Erasmus University Rotterdam
Rotterdam School of Management
Master Thesis Business Analytics & Management

Enhancing Category Marketing Planning by Forecasting SKU Sales Value Share: A Machine Learning Approach

Audrius Šaras (683384)

Coach:	Dr. Sebastian Gabel
Co-Reader:	Dr. J.M.T. Roos
Submission Date:	15th June 2024

Acknowledgments

The process of writing this thesis has been filled with both challenges and triumphs, making it a remarkable learning experience.

I am deeply thankful for Unilever providing me the opportunity to complete my Master Thesis and experience the life of a Data Analyst. A special thanks to my manager, Ross Olivier, and my colleague, Onkar Pratik, for offering me unwavering support and having the confidence to carry out this project. Your belief in my abilities has been a source of motivation and has greatly contributed to the completion of this thesis.

I would like to express my deepest gratitude to my coach, Sebastian Gabel, and co-reader, Jason Roos. I am incredibly grateful for all the support and feedback you have provided me within this process. Your guidance has not only enabled me to learn countless new things but also helped me explore potential applications that I had never considered. Your expertise and suggestions were invaluable during this process and I truly appreciate having the opportunity to work with you.

I am also profoundly grateful to my parents for allowing me to express my doubts and talk about my thesis non-stop. Finally, to my partner, thank you for your patience, understanding, and constant support during the ups and downs of this process. Your unwavering support has been my anchor.

Preface

Executive Summary

Finding the competitive advantage within the Fast Moving Goods Consumer (FMCG) industry is a complex task for Consumer Packaged Goods (CPG) manufacturers. Gaining an edge within the sector depends on accurately forecasting Stock-Keeping-Unit (SKU) sales value share, which enables to optimize product portfolios and amplify category management. Majority of developed forecasting systems rely on statistical time-series methods. However, these methods face multiple limitations, making it difficult to capture complex effects. Recently, a new wave of research utilized ensemble machine learning for time-series forecasting problems.

This study further expands this research field and explores how ensemble ML models, Random Forest (RF) and Gradient Boosting Regressor (GBR), can be applicable to SKU sales value share forecasting and enhance category marketing planning. By producing a 12 week forecast, this thesis compares the predictive performance of ensemble ML models with a statistical time-series method, Prophet. Moreover, the study investigates how to improve RF performance by transforming SKU descriptions into text embeddings to capture latent features, and testing the Targeted Random Forest (TRF), which uses LASSO regularization to target predictors. Lastly, the research explores SHAP values, a novel explainable AI method, to address the black-box nature of ensemble ML models.

The study concludes that GBR achieves the best predictive performance among the tested methods. However, TRF significantly reduces the computational complexity and achieves comparable results. Ultimately, SHAP values are recommended for interpreting ML outputs, as they provide a glance into the feature importances for the global model output and individual predictions.

Abbreviations and their Explanations

Abbreviation	Explanation
ADL	Autoregressive Distributed Lag
ARIMA	Auto-Regressive Integrated Moving Average
CM	Category Management
CPG	Consumer Packaged Goods
ES	Exponential Smoothing
ETS	Exponential Trend Smoothing
FMCG	Fast Moving Consumer Goods
GAM	Generalized Additive Model
GBR	Gradient Boosting Regressor
LASSO	Least Absolute Shrinkage and Selection Operator
MAE	Mean Absolute Error
ML	Machine Learning
PCA	Principal Component Analysis
POS	Point of Sales
RF	Random Forest
RMSE	Root Mean Squared Error
SARIMAX	Seasonal Auto-Regressive Integrated Moving Average with eXogenous variables
SHAP	SHapley Additive exPlanations
SKU	Stock Keeping Unit
SMAPE	Symmetric Mean Absolute Percentage Error
TDP	Total Distribution Points
TRF	Targeted Random Forest

Contents

1. Introduction

1.1 Problem background

Within the complex and dynamic world of retail, stakeholders are constantly seeking ways to simplify processes and achieve impactful results. One powerful strategy is category management (CM), a cornerstone process for retail businesses that involves coordinating product categories as distinct business units (Zenor, 1994). It allows retailers to manage merchandise assortment, shelf-space allocation, promotions and pricing more efficiently, as they can make decisions across categories (Basuroy et al., 2001; Dupre & Gruen, 2004). In fact, German retailers rated "optimization of product portfolio and category management" the most important task for achieving performance goals (Hübner & Kühn, 2012), and this perspective is echoed in recent research (Heger & Klein, 2024). This view marks a prominent deviation from traditional brand-centric merchandising strategies and is increasingly embraced by retailers, particularly in fast moving consumers good (FMCG) industry (Wedel et al., 2015). Thus, considering the competitiveness of the retail market, large suppliers[1] who have been heavily relying on their brand-power must also shift their approach to become more category-centric.

One such supplier is Unilever, a multinational FMCG company with a diverse portfolio of brands within the food, personal care, home care, beauty, and ice cream industries. In July 2022, the company made a significant operational change, abandoning its previous structure of having the combined divisions of Beauty & Personal Care, Home Care, Foods & Refreshment (Unilever, 2022). The focus shifted to the creation of 5 new business groups (Nutrition, Personal Care, Home Care, Beauty & Wellbeing, Ice Cream) that would be responsible for their own strategy and growth. One of the reasons cited for this restructuring was to make Unilever a simpler, more category-focused organization (Unilever, 2022). Yet, a critical aspect of achieving success as a category-focused organization is optimizing the product portfolio to enable sustained growth across various retailers (Timonina-Farkas et al., 2020). Central to this is developing a forecasting system that can accurately predict the sales value share of its Stock Keeping Units (SKU), ensuring that the right product is available at the right place, at the right time.

Sales value share is a key metric in Unilever, as it measures how a product, brand, or category is performing against competitors in the retailer. Changes in share directly mirror consumer trends and preferences, which makes it even more important to ensure that products on the shelf align with evolving consumer desires (Huang et al., 2014). Accurately predicting which products will perform best at the SKU level within a retailer offers numerous benefits, as precise forecasts can provide valuable insights when formulating retailer strategy (from supplier side) whilst also helping marketing professionals plan promotional campaigns more effectively

[1]Consumer Packaged Goods (CPG) manufacturer's will be referred as suppliers within this thesis.

(Ali et al., 2009). Therefore, it is necessary to incorporate accurate and interpretable sales value share forecasts at the SKU-level to generate further insights about products that succeed and the factors that cause them to do so.

In marketing practice and scholarly literature, there is a rich history of standard econometric models that forecast demand and share. State-of-the-art models such as Exponential Smoothing (ES), Auto-Regressive Integrated Moving Average (ARIMA), as well as more complex models, SCAN*PRO and CHAN4CAST, have been widely tested and applied within the FMCG retail industry (Falatouri et al., 2022; Fildes et al., 2019; Spiliotis et al., 2020). However, these models come with a set of limitations that could introduce bias, diminishing their performance. In academic literature, new approaches of using Machine Learning (ML) models for forecasting are gaining significant traction, as they can overcome challenges that traditional methods pose. Firstly, traditional methods often assume linearity, while ML models can capture non-linear patterns, which enables them to detect more complex effects (Barker, 2020; Petropoulos et al., 2022). This is especially relevant in the FMCG retail industry, as changes in competitor pricing and promotional effects could vary greatly in magnitude, influencing share in complex ways. Secondly, more complex ML models, such as Random Forest or Gradient Boosting are inherently designed to handle sparse datasets, while traditional models often require a complete time-series with potential imputation of data beforehand (Hapfelmeier & Ulm, 2014). This can be beneficial when forecasting at the SKU level, as such granularity can introduce inconsistencies and gaps, like the lack of sales in certain periods. Using ML also decreases the potential bias introduced by imputing specific data that is necessary for traditional models, as no assumptions about the data generating process is made. Lastly, traditional methods often suffer from ”the curse of dimensionality”, where the performance and interpretability of the model degrade as the number of dimensions (features) increases (Ma et al., 2016). ML methods, on the other hand, are equipped to reduce dimensionality and capture interactions between features within high dimensional data without sacrificing the predictive power (Spiliotis, 2023, pp. 55-56). This becomes particularly important considering the cross-product effects that can exponentially increase the number of features in the dataset (Ma et al., 2016). Ultimately, multiple studies showcase ML models superior performance compared to traditional forecasting models (Fildes et al., 2022; Huber & Stuckenschmidt, 2020). Researchers suggest that the observed effectiveness stems from employing global learning, a method where a single model is utilized to predict multiple time series, allowing identification of patterns across them (Spiliotis, 2023, pp. 55-56). However, many of these models operate as black boxes, which means their outputs can be difficult to interpret. Due to this constraint, practitioners in the field opt for the aforementioned simpler, more interpretable models as the outputs from ML models can appear untrustworthy (Ali & Pinar, 2016). Nevertheless, more methods of interpretable ML are being developed, such as SHAP values, and it is imperative to investigate their efficacy (Fildes et al., 2022). Exploring these methods could bridge the gap between advanced forecasting accuracy and the transparency needed for strategic decision-making, enabling suppliers to not only predict

but also understand why particular products are expected to perform well within the retailers.

1.2 Problem statement and research questions

Henceforth, the primary challenge of this research is to explore how can ensemble ML models compare to time-series forecasting methods in terms of accuracy when predicting sales value share at the SKU level, and explore how can explanaibility methods help identify the drivers of the prediction. Thus, the following problem statement arises: *"How can ensemble machine learning models help develop a forecasting system to predict SKU-level sales value share, and thereby enhance category-level marketing planning in the FMCG retail sector?"* To answer this overarching question, multiple sub-questions are derived that will help to break it down:

1. How do ensemble Machine Learning models (Random Forest, Gradient Boosting Regressor) compare to statistical time-series forecasting methods (Prophet) in terms of accuracy for predicting weekly SKU level sales value share?

2. How does applying different high-dimensionality reduction techniques, such as a) transforming SKU descriptions into text embeddings, b) targeting predictors with LASSO, impact ensemble model performance and its interpretability?

3. What are the most important factors that predict sales value share and how do SHAP values help interpret the model output?

1.3 Academic relevance

This research intersects with multiple streams within the academic literature, primarily focusing on forecasting in the retail sector at the SKU level, the application of machine learning in retail analytics, and interpretability in machine learning models. This paper seeks to contribute to this body of knowledge primarily from a methodological perspective in four ways: a) examining forecasting from the supplier's standpoint instead of the retailer's, b) employing SKU descriptions as text embeddings and using them as features in the modeling process, c) evaluating a Targeted Random Forest (TRF), and d) analyzing feature importance with SHAP values.

Firstly, forecasting at the SKU-level is an extensively covered topic in academic literature from various fields, including supply chain management and marketing management (Ali et al., 2009; Spiliotis et al., 2020). However, much of the literature addressing SKU-level forecasting embraces a retailer's viewpoint, focusing primarily on developing models that accurately predict sales volume for restocking purposes (Babai et al., 2021; Ma, 2024). This study adopts a supplier viewpoint, which transpires in two following ways. Firstly, the study specifically focuses on forecasting at the retailer-chain level instead of the store-level. Majority of the studies of retail forecasting focus on predictions on the store-level, which are relevant for retailers in terms of inventory management and producing specific marketing-mix per store (Andrade & Da Cunha, 2023; Gupta et al., 2021; Jin et al., 2015; Ma et al., 2016). However, professionals

working in customer strategy department within the supplier organization have to make high-level marketing decisions regarding the whole retailer chain. Few papers have applied forecasting at the chain level, highlighting the need for more research in this area to provide insights at a higher aggregation level (Baltas, 2005; Curtis et al., 2014). Secondly, this study seeks to predict the sales value share. While a substantial portion of existing literature emphasizes estimating demand as the primary dependent variable (Aburto & Weber, 2007; Babai et al., 2021), the sales value share may prove more useful for marketing purposes. This is because it offers a comparative view of performance relative to competing products, something that demand alone may not fully capture. Thus, this research has the potential to provide new insights into how forecasts can help manufacturer's gain a deeper understanding that informs and enhances their marketing strategies.

Secondly, one of the key aspects of modelling an SKU-level forecast is creating a feature space that can account for granural effects. Capturing all potential effects at the SKU level can lead to a large feature space, especially when accounting for intra-category pricing and promotional effects (Ma et al., 2016). Including product attributes as predictors adds further complexity, resulting in a high-dimensional modeling environment. Many studies have focused on developing models capable of navigating this intricate space effectively (Ma et al., 2016; Ma & Fildes, 2017). This study aims to expand upon existing research by leveraging a method that has recently been employed in product-choice modeling - incorporating SKU descriptions as predictors through text embeddings, potentially offering three significant benefits (F. Chen et al., 2020; Gabel & Timoshenko, 2022). Firstly, embeddings provide dense information about product attributes, enabling the extraction of rich features. Secondly, they facilitate scalability by efficiently summarizing these features, allowing the model to handle diverse products and categories effectively. In product-choice modeling literature, embeddings have proven effective at creating scalable models that capture nuanced product variations (F. Chen et al., 2020; Gabel & Timoshenko, 2022). Lastly, text embeddings enable to capture semantic similarity between products. Within the choice-modelling literature, it is widely acknowledged that the similarity of products impacts decision-making of consumers (Medin et al., 1995; Ye et al., 2019), due to the potential substitution and complementary effects. While embeddings have been applied to sales forecasting in various fields, such as predicting hotel booking cancellations and home appliance sales (S. Chen et al., 2023; D. Li et al., 2022), only one paper has utilized them specifically for demand forecasting in the FMCG retail sector (Vallés-Perez et al., 2022). Therefore, this study will compare to what extent can text embeddings minimize the error rate as compared to models with one-hot encoded product attribute features, offering an insight into their effectiveness.

Thirdly, to further tackle the high-dimensional feature space, this study aims to address a methodological research gap by testing a novel method of targeting predictors with LASSO before applying the Random Forest (RF) algorithm. While this approach has been partially explored in other contexts, such as using LASSO to select predictors before fitting models like the Autoregressive Distributed Lag (ADL) model (Ma et al., 2016), only one study has applied

regularization before utilizing RF (Borup et al., 2023). This is important because the study proved that the two-step approach, denoted as Targeted Random Forest (TRF), was successful in macroeconomic forecasting, with a 13% increase in accuracy compared to regular RF (Borup et al., 2023). However, its generalizability to other fields, such as predictions at the SKU level, remains unproven. By targeting predictors before applying RF, this study seeks to demonstrate how such an approach can improve computational efficiency and enhance forecast accuracy as compared to the regular RF.

Finally, expanding the literature on interpretable ML methods, this study examines how SHAP values can make models understandable and actionable for decision-makers within the FMCG industry. While there have been some studies how SHAP values can aid interpretation of feature importance within other industries (Nohara, 2022; Wang et al., 2022), only one study in the retail forecasting space has explored how SHAP values can help with ensemble machine learning model interpretation (Antipov & Pokryshevskaya, 2020). Moreover, Fildes et al. (2022) in their retail forecasting literature review have identified that interpretability of ML in retail forecasting is limited and is could be a promising direction for future studies. Thus, this research not only contributes to the theoretical understanding of interpretability of ML but also provides practical insights into how such methods can be applied in real-world retail scenarios.

1.4 Managerial relevance

The ability to accurately forecast SKU-level sales value share within a retailer and interpret these forecasts holds significant managerial relevance for Unilever and the broader FMCG sector, impacting strategic decision-making across multiple levels of the organization. This relevance is underlined by three key factors. First, the capacity to anticipate SKU-level performance within the retailer enables a shift from reactive to proactive marketing strategies, allowing Unilever to optimize their product mix and promotional efforts. Second, in an industry characterized by fierce competition and slim profit margins (Kenton, 2024), accurate forecasts offer a substantial competitive advantage. They can enable to make decisions informing strategies that can enhance market share and profitability (De Almeida & Da Veiga, 2022). Third, the application of explainability methods that help interpret the output of complex ML models ensures that the strategic recommendations made are not only accurate but also transparent and understandable. This clarity fosters trust both within the organization and in its interactions with retail partners, facilitating collaborative efforts to capitalize on forecasted market opportunities. Overall, accurate predictions of market outcomes are of utmost importance for management decisions. The necessity for retail analytics to guide retail strategy is evident, as poorly informed decisions can lead to significant repercussions, including declines in sales that may prompt decisions to delist or discontinue products (Fisher, 2020).

Multiple stakeholders can significantly benefit if the solution is successful. Firstly, category managers can gain significantly from leveraging these forecasts in product assortment decisions. For instance, if the forecast for the next 12 weeks indicates that certain SKUs are likely to have

a declining sales value share, category managers can use this insight during negotiations with retailers to adjust product assortments, focusing on removing or replacing underperforming products with those predicted to excel, thereby maximizing the sales potential of the category. Secondly, marketing teams can refine their promotional strategies with SKU-level forecasts, targeting their efforts on products expected to lead the category in performance. This enables the creation of focused marketing campaigns that highlight the benefits and availability of these outperforming SKUs. For example, if forecasts predict growth for specific SKUs due to seasonal demand or emerging trends, marketing teams can initiate targeted promotions campaigns ahead of these periods, capitalizing on such momentum. Finally, the customer strategy and planning teams, tasked with crafting overarching strategies for retailer engagement, can utilize these forecasts for building strategies per category and retailer basis. For example, if the forecast reveals that select SKUs are likely to significantly outperform competition in the coming months, these teams can negotiate for better shelf positioning or enhanced promotional activities for these products, ensuring they capture maximum consumer attention and drive category growth.

1.5 Research Approach

The data for this study was obtained from NielsenIQ, which spans from March 2021 to April 2024, offering 164 weeks of Point-of-Sales (POS) data for the UK market within two major retailers. Utilizing this data, this research aims to examine and develop a thorough and scalable model system that can be used to predict SKU-level sales value share. There are five key nuances to the forecasting approach.

Firstly, the forecast occurs at the SKU-chain-weekly aggregation level, aiming to predict the next 12 weeks. The level of aggregation is extremely important to consider in terms of model choice, as data granularity directly impacts the number of observations and cross-product effects (Zotteri et al., 2005; Zotteri & Kalchschmidt, 2007). Moreover, forecasting at the SKU-weekly level allows to aggregate to higher levels of time and space dimensions. In other words, this means that predicting at such granularity allows to accumulate predictions to a brand attribute-level or a monthly-level prediction.

Secondly, it is considered that accurately predicting products with larger sales value share is more important than predicting products with smaller shares. Products with large share make-up a larger shelf-space and receive tremendous marketing investment, which signifies their importance to the retailer and the manufacturer. This has an implication for the choice of the primary evaluation metric of this research - Root Mean Squared Error (RMSE). RMSE, with its sensitivity to larger errors, emphasizes the importance of minimizing significant forecasting discrepancies. Achieving a low RMSE is crucial for avoiding major strategic and operational consequences, thereby aligning model evaluation with the goal of ensuring robust and reliable share forecasts (Mehdiyev et al., 2016).

Thirdly, considering the granularity necessary to capture effects at the SKU level, the final dataset might end up being high-dimensional and sparse, which could lead to potential

overfitting due to the noise (Ma et al., 2016). For example, when accounting for cross-product promotional effects, this would mean creating a predictor for each SKU and its marketing type, leading to sparse dataset. This is tackled in two ways. Firstly, this is addressed by the introduction of text embeddings as features, as they can transform categorical variables into a lower-dimensional, dense vector space. The pre-trained OpenAI *text-embedding-3-large* model will be used to transform SKU descriptions into embeddings, which is particularly useful since it reduces the necessary computational resources required for training. Ultimately, the generated embeddings can be used as input features within the ML model. Secondly, ensemble methods, such as Random Forest and Gradient Boosting Regressor are well suited to evaluate such data. However, specifically for RF, some studies suggest that their benefits may be lessened in sparse settings due to weak predictors (Borup et al., 2023; Gentzkow et al., 2019). Therefore, this study will utilize and evaluate TRF, a two-stage RF approach that utilizes LASSO prior to applying the ensemble method.

Finally, considering that one of the study's aims is to make black-box ML models more explainable, the interpretability methods are crucial to consider. Thus, ML models will not only be interpreted with feature importance analysis, but also with SHAP(Shapley Additive exPlanations) values. SHAP values is a recently introduced methodology that offers in-depth insights regarding feature importance given a particular model output (Linardatos et al., 2020). SHAP values offer local interpretability (understanding the prediction for an individual instance) and global interpretability (understanding overall model behavior). It discloses how much each feature contributes to each prediction, and it provides the direction of the impact (positive or negative) (Antipov & Pokryshevskaya, 2020). Thus, understanding how SHAP values can help to identify relevant features can enhance marketing planning significantly. On the other hand, feature importance analysis only provides a global view and tells which features are generally most important for making accurate predictions. Thus, utilizing SHAP values could be imperative to increase interpretability of ensemble ML models.

1.6 Overview of the Thesis

This thesis is organized into six chapters besides the current one. Chapter 2 presents a literature review on share modeling, the relevant methodologies utilized, and the theoretical foundations of forecasting at the SKU level. Chapter 3 discusses the data utilized for this research. Chapter 4 details the methodological approach and analyses employed in this study. Chapter 5 assesses the results of these analyses and provides an answer to each research question. Chapter 6 provides a discussion of the findings, their academic and managerial implications, as well as the study's limitations, and directions for future research.

2. Literature Review

The FMCG retail industry is a dynamic and competitive market which is increasingly complex due to the entangled and competing interests between suppliers and retailers. Due to its low profit margins, both retailers and suppliers are in a constant search for a competitive advantage that could put them ahead against their peers (Jackson et al., 2023). Accurate SKU-level sales forecasting emerges as a critical tool in this environment, aiding in assortment planning, inventory management, and marketing strategy formulation. However, the complexity of forecasting in this sector is magnified by the intricate interplay of factors influencing product share of sales, making it a formidable challenge.

The literature review that follows will provide a holistic overview that spans across three key areas: 1) the intricacies of modelling sales share and identifying its key predictors, 2) the application of statistical and ML forecasting methods at the SKU-level, assessing their strengths and limitations, and 3) high-dimensionality reduction techniques and the trade-off between accuracy and interpretability. These sections will collectively address the theoretical underpinnings of each research question and provide a lens to evaluate the study's findings.

2.1 Modelling Sales Value Share: Background and Influential Features

This section will provide a broad overview of practice and literature that has focused on product share predictions. Section 2.1.1 will provide the literature review on modelling share, whilst section 2.1.2 will explore the potential features that need to be accounted for when forecasting share.

2.1.1 Background of modelling share

Gaining market share is a key measure of business success in the FMCG industry (Geurts & Whitlark, 1993). Predicting product sales value share, rather than pure demand, offers several advantages, especially in highly competitive and dynamic retail environments. Firstly, modelling share allows to gain insight from a competitive lens, as it showcases product's performance relative to its competitors (Brodie et al., 2001). This focus on market share predictions allows managers to adapt more effectively to market changes and plan interventions that could enhance product performance. Additionally, predicting share offers a better understanding of consumer preferences and competitive dynamics than pure demand forecasting. This study specifically focuses on forecasting share within a retailer, rather than the entire market, to provide insights at the retailer-chain level.

Within scholarly literature, there have been theoretical and practical efforts to understand when building a share forecasting model is necessary, and which models are the most effective

to forecast (market) share (Geurts & Whitlark, 1993; Brodie et al., 2001; Cain, 2005). Theoretically, one of the influential guiding papers is by Brodie et al. (2001), which summarizes the key criteria under which forecasting market share via an econometric model is more useful than simply predicting naively. According to the authors, such modelling techniques are strongly favored over naive models in the following situations: (1) when there is a substantial sample size, typically around 100 periods for training and validation; (2) when the marketing instruments have significant immediate effects; (3) when models are estimated with brand-specific response parameters; and (4) when using store-level scanner data (disaggregated data) instead of aggregated market-level data (Klapper & Herwartz, 2000). However, this research slightly deviates from these principles. Whilst (1), (2), and (3) are followed, this study does not have store-level scanner data, but rather only the POS data within a retailer-chain. Recognizing this higher aggregation level is important, because it does not allow us to capture store-level insights or identify in which specific stores individual marketing-mix effects occurred. Nevertheless, the decisions made from a supplier's point of view are rarely on a store basis; rather, they are planned across the entire retailer chain or at most the channel level, which requires aggregated insights across the chain (Alldredge et al., 2017).

Moreover, a large part of studies that evaluate market share modelling techniques are within the scholarly scope of product choice modelling. Within these studies, there is a huge focus on applying multinomial-logit (MNL) models and their variations, specifically at a higher product aggregation level - the brand. For example, Fader (1993) combined the Dirichlet-multinomial model with the multinomial logit model for brand choice, demonstrating that this integrated approach offers nuanced explanations of the impact of marketing mix variables across various product categories. In addition, Agrawal and Schorling (1996) evaluated the forecasting capabilities of artificial neural networks against multinomial logit models, concluding that while neural networks performed well, the multinomial logit model remained a reliable and interpretable method for predicting brand shares in grocery product categories. However, there is one specific issue with the basic multinomial-logit models and it is the assumption of independent and identically distributed (i.i.d) error terms, which leads to the Independence of Irrelevant Alternatives (IIA) property, causing unrealistic substitution patterns based solely on market share (Cain, 2005). Such a limitation makes MNL inefficient in predicting real-world choices, as it does not allow flexible substitution patterns to emerge. This is important to consider for this paper, as it focuses on a forecast at the SKU-level, where products can vary significantly in attributes and consumer perception, unlike a specific brand (Fader & Hardie, 1996). Yet, this study suffers from the absence of household-level panel data, which limits the ability to perform detailed choice modeling, as we rely solely on POS data within a retailer-chain. This restricts the analysis of individual consumer preferences and purchase behaviors over time, which are critical for accurate choice modeling. Nevertheless, this study will utilize insights from the product-choice modelling literature in order to develop an effective approach for SKU-level sales value share forecasting.

2.1.2 Influential Features

One of the key insights generated from product-choice modelling literature is how different explanatory variables influence the sales value share. There are three key variable clusters that encompass the different effects: marketing mix, cross-product effects, and seasonality and trend.

Marketing Mix

The Marketing Mix consists of the 4Ps: Product, Price, Place and Promotion, which all have proven to be imperative when considering forecasting share of sales. Firstly, the product aspect relates specifically to the product observed attributes like brand, pack size, pack type and variants. These attribute can all potentially influence SKU share of sales, shaping consumer perceptions and purchase decisions. This was successfully showcased by Fader and Hardie (1996), who theorized that consumers are choosing on the basis of the product attributes, which tend to be discrete and tangible. Secondly, the price factor has a substantial role in influencing consumer buying behavior. Price sensitivity varies across different FMCG product categories, which can influence how price changes (discounts) can impact the share of sales at the SKU. Multiple studies have showcased that consumers significantly respond to price discounts and perceived value, which can lead to an increase in the share of sales for promoted SKUs (Mamuaya, 2024; Venkatesan & Farris, 2012). This is important, as a survey last year showed that groceries is the sector where consumers have been the most price-conscious (Bansal, 2023). Thirdly, place, or distribution is essential for ensuring that products reach the retail shelves on time and in optimal condition, guaranteeing product availability within stores. Experts and multiple scholarly papers claim that distribution could be one of the most important factors that impact market share (Hirche, Greenacre et al., 2021; Wilbur & Farris, 2014). In fact, one paper suggested that 84% of the variation in market share can be statistically attributed to distribution (Kruger & Harper, 2006). This is supported by Broniarczyk et al. (1998), who demonstrated that wider distribution and better shelf positioning can increase visibility and sales, indicating that place is a vital element in the marketing mix. This is especially relevant for share forecasting, as accounting for the changes in distribution levels can help identify an additional driver in the performance of an individual SKU. Lastly, promotions in the FMCG retail grocery sector are a common strategy employed by retailers to clear inventory and boost short-term sales (Krafft & Mantrala, 2010). These can range from price discounts, multi-buy offers, features, and extra Point of Sales (POS) displays (Krafft & Mantrala, 2010). This paper will cover multiple type of promotions, including multi-buy offers, features, displays and temporary price reductions. The effectiveness of promotions varies significantly, influenced by factors such as timing, market saturation, and consumer demand (Chandon et al., 2000; Huang et al., 2019).

