A Work Project, presented as part of the requirements for the Award of a Master's degree in

Business Analytics from the Nova School of Business and Economics.

FIELD LAB JERÓNIMO MARTINS – INTELLIGENT MANAGEMENT INFORMATION SYSTEMS FOR NOVA & GO STORE

Optimizing the Replenishment Process for Slow-moving Products

WANDER PEREIRA SOUZA

Work project carried out under the supervision of:

Professor Qiwei Han
Mr. Rui Tomás

25/01/2023

Acknowledgments: To Professor Qiwei Han, who expertly guided us through this Field Lab, for such insightful feedback and relentless supervision. To the JM Group, especially to Mr. Rui Tomás, for the challenging topic proposed, the provided data, and the utmost guidance through the development of this project. Finally, to my colleagues, Maria and Mariana, for their total dedication and support since the first day of this thesis. My deep gratitude goes to all of them, who, without, this dissertation would not have been possible

Abstract: Taking advantage of the latest technologies, the retail industry has drastically evolved in recent years. Launched by Jerónimo Martins, Nova & Go store is an example of a revolutionary supermarket. With the purpose of further improving the Lab store's daily operations, the slow-moving products stockouts were addressed in the present thesis. The replenishment process for slow-moving products was optimized, based on a demand forecast.

Keywords: Demand Planning, Internet of Things, Logistics, Management Information Systems, Retail 4.0

This work used infrastructure and resources funded by Fundação para a Ciência e a Tecnologia (UID/ECO/00124/2013, UID/ECO/00124/2019 and Social Sciences DataLab, Project 22209), POR Lisboa (LISBOA-01-0145-FEDER-007722 and Social Sciences DataLab, Project 22209) and POR Norte (Social Sciences DataLab, Project 22209).

INDEX

1. Introduction

In today's world, data is viewed, by most organizations, as a strategic asset *(Delloite, 2021).* In the words of Clive Humby, a British mathematician, it can even be described as "the new oil of the 21st century", due to its tremendous potential and value when realized through business applications. However, just like oil, data needs to be refined. Raw data is worthless, if valuable insights cannot be extracted from it *(McKinsey, 2018).* Rather, it's through the acquisition of information that companies are able to respond proactively and intentionally to market conditions. Accordingly, in 2011, Peter Sondergaard, a Gartner's board member, corrected Humby's famous quote by stating: "Information is the oil […] and analytics is the combustion engine. Today, analytics is the new oil."

Among a wide range of industries, Big Data has opened up new opportunities, and Retail is no exception. In fact, the concept of Retail 4.0 was introduced in 2010, with the intention of describing the sector transformation based on Industry 4.0 technologies, such as Internet of Things, Cloud Computing or Artificial Intelligence *(Har et al, 2022).* In fact, technology has enabled the retail sector to create a brand-new shopping experience *(Gazzola et al., 2022).* From self-service kiosks to e-commerce platforms or AR-based tools aimed to enhance customer experience, this new era is characterized by innovation.

An example of an innovative project aligned with this new trend is a cashier-less store created by Jerónimo Martins (JM), the retail leader in Portugal. In October 2019, JM launched Nova & Go store, at the Nova School of Business and Economics campus. Taking advantage of a group of customers predominantly made up of young students, unlike what happens in the regular stores, this project aims to test new technologies and new products. Being primarily visited by Gen Z customers, it is an ideal opportunity to experiment with innovative technologies and new business models. Apart from traditional groceries, the store offers various freshly prepared foods and other novel products, all offered with the students in mind.

Customers' experience in this Lab store starts with the download of the Pingo Doce & Go Nova app, available for both iOS and Android. After registering, a QR code should be displayed to enter the store. Once inside, purchases can be recorded in two ways: either through NFC technology or by pointing the cell phone's camera at the product code. Later, to be able to pay, customers can associate their bank card to the app or, alternatively, use the payment towers. Yet, even in the latter case, only credit or debit cards are accepted, not notes or coins.

During this process, data is generated at various points. So, in order to handle the large volume of data produced by the store, Management Information Systems (MIS) must be incorporated into its daily activities. Essentially, MIS collects and processes data from all sources within an organization to assist managers in their decision-making processes *(Berisha-Shaqiri, 2014)*. As a result, JM has access to a detailed database that helps it explore the habits of its customers and analyze critical aspects of its processes.

Based on the exceptional opportunities provided by the Lab store, JM has identified a topic to be developed throughout this thesis. Sales data will be used to calculate a forecast demand for slow-moving products, which will in turn be used to generate a picking list. Improvements in the replenishment process can lead to a reduction in stock-outs.

Lastly, it should be noted that Design Science Research (DSR) methodology was chosen to structure the present thesis, since it has proven to offer outstanding benefits when it comes to solving real-life problems within organizations *(Hevner et al., 2004).* In the appendix, it is possible to find the different steps of this approach (*Fig 1. in the appendix).*

2. Optimizing the Replenishment Process for Slow-Moving Products

2.1. Problem Statement and Motivation

Several unpredictable factors can influence the level of demand faced by a retail business, highlighting the importance of demand planning. As an example, Nova & Go store's demand can fluctuate throughout the week or even according to seasonality. Furthermore, it is also important to take into account external factors, such as weather conditions and university events. For example, barbecues and sunset parties are typical events that occur at the beginning of each semester.

In this sense, it is clear that developing an appropriate demand planning system and optimizing the logistics process is crucial to the success of any business, as it impacts not only the warehouse's performance, but also the entire supply chain. As a result, a more efficient and effective integrated flow can be achieved, as well as quicker responses to customer requests. Until now, there is no such thing as an optimal supply chain, regardless of all efforts to plan, control, and utilize tools to assist supply chain management. Nevertheless, it is mainly during restocking stores that the insatiable challenge of avoiding product shortages arises.

Nova & Go store is no exception to the relentless challenge of preventing out-of-stocks. Despite the efforts and rigor put into this store every day, stock-outs events are still frequent. This can be attributed directly to the perception of a low level of service by the customer. In fact, out-of-stock situations can adversely affect customer loyalty and create a negative image and reputation for a company. Therefore, Jerónimo Martins (JM) must play a proactive, attentive, committed, and determined role in this area.

Specifically, there is a primary concern regarding the Nova & Go store's logistical process. Despite the access to a wide variety of data recording product movements and sales, replenishment activities are not fully informed and data-driven. Rather, the replenishment process is largely determined by the instincts of workers who base their decisions on what

usually sells the best, as well as on what is displayed on the shelves. Consequently, products with a low turnover rate are often neglected. This results in frequent stock-outs for these slow-moving items and unnecessary trips to the warehouse, due to a miscalculation of the quantity required.

2.2. Objectives of a Solution

Essentially, this thesis aims to optimize replenishment activities, by reducing stock-outs and providing relevant insights to support informed decision-making. Accordingly, we believe that the work developed will contribute to the improvement of the replenishment process, making it more data-driven and efficient. As mentioned above, the in-store replenishment process still represents a blind spot at Nova & Go store, due to a lack of insights derived from the available data. Moreover, particular attention should be paid to the case of slow-movers, which are often overlooked when planning daily replenishments.

As stated by *Ross (2004),* the picking process is indeed one of the most intensive and costly in a warehouse, since it requires strenuous manual labor and a substantial amount of workers' time, having a significant impact on logistics costs. As discussed in the previous section, supply chain management involves a cumbersome process that becomes increasingly complex over time. Thus, in order to avoid out-of-stock situations, it is essential to develop an intelligent logistical plan. For instance, it is typical for employees to use picking lists in order to manage warehouses and inventory for specific products.

Furthermore, it is imperative to take into consideration that poor picking list performance can negatively impact customer service and increase operational costs for their warehouse and, therefore, for their entire supply chain. In addition, it is also important to note that we are aware of the challenges, problems, and constraints that will arise in our model, especially in regard to

slow-moving products. As mentioned by *Cavalieri et al. (2008)*, slow-moving articles tend to have low demand, making it difficult to define a pattern of demand.

Due to the store's revolutionary concept, JM can have access to a large amount of valuable data. Using these data, logistics processes can be improved. According to Rui Tomás (JM's technology director), the current solution they have in place is very inefficient, resulting in repeated trips to the warehouse to collect items that are, in reality, not out-of-stock. To address these problems, an algorithm will be developed to generate an optimized picking list, which will enable store managers to track stock levels and manage replenishment activities. As a result, we will attempt to answer the following research question throughout the thesis:

> ***Research Question:*** *How can Machine Learning be used to generate a picking list for slow-movers?*

2.3. Literature Review

2.3.1. Introduction

In this section, the main concepts, definitions, methodologies, and previous research results on this field are described, considering the scope of this thesis. Firstly, a brief approach to the retail industry and the supply chain concept is carried out, as well as an introduction to inventory management and the replenishment process. Furthermore, stock-outs, namely of slow-moving items, are highlighted as the main challenge to be addressed. Finally, the XGBoost Machine Learning model is presented as a solution to identify demand patterns based on historical data. Moreover, it is able to generate a demand forecast, while ensuring a robust and accurate performance.

2.3.2. Retailing

An entity that belongs to the retail industry is described as an individual or organization that buys products from a producer or distributor and, in turn, resells them to the final consumer *(Council of Supply Chain Management, 2013).* These organizations establish processes to ensure that each activity is performed effectively and efficiently, considering factors such as time and resources *(Graafmans et al., 2020).* Currently, we are living in a customer-driven economy, where the focus is on customer needs, convenience, and satisfaction *(Man et al., 2017).* In fact, with the advent of the Internet, customers have become more demanding, informed, and empowered when it comes to making decisions. Furthermore, our society is becoming increasingly interconnected. Therefore, consumers are no longer dependent on businesses, but are rather able to make their own decisions based on their judgement and available information *(Prabhalad & Ramaswamy, 2004).*

Simultaneously, advances in technology have enabled retailers to explore new business models and processes that are capable of simplifying and accelerating their operations *(Hopping, 2000).* Hence, technology dependence has become a reality in the global context, and retail industry is no exception *(Aalst & Dustdar, 2012).*

The Internet of Things (IoT) consists of the global network of interconnected objects, information and people, which are mainstreamed through networks, such as the Internet. It has been recognized as a key disruptive technology for supply chain management progress. In fact, the concept of IoT was introduced by Kevin Ashton in 1999, aiming at a solution that would facilitate monitoring the high volume of products in the supply chain. From production to distribution and store shelving, IoT provides real-time data at all stages and for all players in the supply chain *(Lee, 2021).* Therefore, it has become pivotal to have tools and systems capable of automatically and intelligently exploring and interpreting this data *(Fayyad, 1996).*

2.3.3. Supply Chain Management

With the increased market competitiveness, organizations are required to promptly and efficiently respond, making supply chain performance an increasingly key factor to organizations *(Beamon, 1998).* Accordingly, in order to achieve competitive advantage and differentiate themselves from their competitors, organizations must improve their logistical processes' response to consumer demands *(Monczka et al., 1998).*

Supply chain management is responsible for designing, implementing, and monitoring the storage of products, services and related knowledge from the point of production to consumption *(Ballou, 2004),* as well as all the processes present along this route. According to another definition, the objective of logistics processes is to plan, manage, and optimize an organization's processes to meet the customer's needs while achieving the desired service and quality *(Christopher, 2005).* Supply chain planning and control, which comprises inventory management, production, materials and distribution planning, from the procurement stage to the distribution of products on the shelves, is indispensable (*Manzini & Bindi, 2009*; *Ivanov et al., 2014*). In this context, the need for intelligent tools and information systems capable of supporting these processes is reinforced.

