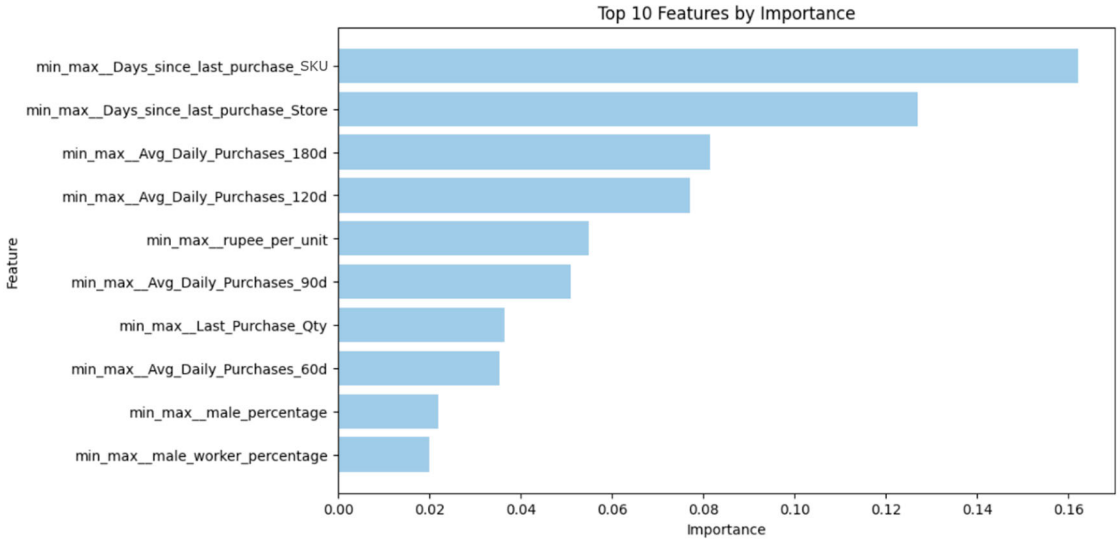
Model 2 Results - Scenario 1

Model 2 – Scenario 1	Accuracy	False positives	False negatives	Precision	Recall	F1-Score
Logistic Regression	71%	13%	16%	Not Available: 0.76 Available: 0.61	Not Available: 0.79 Available: 0.55	Not Available: 0.77 Available: 0.58
XGB Classifier	86%	7%	7%	Not Available: 0.88 Available: 0.81	Not Available: 0.89 Available: 0.80	Not Available: 0.89 Available: 0.80
Random Forest Classifier	87%	6%	7%	Not Available: 0.90 Available: 0.82	Not Available: 0.90 Available: 0.82	Not Available: 0.90 Available: 0.82

When examining the top 10 features ranked by their impact, it's evident that the most influential factor is the duration since the last purchase at the SKU-store combination level. Following this, the days elapsed since the last purchase at the store level also proves to be significant. Additionally, the third and fourth most influential factors are the average daily purchases over 180 and 120 days, respectively, with the average price per unit (average) coming in fifth (See Figure 6). These findings highlight the critical role of temporal factors, such as purchase history and frequency, in predicting On-Shelf Availability (OSA) within the model's framework.

Figure 6

Model 2 – Scenario 1 with Demographic Features Results – Top 10 Features by Importance



4.4. Model 2 Results – Machine Learning Model – Scenario 2 without Demographic Features

Since the demographic features didn't rank among the top five most important factors in the model, we concluded that they likely did not significantly impact the model's accuracy. As demographic data might be difficult to gather for other regions of India, we therefore decided to test the models without demographic features, solely relying on data from the sponsor company.

We evaluated the same classifiers as used with demographic data —Logistic Regression, XGB Classifier, and Random Forest Classifier—and found accuracies of 0.71, 0.82, and 0.85, respectively (See Table 3). Although there was a slight decrease in accuracy compared to the previous model, the random forest accuracy, which was the preferred one before, remained above 0.80 indicating a strong performance overall. This outcome is particularly advantageous for the sponsor company, as they can utilize the model exclusively with their own data without the need for additional demographic information from various wards. By relying solely on sponsor data, the model retains scalability and generalizability, making it a robust solution for the company's needs. Of the top 10 most important features, the first four retained their positions from Scenario 1 (See Figure 7). However, the fifth shifted to average daily purchases over 90 days.