Intra-category effects

Secondly, the changes in the aformentioned factors does not only affect the market share of one SKU, but it can have wide-ranging influence. This is otherwise known as intra-category product effects, which play a pivotal role in influencing the market share of products. The

dynamics within and between product categories significantly impact consumer choice and, subsequently, share. Scholarly research underscores the profound influence that changes within one product have on sales in others, through mechanisms of substitution or complementarity (Leeflang & Parreño-Selva, 2011). These intra-category dynamics, where products within the same category compete or complement each other, are central to understanding market behavior. This is especially relevant for forecasting, as incorporating an analysis of within and cross-category effects has been shown to enhance the accuracy of sales forecasts (Ma et al., 2016). This thesis specifically incorporates only within-category effects and does not account for cross-category effects, as previous research has shown that the most accuracy improvement comes from incorporating within-category competitive effects (95%) and only (5%) from inter-category (Ma et al., 2016). However, when accounting for these effects in the model, one must note that it often leads to a high-dimensional and sparse data environment. This looks like feature-space being in thousands of variables, which can lead to a biased and overfitting model. Nonetheless, this is one of the primary reasons why ensemble ML models are being evaluated, as they can deal with such a high-dimensional space efficiently.

Seasonality, Occassions and Weather

Finally, seasonality and special occasions significantly shape market dynamics and play a crucial role within the FMCG retail grocery sector when considering product market share. The cyclical nature of consumer demand, driven by seasons, holidays, and cultural events, profoundly impacts purchasing patterns, thereby affecting market share across various product categories (Fildes et al., 2019; Huang et al., 2019). Consequently, majority of forecasting models incorporate seasonality and occasion-based variables to predict market share fluctuations accurately (Dekker et al., 2004; Huber & Stuckenschmidt, 2020). This involves analyzing historical sales data to identify trends and patterns associated with different times of the year and special events. Specifically for Unilever, incorporating seasonality variables, such as holidays, is very important, as some categories are heavily impacted by seasons and their nature, such as weather during the summer. In the context of our forecasting category, dressings, weather plays a significant role, particularly in the summer months, which are critical for sales due to their association with barbecues and outdoor gatherings. Dressings, such as salad dressings, marinades, and sauces, are essential accompaniments to grilled meats and vegetables, which can spike the overall sales. Therefore, this thesis will incorporate the weekly average temperature as well as the average precipitation within its features.

2.2 Forecasting at the SKU level

This section will synthesize literature of the methods that have been applied when forecasting at the SKU level. Section 2.2.1 will overview the econometric models, while section 2.2.2 will review the machine learning models.

2.2.1 Econometric Models

Econometric models can be divided in two forms: univariate and multivariate. Firstly, univariate models focus on forecasting based on a single time series data of the SKU's past sales or demand. These models do not account for external variables and are primarily used when the interest lies solely in the historical data of the product itself. The most popular methods for forecasting SKUs are ARIMA (Autoregressive Integrated Moving Average) and ES (Exponential Smoothing), as well as its multitude of variations. ARIMA models are synonimous to the Box-Jenkins method and are widely applied for time series forecasting, particularly valued for their flexibility (Petropoulos et al., 2022). ES models are favored for their simplicity and efficiency in capturing trends and seasonal patterns. They are quite robust and not as prone to overfitting as more complex methods (Petropoulos et al., 2022). Ramos and Oliveira (2016) has identified that such models also work better at higher aggregation levels, which is important to take into account when measuring their performance. However, one of the limitations of both ES and ARIMA models is that a continuous time-series is required, meaning that your data cannot have gaps and inconsistencies. This limits their applicability to forecast SKU share of sales, as SKU sales can be intermittent, meaning that they lack sales/are not available during certain periods. Most univariate methods are mostly used as benchmarks, meaning they are the comparative baseline for other tested model performance (Fildes et al., 2019).

Secondly, multivariate econometric models analyze multiple variables to predict outcomes and understand relationships between them, enabling more complex and accurate forecasts. Firstly, Vector Autoregression (VAR) has been a pretty popular method to account for endogenous variables. For example, Curry et al. (1995) forecasted sales of canned soup at the brand level using Bayesian VAR and found it superior to univariate methods such as ES and ARIMA. However, the application was limited, as the signs and magnitudes of parameters could not be readily interpreted. Moreover, extensions and modifications of univariate ARIMA models have been applied. For instance, SARIMA and SARIMAX have been applied to forecast daily sales of food products. Falatouri et al. (2022) ran a comparative study comparing SARIMA to Long Short Term Memory neural network approach and found that SARIMA performs better when predicting seasonal product sales, and extending the model to SARIMAX by adding promotions improves the performance even more. Moreover, a cointegration model family is frequently utilized in retail forecasting, including such models as Autoregressive Distributed Lag (ADL) and Johansen's method. Such models are particularly useful in non-stationary time series data where it is assumed that a long-term equilibrium relationship exists, in other words, a stable, consistent association between two or more variables. For example, Huang et al. (2014) has applied ADL to forecast product sales at the SKU-chain level whilst including competitive pricing and promotion effects and has found that it outperforms all benchmarks. The ADL model has been widely applied in promotional forecasting as it can take into account endogenous variables and it is based on a simple regression style model structure, which makes it easy to interpret (Ma et al., 2016). Finally, models such as SCAN*PRO, PROMOCAST, and CHAN4CAST have been

widely examined in the academic literature (Cooper et al., 1999; Divakar et al., 2005; Van Heerde et al., 2002). These models are decision support systems that offer distinct advantages. For example, SCAN*PRO is renowned for its ability to disaggregate the effects of promotional elements and offer precise insight into promotional effectiveness (Van Heerde et al., 2002). CHAN4CAST differentiates itself by allowing to focus on distribution channel-level forecasting and can provide a nuanced view how different distribution channels contribute to overall sales performance (Divakar et al., 2005). However, both SCAN*PRO and CHAN4CAST are usually implemented and work within a store/channel-basis, making them unsuitable for this study, as the forecast is implemented at the retailer-chain level. Lastly, one of the newer forecasting methods is Prophet. It is a forecasting model developed by Facebook (Taylor & Letham, 2018) and there have been several applications of this model within the scientific literature (Jiang et al., 2021; Jha & Pande, 2021). At its core, this method is a generalized additive regression model (GAM) that can capture non-linear effects. Whilst being fundamentally univariate, this method also allows to add external regressors and capture multivariate effects. The primary benefit of this system is its robustness to outliers and missing data (Hyndman & Athanasopoulos, 2018; Lin et al., 2021; Taylor & Letham, 2018), which is key in SKU-forecasting, since the demand can be intermittent. Therefore, the Prophet model will be the primary statistical time-series forecasting model which will be compared to the Machine Learning models.

2.2.2 Machine Learning Models

When Fildes et al. (2019) published a retail forecasting review, they claimed there has not been enough wide-ranging evidence of the benefits that ML algorithms produce whilst forecasting. In three years, the landscape has shifted, and a post-scriptum release of the review states that research in ML has significantly advanced, and ML applications are signaling superior performance compared to traditional time series or linear regression methods (Fildes et al., 2022). This section will provide a succinct overview of application of ML methods within the FMCG retail space.

One of the first studies that forecasted SKU sales with ML techniques evaluated Support Vector Machines (SVMs) and Regression Trees (RT) (Ali et al., 2009). This study found that regression trees with rich input data of derived explicit features improved forecast accuracy by 24%, especially during promotion periods. More importantly, they found that increasing the scope of the model, also known as pooling more data, resulted in the best accuracy. This is particularly relevant for our study, since one of the aims is to build a model that can predict for different retailer chains across multiple categories. Moreover, ensemble methods, such as Random Forest (RF) and Gradient Boosting (GB), are becoming more popular due to their innate ability to deal with high-dimensional data and capture non-linear patterns. In one large comparative study of forecasting daily demand, where 11 statistical and 7 ML methods were compared, RF ranked second in terms of accuracy when trained in a series-by-series fashion, illustrating its ability not to overfit due to noise (Spiliotis et al., 2020). However, the study

found that cross-learning, which references the model's ability to learn from concurrent SKUs, only increased the accuracy of Neural Networks and did not improve its accuracy across all ML models, such as RF, SVM and RT. Authors cite that potential causes of this could be the lack of explanatory and hierarchical data. This finding is particularly relevant for our research, as it aims to incorporate explanatory data, such as product attributes and within-category effects, and implement a model that benefits from cross-learning. Furthermore, another paper forecasting SKU sales compared Elastic Nets, RF and Gradient Boosting Machines (GBM) and incorporated within-category price and promotional effects to account for potential non-linear impact(Antipov & Pokryshevskaya, 2020). Both non-linear methods outperformed the regularized regression, with Gradient Boosting being the most accurate, illustrating the potential of both methods within SKU-level forecasting. However, one of the limitations of the study was that it only predicted the sales of 13 products, which limits the findings' generalizability and robustness to other products with different seasonalities and statistical properties. This study will evaluate the performance of Random Forest and Gradient Boosting Regression on 583 unique products, aiming to provide a more comprehensive and robust analysis across diverse product categories.

Neural Networks and other deep learning approaches also have received tremendous attention within research. A large body of literature has examined different types of neural networks, such as fuzzy neural networks, back-propogation neural networks and LSTM neural networks, and majority of these studies have found that these non-linear models perform better than linear regressions (Aburto & Weber, 2007; Falatouri et al., 2022; Kuo, 2001). However, there are several challenges when implementing neural networks. One significant challenge is the requirement for large datasets to train effectively, which may not always be available or feasible to collect. Even if the data requirement is achieved, deep learning models can be resource-intensive to train to achieve reasonable training times. Finally, these models are often viewed as "black boxes" due to their complex internal mechanisms, making it difficult to interpret how predictions are made. For instance, Punia and Shankar (2022) combined RF and LSTM to get a more interpretable model output, as they claim that RF is a more interpretable method. Due to these constraints, this thesis will not investigate neural network applications to forecasts and focus on ensemble machine learning methods.

2.3 High Dimensionality Reduction Techniques

One of the key issues encountered within SKU-forecasting literature is the high-dimensional feature space that arises from primarily two sources. Such high-dimensionality is introduced when accounting for cross-product effects (Ma et al., 2016) as well as when utilizing the one-hot encoding technique to represent product attribute categorical variables (Mezzogori & Zammori, 2019). This section will provide a short overview of how studies dealt with this high-dimensionality and its implications.

Several papers have investigated different methodologies to reduce the high dimensional feature space. One of the most popular techniques was to use stepwise regression to select

important variables (Fildes et al., 2019; Huang et al., 2014). However, as research progressed, LASSO shrinkage, a regularization technique became more popular, as some studies showcased that stepwise regression picked irrelevant variables (Flom & Cassell, 2007). Thus, multiple papers have adopted different shrinkage methods, such as LASSO, Ridge and Elastic Net when reducing the features. For instance, Huang et al. (2014) combined stepwise regression and LASSO prior to applying ADL to forecast SKU sales. Moreover, Ma et al. (2016) developed a thorough methodological framework of a multistage PCA-LASSO regression to reduce dimensionality. Their study considered both within and cross-category price and promotional effects, and they had found that the multi-stage LASSO, a 3 step approach which modelled SKU sales on own product effects, within-category effects and cross-category effects, performed significantly better than one-stage LASSO. Lastly, Ramos et al. (2022) corroborated and extended these findings by exploring the use of both the PCA and Ridge regression to handle the many drivers present at the SKU-store level. They found that both PCA and the shrinkage method improved forecast accuracy by 10% over benchmark models, which is a significant improvement. Thus, it is clear that reducing dimensionality is at utmost importance when forecasting at the SKU-level.

However, there are several implications that can be drawn from the studies. Firstly, majority of these papers that apply a high-dimensionality reduction techniques use a statistical model for forecasting, such as ADL and SARIMAX, instead of a Machine Learning model, like Random Forest. Recent research has shown that targeting predictors with LASSO before applying a RF can achieve better accuracy by 13% as compared to just applying a regular RF (Borup et al., 2023). This study showcased that the initial targeting step enhances performance of RF in high-dimensional settings, potentially due to improving the strength of individual trees. However, this study occurred within the macroeconomic and financial fields. Thus, this study will evaluate whether this novel approach, Targeted Random Forest (TRF), can also be efficient within forecasting at the SKU-level, and how does it perform against regular RF. Secondly, what is recognized within the studies is the potential impact that dimensionality reduction techniques have on maintaining model interpretability. For example, LASSO shrinks some feature coefficients down to zero, and whilst this simplifies the model, some potential subtle interactions amongst variables might be missed, which can lead to a loss of interpretability. This research will aim to examine this trade-off between potential improved accuracy and interpretability, and evaluate how different variations of machine learning models select the most important features.

The interpretability problem is further amplified when applying ensemble machine learning models due to their black-box nature. Within retail forecasting research, this was often their biggest criticism, despite often achieving higher prediction accuracy than their statistical counterparts (Fildes et al., 2022). The inherent methods, such as permutation feature importance provide a general sense of the most important variables, yet they do not explain the nature or the direction of the effect, making them insufficient for comprehensive insights. However, research into explainable ML methods is expanding, and one of the most popular methods

is using SHapley Additive exPlanations (SHAP) values. SHAP values are based on Shapley values from cooperative game theory and provide a unified measure of feature importance by attributing the change in the expected model prediction to each feature (M. Li et al., 2024). This method offers several advantages over traditional feature importance scores. Firstly, SHAP values ensure that the contribution of each feature is fairly distributed, reflecting its true impact on the model's predictions, including the direction of the effect. Secondly, it offers both local and global interpretability, meaning that local individual predictions can be interpreted. There have been limited studies incorporating SHAP values when forecasting sales, however, they demonstrate the use-case of SHAP values well. For instance, Antipov and Pokryshevskaya (2020) applied SHAP values to understand the contributions of various features in predicting sales. This approach provided 500 unique predictors and allowed to explain the model's predictions more effectively. Similarly, J. Chen et al. (2021) utilized SHAP values to interpret a neural network model for predicting Walmart's sales. They found that SHAP values helped identifying the most influential attributes across different dimensions, improving the model's transparency. Therefore, this study will utilize the SHAP library in Python and use it to interpret the effects of the tested ensemble machine learning models.

Table 2.1.
Summary of Research in FMCG Retail Sector Forecasting

Reference	**Forecasting Level**	**Method to Reduce HD[1]**	**Forecasting Method**	**Interpretation Method**
Ali et al. (2009)	SKU-store	-	Stepwise Regression, SVM, RT	
Huang et al. (2014)	SKU-chain	Stepwise regression and LASSO selection	ADL	
Ma et al. (2016)	SKU-store	Multistage LASSO regression	ADL	
Ma and Fildes (2017)	SKU-store	Stepwise linear regression	RT	
Antipov and Pokryshevskaya (2020)	SKU-store	-	Elastic Net, RF, GBM	SHAP
Ramos et al. (2022)	SKU-store	PCA, Ridge	SARIMA, ES	
Wellens et al. (2024)	SKU-store	-	GBDTs	SHAP

2.4 Academic Contributions

Given the described literature streams, this thesis will primarily focus on providing methodological contributions. The potential additional inputs can be utilized when constructing and evaluating forecasting systems. There are three key ways of how this study aims to add to the existing literature.

Firstly, one of the aims of the paper is to predict SKU share of sales value by employing ensemble machine learning models. Majority of the studies that examine (market) share forecasting apply econometric models, which presents a research gap. Whilst econometric models are interpretable, they have some key limitations, such as suffering from the curse of dimensionality when accounting for a large number of variables. On the other hand, ensemble machine learning models are more capable to understand complex interactions and handle high-dimensional data. Therefore, building on the recent success that machine learning applications had when forecasting pure demand, this study will evaluate whether these results also translate when forecasting share of sales value.

Secondly, this research will evaluate potential methods that can enhance performance of the ensemble machine learning models. Despite their inherent ability to handle high-dimensional data, research has showcased that sparsity can hinder them from reaching their true potential (Borup et al., 2023; Gentzkow et al., 2019). However, there is a lack of extensive research of methods that tackle this hindrance and amplify their performance. Thus, this thesis aims to fill this gap by adapting two methods. First, it will explore the use of text embeddings when transforming SKU descriptions to reduce high-dimensionality of data and potentially capture intricate patterns that one-hot encoded categorical variables cannot. Second, this study will assess the effectiveness of a Targeted (LASSO-regularized) Random Forest as compared to a non-regularized RF. Since one study has showcased this method's value when forecasting within the macroeconomic and financial fields (Borup et al., 2023), it is imperative to understand whether this is applicable to SKU share predictions, where ultra-high dimensional space is often present due to high level of granularity.

Finally, this study will aim to counter one of the limitations of ensemble machine learning models, which is their black-box nature. This aspect has been frequently their biggest criticism within retail forecasting research, despite the high prediction accuracy they have proven to achieve (Fildes et al., 2022). Thus, a key new stream of research is to utilize explainable AI methods to better understand and explain predictions. This study will utilize SHAP values to analyze feature importances and provide a view how both global model output's and local SKU-level predictions can be interpreted.

3. Data

3.1 Data Description

To conduct the forecast, Point-of-Sales (POS) data from the leading consumer intelligence company Nielsen has been provided by Unilever. NielsenIQ Discover is a large database that stores vast information on sales, pricing, promotions, and distribution data. This dataset provides detailed insights into whether a product was promoted (through features and displays), the extent of price changes, and its distribution across stores. The data spans from March 2021 to April 2024 on a weekly basis, offering 164 weeks of records. NielsenIQ data is available for at least one market, the United Kingdom. In this market, data from two major retailers is accessible within the food segment dressings, which comprises of six product categories: Mayonnaise, Mustard, Ketchup, Meat & Fish Sauces, Salad Dressings, and Other Dressings. Besides Unilever products, each category contains multiple competitors, and their respective POS data as well. This study will assess and evaluate the model across each available category and make predictions at the retailer level. Overall, the volume of the data for all the aforementioned categories across two retailer chains yielded over 1087 unique SKUs and 122,236 observations. Further inspection regarding the data description is available in the table below.

Table 3.1.
Initial Data Description Prior to Cleaning

Subcategories	**# Items**	**# Companies**	**# Brands**	**# Observations**
Ketchup	101	25	40	10,408
Mayonnaise	209	35	69	23,182
Meat & Fish Sauces	521	82	325	59,959
Mustard	71	17	37	7,525
Other Dressings	69	16	29	8,447
Salad Dressings	116	21	42	12,715
Total	1,087	125	535	122,236

3.2 SKU Selection

Moreover, there is some important context surrounding SKU-level weekly data. Within our dataset, SKU demand is rather inconsistent, meaning that an SKU can be sold during multiple weeks and then be de-listed from the retailer chain in the other weeks (see Figure A.1) [1]. Such demand is referred to as intermittent, meaning that it can be volatile, with random spikes and multiple zero values throughout its product life-cycle (Spiliotis et al., 2020). This could be due to many reasons, including, but not limited to distributional issues, de-listing from the retailer,

[1] Please click on the number to reach the Appendix, where the visualization is present. All numbers should be clickable

supplier changing the product attributes and formulation, and the product being released during a particular season only. Within this particular dataset there were 401 consistent SKU and retailer combinations and 963 inconsistent ones[2]. Despite these inconsistencies, there is also a need to forecast such data, as seasonal SKUs can be important to understand as they contribute to sales. This can also be seen in the the table below, where notably, both Ketchup and Mayonnaise categories have 35.1% and 23.6% of their total sales generated from inconsistent products.

Table 3.2.
Consistent and Inconsistent SKU Sales Figures by Segment

Subcategory	Consistent Sales ($)	Inconsistent Sales ($)	Inconsistent/ Total Ratio (%)
Ketchup	181,612,600	98,219,700	35.1
Mayonnaise	271,229,600	83,645,600	23.6
Meat & Fish Sauces	369,746,800	54,634,800	12.8
Mustard	55,296,700	3,644,520	6.2
Other Dressings	70,681,400	8,257,500	10.5
Salad Dressings	136,995,600	10,309,690	7.0

To select the relevant SKUs to forecast, a quantile analysis of the subcategory sales value was performed (see Table 3.3). Since the subcategories have non-uniform demand, meaning that some sell on average less than others, it was important to remove SKUs based on their average sales per week. Therefore, for every subcategory, the 25% quantile was found, which indicated a low level of sales, and the SKUs that had their average sales per week lower than the 25% quantile of their subcategory sales were removed from the dataset. This resulted in 559 Retailer-SKU combinations being removed. Nonetheless, it had a marginal impact on the total sales of the subcategories, as the sales value contribution of these Retailer-SKU combinations was less than 1% to their respective subcategory. Removing these items is not only beneficial from the computational perspective, but also from the strategic, as analyzing larger SKUs provides more actionable insights to the business.

Table 3.3.
Quantile Analysis for SKU Targeting

Subcategory	Total Sales Value ($)	25th Percentile per Week($)	25th Percentile Total Sales Value ($)	Contribution Percentage (%)	Consistent SKUs (Units)	Inconsistent SKUs (Units)
Ketchup	279,517,200	615.2	406,274.7	0.15	2	50
Mayonnaise	354,847,200	919.8	1,125,916.8	0.32	6	103
Meat & Fish Sauces	424,380,700	609.2	1,903,107.0	0.45	19	248
Mustard	58,941,200	84.0	24,296.6	0.04	1	36
Other Dressings	78,944,900	268.4	76,215.9	0.10	0	29
Salad Dressings	147,305,200	562.5	223,931.1	0.15	0	65
Total	1,344,936,400		3,759,742.1	0.28	28	531

[2]This number does not equal to the total number of unique items due to the fact that it is Retailer and SKU combination, as in one chain the SKU might be consistent, but in the other chain it is not.

4. Methods

The Methods chapter aims to explain the key aspects of methodologies that were used within the study. The chapter is divided into three critical sections to explain how the problem was approached. Firstly, section 4.1 delves into the data preprocessing techniques as well as the feature engineering process. Secondly, section 4.2 provides an overview of the models tested, approach to cross-validation and the reasoning behind the selection of key evaluation metrics. Lastly, section 4.3 explains the methods that were utilized to deal with the high-dimensional feature space.

4.1 Feature Engineering

This section will overview the data wrangling, including a missing value analysis, validity and outlier evaluation and the nuances behind creating the feature space.

4.1.1 Data Quality

Missing Value Analysis

The data extraction from the NielsenIQ Discover portal resulted in 77 features. Several steps were taken to ensure high data quality within these features. The data extracted had some inconsistencies and gaps, with a large number of features having more than 1000 missing values. The following paragraphs will describe the source of these missing values and what was done to amend these inconsistencies.

Firstly, the column 'Sales Value' had 4409 missing values. Since the data contained primarily of 1 week lag values[1], they were inspected to help infer whether these missing values arose from potential lack of sales within that time period. This was indeed the case, as the mean of those lags were approximately 6 euros, and 4351 of such values had sales less than 50 euros (see Figure A.5) . Thus, those values were filled with zeros. Variable 'Unit Weight (KGS)' had a similar problem, as the variable did not have a value for every missing value of the 'Sales Value'. However, since the SKU description contained the size of the item, the size was extracted from there using the 'regex' library in Python, as each size had either 'G', representing grams, or 'ML', representing milliliters, next to it. Only 4 items did not have any specific size in their description, thus, they were imputed with the mean size of the items from the particular subcategory they belonged in.

Secondly, it is important to take into account that the majority of the features that were extracted were 1-week lags of particular features, such as 'Sales Value', 'Incremental Sales Value', 'Price per KGs' and 'Total Weighted Distribution Points'. Thus, the first week had no

[1]When the data was extracted from Nielsen, it was extracted with a purpose to make a one-week-ahead forecast, and thus one week lags were selected from the database. However, as the thesis progressed, the objective changed to forecast for the next 12 weeks, and thus, the data had to be adjusted.

observations in these columns, which meant that this week had to be deleted from the dataset. Moreover, there were 28 of these lagged columns that had a very similar number of missing values (4000-5000), thus, they were inspected for a common cause. As seen from the missing value correlation analysis (see Figure A.6), majority of these values coincide. The primary reason for these missing values was the introduction or re-listing of SKUs (see Figure A.7). As seen from the graph, majority of missing values came at the start of the product life-cycle, whilst a small minority was during it, potentially indicating a relisting of the product. Thus, to mirror the reality and ensure that these lags are correctly represented, they were filled with zeros.

Thirdly, a lot of features that related to promotions had a significant amount of missing values, ranging from 78000 to 112000 observations. These missing values occurred systematically, as after inspecting the column 'Sales Value Any Promo Prev', which indicates the Sales Value generated from any type of promotion, it contained of 78612 observations that were equal to zero. This was very similar to the value of missing values (78607) for the column, 'Incremental price per Sales (KGS) Any Promo Prev', which indicated the price change due to a promotion. Therefore, it could be inferred that these missing values were due to the fact that no promotion occurred for that particular product within that period (see Figure A.8). Thus, these values were filled with zeros.

Fourthly, the missing values occurred within the distribution columns, with around 340 missing values in each of the Total Weighted Distribution Points columns. After inspecting the data, it was found that only 5 unique SKUs had this data missing. The data was missing throughout every week they were present in the retailer-chain, potentially indicating that this is missing data at random. Therefore, to cover these values, the data was imputed with the mean distribution of the brand that these items belonged to. This was done to ensure that the distributions would be similar to those of that particular brand that had values within the store. Finally, the last column that had more than 1400 missing values was 'Value/Store Prev'. After doing a root-cause analysis, it was found that this was due to the re-listing and de-listing of SKUs, which meant that for particular weeks, the value was NA, and thus, those rows were filled with zeros.

Data Validity and Outlier Analysis

There were some outliers within the dataset that could have invalidated the data. Firstly, considering majority of the metrics had to be non-negative (excluding the Incremental Sales cluster), it was of utmost importance to ensure that all values were 0 or above. Therefore, the descriptive statistics for each variable was inspected (see Table A.1). Multiple columns were found to have negative values, yet the number of impacted rows was marginal. For example, Sales Value Prev and Sales Value had 4 rows with negative values, thus, they were treated as incorrect inputs and converted to zeros. Moreover, some columns that related to the Sales Value and Baseline Sales Value generated from promotions had negative values, such as Sales Value Disp or Feat, Sales Value Total Disp and Sales Value TPR Only, Baseline Sales Value Total Disp and Baseline Sales Value Disp or Feat. This, unlike incremental sales value columns, which

are compared to the baseline and can be negative, should be non-negative, as if no sales were generated from promotion it should be zero. Thus, those negative values were imputed with zeros. Moreover, there were another four rows within Price (KGS) that had negative values, and it was all within products that had very low sales. It was inspected whether the negative sign was accidental, and whether other non-negative prices within these products were similar, however, this was not the case. Thus, this could have potentially been a data entry error, and therefore those rows were dropped from the data. Lastly, within the distribution columns, Sales Value/ACV TDP Prev had some negative values, however, they were the same rows as the Price negative values. This ensured that all sales columns had non-negative values.