2.3.4. Stocks

Despite all the efforts made towards supply chain planning and the development of technologies such as RFID (Radio-Frequency Identification) *(Aastrup & Kotzab, 2010),* having a product out-of-stock, meaning not readily available, is a common issue *(Grewal & Levy, 2007).* In fact, globally, on average, about 8.3% of a store's products are not available on the intended shelves. This average value is larger for European stores, compared to other regions *(Corsten & Gruen, 2005).*

Stock-outs can be viewed from two perspectives: distribution inefficiency on the one hand, and consumer behavior-demand on the other hand *(Aastrup & Kotzab, 2010).* For instance, not only

does it directly result in lost sales *(Avlijaš et al., 2001),* but the absence of the desired product can lead to a feeling of consumer dissatisfaction, which can deteriorate brand loyalty *(Sanchez-Ruiz et al., 2018).* Hence, there are several possible behaviors that consumers can adopt when faced with a stock-out, including buying a substitute item, postponing the purchase, giving up the purchase, or looking for the item in another store *(Corsten & Gruen, 2003).* Furthermore, the causes of stock-outs can be due to factors before the goods arrive at the store (pre-store), or, alternatively, when they are already in the store (in-store) *(Ehrenthal & Stölzle, 2013)*

No matter whether pre-store disruptions are caused by suppliers or distributors, it can be attributed to non-compliance with dates or quantities. As for in-store disruptions, they are often caused by errors in the ordering of quantities by the stores, or by problems associated with the movement of products from the backroom to the shelves. In fact, the major causes of in-store disruptions are the misallocation of available space for each item on the shelves (store planogram), and the inadequate shelf replenishment routines *(Aastrup & Kotzab, 2010).*

2.3.5. Demand Patters & Predictions

In order to have an efficient supply chain, in-store inventory and shelf replenishment management, it is necessary to understand the existing demand patterns. In *Syntetos et al., 2005*'s view, demand can be categorized into four patterns: continuous/smooth (relatively consistent throughout time), erratic (constant in demand periods, but with high variability in size), intermittent (randomly and with several periods of zero demand), and irregular (intermittent and erratic demand, meaning demand sizes show great variability and appear randomly with several periods of no demand).

Accordingly, managing products in a warehouse should be done in accordance with their type: fast-movers or slow-movers. The demand for fast-movers tends to be stable, while the demand for slow-movers tends to be intermittent. This makes it difficult to predict the reorder point and

manage the inventory *(Nenes, et.al., 2010).* Food retailers and wholesalers operate in a fast-moving industry, but some of their products aren't as popular as others *(Garry, 2011).*

Slow-moving goods were initially investigated by *Whitin and Youngs*. Although supply chain and stock management have progressed for continuous/smooth demand, most developments are not applicable to slow-moving items yet *(Williams, 1984).*

As *Cavalieri et al. (2008)* point out, the main problems of slow-movers are the lack or poor historical data about demand, since there are many zero values and an extreme variation of the quantity demanded. Thus, for slow-movers, a regular reorder-based inventory control system is not appropriate, since in many cases the demand rate will lead to zero replenishment *(JenYeh et al., 1997).* To determine how to forecast slow-movers' future demands, choosing the right periodic inventory system is crucial *(Sani & Kingsman, 1997).* In fact, demand forecasting is one of the foundations for planning a company's operations, which is a fundamental task in supply chain management, especially when facing such intermittent and variable demands *(Wong & Guo, 2010).* It is the basis for the company to arrange the sales and allocation plans at each store and shelf *(Ni & Fan, 2011).*

2.3.6. XGBoost

Machine learning has great potential regarding demand forecasting, as it provides flexible models. However, those methods have not been properly investigated in the context of supply chain management yet *(Abolghasemi et al., 2019).* For instance, XGBoost (Extreme Gradient Boosting) is an improved boosting algorithm belonging to supervised learning, which, using CART (Classification and Regression Trees), builds a strong model by integrating weak classifiers *(Thongsuwan et al., 2020)*. In essence, XGBoost performs feature splitting and tree addition iteratively to develop a tree model. As new additions are made, a new function can be learned in order to fit the residual from the previous prediction. Based on the characteristics of

the sample, each tree calculates the score for each child node, which when added up gives the predicted sample score *(Chen & Guestrin, 2016).*

Being a highly efficient implementation of Gradient Boosted Decision Trees (GBDT), XGBoost supports not only decision trees, but also weak learners such as GBtrees, GBblinear, and Dart. Moreover, although GBDT's loss function only performs negative gradient (first-order Taylor) expansion on the error, XGBoost's applies a second-order Taylor expansion, which improves the accuracy of the prediction. In addition, XGBoost models have the advantage of being easily scalable for a variety of scenarios, as well as requiring fewer resources than other models. Parallel and distributed computation within XGBoost speeds up model learning and accelerates model exploration. Notably, XGBoost is a robust and effective Machine Learning method in prediction *(Chen & Guestrin, 2016).*

2.4. Design And Development

2.4.1. Data Understanding

At the beginning, different data sets were provided, in order to understand the flow of products within the store. Even so, for the purposes of developing the picking list, sales data provided the most relevant, comprehensive, and useful information. As a matter of fact, it contained relevant information, such as the unique identifier of the customer's basket, the date on which the product left the store after a sale, or even the number of units purchased.

Additionally, the chosen data set included a variable that was common to both *Planogram* and *Product Lookup* datasets, allowing them to be merged and our analysis expanded. Regarding the *Product Lookup* dataset, we will use it to obtain a deeper level of contextualization of the products purchased, such as their category, division, family, and brand. On the other hand, the *Planogram* data consists of information regarding the organization of the products in the store.

For instance, the different products that should be displayed on each shelf, as well as its capacity, were considered.

For this reason, our analysis will be primarily based on the *Sales* dataset, which contained information from September 1, 2022, to September 30, 2022. Plus, it comprised 123,261 rows and 37 columns, where each row represents a product's sale.

2.4.2. Data Curation & Feature Engineering

Once the most relevant data sets have been chosen, it is essential to examine the quality of the data, which is the foundation of any data-driven project. As a matter of fact, data curation plays a vital role in the preprocessing steps. As mentioned above, three different datasets were selected to conduct our analysis: *Sales*, *Product Lookup*, and *Planogram*. In spite of that, some transformations were supported by the *Events* dataset, as well.

Starting with sales, we transformed existing variables and added new ones, to ensure a clean and clear dataset. First and foremost, we discovered that there were no duplicated rows, nor columns with null values, which greatly expedited our analysis. Then, we proceed by eliminating all variables that did not prove to be relevant and pertinent, such as *Id Loja*, since all sales were conducted at the Nova & Go store, or *UMV*, which just provided additional information regarding the units of measurement of each products sold.

Furthermore, we also created some variables that would later facilitate the visualization and interpretation of the data. For instance, a variable called *Time of the Day* was added, describing the distribution of sales throughout the day, as well as a new column *Weekday*, that identified the day of the week when the sale took place. Plus, from the use of this variable, 12 lines of data were eliminated. These corresponded to entries mistakenly recorded as Sunday purchases, when, in fact, the store is closed. Similarly, and with the support of the *Hour TXS* variable, we were able to eliminate all sales that occurred outside of the store's operating hours. Moreover,

all negative quantities and prices (*QTE* and V*alue PV*) were removed, which are the result of returns and therefore are not useful for analysis.

In addition, we wanted to analyze revenues over the course of the month. Then, to calculate the values of the *Revenues* column, the price, V*alue PV*, was multiplied by the *QTE* variable, which represented the number of units sold.

As explained before, in order to obtain more information about the products' specificities, we aimed to merge the *Sales* data with the *Product Lookup* dataset. To accomplish this, however, certain adjustments had to be made beforehand, such as the removal of products without *Article Code*, which would make the merger impossible. In fact, different granularities of the products sold, from categories to subcategories, had more than one unique code per category description. Similarly, as a result of spelling errors, some categories, subcategories, and products had different descriptions.

Starting with the cases in which there were two codes for the same description, we assumed that the appropriate solution would be to keep the code presented in the *Events* dataset. This way, according to the data from last September, this code was just recently used, whereas the others might have been deprecated. Moreover, to correct spelling errors in product names, the same reasoning was followed, which led us to always verify that the names matched those in the *Events* dataset. Whenever detecting these cases was not possible, we proceeded to eliminate them. After these changes, we were then able to merge the *Sales* dataset with the *Product Lookup* one, through *Article Code*.

After obtaining a complete dataset that included all the product information, it was necessary to obtain specific information from the planogram dataset. Initially, similar to what was done before, some data problems were addressed in the planogram. For instance, we eliminated all negative values for *Units per Box* and *Filling*, as they were not realistic. Then, once the variables had been curated, the dataset was ready to be merged with the *Sales* dataset using the

EAN column. Yet, due to the planogram covering primarily non-perishable items, the number of rows was reduced by 57%, when merging it.

As a result of the merger, it became possible to begin thinking about a strategy for resolving the issue at hand. In this way, as a first step, the fast-movers' products were identified. According to JM, a product is considered a fast-mover if it is sold in quantities exceeding the store's capacity sixteen or more times per month. Additionally, it should be noted that a product is considered sold out when the quantity purchased exceeds the shelf's capacity.

Following these definitions, we have developed a method that calculates whether a product is out-of-stock, based on the number of units sold per day and the capacity. Additionally, through this method, we were also able to record the precise time when the out-of-stock occurred, which can be explored to determine when these situations occur more frequently. After completing the data curation, the cleaned dataset included 52,931 rows.

2.4.3. Exploratory Data Analysis

After a first look at the data, we conducted an exploratory data analysis (EDA), which identified potential insights from the data beyond the formal modelling task.

To begin with, we explored how sales flowed throughout different granularities of time, such as monthly, daily or even hourly. We were interested in understanding whether client behavior differed depending on seasonality. Considering the information provided in *Table 18*, it is possible to follow most of the conclusions presented in this section.

First of all, as shown in *Fig 2 in the appendix*, the day that registered the largest number of sales, measured in total units sold, was the 20th of September. In fact, 3468 total sales occurred on that day. However, in contrast to our expectations, this did not coincide with the day with the largest number of visits. Rather, it was September 27th the day with the most entries, when 3,302 transactions were recorded. This implies that the total entries are not an accurate indicator

of a day's performance, since each person can purchase more than one item, which will result in a higher revenue result for the firm.

Moreover, we were able to conclude that all Saturdays were relative minimums in terms of total sales. This information is aligned with the information on the number of entries, as it is also the day of the week with less visits. This can be explained by the fact that there are no classes taking place, so there are usually fewer students on campus.

Moreover, we intended to examine how sales were distributed throughout the day. According to our analysis, the highest sales peak occurred during lunch time, which is not surprising. This is when most people are moving through the store *(Fig 3 in the appendix).*

Similarly, by analyzing the distribution of total sales by day of the week, on average *(Fig 4, 5, 6 and 7 in the appendix),* it was possible to determine that, on average, the best performer, in average total units sold, was Tuesday. Coincidentally, this was also the day when the greatest number of customers entered the store. It is estimated that on average 2,990 units were purchased on Tuesdays, resulting in an average revenue of 4,397 €. Besides, the only days on which average total sales, both in units and euros, were below the average were Friday and Saturday. Some other insightful statistics can be drawn from the plots. For instance, the average number of units sold per day was equal to 2,244 units. Additionally, the average daily revenue corresponded to 3,597 €.

The next step was to analyze the products sold in more detail. First, we identified the top products and divisions, which allowed us to realize that beverages were mainly the best-selling products *(Fig 8 in the appendix).* Additionally, as shown in *Fig 9 in the appendix,* none of the top 30 products had a selling price superior to 3.5 €.

Due to the fact that we had already defined the fast and slow-movers in the data curation step, we were able to determine, in more detail, which were the best sellers for each of these categories. In fact, while slow-moving products are dominated by beverages *(Fig 10 in the*

appendix), fast-movers' best-selling products *(Fig 11 in the appendix)* include beverages (more specifically water), sandwiches and salads. In terms of the total sales performed by both fast-movers and slow-movers over the course of the week *(Fig 12 and 13 in the appendix),* it appears that there were more slow-movers sold on average than fast-movers. A possible explanation for this is the fact that there are significantly more fast-movers (19 unique products) than slow-movers (1062 unique products). Additionally, Fridays and Saturdays were the days of the week on which more fast-movers were sold than slow-movers.

In this phase, we dive into the issue of out-of-stock products. As expected, these situations are most likely to arise during lunch time, when the store is more crowded, and the number of sales is also greater *(Fig 14 in the appendix).* Additionally, we examined the number of stock-outs events on a daily basis *(Fig 15 in the appendix).* The first thing we discovered was that there was at least one case of stock-out every day in the store, which indicates the severity of the problem. In fact, on average, there were 27 daily out-of-stock cases. As a result, there is evidence that logistics failed to replenish shelves on a timely basis. In addition, we were able to determine that the 20th of September experienced the highest rate of out-of-stocks, with a total of 50 cases, as this was also the day with the largest number of products sold *(Fig 16 in the appendix).*

Afterwards, we examined the Top 30 best-selling products that went out-of-stock during September *(Fig 17 in the appendix).* As a result, we realized that the maximum number of times that a specific product went into stock-out was 23. This reveals an extremely large demand for the two products in this situation, namely "*maçã royal*" and "*água monchique*". Plus, among these 30 products, 63% correspond to fast-movers. This makes sense, since the definition is based on the number of times a product is out-of-stock.

