NOVA IMS Information Management School

MDSAA

Master Degree Program in

Data Science and Advanced Analytics

TRANSFORMING RETAIL DYNAMICS

Exploration of a Machine Learning Approach for Sales Forecasting:
Case Study of an Athletic Digital Retailer

SARRA JEBALI

Master Thesis

Presented as partial requirement for obtaining a Master's Degree in Data Science and Advanced Analytics

NOVA Information Management School
Instituto Superior de Estatística e Gestão de Informação

NOVA Information Management School

Instituto Superior de Estatística e Gestão de Informação

Universidade Nova de Lisboa

TRANSFORMING RETAIL DYNAMICS

Exploration of a Machine Learning Approach for Sales Forecasting: Case Study of an Athletic Digital Retailer

by

Sarra Jebali

Master Thesis presented as partial requirement for obtaining the Master's degree in Data Science and Advanced Analytics, with a specialization in Business Analytics

Supervised by

Roberto André Pereira Henriques

November, 2023

STATEMENT OF INTEGRITY

I hereby declare having conducted this academic work with integrity. I confirm that I have not used plagiarism or any form of undue use of information or falsification of results along the process leading to its elaboration. I further declare that I have fully acknowledged the Rules of Conduct and Code of Honor from the NOVA Information Management School.

Sarra Jebali

[Lisbon, 12-11-2023]

DEDICATION

In dedication to my dear parents, Ezzeddine Jebali and Olfa Ferjani, whose love, support, and prayers have kept me going. Thank you for your efforts, encouragement, and all the sacrifices you have made for me to succeed.

I also dedicate this thesis to my dear friends and siblings. Thank you for all your support. I will always appreciate you being in my life.

ACKNOWLEDGMENTS

I wish to express my sincere gratitude to Professor Roberto Henriques, my academic supervisor for the engaging Machine Learning class that sparked my interest in choosing this thesis topic, as well as his guidance throughout this project.

I would also like to extend my sincerest appreciation to my professional supervisor, Mr. Jamie Ogier for granting me the freedom to choose my project focus and to the whole team especially Lisa Hull, Charlotte Bateson, Freya Vermander, Juliette Louatron, and Santiago Ruffini for their support and warm welcome.

ABSTRACT

This project studies the possibility of retail sales forecasting through an artificial intelligence approach, using an in-depth analysis of a prominent sportswear company as a case study. To build a robust foundation, a thorough review of past research in this field, the approaches adopted, and the results reached was made. We extracted real-case observations and addressed their issues to ensure they were satisfactory for a machine-learning model application. Finally, granular forecasts were built starting on a product level and then aggregating to derive an overall forecast for the whole retailer. This project aims to highlight the importance of predictive analytics in decision-making and emphasize the ongoing need for improvement in dynamic retail through collaborative efforts between data science and business intuition. It serves as a testament to the potential of enhancing sales forecasting methodologies, paving the way for more accurate and adaptive predictions in the future.

KEYWORDS

Retail Forecasting; Machine learning; Artificial Intelligence; Digital Sales

TABLE OF CONTENTS

1. INTRODUCTION 1
2. LITERATURE REVIEW 3
   2.1. Introduction 3
   2.2. Forecasting in Retail 3
   2.3. Forecasting Granularity 3
   2.4. Retail Forecasting Factors 4
      2.4.1. Seasonality 5
      2.4.2. Discounts and Promotions 5
      2.4.3. Online Product Reviews 5
      2.4.4. Other Factors 5
   2.5. Approaches to Forecasting 6
   2.6. Retail Forecasting Challenges 7
   2.7. Conclusion 7
3. PROJECT FRAMEWORK 8
   3.1. Internship Context 8
   3.2. Business Context 8
   3.3. Problem Statement 8
   3.4. Solution Envisioned 9
   3.5. Business Value 9
4. METHODOLOGY 10
   4.1. Methodological Framework 10
      4.1.1. Tools 10
      4.1.2. Data 11
      4.1.3. Approach 11
   4.2. Applied Methodology 13
      4.2.1. Business Understanding 13
      4.2.2. Data Collection 14
         4.2.2.1. Table Identification and Transformation 14
         4.2.2.2. Data Integration and Table Joins 18
      4.2.3. Data Understanding 20
         4.2.3.1. Data Description 20
         4.2.3.2. Assessing Data Quality 20

4.2.3.3. Data Visualization 22
4.2.4. Data Preparation 24
4.2.4.1. Data Consistency Improvement 24
4.2.4.2. Missing Values 24
4.2.4.3. Outliers 25
4.2.5. Data Preprocessing 25
4.2.5.1. Feature Engineering 25
4.2.5.2. Feature Selection 26
4.2.5.3. One-hot Encoding 27
4.2.6. Model Implementation 27
4.2.6.1. Data Split 28
4.2.6.2. Model Building 28
4.2.6.3. Model Evaluation 28
4.2.6.4. Hyperparameter Tuning 29
4.2.6.5. Feature Importance 29
4.2.7. Project Deployment 30
4.2.7.1. Scenario Mapping 30
4.2.7.2. Project Workflow 31
5. RESULTS AND LIMITATIONS 33
5.1. Results and Findings 33
5.2. Limitations 33
5.3. Future Projects 34
6. CONCLUSION 35
7. REFERENCES 36

LIST OF FIGURES

Figure 1 - Model of the CRoss Industry Process for Data Mining 12
Figure 2 - CTE Created Tables 19
Figure 3 - Numerical Data Description 21
Figure 4 - Variation of Demand throughout the Year 22
Figure 5 - Average Demand for each day of the month 23
Figure 6 - Average Demand for each day of the week 23
Figure 7 - Transaction Count per Gender 23
Figure 8 - Total Demand per Gender 23
Figure 9 - Total Count per Division 24
Figure 10 - Total Demand per Division 24

LIST OF TABLES

Table 1 - Numerical Features Extracted 15
Table 2 - Additional Columns Created 16
Table 3 - Attribution Columns Extracted 16
Table 4 - Data Columns Categorization 20
Table 5 - Excel Input Sheet Columns 31

LIST OF EQUATIONS

Equation 1 - MAPE 29
Equation 2 - MAE 29
Equation 3 - SMAPE 29
Equation 4 - WSMAPE 29
Equation 5 - Accuracy 29

1. INTRODUCTION

In the evolving field of global commerce, online retail has experienced significant innovation and expansion. Marked by its dynamic nature and constant shifts in consumer preferences, it has created an interesting ground for competition (Ferrera & Kessedjian, 2019). To stay ahead of the curve, retailers have long sought accurate methods for sales forecasting, and in the past few years, we witnessed the implementation of advanced analytics in this process (Lalou, Ponis, & Efthymiou, 2015). However, It is crucial to recognize that this digital revolution in retail did not happen overnight. In the earlier chapters of retail history, the industry operated on traditional models lacking sophisticated IT infrastructure. Retail companies, based on physical stores, lacked the technical skills required for advanced IT solutions. Consequently, the adoption of cutting-edge technologies, especially in areas like machine learning-based forecasting, has gradually evolved (Har et al., 2022).

A few decades back, the retail landscape mainly focused on physical stores, guided more by intuition than by data-driven decisions. However, with the rise of the digital age, retail began to shift, leading to unprecedented growth and transformation. The digital changed significantly, requiring retail enterprises to reimagine their operations. The traditional reliance on intuition and manual processes led to a more data-centric approach. The need for advanced IT solutions, particularly for forecasting purposes, became apparent in this transformation. However, As retail companies underwent this transformation, they faced a dual challenge: adapting to the demands of the digital marketplace and simultaneously constructing the technical infrastructure required for advanced data analysis (Shankar et al., 2021).

The absence of a historical foundation in IT solutions for retail poses unique challenges, especially when exploring advanced methodologies like machine learning. This thesis tackles this forecasting topic by considering the example of a digital athletic retailer perspective and leveraging machine learning techniques to attempt and forecast sales with precision. As we explore the complexities of this process, we will focus on the historical context and acknowledge the challenges retail enterprises face in integrating and optimizing these sophisticated IT solutions.

This project aims to optimize the sales forecasting process, promoting informed inventory management and strategic marketing campaigns. We will focus not only on the model application but on the whole process adopted starting from data collection, and data preparation to feature engineering bringing the data to a state of readiness for predictive modeling. This task will be challenged by numerous potential issues, including the complexity of managing a vast product catalog encompassing numerous product attributes, considering multiple factors, as well as understanding the complexity of retail data.

Since sales forecasting can highly influence retail business success, our main goal throughout this project would be to achieve the highest accuracy possible. We also aim to better understand the dynamic field of retail forecasting and be able to offer valuable insights to all the stakeholders in the digital retail landscape. Our final objective would be maintaining and sustaining this project. Since the retail field is known for its unpredictability in consumer

behavior, we will focus on building a project that is adaptable to the changes in the market, promoting innovation and revolutionizing old processes.

2. LITERATURE REVIEW

2.1. INTRODUCTION

Online retail has experienced unprecedented growth in recent years. In 2023, the global e-commerce growth rate reached almost 9%, which resulted in 5.8 trillion dollars of global e-commerce sales (Statista Search Department, 2023). With this rapid expansion that is still expected to grow more in upcoming years, assessing the business performance, efficiently planning, and predicting in advance the sales became increasingly crucial. According to Fildes et al.(2022), forecasts are a vital component of many decision-making processes and are significantly and directly linked to higher profitability. This approach became imperative, particularly when optimizing discount and markdown strategies, as these areas are prone to a higher frequency of errors. In the context of price promotion campaigns, Wolters and Huchzermeier (2021) found that demand forecasts show more inaccuracies, resulting in significantly higher forecast errors when compared to the total sales volume.
This literature review starts by introducing the concept of retail forecasting in a general sense. However, our main focus throughout this project will be on digging deeper into its application in digital retail specifically. We will also explore the level of detail in forecasting, the different factors, and various methods for making forecasts, particularly emphasizing the more advanced approaches and the associated challenges.

2.2. FORECASTING IN RETAIL

The digital revolution in the retail sector has given companies easier and faster access to an abundance of information about customers preferences and behaviors. This has given them the ability to make proper usage of this data, allowing them to respond to the dynamic marketplace much more rapidly than was possible before. And to effectively meet the changing customer demand, it became necessary to remodel traditional retail processes and integrate the massive data flow in decision-making. Therefore, forecasting became an essential part of the systems, playing a vital role in ensuring the flexibility and adaptability of retail businesses in the current dynamic marketplace. Forecasting, as a concept, involves predicting future customer demand and sales patterns based on historical data and various influencing factors (Petropolous, et al., 2022). Accurate forecasts can help retailers optimize inventory management, pricing strategies, and resource allocation, ultimately enhancing overall operational efficiency. However, given that many decisions are made at a detailed level, such as setting prices or managing inventory for individual products, the scope of the challenge becomes considerably larger. According to Brian Seaman (2018), if the sales of an online business are forecasted by geographical location, the number of forecasts can exceed 1 trillion on average. In light of this, to manage this process smoothly, systems must be developed to run these forecasts swiftly and efficiently.

2.3. FORECASTING GRANULARITY

The level of granularity in retail forecasting refers to the degree of depth and precision of predictions, and the level of detail at which forecasts are generated. This extent of granularity is determined by the business needs and figuring out the right balance is not an easy task. At a high level of aggregation, forecasts may include general categories, regions, or time frames. At its most granular level, forecasts can consist of product details, used to target more complex retail operations, such as specific product sales or stores demand. Achieving the right level of

granularity allows retailers to adapt easily to the dynamic market changes through efficiently managing inventory, pricing strategies, and building a solid understanding of customer preferences. Fildes, et al. (2022) have defined three different dimensions when it comes to product-level demand forecasting:

The first dimension is the time dimension. The demand forecasts can be needed at different time granularities depending on the managerial decisions. In general, the forecasting time granularity decreases as the decision level rises, from the operational to the strategic. For example, while digital sales may rely on an initial estimate of total seasonal sales that is updated just once mid-season, store replenishment forecasts may require forecasts at a daily level. On another hand, promotion planning, and allocation planning may require forecasts at a weekly level.