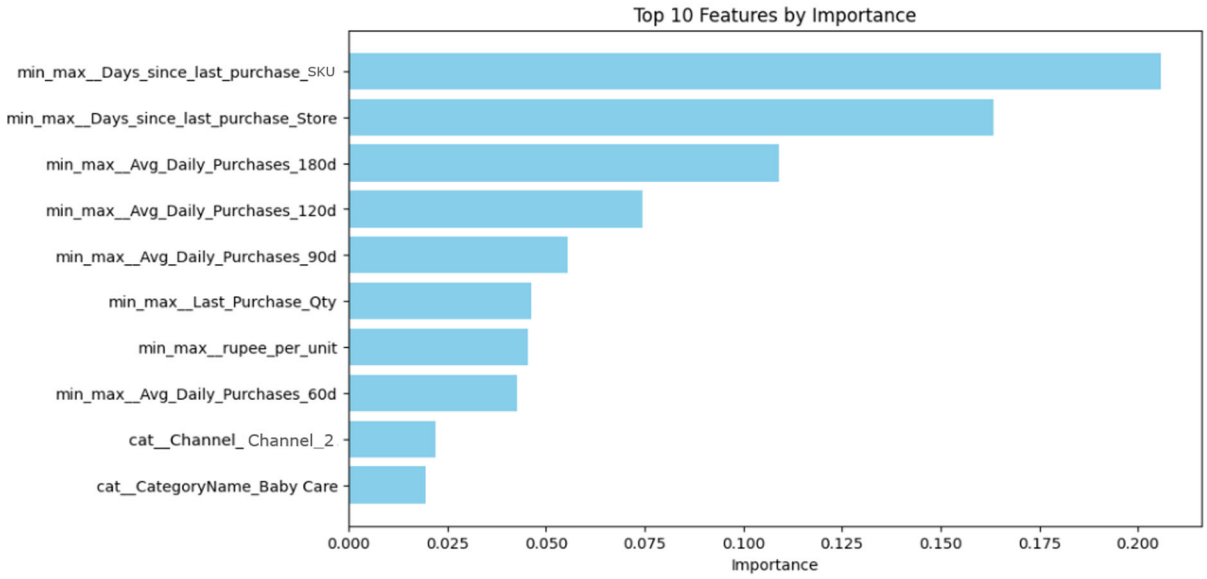
Table 3

Model 2 Results - Scenario 2

Model 2 – Scenario 1	Accuracy	False positives	False negatives	Precision	Recall	F1-Score
Logistic Regression	71%	12%	17%	Not Available: 0.75 Available: 0.62	Not Available: 0.81 Available: 0.54	Not Available: 0.78 Available: 0.58
XGB Classifier	82%	9%	9%	Not Available: 0.85 Available: 0.76	Not Available: 0.87 Available: 0.74	Not Available: 0.86 Available: 0.75
Random Forest Classifier	85%	8%	7%	Not Available: 0.89 Available: 0.79	Not Available: 0.88 Available: 0.81	Not Available: 0.88 Available: 0.80

Figure 7

Model 2 – Scenario 2 without Demographic Features Results – Top 10 Features by Importance



5. DISCUSSION

In this chapter we present the main insights derived and the limitations encountered when building and validating the models.

When validating the accuracy of the probabilistic model (Model 1), we found that the limited visibility of the data used for its creation constrained us in developing a robust model that could actually estimate OSA. The model was created by combining different assumptions; when any of these were violated, the model's ability to estimate OSA was severely weakened. The robustness of the model relied on all five assumptions to hold and to be holistic (i.e. we had incorporated every relevant driver of OSA). Since we saw two assumptions only partially confirmed and realized that the wholesaler played a larger than anticipated role on OSA for nanostores, Model 1 did not perform as expected.

Regarding the Machine Learning model (Model 2), it had a significant advantage over Model 1 as it was built with actual OSA data collected from the nanostores during the field study. This data allowed us to create a model that relied on labeled testing and training data instead of assumptions. We chose to use widely used machine learning algorithms for binary classification, as these seemed a natural application to predict the binary stockout vs. on-shelf status. When performing a classification algorithm, we had to integrate features that could impact the prediction of OSA. Outside of store-specific sales data, and product category data, we added demographic data in Scenario 1 with the belief it could shape the

behavior of the final consumer. However, when validating the feature importance, the demographic features did not appear to be highly significant in terms of OSA prediction. This led us to think of the possibility of working with only the sponsor's sales data, which would deliver a simpler and robust model to train. With this idea in mind, Scenario 2 was built. It had a similarly good performance as Scenario 1 (87% accuracy vs. 85% accuracy), having the advantage of relying only on the sponsor's data, which makes it more easily scalable for other regions.