As a result of the wide variety of products, which have varying requirements, replenishment is a very complex task that requires an automated tool. The use of data can facilitate this process.

Based on the results of this exploratory analysis, it is clear that it will be necessary to optimize more than just the replenishment process. Rather, to redefine the shelves available for each product is also an important step. In this sense, based on *Fig 18 in the appendix*, a comparison was made between average sales and shelf capacity for the most popular products in September. We chose to explore these cases in greater depth because of their importance. Thus, we found that average sales were significantly higher than shelf capacity, for all the top stock-out products.

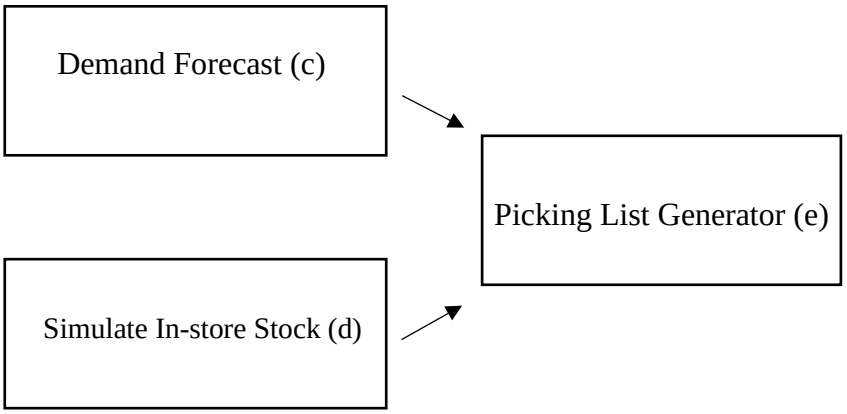
2.4.4. Model

a) Overview

Throughout this section, we provide an overview of the phases involved in developing the algorithm, in order to generate a picking list. Using the demand forecast model (c) and the simulation of the in-store quantities of products (d), we were able to develop the list (e) later. Essentially, the purpose of the picking list is to ensure that the store manager knows how many slow-moving products need to be replenished every day. In this way, store restocking decisions will become more informed and data-driven.

Therefore, we will follow these three main steps throughout this section:

Fig 19– Model Overview



b) Data selection and preparation

An effective model requires input data that is aligned with the desired outcome. This section, then, will provide a forecast of demand that is informative, i.e., a forecast of sales volumes. In order to accomplish this, the data was grouped by product and by day of the month. In addition, the number of boxes sold on a daily basis, per product, was computed and included in the dataset as a feature.

Furthermore, since not all products were purchased every day, we develop a method to fill this gap, so that this would not have an adverse effect on the model. Our objective was to include a row for each product, every day, regardless of whether it has been sold In this sense, we identified all the available products that did not have a complete list of all the days that the store was open. In these cases, a new row was created, where the quantity sold (*QTE*) was zero. In this way, all products ended up having the same number of days in the time series.

c) Model Training & Data Product

i) Demand Forecast

Time Series Forecasting is an important subject in Statistics, Machine Learning, Operations Research, and Data Mining *(DT Wiyant et al., 2021)*. For the forecasting of time series data, we decided to use XGboost. In fact, several papers reinforce the numerous advantageous characteristics of this model, which have already been contextualized in the Literature Review. Among these, *Zhang et al., 2021* developed a study that demonstrated that this algorithm was able to outperform Deep Learning, when it comes to Time Series Forecasting.

As mentioned above, one of the most significant factors contributing to the success of XGBoost is its ability to scale across a wide range of scenarios. In fact, it is ten times faster than the existing popular solutions and scalable to billions of examples *(Chen & Guestrin, 2016).*

To forecast demand, time series are usually divided into two continuous parts. So, for our model, we used the first three weeks of September to train the model, and later, the last week

to test it. Additionally, due to the limited number of relevant attributes, the input features for the model were trivially the daily number of boxes sold per product, and a time variable, in this case, days. Furthermore, due to the limited amount of data, our time horizon was quite short, as we only had one month of data.

It is pertinent to note that the main objective was to obtain a forecast demand by product. As a starting point, in order to verify the validity of the model, a demand forecast for only one single product was generated. As will be discussed in the evaluation section, the model showed promising results and a high degree of predictability. As a next step, it was later generalized to all products.

ii) Fine tunning

Although the algorithm provides solutions for the parameters, the performance of the model is strongly influenced by the hyperparameters selected *(Shekhar et al., 2022).* Therefore, this choice is one of the most critical factors in the fitting of supervised Machine Learning algorithms *(Bergstra & Bengio, 2012).*

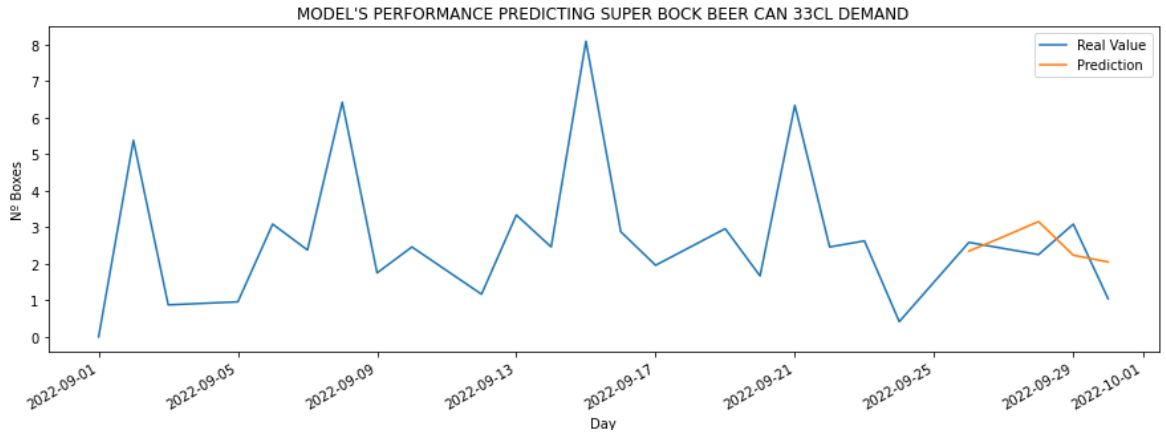
In essence, the main difference between parameters and hyperparameters is that, while the former are learned by optimizing loss functions or gradients; hyperparameters, on the other hand, cannot be inferred and will control this learning process *(Shekhar et al., 2022).*

The best hyperparameters can be obtained through hyperparameter optimization. The same can be achieved using scikit-learn, a Python-based Machine Learning library. In fact, a variety of algorithms can be applied. Specifically, to adjust our model's hyperparameters, we decided to use Grid Search. It is a technique that analyzes all possible combinations of hyperparameters until the one that results in the best model's performance is found *(Shekhar et al., 2022).* For example, to avoid overfitting, the hyperparameter *max_depth* was adjusted at this stage. Furthermore, the learning rate that is used to define the step size in each iteration has also been adapted to meet the requirements of the model.

iii) Model Evaluation

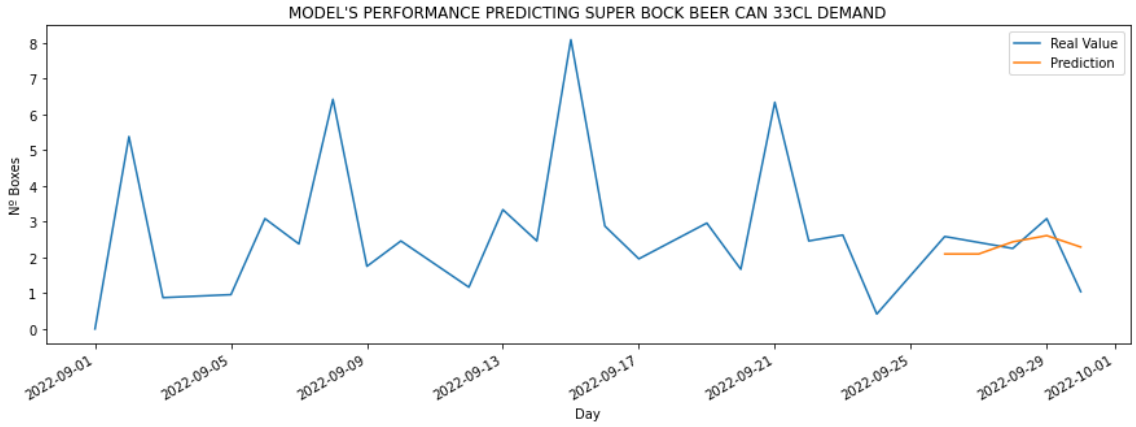
Having trained the model and generated predictions, the next step was to analyze and evaluate the results. As discussed in the Model section, we initially applied the model to just one product, in order to simplify the interpretation and visualization of its performance. Furthermore, the once-case scenario was then trained without and with Grid Search, optimizing its hyperparameters. To determine the impact of this on prediction accuracy, we used a specialized testing technique called Walk-Forward Validation. Since we are testing the model only on one product, one was randomly chosen. In this case, Super Bock beer was used as an example.

Fig 20– Boxes sold vs Model predictions (without grid search)



This first plot *(Fig 20)*, which corresponds to forecasts made without Grid Search optimization, showed a great deal of closeness between the actual values (blue line) and those predicted by the model (orange line), confirming the model's accuracy.

Fig 21– Boxes sold vs Model predictions (with grid search)



However, in the second graph *(Fig 21)*, which represents the performance of the model using the Grid Search technique, we can observe an even greater proximity between the orange line and the blue than in *Fig 20*. This reflects an improvement in the model's performance. Therefore, through Walk-Forward Validation, we were able to verify that Grid Search improved the model for individual products (in this case, *Super Book beer*). Yet, the entire dataset still needs to be evaluated, to confirm that the model provides reliable predictions.

Considering the wide variety of products and patterns across the month, it was necessary to utilize different metrics, when analyzing the model's performance. This allowed us to examine not only how Grid Search affected individual products, but how it affected the model as a whole, as well.

The overall performance of the model was measured using mean absolute error (MAE) and Mean Squared Error (MSE) metrics. When evaluating regression models, the mean absolute error represents the mean absolute difference between the prediction of the model and the target value *(Mitra et al., 2022).* This value is then calculated by summing the absolute errors and dividing it by the number of errors:

Equation 1- Mean Absolute Error Formula

$$MAE = \frac{1}{n}\sum_{i=1}^{n}\left|y_i - \hat{y}_i\right|$$

Similarly, the Mean Squared Error (MSE) is also used to measure the quality of an estimator, but rather it gives the squared difference between the model's prediction and the target value (*Mitra et al., 2022*).

Equation 2- Mean Squared Error Formula

$$MSE = \frac{1}{n}\sum_{i=1}^{n}\left(y_i - \hat{y}_i\right)^2$$

It is important to note that the MAE measures the absolute distance between the observations and the regression predictions, taking into account the average of all observations. Then, to account for negative errors, the absolute value of distances is used. In turn, the MSE performs this function by squaring the distance, thereby obtaining a positive result. Consequently, larger errors (or distances) have a greater impact on the measurement than smaller ones. Even so, in both metrics, the closer the error value is to zero, the more accurate the model is.

Since we have developed a single model per product, we had to calculate the mean of the results obtained for each product, in order to evaluate the model's overall performance. It is noteworthy that the model achieved excellent results, even without fine-tuning, with an average MSE of 0.2277 and MAE of 0.1915. This indicates that the model had a high degree of quality, since there were not excessively disparate predictions, compared to the observed values.

Using Grid Search, as expected, had a positive impact on the forecast performance, because of the advantages of this type of technique. Thus, Grid Search improved both MAE and MSE to 0.1625 and 0.1311, respectively. In this way, we were able to conclude that the error was reduced, due to the optimization of the hyperparameters. These results confirm all the claims made regarding the benefits and impact of this technique on ML models, such as the XGboost, used in this proposed solution.

b) Simulate In-Store Stock

As previously stated, the replenishment process has not been adequately studied by JM before. Consequently, at Nova & Go store, this task is still performed without much data support, relying primarily on human intuition. To later develop a picking list, it was first necessary to understand how replenishment works, and in what situations it is necessary.