The second dimension to consider is the product dimension. When forecasting demand for products, different approaches are used. On a high level, the retailer can consider just the category or even the division of the product. On a more detailed level, it can go as far as forecasting for each product while taking into account the differences in color and size. In our case, the latter will be considered, which will create issues with a large dataset. A solution for this high-volume data that has been found to be useful is clustering. It has been proven that cluster analysis can improve forecast performance (Boylan, Chen, Mohammadipour, & Syntetos, 2014). For example, we can cluster products based on whether they have similar demand patterns rather than according to their similar characteristics (color, category, franchise, etc.) and reduce a large amount of data while also maintaining or even increasing the forecast accuracy.

The final dimension is the supply chain dimension. Often, retailers find themselves in need of forecasts at different levels in the supply chain. Not only do planners need data forecasts at the current level, but they often want the knowledge of what to expect at the different distribution centers, on a chain level, or in factories. This information becomes useful when discussing preordering, supplier negotiations, and updating the manufacturing decisions. Although in digital forecasts no retail stores are accounted for, digital retail still requires updates on the manufacturers, the wholesalers, or other digital partners.

For this business case, the scope of the problem will be on a digital level which will make the process simpler compared to store forecasts. The reason for stores' complexity is that many other factors should be accounted for as opposed to the digital business, for instance, the different stores for the retailer, their locations, and the different customer segments in each of those locations. However, our goal will also be to forecast at the most granular level possible. We will be considering each product's daily sales separately and account for its detailed characteristics such as color. We will be aiming at generating forecasts for a large number of products over a short forecasting horizon which can come with its complexities, but it will also support the survival and growth of the retailer since many operational decisions are related directly to those forecasts.

2.4. RETAIL FORECASTING FACTORS

Numerous factors might have an impact on a product's observed sales and potential demand. We can categorize these factors into three elements. Factors that would be known in advance such as an upcoming event, the different seasons, and holidays. Factors that the retail planner will decide, for instance, pricing and promotions. Finally, some factors that are unknown but can be predicted to be included in the forecasts; these elements include regional and national economies, competitive behavior, and even weather. Apart from the historical sales, which are the main drivers for retail demand forecasting, we will go over some of the important factors that were included in other research papers.

2.4.1. Seasonality

In the retail industry, product sales data have a strong seasonality and usually contain multiple seasonal cycles. Depending on the specific nature of the business, various times throughout the year can exert varying effects on demand, either augmenting or attenuating it. For the fashion industry, for example, holidays such as Christmas or the back-to-school season can have a huge impact on the demand. For this reason, forecasting models must be able to handle the multiple seasonal patterns. However, one challenge that can be encountered is the regularity of certain events. While some holidays, like Christmas or the Fourth of July in the US, occur at fixed intervals and can be considered part of seasonality, other holidays do not follow the Western calendar and have varying dates each year. For instance, Jewish or Muslim holidays, such as Ramadhan or Eid follow lunar calendars (Fildes, Ma, Kolassa, 2022). To account for these unpredictable holidays, dummy variables can be employed to model them (Cooper, Baron, Levy, Swisher, & Gogos, 1999). Nonetheless, effectively modeling seasonality is a challenge for the models whether they are nonlinear or traditional.

2.4.2. Discounts and Promotions

The original price and the price changes throughout the forecasted period are important factors that should be accounted for, especially when focusing on short-term forecasts. The effects of the promotions and discounts will vary heavily depending on the approach adopted by the company. While some promotions occur regularly throughout the year such as Cyber Week and black Friday, others occur on random periods when the business sees the need for it such as trying to liquidate the inventory stock and therefore, they can have different effects that are not captured just by the unit price. The impact of the price changes can also impact other demands, for instance, a promotion on one item can affect the sales of another item, or what is known as cannibalization, and incorporating this phenomenon in the forecasts has proven to improve accuracy substantially (Srinivasan, Ramakrishnan, & Grasman, 2005).

2.4.3. Online Product Reviews

In digital retail, online product reviews have been found to have a big impact on future sales. Consumers often take product reviews into consideration when making their own purchasing decisions since they believe these reviews are a voluntary expression of consumers' experiences and beliefs (Zhao, Yang, Narayan, & Zhao, 2013). However, product reviews are mainly textual data, which cannot be used directly in a sales forecasting model. Therefore, two approaches can be adopted. The first is to include only the numerical ratings since the retailers often adopt a simple five-star rating mechanism (for example, Amazon) which is considered as a simple and fast method. The second approach that Lyu and Choiis (2020) adopted is to incorporate the reviews using tools such as text mining and sentimental analysis, which allows retailers to also use them as explanatory variables to identify the reasons for the drop or rise in sales.

2.4.4. Other Factors

Sales can be affected by various factors, often ones that we do not have any control over or know about. Abnormal occurrences such as health concerns or natural disasters can cause random disturbances affecting retail sales. Other factors we don't control can still be forecasted and included in the models. For example, Steinker, Hoberg, and Thonemann (2017) studied the

effects of weather effects on sales in an online fashion retailer and concluded that including weather factors substantially increased the forecast accuracy by almost 10%

2.5. APPROACHES TO FORECASTING

In recent decades, significant attention and work have been dedicated to advancing and refining demand forecasting techniques within the retail sector. Two different schools have emerged in sales forecasting: the linear and non-linear approaches. Out of the linear approaches, one of the most basic demand forecasting models on a product level is the univariate method. This technique would only use the past sales history as a variable, and it can vary from the traditional time-series approaches such as a simple moving average to an ARIMA or a state space model. In comparing these methods, research has found that simple time-series techniques performed well for periods with no product promotions, however, methods that included the promotion factor improved the accuracy substantially for periods with promotion (Ramos & Fildes, 2017). In that sense, the univariate method can be useful for higher granularity demand forecasting or for products with low promotional activity. Another exceptional linear approach is the econometric model. In research done by Bechter and Rutner (1978) comparing the performance of ARIMA and econometric models, it was shown that simple single-equal economic models had an advantage and generally performed better than ARIMA models. Although, it is worth noting that the mixed model had a better record than any of the other two models. For the non-linear methods, the most common non-linear approach that received extensive attention is the neural network model. With its way of processing information that is inspired by the human brain, NN models are able to learn from the data, identify a pattern, and make conclusions for the future. With the appropriate architectures, NN is capable of simulating a wide variety of non-linear behaviors and approximating any type of problem with a high level of accuracy compared to other models (Chu, Zhang, 2003). When building this model, a few assumptions would be needed, since the model is adaptively formed with real data. This flexibility has made the NN model an attractive tool for many sales forecasting retailers especially since the data is often abundant and endures many changes in the real-world environment. When comparing both of these schools, one of the major limitations of the traditional methods is the assumption of linearity. For them to be applied, the analyst must specify the model form without genuine knowledge about the actual complex relationship in the data. For that reason, some researchers have found that standard time series models are sometimes inadequate for forecasting sales, after identifying evidence of nonlinearity and volatility in retail sales (Chu & Zhang, 2003). However, if the linear models can generate high-accuracy forecasts, then they should be considered as the preferred models over the more complicated ones since they have the practical advantage of easy interpretation and implementation. In general, there is no unique model that performs best across all scenarios (Fildes, Ma, Kolassa, 2022). The research has shown that traditional time series models with a stochastic trend, such as exponential smoothing and ARIMA, performed well when the macroeconomic conditions were relatively stable. However, when factors are highly volatile with rapid changes then NNs have been claimed to outperform the linear methods (Alon et al., 2001).
In our project, we will also be focusing on a non-linear method when building the model. However, our approach will focus on tree-based models as we are still looking for some explainability to the way the different factors affect sales, a feature that is not possible with NN models that are considered black-box models.

2.6. RETAIL FORECASTING CHALLENGES

Retail sales forecasting can impose many challenges since it is dependent on a wide range of volatile factors. No matter how much the analyst will try to account for the factors, many other unpredictable elements highly influence retail sales but can not be included in the forecasts. These factors can include a terrorist attack or a pandemic such as COVID-19 that shook the whole retail process. Moreover, the forecasting models may require more information than what is readily available. For instance, inventory always poses an issue in forecasting due to how hard it is to always keep track of its levels, and its systems are notoriously inaccurate (DeHoratius & Raman, 2008). When we forecast on a low granularity, we would need inventory levels for every product, and keeping track of every change in the inventory shelves is costly for retailers. Another issue that is also only faced in lower granularity is the metrics used to evaluate the forecasts. On a product level, some of the standard forecast accuracy metrics may become misleading or even unusable. For example, a very commonly employed accuracy metric, the mean absolute percentage error (MAPE), can only be used on sufficiently highly aggregated data. Its formula includes dividing by the actual demand and as the granularity gets lower it becomes more undefined when some products have zero sales on a particular week for example. Therefore, choosing the right evaluation metric will require more research and assessment.

2.7. CONCLUSION

This literature review explored the main papers that covered retail forecasting and the different approaches adopted. The complexity of this problem is a common element that was addressed by all the researchers. The choice of the level of granularity, the factors to be included in the model, the approach to adopt, and the method by which the model is evaluated will influence the success of the forecasts. Despite its complexity, retail sales forecasting has proven to drive businesses forward. The forecasts provide essential insights allowing companies to make more data-based decisions about production planning, inventory management, supply chain operations, and promotions strategies.

3. PROJECT FRAMEWORK

In this section of the thesis, we will contextualize the project idea by diving into the problem and need behind it, the envisioned solution, and finally its value and importance to the business.

3.1. INTERNSHIP CONTEXT

This practical project is a mandatory component of my second-year internship at NOVA IMS, aimed at achieving a Master's degree in Data Science and Advanced Analytics. During this internship, which spanned from September 1, 2022, to August 31, 2023, I was part of a prominent sportswear company known for its global impact and innovative contributions to athletic gear and apparel. I joined their planning team, focusing on the analytics aspect of their operations. To ensure confidentiality and protect sensitive company information, I will obscure specific details and transactional Point of Sale (POS) data. However, the results presented will retain the same pattern detected in the original data, allowing us to draw meaningful conclusions.

3.2. BUSINESS CONTEXT

In this retail company, an inventory management department for the digital part of the business plays an important role. Its objective is to keep healthy levels of inventory stocks and ensure an optimal balance within the organization's warehouses. This goal is essential to meet the ever-changing demands of the market and to navigate the dual challenges presented by inventory surpluses and shortages.

In the first scenario, when inventory levels exceed expected demand, the company faces the issue of managing surplus stock and the associated warehousing costs. To solve this, the company has two strategic options. Firstly, they may choose to cancel future orders, thereby reducing the upcoming inventory. Alternatively, they can adopt a dynamic pricing strategy, incorporating markdowns and promotional periods to stimulate and increase customer demand.

The second scenario unfolds when market demand surpasses available inventory. In this case, the department suggests ordering more materials to bridge the gap and meet the surging demand. In essence, the general goal of the Inventory Management department revolves around finding the perfect inventory levels balance, ensuring they neither fall short nor surpass demand, thus optimizing the efficiency of the entire supply chain.

3.3. PROBLEM STATEMENT

To achieve business objectives, accurate forecasts are a crucial component of effective decision-making. However, with the current tools available in the company, forecasting is increasingly challenging. The usual forecasts done in this aspect often stem from a purely financial standpoint, thereby lacking the essential granularity required by our inventory management team. Furthermore, they often fall short of addressing the full scope of our business landscape. For instance, certain forecasting models within the team tend to be more explanatory than predictive, leaving us with limited foresight. This lack of visibility leads the team to often make decisions blindly and without full context. Consequently, we find ourselves without a comprehensive understanding of the effectiveness of such decisions, leaving questions unanswered, such as the impact of price adjustments on product performance.