Delving deeper into the outlier analysis, the AutoViz package was utilized to automatically visualize the data. In terms of the Sales Value columns, the data was pretty consistent, and all the large values that occurred were within the right context, i.e, they were either during the Christmas Period, where sales spike in general and they belonged to the largest selling products (see Figure A.9). On the other hand, the price columns had some outliers. Firstly, the Price (KGS) Prev and Avg Base Price (KGS) Prev had one outlier, with a price of 3000 euros, and that ITEM was removed from the data. Moreover, two columns stood out: Incremental Price (KGS) and Incremental Price Any Promo Prev (KGS), which both had huge potential outliers. As seen from the descriptive statistics, their 75% quantile was just 3.08, however, it was found that more than 2500 rows had values larger than 20 or -20, which indicated an infeasible significant price change (see Figure A.10). Therefore, these outliers were imputed with the median of the particular Retailer-ITEM combination in order to ensure consistency. Besides the price columns, no other column possessed an unconventional distribution that could be impacting the robustness of the analysis. Therefore, in the table below the main summary statistics of the primary predictors from the cleaned data are presented.

Table 4.1.
Descriptive statistics of the (clean) data sample

Subcategory	# Items	Mean Unit Weight (KGS)	Mean Share of Sales Value	Mean Sales Value	Mean Sales Value on Promo	Mean Price per Unit	Mean TDP
Ketchup	57	0.58	0.04	33481.34	9781.35	2.29	62.94
Mayonnaise	115	0.40	0.02	19030.99	6181.50	2.08	66.99
MFS	288	0.29	0.01	8848.72	2278.64	1.81	62.34
Mustard	41	0.19	0.05	10009.01	1637.01	1.22	64.67
Other Dressings	41	0.26	0.05	11546.71	3434.74	1.74	65.52
Salad Dressings	58	0.31	0.03	14782.39	3065.54	1.90	110.33

4.1.2 Exploratory Analysis

Once the data was cleaned, an exploratory data analysis followed. There are two key nuances regarding the data, which will be described in this section: a) subcategory characteristics, b) temporal trends.

Subcategory Characteristics

Firstly, a table to summarize the key metrics as well as the dependent variable is provided across all categories (see Table 4.1). Ketchup category stands out with the highest mean sales value, and a significant amount of it being generated being on promotion. The category also has the largest average price per unit, as well as the largest pack size. Considering the small number of items and the relatively high average share, this subcategory can be seen as relatively concentrated. Mayonnaise is similar across all metrics, except for the number of items, which are double of Ketchup. This can explain the relatively small mean share of sales value, as the market contains more items, and it can be seen in the density plot (see Figure A.2). Meat & Fish Sauces have the largest number of items, indicating a wide variety of products, however, this also leads to a large number of products with a very small share of sales (see Figure A.2). Mustard and Other Dressings have the same low number of items and a very similar share distribution, which indicates they are concentrated markets. However, they differ on key metrics, such as mean price, where Mustard's average price per unit is far lower, whilst within Other Dressing's price per unit is higher, and far more sales are generated from promotions. Lastly, Salad Dressings stand out for their high total distribution points. This could indicate that the products within this category appear in very successful stores with a wide availability. Overall, this table as well as the density plots are key to understanding the diverse subcategory characteristics, which can later aid the result interpretation.

Temporal Dynamics

To further enhance the understanding of data, the time-series sales data of the separate categories was inspected. Firstly, the pattern for both Ketchup and Mayonnaise is pretty similar, and the time-series exhibits large holiday effects (see Figure 4.1 & 4.2). There is a tremendous spike during Christmas, a smaller increase during Easter, and a huge decrease during the first week of January. For Mayonnaise, there seems to be a larger increase during the summer period, which coincides with the Barbecue season, where majority of items are promoted. This entails that accounting for holiday and summer effects will be crucial. Similar effects also applies to Meat & Fish Sauces, Mustard, and Other Dressings categories, which indicates a spike during Christmas, yet the sales seem more consistent and less volatile during the off-season (see Figure 4.3 & 4.4 & 4.5). Finally, for Salad Dressings, the sales are far higher during the summer time rather than winter (see Figure 4.6). This indicates that weather could be a crucial determinant for this category.

4.1.3 Creation of the Feature Space

Feature Description

Building on the factors discussed in the theoretical framework and an exploratory data analysis, a comprehensive feature space was constructed. The individual variables can be summarized into three groups: marketing mix, seasonality and temporal dynamics, cross-product effects (see Table A.3 & A.4 & A.5).

Figure 4.1.
Ketchup

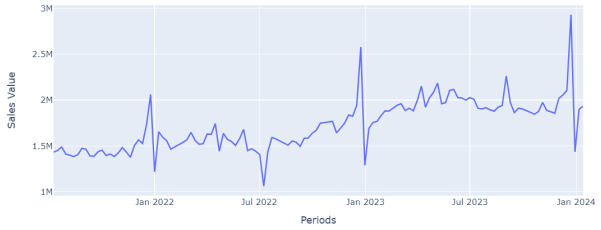


Figure 4.2.
Mayonnaise

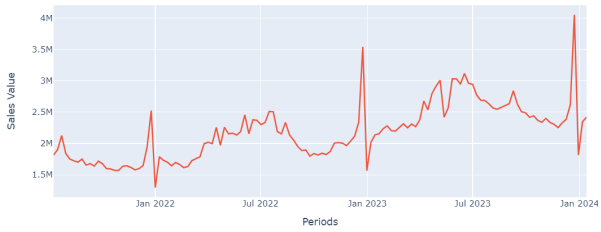


Figure 4.3.
Meat & Fish Sauces

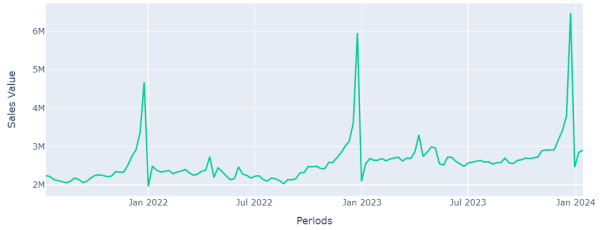


Figure 4.4.
Mustard

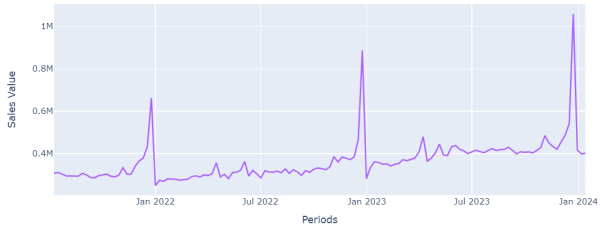


Figure 4.5.
Other Dressings

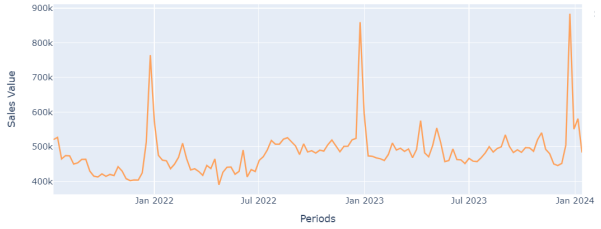
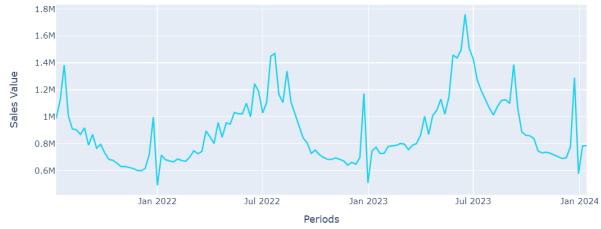


Figure 4.6.
Salad Dressings



Firstly, the marketing mix group contains all variables regarding the product descriptive variables, price metrics, place(distribution metrics) and promotional metrics, and majority of them were derived from Nielsen itself. Key product descriptive variables such as 'Segment', 'Brand', 'Type', 'Pack Type', 'Variant' provide detailed insights into the product's characteristics, and will be utilized as dummies. In terms of numerical columns, pricing metrics, such as 'Price per Sales (Unit)', and 'Incremental Price per Sales (Unit) Any Promo', will capture the price and price changes during promotion for each respective SKU. Moreover, considering that Price is not weighted, the weighted version of price was computed (weighted by sales value). Distribution metrics like 'Total Weighted Distribution Points (TDP)' and 'Number of Items' indicate the availability and market reach of the products. Promotional metrics, such as 'Sales Value Any Promo' and 'Sales Value Feat W/o Display', track sales generated through various promotional activities, offering a clear view of the impact of promotions on sales performance.

Secondly, the seasonality and temporal dynamics features aim to capture the time-related factors on share of sales. From the data exploration, it is evident that the sales within certain subcategories are driven by seasonality. For example, the Sales Value Trend for Ketchup and Mayo is quite seasonal. There is an obvious spike during the Christmas, a small spike around

Easter and a some bumps during the summer time (see Appendix A, Figure 4.1 & Figure 4.2). To account for the holiday effects, the UK holidays dates for 2021-2024 were extracted from the Python's 'holidays' package and utilized as dummies. Moreover, from the trend data exploration, it appeared that Salad Dressings became far more popular during summer time rather than winter, as their sales spikes during this period. To account for this effect, weather data from the London's Weather Centre was extracted from Meteostat, which included the weekly average temperature and precipitation (Meteostat, 2024). The researcher could not obtain the average across the whole of UK, however, it is assumed that the primary purchasing power is within London, thus the weather data from there would be the best alternative. Lastly, to account for the temporal dynamics, several variables were computed. The calendar variables, such as Week, Month, Quarter and Year were extracted from the date column, used as dummies. Moreover, a consistency feature, which indicated whether the data was consistent (appeared in the whole dataset) was computed in order to rationalize which SKUs are seasonal and not seasonal. Lastly, the feature 'Data Age' was introduced, showcasing the number of days since the introduction of the product, effectively indicating the product life-cycle.

To elucidate within-category cross-product effects, there were five specific variables types addressed. Due to computational constraints, however, these effects were only calculated for the top 5 items by sales value within each of the 6 segments. This is similar to the approach of Ma et al. (2016), who considered the top 5 items. Specifically, I considered variables such as 'Total Weighted Distribution Points (TDP)', 'Weighted Price per Sales (Unit)', and the Sales Value generated from each existing promotional type. Furthermore, the number of products on promotion within the same category was computed to capture the promotional intensity. Moreover, the average brand price was computed in order to capture brand-specific effects. Lastly, based on the text-embeddings, a product pairwise cosine similarity matrix was computed. From this matrix, it was possible to extract the average similarity to 10 most similar items for each item, and calculate the what is the average share of sales across these 10 items.

Lags

What is crucial to keep in mind when creating feature space is the forecasting horizon of 12 weeks. This is a direct multi-output ahead forecast, as it produces a prediction for each week over the next 12 weeks in the test set. Therefore, it is imperative to not introduce data leakage into the model when creating the lags of the primary metrics, and only use features that would be known at the time of prediction.

Creating lags in forecasting is a standard practice that can capture temporal dependencies. For each numerical variable (except weather), 7 lags were created, from the previous 12 to 18 weeks. Research claims that introducing lags can help the model better understand trends and recurring patterns, and picking effective lags is crucial to do so (Surakhi et al., 2021). However, introducing lags also comes with a limitation - they introduce NaN values at the beginning of the dataset. Nonetheless, to not introduce potential bias into the model and make assumptions of past values, the NaNs were removed, leading to a loss of 18 weeks of data.

4.2 Models and Model Evaluation

This section will overview the chosen models, hyperparameter tuning and cross-validation approaches, as well as the utilized evaluation metrics.

4.2.1 Statistical time-series forecasting

The Prophet forecasting method has been chosen as the statistical forecasting method. Prophet is a univariate method that has been developed by Facebook, which is designed to handle time series data with strong seasonality and multiple historical seasons (Hyndman & Athanasopoulos, 2018; Taylor & Letham, 2018). Similarly to such models as Holt-Winters method, it decomposes time series data into trend, seasonality, as well as holiday components, which makes it a flexible method to be applied to a variety of forecasting tasks. The modelling process of Prophet can be expressed in the following way:

$$y_t = g(t) + s(t) + h(t) + \epsilon_t,$$

where $g(t)$ describes a piecewise-linear trend (or "growth term"), $s(t)$ describes the various seasonal patterns, $h(t)$ captures the holiday effects, and ϵ_t is a white noise error term (Hyndman & Athanasopoulos, 2018; Taylor & Letham, 2018).

Prophet is unique in its ability to automatically capture different types of seasonality, which includes daily, weekly and yearly without requiring extensive tuning. Seasonality $s(t)$ is modeled using Fourier series, allowing the model to capture periodic patterns in the data. The trend is modelled either using the linear model or a logistic growth model, which can be expressed in the following way:

$$T(t) = \begin{cases} k + mt & \text{(linear growth)} \\ \frac{C}{1+\exp(-k(t-t_0))} & \text{(logistic growth)} \end{cases}$$

where k is the growth rate, m is the trend change rate, and C and t_0 are parameters in the logistic model, representing the growth. Implementing logistic growth allows to specify the expected capacities at any point at time, as well as the expected floor value, which can prove to be useful in the case of forecasting share due to its inherent bounds. On the other hand, since the forecast is univariate, which per SKU, the logistic growth may not necessarily be represented. Nevertheless, both linear and logistic growth rates will be tested. Lastly, Prophet incorporates the effect of holidays $h(t)$, which can be specified per country and dummy variables are created in order to capture their effect. This is particularly useful for our data, since the data exploration revealed significant demand shocks during Christmas and Easter, and therefore UK holidays are added to the model.

Ultimately, the primary advantage of Prophet is its ability to handle missing data and deal with outliers. Considering that SKUs are often in and out of the market, which creates inconsistency within their time series, this is an essential feature that allows to handle irregularities

without extensive preprocessing. Therefore, Prophet is a strong univariate statistical model that will be used to forecast SKU sales value share and benchmark against ML models.

4.2.2 Machine learning methods

Random Forest

Random Forest is a popular ensemble ML model that was first introduced by Breiman (2001) and now has been applied within a variety of data problems. It is a method that is developed based on building multiple decision trees and merging their results to improve overall prediction accuracy and reduce overfitting.

RF works by generating a number of decision trees, and each tree is trained using a randomly chosen subset of training data using a different bootstrap sampling, which selects random number of features. This is a key aspect of RF, which allows to decorrelate trees, by selecting a random sample of m predictors from the full set of p predictors at each split, where m is chosen to be approximately $\sqrt{p}$. This introduces a randomness aspect, which allows each tree to be unique, results in a diverse set of individual models that can generalize better than one model, improving the model's robustness and accuracy (James et al., 2023, pp. 346-347).

Despite this, RF can still overfit, which entails that it learns the training data too well, potentially impacting its performance on the out of sample hold out test data. This can occur when the the trees are very deep (Segal, 2004). Therefore, multiple hyperparameters have been developed that can limit overfitting with RF. For example, *max_depth* allows to specify the maximum depth of the tree, preventing it to learn too many specific details about the data. Moreover, limiting the number of features that are considered within each split with *max_features* can be useful, restricting the model to not become specialized in particular features. Thus, these techniques can be employed to control the fitting process of RF on the training set.

What makes it particularly applicable to forecast SKU-sales is its ability to capture non-linear patterns and handle high-dimensional data (Breiman, 2001). Considering that share of sales can be effected in complex ways with regards to competitive effects, RF can be a versatile and a powerful method for this task. Therefore, it is chosen as a model to predict share of sales within a retailer.

Gradient Boosting Regressor

Similarly to Random Forest, Gradient Boosting Regressor is another ensemble ML model that is based on weak learners, which are typically also decision trees. This method works in an iterative manner, by adding models to correct the errors that are made by the previous models.

Specifically referencing decision trees, Gradient Boosting Regressor trains a sequence of these, where each new tree is trained to predict the errors of the previous collection of trees, and the predictions are added to the existing model to improve its robustness (James et al., 2023, 337-348). This is unlike random forest, which builds multiple independent trees on bootstrapped data samples. This particular process is repeated for a specific number of iterations, with each tree improving the model's performance in areas where it previously performed poorly. Thus,

instead of fitting the total data, the model learns slowly by gradually fitting trees. This can prevent overfitting as it has an opportunity to flexibly adapt to changes.

GBR also has similar hyperparameters as Random Forest that can be tuned during the cross-validation process, such as *max_depth* and *n_estimators*. A unique hyperparameter that can be tuned for GBR is *learning_rate*, which can control the speed of the learning process (James et al., 2023, p.349). It works by shrinking the contribution of each new tree, and multiplying the residuals in previous tree, which is further used in modelling. A lower *learning_rate* can be more computationally expensive, however, may capture the patterns within the data better, whilst a higher learning_rate may pass the optimal solution as it speeds up the learning process. Ultimately, GBR serves as an efficient model that can prevent overfitting, yet due to described processes can be very expensive to tune.

Due to the computational and memory constraints, a variant of GBR was used - Histogram-based GBR. This model is optimized to speed up computation by binning continuous features into discrete intervals, which diminishes the number of potential split points. Whilst this is a trade-off, since the continuous scale can reveal more granular detail, it is proven to be efficient and still predict with similar accuracy (Guryanov, 2019).

4.2.3 Hyperparameter selection and tuning

Tuning hyperparameters within ensemble machine learning methods directly impacts their performance and ability to generalize from the training data. There are three hyperparameters considered, two of which both RF and GBR share, which are number of estimators (*n_estimators*) and maximum depth of trees (*max_depth*). The chosen values of those hyperparameters are describe in the table below (see Table 4.2).

Firstly, the selected range of *n_estimators* allows to balance between model accuracy and computational complexity. The model's performance can improve model performance by reducing variance, as it allows more decision trees to fit. However, it increases computational time and memory usage significantly. Therefore, selecting 100 and 200 as the values allows to compare the model's performance within different complexities, without introducing too much computational complexity with a high number of estimators.

Secondly, the *max_depth* parameter is chosen due to its innate ability to prevent overfitting. The range from 5 and 11 (inclusive) with a step of 2 is selected to capture different model complexities. Shallower trees (e.g., *max_depth* of 5) can aid to prevent overfitting by restricting the model's ability to capture noise within the training data. On the other hand, it can oversimplify the model, and deeper trees can capture more complex patterns. However, the deeper trees introduce a greater risk of overfitting, therefore, choosing the right value for *max_depth* is crucial to manage the trade-off of bias and variance.

Lastly, within GBR, a key hyperparameter is *learning_rate*, which controls the decision trees contribution to the final model. The *learning_rate* values (0.01, 0.05, 0.1) allows to inspect how model perform with lower and higher values. Lower values make the learning process slower

but could potentially lead to better generalization by preventing the model from overreacting to noise in the data. A higher *learning_rate* speeds up learning but can risk overshooting the optimal solution, making it crucial to strike a balance.

Whilst the hyperparameter range could be broader, the choices were constrained by computational limits. Therefore, these hyperparameter values were chosen carefully to observe the trade-off between the parameters that bring complexity to the model, as compared to those that make the model simpler.

Table 4.2.
Hyperparameters Tuned for Gradient Boosting Regressor and Random Forest

Model	Hyperparameter	Values
Gradient Boosting Regressor	n_estimators	100, 200
	max_depth	5, 7, 9, 11
	learning_rate	0.01, 0.05, 0.1
Random Forest	n_estimators	100, 200
	max_depth	5, 7, 9, 11

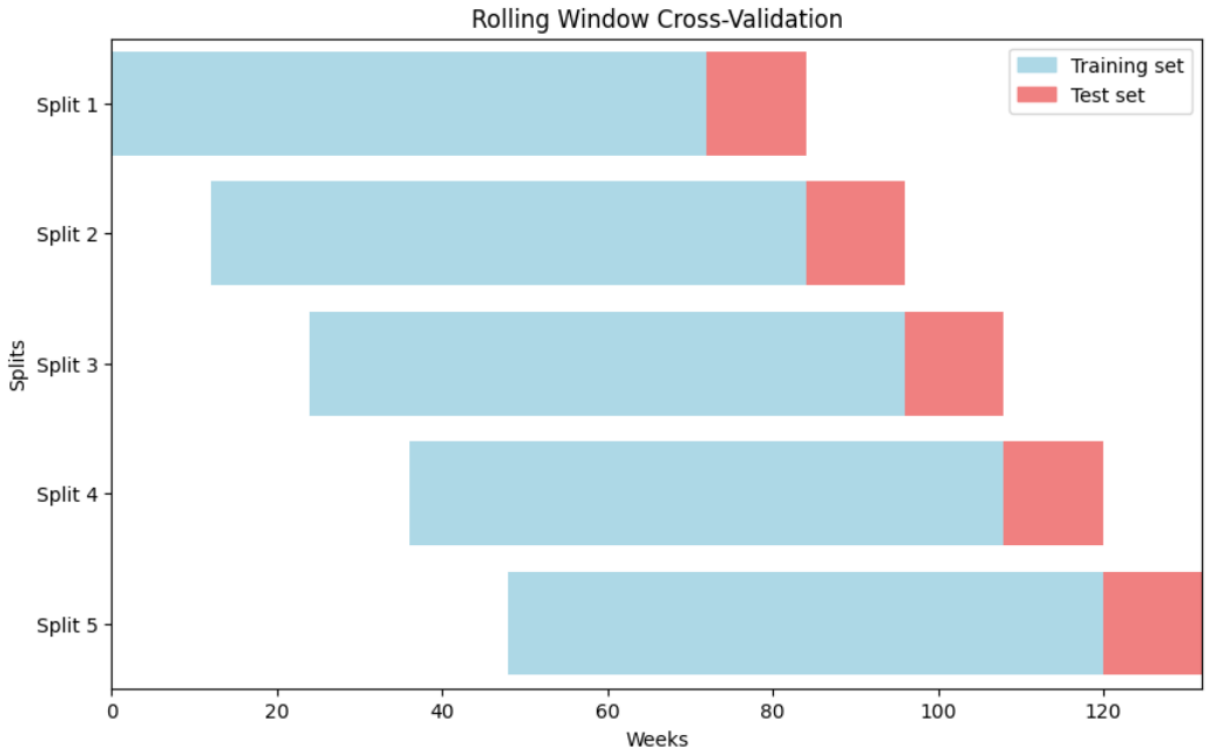
Train-test split and Cross-Validation

After removing the missing data that was due to the introduction of lags, the dataset contains 144 weeks of data. Since the primary objective is to forecast 12 weeks into the future, the out-of-sample hold out test set will be 12 weeks. Therefore, this leaves 132 weeks for the training set.

Within this training set, cross-validation will be applied to tune the hyperparameters and choose the best model. Cross validation within time-series data has to respect the nature of the data (Bergmeir & Benítez, 2012). Regular CV techniques, such as K-Fold or LOOCV, cannot be applied to time series data, as they assume the data points are independent and identically distributed. Since this is not the case with time series data, one needs to adopt a time-series specific cross validation technique. To tune the aforementioned hyperparameters and validate model's performance, the rolling window approach is adopted, which works in the following way. At each step, the model is retrained on the most up-to-date training set and then tested on the next block to assess performance across these specific forecasting periods. In our case, this looks like having the training set of week 1-72 and testing it on week 73-84, and within the next iteration the test set integrates into the training set, yet the origin moves forward, which means that the training set becomes week 12-84 and test set becomes 85-96. This is summarized in Figure 4.7.

This approach is appropriate for two reasons. Firstly, Nielsen provides Unilever a fixed 164 weeks of data. This means that the cross-validation mimics the real-world scenario, as there would be a fixed training set, rather than an increasing one. Secondly, such an approach is more robust and potentially could prevent overfitting, as it discards previous data, which may not be

Figure 4.7.
Rolling Window Time Series Cross Validation



useful anymore due to changing trends, keeping training data stable. Therefore, this approach is well-suited for our time-series problem, ensuring robust and realistic model validation.

4.2.4 Evaluation Metrics

The models will be rigorously evaluated using three metrics: mean absolute error (MAE), symmetric Mean Average Percentage Error(sMAPE) and root mean squared error (RMSE).

RMSE will be the primary evaluation metric due to its sensitivity to large errors. It emphasizes the importance of minimizing significant forecast inaccuracies, which is crucial for avoiding major strategic and operational consequences, thereby aligning model evaluation with the goal of ensuring robust and reliable market share forecasts (Mehdiyev et al., 2016). Moreover, MAE provides a straightforward measure of the average error magnitude, making it particularly useful for assessing the baseline accuracy of SKU market share predictions and ensuring that forecasts are reliable on an operational level (Mehdiyev et al., 2016). sMAPE, by expressing errors as a percentage of actual values, offers a clear view of the relative forecast accuracy across different SKUs (Kim & Kim, 2016). sMAPE is also used since it avoids the issue of division by zero, as some products had weeks where they did not sell. This makes it superior to MAPE and other variations, as calculating MAPE with zero actuals provides undefined values.

Together, these metrics provide a comprehensive framework for evaluating model performance, directly addressing the research objectives and the potential forecasting errors.

4.3 High Dimensionality Reduction Techniques

Considering the large feature space that arose due to the introduction of lags, one-hot encoding categorical variables and cross-product effects, it introduces severe high-dimensionality. High-dimensional data can present challenges for modelling, and thus, this section will overview the multiple measures that have been implemented to control high-dimensionality and reduce multi-collinearity within the data.

4.3.1 Multicollinearity Inspection

Firstly, Pearson's correlation coefficients were computed for all numerical predictor variables (see Figure A.18 & A.19 & A.20 & A.21) . The primary aim was to remove variables which coefficients were above 0.9, as that indicated high level of multicollinearity. This was done to ensure that the feature space contained only distinct variables with different meanings. For example, variables such as 'Incremental Sales Value Any Promo' had a very high correlation with 'Sales Value Any Promo' (0.93), and thus were removed from the data. In total, 19 unique variables and their lags were removed, which minimized the feature space by 195 variables. This step ensured that all variables remaining in the dataset were interpretable and contained independent information from other variables.

4.3.2 Text embeddings

Secondly, another source of high-dimensionality and sparsity is introduced by one-hot encoding categorical variables, such as brand, type, pack type and variant, which represent product attributes. To address this issue, text embeddings were introduced to capture the latent information that exists within the categorical columns. Essentially, embeddings are vectors of floating point numbers that transform the data into a lower-dimensional dense vector space (James et al., 2023, pp. 419-420). They can be described as a function $f : V \rightarrow \mathbb{R}^d$, where V is the set of all possible data points and d is the dimensionality of the embedding space. Embeddings primarily capture the semantic similarity, as the distance between two vectors capture, in our case, item description relatedness. Small distances between vectors can suggest high relatedness between items, whilst larger distances entails that products are not similar (*OpenAI Platform*, n.d.).

These text embeddings were introduced as features by utilizing OpenAI's *text-embedding-3-large* model. This model is the most powerful OpenAIs embedding model, generating 3072-dimensional embeddings, which allows to capture deep underlying semantic patterns. However, due to memory constraints, the dimensionality was reduced to 256 dimensions. According to OpenAI, the reduced dimensional space still outperforms their third best model, *text-embedding-ada-002*, which has 1536 dimensions (*OpenAI Platform*, n.d.). Therefore, this dimensionality reduction of embeddings is justified, as it allows for efficient information capture whilst also minimizing computational complexity and potential overfitting. Ultimately, embeddings offer scalability since they are able to capture each data point's essential features in a compact, continuous vector space, facilitating efficient processing and retrieval in large-scale applications.

4.3.3 Targeting Predictors with LASSO

A key aim of this research is to examine Targeted Random Forest (TRF) as a method that could enhance predictive accuracy. To do this, it is key to pick a targeting, otherwise known as feature selection, technique that can be suitable. This study follows the methodology of Borup et al. (2023), who implemented the TRF originally.