Firstly, to monitor the stock levels of each product, we have created a new variable: *Units Sold*. As the product's sales evolved throughout the day or even week, the value of this feature was then updated, increasing for each unit sold. Once the shelf capacity minus the total number of units sold exceeds the recommended minimum quantity, products require replacement, and *Units Sold* is reset to a 0 value. Otherwise, *Units Sold* will remain the same as the previous day, as no units were restocked. By analyzing the values taken, we can estimate the number of products that have not yet been replenished.

As a result, we were able to develop a methodology capable of defining and understanding the situations that would require a replacement, as well as the number of units that were left on the shelf, at the end of a day. Yet, as this is a first step on transforming the replenishment process into a data-driven activity, a daily-based analysis has been assumed. All replenishment needs were defined as those in which the shelf capacity (*Filling*) minus the number of units sold, *Units Sold*, is less than the minimum quantity required (*Minimum*).

Furthermore, it was crucial to outline the replenishment strategy adopted by the Nova & Go store, so that it could be reflected in the model afterwards. During the meetings with JM, it was explained that the process revolved around boxes of products, rather than units. As a result, from the picking list's viewpoint, a box could not be partially used for restocking. Instead, for products with a small shelf capacity, the entire box would be used when refilling. Likewise, high-demand products, which are commonly kept on larger shelves, would be replenished based on the maximum number of full boxes that can be accommodated on the shelf capacity. This

assumption, however, has some flaws and evidences the lack of maturity in the study of the store's replenishment process.

The best way to understand this is by taking an example into account. Suppose that a given product had a minimum volume of 3 units and a capacity of 10. Once the quantity available on the store’s shelf fell below the threshold, the product was flagged with a need for replenishment. Assume that, at the end of the day, there were only 2 units left. Moreover, each product’s box contained 10 units. In this case, the entire box would be used to restock the shelf, resulting in 12 units available to be sold. However, this is not realistic, since it exceeds the shelf's capacity. Hence, even though the entire box was included on the picking list used by the employee when restocking, these additional 2 units would be stored once again in the warehouse. There was an overestimation of the quantity needed.

In contrast, for a given product, if the capacity is greater than the quantity of a single box, underestimation may occur. For example, suppose a product's threshold was 4 units and the shelf's capacity equaled 25. Plus, each box contained 5 units. At the end of the day, there were 3 units left. Thus, only 4 boxes would leave the warehouse when replenishing for that product, corresponding to 20 units. As a result, 23 units would be displayed, even though the shelf's capacity was not fully utilized. There was an underestimation of the quantity needed.

Looking at the examples above, it is evident that the strategy used to base the picking list reasoning is not perfect. Nevertheless, such a tool can already result in a performance improvement, as employees are able to estimate the number of boxes necessary, rather than relying solely on intuition. Plus, by looking at the product's boxes rather than its units, it is possible to avoid the presence of many individual packages in the warehouse, which is a concern for the store's manager.

c) Picking List Generation

After identifying the instances in which replacement is required, a picking list can be developed. On the basis of the method outlined in the previous section for estimating stock quantities (section d) and forecasted demand (section c-i), we were able to identify all the products and their respective quantities that will require replenishment each day. As an illustrative example of the algorithm's output, a suggested picking list was generated for September 26. In *Fig 22 in the appendix,* we can see its structure, including the day on which the replenishment will take place, the names of the products, as well as the number of boxes that need to be replenished. However, it should be noted that, as a result of the large size of the list, only 25 products were shown, out of 94 existing products.

Furthermore, in the model's evaluation section, it was determined that the errors associated with forecasts had a negative impact on the estimated demand. As a result, this would affect the quantity of the suggestions for product replacements. In this regard, once again, we realize the importance of fine-tuning to optimize the hyperparameters and, consequently, improve the model. An important aspect of this step is that the quality of the model plays a significant role in the generation of the picking list.

2.5. Challenges And Limitations

Although XGboost was evaluated positively and the picking lists were created, it is important to reiterate some of the limitations we encountered during the development of this thesis. Particularly, because a data-driven replenishment process was a blind spot in Nova & Go store's daily activities, we had to start from scratch. As a consequence, in previous sections, it was necessary to make several assumptions and develop hypothetical scenarios. For instance, it was unavoidable to simulate the number of units displayed on the store' shelves. Moreover, additional detailed information concerning the replenishment activity, such as when it occurred,

the products that were replenished, or which products required redundant trips to the warehouse, would have been extremely beneficial for us to be able to generate more accurate picking lists. As discussed in the development section of the model, a significant limitation was the amount of data. Specifically, models that use time series as a basis for solving problems require extensive historical data to operate. For example, many recommend a minimum of one year of data. However, in our case, only September data was available. In this way, the limited amount of data not only hindered the selection of an appropriate model, but also adversely affected its performance. Because these models are based on patterns in the data, larger data sets are likely to exhibit more clearly defined patterns.

2.6. Recommendations For Future Steps

Using more data is the most immediate next step to improve the performance and generalization abilities of the model. In addition, in the future, the planogram should also be revised, since the exploratory analysis of the data suggested that some products had a storage capacity lower than the average sales volume. Finally, as part of this effort, it would also be beneficial to develop a better solution sustained on clearly defined metrics and KPIs, defined by JM, so that replenishment can become even more data-driven.

3. Conclusion

In *Hopping's (2000)* view, the history of retail industry reflects the progress of technology. In fact, retailers have been able to streamline and accelerate their processes over the years by implementing technology, from ATM card payments to RFID systems.

Nova & Go store exemplifies this increased ability to modernize and improve the shopping experience. Among other things, this store can gather various types of data, from sales, in-store product placement, inventory levels or even the cooking process of takeaway meals. The adoption of tools that can then synthesize, standardize, analyze, and visualize all these data collected is, therefore, essential to generate value. For the purpose of this thesis, Jerónimo Martins identified the slow-moving products stockouts as a challenge.

To prevent stock-outs, an inventory management tool was designed, focusing on slow-movers' products. Based on the sales data, demand was first forecasted. Then, if the estimated demand for a given product was exceedingly large, resulting in less than the recommended threshold for the shelf's capacity, the item would be flagged. At the end, all the products in these circumstances were added to a picking list. By identifying products that are likely to be out-of-stock, replenishment activities can be optimized, reducing forgetfulness and unnecessary trips to the warehouse. Given the large number of products offered by the store and the fact that some products are not replenished so frequently, the use of automated tools that enable data-driven decisions are critical.

3. References

Aalst, Wil M.P. van der, and S. Dustdar. 2012. “Process Mining Put into Context.” Undefined. January 2012. https://www.semanticscholar.org/paper/Process-Mining-Put-into-Context-Aalst-Dustdar/68fe733f2945297ff96509af8af257b95db49d11.

Aalst, Wil M.P. van der. 2013. “‘Mine Your Own Business’: Using Process Mining to Turn Big Data into Real Value.” 2013. https://www.researchgate.net/publication/283581576_Mine_your_own_business_Using_process_mining_to_turn_big_data_into_real_value.

Abolghasemi, Mahdi, Rob Hyndman, and Garth Tarr. 2019. “Machine Learning Applications in Time Series Hierarchical Forecasting.” 2019. https://arxiv.org/pdf/1912.00370.pdf.

Avlijaš, Goran, Ana Simićević, Radoslav Avlijaš, and Marijana Prodanovic . 2001. “Measuring the Impact of Stock-Keeping Unit Attributes on Retail Stock ...” 2001. https://www.researchgate.net/publication/281441888_Measuring_the_impact_of_stock-keeping_unit_attributes_on_retail_stock-out_performance.

Ballou, R. 2004. “Business Logistics/Supply Chain Management : Planning, Organizing, and Controlling the Supply Chain: Semantic Scholar.” Undefined. January 1, 2004. https://www.semanticscholar.org/paper/Business-logistics%2Fsupply-chain-management-%3A-and-Ballou/6efcdd5aead4d449111598f5c62f2d09afe1e521.

Beamon, B. M. 1998. Supply Chain Design and Analysis Models and Methods. International Journal of Production Economics, 55, 281-294. - References - Scientific Research Publishing. 1998. https://www.sciencedirect.com/science/article/pii/S0925527398000796.

Bergstra, James, and Yoshua Bengio. 2012. “Random Search for Hyper-Parameter Optimization.” 2012. https://www.jmlr.org/papers/volume13/bergstra12a/bergstra12a.pdf.

Berisha-Shaqiri, Aferdita. 2014. “Management Information System and Decision-Making.” 2014.https://www.researchgate.net/publication/287205806_Management_Information_System_and_Decision-Making.

Cavalieri, Sergio, Marco Garetti, Roberto Pinto, and Marco Macchi. 2008. “A Decision-Making Framework for Managing Maintenance Spare Parts.” 2008. https://www.researchgate.net/publication/230815391_A_decision-making_framework_for_managing_maintenance_spare_parts.

Chen, Tianqi, and Carlos Guestrin. 2016. “XGBoost: Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining.” ACM Conferences. August 1, 2016. https://dl.acm.org/doi/10.1145/2939672.2939785.

Christopher, Martin. 2005. *Logistics and Supply Chain Management: Creating Value-Added Networks*. Harlow: FT Prentice Hall.

Corsten, Daniel, and Thomas Gruen. 2003. “Desperately Seeking Shelf Availability: An Examination of the Extent ...” 2003. https://www.researchgate.net/publication/36385147_Desperately_Seeking_Shelf_Availability_An_Examination_of_the_Extent_the_Causes_and_the_Efforts_to_Address_Retail_Out-of-Stocks.

Council of Supply Chain Management. 2013. “CSCMP Supply Chain Management Definitions and Glossary.” SCM Definitions and Glossary of Terms. 2013. https://cscmp.org/CSCMP/Academia/SCM_Definitions_and_Glossary_of_Terms/CSCMP/Educate/SCM_Definitions_and_Glossary_of_Terms.aspx?hkey=60879588-f65f-4ab5-8c4b-6878815ef921.

Deloitte. 2011. “Data as a Strategic Asset.” Deloitte United States. 2011. https://www2.deloitte.com/us/en/pages/consulting/articles/data-strategic-asset.html.

DT Wiyant, Kharisudin, AB Setiawan, and AK Nugroho. 2021. “Machine-Learning Algorithm for Demand Forecasting Problem.” 2021. https://iopscience.iop.org/article/10.1088/1742-6596/1918/4/042012.

Ehrenthal, J., and Wolfgang Stölzle. 2013. “An Examination of the Causes for Retail Stockouts: Semantic Scholar.” Undefined. January 1, 2013. https://www.semanticscholar.org/paper/An-examination-of-the-causes-for-retail-stockouts-Ehrenthal-St%C3%B6lzle/ce89a2da4f08e90a02fdd33273fa3daf6f6c06a1.

Fayyad, Usama M. 1996. *Advances in Knowledge Discovery and Data Mining*. Menlo Park (Calif.): AAA/MIT Press.

Garry , Michael. 2011. “Distributors Grapple with Slow-Moving Items.” Supermarket News. November 7, 2011. https://www.supermarketnews.com/supplier-news/distributors-grapple-slow-moving-items.

Gazzola, Patrizia, Daniele Grechi, Ilaria Martinelli, and Roberta Pezzetti. 2022. “The Innovation of the Cashierless Store.” MDPI. Multidisciplinary Digital Publishing Institute. February 11, 2022. https://www.mdpi.com/2071-1050/14/4/2034/htm.

Graafmans, T., O. Turetken, J. Poppelaars, and Dirk Fahland. 2020. “Process Mining for Six Sigma: Semantic Scholar.” Undefined. January 2020. https://www.semanticscholar.org/paper/Process-Mining-for-Six-Sigma-Graafmans-Turetken/1947cbfc31d3321366f8809eb804457ef32e60db.

Grewal, Dhruv, and Michael Levy. 2007. “Retailing Research: Past, Present and Future.” Dr. Dhruv Grewal. 2007. https://www.sciencedirect.com/science/article/pii/S0022435907000644.