These Machine Learning model scenarios (Model 2), while not relying on assumptions, rely on data collection. This creates an inherent limitation, if data cannot be collected regularly. Additionally, our specific survey data was collected from only 40 stores, leaving some concerns about an unrepresentative sampling.

6. RECOMMENDATIONS

Recognizing how the Machine Learning model yields significantly higher accuracy compared to the probabilistic model we strongly advocate for its adoption, with the suggestion to recalibrate it annually for optimal performance. As the sponsor company currently lacks methods to estimate OSA at nanostores we recommend that this model should be implemented with the existing data available. If its usage would be expanded to significantly different regions, we recommend to survey a larger number of stores from diverse locations for the data collection, aiming to have a representative sample of stores in the training data. This could be achieved by visiting more stores but less frequently (e.g., weekly instead of 2-3 times per week), allowing for a more comprehensive survey across a larger number of locations.

Another recommendation comes from the "regular" inventory and sales data collection systems from other retail channels to overcome challenges pertaining to data collection. To improve data collection, the investment in automated technologies like barcode scanners or mobile apps for nanostores to work as POS trackers would be ideal. These tools streamline the process, enabling better On-Shelf Availability (OSA) estimates. The integration of POS tracker would be advantageous for store owners', allowing them to track their own sales and therefore potentially increasing the willingness to collaborate with the sponsor company.

An option could be to combine the two recommendations mentioned above by installing a tracking system only for the sample nanostores that need to be surveyed to gather the necessary data for

a Machine Learning algorithm. This way the data gathered would be reliable, reducing the likelihood of human error in data collection and sampling a wider geographical area would be possible.

A final recommendation aims to address the challenge of OSA visibility loss when nanostore owners purchase from wholesalers instead of directly from the sponsor company. Collaboration with wholesalers seems essential. By initiating a visibility strategy, wholesalers can be incentivized to share information about OSA at nanostores when nanostore owners purchase products from them. This collaborative effort may involve implementing a system where wholesalers provide OSA data for nanostores to the sponsor company based on nanostore purchases. Incentives such as exclusive promotions or rewards programs can be offered to wholesalers who participate in this data-sharing initiative. Through this collaborative approach, the sponsor company can expand its existing nanostore data from distributors with additional insights from wholesalers, enhancing OSA visibility at nanostores.

7. CONCLUSION

This capstone project addressed the challenge of estimating On-Shelf Availability (OSA) within nanostores, recognizing the essential role of these stores in the sales ecosystem of Consumer-Packaged Goods (CPG) companies, particularly in emerging markets such as India. In the retail market, where traditional and modern channels coexist, nanostores emerge as crucial players. These small neighborhood shops, unlike larger retail outlets, offer close proximity to customers and often extend informal credit services. However, they pose distinct challenges, such as no visibility into inventory data at nanostores and the impossibility to track final consumers demand and understand purchase behaviors.

We faced the challenge of estimating OSA with no inventory or POS data available by developing two different models —a probabilistic model and a Machine Learning model. The probabilistic model (Model 1) relies on key assumptions that need to be true and holistic to ensure the model's robustness. The Machine Learning model (Model 2) relies on sample inventory information gathered through a field study to collect the necessary training dataset.

With the help of a field study, we validated the probabilistic model (Model 1) by testing the key assumptions, elaborated for the creation of the model, through a qualitative survey. We found that two key assumptions – "store owner only orders when the stock of a product is close to or equal to zero" and "store owner will wait for a salesperson visit before replenishment" – were not entirely confirmed. The

violation of the latter assumption brought to light that store owners frequently purchase from wholesalers and not as originally expected directly from the sponsor company's distribution network.

We validated the probabilistic models' OSA estimates, with the quantitative inventory data from the field study. As a result, the model proved not to be as strong as expected, showing useful results only for models' OSA estimates <= 10%.