3.4. SOLUTION ENVISIONED

To enhance inventory-related decisions, a weekly granular bottom-up demand forecast is required. Such a forecast, providing insights into anticipated demand for the upcoming week, empowers our team to make well-informed choices concerning inventory management. Moreover, the team also hopes that by having visibility on the forecasts they can also assess the impact of their decisions on the sales and see how the alterations in price can improve their planning. This solution is attained by leveraging material-level historical data, analyzing it, and using it to get accurate weekly predictions. The approach that will be followed will rely on modern forecasting methods such as machine-learning solutions to handle this problem. The purpose of this project is also to assess how the different factors affect the demand, therefore, assessing the historical sales alone won't be sufficient. As proven by various research, multiple variables can significantly impact product sales and we will try to account for those different elements. The objective is to improve the conclusions reached through the model. For instance, when we observe that increasing the price of a product doesn't lead to a decline in sales, we cannot immediately conclude that the product is price inelastic. Various other factors, such as inventory levels or digital outreach, could be influencing sales stability. In conclusion, the envisioned model serves as not only a forecasting tool but also as a comprehensive mechanism for evaluating how our team's decisions impact sales, accounting for different variables to provide a full understanding of the dynamic retail landscape.

3.5. BUSINESS VALUE

The implementation of a simplified, automated, and more accurate forecasting capability will allow for scalable and data-driven decision-making within this retail business. It will equip retailers with the information needed to optimize inventory levels, reduce carrying costs, and enhance overall supply chain management. These operational improvements, in turn, yield substantial cost reductions and improve financial performance. Additionally, precise sales forecasts allow retailers to better plan inventory management and pricing strategies, which can further lead to enhanced customer satisfaction and revenue growth. Through this tool, retailers gain a deeper understanding of how different elements are affecting the products' sales, enhancing the quality of their decision-making. Additionally, having an automated tool will save the team time, costs, and resources and improve overall operational efficiency.

4. METHODOLOGY

In this chapter, we will explain the methodology that will be followed throughout this project from a theoretical perspective and a practical one. In the first section, we will describe the tools including the different environments that will be used, the data collection, and the modeling approach followed. In the second section, we will take the example of a digital retailer as a user case illustrating the real-world application of the methodology.

4.1. METHODOLOGICAL FRAMEWORK

4.1.1. Tools

For this project, we will be working with two different languages for two different purposes. First, we will be using SQL, a specialized programming language, to retrieve data from multiple databases, manage it, and perform the necessary changes to make it ready for the next phase of modeling. This step will be carried out through Snowflake, as described on their official website Snowflake (n.d.), it is a cloud-based data warehousing platform allowing users to store, manage, and analyze data efficiently simplifying data management.

The second language is Python which will be used to deploy a prediction model. This programming language is the number one language and the most used tool for data scientists according to KDNuggets's annual poll. This is because Python is an easy-to-use tool, lightweight, efficient at executing codes, and multifunctional. We will use it to further understand the data, visualize it, alter it, and utilize it to train a model and assess it. This will be feasible through the variety of libraries that can be used which are described below (GeeksforGeeks, n.d.):

- Pandas: a powerful, fast, flexible, and easy-to-use tool for data manipulation and analysis.
- NumPy: a numerical computing tool that handles a variety of mathematical and statistical operations. It also handles mathematical operations on multidimensional arrays.
- Matplotlib: a comprehensive visualization library for creating static, animated, and interactive charts for the data.
- Seaborn: it is a Matplotlib-based Python data visualization library. It is great at creating visually appealing and insightful statistical graphics.
- Scikit-learn: the most popular library for machine learning in Python due to its simplicity and efficiency.
- Statsmodels: it conducts statistical tests, data exploration as well and estimations of many models.

I will work on Python on Jupyter Notebook through Anaconda, a large distribution platform of programming languages, such as R and Python. As for Jupyter Notebook, it is an open-source web application for Python editing. It enables the user to create and share documents containing equations, visualization graphs, and narrative text. The output of this notebook is a human-readable file with a customized interactive dashboard (Jupyter, n.d.).

4.1.2. Data

First, when deciding what data to use we had to consider which variables affect retail sales. According to the different research done in this aspect that was mentioned in the literature review, many different elements can be a factor in the rise or fall of sales whether directly or indirectly. We had to think carefully of not only what we wanted to include but also within the constraints posed by the data system given that the data warehousing in this company is not well structured. The company, being a large-scale retail enterprise, operates with a multitude of warehouses, databases, and data collection methods across different departments. This operational diversity resulted in discrepancies in data values, granularity, and structure. When assessing the data and its validity we decided to extract these variables:

- Historical Sales: This variable acts as the cornerstone of our analysis, providing essential insights into past sales trends. By studying historical sales data, we can identify patterns, seasonality, and fluctuations that serve as a foundation for predicting future sales.
- Product Attributions: Understanding the characteristics and attributes of products is crucial. Product attributions encompass details like name, color, category, franchise, and product specifications. This information aids in segmenting products and tailoring forecasting models to different product groups.
- Markdown Change: Detecting when the price of a product changes will be an important factor in sales predictions. We will monitor markdown changes by date to allow the model to detect if the new product price fell from its original price or a previous markdown in comparison to prior dates.
- Inventory: Inventory levels are a vital factor in sales forecasting. Knowing the stock levels helps in estimating the availability of products and, consequently, their sales potential. It allows for more precise predictions, particularly during peak demand periods or clearance sales.
- Digital Traffic: Monitoring digital traffic data, such as website visits for each specific product provides valuable insights into consumer behavior. The correlation between digital traffic and sales can be leveraged to adjust forecasts based on online activities.
- Promotion Indicator: The presence of promotions can significantly impact sales. It allows us to see how special promotions and discounts can affect sales.
- Retail Calendar: retail companies do not follow the conventional calendar. Instead, they have their unique retail calendars with indicators for the retail season and retail week numbers. Incorporating this table with the rest ensures accurate time alignment and facilitates precise tracking of sales patterns, trends, and seasonality.

We believe that each of these variables serves a distinct purpose in enhancing the accuracy and effectiveness of the forecast. However, collecting all this data and joining it into one single dataset won't be an easy task due to the warehousing issue mentioned earlier. We will get more into the technical aspect of it in the data collection section.

4.1.3. Approach

For this project, we will be following the CRISP-DM approach. CRISP-DM stands for CRoss Industry Standard Process for Data Mining, and it's a six-phase process model that represents the data science life cycle (Wirth & Hipp, 2020).

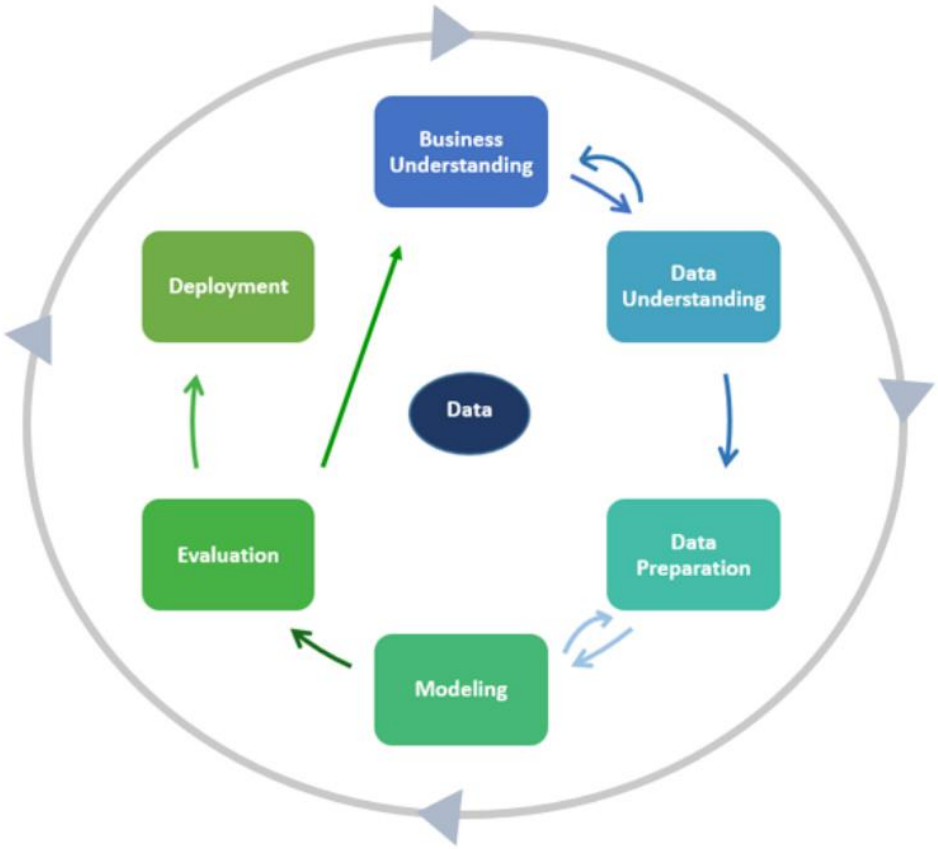

Figure 1 - Model of the CRoss Industry Process for Data Mining

In the following, we will go through each phase briefly:

- **Business Understanding:**

This first step consists of fathoming the business problem and its objectives and purposes. It represents the foundation and basis of the whole project. During this phase, we identify the key variables related to the business problem, classify the lead indicators, formulate problem hypotheses, and set an introductory plan. This is known as converting a business problem into a data mining problem.

- **Data Understanding:**

This phase succeeds in the data collection phase. And it consists of activities to get familiar with the dataset. We identify data quality problems, discover new insights, and detect interesting patterns to form other hypotheses from hidden information. This is why this step has a two-way link with business understanding.

- **Data Preparation:**

This phase consists of transforming and manipulating raw data to construct a dataset fit for the modeling tools. This phase can involve data cleaning, feature selection, feature engineering, and data transformation, based on the needs of the model we are going to use. For instance, some algorithms require categorical one-hot encoding or feature normalization. This step's purpose is to meet the needs of each model being used.

- **Modeling:**

During this phase, modeling algorithms are applied, improved, and selected. Since several techniques can work for the same problem, we must try different approaches and choose the best one. There is a strong link between this part and the previous one because we might realize certain algorithm requirements while working on the model.

- **Evaluation:**

Upon reaching this stage, we should have built one or more high-quality models. However, before proceeding with the deployment, it is essential to evaluate the model and review the steps executed in constructing it. An important step is to check if there is a business issue that has not been considered.

- **Deployment:**

Building the model is not the end of the project. The knowledge gained and results reached need to be arranged and displayed in a way understandable for the user. This phase can go from writing a simple report to implementing a data mining process.

4.2. APPLIED METHODOLOGY

4.2.1. Business Understanding

The purpose of this initial phase is to work on a more in-depth analysis of the business problem at our hands, from a technical perspective. As mentioned in the Project Framework chapter, the goal of this project is to forecast weekly sales on a product level. In technical terms, our goal would be to achieve the highest level of accuracy possible.
We will begin by assessing the validity of our assumptions. One of the hypotheses that we established is the importance of various contributing factors in retail sales forecasting. We believe that elements such as inventory levels, web traffic, pricing strategies, and promotional events display significant influence over the sales performance of individual products. Despite the certainty of the business specialists on the accuracy of this assumption, we will still need to validate it through a data-driven analysis.
The second thing that we will address is the magnitude of the problem at hand. These forecasts would have to be run every week, and with a multitude of products, numerous influencing variables, and frequent fluctuations in customer demand, this challenge is of substantial proportions. However, it is promising when addressed effectively. A pivotal decision revolves around the granularity of forecasting, whether to predict sales for each product individually or collectively for the entire inventory. This decision has significant effects on model complexity, computational requirements, and forecasting precision. Furthermore, we need to consider the

feasibility of training a new model each week or leveraging a previously trained model. Achieving a balance between granularity, computational resources, and model maintenance is a strategic challenge.
Finally, to tackle the vast volumes of data inherent to the retail sector, the selection of an appropriate model is of paramount importance. We need to focus on choosing a model that exhibits the capacity to efficiently process and analyze extensive datasets, adapts to evolving patterns, and provides accurate forecasts.