Har, Loh Li, Umi Kartini Rashid, Seah Choon Sen, Loh Yin Xia, and Lee Te Chuan. 2022. “Revolution of Retail Industry: From Perspective of Retail 1.0 to 4.0.” Procedia Computer Science. Elsevier. March 8, 2022. https://www.sciencedirect.com/science/article/pii/S1877050922003714?fr=RR-2&ref=pdf_download&rr=777a9fa62f08866f.

Hevner, Alan R., Salvatore T. March, Jinsoo Park, and Sudha Ram. 2004. “Design Science in Information Systems Research.” University of Arizona. Management Information Systems Research Center. 2004. https://experts.arizona.edu/en/publications/design-science-in-information-systems-research.

Hopping, Dan. 2000. “Technology in Retail: Semantic Scholar.” Undefined. January 1, 2000. https://www.sciencedirect.com/science/article/pii/S0160791X99000421?via=ihub.

Ivanov, Dmitry, Alexandre Dolgui, Marina Ivanova, and Boris Sokolov. 2017. “Literature Review on Disruption Recovery in the Supply Chain.” 2017. https://www.researchgate.net/publication/317297853_Literature_Review_on_Disruption_Recovery_in_the_Supply_Chain.

JenYeh, Q., T.P. Chang, and H. Chang. 1997. “An Inventory Control Model with Gamma Distribution.” Microelectronics Reliability. Pergamon. May 17, 1997. https://www.sciencedirect.com/science/article/abs/pii/S0026271496002958.

Lee, In. 2021. “The Internet of Things (IoT) for Supply Chain Management.” 2021. http://resources.css.edu/sais/2015_proceedings/pdfs/mbaa-2015_i_lee_internet_of_things.pdf.

M. Whitin, T., and J. W. T. Youngs. 1995. “A Method for Calculating Optimal Inventory Levels and Delivery Time ...” 1995. https://onlinelibrary.wiley.com/doi/pdf/10.1002/nav.3800020306.

Man, Jimmy, Atilla Terzioglu, Kumar Ranjan, Ritesh Biswas, and Chuck Dean. 2017. “Retail the Customer-Driven Cloud Economy - Deloitte.” 2017. https://www2.deloitte.com/content/dam/Deloitte/us/Documents/technology/us-technology-retail.pdf.

Manzini, R., and D. Bindi. 2009. “Strategic Design and Operational Management Optimization of a Multi ...” 2009. https://www.researchgate.net/publication/223267950_Strategic_design_and_operational_management_optimization_of_a_multi_stage_physical_distribution_system.

Mckinsey. 2018. “Achieving Business Impact with Data - Mckinsey & Company.” 2018. https://www.mckinsey.com/~/media/mckinsey/business%20functions/mckinsey%20analytics/our%20insights/achieving%20business%20impact%20with%20data/achieving-business-impact-with-data_final.ashx.

Mitra, Arnab, Arnav Jain, Avinash Kishore, and Pravin Kumar. 2022. “A Comparative Study of Demand Forecasting Models for a Multi-Channel Retail Company: A Novel Hybrid Machine Learning Approach - Operations Research Forum.” SpringerLink. Springer International Publishing. September 27, 2022. https://link.springer.com/article/10.1007/s43069-022-00166-4.

Monczka, Robert M., Robert J. Trent, and Robert B. Handfield. 1998. *Purchasing and Supply Chain Management*. Cincinnati: South-Western College Pub.

Nenes, George. 2021. “Inventory Management of Multiple Items with Irregular Demand: A Case Study.” European Journal of Operational Research. Elsevier BV. May 4, 2021. https://www.academia.edu/48120151/Inventory_management_of_multiple_items_with_irregular_demand_A_case_study.

Ni, Yanrong, and Feiya Fan. 2011. “A Two-Stage Dynamic Sales Forecasting Model for the Fashion Retail.” Expert Systems with Applications. Pergamon. August 6, 2011. https://www.sciencedirect.com/science/article/abs/pii/S0957417410006937?via%3Dihub.

Prahalad, C. K, and Venkatram Ramaswamy. 2004. The New Frontier of Experience Innovation. July 15, 2004. http://socialmediaclub.pbworks.com/f/cocreation.pdf.

Reinartz, Werner, and V. Kumar. 2000. “On the Profitability of Long-Life Customers in a Noncontractual Setting: An Empirical Investigation and Implications for Marketing: Semantic Scholar.” Undefined. October 2000. https://www.researchgate.net/publication/229819388_On_the_Profitability_of_Long-Life_Customers_in_a_Noncontractual_Setting_An_Empirical_Investigation_and_Implications_for_Marketing. Ross, David Frederick. 2015. *Distribution Planning and Control: Managing in the Era of Supply Chain Management*. Berlin: Springer.

Sanchez-Ruiz, Lidia, Beatriz Blanco, and Asta Kyguoliene. 2018. “A Theoretical Overview of the Stockout Problem in Retail: From Causes ...” 2018. https://www.researchgate.net/publication/328337965_A_Theoretical_Overview_of_the_Stockout_Problem_in_Retail_from_Causes_to_Consequences.

Sani, B, and B G Kingsman. 1997. “Selecting the Best Periodic Inventory Control and Demand Forecasting Methods for Low Demand Items - Journal of the Operational Research Society.” SpringerLink. Palgrave Macmillan UK. December 18, 1997. https://link.springer.com/article/10.1057/palgrave.jors.2600418.

Syntetos, M, John Boylan, and JD Croston. 2005. “On the Categorization of Demand Patterns.” 2005.https://www.researchgate.net/publication/28578603_On_the_categorization_of_demand_patterns.

Thongsuwan, Setthanun, Saichon Jaiyen, Anantachai Padcharoen, and Praveen Agarwal. 2020. “ConvXGB: A New Deep Learning Model for Classification Problems Based on CNN and XGBoost.” Nuclear Engineering and Technology. Elsevier. August 2, 2020. https://www.sciencedirect.com/science/article/pii/S1738573319308587.

Shekhar, Shashank, Adesh Bansode, and Asif Salim. 2022. “A Comparative Study of Hyper-Parameter Optimization Tools - Arxiv.” 2022. https://arxiv.org/pdf/2201.06433.pdf.

Williams, Terry. 1984. “Stock Control with Sporadic and Slow-Moving Demand.” 1984. https://www.researchgate.net/publication/245279534_Stock_Control_with_Sporadic_and_Slow-Moving_Demand.

Wong, W. K., and Z. X. Guo. 2010. “A Hybrid Intelligent Model for Medium-Term Sales Forecasting in Fashion Retail Supply Chains Using Extreme Learning Machine and Harmony Search Algorithm.” International Journal of Production Economics. Elsevier. July 23, 2010. https://www.sciencedirect.com/science/article/abs/pii/S0925527310002331.

Zhang, Lingyu, Wenjie Bian, Wenyi Qu, Liheng Tuo, and Yunhai Wang. 2021. “Time Series Forecast of Sales Volume Based on XGBoost.” 2021. https://iopscience.iop.org/article/10.1088/1742-6596/1873/1/012067.