The Machine Learning model (Model 2) was validated by a widely used train and test data split, performed on the collected inventory data. For this model we used three different classifiers and two different scenarios. The three classifiers were Logistic Regression, XGBoost and Random Forest. The two scenarios were built to determine the importance of the different features the dataset contained when predicting OSA. The first scenario contained both demographic and sales features, while the second scenario only used sales features. The results were promising: the Random Forest in both scenarios yielded an accuracy of 87% for Scenario 1 and 85% for Scenario 2. The model is robust and, since it has the capability of working only with data from the sponsor, is scalable. The project's geographical focus on the Mumbai region serves as a sample for broader implications. There is potential to extend its findings and methodologies to the broader Asia, Middle East, and Africa (AMA) region where the sponsor covers approximately 4.9 million nanostores.

Our key recommendation is to implement the Machine Learning model to estimate OSA levels with the current data available for India's regions. In the case of expansion, this implementation should acknowledge the caveat that data needs to be gathered from a larger sample of stores, across a wider area. The results of the probabilistic model were unsatisfactory, so installing a widespread tracking systems technology through all relevant nanostores seems to be unfeasible in the near term.

This capstone project represents a pioneering attempt to address the complexities of OSA estimation within nanostores, offering an insightful understanding of the challenges and opportunities inherent in this domain. We encourage future researchers to look for ways to improve this Machine Learning model through hyperparameter tuning and potentially incorporating additional features as well as neural networks.

REFERENCES

Andaur, J.M.R., Ruz, G., & Goycoolea, M. (2021). Predicting Out-of-Stock Using Machine Learning: An Application in a Retail Packaged Foods Manufacturing Company. *Electronics*, 10(22), 2787. https://doi.org/10.3390/electronics10222787

Atali, A., Lee, H., & Özer, Ö. (2009). If the Inventory Manager Knew: Value of Visibility and RFID under Imperfect Inventory Information. *Stanford University*. Available at SSRN: https://ssrn.com/abstract=1351606 or http://dx.doi.org/10.2139/ssrn.1351606

Balkrishan, S., Ashutosh, K., & Avinash, P. (2019). Competitive Strategies for Unorganised Retail Business: Understanding Structure, Operations, and Profitability of Small Mom and Pop Stores in India. *International Journal on Emerging Technologies*, 10(3), 253-259.

Bisen, A., Tiwari, M., Yadav, R., Kalia, P., Gupta, R., & Abrol, P. (2020). Impact of Indian Retail on Employment & Taxation. *Technopak*. https://www.technopak.com/wp-content/uploads/2021/08/Retail-Impact-Assessment-2.pdf

Blanco, E. E., & Fransoo, J. C. (2013). Reaching 50 million nanostores : retail distribution in emerging megacities. (BETA publicatie : working papers; Vol. 404). *Technische Universiteit Eindhoven.*

Caplice, C. (2016). SC1x: Supply Chain Fundamentals. Single Period Inventory Models. *MIT Center for Transportation & Logistics.* https://studio.edx.org/asset-v1:MITx+CTL.SC3x+2T2016+type@asset+block@SC1x_KeyConcepts_v6.pdf

Caplice, C., & Ponce, E. (2021) MITx MicroMasters Program in SCM - Key Concepts*. MIT Center for Transportation & Logistics.* https://ctl.mit.edu/sites/ctl.mit.edu/files/attachments/S2%20-%20Aug%204%20-%20W2%20-%20SCx_KeyConcepts.pdf

Chen L., & Mersereau, AJ. (2015). Analytics for operational visibility in the retail store: The cases of censored demand and inventory record inaccuracy. In: Agrawal, N., Smith, S. (eds) *Retail Supply Chain Management. International Series in Operations Research & Management Science*, vol 223. Springer, Boston, MA. https://doi.org/10.1007/978-1-4899-7562-1_5

Das, L. (2018). Backroom space allocation in retail stores [Doctoral Thesis, Massachusetts Institute of Technology]. *Dspace*. https://dspace.mit.edu/handle/1721.1/120662

DeHoratius, N., Mersereau, A., & Schrage, L. (2008). Retail inventory management when records are inaccurate. *Manufacturing Service Operations Management*, 10(2), 257–277.