4.2.2. Data Collection

In this particular project, the data collection phase is anticipated to be the most time-consuming and critical element. The complexity arises from the warehousing issues previously mentioned. Within a sizable retail company that lacks a robust IT infrastructure, data collection can be somewhat chaotic. The Various departments employ different data collection systems, leading to inconsistencies in data quality, granularity, and structure across different platforms. To address this challenge, a significant portion of our effort will be directed toward creating a robust and reliable dataset tailored to the specific needs of this project. These databases are housed in Snowflake warehouses, and the data will be extracted from there, with subsequent manipulation using SQL.
As previously discussed, our primary objective is to combine the identified variables into a unified and coherent dataset. This will involve a multi-step process, commencing with the identification of the essential tables required for each variable. Subsequently, we will determine the necessary data transformations and manipulations to achieve the desired output of joining all the tables into a coherent one. This approach is fundamental to ensuring the quality and integrity of the dataset and laying a solid foundation for the subsequent phases of our project.

4.2.2.1. Table Identification and Transformation

Historical Sales:

To get the historical sales, we will be drawing from a table that records every transactional sale spanning the entirety of the company's product catalog. This table encompasses 72 columns that include both descriptive columns showcasing the product characteristics and numerical columns for the number of sales in units and dollars. This table includes the primary numerical columns that will serve as the foundation for our analysis. Therefore, we will focus, in a little bit of detail, on the key modifications that significantly impact our approach.
The first thing we checked for in this dataset is the level of granularity. We noticed that the dataset records data at an exceptionally granular level, capturing every individual transaction for each product, even specifying its size. This means that each product can have multiple transactions per day. To streamline our analysis and make it more manageable, we aggregated this data, elevating the granularity to a daily basis for each product, while omitting size-related distinctions. The decision to refrain from aggregating data at the size level is related to future analysis. This level of granularity, while valuable for the forecast, would potentially introduce complexities during subsequent stages of data manipulation. Other tables in our dataset, such as inventory and traffic data, are structured and documented at a different, non-size-specific level of granularity which would make it difficult to join them with the rest of the tables.
The dataset also has product attribution, however, it's essential to note that this information was collected in a different geographical location, incorporating specific and distinct product attributions for that region. Therefore, while we will incorporate some attributions from this

source, we have chosen not to utilize these columns for the majority of characteristics. Instead, we are relying on other alternative datasets to provide the necessary attributes. This decision is aimed at mitigating potential mapping issues and ensuring the reliability and consistency of our dataset within the team. From the characteristic columns, we will keep the product name, its color description, and a line of business (LOB) column indicating if the transaction was made on a product at its full price or on clearance. For the numerical features, we will be interested in eight main columns that we summarized in this table:

Table 1 - Numerical Features Extracted

Original Column Name	Description	Updated Column Name
TOTAL_DEMAND	This is a column showing the total sales for the product in monetary value (USD). After aggregating the data, this column shows the daily sales for each product.	DEMAND
TOTAL_UNITS	This column shows the total sales for the product in number of units.	TOTAL_UNITS
CANCEL_DEMAND	This column shows the amount of canceled demand in monetary value (USD)	CANCELED_DEMAND
CANCEL_UNITS	This column shows the amount of canceled demand in the number of units.	CANCELED_DEMAND
RETURN_DEMAND	This column shows the amount of returned demand in monetary value (USD).	RETURNED_DEMAND
RETURN_UNITS	This column shows the amount of returned demand in the number of units.	RETURNED_UNITS
TOTAL_GROSS_MARKDOWN	These columns explain the amount of money the company lost in markdowns or promotions. For example, if a product's original price is 100$ and the markdown for it is 10%, making the current price 90$, then if one unit is sold of that item, the TOTAL_GROSS_MARKDOWN column would show 10$ and the TOTAL_DEMAND column would show 90$.	MD_AMOUNT
TOTAL_GROSS_PROMO		PROMO_AMOUNT

In addition to these existing columns, we've also chosen to create extra columns to further enrich our analysis:

Table 2 - Additional Columns Created

Column Name	Description	Formula
NOMINAL_AMOUNT	This column shows the sales amount that the company would generate if there were no markdowns or promotions.	DEMAND + MD_AMOUNT + PROMO_AMOUNT
MD_PERC & PROMO_PERC	Given the absence of columns showing the percentages of markdowns and promotions, we've created additional columns for these factors.	DIV0(MD_AMOUNT,NOMINAL_AMOUNT) DIV0(PROMO_AMOUNT,NOMINAL_AMOUNT)
DISCOUNT_DEPTH	This column would show the total effect of markdowns and promotions compared to the original price sales.	DIV0((SUM(TOTAL_GROSS_MARKDOWN) + SUM(TOTAL_GROSS_PROMO)), (SUM(TOTAL_GROSS_PROMO) + SUM(TOTAL_GROSS_MARKDOWN)+ SUM(TOTAL_DEMAND)))
FPR	Full price rate is another metric used in retail to show the fraction of the actual sales compared to the original price sales.	DIV0(SUM(TOTAL_DEMAND), (SUM(TOTAL_GROSS_PROMO) + SUM(TOTAL_GROSS_MARKDOWN)+ SUM(TOTAL_DEMAND)))

Product Attributions:

The selection of precise product attributes is significantly important in the context of this project. To avoid potential mapping issues, we have opted to align our attributions with those currently in use by our team, following their established selection criteria. However, the complexity arises from the absence of a single table that houses all the required characteristics. Therefore, we anticipate the need to collect data from multiple tables to obtain the desired attributions. We won't go into detail about these table joins in this section, but a comprehensive description of each selected column can be found in the table below:

Table 3 - Attribution Columns Extracted

Column	Description
GENDER	The gender of the intended target consumer: Men, Women, and Kids
SEGMENT	Products are split between a performance segment and a lifestyle one.
DIVISION	Products are split between three divisions: footwear, apparel, and accessories & equipment.
CATEGORY	Products fall under 24 different categories.
DIMENSION	Products fall under 14 different dimensions.
FAMILY	Each product belongs to one of 415 families.
MODEL	Each product is one of 977 models.
FRANCHISE	Products are divided into different franchises that are tied together by a cohesive design characteristic.
SUB_FRANCHISE	Each franchise has sub-franchises based on the family.
MERCHANDISING_CLASSIFICATION	A different classification given by another department.

SILHOUETTE	Each product belongs to one of the 116 silhouettes.
SILO	A distinct way to segment the line that represents the product's inspiration, intended use, or heritage.
EPOD	Each product has a defined lifespan, and the EPOD date signifies the date at which the product will be withdrawn from the market.

Markdown Change:

Price changes will be extracted from a table named 'Prodigy,' which records the daily pricing information for every product featured on the digital website. This dataset includes both the original prices and the current prices. However, given the variation in prices across different countries and our focus on the European market, we have opted to filter the data for one specific country. This approach allows us to narrow our analysis to a region where price fluctuations align more closely with the common trends observed across the entire European market.

Inventory:

Within this retail company, inventory is sourced from two distinct warehouses, each requiring a separate approach due to the data also being presented across two separate tables. The first warehouse serves a multitude of purposes, catering to different stores, partners, and company-owned outlets. To leverage our data, we've implemented specific filters to ensure that we exclusively consider stock levels originating from contracts established through Digital channels. This filtering enables us to focus solely on the inventory relevant to our analysis, enhancing data clarity and relevance.
On the other hand, the second warehouse exclusively serves digital operations, making the extraction of stock levels a straightforward task. However, the challenge emerges from the fact that this second table does not include a date column only a week indicator that follows a unique calendar, distinct from the conventions observed by other data tables. To harmonize our data sources and facilitate integration with other datasets, we've undertaken the step of joining it with the retail calendar table.

Digital Traffic:

The retrieval of digital traffic data was a relatively straightforward query, with the sole requirement being the application of a filter specifying the region from which this traffic originates. This targeted filter effectively narrows down the dataset, allowing us to get the specific digital traffic relevant to our analysis.

Promotion Indicator:

Determining a reliable promo indicator presented a considerable challenge, as the absence of a dedicated calendar specifying promotional dates complicated the task. Promotions within the company's operations can vary widely in terms of lead time, with some planned well in advance and others arranged only a few days prior, often without documented changes. To overcome this issue, we needed to devise an alternative approach to detect promotions.
Our initial consideration was to examine the sales table and identify transactions where the promotion amount exceeded zero. However, a closer examination revealed that almost all daily aggregated transactions contained some promotional amount. This phenomenon stemmed from the company's policy of granting employees discounts on their purchases, with these discounts

being recorded as promotional amounts in transaction records.
As a more effective alternative, we turned to an orders table. This table offers a daily breakdown of every order placed and includes a promotional ID whenever a promotional code is employed. We used this code along with a threshold for the discounts to determine if the company is in a promotional period or not. This shift in our approach to promotions detection ensures that we accurately capture promotional activities, bypassing the intricacies of employee discounts in the sales data and enabling us to maintain precision in identifying promotion-related events. Despite the effectiveness of this approach, we noticed that it was not able to detect a very important promotional period that occurs during the transition between seasons, specifically, in the last week of each season and the first week of the subsequent one. This issue will be fixed when joining the tables together.

4.2.2.2. Data Integration and Table Joins

When writing the query to join these tables into one cohesive dataset, we decided to employ the common table expressions, commonly referred to as CTEs in SQL.
CTEs are similar to temporary tables that help simplify complex SQL queries by making queries easier to read and maintain. The concept consists of defining a CTE at the beginning of a query and then referencing it within that query to break down complicated tasks into smaller, more understandable pieces.
After writing the queries for those various tables separately, we ended up with these temporary tables and relationships illustrated in Figure 2.