4. Appendix

Appendix Formulas

Equation 1- Mean Absolute Error Formula

$$MAE = \frac{1}{n}\sum_{i=1}^{n}\left|y_i - \hat{y}_i\right|$$

Equation 2- Mean Squared Error Formula

$$MSE = \frac{1}{n}\sum_{i=1}^{n}\left(y_i - \hat{y}_i\right)^2$$

Appendix Figures

Fig 1– Data Science Research

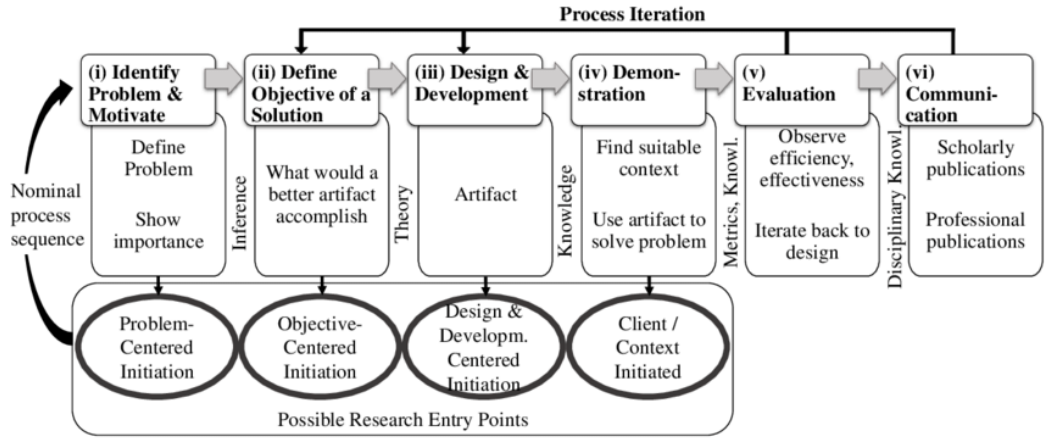


Fig 2 – Total Sales by Day, in Units Sold

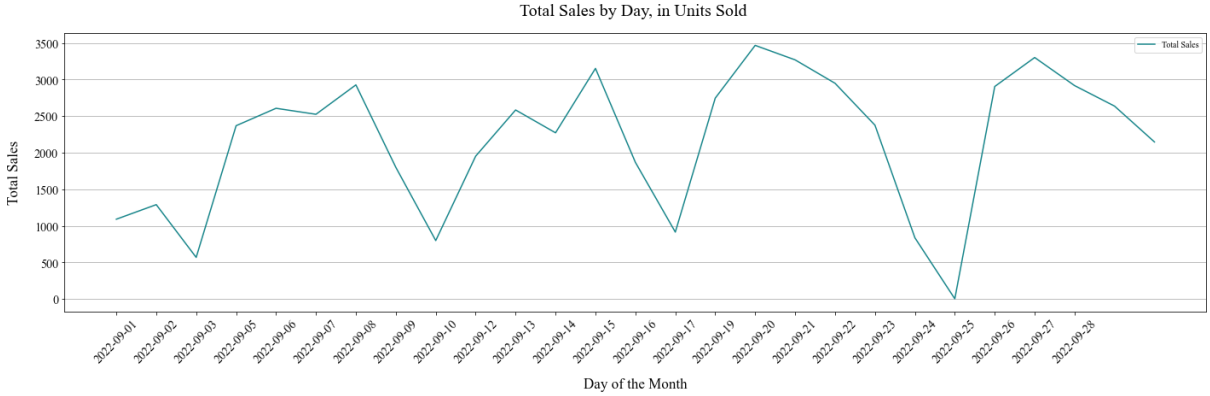


Fig 3 – Total Sales throughout the Day, in Units Sold

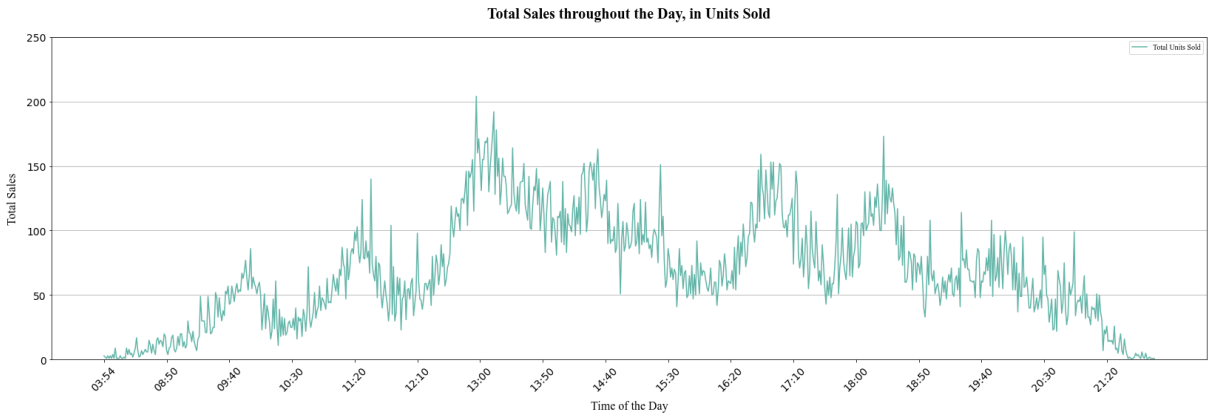


Fig 4– Total Sales by Weekday, in Units Sold

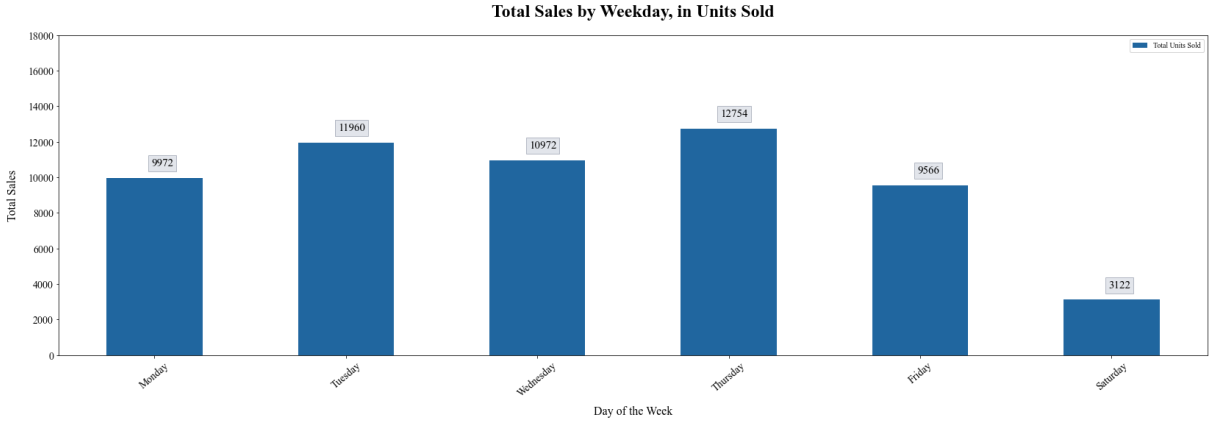


Fig 5– Average Total Sales by Weekday, in Units Sold

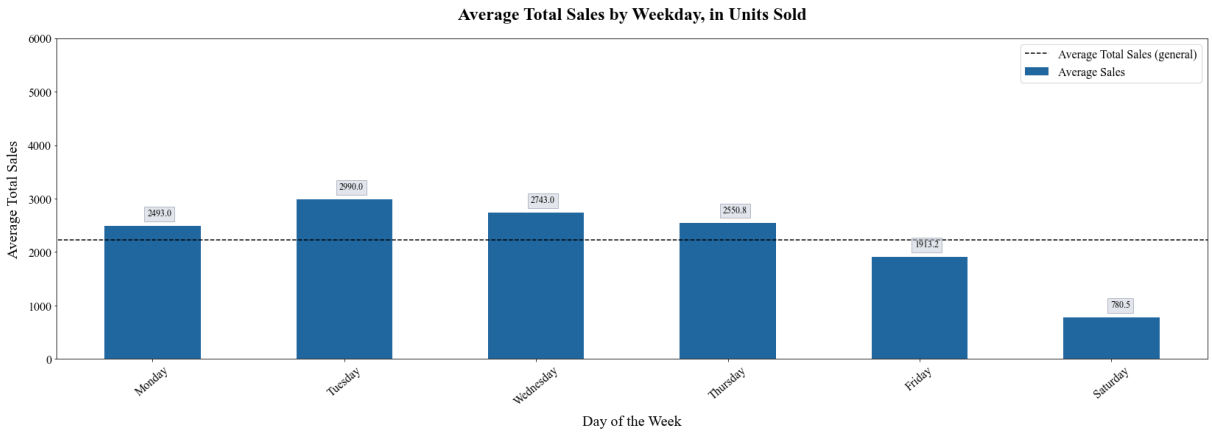


Fig 6– Total Sales by Weekday, in Euros (€)

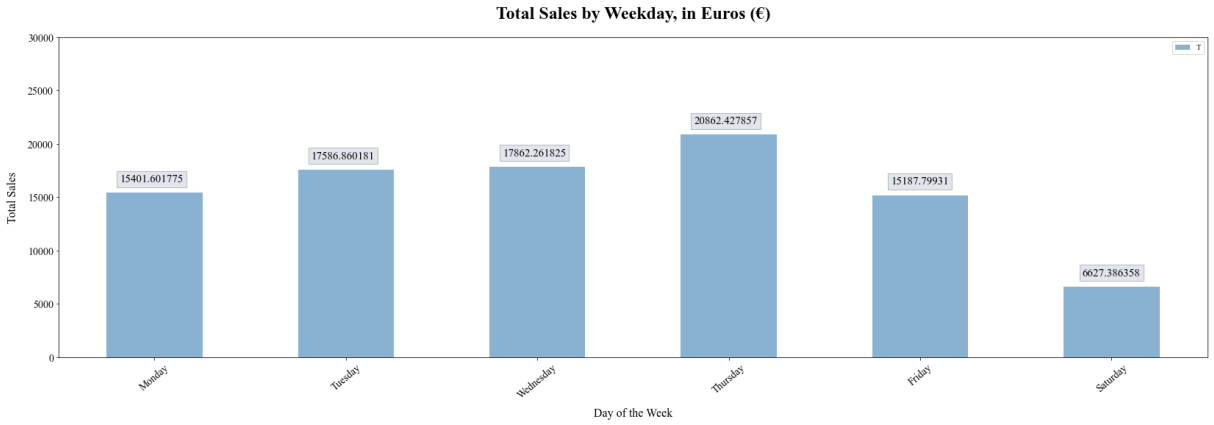


Fig 7– Average Total Sales by Weekday, in Euros (€)

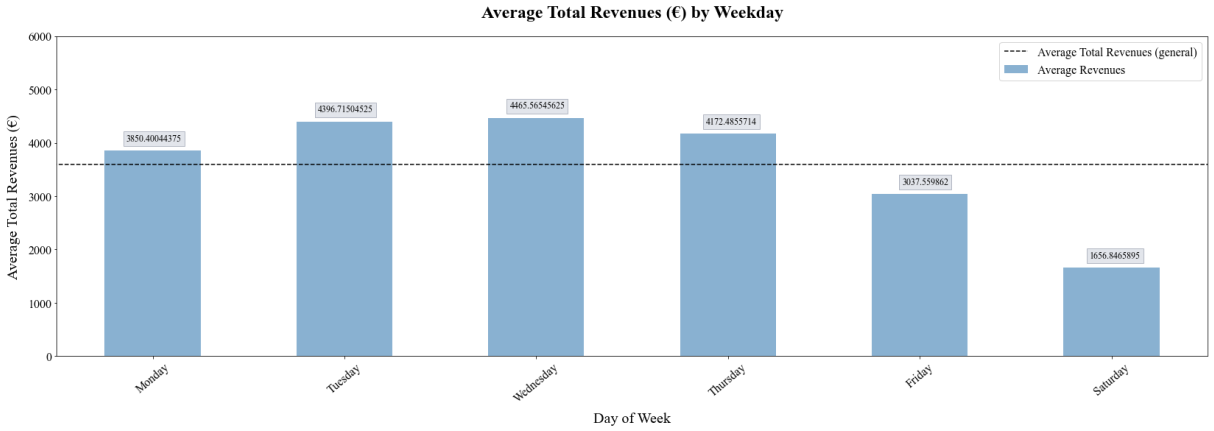


Fig 8– Top 10 Most Sold Products, in Units

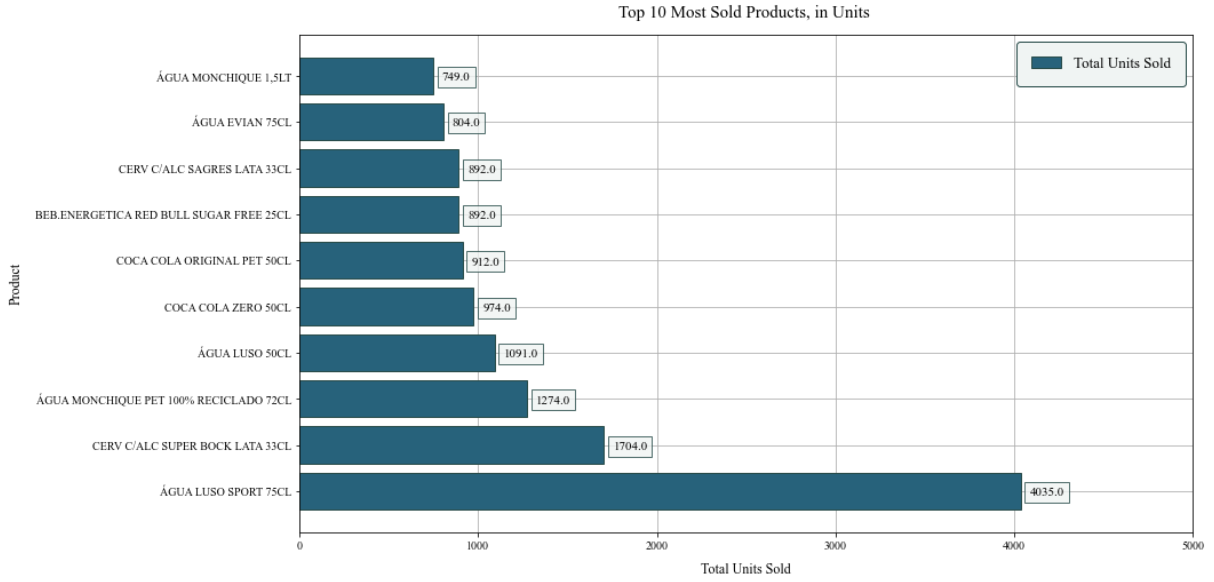


Fig 9– Total Units Sold and respective Price per Unit, for the Top 30 Most Sold Products