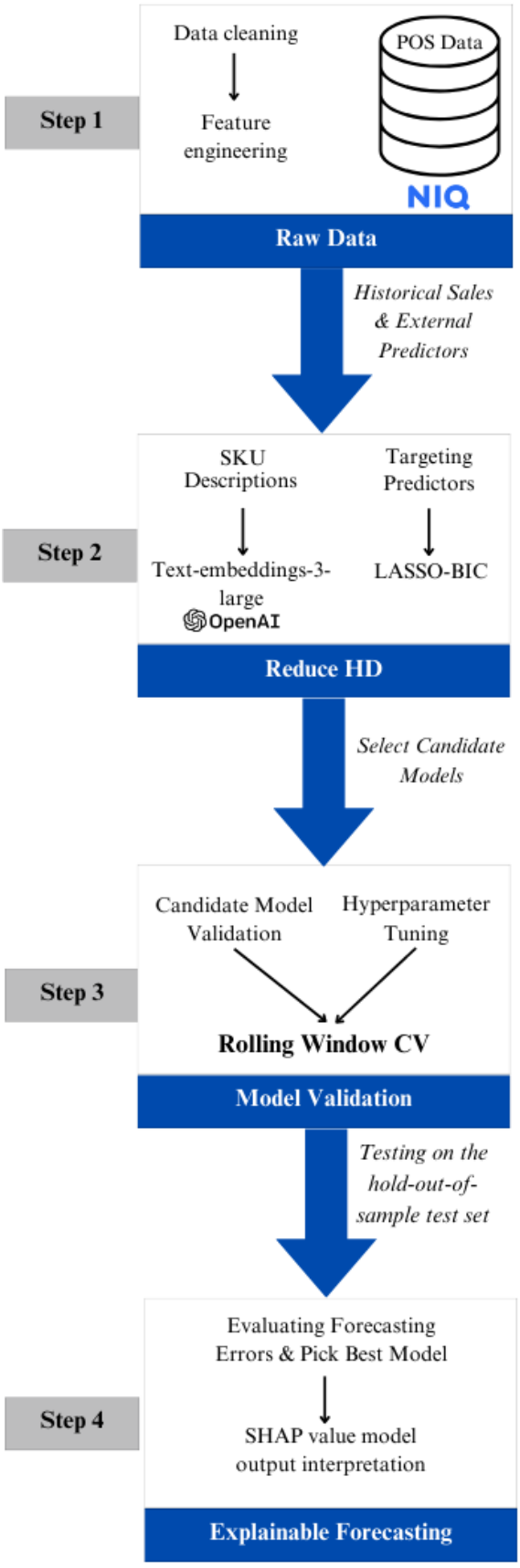
The study suggest to employ LASSO regression with Bayesian Information Criterion (BIC) for feature selection. LASSO is a regularization method that penalizes the regression efficients by shrinking them to zero if they are statistically insignificant, which effectively performs feature selection (James et al., 2023, pp. 246-247). This results in a model that includes only the most relevant predictors. The regularization penalty is controlled by parameter λ, which becomes stronger as the number increases, meaning that the feature space becomes more suppressed as penalty increases. To select the correct penalty, BIC criterion is utilized, as it allows to select the best λ based on the model fit and complexity. BIC is based on the likelihood function and introduces a penalty term for the number of parameters in the model to avoid overfitting. The BIC is defined as follows:

$$\text{BIC} = -2\log(\hat{L}) + (p+1)\log(n) \tag{4.1}$$

where $\hat{L}$ is the maximized value of the likelihood function of the model, p is the number of parameters in the model, n is the number of observations (Clyde et al., 2022). BIC will be used to compare the different models of LASSO that have been selected with different λ values. This is implemented with the class of *LassoLarsIC* from *scikit-learn*. This class allows to automatically go through multiple λ values and using BIC evaluate their performance. Therefore, using LASSO with BIC is a data-driven approach that has proven to be effective in Borup et al. (2023) study, as the selected degree of targeting with this method yielded the best results for the majority of the cases.

These approaches help to deal with our feature space, that can be defined as highly dimensional and sparse, as more than 1500 features are present in the model. This could potentially result in enhanced predictive power of the examined models. Consequently, this research will address to what extent do these high-dimensionality reduction techniques contribute to the model performance.

Figure 4.8.
Proposed Modelling Framework

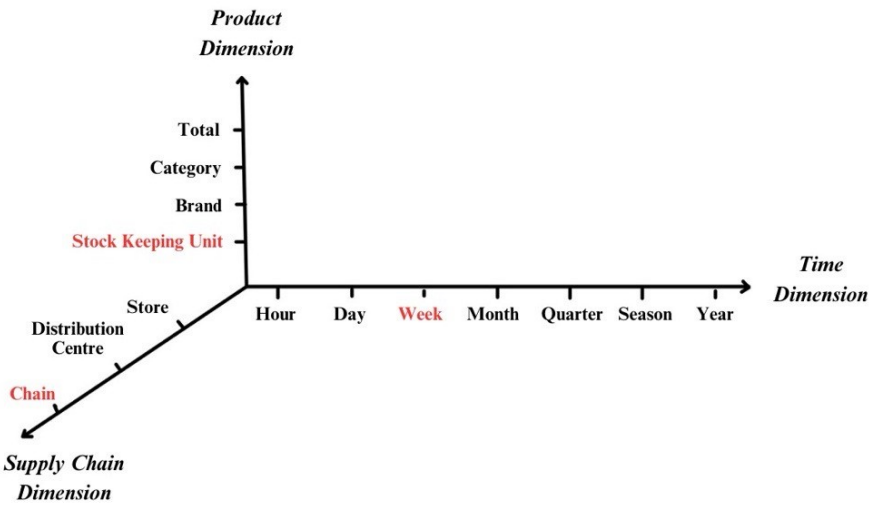


Adapted from S. Chen et al. (2024)

5. Results

The results section will provide three level insights. Firstly, section 5.1 will compare the statistical forecasting method and machine learning methods, evaluating their performance with RMSE, MAE and sMAPE. Secondly, the second section will explore how dimensionality reduction techniques impact the accuracy and interpretability trade-off. Thirdly, the last section will focus on interpreting the results of the best model using SHAP values and provide a category-overview. It is important to note that all models were trained and tested at the same aggregation level, predicting the weekly SKU-share of sales at the retailer-chain level for the next 12 weeks (see Figure 5.1). The results will serve as a presentation of the findings, which will be contextualized in Chapter 6.

Figure 5.1.
Forecasting Dimensional Hierarchy



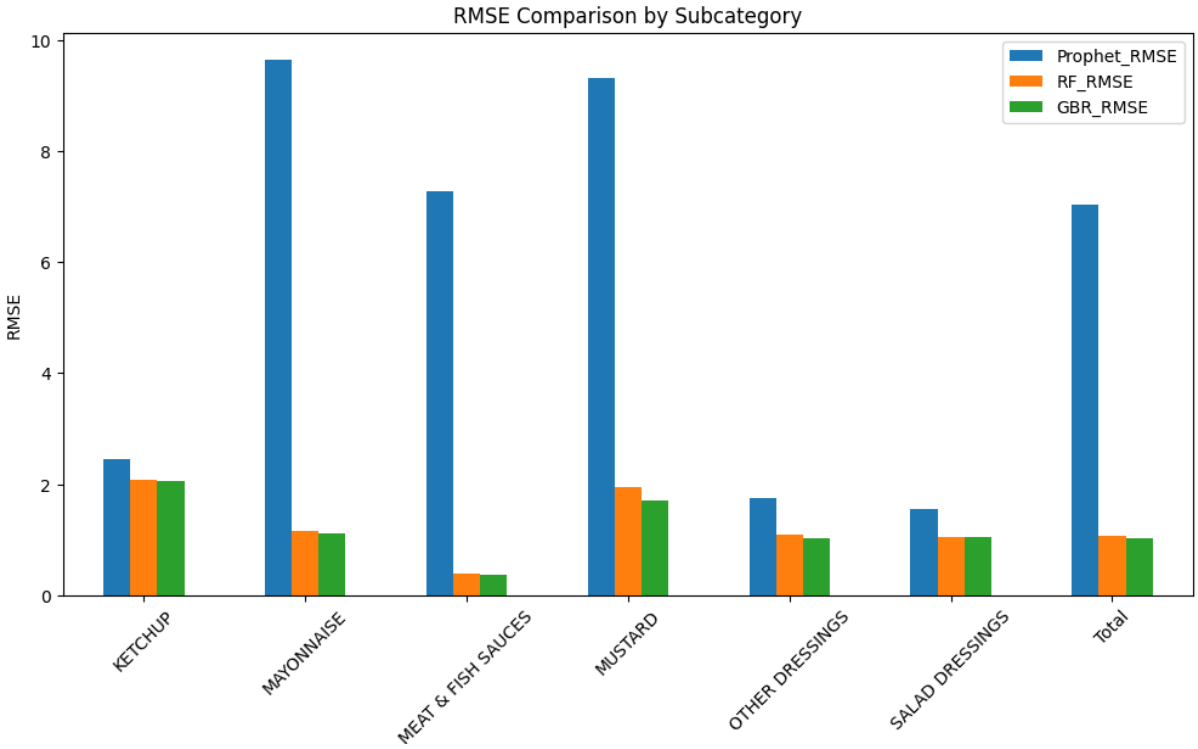
Adapted from Fildes et al. (2019) and Syntetos et al. (2016)

5.1 Comparison of Statistical and Machine Learning methods

To answer the question ”How do ensemble Machine Learning models compare to statistical time-series forecasting methods in terms of accuracy for predicting weekly SKU level sales value share?” three separate models were evaluated on the hold-out-of sample test set. One of these models were statistical (Prophet), whilst the other were ensemble machine learning techniques (Random Forest and Gradient Boosting Regressor). All of these models were ran without any regularization or feature selection applied.

Amongst the three models, Gradient Boosting Regression emerged as the superior model when evaluating overall metrics (see Figure 5.2). It achieves the lowest overall RMSE, with a value of 1.03, compared to Random Forest’s RMSE of 1.08, and Prophet’s RMSE of 7.03.

Figure 5.2.
Model Performance on the Hold-out-of-Sample Test Set



*NOTE: RMSE is scaled to 0-100% for better interpretation purposes.

Whilst its' sMAPE (45.51%) is slightly worse than Random Forest's sMAPE (44.60%), Gradient Boosting showcases strong consistency across all error metrics (see Table B.4). A key difference between the predictions is that both Random Forest and Gradient Boosting have more stable predictions (see Figure B.2 & B.3). Prophet, on the other hand, has several really large errors, which deeply impacts its RMSE (see Figure B.1). Nevertheless, it can be observed that its sMAPE (44.69%), which is less sensitive to outliers, is comparable to both Random Forest and Gradient Boosting. However, a critical insight about Prophet is observed - it cannot forecast SKUs that were not part of the training set, unlike RF or GBR, which means that for 241 observations it predicts NaN values. This is important, as it showcases the scalability of Random Forest and GBR, which enable predictions for products that had no past data.

When delving within the categories, it can be seen that Gradient Boosting consistently shows lower RMSE values. For example, in the Mayonnaise subcategory, Gradient Boosting achieves an RMSE of 1.11, outperforming Random Forest's RMSE of 1.16, and Prophet's RMSE of 9.65 (see Figure 5.2). On the other hand, the sMAPE is lower for almost all categories within Random Forest, which could entail that Random Forest can better predict values closer to zero (see Table B.4). This is due to the fact that sMAPE is known to inflate when there are significant number of actual values that are zero or close to zero within the dataset. Prophet, despite being a far simpler model, does not lag behind much across all categories in terms of sMAPE, yet it has large errors in terms of RMSE. The highly significant difference is within Mayonnaise,

Mustard's and Meat & Fish Sauces, where Prophet's RMSE is 9.65, 9.32, and 7.27 respectively, which are significantly higher than the ensemble machine learning models (see Table B.4). This is due to the fact that it predicts a share of 1 in some weeks for certain products, which destabilizes its result.

What is common between the ensemble ML models is that the worst predictions in terms of RMSE occur within the Ketchup subcategory and the best predictions pertain to the Meat & Fish Sauces subcategory. Referring to the descriptive analysis in the Methods section, this could potentially explained by the nature of these subcategories. For example, Ketchup has a lot of inconsistent SKUs that generate over 35% of the categories total sales, potentially introducing high-volatility within the data that makes it difficult to learn. Moreover, it has larger average share sales value and a comparatively small number of items (57), making it a concentrated segment. In terms of Meat & Fish Sauces, it is a subcategory that has the largest number of total SKUs, with 288. This abundance of SKUs can help model performance in two ways. Firstly, by having more examples of the SKUs within the subcategory, the model can better learn the signal within the data. Secondly, the share of sales value due to the large number of SKUs is low, and thus, less volatile.

Table 5.1.
Post-Hoc Comparisons using Tukey HSD

Comparison	Mean Difference	Std. Error	p-adj	Lower CI	Upper CI
GBR vs. Prophet	5.94	3.03	0.000***	5.91	5.98
GBR vs. RF	0.05	0.03	0.002**	0.02	0.08
Prophet vs. RF	-5.89	3.01	0.000***	-5.93	-5.86

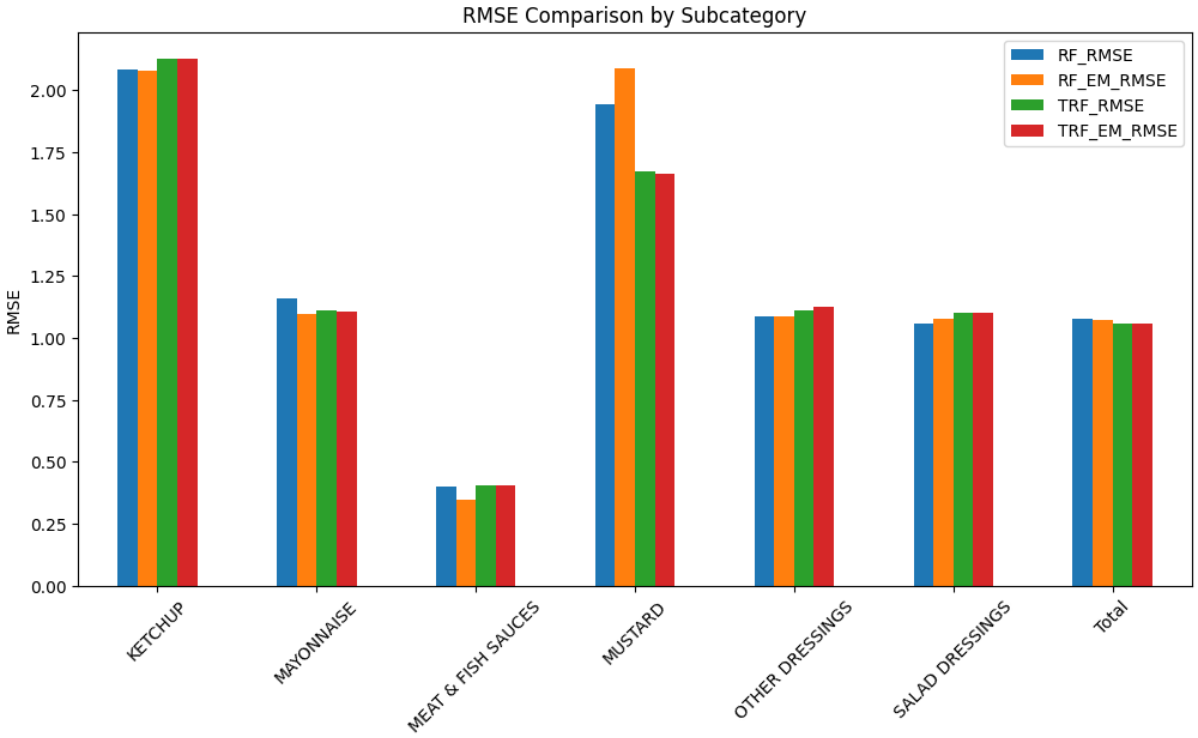
*NOTE: RMSE is scaled to 0-100% for better interpretation purposes. The significance level is ($p < .05$)

To understand whether there are significant differences between model performance in terms of RMSE, a non-parametric bootstrap approach was employed. Using 1000 bootstrap samples for each model, the test set was resampled with replacement and the RMSE was computed for each sample. This approach produced an empirical distribution that can help with statistical inference. Considering that there are three models, a One Way Analysis of Variance (ANOVA) was conducted to inspect for statistical significance. To evaluate the difference between individual models, post-hoc comparisons using the Tukey Honest Significant Difference (HSD) test were ran (see Table 5.1). ANOVA results disclose that there is a significant difference between the three model performances ($F(2, 1998) = 114097.78, p < .001$). GBR significantly outperformed both Prophet ($p < .001$) and RF ($p < .05$). RF model also had a significant difference compared to the Prophet RMSE performance, with a mean difference of -5.89 ($p < .001$). Overall, these results showcase that the GBR model is significantly different from both RF and Prophet, indicating superior performance (see Table B.5).

5.2 Text Embeddings and Targeted Random Forest

To answer the second question ”How does applying different high-dimensionality techniques, such as a) transforming SKU descriptions into embeddings, b) targeting predictors with LASSO, impact ensemble model performance and interpretability?”, several aspects will be covered. Firstly, predictive performance in terms of RMSE will be compared. Secondly, feature importances will be inspected to compare whether the model captures features differently under the two techniques. Thirdly, computational performance will be depicted to understand each methods computational need. Ultimately, Random Forest will serve as the baseline model, since one of the aims of the study is to examine Targeted Random Forest.

Figure 5.3.
Model Performance on the Hold-out-of-Sample Test Set



*NOTE: RMSE is scaled to 0-100% for better interpretation purposes.

5.2.1 Impact of Embeddings

Embedding the SKU descriptions generally enhances the predictive performance of Random Forest models, as the total RMSE is down from 1.08 to 1.07 (see Table B.6). However, RF with embeddings only achieves better performance in three categories - Ketchup, Mayonnaise and Meat Fish Sauces. For instance, the Mayonnaise subcategory, RF with Embeddings achieves a lower RMSE (1.09) compared to the RF Baseline (1.16), and similar improvement are seen Meat & Fish Sauces segment (see Table B.6). However, in other categories, the RF baseline performs better, such as Mustard (1.94 vs 2.09) and Salad Dressings (1.06 vs 1.08). Thus, within categories, there are some performance trade-offs that exist between applying random

forest without and with embeddings. Nonetheless, the overall performance across the tested time-period improves (see Figure 5.3).

When inspecting the feature importances of the Baseline Random Forest and Random Forest with Embeddings, one aspect stands out (see Figure B.4 & Figure B.5). *Embedding_5* and *Embedding_181* appear within the top 10 most important predictors in the Random Forest, whilst only one one-hot-encoded dummy appears in the Baseline Random Forest. This could indicate that the embeddings capture a latent feature that is not described in the one-hot encoded features, potentially explaining the model's enhanced performance. When inspecting the data, it was found that Embedding_5 did have a commonality - it depicted the claim that the SKU had primarily within the Mayo and Ketchup subcategories (see Figure B.6). For example, items had such claims as 'LOW CALORIE CLAIM' or '50% LESS SUGAR & SALT' or 'ORGANIC CLAIM', which was not part of the product attributes one-hot encoded dummies. When comparing the RMSE on these items between the two models, there was a difference, with the Baseline RF of 1.32 RMSE and RF with Embeddings achieving 1.19 RMSE. This is important, as it potentially explains the observed performance increase seen within Ketchup and Mayonnaise, depicted in Figure 5.3. Whilst Embedding_181 did not have an obvious pattern, majority of the products were related to either VEGAN or CHILLI variant mayonnaise. Thus, this showcases how embeddings can be superior to one-hot encoded features by determining intricate latent patterns within categorical data.

Nonetheless, one has to factor in the additional computational constraint - the baseline RF Grid Search took 285 minutes, whilst RF with Embeddings took 393 minutes to tune (see Table B.7). Considering that this thesis operates within a limited hyperparameter grid, this can be an important implication if the model hyperparameter tuning occurs on a more extensive grid.

5.2.2 Impact of Targeting Predictors with LASSO

Applying LASSO with BIC criterion reduced the feature space from 1662 features to 101 features. This strategy to target predictors prior to applying RF showed some improvements over the baseline. Targeted RF consistently achieves lower RMSE and MAE values across several subcategories (see Table B.6). For example, in the Mustard subcategory, Targeted RF outperforms the baseline with an RMSE of 1.67 compared to the RF Baseline's RMSE of 1.94. Another segment where predictions improve is Mayonnaise, where the RMSE decreases from 1.16 to 1.11. Performance within these categories are the primary sources of improvement, as none of the other categories have better performance over the baseline. Moreover, combining both embeddings and regularization (Targeted RF with Embeddings) does not result in significant changes from LASSO regularized forest. The performance across all categories in terms of RMSE stays very similar with no substantial improvements (see Figure 5.3). Nevertheless, this can be explained due to the features that LASSO selects. It shrinks all the embedding features, and prefers to select one-hot encoded dummies. Therefore, this results in a very similar model to LASSO Random Forest, and it can be observed through its feature importances (see Figure

B.8 & B.9).

What is particularly beneficial about the the TRF method is the reduced computational cost and complexity. The reduction of number of features resulted in a 5 time decrease in time to tune the model, with only 57 minutes runtime (see Table B.7). Considering that the LASSO-BIC tuning time was negligible, this is an imperative method to consider if reducing computational costs is an important priority.

Table 5.2.
Post-Hoc Comparisons using Tukey HSD

Comparison	Mean Difference	Std. Error	p-adj	Lower CI	Upper CI
RF_Embeddings vs. RF	0.0060	0.0031	0.004	0.0014	0.0106
RF_Embeddings vs. RF	-0.0129	0.0066	0.000	-0.0175	-0.0083
RF_Embeddings vs. TRF	-0.0120	0.0061	0.000	-0.0167	-0.0074
RF vs. TRF_Embeddings	-0.0189	0.0096	0.000	-0.0235	-0.0143
RF vs. TRF	-0.0180	0.0092	0.000	-0.0227	-0.0134
TRF_Embeddings vs. TRF	0.0009	0.0005	0.964	-0.0038	0.0055

*NOTE: RMSE is scaled to 0-100% for better interpretation purposes. The significance level is ($p < .05$)

The same non-parametric bootstrap sampling approach was used to find whether the models differed in terms of statistical significant. ANOVA results showcase that there is a significant difference between the four model performances ($F(3, 2997) = 654.59, p < .001$) (see Table B.8). All models significantly differ in their performance in terms of RMSE, besides LASSO regularized Random Forest and LASSO regularized Random Forest with Embeddings ($p < .001$). Overall, Targeted RF emerges as the best model in terms of predictive performance due to its consistent lower metrics across various categories, achieving the lowest total RMSE (1.06) and MAE (0.45), and its result is significantly different from all other approaches (see Table B.6). This demonstrates that regularization performed prior to applying Random Forest can enhance model performance by reducing overfitting and improving prediction accuracy.

5.3 Model Output Interpretation with SHAP Values

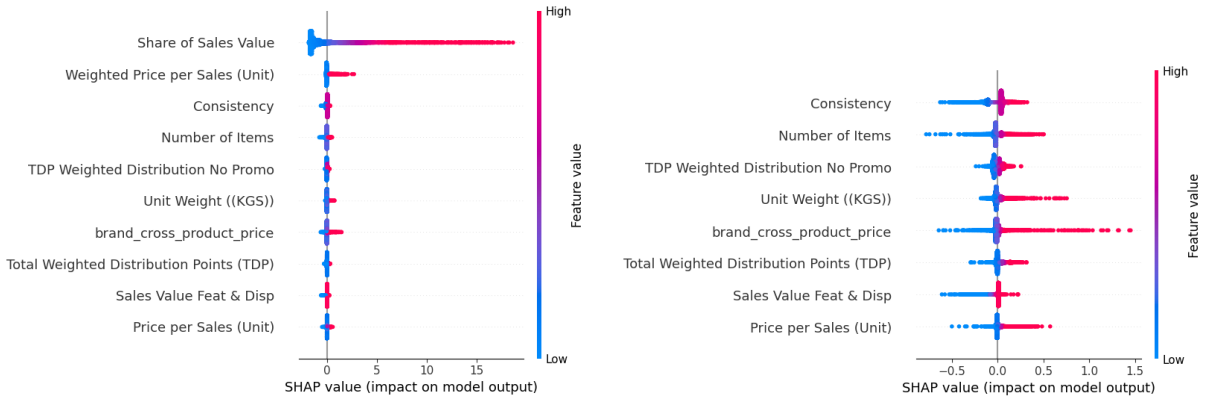
This section will shed a light how utilizing SHAP values can be beneficial to explain ensemble machine learning model output. GBR will be interpreted since it achieved the best results. The section is structured in the following manner. Firstly, global feature importances will be interpreted and a subcategory analysis will be carried out to understand the potential differences. Secondly, a case-study of local interpretations will be provided to understand how SHAP values help understand individual predictions

5.3.1 Global Feature Importances

Firstly, SHAP allows to inspect global feature importances of the predictions. Considering the vast number of features present in the data, the effects of the lags were aggregated to

their respective variable, and the same was done for the item and brand-cross product effects. Whilst individually those variables may have not had a large effect, their cumulative impact on predictions are larger. Every dot within this particular plot corresponds to an individual observation in the test set. This positioning of each dot is important, since it represents the SHAP value for that particular observation. Positive SHAP values indicate that the feature increased the predicted value against the baseline, whilst negative SHAP values indicate a decrease. The color of each dot represents the whether the actual value of the feature is high or low, with red dots indicating higher values and blue dots indicating smaller values. The plot can be inspected below (see Figure 5.4).

Figure 5.4.
Summary Plot of the Top 10 Most Important Features



The summary plot discloses the top 10 most impactful features on the model's output, ranked by their importance. The lags of *Share of Sales Value* are the most significant predictors of Share of Sales Value, followed by *Weighted Price per Sales (Unit)*. This is expected since both features are constructed based of the dependent variable. Intriguingly, within feature *Share of Sales Value*, the most substantial effect comes from higher values (red points) as they significantly amplify the model's predictions. Similarly, yet on a smaller scale, this also applies to *Weighted Price per Sales (Unit)*. Considering the large impact of *Share of Sales Value* and *Weighted Price per Sales per Unit*, a supportive figure is provided to investigate the smaller effect distributions. A considerable feature is *Consistency*, which distribution indicates that with low values (meaning when the product is inconsistent), it has a negative effect on the predicted sales value share. Moreover, three distribution metrics are deemed important by the model, which are *Number of Items*, *TDP Weighted Distribution No Promo*, and *TDP*. All three metrics have a positive impact when their values are high, and a negative one when they are low as indicated by the dense concentration of positive SHAP values. This result showcases how important it is to penetrate high-value stores and ensure a large allocation of shelf-space within them. Additionally, *Unit Weight (KGS)* signifies an important impact, with larger products generally indicating a more positive impact on the model output, unlike smaller products, which do not have a large negative effect. Lastly, the *brand_cross_product_price* effect is notable, with high

feature values indicating a positive impact on the model output.

To examine whether these importances are homogeneous across the different product subcategories, summary plots were made for each subcategory. Across all subcategories, *Share of Sales Value* lag emerges as the most important predictor alongside *Weighted Price per Sales (Unit)*. Thus, other feature distributions will be inspected (see Appendix B, Figure B.10). *Consistency* is the only feature that is the most important across all subcategories. The aforementioned distribution metrics also keep their positioning across most metrics, displaying that distribution breadth and non-promotional availability can drive share. On the other hand, the influence of *brand_cross_product_price* varies, with Mustard and Ketchup being affected the most by it, meaning that the competitor average brand price impacts those categories the most. This is also supported since the dummy variables of *Brand* is important in the Mustard category, showcasing that the brand dummies impact the model output. Intriguingly, within other categories, brand is not as important as product type or product variant. For example, in both Mayonnaise and Salad Dressings, the product *Type* does play a role in the top 10 most important features, and product *Variant* plays a role in Other Dressings. Lastly, across all categories *Unit Weight (KGS)* is present, with notably having the most impact on Mayonnaise share prediction, with bigger pack sizes indicating a more positive impact on the share predictions. Therefore, these findings reveal the most important predictors for each category. Nonetheless, one has to keep in mind that the impact of these variables is still relatively marginal as compared to the *Share of Sales Value* lags. Yet, understanding what drives share besides past performance can be important as it allows to make changes based on the category needs.