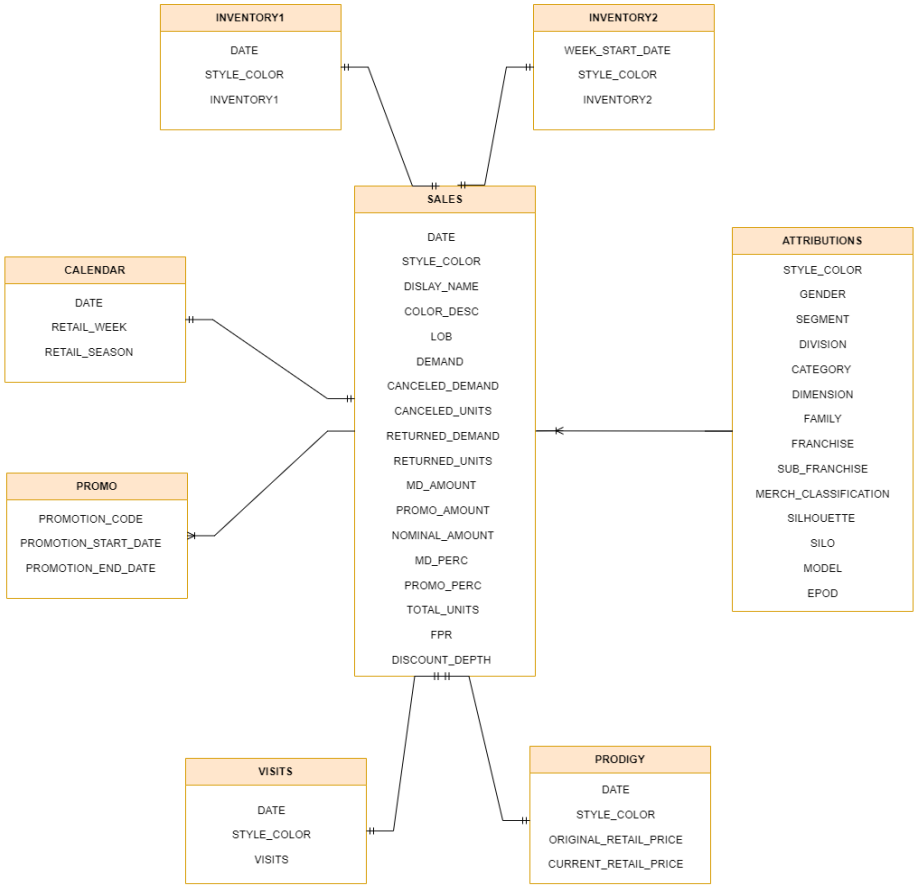

Figure 2 - CTE Created Tables

Most of these tables share two columns in common, the date and the product code, also known as the style-color. This notation came from the combination of the product style code and a color code making it a unique identifier for each product. Following this, the process of joining the tables became a straightforward task. We accomplished this by utilizing left join criteria, specifically, the product code and date. It is important to note that with the exception of the 'Inventory2' table, all tables maintained a consistent level of granularity. The 'Inventory2' table presented a unique case as it was aggregated every week rather than daily. To integrate it with the 'sales' table, we employed the week's start date as a common reference point with the date in the ‘Sales’ table. This method resulted in the emergence of missing values, which we will address during the data preparation phase. Additionally, we integrated the retail calendar into the entire dataset, enabling us to access valuable calendar-related attributes like the retail week and the current season. For the promotional calendar, we checked each sales transaction to determine if the sale date fell within the specified promotion start and end dates. If it did, we created a corresponding indicator to mark it as a promotional event. Furthermore, we implemented a condition to identify and mark sales occurring during the last week of a season and the first week of the subsequent season as promotions.

4.2.3. Data Understanding

Before testing a model, it's a good idea to first comprehend the data and try to get as many insights as possible. For this part, we will be trying to check data quality, discover patterns, spot anomalies, and test hypotheses with the help of summary statistics and visualization tools.

4.2.3.1. Data Description

The athletic retail business dataset consists of 3,472,791 transaction records, and 37,676 different products, covering a period of one year. Specifically, this test dataset encompasses transactions from May 8, 2022, to May 7, 2023. Within this dataset, there are 36 columns, with 18 of them being categorical and the remaining 18 being numerical. After analyzing these columns, we decided to categorize them as follows:

Table 4 - Data Columns Categorization

Group	Variables
Metric features	DEMAND, CANCELED_DEMAND, CANCELED_UNITS, RETURNED_DEMAND, RETURNED_UNITS, MD_AMOUNT, PROMO_AMOUNT, NOMINAL_AMOUNT, MD_PERC, PROMO_PERC, TOTAL_UNITS,FPR, DISCOUNT_DEPTH, ORIGINAL_RETAIL_PRICE, CURRENT_RETAIL_PRICE, INVENTORY1, INVENTORY2, VISITS
Non-metric features	STYLE_COLOR, DISPLAY_NAME, COLOR_DESC, GENDER, SEGMENT, DIVISION, CATEGORY, DIMENSION, FAMILY, FRANCHISE, SUB_FRANCHISE, MERCHANDISING_CLASSIFICATION, SILHOUETTE, SILO, MODEL, LOB, PROMO_PERIOD
Date Data	DATE, RETAIL_SEASON, RETAIL_WEEK

An initial analysis of the dataset showed that columns related to cancellations and returns displayed only negative values, indicating a potential adverse effect on sales. Furthermore, we noted that not all products are consistently present in the dataset throughout the year. While this absence could imply that these products were not sold during certain periods, it may also reflect their absence from the market entirely. Unfortunately, at this stage, we lack the means to definitively determine a product's presence or absence in the market during different timeframes.

4.2.3.2. Assessing Data Quality

Our first step in assessing data quality involves checking for the presence of duplicates or missing values. No duplicates were present in the dataset. However, a substantial number of missing values were identified. All columns originating from the attribution table showed missing values, with 'MERCHANDISING_CLASSIFICATION' having more than 27% of its values missing. This can be explained by the fact that not all products featured in the sales table were present in the attribution table. Additionally, missing values were detected in columns from the sales table, where additional calculations had to be made. These missing values might have arisen from divisions by zero during calculations. We also noticed missing values in the traffic and inventory columns. In the case of inventory2, the presence of missing values was understandable, as it was aggregated weekly, and merging it with a daily table naturally resulted in gaps. However, missing values in the first inventory column and traffic were due to various reasons. One possibility is that these tables did not encompass every product featured in sales,

leading to missing values upon joining. Another hypothesis is that when a product registered a value of 0, it might not have been recorded in these tables, indicating the potential absence of daily entries for certain products. It's worth noting that the remaining columns from sales, as well as those from the retail and promotional calendars, exhibited no issues in terms of missing values.

The second step is to check for the existence of outliers. Our first approach to check was to analyze the descriptive statistics table of our data shown in Figure 3. By examining the difference between the max and the 75% percentile, we can suspect the presence of extreme values for almost all the variables. On the other hand, by examining the min and the 25% percentile, it seems that we will not be dealing with lower outliers, except in the case of the cancellations and returns variables, where the values are recorded with a negative sign. Our second approach was to build histograms for all the numeric variables. The graphs obtained showed highly skewed data, which confirmed the existence of outliers. Our third and final approach was to build boxplots for the numeric variables. This final method affirmed that we have some extreme values that need to be dealt with.

	count	mean	std	min	25%	50%	75%	max
DEMAND	3472790.0	1179.406185	12929.973185	0.000000	48.584349	183.739500	611.130619	6.602806e+06
CANCELED_DEMAND	3472790.0	-35.836472	987.769954	-923501.269537	0.000000	-0.000000	-0.000000	-0.000000e+00
CANCELED_UNITS	3472790.0	-0.507074	15.780644	-16074.000000	0.000000	-0.000000	-0.000000	-0.000000e+00
RETURNED_UNITS	3472790.0	-3.023236	11.445594	-1464.000000	-3.000000	-0.000000	-0.000000	-0.000000e+00
MD_AMOUNT	3437362.0	210.559664	2421.138159	0.000000	0.000000	0.000000	32.400000	1.234329e+06
PROMO_AMOUNT	3437362.0	86.566095	1367.816918	0.000000	0.000000	0.015000	32.805000	1.344576e+06
NOMINAL_AMOUNT	3437362.0	1487.432290	14276.678971	0.000000	59.049000	227.924646	769.775111	6.623112e+06
MD_PERC	3437362.0	0.090528	0.147300	0.000000	0.000000	0.000000	0.192901	7.887953e-01
PROMO_PERC	3437362.0	0.045204	0.081442	0.000000	0.000000	0.000034	0.054915	9.999882e-01
TOTAL_UNITS	3472790.0	18.721317	138.548115	0.000000	1.500000	4.500000	13.500000	6.476550e+04
FPR	3437362.0	1.026028	2.035917	0.000000	0.839694	1.199073	1.456165	4.215730e+02
DISCOUNT_DEPTH	3437362.0	0.181601	1.974354	-420.073034	0.000000	0.080649	0.375044	1.500000e+00
INVENTORY1	1860201.0	1875.658989	4906.724560	1.500000	183.000000	627.000000	1854.000000	6.552750e+05
INVENTORY2	386833.0	3342.381378	37300.521283	0.000000	52.500000	313.500000	1579.500000	1.500043e+07
VISITS	3357227.0	813.205403	4182.674019	1.500000	85.500000	228.000000	616.500000	2.180158e+06

Figure 3 - Numerical Data Description

In the final step of assessing the data quality, we decided to run a coherence check. We wanted to make sure that the data is logically consistent and can be reliably combined for analysis. Most columns seemed to follow a logical pattern. In cases of high sudden demand for specific products, it was usually highly linked to promotional periods. However, we identified a unique case where certain products appeared in the dataset for only a couple of weeks but showcased a high demand compared to the rest of the products during that period. After consulting the team, we learned about a different part of the business dedicated to limited-edition products. These products are usually collaborations with celebrities and are very exclusive and more expensive. Buyers often have to enter a lottery to secure the chance to purchase these products, which can be available for as little as a couple of days. Hence, this explains the big difference in distribution and demand compared to the rest of the products. Added to that, we also encountered cases of products present in the dataset but showed no recorded demand. These situations could result from different scenarios. For instance, we noticed that those products always had a value in the canceled or returned columns, which signifies that these transactions were canceled, or the products were returned at one point during the year without actually registering any demand. Alternatively, there might be issues with data recording or potential

entry errors. Furthermore, the dataset contained daily transactions with unidentified or unknown product codes but still showcased valuable demand. And since the predictions will be made on a product level, those entries would be irrelevant to the model. Finally, we also observed products with a low number of transactions, which could pose a challenge for our forecasting model. For example, products with three transactions throughout a whole year, make it impossible to detect patterns and trends, which affects the model's accuracy. To face these challenges, we may consider addressing them as a separate category in our modeling process. After conducting a detailed exploration of the dataset, it is evident that substantial work lies ahead. The dataset exhibits numerous missing values and outliers that necessitate attention. However, it is worth noting that with the necessary corrections and enhancements, we can transform this dataset into a satisfactory and suitable resource to fulfill the project's objectives.

4.2.3.3. Data Visualization

In this section, our primary objective was to gain a deeper understanding of the dataset by employing data visualization techniques. We aimed to identify patterns and insights that would be valuable during the model-building phase of the project.
We initiated our exploration by focusing on the 'sales' variable, visualizing the sales throughout the year, and trying to detect any pattern from the date variables perspective.

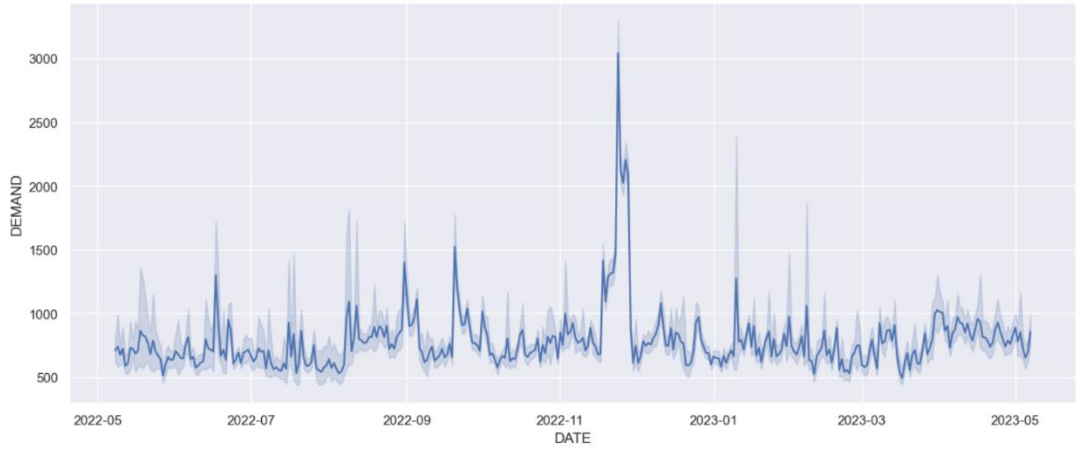

Figure 4 - Variation of Demand throughout the Year

As shown in Figure 4, the first thing we noticed is that sales fluctuate a lot throughout the year. However, we did notice remarkable surges in sales during certain weeks, specifically towards the end of November. And after checking with the team, it turned out that those spikes in sales coincided with a significant promotional period. In the last week of November, the company undergoes one of its most influential promotion periods known as Cyber Week, during which many promotional codes are distributed, and markdown strategies are implemented. Consequently, November was shown as the highest month for sales due to its massive impact on overall sales throughout the year. This conclusion confirms that the sales vary with the 'Date' variable and therefore there is a seasonality factor present in our data.
We have also visualized the sales for the different days of the week and days of the month. As shown in Figure 5 and Figure 6, we observed a consistent trend that day 25 of each month is the highest in terms of sales. This pattern can be explained by the fact that most people receive their salaries on that day and would probably freely engage in purchases right away. We have also noticed that Sunday is the highest day of the week in sales, while Wednesday is the lowest.