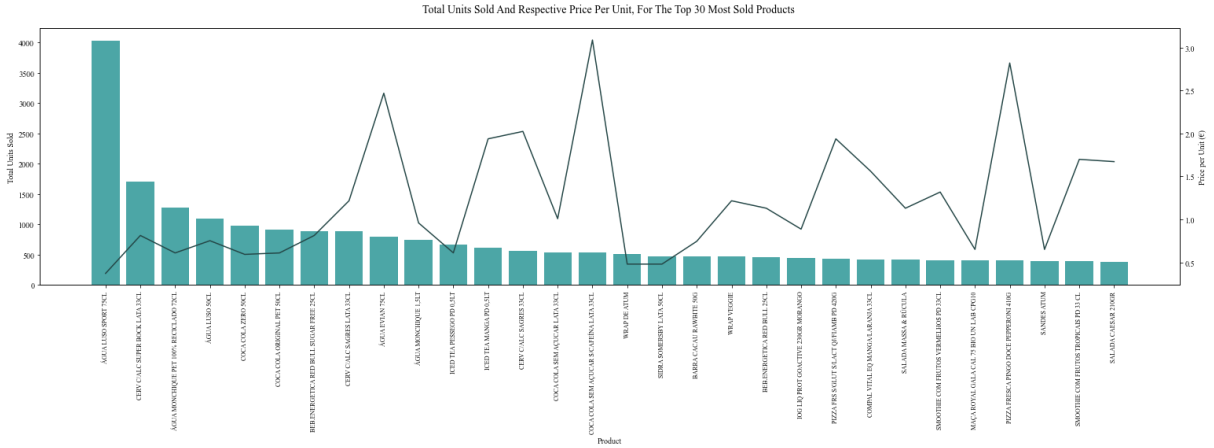


Fig 10– Top 10 Slow-mover Products, in Total Units Sold

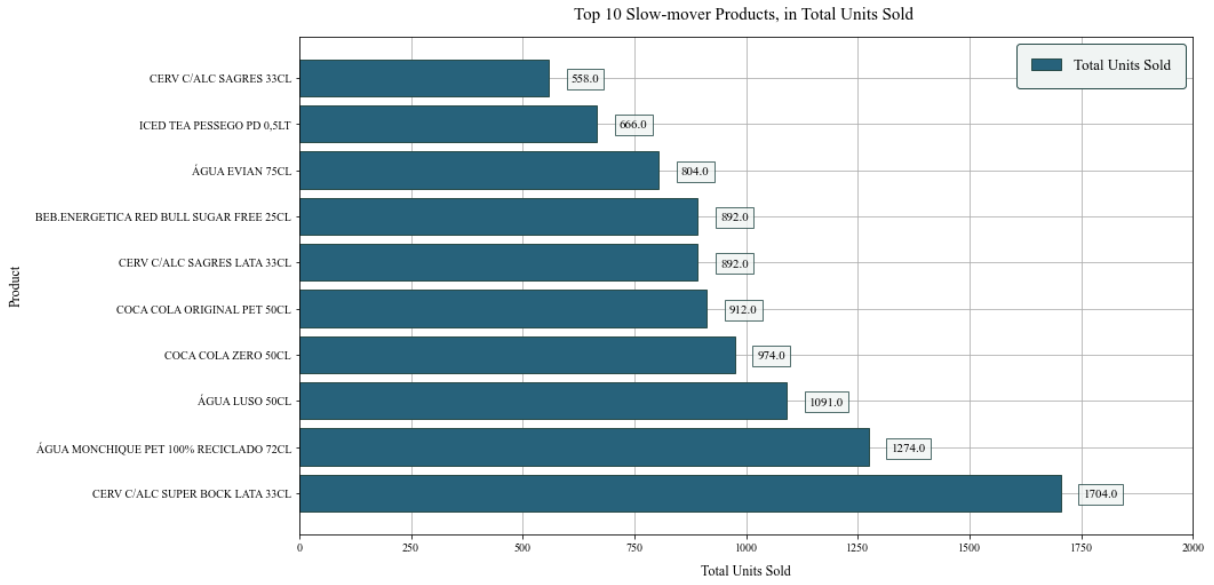


Fig 11- Top 10 Fast-mover Products, in Total Units Sold

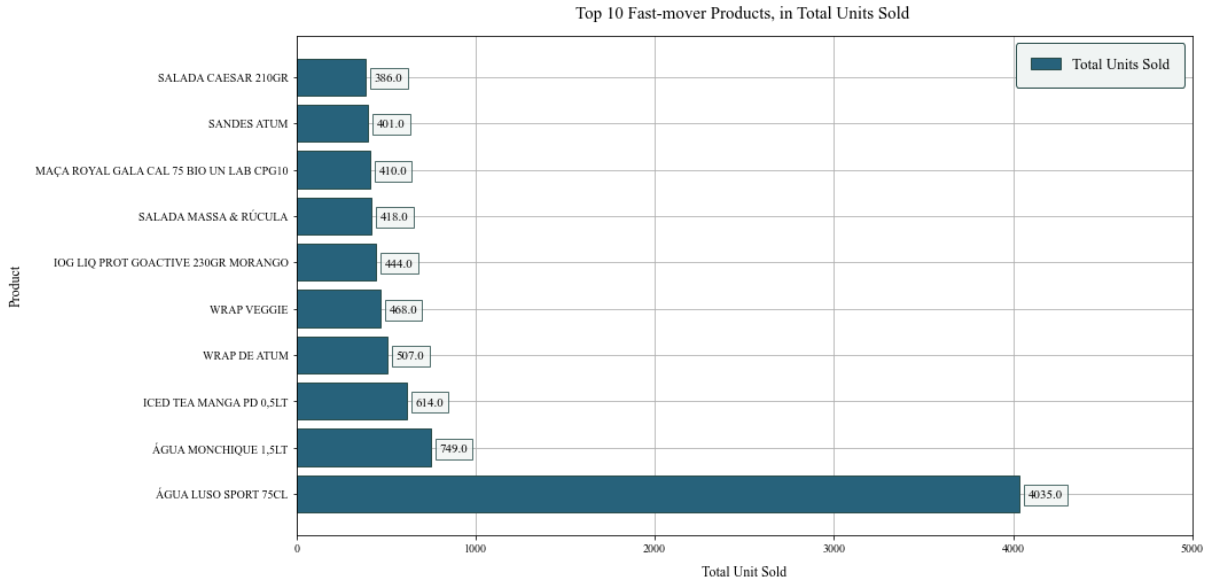


Fig 12– Average Sales by Weekday, in Units Sold, considering Fast-Movers Products

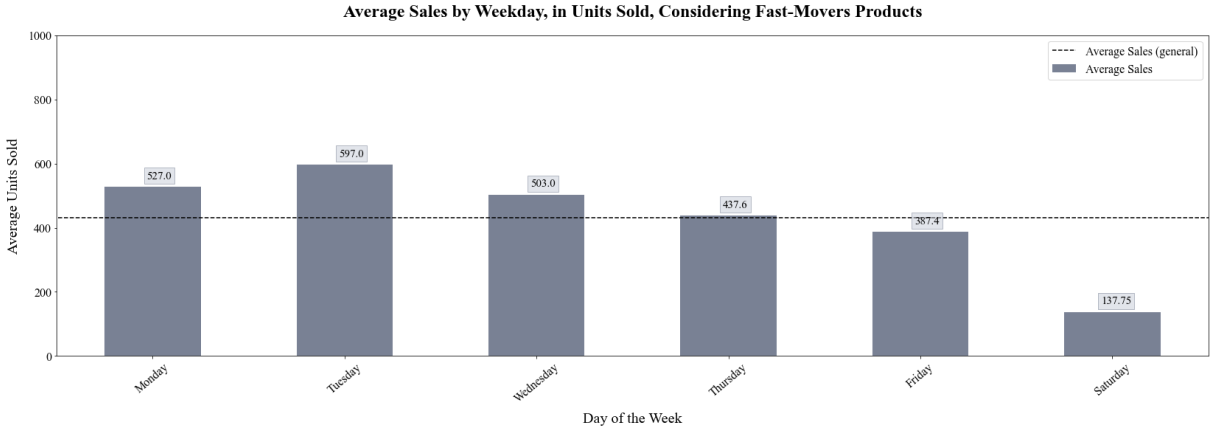


Fig 13– AverageSales by Weekday, in Units Sold, considering Slow-movers' products

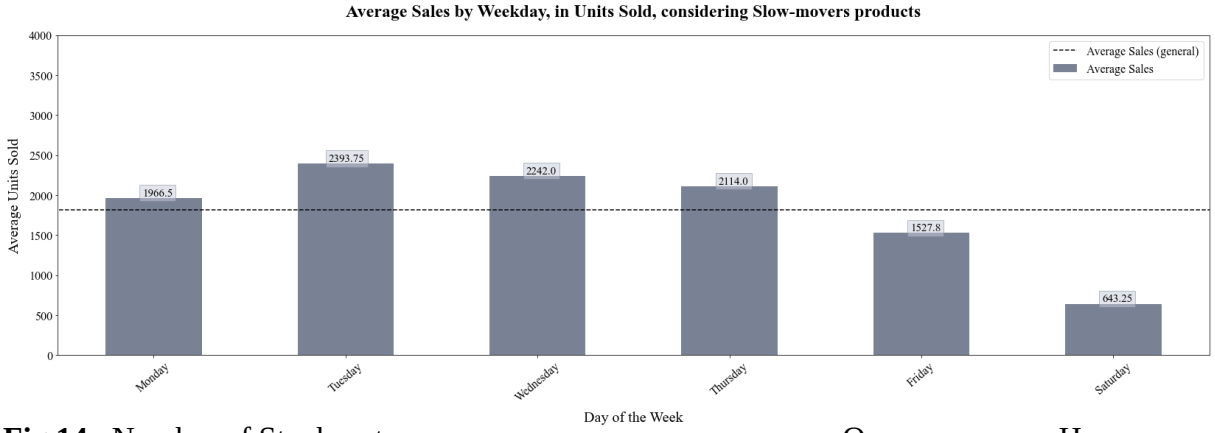


Fig 14– Number of Stock-outs Occurrences per Hour

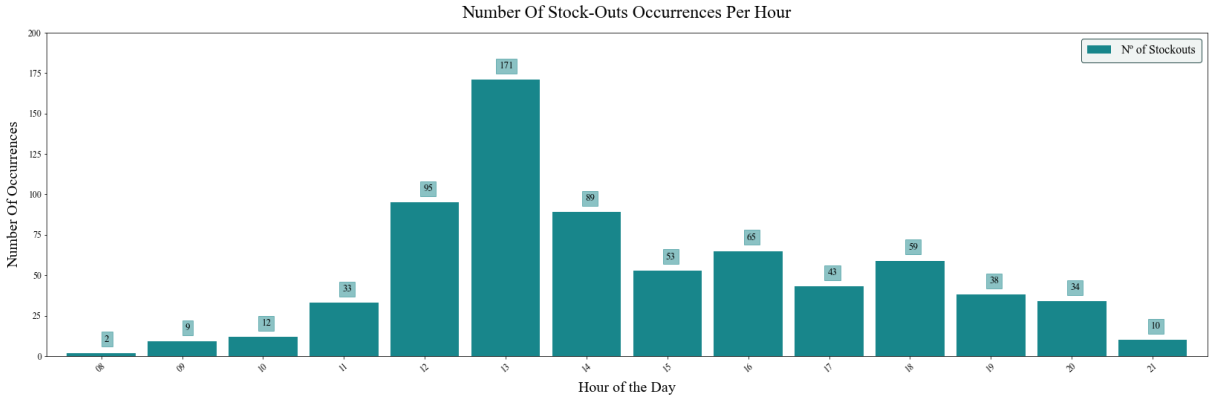


Fig 15– Number of Stock-outs Occurrences per Day

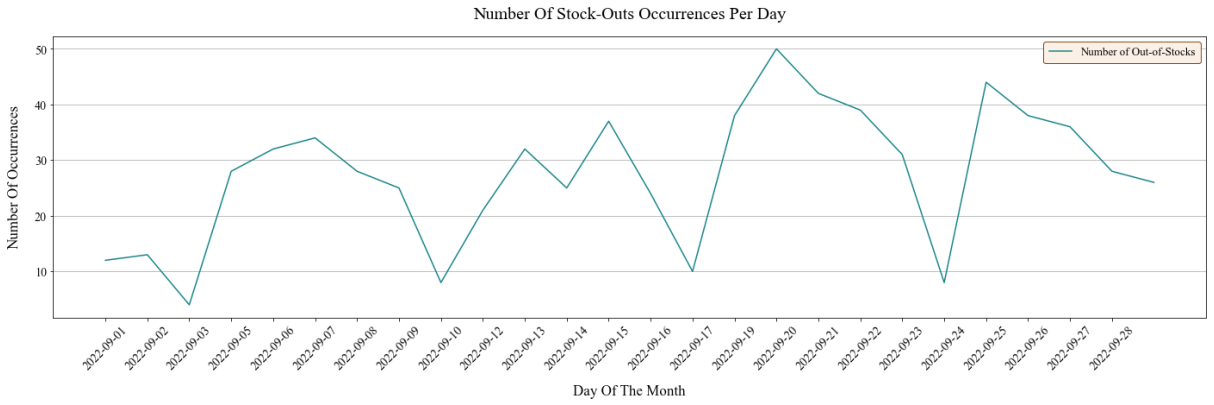


Fig 16– Top 30 Products with more occurrences of Stock-out

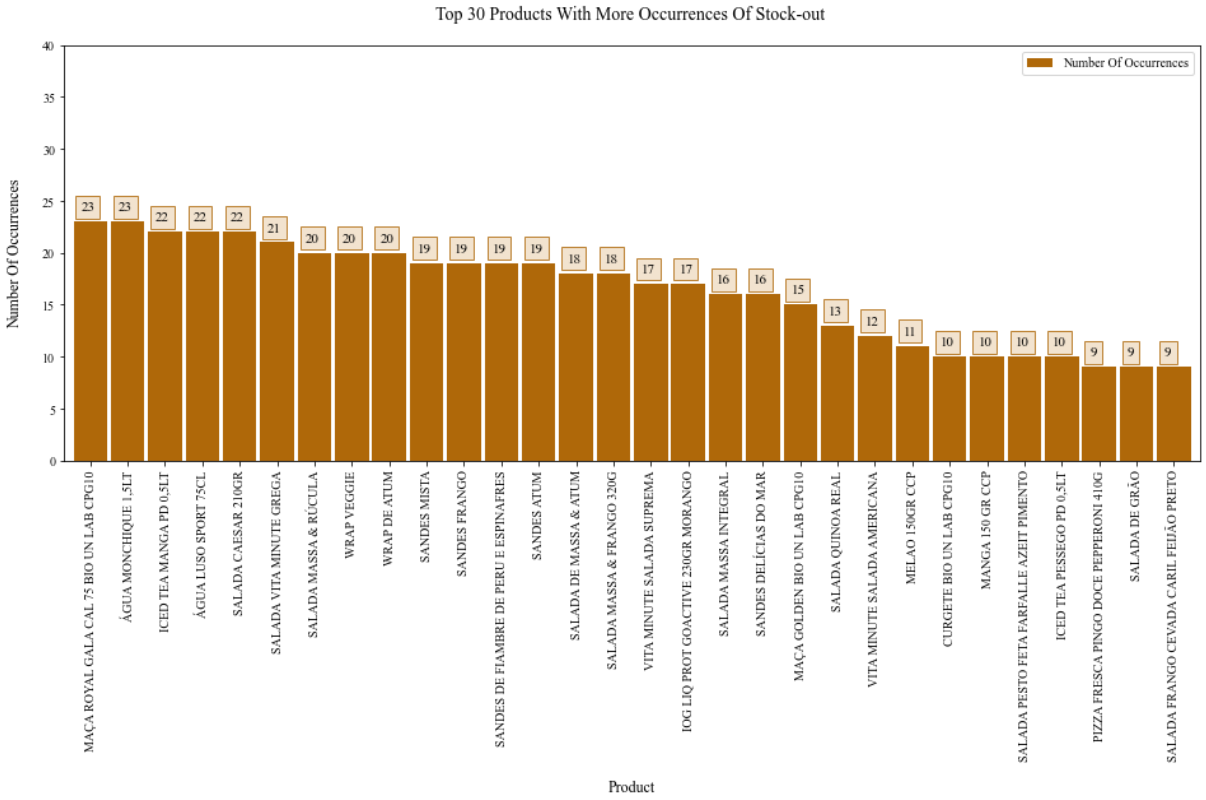


Fig 17– Differences Between Sales Quantity and Capacity, considering Out-of-Stock Products