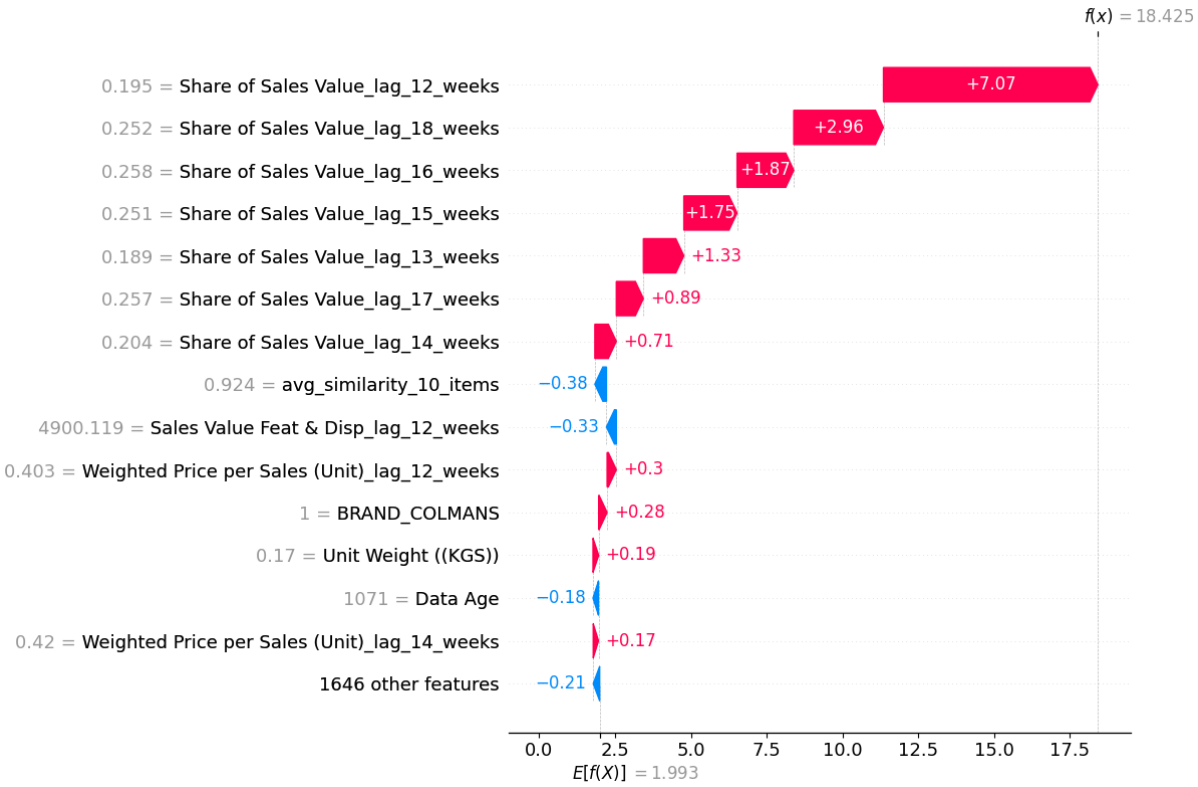
5.3.2 Local Feature Importance

A primary benefit of SHAP values is its ability to explain individual predictions. This can be particularly useful when considering granural decisions, such as putting an SKU on promotion during a particular week, as it helps to understand the decision making process of the model at hand. This section will provide a use-case how this can be done through the use of waterfall plots.

Figure 5.5 provides an interpretation of the model's predictions of the top performing SKU with the highest share within the Mustard subcategory during a particular week. The *f(x)* represents the prediction, and *E[f(X)]* represents the baseline value, which is the predicted share of sales value if no features were present. The plot displays the feature importance of 15 features to the predicted outcome, which is 18.45%, and showcase what is the actual value of the feature when the prediction was being made. Supporting the global feature importance analysis, this plot displays that high values of *Share of Sales Value* lags have a tremendous positive impact on the on the output, with *Share of Sales Value_lag_12_weeks*, *Share of Sales Value_lag_18_weeks*, and *Share of Sales Value_lag_16_weeks*, adding 7.07, 2.96, and 1.87 to the prediction, respectively.

Moreover, *Weighted Price per Sales (Unit)_lag_12_weeks*, *BRAND_COLMANS*, and *Unit Weight (KGS)* signify smaller, yet positive contributions to the model output. These three

Figure 5.5.
Waterfall Plot of Feature Importance for the Highest Share SKU within Mustard Subcategory



variables, however, are imperative to consider, since past performance is intangible, whilst these variables can be indicative of something that could be potentially altered. For example, knowledge that *BRAND_COLMANS* correlates with positive model output (increased share) could indicate that brand awareness correlates with higher share of sales. Moreover, 'Unit Weight' of 0.17KG could be indicative of the optimal size within this subcategory, which could contribute to increased shares. Nevertheless, it is also important to consider the features that are negatively impacting the model output. Variables, such as *avg_similarity_10_items*, *Sales Value Feat Disp_lag_12_weeks* and *Data Age* have negative impacts, -0.38, -0.33, and -0.18, respectively. Firstly, this indicates that a high similarity to the most similar 10 items could be correlated with negative output, potentially indicating that the item is highly substitutable. Another insight derived is that the level of promotional activity associates with a negative predicted share, which indicates more sales value generated from Feature & Displays is correlated with a higher output. Lastly, the *Data Age*, which denotes the length of time since the product has been introduced (in days), showcases a negative effect, potentially disclosing that products are expected to have smaller shares at such a stage of their product lifecycle. This analysis provides an example, which can be crucial for transparency and understanding how the model arrives at the output that it does. Whilst analyzing more individual feature interpretations is out of scope of this thesis, waterfall plots of SKUs with the highest share can be inspected in Appendix B (see section B.4.1).

6. Discussion

6.1 Conclusion

The research question of this study was to evaluate *"How can ensemble machine learning models help develop a forecasting system to predict SKU-level sales value share, and thereby enhance category-level marketing planning in the FMCG retail sector?"*. Considering the versatility of this problem statement, three sub-questions were derived to answer it.

The first sub-question *"How do ensemble Machine Learning models (Random Forest, Gradient Boosting Regressor) compare to a statistical time-series forecasting method (Prophet) in terms of accuracy for predicting weekly SKU-level sales value share?"* intended to assess the predictive performance of ensemble ML models compared to a statistical approach. Prophet was chosen as the statistical benchmark, as this forecasting method is able to handle missing values and inconsistencies within the data, which is something that naturally occurs within real POS data. It was found that both RF and GBR achieved a smaller RMSE across all categories and the entire test-set compared to Prophet, with GBR minimizing RMSE loss the most. When inspecting statistical significance, RF and GBR were significantly different than the statistical method Prophet in terms of their RMSE performance. Such results are inline with the findings of Spiliotis et al. (2020) and Antipov and Pokryshevskaya (2020), who found that Random Forest and Gradient Boosting outperformed their statistical counterparts, whilst also finding GBR as the better method. On the other hand, Moroff et al. (2021) found the statistical models, SARIMAX and ETS, outperformed RF and a form of Gradient Boosting, XGBoost, when comparing their performance. However, the aforementioned study forecasted demand for only 5 products in a series-by-series fashion, which could limit ML effectiveness. Ultimately, the findings of this sub-question contribute to the existing literature in two ways. Firstly, it showcases that whilst Prophet can be a good statistical benchmark, since it is a scalable method that can deal with inconsistencies and gaps within the data, it can introduce large errors that may require expert judgment intervention. Secondly, it adds further evidence that global ML models, specifically GBR, have a stronger predictive performance than another statistical benchmark, that has not been compared against ensemble models within the literature. Therefore, the results indicate that ensemble machine learning models have strong potential to improve forecasting accuracy in SKU-level predictions in a scalable manner.

The second sub-question will provide a twofold interpretation that delves into *How does applying different high-dimensionality reduction techniques, such as a) transforming SKU descriptions into embeddings, b) targeting predictors with LASSO, impact ensemble model performance and its interpretability?*

Firstly, it was found that embeddings did improve the performance on the test-set and proved to be significantly different than the baseline RF model. Moreover, when inspecting the feature

importances, it was found that two embeddings were within the top 10 most important features. After further analysis, it was revealed that one of the embeddings depicted a commonality that was not captured by the one-hot encoded categories, which was the product claim. When inspecting the difference in RMSE for these observations, it was found that for these product observations the RMSE was lower by 0.13, which could potentially be a reason for the increased performance. This showcased the ability of OpenAIs *text-embeddings-3-large* model to detect a latent feature for increased interpretability, and, sequentially enhance model performance. Majority of studies that utilize product embeddings do so through a neural network deep learning approach. For example, Mezzogori and Zammori (2019) transformed product attributes into product embeddings and used Recurrent Neural Networks (RNN) to forecast demand, finding an improvement over the baseline model across all product classes. Another research that supports our finding is by F. Chen et al. (2020), who utilized product embeddings within shopper basket data. They found that integrating product embeddings within their used choice model helped to derive precise estimates of price elasticities, and uncover latent patterns within product attributes. This thesis contributes to the SKU-forecasting field by showcasing that product embeddings can serve as essential features even within a tree-based model. They are useful in reducing sparsity within the data that is often induced due to one-hot encoded features, as well as to observe latent patterns. Ultimately, the most important part is they are scalable method to reduce dimensionality. Whilst this study contained only 583 products, as the number of items grow, their effectiveness increases. This is because they are able to represent key latent features of the product description with a fixed number of dimensions.

Furthermore, tackling the second aspect of the sub-question, it was revealed that TRF did marginally improve the performance in terms of RMSE, with a 1.85% decrease over the baseline. The lambda selected for the LASSO regularization significantly reduced the feature space, from 1662 features to 101 features. Whilst the important features did not change drastically, the computational complexity severely decreased, which in turn diminished the time needed to tune the hyperparameters and train the model by 5 times. This is important, since if the model were to be trained on more categories or on a more extensive hyperparameter grid, this would save significant resources, whilst retaining the most important features. On the other hand, however, LASSO did not recognize the importance of embeddings the same way that Random Forest without regularization did. Thus, the performance remained consistent with the TRF without embeddings, which could entail that utilizing embededdings and applying regularization may not be very effective. When relating these findings to the original study, the improvement in accuracy is not as pronounced, yet the general conclusion still aligns with the findings of Borup et al. (2023). That study found that applying LASSO prior to Random Forest achieved a 13% accuracy improvement over the baseline Random Forest. Overall, these findings contribute to the existing research by offering evidence that TRF can also be useful within a different context - forecasting SKU sales value share.

Lastly, the third sub-question of this study investigated *"What are the most important*

factors that predict share of sales value and how do SHAP values help interpret the model output?" Firstly, global feature importances were inspected. As expected, the lags of "Share of Sales Value" and "Weighted Price per Sales (Unit)" were the most significant, since they directly relate to the past performance of the dependent variable, potentially indicating the trend of the SKU performance. When a dot-plot was inspected, showcasing the distribution of observations with their SHAP values, it revealed that the most substantial effects come from higher values of 'Share of Sales Value', amplifying the model's predictions. Other features, which effects are smaller, yet important, include "Consistency", which has a negative impact when products are inconsistent (i.e., have missing periods). A key contributor were various distribution metrics like "Number of Items," "TDP Weighted Distribution No Promo," and "TDP," which all positively impact the model when their values are high. Additionally, "Unit Weight (KGS)" and "brand_cross_product_price" showcase smaller, yet also important effects, with larger product weights and higher values of cross_product_effects generally contributing positively to the model's output. Moreover, this study showcased how SHAP values can be useful to understand and increase transparency of individual predictions. The waterfall plot, which showcased each features individual contribution to the prediction of the highest share of SKU had multiple revelations. Firstly, the insights echoed the global feature importance analysis, yet it also helped uncover some effects that are unique to the item. This can be helpful for category marketing planning, since understanding why the model predicts the outcome for a particular SKU in a given week can give a justification for making specific decisions. One study that used SHAP values to interpret SKU level predictions found that 'Brand Price and Promotion Features within period t' was the most important predictor Antipov and Pokryshevskaya (2020). However, the authors assumed to know these price and promotion features at the time of prediction, which is a key difference from this research, since we do not make that assumption and only use historical data. Moreover, the importance of distribution metrics, such as TDP, aligns with the previous research that highlights the strong relationship between (market) share and distribution (Hirche, Farris et al., 2021; Wilbur & Farris, 2014). The finding of larger 'Unit Weight (KGS)' positively impacting the share also supports previous research, which found that consumers are more likely to buy larger pack sizes and hence consume more (Hieke et al., 2016; Hirche, Greenacre et al., 2021). Yet, this study also provides evidence that 'Unit Weight (KGS)' is not uniformly important through all categories, which could entail that there is a threshold of an attractive pack size. Thus, adjusting pack size of SKUs has to be done carefully, with respect to each category's needs. Most importantly, this study contributes to the existing research gap of lack of studies that use explainability methods to interpret black-box models (Fildes et al., 2022). It showcases that SHAP values is a far superior tool than just permutation importance in multiple ways. Firstly, they provide an indication whether the feature is affecting the model negatively or positively. Secondly, SHAP values can provide interpretations both on the global and the local level. This thesis showcases how investigating local effects can reveal how specific unique insights, which can be relevant to practitioners when making decisions about SKUs.

Therefore, synthesizing the insights from these three sub-questions, this thesis shows that ensemble machine learning models can be a backbone of a forecasting system to predict SKU-level sales value share. This research demonstrates that both ensemble models tested outperform the statistical time-series benchmark, which showcases their applicability within this problem. More importantly, applying ensemble machine learning is scalable, since the method utilizes cross-learning, which means that the only one model is developed for all the predictions, instead of the usual series-by-series fashion, which can require tuning the model for each product. Moreover, this research explores multiple methods of how to improve predictions by dealing with the high-dimensional feature space which results from the need to capture granural effects. Transforming product descriptions into text embeddings and using them over one-hot encoded categorical features showcased an improved performance, yet the more significant effect was applying Targeted Random Forest with LASSO regularization as the initial step. Thus, both methods showcase the importance of addressing multicollinearity and sparsity even within ensemble machine learning models, that are often lauded for their ability to deal with such data. Ultimately, despite the challenge of the black-box nature of these models, category marketing planning can be enhanced with the application of explainable AI methods. The use of SHAP values provides valuable insights, revealing both global feature importance and local individual prediction explanations. This transparency is essential for decision-makers to ensure increased trust in the model's outputs, ultimately aiding planning efforts and data-driven decision making.

6.2 Managerial Implications

Based on these insights, there are three key managerial implications. It is important to note that these implications are primarily relevant for managers within the suppliers side, such as Unilever.

Firstly, category managers can use a forecasting system of SKU sales value share to make informed price, promotion and distribution related decisions. The forecast provides a 12 week view into the future, which is a strategically important period, as it gives enough time to gather further insights, engage with the retailer and make a decision. For example, if the forecast indicates that the share is expected to decrease over the next 12 weeks, category managers could introduce a promotion that could counteract this expectation. Thus, it enables managers to act proactively, rather than reactively on past information.

Secondly, the setup of the forecasting system can be relevant for customer strategy & planning (CS&P) managers. An advantage of forecasting at the SKU-level is that the predictions can be aggregated to a higher level. Thus, managers could investigate the predictions for relevant product attributes or even brands. This flexibility can enable them to make decisions at a higher-aggregate level, and the ability to drill down to an SKU level. This is particularly relevant for CS&P managers, as they construct customer strategies on a higher granularity. Moreover, this study builds a model using the data from two retailers, and sales value share is calculated on a retailer-chain basis. This means that managers can derive a forecast that is specific to a retailer,

which enables them to make a more specific decision, which could benefit the relationship with the retailer.

Lastly, utilizing SHAP values for interpreting model outputs can be an imperative aspect for managers. Whilst SHAP values do not prove causality, they provide valuable transparency into how the model makes predictions and what features are affecting those predictions the most. Considering the cross-functional alignment needed to make decisions in FMCG, with category managers, marketing managers and supply chain managers often engaging, interpretability is crucial for justifying decisions to these different stakeholders. Besides the overall predictions, this thesis also showcases that individual predictions can be interpreted, revealing SKU specific factors. This is relevant, as inspecting local predictions and their feature importance can enable more targeted planning by the managers.

6.3 Limitations

This study encountered several limitations that ought to be discussed. These limitations were primarily related to data characteristics, methodological approaches, and computational resources, which are described in detail below.

There were several issues regarding the data characteristics that prevented gaining deeper insights. Firstly, the data that this study obtained was on a retailer-chain level rather than the store level. One of the implications of forecasting at a higher level is that it might be harder to capture granural and localized effects. For example, a store-level model could reveal which stores are particularly driving the share of products, allowing for more targeted decisions. Another data limitation is the access to only the data of two retailers. Considering that models tested within this study employ cross-learning - where one model is trained on all observations rather than in a series-by-series fashion per product - extra observations could be particularly useful to improve predictive performance. Finally, this study does not implement macroeconomic trends as external variables due to the lack of access at the weekly level. A survey has showcased that grocery retail has been a sector where consumers have been the most price sensitive, which could explain the rise in share of particular products, such as Private Label (Bansal, 2023). Thus, incorporating such features as consumer price inflation could be particularly useful.

Another limitation within this study lies in one of its key methodological characteristics - the direct multi-output forecasting of the next 12 weeks. This study took a conservative approach and assumed that pricing, promotional and distributional plans were not known in advance. This was done to mimic a real-world scenario, since SKUs of competitors were also forecasted, and it would be impossible to obtain their pricing, promotion and distributional plans. Therefore, this study purely relied on historical data that was available, which came with a set of implications. Since the forecast was set up to predict the next 12 weeks in one-go, this meant that it was difficult to capture immediate marketing effects. This can explain why cross-product effects were not highly prominent within the feature importance analysis. Whilst it was attempted to mitigate this limitation by incorporating lags of promotional variables to account for any potential trend,

it still proved to be limiting. On the other hand, alternative forecasting methods had multiple disadvantages. For instance, recursive multi-step ahead forecasting relies on predicting every next week based on the past week's prediction. This, however, means that for every external regressor used, a prediction would also need to be generated. The drawback of this is the potential compounding error, also known as error propagation (Cerqueira et al., 2024). This implies that if one prediction is far off from the target for that week, it can distort the next set of predictions heavily. Moreover, direct multi-step forecasting implies making a separate model for each predicted week, which can be very computationally expensive (Livieris & Pintelas, 2022). Nevertheless, this is a disadvantage that does not allow to reveal the immediate effects of changes in prices and promotion.

Lastly, this study suffered from limited memory constraints. The PC on which the models were ran had 16GB of RAM. This proved to be an issue when initially constructing the feature space and accounting for a more significant number of cross-product effects, which significantly increased the number of external regressors. Due to this highly dimensional feature space, the PC would simply crash and would have to be restarted. The primary effort to counteract this issue was a function to change the data-types of the variables, from *float64* to *float16*, and *int64* to *int8* (for dummy variables). Whilst this helped, the model could not contain more cross-product-effects than it currently does, which potentially means that some valuable features were left out of the model.

6.4 Directions for future research

Based on the findings, described limitations and the current landscape of retail scholarly literature, the directions for future research will be described.

Firstly, future research should aim to to obtain not only more, but also higher quality data. This thesis suffers from the limitation of having complete data of only two retailer chains, which has both theoretical and practical limitations. Ensemble machine learning models can benefit from more data, as more examples of SKUs can increase accuracy. One of the ways this can be done is obtaining store-specific data instead of retailer-chain data, as having individual store data would expand the number of observations significantly. Moreover, store-level data can help capture individual store-effects, which have been showcased to have an impact when forecasting sales metrics. Lastly, majority of the SKU forecasting research only incorporates marketing mix effects as the primary features. Future research should aim to incorporate more external factors to account for potential macroeconomic effects, such as consumer price index. Macroeconomic outlook can influence consumers price sensitivity, and multiple papers that forecasted other products have showcased effectiveness incorporating them into models (Gao et al., 2018; Zhang et al., 2019).

Thirdly, another insightful research area would be to further evaluate the effect on targeting predictors as an initial step prior to applying ensemble machine learning models. Whilst this thesis follows Borup et al. (2023) methodology and applies LASSO, other feature selection

methods can be tested and applied. For example, a comparative study exploring the effectiveness between applying Elastic Net and LASSO could prove fruitful, as it can highlight the differences how both algorithms select features. Elastic Net can potentially be even more beneficial for feature selection and interpretability, as LASSO is known to struggle with highly correlated features, potentially selecting a less interpretable feature from a group of correlated ones (Ali & Gürlek, 2020).

Lastly, future research should investigate under what conditions applying global machine learning models is more effective than series-by-series models. For example, this research tested subcategories that have quite similar product types, which can be helpful for cross-learning, since the model is able to generalize from similar examples that reside in the training data. Examining whether cross-learning is still valuable when forecasting for significantly different product categories could provide an understanding to what extent applying global models is applicable. For example, Spiliotis et al. (2020) that when forecasting intermittent demand, cross-learning was less effective than forecasting for series by series fashion with Random Forest and Gradient Boosting. Thus, understanding under which conditions applying global models is beneficial can prove to be an insightful avenue for future research.

References

Aburto, L. & Weber, R. W. (2007). Improved supply chain management based on hybrid demand forecasts. *Applied Soft Computing*, *7*(1), 136–144. https://doi.org/10.1016/j.asoc.2005.06.001

Agrawal, D. & Schorling, C. (1996). Market share forecasting: An empirical comparison of artificial neural networks and multinomial logit model. *Journal of Retailing*, *72*(4), 383–407. https://doi.org/10.1016/s0022-4359(96)90020-2

Ali, G. & Gürlek, R. (2020). Automatic interpretable retail forecasting with promotional scenarios. *International Journal of Forecasting*, *36*(4), 1389–1406. https://doi.org/10.1016/j.ijforecast.2020.02.003

Ali, G. & Pinar, E. (2016). Multi-period-ahead forecasting with residual extrapolation and information sharing — Utilizing a multitude of retail series. *International Journal of Forecasting*, *32*(2), 502–517. https://doi.org/10.1016/j.ijforecast.2015.03.011

Ali, G., Sayın, S., Van Woensel, T. & Fransoo, J. J. (2009). SKU demand forecasting in the presence of promotions. *Expert Systems with Applications*, *36*(10), 12340–12348. https://doi.org/10.1016/j.eswa.2009.04.052

Alldredge, K., Lowrie, J. & Schmutzler, R. (2017). *Global customer and channel management: What the best CPG companies do.* Retrieved from `https://www.mckinsey.com/industries/consumer-packaged-goods/our-insights/global-customer-and-channel-management-what-the-best-cpg-companies-do`

Andrade, L. A. C. & Da Cunha, C. B. (2023). Disaggregated retail forecasting: A gradient boosting approach. *Applied Soft Computing*, *141*, 110283. https://doi.org/10.1016/j.asoc.2023.110283

Antipov, E. & Pokryshevskaya, E. (2020). Interpretable machine learning for demand modeling with high-dimensional data using Gradient Boosting Machines and Shapley values. *Journal of Revenue and Pricing Management*, *19*(5), 355–364. https://doi.org/10.1057/s41272-020-00236-4

Babai, M. Z., Boylan, J. E. & Rostami-Tabar, B. (2021). Demand forecasting in supply chains: a review of aggregation and hierarchical approaches. *International Journal of Production Research*, *60*(1), 324–348. https://doi.org/10.1080/00207543.2021.2005268

Baltas, G. (2005). Modelling category demand in retail chains. *Journal of the Operational Research Society*, *56*(11), 1258–1264. https://doi.org/10.1057/palgrave.jors.2601972

Bansal, B. (2023, 2). *Global: Where are consumers most price conscious about clothes and shoes?* Retrieved from `https://business.yougov.com/content/46216-global-where-are-consumers-most-price-conscious-about-clothes-and-shoes`

Barker, J. (2020). Machine learning in M4: What makes a good unstructured model? *International Journal of Forecasting*, *36*(1), 150–155. ht-

tps://doi.org/10.1016/j.ijforecast.2019.06.001

Basuroy, S., Mantrala, M. & Walters, R. (2001). The impact of category management on retailer prices and Performance: Theory and evidence. *Journal of Marketing*, *65*(4), 16–32. https://doi.org/10.1509/jmkg.65.4.16.18382

Bergmeir, C. & Benítez, J. M. (2012). On the use of cross-validation for time series predictor evaluation. *Information Sciences*, *191*, 192–213. https://doi.org/10.1016/j.ins.2011.12.028

Borup, D., Christensen, B. J., Mühlbach, N. S. & Nielsen, M. S. (2023). Targeting predictors in random forest regression. *International Journal of Forecasting*, *39*(2), 841–868. https://doi.org/10.1016/j.ijforecast.2022.02.010

Breiman, L. (2001). Random Forests. *Machine Learning*, *45*(1), 5–32. https://doi.org/10.1023/a:1010933404324

Brodie, R. J., Danaher, P. J., Kumar, V. & Leeflang, P. S. H. (2001). *Econometric models for forecasting market share*. https://doi.org/10.1007/978-0-306-47630-3{27

Broniarczyk, S. M., Hoyer, W. D. & McAlister, L. (1998). Consumers' perceptions of the assortment offered in a grocery category: the impact of item reduction. *Journal of Marketing Research*, *35*(2), 166–176. https://doi.org/10.1177/002224379803500203

Cain, P. (2005). Modelling and forecasting brand share: A dynamic demand system approach. *International Journal of Research in Marketing*, *22*(2), 203–220. https://doi.org/10.1016/j.ijresmar.2004.08.002

Cerqueira, V., Torgo, L. & Bontempi, G. (2024). Instance-based meta-learning for conditionally dependent univariate multi-step forecasting. *International Journal of Forecasting*. https://doi.org/10.1016/j.ijforecast.2023.12.010

Chandon, P., Wansink, B. & Laurent, G. (2000). A benefit congruency framework of sales promotion effectiveness. *Journal of Marketing*, *64*(4), 65–81. https://doi.org/10.1509/jmkg.64.4.65.18071

Chen, F., Liao, X., Proserpio, D. & Troncoso, I. (2020). Product2VEC: Understanding product-level competition using representation learning. *Social Science Research Network*. https://doi.org/10.2139/ssrn.3519358

Chen, J., Koju, W., Xu, S. & Liu, Z. (2021, 3). Sales forecasting using deep neural network and SHAP techniques.. https://doi.org/10.1109/icbaie52039.2021.9389930

Chen, S., Ke, S., Han, S., Gupta, S. & Sivarajah, U. (2024). Which product description phrases affect sales forecasting? An explainable AI framework by integrating WaveNet neural network models with multiple regression. *Decision Support Systems*, *176*, 114065. https://doi.org/10.1016/j.dss.2023.114065

Chen, S., Ngai, E. W., Ku, Y., Xu, Z., Gou, X. & Zhang, C. (2023). Prediction of hotel booking cancellations: Integration of machine learning and probability model based on interpretable feature interaction. *Decision Support Systems*, *170*, 113959. https://doi.org/10.1016/j.dss.2023.113959

Clyde, M., Çetinkaya Rundel, M., Rundel, C., Banks, D., Chai, C. & Huang, L.