Eustis, E. & Sonnenberg, H. (2023). Transforming Micro-Retailing in Emerging Markets [Capstone Project, Massachusetts Institute of Technology]. *Dspace*. https://dspace.mit.edu/handle/1721.1/152063

Eroglu, C., Williams, B., & Waller, M. (2013). The backroom effect in retail operations. *Production and Operations Management*, 22(4), 915-923

Escamilla, R., Fransoo, J.C., & Tang, C. (2021). Improving Agility, Adaptability, Alignment, Accessibility, and Affordability in Nanostore Supply Chains. *Production and Operations Management*, 30(3), 676-688.

Fransoo, J. C., Blanco, E., Mejia-Argueta, C. (2017). Reaching 50 Million Nanostores: Retail Distribution in Emerging Megacities. *CreateSpace Independent Publishing Platform*, Cambridge, MA.

Guo, X., Yu, Y., & De Koster, R. (2016). Impact of required storage space on storage policy performance in a unit-load warehouse*. International Journal of Production Research*, 54(8), 2405-2418.

Jain, A. Rudi, N., Wang, T. (2014). Demand Estimation and Ordering Under Censoring: Stock-Out Timing Is (Almost) All You Need. *OPERATIONS RESEARCH.* Vol. 63, No. 1. ISSN 1526-5463

Khare, A. (2013). Retail service quality in small retail sector: the Indian experience. *Facilities*, 31(5/6), 208-222. https://doi.org/10.1108/02632771311307089

Montoya, R., Gonzalez, C. (2019) A Hidden Markov Model to Detect On-Shelf Out-of-Stocks Using Point-of-Sale Data. *Manufacturing & Service Operations Management*, 21(4), 932-948. https://doi.org/10.1287/msom.2018.0732

Mora, C., Cárdenas, L., Velázquez, J., & Gámez K. (2021). The Coexistence of Nanostores within the Retail Landscape: A Spatial Statistical Study for Mexico City. *Sustainability*, 13(19). https://doi.org/10.3390/su131910615

Ortega, C., Amador, A., Parada, J., Zavala, D., & Alvarado, S. (2022). A meta-analysis of Nanostores: A 10-year assessment*. Exponential Technologies and Global Challenges: Moving toward a new culture of entrepreneurship and innovation for sustainable development*. https://laccei.org/LEIRD2022-VirtualEdition/full-papers/FP101.pdf

Shukla, S., & Madhusudanan, V. (2022). Stockout Prediction in Multi Echelon Supply Chain using Machine Learning Algorithms. *2nd Indian International Conference on Industrial Engineering and Operations Management Warangal, Telangana, India, August 16-18, 2022*. https://ieomsociety.org/proceedings/2022india/368.pdf

Steeneck, D., Eng-Larsson, F., & Jauffred, F. (2022) Estimating Lost Sales for Substitutable Products with Uncertain On-Shelf Availability. *Manufacturing & Service Operations Management*, 24(3):1578-1594. https://doi.org/10.1287/msom.2021.1015

Zhang, W., & Rajaram, K. (2017). Managing limited retail space for basic products: Space sharing vs. space dedication. *European Journal of Operational Research*. 263(3), 768-781.

APPENDIX

Appendix A - Questions from Field Study Questionnaire for Validating Key Assumptions

	Key Assumption	Questions to Store Owners from Field Study Questionnaire for Validating Key Assumptions
A	Store owners only order when their stock a product is close to or equal to 0	How many units of a sponsor's product do you have at the moment of reordering the same item? (This question came with a list of the Top 20 SKUs to be filled by product)
		Do you keep stock of sponsor's products for more than one replenishment period? This means you still have sponsor's products when the salesperson visits you
B	Distributor can always fulfill store owner's demand (no backlog)	Do you always receive everything that you have ordered?
C	Lead Time is negligible (fulfillment from sponsor company within 1 day)	How long does it take for you to receive the order?
D	Store owner will wait for Salesperson visit before replenishing	Within the same month, how often do you run out of stock of sponsor's products? Would you say it's: never, rarely, often, or always?
		What action do you normally take when you run out of stock of a sponsor's product within the same month?
E	End customers and shop owners place significant value on purchasing sponsor company's brand (they do not purchase by product category)	How do you, as the store owner, normally decide which products to buy? Is the brand more important than the product category?
		How do your customers normally decide which products to buy? Is the brand more important than the product category?