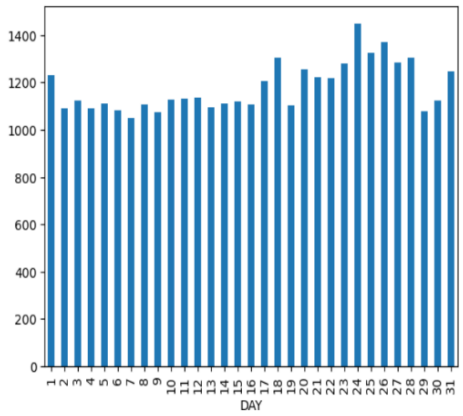

Figure 5 - Average Demand for each day of the month

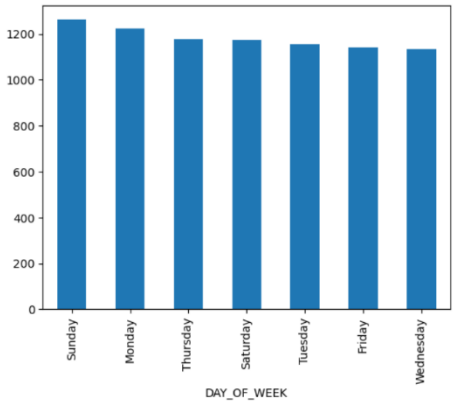

Figure 6 - Average Demand for each day of the week

From the above plots, we can see that there are seasonality and trends present in our data. Therefore, both of these factors must be taken into consideration in the modeling phase. Next, we focused on the categorical variables within the dataset. Our analysis entailed examining the distribution of these variables in relation to their impact on the sales column. We visualized the transaction count for each categorical variable and the sales effect. This analysis led to the discovery of meaningful conclusions. Firstly, we encountered scenarios where a specific element within a variable significantly dominated the number of transactions, and this dominance was consistent with its impact on sales. One such example is the 'gender' variable as shown in Figure 7 and Figure 8.

On the other hand, there were instances where certain variables had elements with the highest transaction counts, but this did not correlate with their contribution to sales. For instance, as shown in Figure 9 and Figure 10, the 'division' column exhibited this discrepancy, which can be attributed to the influence of pricing factors.

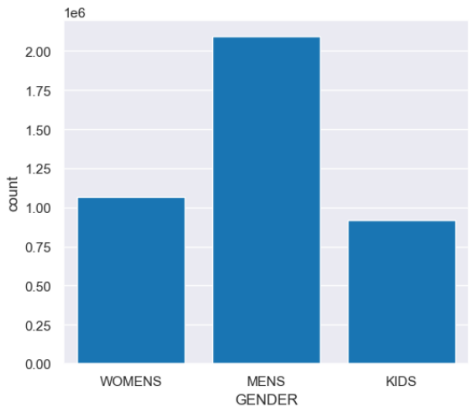

Figure 7 - Transaction Count per Gender

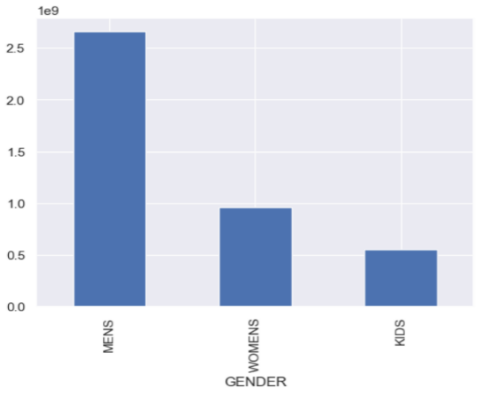

Figure 8 - Total Demand per Gender

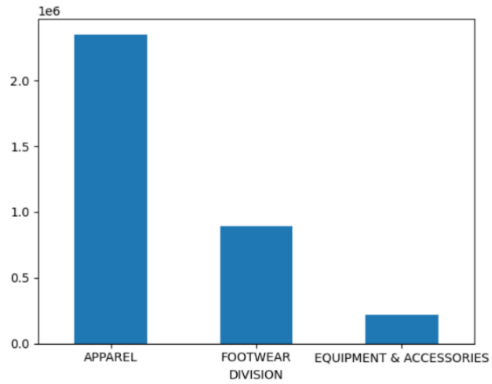

Figure 10 - Transaction Count per Division

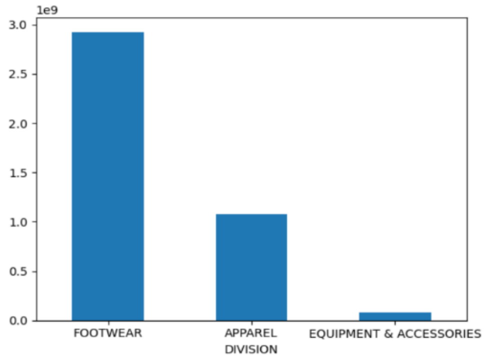

Figure 9 - Total Demand per Division

4.2.4. Data Preparation

According to a survey done by CrowdFlower conducted on 80 data scientists, 60% of them agreed they spend 80% of their time on data preparation. This is understandable due to the importance of this phase in building the model later on. The complexity comes from the effort it takes to accommodate the database to the model and fix its issues. During this phase, we will be manipulating and transforming data elements into a useful dataset to match the model's needs.

4.2.4.1. Data Consistency Improvement

Before dealing with outliers and missing values, we decided to get rid of rows that would affect the rest of the process. We started by deleting rows with an unknown product identifier as well as products that had no demand throughout the year.
Finally, to tackle the issue of products with a low transaction count, we had to create a threshold for the number of transactions for each product. A good rule of thumb that we deemed effective in similar retail forecasting problems was to set the lowest threshold as the 25th percentile of the total transactions count. This threshold was tested throughout various weeks, and it has always been in the range of 30-40 transactions. Therefore, we have created a separate dataset where we stored product codes having transactions below the established threshold. These products will not be included in the modeling phase, but they will still be considered for predictions using other methods that will be discussed later.

4.2.4.2. Missing Values

First, for the product characteristic columns with missing data which come from both the attribution table and the sales table: gender, segment, division, category, dimension, family, franchise, sub-franchise, silhouette, silo, model, EPOD, display name, and color description, we decided to fill the missing values with ‘Unknown’. For the merchandising classification column which exhibited the biggest volume of missing values, we decided to exclude it in the modeling phase. First, it would be difficult to rectify the missing data issue without compromising the overall data quality. Second, that specific classification is not an important one for the inventory management team and doesn’t align well with the remaining attributions.
Next, for the missing numeric variables coming from the sales table, we replaced the missing values with 0. It is important to note that many of the numeric variables extracted from the sales

table were pulled purely for analysis purposes and were not to be used in the modeling process. Therefore, we only devoted our attention to resolving issues with the columns that can potentially be used in the modeling phase and align with the project's objectives. For the missing values coming from the pricing, inventory, and visits columns, we had a common approach. For this group of columns, we used the forward-fill and backward-fill methods. Forward-filling captures the most recent non-null value preceding the missing entry, while backfilling uses the subsequent non-null value to replace the missing values. The remaining gaps were filled with zeros to maintain data consistency.

4.2.4.3. Outliers

"Data often contain noise, errors, or exceptions. Errors and noise may confuse the data mining process, leading to the derivation of erroneous patterns" (Tallón-Ballesteros & Riquelme, 2014). However, determining whether a value is an outlier that should be removed or not is very subjective. While there are certainly valid reasons for throwing away outliers if they are the result of a computer glitch or a human error, eliminating every extreme value is not always a good idea. Therefore, we must first study each column separately to decide about the nature of outliers. In our case, the inventory 1 and 2 show extreme values that are probably a result of a glitch. It would be impossible to accommodate 10 million for just one product in our warehouses. For this case, we decided to use clipping methods, in which we use the 99th percentile as a threshold and replace the values above that value with the threshold. For the rest of the columns, the outliers are not unusual values that are impossible to occur. On the contrary, the range of these extreme values is very normal in high promotional periods for example, and their existence may show some hidden meaning about the demand pattern. Instead of removing them, they will be kept to further investigate their meaning. Nevertheless, to not disturb the process of the modeling process, we will try to reduce their effect by applying a square root transformation.

4.2.5. Data Preprocessing

4.2.5.1. Feature Engineering

The purpose of this part is to make patterns in the data more explicit for the model. Usually, when dealing with time series problems, there are four types of features to be extracted for the modeling process. We will go through these extracts separately.
The first type is the date component. These are values extracted from a specific date or timestamp to help the model identify seasonal and cyclical patterns in the time series. For our project, we've already extracted date components in the data understanding part to better analyze the variation of sales. The variables that we extracted are the following:

- Year
- Month
- Quarter
- Day
- Day of week

The second type is the lags. These are simply past values of the time series that are shifted forward to use as features in the model. These variables are used based on the assumption that past values can define the present and future predictions. For example, predictions for the

demand for a product next week can rely on the demand of the same weekday in the previous week as a feature. For this project, we tested different lags to find the ones that work best for this specific problem, and we identified them as follows:

- A 9-day sales lag
- A 30-day sales lag
- A 45-day sales lag

The third type is the rolling window aggregations. These are statistical functions applied to records in a sliding window to help the model identify trends in the time series. For this project, these variables were created:

- A 7-day sales moving average
- A 30-day sales moving average
- A 45-day sales moving average

Finally, the last type is the differences. These are simply the subtraction of two values in a time series. For example, the difference between the demand of today and the demand of yesterday. For this project, we wanted the model to be able to detect if the price and the markdown level fell from its original price or a previous markdown in comparison to prior dates. Therefore, we created a difference variable to showcase this change based on the markdown percentage column.
In addition to these variables created, we have also created average variables as follows:

- Average sales for each product
- Average sales per month for each product

4.2.5.2. Feature Selection

This part of the project was a tricky one. When deciding which variables to incorporate into the modeling process, we had to consider which variables we would know in advance. For instance, our current database contains inventory levels and website traffic, however, when we predict for the upcoming week, we wouldn't have the knowledge of the inventory levels for that specific product during that time. That knowledge would require a separate prediction model. Therefore, we had to find an alternative approach for including those variables. We decided to create additional variables that could encompass certain aspects of the information found in these columns and they were as follows:

- Average Inventory1 for each product
- Average Inventory1 per month for each product
- Average Inventory2 per month for each product
- Average Inventory2 for each product
- Average Visits for each product
- Average visits per month for each product

However, after a trial-and-error method of trying those different variables with the model and checking for accuracy, we found out that average inventory per month variables did not contribute well to the model performance and were therefore removed. For the rest of the variables, we followed a correlation-based approach. This involved evaluating their correlation with the target variables and their correlation with other variables in the dataset. With this

method, we concluded that canceled demand/units, returned demand/units, FPR, and discount depth had a low correlation with demand, and they were removed from the dataset.
The nominal demand column was derived from the demand column and it was also removed. We have also decided to keep only one of these columns to avoid redundancy in the modeling process:

- MD amount and MD perc
- promo amount and promo perc

4.2.5.3. One-hot Encoding

One-hot encoding is the process of changing categorical data into numerical ones. When dealing with one-hot encoding, the two, most common options that might be considered are:

- get_dummies: This function turns categorical values into dummy values, meaning binary values of either 0 or 1.
- Encoding: This method is useful for obtaining a numeric representation of an array when all that matters is identifying distinct values. It gives each option a numeric value starting from 0 or 1. The drawback of this method is that the algorithm might prioritize and give more importance to options with the highest number.