(2022). *Chapter 7 Bayesian Model Choice — An Introduction to Bayesian thinking*. Retrieved from https://statswithr.github.io/book/bayesian-model-choice.html#definition-of-bic

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

Curry, D. J., Divakar, S., Mathur, S. K. & Whiteman, C. H. (1995). BVAR as a category management tool: An illustration and comparison with alternative techniques. *Journal of Forecasting*, *14*(3), 181–199. https://doi.org/10.1002/for.3980140304

Curtis, A., Lundholm, R. J. & McVay, S. E. (2014). Forecasting sales: A model and some evidence from the retail industry. *Contemporary Accounting Research*, *31*(2), 581–608. https://doi.org/10.1111/1911-3846.12040

De Almeida, W. M. & Da Veiga, C. P. (2022). Does demand forecasting matter to retailing? *Journal of Marketing Analytics*, *11*(2), 219–232. https://doi.org/10.1057/s41270-022-00162-x

Dekker, M. M. J., Van Donselaar, K. K. & Ouwehand, P. (2004). How to use aggregation and combined forecasting to improve seasonal demand forecasts. *International Journal of Production Economics*, *90*(2), 151–167. https://doi.org/10.1016/j.ijpe.2004.02.004

Divakar, S., Ratchford, B. T. & Shankar, V. (2005). Practice prize article—CHAN4CAST: A multichannel, multiregion sales forecasting model and decision support system for consumer packaged goods. *Marketing Science*, *24*(3), 334–350. https://doi.org/10.1287/mksc.1050.0135

Dupre, K. & Gruen, T. W. (2004). The use of category management practices to obtain a sustainable competitive advantage in the fast-moving-consumer-goods industry. *Journal of Business Industrial Marketing*, *19*(7), 444–459. https://doi.org/10.1108/08858620410564391

Fader, P. S. (1993). Integrating the Dirichlet-multinomial and multinomial logit models of brand choice. *Marketing Letters*, *4*(2), 99–112. https://doi.org/10.1007/bf00994069

Fader, P. S. & Hardie, B. G. S. (1996). Modeling consumer choice among SKUs. *Journal of Marketing Research*, *33*(4), 442–452. https://doi.org/10.1177/002224379603300406

Falatouri, T., Darbanian, F., Brandtner, P. & Udokwu, C. (2022). Predictive analytics for demand forecasting – A comparison of SARIMA and LSTM in retail SCM. *Procedia Computer Science*, *200*, 993–1003. https://doi.org/10.1016/j.procs.2022.01.298

Fildes, R., Kolassa, S. & Ma, S. (2022). Post-script—Retail forecasting: Research and practice. *International Journal of Forecasting*, *38*(4), 1319–1324. https://doi.org/10.1016/j.ijforecast.2021.09.012

Fildes, R., Ma, S. & Kolassa, S. (2019). Retail forecasting: Research and practice. *International Journal of Forecasting*, *38*(4), 1283–1318. https://doi.org/10.1016/j.ijforecast.2019.06.004

Fisher, M. (2020, 10). *Which products should you stock?* Retrieved from https://hbr.org/2012/11/which-products-should-you-stock

Flom, P. L. & Cassell, D. L. (2007, 1). Stopping stepwise: Why stepwise and similar selection methods are bad, and what you should use.. Retrieved from http://denversug.org/presentations/2010CODay/StopStepPresntn.pdf

Gabel, S. & Timoshenko, A. (2022). Product Choice with large assortments: A scalable deep-learning model. *Management Science*, *68*(3), 1808–1827. https://doi.org/10.1287/mnsc.2021.3969

Gao, J., Xie, Y., Cui, X., Yu, H. & Gu, F. (2018). Chinese automobile sales forecasting using economic indicators and typical domestic brand automobile sales data: A method based on econometric model. *Advances in Mechanical Engineering*, *10*(2), 168781401774932. https://doi.org/10.1177/1687814017749325

Gentzkow, M., Kelly, B. T. & Taddy, M. (2019). Text as data. *Journal of Economic Literature*, *57*(3), 535–574. https://doi.org/10.1257/jel.20181020

Geurts, M. D. & Whitlark, D. (1993). Forecasting market share. *The Journal of Business Forecasting Methods Systems*, *11*(4), 17–. Retrieved from https://www.questia.com/library/journal/1P3-574882/forecasting-market-share

Gupta, P., Ladia, H., Kakkar, K., Rai, K., Agrawal, Y. C., Mamgain, R. & Gaur, N. (2021). Implementation of demand forecasting – a comparative approach. *Journal of Physics: Conference Series*, *1714*(1), 012003. https://doi.org/10.1088/1742-6596/1714/1/012003

Guryanov, A. (2019, 1). Histogram-Based algorithm for building gradient boosting ensembles of piecewise linear decision trees. In (pp. 39–50). https://doi.org/10.1007/978-3-030-37334-4{4

Hapfelmeier, A. & Ulm, K. (2014). Variable selection by Random Forests using data with missing values. *Computational Statistics Data Analysis*, *80*, 129–139. https://doi.org/10.1016/j.csda.2014.06.017

Heger, J. & Klein, R. (2024). Assortment optimization: a systematic literature review. *OR spectrum/OR-Spektrum*. https://doi.org/10.1007/s00291-024-00752-4

Hieke, S., Palascha, A., Jola, C., Wills, J. & Raats, M. M. (2016). The pack size effect: Influence on consumer perceptions of portion sizes. *Appetite*, *96*, 225–238. https://doi.org/10.1016/j.appet.2015.09.025

Hirche, M., Farris, P., Greenacre, L., Quan, Y. & Wei, S. (2021). Predicting under- and over-performing SKUs within the distribution–market share relationship. *Journal of Retailing*, *97*(4), 697–714. https://doi.org/10.1016/j.jretai.2021.04.002

Hirche, M., Greenacre, L., Nenycz-Thiel, M., Loose, S. M. & Lockshin, L. (2021). SKU performance and distribution: A large-scale analysis of the role of product characteristics with store scanner data. *Journal of Retailing and Consumer Services*, *61*, 102533. https://doi.org/10.1016/j.jretconser.2021.102533

Huang, T., Fildes, R. & Soopramanien, D. (2014). The value of competitive information in fore-

casting FMCG retail product sales and the variable selection problem. *European Journal of Operational Research*, *237*(2), 738–748. https://doi.org/10.1016/j.ejor.2014.02.022

Huang, T., Fildes, R. & Soopramanien, D. (2019). Forecasting retailer product sales in the presence of structural change. *European Journal of Operational Research*, *279*(2), 459–470. https://doi.org/10.1016/j.ejor.2019.06.011

Huber, J. & Stuckenschmidt, H. (2020). Daily retail demand forecasting using machine learning with emphasis on calendric special days. *International Journal of Forecasting*, *36*(4), 1420–1438. https://doi.org/10.1016/j.ijforecast.2020.02.005

Hyndman, R. J. & Athanasopoulos, G. (2018). *Forecasting: principles and practice*. OTexts.

Hübner, A. & Kühn, H. (2012). Retail category management: State-of-the-art review of quantitative research and software applications in assortment and shelf space management. *Omega*, *40*(2), 199–209. https://doi.org/10.1016/j.omega.2011.05.008

Jackson, I., Sáenz, M. J., Li, Y. & Moreno, M. S. R. (2023). Synchromodal supply chains for Fast-Moving consumer goods. *Applied Sciences*, *13*(5), 3119. https://doi.org/10.3390/app13053119

James, G., Witten, D., Hastie, T., Tibshirani, R. & Taylor, J. (2023). *An introduction to statistical learning*. Springer Nature.

Jha, B. K. & Pande, S. (2021, 4). Time Series forecasting model for supermarket Sales using FB-Prophet.. Retrieved from https://doi.org/10.1109/iccmc51019.2021.9418033 https://doi.org/10.1109/iccmc51019.2021.9418033

Jiang, H., Ruan, J. & Sun, J. (2021, 3). Application of machine learning model and hybrid model in retail sales forecast.. https://doi.org/10.1109/icbda51983.2021.9403224

Jin, Y., Williams, B. D., Tokar, T. & Waller, M. A. (2015). Forecasting with temporally aggregated demand signals in a retail supply chain. *Journal of Business Logistics*, *36*(2), 199–211. https://doi.org/10.1111/jbl.12091

Kenton, W. (2024, 5). *Fast-Moving Consumer Goods (FMCG) Industry: Definition, types, and Profitability.* Retrieved from https://www.investopedia.com/terms/f/fastmoving-consumer-goods-fmcg.asp#:~:text=Because%20fast-moving%20consumer%20goods%20have%20such%20a%20high,Unilever%2C%20Procter%20%26%20Gamble%2C%20Nestl%C3%A9%2C%20PepsiCo%2C%20and%20Danone.

Kim, S. & Kim, H. (2016). A new metric of absolute percentage error for intermittent demand forecasts. *International Journal of Forecasting*, *32*(3), 669–679. https://doi.org/10.1016/j.ijforecast.2015.12.003

Klapper, D. & Herwartz, H. (2000). Forecasting market share using predicted values of competitive behavior: further empirical results. *International Journal of Forecasting*, *16*(3), 399–421. https://doi.org/10.1016/s0169-2070(00)00052-2

Krafft, M. & Mantrala, M. K. (2010). *Sales Promotions*. Springer.

Kruger, M. W. & Harper, B. (2006, 1). Market share and product distribution: Re-Tested and extended. *Social Science Research Network*. https://doi.org/10.2139/ssrn.2157528

Kuo, R. J. (2001). A sales forecasting system based on fuzzy neural network with initial weights generated by genetic algorithm. *European Journal of Operational Research*, *129*(3), 496–517. https://doi.org/10.1016/s0377-2217(99)00463-4

Leeflang, P. S. H. & Parreño-Selva, J. (2011). Cross-category demand effects of price promotions. *Journal of the Academy of Marketing Science*, *40*(4), 572–586. https://doi.org/10.1007/s11747-010-0244-z

Li, D., Lin, K., Li, X., Liao, J., Du, R., Chen, D. & Madden, A. D. (2022). Improved sales time series predictions using deep neural networks with spatiotemporal dynamic pattern acquisition mechanism. *Information Processing and Management*, *59*(4), 102987. https://doi.org/10.1016/j.ipm.2022.102987

Li, M., Sun, H., Huang, Y. & Chen, H. (2024). Shapley value: from cooperative game to explainable artificial intelligence. *Autonomous Intelligent Systems*, *4*(1). https://doi.org/10.1007/s43684-023-00060-8

Lin, X., Zhang, B., Zhang, J., Qi, Y. & Hu, H. (2021, 10). A practical framework for forecasting stock keeping unit level seasonal sales.. https://doi.org/10.1109/bigdia53151.2021.9619627

Linardatos, P., Papastefanopoulos, V. & Kotsiantis, S. (2020). Explainable AI: A review of machine learning interpretability methods. *Entropy*, *23*(1), 18. https://doi.org/10.3390/e23010018

Livieris, I. E. & Pintelas, P. (2022). A novel multi-step forecasting strategy for enhancing deep learning models' performance. *Neural Computing Applications*, *34*(22), 19453–19470. https://doi.org/10.1007/s00521-022-07158-9

Ma, S. (2024). Retail store-SKU level replenishment planning with attribute-space graph recurrent neural networks. *Expert Systems with Applications*, *249*, 123727. https://doi.org/10.1016/j.eswa.2024.123727

Ma, S. & Fildes, R. (2017). A retail store SKU promotions optimization model for category multi-period profit maximization. *European Journal of Operational Research*, *260*(2), 680–692. https://doi.org/10.1016/j.ejor.2016.12.032

Ma, S., Fildes, R. & Huang, T. (2016). Demand forecasting with high dimensional data: The case of SKU retail sales forecasting with intra- and inter-category promotional information. *European Journal of Operational Research*, *249*(1), 245–257. https://doi.org/10.1016/j.ejor.2015.08.029

Mamuaya, N. C. (2024). Investigating the impact of product quality, price sensitivity, and brand reputation on consumer purchase intentions in the FMCG sector. *International Journal of Business, Law, and Education*, *5*(2), 1576–1583. https://doi.org/10.56442/ijble.v5i2.614

Medin, D. L., Goldstone, R. L. & Markman, A. B. (1995). Comparison and choice: Relations between similarity processes and decision processes. *Psychonomic Bulletin Review*, *2*(1), 1–19. https://doi.org/10.3758/bf03214410

Mehdiyev, N., Enke, D., Fettke, P. & Loos, P. (2016). Evaluating forecasting methods by

considering different accuracy measures. *Procedia Computer Science*, *95*, 264–271. https://doi.org/10.1016/j.procs.2016.09.332

Meteostat. (2024, May). *Meteostat.* Retrieved from https://meteostat.net/en/place/gb/london?s=03779&t=2024-05-30/2024-06-06

Mezzogori, D. & Zammori, F. (2019). An entity embeddings deep learning approach for demand forecast of highly differentiated products. *Procedia Manufacturing*, *39*, 1793–1800. https://doi.org/10.1016/j.promfg.2020.01.260

Moroff, N. U., Kurt, E. & Kamphues, J. (2021). Machine Learning and Statistics: A Study for assessing innovative Demand Forecasting Models. *Procedia Computer Science*, *180*, 40–49. https://doi.org/10.1016/j.procs.2021.01.127

Nohara, Y. (2022). Explanation of machine learning models using shapley additive explanation and application for real data in hospital. *Computer Methods and Programs in Biomedicine*, *214*, 106584. https://doi.org/10.1016/j.cmpb.2021.106584

OpenAI Platform. (n.d.). Retrieved from https://platform.openai.com/docs/guides/embeddings

Petropoulos, F., Apiletti, D., Assimakopoulos, V., Babaï, M. Z., Barrow, D. K., Taieb, S. B., ... 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

Punia, S. & Shankar, S. (2022). Predictive analytics for demand forecasting: A deep learning-based decision support system. *Knowledge-Based Systems*, *258*, 109956. https://doi.org/10.1016/j.knosys.2022.109956

Ramos, P. & Oliveira, J. M. (2016). A procedure for identification of appropriate state space and ARIMA models based on Time-Series Cross-Validation. *Algorithms*, *9*(4), 76. https://doi.org/10.3390/a9040076

Ramos, P., Oliveira, J. M., Kourentzes, N. & Fildes, R. (2022). Forecasting Seasonal Sales with Many Drivers: Shrinkage or Dimensionality Reduction? *Applied System Innovation*, *6*(1), 3. https://doi.org/10.3390/asi6010003

Segal, M. R. (2004). Machine learning benchmarks and random forest regression. *UCSF: Center for Bioinformatics and Molecular Biostatistics*. Retrieved from https://escholarship.org/content/qt35x3v9t4/qt35x3v9t4.pdf

Spiliotis, E. (2023). *Time Series Forecasting with Statistical, Machine Learning, and Deep Learning Methods: Past, Present, and Future*. Retrieved from https://doi.org/10.1007/978-3-031-35879-1_3 https://doi.org/10.1007/978-3-031-35879-1{3

Spiliotis, E., Makridakis, S., Semenoglou, A.-A. & Assimakopoulos, V. (2020). Comparison of statistical and machine learning methods for daily SKU demand forecasting. *Operational Research*, *22*(3), 3037–3061. https://doi.org/10.1007/s12351-020-00605-2

Surakhi, O., Zaidan, M. A., Fung, P. L., Motlagh, N. H., Serhan, S., AlKhanafseh, M., ... Hussein, T. (2021, 10). Time-Lag selection for Time-Series forecasting using neural network and heuristic algorithm. *Electronics*, *10*(20), 2518. ht-

tps://doi.org/10.3390/electronics10202518

Syntetos, A., Babai, Z., Boylan, J. E., Kolassa, S. & Nikolopoulos, K. (2016). Supply chain forecasting: Theory, practice, their gap and the future. *European journal of operational research*, *252*(1), 1–26. https://doi.org/10.1016/j.ejor.2015.11.010

Taylor, S. J. & Letham, B. (2018). Forecasting at scale. *The American Statistician*, *72*(1), 37–45. Retrieved from https://doi.org/10.1080/00031305.2017.1380080 https://doi.org/10.1080/00031305.2017.1380080

Timonina-Farkas, A., Katsifou, A. & Seifert, R. W. (2020). Product assortment and space allocation strategies to attract loyal and non-loyal customers. *European Journal of Operational Research*, *285*(3), 1058–1076. https://doi.org/10.1016/j.ejor.2020.02.019

Unilever. (2022, 9). Unilever simplifies organisation. Retrieved from https://www.unilever.com/news/press-and-media/press-releases/2022/unilever-simplifies-organisation/

Vallés-Perez, I., Soria-Olivas, E., Martínez-Sober, M., Serrano-López, A. J., Gómez-Sanchís, J. & Mateo, F. (2022). Approaching sales forecasting using recurrent neural networks and transformers. *Expert Systems with Applications*, *201*, 116993. https://doi.org/10.1016/j.eswa.2022.116993

Van Heerde, H. J., Leeflang, P. S. H. & Wittink, D. R. (2002). How promotions work: SCAN*PRO-Based Evolutionary Model Building. *Schmalenbach Business Review*, *54*(3), 198–220. https://doi.org/10.1007/bf03396653

Venkatesan, R. & Farris, P. W. (2012). Measuring and Managing Returns from Retailer-Customized Coupon Campaigns. *Journal of Marketing*, *76*(1), 76–94. https://doi.org/10.1509/jm.10.0162

Wang, D., Thunéll, S., Lindberg, U., Jiang, L., Trygg, J. & Tysklind, M. (2022). Towards better process management in wastewater treatment plants: Process analytics based on SHAP values for tree-based machine learning methods. *Journal of Environmental Management*, *301*, 113941. https://doi.org/10.1016/j.jenvman.2021.113941

Wedel, M., Zhang, J. & Feinberg, F. M. (2015). Implementing retail category management: a model-based approach to setting optimal markups. *Customer Needs and Solutions*, *2*(2), 165–176. https://doi.org/10.1007/s40547-015-0041-4

Wellens, A., Boute, R. & Udenio, M. (2024). Simplifying tree-based methods for retail sales forecasting with explanatory variables. *European Journal of Operational Research*, *314*(2), 523–539. https://doi.org/10.1016/j.ejor.2023.10.039

Wilbur, K. C. & Farris, P. (2014). Distribution and market share. *Journal of Retailing*, *90*(2), 154–167. https://doi.org/10.1016/j.jretai.2013.08.003

Ye, C., Xiong, Y., Li, Y., Liu, L. & Wang, M. (2019). The influences of product similarity on consumer preferences: a study based on eye-tracking analysis. *Cognition, Technology Work*, *22*(3), 603–613. https://doi.org/10.1007/s10111-019-00584-1

Zenor, M. J. (1994). The profit benefits of category management. *Journal of Marketing*

Research, *31*(2), 202–213. https://doi.org/10.1177/002224379403100205

Zhang, C., Tian, Y.-X., Fan, Z.-P., Liu, Y. & Fan, L.-W. (2019). Product sales forecasting using macroeconomic indicators and online reviews: a method combining prospect theory and sentiment analysis. *Soft Computing*, *24*(9), 6213–6226. https://doi.org/10.1007/s00500-018-03742-1

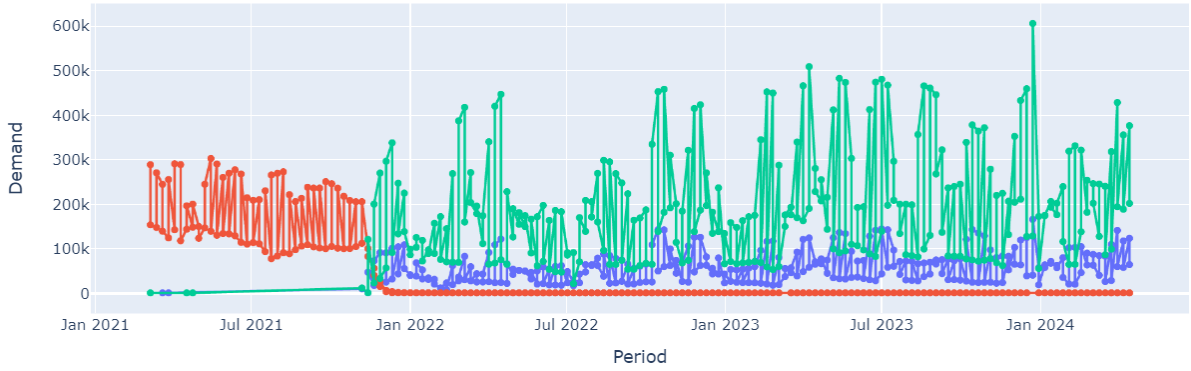
Zotteri, G. & Kalchschmidt, M. G. M. (2007). A model for selecting the appropriate level of aggregation in forecasting processes. *International Journal of Production Economics*, *108*(1-2), 74–83. https://doi.org/10.1016/j.ijpe.2006.12.030

Zotteri, G., Kalchschmidt, M. G. M. & Caniato, F. (2005). The impact of aggregation level on forecasting performance. *International Journal of Production Economics*, *93-94*, 479–491. https://doi.org/10.1016/j.ijpe.2004.06.044

A. Methods Extensions

A.1 Descriptive Statistics

Figure A.1.
SKU Inconsistency Example within Ketchup Subcategory



Return to the reading spot by clicking the number, Section 3.2

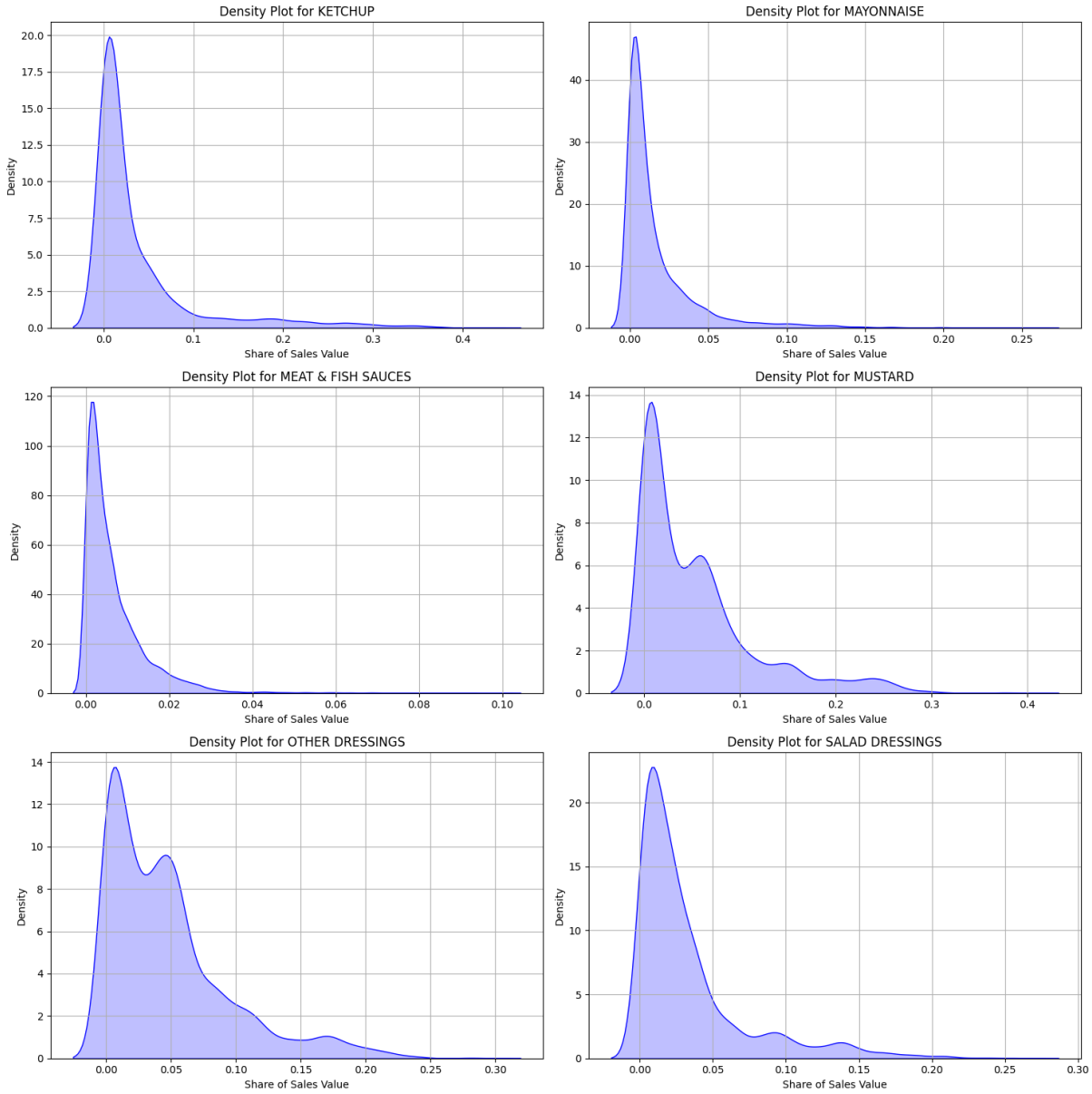
Table A.1. Descriptive Statistics of Numerical Columns Prior to Cleaning

	count	mean	std	min	25%	50%	75%	max
Sales Value Prev	122236	10983.47	24773.36	-1.90	591.04	3631.91	11243.82	645644.91
Sales (KGS) Prev	122236	2448.99	6530.14	0.00	87.55	558.36	2026.18	168339.99
Baseline Sales Value Prev	122236	9890.46	20629.51	0.00	557.99	3419.30	10766.09	640259.35
Incremental Sales Value Prev	122236	1092.48	8697.32	-152418.17	0.00	0.00	0.03	324395.47
Sales Value	122236	10997.41	24794.48	-1.90	588.65	3635.48	11258.20	645644.91
Sales Value Any Promo Prev	122236	2985.54	16019.12	0.00	0.00	0.00	242.32	605682.55
Sales Value No Promo Prev	122236	7995.96	18000.88	-1.90	143.88	2237.06	8553.66	597967.10
Sales Value Feat & Disp Prev	122236	483.11	5261.07	0.00	0.00	0.00	0.00	355803.85
Sales Value Feat w/o Disp Prev	122236	794.57	4850.47	0.00	0.00	0.00	0.00	284561.30
Sales Value Disp w/o Feat Prev	122236	136.21	1832.02	0.00	0.00	0.00	0.00	124341.69
Sales Value Disp or Feat Prev	122236	1413.89	9669.53	-119.14	0.00	0.00	0.00	538462.32
Sales Value Total Disp Prev	122236	618.98	5874.43	-119.14	0.00	0.00	0.00	355803.85
Sales Value Total Multibuy Prev	122236	43.20	562.90	0.00	0.00	0.00	0.00	49514.38
Sales Value TPR Only Prev	122236	1569.60	8609.94	-113.51	0.00	0.00	83.22	301449.17
Base Sales Value Total Disp Prev	122236	343.97	2906.52	-491.55	0.00	0.00	0.00	124184.45
Base Sales Value Total Multibuy Prev	122236	21.30	259.64	0.00	0.00	0.00	0.00	30111.20
Base Sales Value TPR Only Prev	122236	1040.81	5250.30	0.00	0.00	0.00	72.37	237487.24
Base Sales Value Any Promo Prev	122236	1882.55	8573.95	0.00	0.00	0.00	216.55	351546.54
Base Sales Value No Promo Prev	122236	8006.39	18025.81	0.00	144.00	2240.09	8562.34	597985.01
Base Sales Value Feat & Disp Prev	122236	242.27	2318.46	0.00	0.00	0.00	0.00	124184.45
Base Sales Value Feat w/o Disp Prev	122236	496.00	2690.96	0.00	0.00	0.00	0.00	156564.70
Base Sales Value Disp w/o Feat Prev	122236	101.83	1411.80	0.00	0.00	0.00	0.00	96773.13
Base Sales Value Disp or Feat Prev	122236	840.25	4831.95	-491.55	0.00	0.00	0.00	250379.37
Incr Sales Value Total Disp Prev	122236	275.01	3522.27	-29636.34	0.00	0.00	0.00	233088.53
Incr Sales Value Total Multibuy Prev	122236	21.90	357.85	-1797.57	0.00	0.00	0.00	37881.99
Incr Sales Value TPR Only Prev	122236	528.79	4224.93	-152418.17	0.00	0.00	0.00	168277.85
Incr Sales Value Any Promo Prev	122236	1102.98	8707.04	-152418.17	0.00	0.00	3.08	324395.47
Incr Sales Value No Promo Prev	122236	-10.69	87.45	-10311.07	0.00	0.00	0.00	592.11
Incr Sales Value Feat & Disp Prev	122236	240.84	3302.61	-27002.28	0.00	0.00	0.00	231619.40
Incr Sales Value Feat w/o Disp Prev	122236	298.56	2550.67	-20057.02	0.00	0.00	0.00	159721.84
Incr Sales Value Disp w/o Feat Prev	122236	34.38	704.30	-29636.34	0.00	0.00	0.00	61910.72
Incr Sales Value Disp or Feat Prev	122236	573.64	5566.86	-36725.99	0.00	0.00	0.00	294255.56
Price (KGS) Prev	122236	6.65	9.83	-6.33	3.87	5.94	8.30	3019.36
Avg Base Price (KGS) Prev	122236	6.95	9.94	0.00	3.99	6.29	8.75	3019.36
Incr Price (KGS) Prev	122236	23.93	7711.10	-48432.67	0.00	0.00	3.08	2694840.00
Incr Pice (KGS) Any Promo Prev	122236	24.38	7693.02	-67446.19	0.00	0.00	3.08	2689116.00
Any Promo Base (KGS) Price Prev	122236	2.81	4.79	0.00	0.00	0.00	5.53	70.61
TDP Prev	122236	57.25	46.62	0.00	14.73	63.94	86.49	540.13
Distribution Quality Prev	122236	4.01	20.43	0.00	2.06	2.56	3.34	935.32
ACV TDP Prev	122236	53.98	44.30	0.00	13.49	59.71	81.19	517.63
Number of Items Prev	122236	513.89	566.24	0.00	105.10	404.40	666.54	6597.00
Sales Value/ACV TDP Prev	122236	153.80	293.49	-8.72	35.47	74.21	158.39	9347.06
Value / Store Prev	122236	5.36	11.00	-0.00	0.30	1.96	5.89	235.90
TDP Any Promo Prev	122236	10.81	25.77	0.00	0.00	0.00	3.19	496.46
TDP No Promo Prev	122236	46.40	47.00	0.00	2.38	47.11	81.50	529.42
TDP Feat & Disp Prev	122236	0.73	4.39	0.00	0.00	0.00	0.00	81.20
TDP Disp or Feat Prev	122236	4.18	12.60	0.00	0.00	0.00	0.00	228.94
TDP Feat w/o Disp Prev	122236	3.22	10.24	0.00	0.00	0.00	0.00	228.94
TDP Disp w/o Feat Prev	122236	0.23	1.59	0.00	0.00	0.00	0.00	63.91
TDP Total Multibuy Prev	122236	0.09	0.62	0.00	0.00	0.00	0.00	17.04
TDP TPR Only Prev	122236	6.63	17.28	0.00	0.00	0.00	1.59	351.09