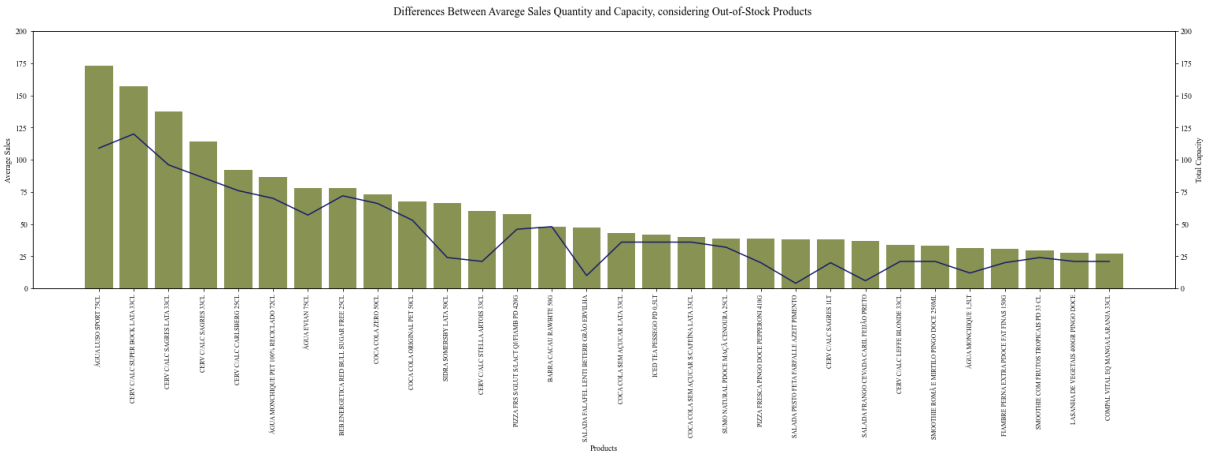


Fig 18– Model Overview

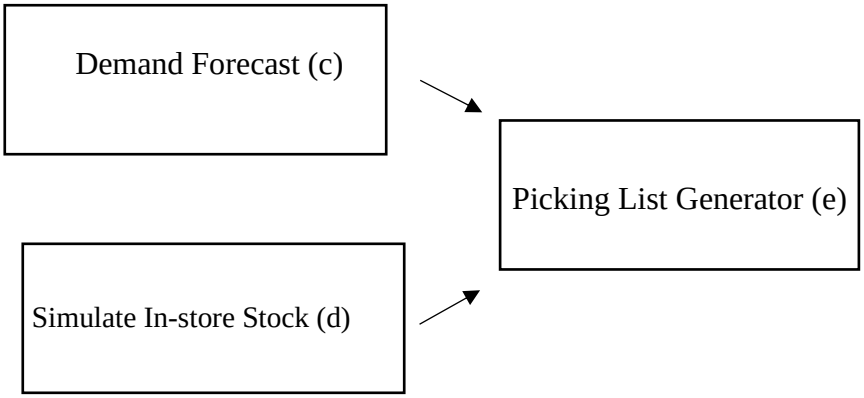


Fig 19– Boxes sold vs Model predictions (without grid search)

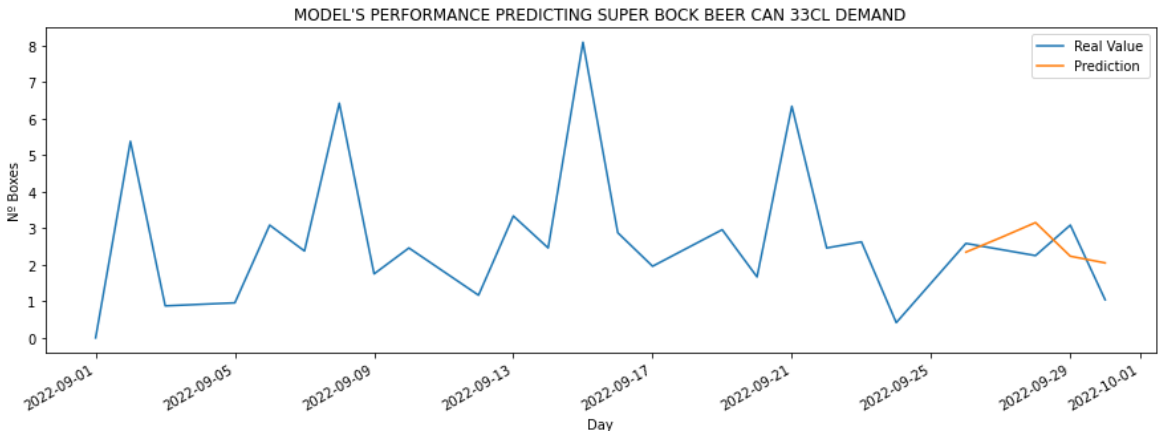


Fig 20– Boxes sold vs Model predictions (with grid search)

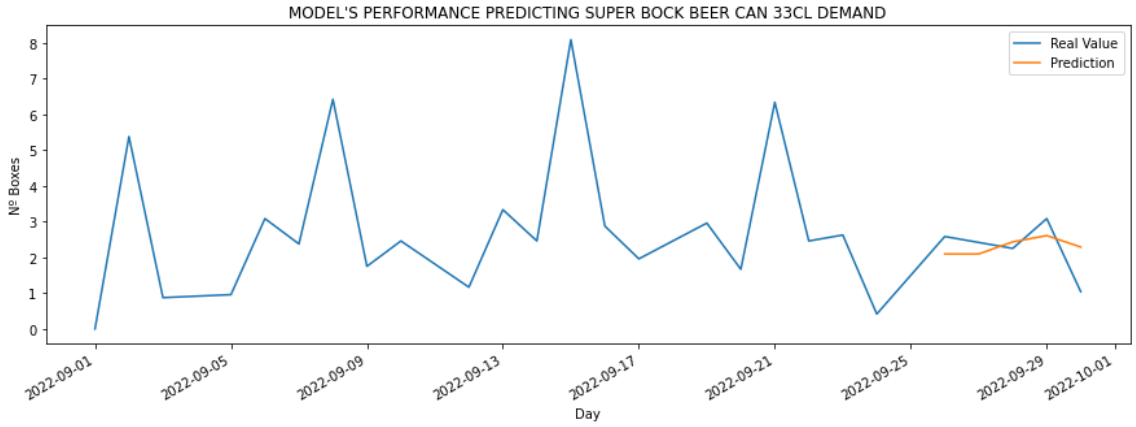


Fig 21– Feature importance

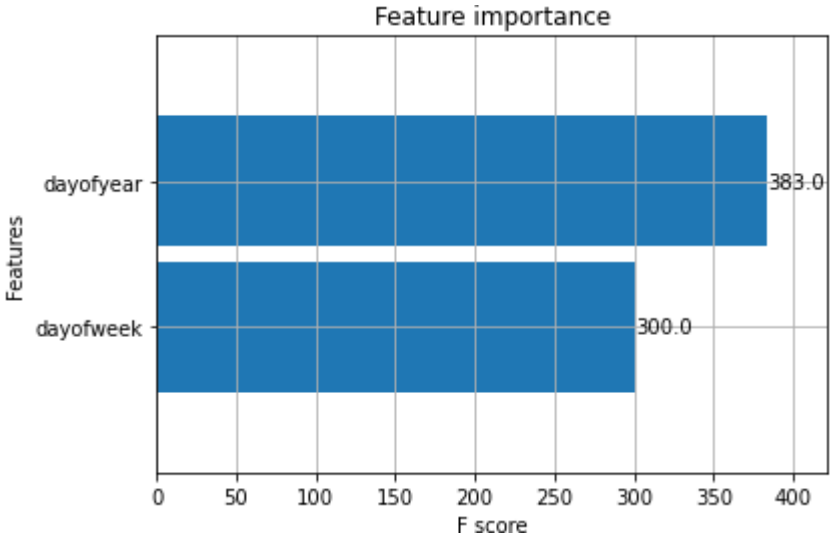


Fig 22– Picking list example

Day	Product	N_Boxes
2022-09-26 00:00:00	COCA COLA ZERO 50CL	7
2022-09-26 00:00:00	ÁGUA MONCHIQUE PET 100% RECICLADO 72CL	5
2022-09-26 00:00:00	COCA COLA SEM AÇUCAR S/CAFEÍNA LATA 33CL	4
2022-09-26 00:00:00	IOG LINDAHLS SÓLIDO STRACCIATELLA 150G	3
2022-09-26 00:00:00	BAR ZERO SNACK BISCOITO PEPIT PROZIS 35G	3
2022-09-26 00:00:00	SUMO NATURAL PDOCE LARANJA MAÇÃ 75CL	2
2022-09-26 00:00:00	ICED TEA PESSEGO PD 0,5LT	2
2022-09-26 00:00:00	IOG LIQ YOPRO DANONE CAFÉ 300G	2
2022-09-26 00:00:00	SMOOTHIE COM LINHAÇA PINGO DOCE 250ML	2
2022-09-26 00:00:00	PUDIM YOPRO CHOCOLATE 180G	2
2022-09-26 00:00:00	IOG LIQ YOPRO DANONE BAUNIL/COOKIES 300G	2
2022-09-26 00:00:00	ICED TEA VERDE ZERO MANGA MARAC PD 1,5L	2
2022-09-26 00:00:00	SMOOTHIE ROMÃ E MIRTILO PINGO DOCE 250ML	2
2022-09-26 00:00:00	BOLACHAS MATINAIS SEM AÇUCAR PD 300GR	2
2022-09-26 00:00:00	WAFFLES PINGO DOCE 165G	2
2022-09-26 00:00:00	SMOOTHIE AÇAÍ MIX PINGO DOCE 250ML	2
2022-09-26 00:00:00	BEB.ENERGETICA RED BULL SUGAR FREE 25CL	2
2022-09-26 00:00:00	DEO ROLL-ON NIVEA MEN COOL KICK	2
2022-09-26 00:00:00	SNACK CHOC KINDER BUENO T2X4 172GR	1
2022-09-26 00:00:00	QJ MOZZARELLA MINI PINGO DOCE 125G	1
2022-09-26 00:00:00	IOG LIQ YOPRO DANONE LIMÃO MENTA 300G	1
2022-09-26 00:00:00	MANGA 150 GR CCP	1
2022-09-26 00:00:00	TORTITAS PURA VIDA S.G ARROZ S/SAL 130GR	1
2022-09-26 00:00:00	TORTIT ARROZ C/CHOC BRAN SAB IOG PD 106G	1
2022-09-26 00:00:00	SNACK APERITIVO DORITOS TEX MEX 120G	1