In our case, with 37,676 unique product codes, 977 different models, and 656 distinct franchises, the first one-hot encoding approach would result in an overwhelming number of additional columns, making the model construction infeasible. Consequently, we opted to employ an encoder to efficiently manage this complexity.

4.2.6. Model Implementation

As mentioned in the data understanding part, this dataset encompasses 37,477 different products, meaning we have to predict 37,477 different values for the same week. This presents us with two different approaches. The first involves treating each product independently, necessitating the creation of a unique model for each. The second approach involves developing a single model to forecast all the sales values collectively.
To make this decision, we took into consideration business opinions that have clarified that they believe product sales have interdependencies. They emphasized that if the sales of one product increase, it can potentially influence the sales of another. And this behavior cannot be detected unless we adopt a single-model approach consisting of all the products.
To validate this decision, we've also decided to acknowledge past similar projects and see their approach to this problem. During our research, we came across a data science challenge launched by Walmart known as the M5 competition (Kaggle, 2020) which was created for the company to enhance its forecasting models. The competition's purpose was to predict future sales at a product level based on historical data. More than 5000 teams of data scientists participated in this competition sparkling discussions about the methods, features, and models that would work best to address this problem. The debates highlighted recurring issues in all retail forecasting and emphasized the importance of one model for all product types of forecasts.

4.2.6.1. Data Split

Unlike regular machine learning problems where the random split is required since there is no dependence from one observation to the other, time series poses a different situation. In time series, the date variable is a viable component of the model, and our purpose is to maintain the timeline of the data to ensure the model can detect the change of patterns over time. For this reason, we sorted our data by date and product code and then extracted the last seven days to be our test dataset throughout the process.

4.2.6.2. Model Building

Throughout this whole project, the main challenges that we encountered were not related to the ML model building, but rather to collecting the data, manipulating it, and enhancing its quality. Now that that is sorted, our purpose is to find a model that is capable of handling large-scale data and deriving insights from it. For this quest, we have also referred to the Walmart M5 competition in which they have concluded that the best model to use for this type of retail forecasting is LightGBM. LightGBM is a light gradient-boosting machine learning model based on decision tree algorithms, originally developed by Microsoft but was made to be free open-source. It is known for its higher efficiency, better accuracy, and lower memory usage due to these two features (Lightgbm, n.d.):

- **Gradient-based One Side Sampling:**

The different data observations have varied impacts on the information gain computation. LightGBM has the capability to determine which instances possess a larger impact and which instances don't. It then sets a threshold dropping the observations with a lower impact and keeping instances with larger gradients and uses them to improve on the accuracy of information gain estimation.

- **Exclusive Feature Bundling:**

LightGBM can detect features that are mutually exclusive and bundle them into one single feature without causing any information loss. This step ensures the reduction of the number of features, the simplification of the data, and the increase in the speed of the training framework without affecting the accuracy.

This model was used to derive daily forecasts on a product level, which were then aggregated to get weekly predictions. The reason we did not aggregate the data on a weekly basis before modeling is to ensure the model is detecting patterns at the lowest granularity possible. We have also used the demand in monetary value as our target variable and used the predictions as well as the price to derive weekly forecasts in units.

4.2.6.3. Model Evaluation

Given the small granular level of forecasting where actual demand can occasionally be zero, it's essential to select appropriate metrics to assess the model's performance effectively. In this context, traditional metrics like Mean Absolute Percentage Error (MAPE), which involve division by actual demand as shown in Equation 1, may not be suitable due to the potential presence of zeros. Instead, metrics such as Mean Absolute Error (MAE) or Percentage

Accuracy could be more fitting. MAE measures the average absolute differences between predicted and actual values, providing a straightforward assessment of forecasting accuracy (Equation 2). We decided to use MAE to assess the general performance of the model. However, we will need a metric to evaluate the model for granular-level assessment. For that, a weighted version of the Symmetric Mean Absolute Percentage Error (SMAPE) will be used (Equation 3). SMAPE is well-suited for these cases. It quantifies the percentage difference between predicted and actual values, with a symmetric approach to account for both overestimations and underestimations. By incorporating a weighting scheme, as shown in Equation 4, we can further customize the metric to align with the specific priorities and characteristics of the forecasting model. Given that not all products in the business have the same importance and we would want to prioritize inclining the model effectiveness towards certain products than the others, this weighted SMAPE offers a more nuanced view of performance, acknowledging that not all granular-level forecasts are equally significant and tailoring the evaluation accordingly. Together, these metrics, MAE and weighted SMAPE, provide a comprehensive and tailored framework for assessing the model's accuracy and effectiveness in the context of granular-level forecasting without encountering issues related to zero demand. The final accuracy metric is shown in Equation 5.

- **MAPE:** $\frac{1}{n}\sum_i^n abs(\frac{A_i - F_I}{A_i})$ *Equation 1 - MAPE*
- **MAE:** $\frac{1}{n}\sum_i^n abs(A_i - F_i)$ *Equation 2 - MAE*
- **SMAPE:** $\frac{100}{n}\sum_i^n \frac{ABS(A_i - F_I)}{A_i + F_I}$ *Equation 3 - SMAPE*
- **WSMAPE:** $\frac{100}{n}\sum_i^n \frac{Wi * ABS(A_i - F_I)}{Wi\ A_i + Wi\ F_I}$ *Equation 4 - WSMAPE*
- **Accuracy:** 1 – WSMAPE *Equation 5 - Accuracy*

With A being the actual value, F is the forecasted value, n the number of observations, and w the weight of each product.

Following these formulas, the initial model evaluation showed 80% accuracy on a product level and 92% for the whole business.

4.2.6.4. Hyperparameter Tuning

Since we are dealing with a dataset of almost 4M observations and a high learning time, using methods such as random search and grid search resulted in high computational costs and an unreasonable waiting time. Instead, we opted for manual hyperparameter tuning following a trial-and-error method, where we tested different parameter combinations and kept a record of the accuracy points. Throughout this process, the only parameter that has shown an effective enhancement is the learning rate parameter, which determines the step size at each iteration while moving toward a minimum of the loss function.
This change has resulted in only a 0.01 decrease in the MAE, but it resulted in a 7ppt increase in the WSMAPE on a granular level, making our final accuracy 87% on a product level and 95% on a topline level (for the whole business)

4.2.6.5. Feature Importance

One of the reasons we resolved to lightGBM as our model is that is it a tree-based algorithm allowing us to assess how the model reached the final predictions and which features

contributed the most to the decision. After drawing one of the decision trees and using the feature importances function, we discovered that the feature engineering process had an important impact on the model performance. The main features behind the predictions were the average product sales per month and the sales lag. The markdown percentage and markdown differences were second on the list, showing that the sales behave differently during and outside of promotions. The date variables have also proven to be efficient, as the model gave an importance value on the day and retail week features. Added to that, the extra features added to the model, such as inventory and visits were also affecting the sales predictions despite not being on the same aggregation as the same dataset. One final conclusion we made out of the feature importance analysis is that the lowest contributing features were the attributions related to the characteristics of the products. This showed that products sharing similar characteristics, like belonging to the same family, do not necessarily behave in the same way.

4.2.7. Project Deployment

4.2.7.1. Scenario Mapping

The model built is useful for products with historical sales that have 30 minimum transactions throughout the year. However, when deploying this model in real cases, we will face scenarios of products with fewer transactions, or cases of products newly put on the market with no historical sales. Therefore, we had to create different scenarios to tackle all the cases that we might come across. The decisions made for these scenarios were agreed upon by the team as the approach they would normally follow in building their manual forecasts.

- **Scenario 1:** The product code has historical sales and more than 30 transactions throughout the year.

In this case, the product code will go through the model for the weekly demand to be predicted using the model built.

- **Scenario 2:** The product code has historical sales but its transactions throughout the year are less than 30.

In this case, we will take the average sales of the available transactions for that product.

- **Scenario 3:** The product does not have any historical sales either because it's a newly launched one, or because of mapping issues.
    - **Scenario 3.1:** The product belongs to the apparel or equipment division.

In this case, we will consider products with similar characteristics. We will consider the weekly predictions to be the average sales of the transactions of products having the same division, gender, category, and family.

    - **Scenario 3.2:** The product belongs to the footwear division.

In this case, we will consider the weekly predictions to be the average sales of the transactions of products having the same division, gender, category, and franchise.

In case we do not have neither the family nor the franchise of the product, we will keep making the clusters bigger. For instance, instead of considering products with the same division, gender, category, and family, we will consider segments of the same division, gender, and category. If the category does not exist, then we will consider just the division and gender.

- **Scenario 4:** The product does not have any matching attributions to the current dataset. This case would be a result of a mapping issue or a glitch as the division and gender are always available for every product. However, in case of this error, we will just mark it as unknown, and no predictions will be made for that product.

4.2.7.2. Project Workflow

Once the model is up and running, the next critical step is to develop an accessible interface that caters to retail planners who may not possess technical expertise. The proposed solution is a seamless fusion of Jupyter Notebook and Excel, ensuring user-friendliness for the entire team. The workload requested from the business team would be in the form of a simple Excel file filled with the products to be forecasted and containing these columns:

Table 5 - Excel Input Sheet Columns

Input Sheet Columns	Description
Style-color	Product code
Division	Product attributions that will be used in case the product code is not found in the training dataset.
Gender	
Category	
Family	
Franchise	
Promo Period	Our current code can not detect a promotional period the day it happens, therefore, we will need an indication from the team to facilitate the predictions.
Original Retail price	Price points for the week of predictions. This is where the team can test their pricing strategy.
Current Retail Price	

The predictions will be made for a retail week starting Sunday and ending Saturday. However, the code for the predictions will be run on Friday (The closest workday to the required predictions). On that day, the team fills out this Excel sheet with the products they want to have predictions for as well as the attributions, and the price they would set for the predicted week. This sheet is then used as input in a Jupyter Notebook where the following changes will be made:

1. We extract the product codes that do fit into the modeling conditions (more than 30 historical transactions throughout the year) and save them in a separate table.
2. A Calendar table is created in Python with the days to be predicted containing: the year, month, quarter, day, and day of week.

3. The retail calendar table is pulled from the snowflake warehouse containing the retail season and the retail week and is merged with the calendar table from Step 2.
4. We merge the calendar table with the Excel input sheet to make sure each of the product code inputs is redundant for every day throughout the prediction week.
5. From the training dataset, we take the last day it was updated (Friday) and create sales lags (9, 30, 45 days lags) to be used as features for the input sheet.
6. From the training dataset, averages for the sales, and sales per month for each product are created to be used as features for the input sheet.
7. The new features created are merged with the input sheet.
8. We create a column for the markdown percentage and markdown difference using the original retail price and the current one.
9. Now that the input sheet looks like our training dataset, we perform the same steps we did in data preparation in dealing with missing values, outliers, resizing, and one hot encoding.
10. We run the model, get the predictions, and then reverse the data resizing and the one-hot encoding.
11. We use the separate dataset created in Step 1 and run through the scenarios that we have created to get predictions for the products with low or no transactional sales.
12. We merge both predictions in one dataset and use the demand in monetary value predicted and the price to get a demand units prediction.
13. We extract the output as an Excel Sheet containing the product code, its attributions, and demand predictions both in units and monetary value.