Return to the reading spot by clicking the number, Section 4.1.1

A.2 Dependent Variable Inspection

Figure A.2.
Density Plot Share of Sales



Return to the reading spot by clicking the number, Section 4.1.2

Figure A.3.
Histogram Share of Sales

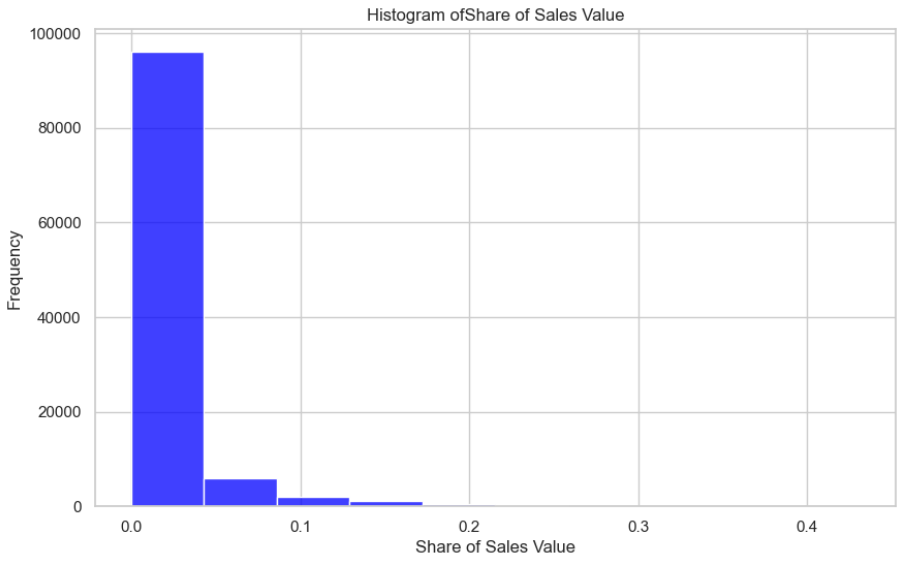
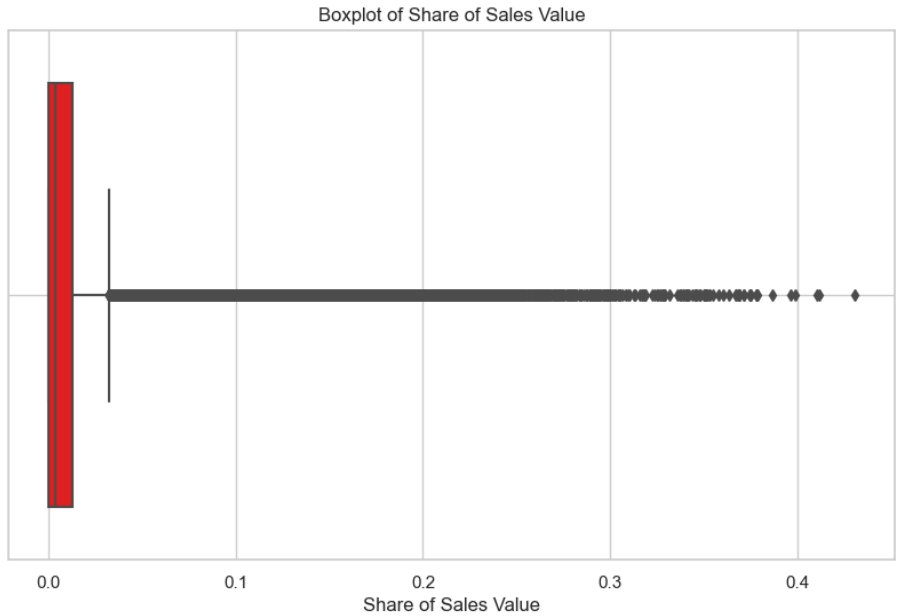
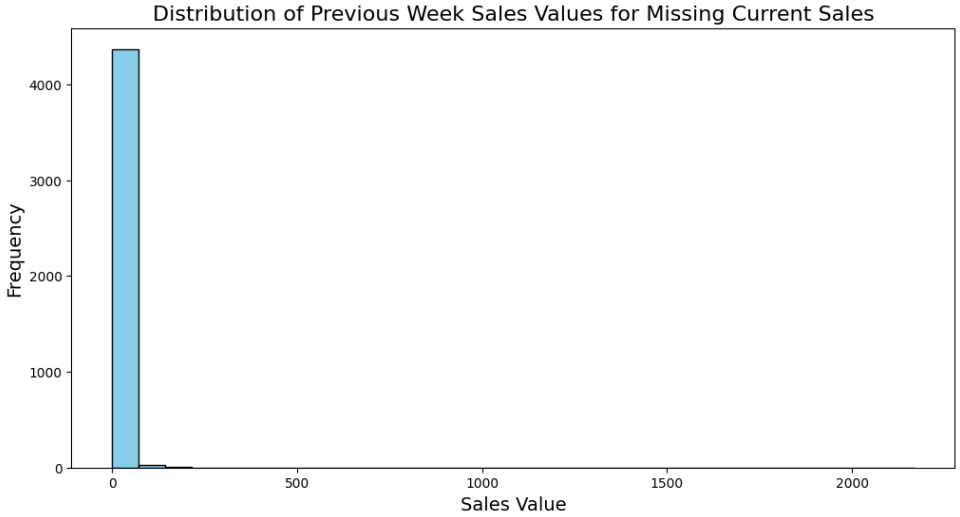


Figure A.4.
Boxplot of Share of Sales



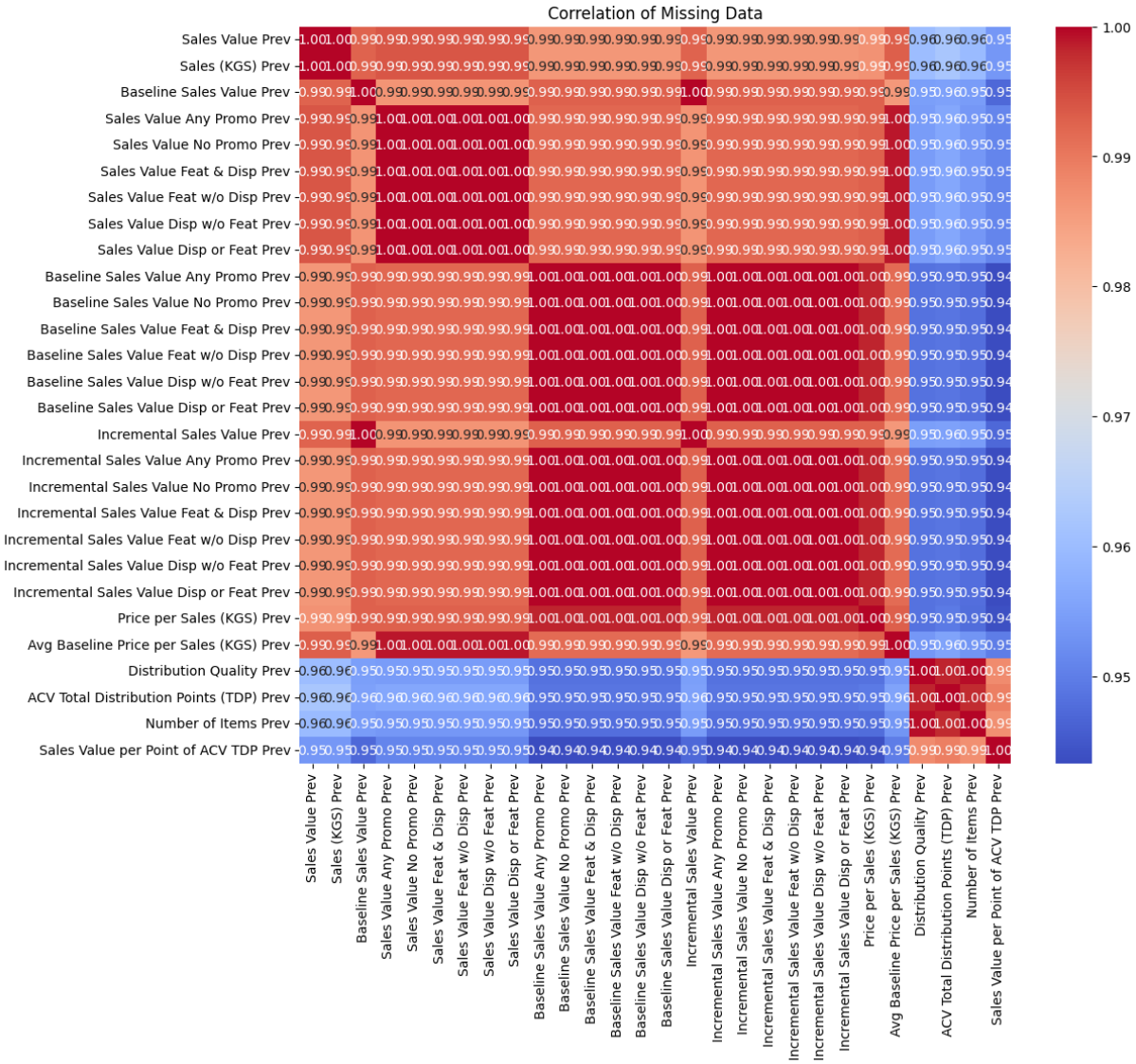
A.3 Missing Value Analysis

Figure A.5.
Sales Value Lag Inspection for Missing Values of 'Sales Value'



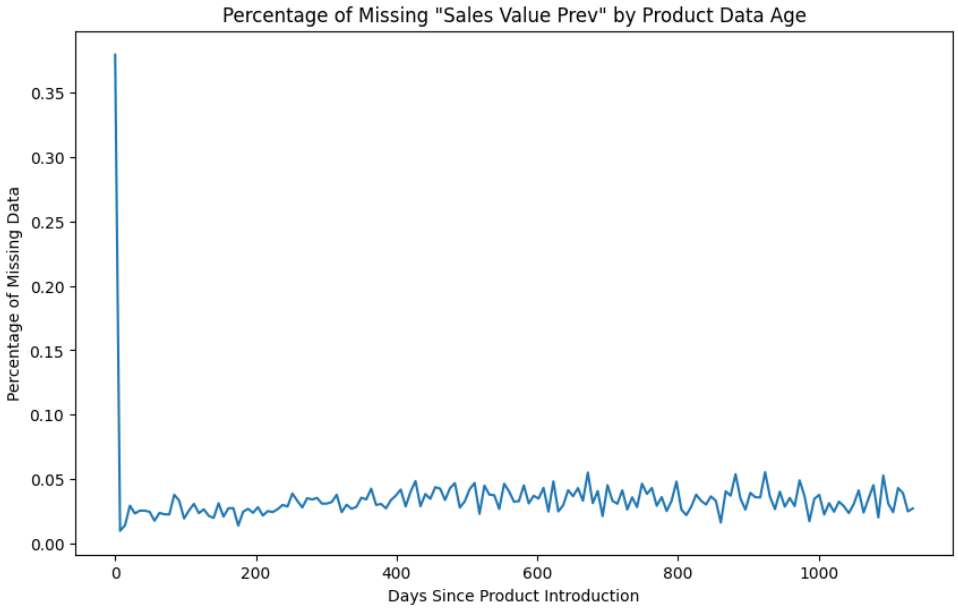
Return to the reading spot by clicking the number, Section 4.1.1

Figure A.6.
Missing Value Correlation Analysis



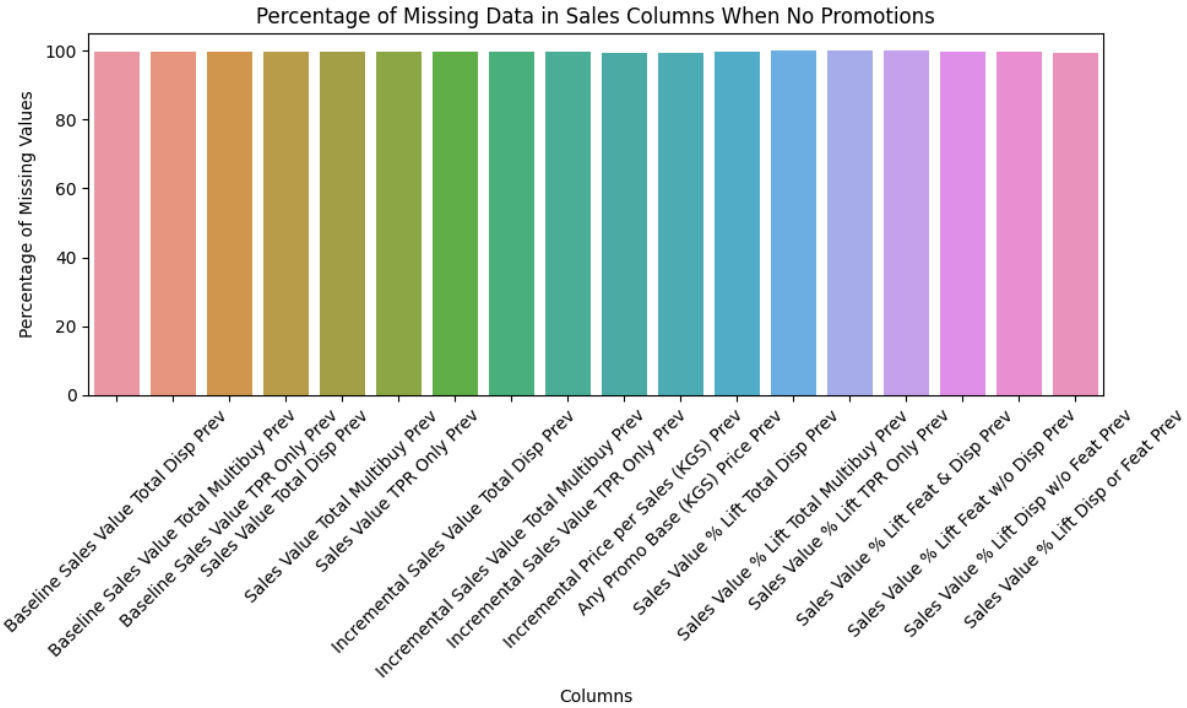
Return to the reading spot by clicking the number, Section 4.1.1

Figure A.7.
Missing Lag Value Cause



Return to the reading spot by clicking the number, Section 4.1.1

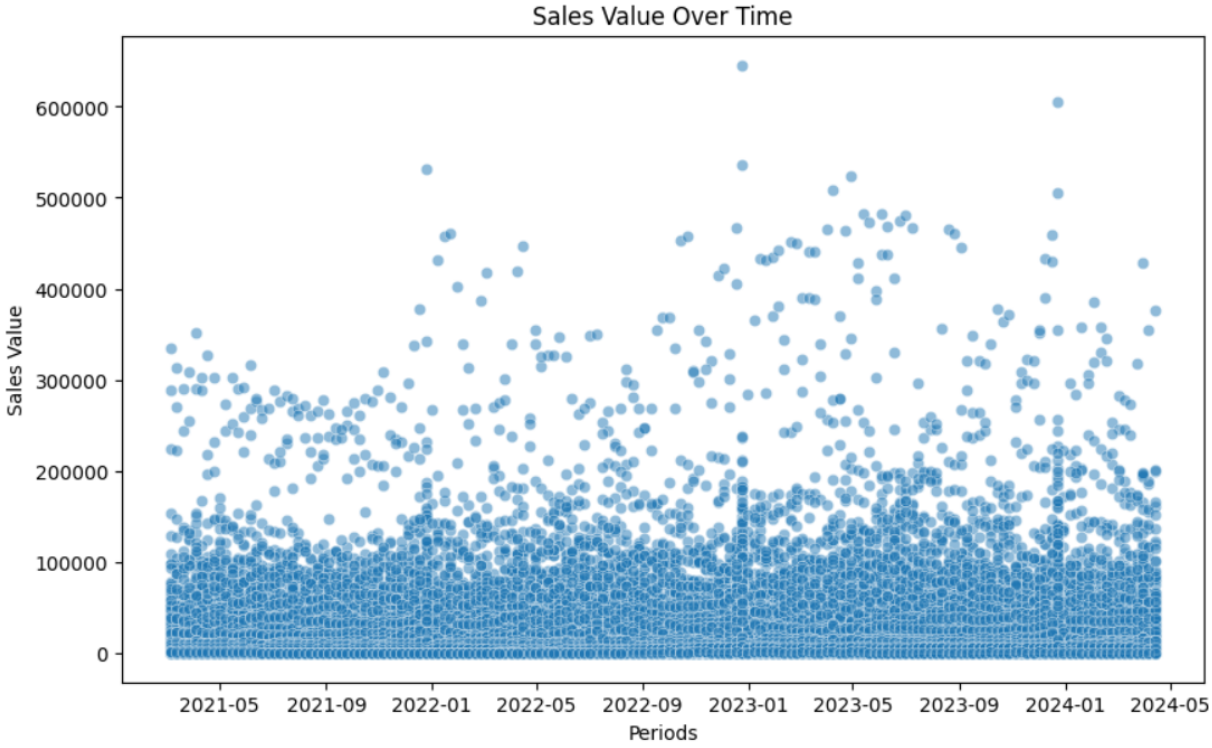
Figure A.8.
Promotion Missing Values Cause