Upon running the code, the system will automatically generate and download an Excel file containing sales predictions for the upcoming week. This Excel file will be intuitive and easy to use. It will allow retail planners to adopt the forecasting insights in their decision-making without the need for any technical skills. This approach guarantees a smooth and efficient workflow for the team, benefiting their daily planning activities.

5. RESULTS AND LIMITATIONS

This chapter is dedicated to the results achieved and the conclusions reached through this implementation. We will also tackle the challenges we faced that have limited this project and how to overcome them for future projects.

5.1. RESULTS AND FINDINGS

Finding the right algorithm for this problem was vital for the success of the project. Light GBM has proven to be one of the best models in similar retail forecasting problems, and that was also confirmed for this project. Before implementing it for real case predictions, we kept on testing it throughout multiple weeks where it had shown a significant accuracy consistency showing that it can be used for future periods. It has proven to be highly efficient, generating quick predictions in the span of a couple of minutes and allowing the team to make data-driven decisions. The results achieved were sufficient and did not call for other models to be tested. However, one of the observations we noticed throughout our analysis is that the accuracy of sales predictions can differ significantly from one product to another. With a dataset consisting of more than 37k products, and how diversified these products are, it is a difficult mission to tailor a model to the different product characteristics. The way we tackled this problem is by basing our model accuracy on a weighted metric favoring products with higher sales throughout the year which gives the team the opportunity to have visibility on more important products that must be prioritized in decision-making. Despite this vast diversity among products, we discovered a common pattern. The inclusion of various features, such as inventory, traffic, and promotional data, played an important role in improving our predictions. These features revealed distinct patterns in product sales, allowing us to better understand the dynamics that influence consumer demand. These findings have the potential to reshape the way we approach retail forecasting problems and magnify the importance of multi-factor data-driven approaches.

5.2. LIMITATIONS

Throughout this project, our main challenge was not with the model creation, but rather with manipulating the data and getting it to the point of being implemented in the model. As a result, our main struggles were encountered during the data collection and preparation phases. Below, we will go through each problem separately:

- **Data Discrepancies and Quality Issues:** As mentioned multiple times throughout this paper, in this retail company, there is not a single source of truth. The databases can be built across different departments and following different methods, updates, and attributions resulting in inconsistency when comparing different sources. In our case, our main issue was the discrepancies between the input sheet the team delivers and the snowflake tables we used for our analysis. Since the team does not use tools like Snowflake, they pull the product codes and attributions from another data source which has resulted in differences in the attributions from both datasets. Added to that, we noticed that not all products coming from Excel exist in the training dataset resulting in these products not being included in the model despite having historical sales, which requires a deeper look into the root causes of these differences and a comprehensive data cleansing.

- **Data Deficiency:** Despite the data abundance in this company, we still found ourselves lacking one of the main factors. For instance, an inventory challenge was presented by the existence of multiple inventory tables with different granularities. While the first table was updated daily, the second one was only updated weekly resulting in data deficiency. Another issue that we had to acknowledge is the dynamic nature of the datasets. Most databases face constant changes, updates, and sometimes even decommissioning which requires us to always be on the lookout for other data sources that can be used as a replacement.
- **Data Scale:** Our dataset scale, characterized by millions of transactions and thousands of products, made the process of hyperparameter tuning a complex and resource-intensive task.

5.3. FUTURE PROJECTS

The project was able to satisfy the team's needs and answer the most complex questions, however, there is always room for improvement. One of the main things we can further push on are listed below:

- **Full Automation:** Our current process now still relies on team input and an analyst to run the code for the predictions to be derived. To enhance the project efficiency, we anticipate transitioning to an automated Tableau solution that can streamline the operations run the necessary steps automatically, and display the results in the form of a dashboard providing the team with a self-service tool for insights. In future years, perhaps there can also be a way to get real-time data updates as well instead of waiting for a daily refresh at a certain hour. Focusing on these real-time data updates will ensure access to the most current insights.
- **Price Change Impact:** Although the model has provided a simple understanding of the variables affecting the sales changes, we still don't have a full understanding of the relationship between these factors and sales. One specific variable that the team needs is the price changes. Fathoming the hidden relationship between price changes and consumer demand can help the team decide on a better pricing strategy for the upcoming period. Instead of getting simple forecast predictions, we can work on getting the price that will incite higher demand.
- **Continuous Improvement:** This project will not remain valid all the time. The business needs and market dynamics are ever-changing and constant and continuous improvement is needed. This is achievable by first keeping up with data source changes. Data quality and consistency will remain the priority for any future projects to ensure our models are built on reliable and accurate data sources. Second, keeping up with the business changes and adapting to them accordingly, for instance, changes in the pricing strategies.

This chapter summarizes the key results and findings from our project, highlights the limitations faced, and outlines the path forward for future research and improvement aimed at advancing retail forecasting.

6. CONCLUSION

This paper studies an artificial intelligence approach for retail sales forecasting through machine learning with a specific focus on an athletic digital retailer as an illustrative example. This process was not an easy task due to the fundamental nature of the retailer's infrastructure. Throughout the process, numerous challenges were encountered in the effort to prepare the data for consumption by the model. The issues included discrepancies in the data, issues related to its availability, and complexities in the maintenance process. However, despite all these complexities, we managed to create a process that was able to answer all the business needs and provide the team with the necessary information to make more data-driven decisions. It has also laid the groundwork for any potential or ongoing improvement aimed at sustaining the project through adaptation.

The central focus of this research was the feasibility of implementing advanced analytics within a retail business with a delicate IT infrastructure. Despite the challenges, a sophisticated machine learning model was built, characterized by both consistent high accuracy and a user-friendly process tailored for the business team. Not only did this model provide the team with valuable insights into expected sales values, but it also provided a nuanced understanding of how various factors influence overall sales dynamics including the impact of their pricing strategies on sales.

One of the main takeaways drawn from this project is the indispensable role of domain expertise. Throughout our work, we encountered many problems that we could not have overcome if not for the business experts. They have offered nuanced understanding that complemented our quantitative aspect of forecasting and integrating these perspectives has been vital in enhancing the robustness of our predictive model. This is a concept they often refer to as the balance between art and science, between the scientific aspect of data analytics and the art of business intuition. The success of this project lies in the perfect integration of these two facets and moving forward, this balance remains important to the evolution of retail forecasting.

In essence, this project not only signifies a milestone in effective granular forecasting but also highlights that there is always room for improvement in the dynamic field of retail. In recognizing the accomplishments achieved, it's crucial to acknowledge the potential for future enhancements. This is achieved through continuous refinement and adaptation and keeping the effective collaboration between data-driven methodologies and the insights of industry experts.

7. REFERENCES

Alon, I., Qi, M., & Sadowski, R. J. (2001). Forecasting aggregate retail sales: Journal of Retailing and Consumer Services, 8(3), 147–156. https://doi.org/10.1016/s0969-6989(00)00011-4

Boylan, J. E., Chen, H., Mohammadipour, M., & Syntetos, A. (2014). Formation of seasonal groups and application of seasonal indices. Journal of the Operational Research Society, 65(2), 227–241. https://doi.org/10.1057/jors.2012.126

Chu, C.-W., & Zhang, G. P. (2003). A comparative study of linear and nonlinear models for aggregate retail sales forecasting. International Journal of Production Economics, 86(3), 217–231. https://doi.org/10.1016/s0925-5273(03)00068-9

Cooper, L. G., Baron, P., Levy, W., Swisher, M., & Gogos, P. (1999). PromoCastTM: A new forecasting method for promotion planning. Marketing Science, 18(3), 301–316. https://doi.org/10.1287/mksc.18.3.301

DeHoratius, N., & Raman, A. (2008). Inventory record inaccuracy: An empirical analysis. Management Science, 54(4), 627–641. https://doi.org/10.1287/mnsc.1070.0789

Ferrera, C., & Kessedjian, E. (2019). Evolution of E-commerce and Global Marketing.

Fildes, R., Ma, S., & Kolassa, S. (2022). Retail forecasting: Research and practice. International Journal of Forecasting, 38(4), 1283–1318. https://doi.org/10.1016/j.ijforecast.2019.06.004

GeeksforGeeks. (n.d.). Libraries in Python. Retrieved from https://www.geeksforgeeks.org/libraries-in-python/

Har, L. L., Rashid, U. K., Chuan, L. T., Sen, S. C., & Xia, L. Y. (2022). Revolution of Retail Industry: From Perspective of Retail 1.0 to 4.0.

Jupyter. (n.d.). Retrieved from https://jupyter.org/

Kaggle. (2020). M5 Forecasting – Accuracy

Koolen, D., Sadat-Razavi, N., & Ketter, W. (2017). Machine Learning for Identifying Demand Patterns

Lalou, P., Ponis, S. T., & Efthymiou, O. K. (2015). Demand Forecasting of Retail Sales Using Data Analytics and Statistical Programming.

LightGBM. (n.d.). Retrieved from https://lightgbm.readthedocs.io/en/stable/

Lyu, F., & Choi, J. (2020). The forecasting sales volume and satisfaction of organic products through text mining on web customer reviews. Sustainability, 12(11), 4383. https://doi.org/10.3390/su12114383

Petropoulos, F., Apiletti, D., Assimakopoulos, V., Babai, M. Z., Barrow, D. K., Ben Taieb, S., Bergmeir, C., Bessa, R. J., Bijak, J., Boylan, J. E., Browell, J., Carnevale, C., Castle, J. L., Cirillo, P., Clements, M. P., Cordeiro, C., Cyrino Oliveira, F. L., De Baets, S., Dokumentov, A., … Ziel, F. (2022). Forecasting: Theory and practice. International Journal of Forecasting, 38(3), 705–871. https://doi.org/10.1016/j.ijforecast.2021.11.001

Ragharan Srinivasan, S., Ramakrishnan, S., & Grasman, S. E. (2005). Incorporating cannibalization models into demand forecasting. Marketing Intelligence & Planning, 23(5), 470–485. https://doi.org/10.1108/02634500510612645

Seaman, B. (2018). Considerations of a retail forecasting practitioner. International Journal of Forecasting, 34(4), 822–829. https://doi.org/10.1016/j.ijforecast.2018.03.001

Shankar, V., Kalyanam, K., Setia, P., Golmohammadi, A., Tirunillai, S., Douglass, T., Hennessey, J., Bull, J. S., & Waddoups, R. (2021). How Technology is Changing Retail.

Snowflake. (n.d.). Retrieved from https://www.snowflake.com/en/

Statista. (2023). E-commerce as percentage of total retail sales worldwide from 2015 to 2027

Steinker, S., Hoberg, K., & Thonemann, U. W. (2017). The value of weather information for e-commerce operations. Production and Operations Management, 26(10), 1854–1874. *https://doi.org/10.1111/poms.12721*

Tallón-Ballesteros, A. J., & Riquelme, J. C. (2014). Deleting or Keeping Outliers for Classifier Training?

Wirth, R., & Hipp, J. (2000). CRISP-DM: Towards a Standard Process Model for Data Mining.

Wolters, J., & Huchzermeier, A. (2021). Joint in-season and out-of-season promotion demand forecasting in a retail environment. Journal of Retailing, 97(4), 726–745. https://doi.org/10.1016/j.jretai.2021.01.003

Zhao, Y., Yang, S., Narayan, V., & Zhao, Y. (2012). Modeling consumer learning from online product reviews. SSRN Electronic Journal. https://doi.org/10.2139/ssrn.1832763

NOVA Information Management School
Instituto Superior de Estatística e Gestão de Informação

Universidade Nova de Lisboa