Return to the reading spot by clicking the number, Section 4.1.1

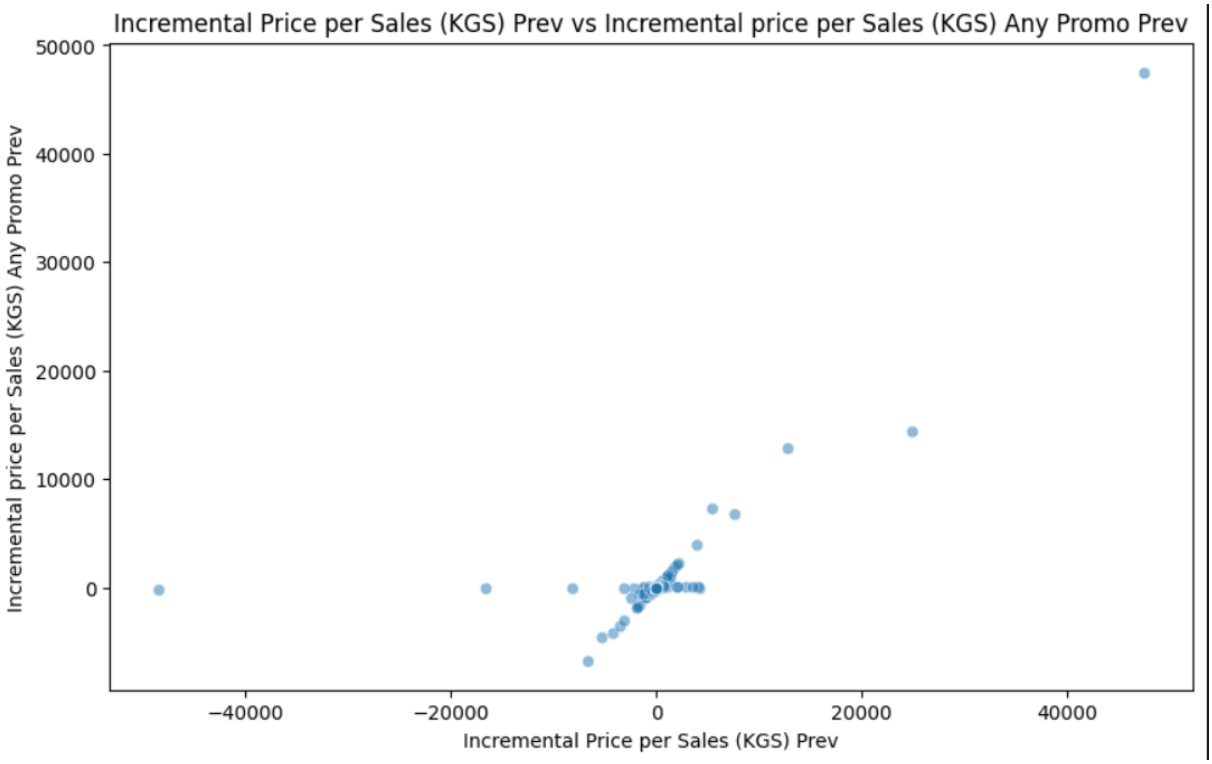
A.4 Validity and Outlier Visualizations

Figure A.9.
Sales Value Inspection



Return to the reading spot by clicking the number, Section 4.1.1

Figure A.10.
Incremental Price Outliers



Return to the reading spot by clicking the number, Section 4.1.1

A.5 Subcategory Quarterly Trends

Figure A.11.
Ketchup Sales Value

Figure A.12.
Mayonnaise Sales Value

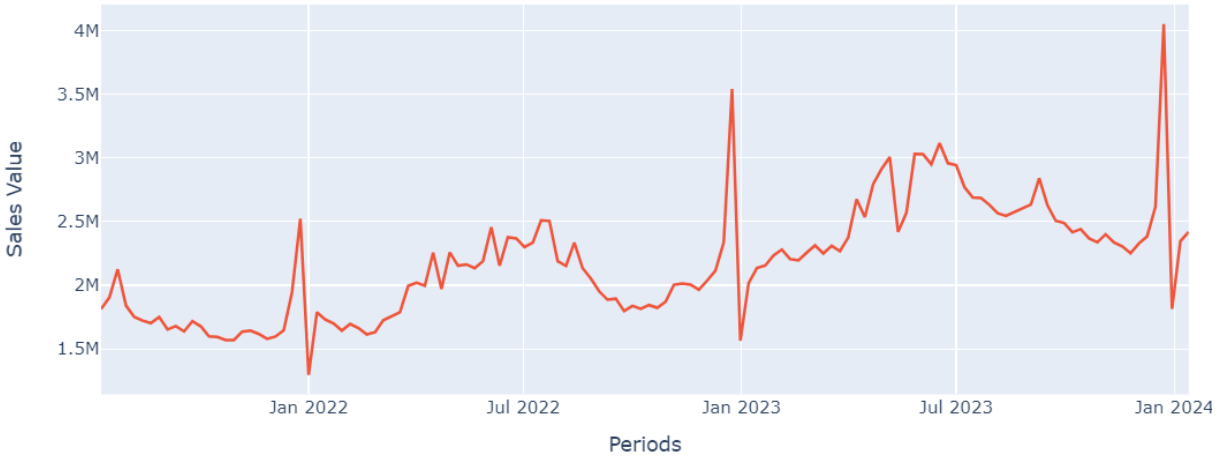


Figure A.13.
Meat & Fish Sauces Sales Value

Figure A.14.
Mustard Sales Value

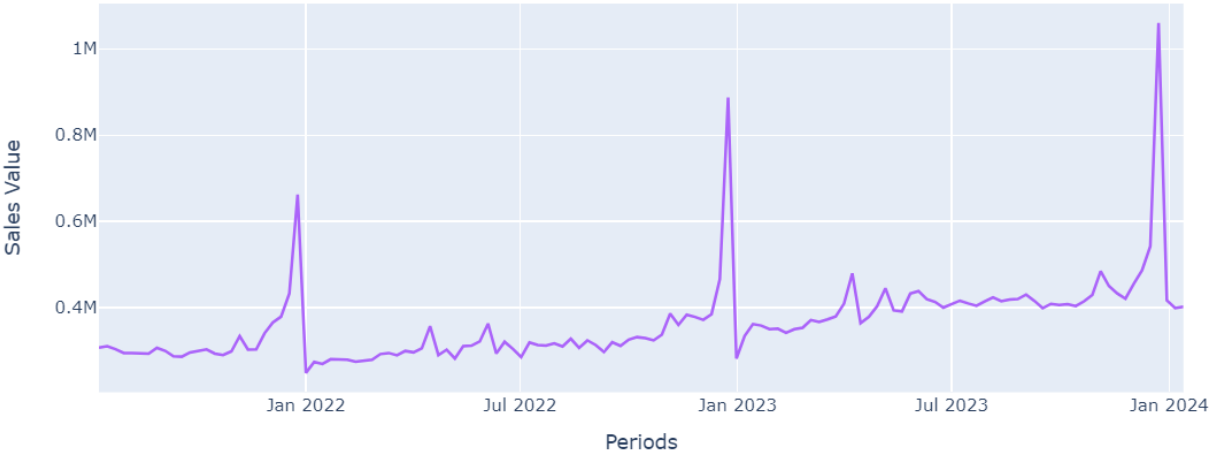


Figure A.15.
Other Dressings Sales Value

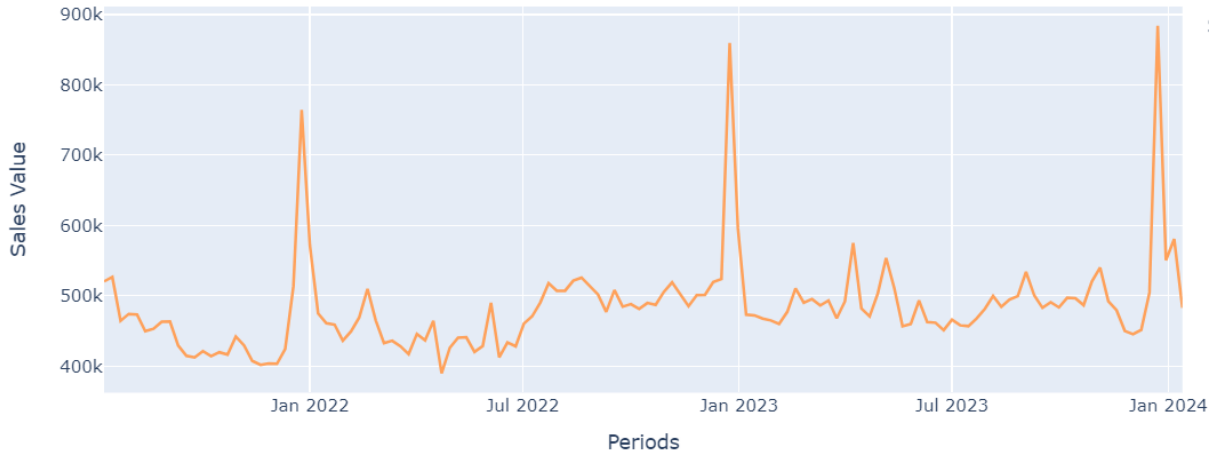
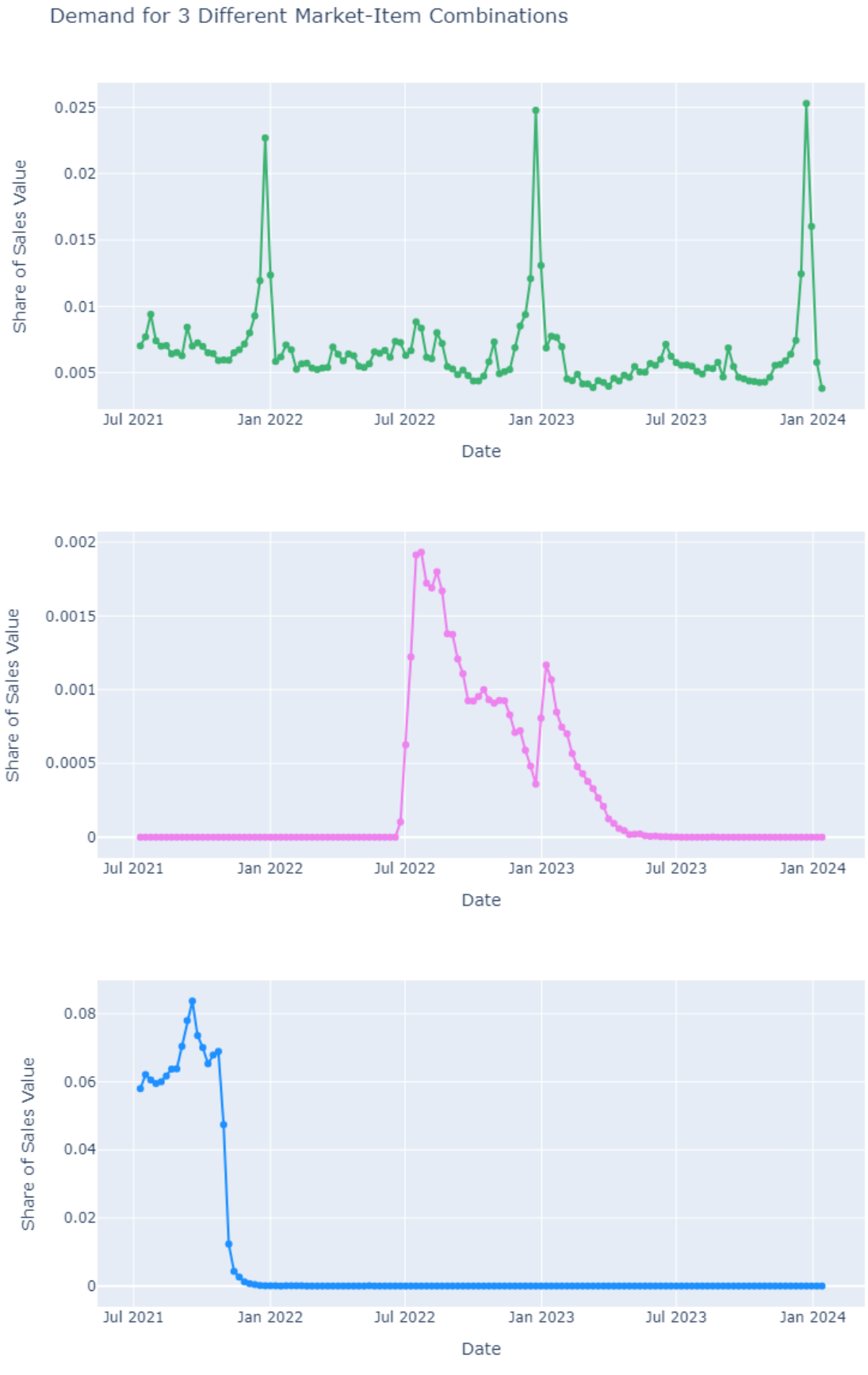


Figure A.16.
Salad Dressing Sales Value

Figure A.17.
3 Random SKU Sales to Illustrate Volatility



A.6 Feature Space

Table A.3.
Marketing Mix

Feature Name	Description
Product-Descriptive Variables	
Markets	This variable denotes the retailer-chain name. There are 2 unique markets, and they are used as one-hot encoded features.
Company	This variable denotes the manufacturer that produced the SKU. There are 125 unique companies, and they are used as one-hot encoded features.
Segment	This variable denotes the subcategory to which the SKU belongs. There are 6 unique subcategories, and they are used as one-hot encoded features.
Brand	This variable denotes the brand to which the SKU belongs. There are 97 unique brands, and they are used as one-hot encoded features.
Type	This variable denotes the type of product. There are 5 unique types, and they are used as one-hot encoded features.
Pack Type	This variable denotes the type of packaging. There are 14 unique pack types, and they are used as one-hot encoded features.
Variant	This variable denotes the variant of the SKU. There are 134 unique variants, and they are used as one-hot encoded features.
Unit Weight	This variable denotes the unit weight of the SKU. It is used as a numerical variable on the scale of the kilograms.
Item	This variable denotes the full description of the specific item SKU. There are 583 unique items.
Price Variables	
To capture the trend and prevent data leakage, the following variables are created with a range of lags from 12 weeks to 18 weeks (7 lags in total)	
Price per Sales (Unit)	The average price per unit sold.
Price per Sales (KGS)	The average price per kg sold
Incremental Price per Sales (Unit) Any Promo	The additional price per unit sold during promotional periods against the baseline.
Incremental Price per Sales (KG) Any Promo	The additional price per kg sold during promotional periods against the baseline.

Table A.3. Marketing Mix (continued)

Feature Name	Description
Any Promo Base (Unit) Price	The baseline price per unit during promotional periods.
Any Promo Base (KGS) Price	The baseline price per kgs during promotional periods.
Weighted Price per Sales (Unit)	The average price per unit sold, weighted by sales value.
Weighted Incremental Price per Sales Any Promo (Unit)	The average baseline price per unit sold, weighted by sales value.
Weighted Any Promo Base (Unit) Price	The base price per unit sold during promotional periods, weighted by sales value.
Promotion Variables	
To capture the trend and prevent data leakage, the following variables are created with a range of lags from 12 weeks to 18 weeks (7 lags in total)	
Promotional Type Indication	Dummy variables indicating whether specific promotional activities were applied. This includes:Any Promo, Feature and Display, Feature without Display, Display without Feature, Multibuy, Temporary Price Reduction.
Sales Value from Promo Type	The total sales value generated from the aforementioned promotion types.
Incremental Sales Value from Promo Type	The additional sales value generated from various promotion types over the baseline sales.
Sales Value Any Promo	The total sales value generated from any promotional activity.
Sales Value No Promo	The total sales value generated without any promotional activity.
Incremental Sales Value Any Promo	The additional sales value generated from any promotional activity over the baseline sales.
Incremental Sales Value No Promo	The additional sales value generated without any promotional activity over the baseline sales.
Distribution Variables	
To capture the trend and prevent data leakage, the following variables are created with a range of lags from 12 weeks to 18 weeks (7 lags in total)	
Total Weighted Distribution Points (TDP)	The total weighted distribution points, indicating the overall distribution coverage of the product. TDP is a measure that evaluates the stores stores a particular item is present and the sales of those stores to identify the reach of the item.

Table A.3. Marketing Mix (continued)

Feature Name	Description
Number of Items	The total number of items/SKUs that were available within the retailer chain.
Value / Store	The average sales value per store.
TDP Weighted Distribution per Promo Type	The weighted distribution points from the aforementioned promotional types.

Return to the reading spot by clicking the number, Section 4.1.3

Table A.4.
Cross-Product Effects

Feature Name	Description
The following cross-product effects only apply within the same category, retailer, and period. Otherwise, they are zeros. To capture the trend and prevent data leakage, the following variables are created with a range of lags of 12, 15, 18 weeks (3 lags in total)	
ITEM_TDP	The effect of Total Weighted Distribution Points (TDP) of the top 5 items within each of the 6 segments. This metric captures the distribution coverage interaction of these top items.
ITEM_Weighted_Price	The effect of Weighted Price per Sales (Unit) of the top 5 items by sales value within each segment.
ITEM_Sales_Promo_Type	The Sales Value generated from each existing promotional type of the top 5 items by sales value within each segment. This includes: • ITEM_Feat_Disp (Feature and Display) • ITEM_Feat_no_Disp (Feature without Display) • ITEM_Disp_no_Feat (Display without Feature) • ITEM_Total_Multibuy (Buy one get one free) • ITEM_TPR_Only (Temporary Price Reduction)
Promo_Intensity per Promo_Type	Number of products on promotion within the same category and market, capturing the promotional intensity of the category for each of the aforementioned promotional types and overall promotion.
Avg_Weighted_Brand_Price	The average weighted price of a brand computed to capture brand-specific pricing effects for all brands.

Table A.4. Overview of Cross-Product Effects (continued)

Feature Name	Description
Avg_Similarity	This variable is constructed based on text-embeddings, from which a product pairwise cosine similarity matrix is extracted. This metric indicates the average semantic similarity to the 10 most similar items for each item.
Avg_Share_Sales	The average share of sales across the 10 most similar items for each item, calculated from the similarity matrix. This indicates the share of sales value that the 10 most similar items for that particular item have. Its lag of 12 weeks is included.

Return to the reading spot by clicking the number, Section 4.1.3

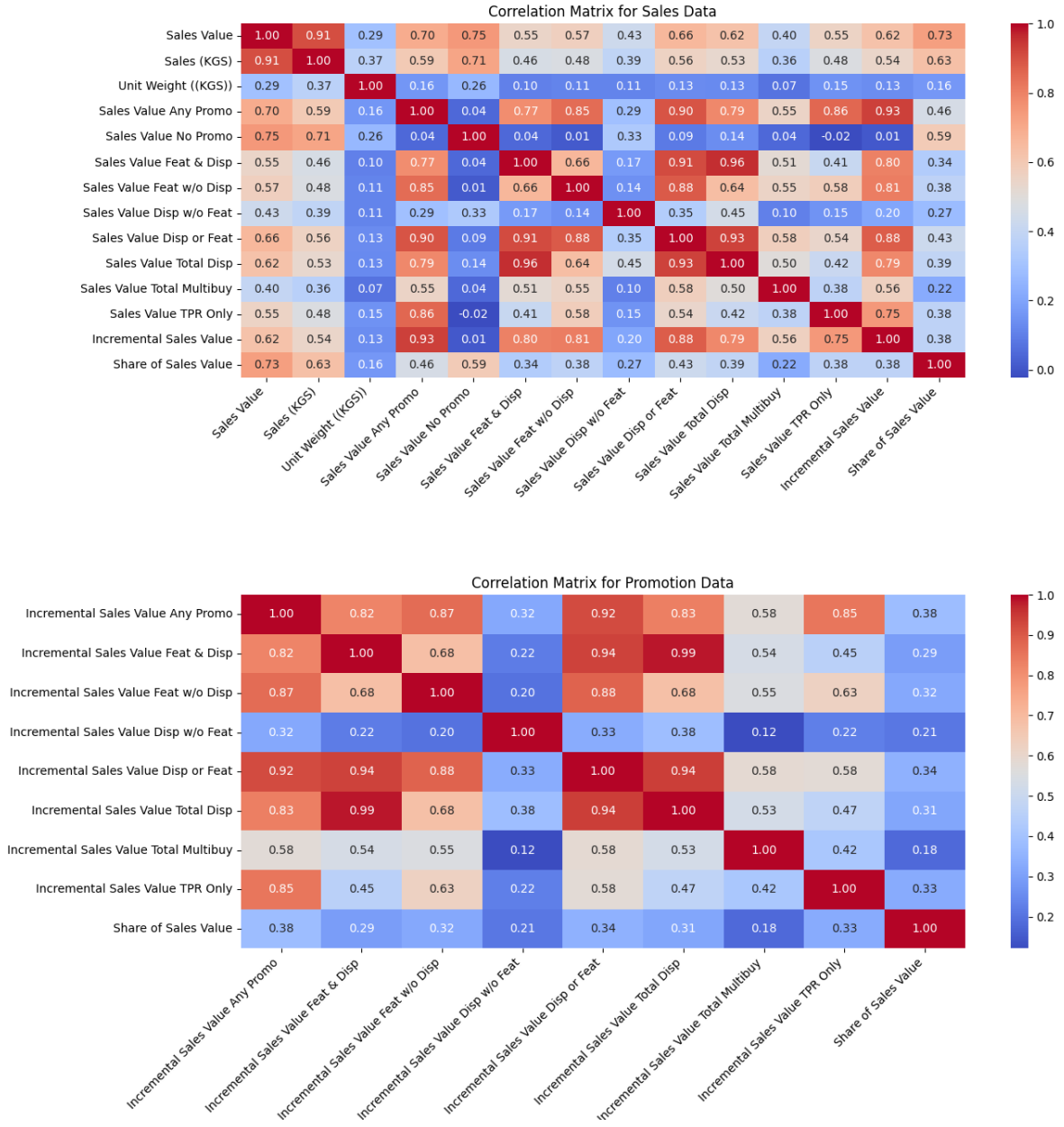
Table A.5.
Seasonality and Trend

Feature Name	Description
UK Holidays	These variables are dummy indicators for specific holidays and significant public events within the United Kingdom. The holidays included are: • Platinum Jubilee of Elizabeth II • Spring Bank Holiday • Christmas Day • Coronation of Charles III • Boxing Day • Good Friday • State Funeral of Queen Elizabeth II • May Day • New Year's Day • Start of Academic Year Each holiday dummy is active including a span of 7 days before the actual holiday date to account for the potential effects prior to the concrete date.
Calendar Variables	These variables capture temporal information, including month, quarter, and year. They are used as dummy variables.
Weather Variables	These variables include weather-related metrics such as average temperature (tavg) in Celsius and precipitation (prcp) in mm. These metrics were extracted from open-source weather database Meteostat from the London's weather centre.
Consistency	Dummy variable that indicates whether the SKU has periods where it does not sell.
Data Age	Numerical variable tht indicates how old the SKU is, based on the date it was introduced.

Return to the reading spot by clicking the number, Section 4.1.3

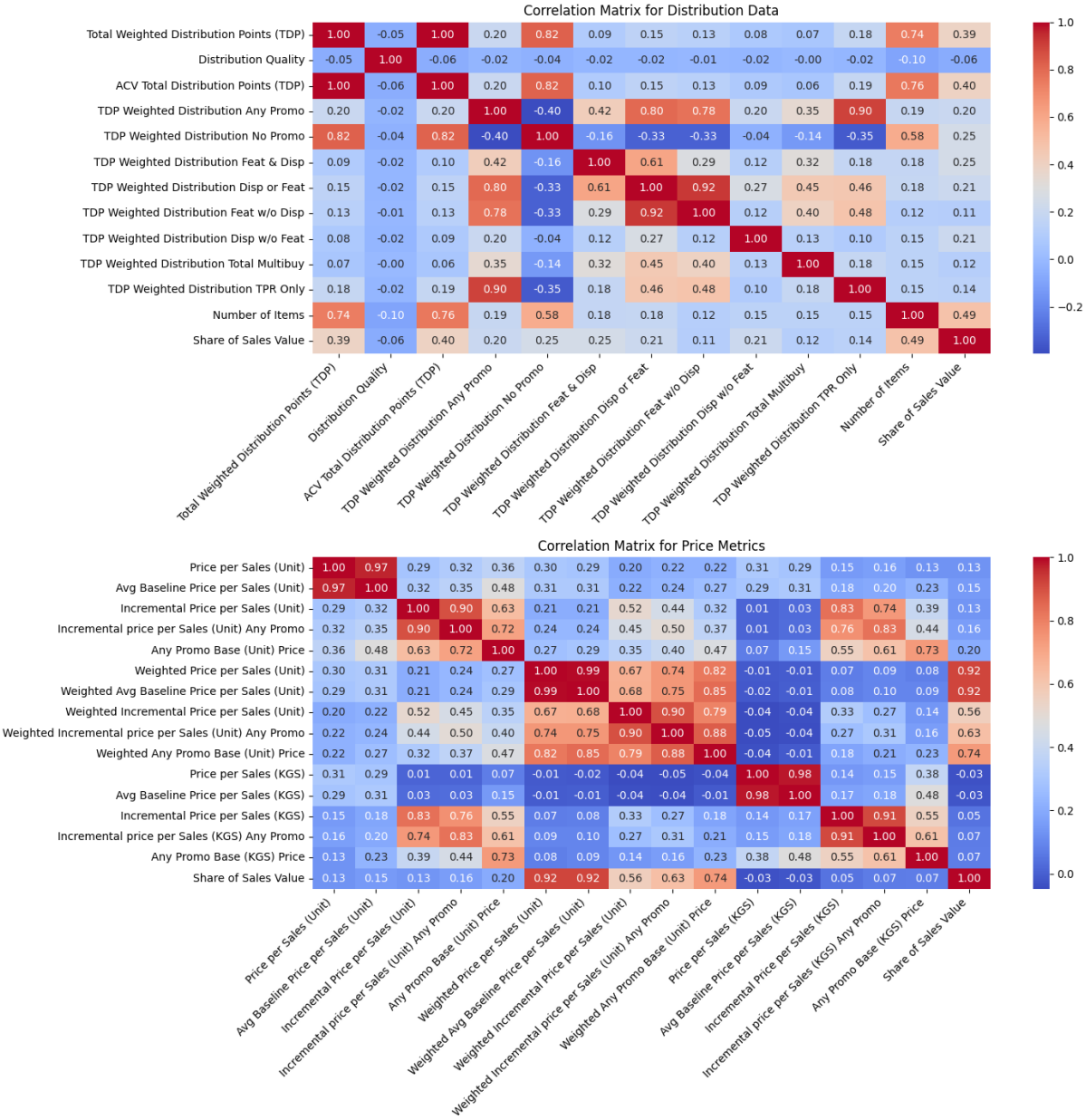
A.7 Multicollinearity Inspection

Figure A.18.
1st Multicollinearity Inspection



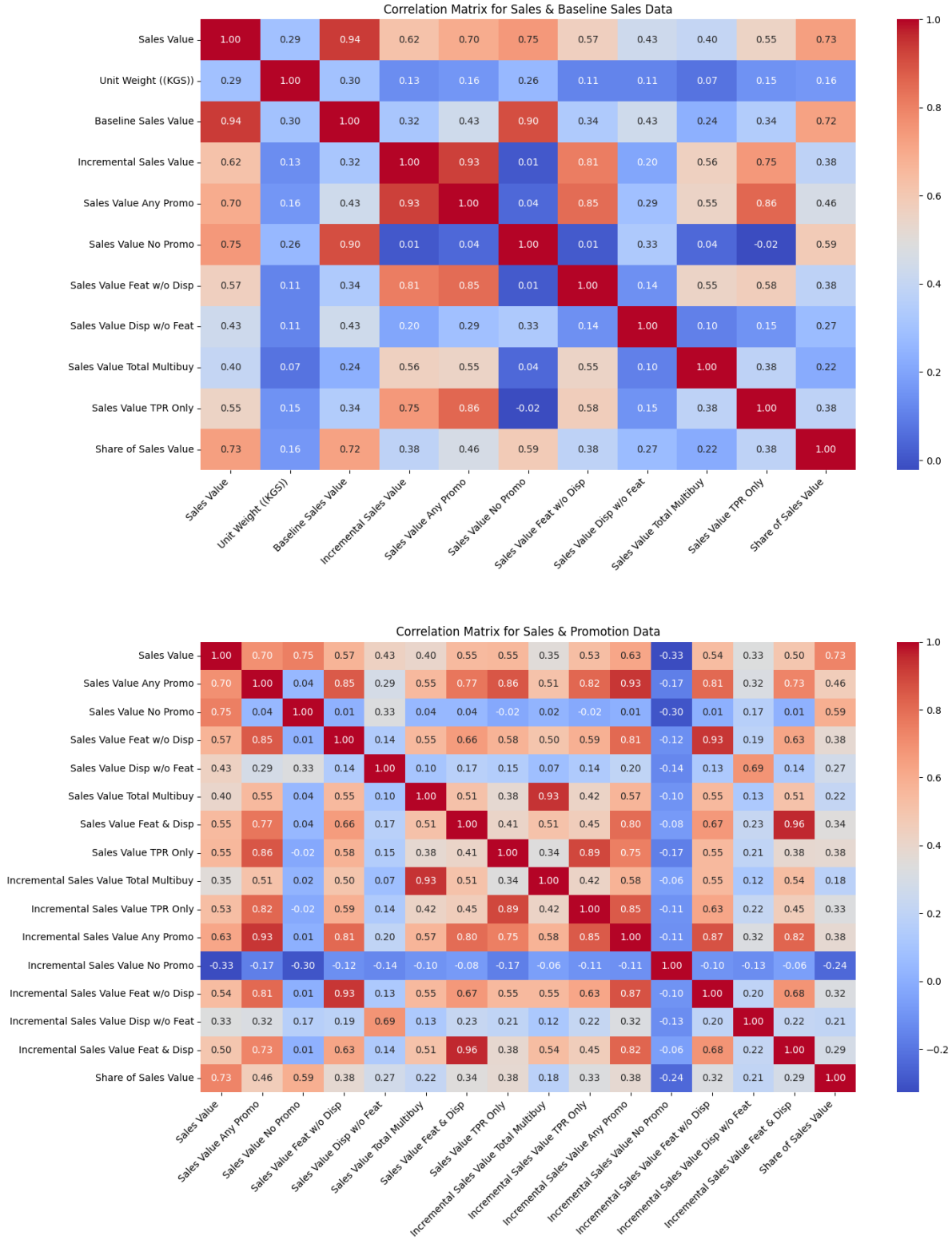
Return to the reading spot by clicking the number, Section 4.3.1

Figure A.19.
2nd Multicollinearity Inspection



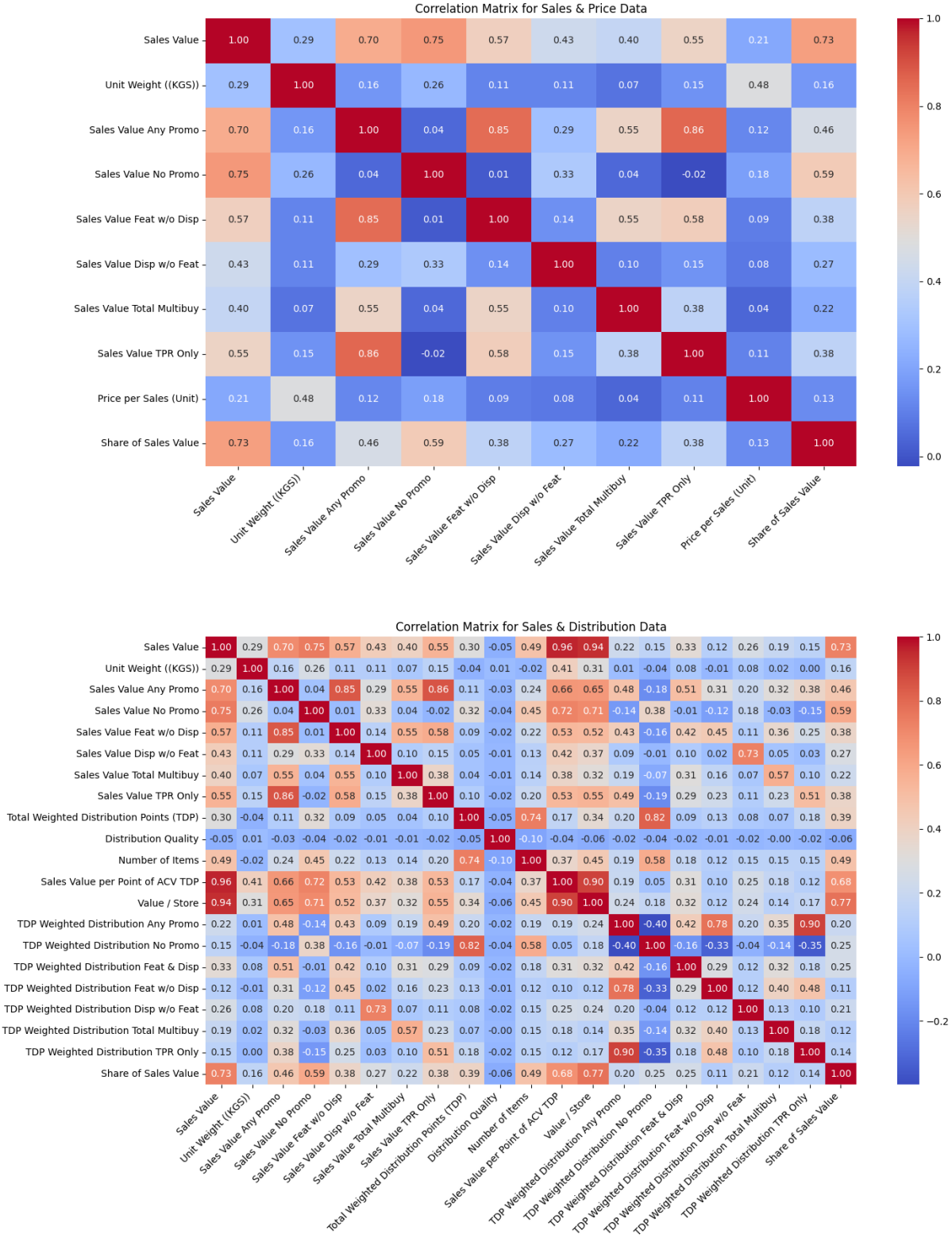
Return to the reading spot by clicking the number, Section 4.3.1

Figure A.20.
3rd Multicollinearity Inspection



Return to the reading spot by clicking the number, Section 4.3.1

Figure A.21.
4th Multicollinearity Inspection



Return to the reading spot by clicking the number, Section 4.3.1

B. Results Extensions

B.1 Grid Search Cross Validation Results

Table B.1. Grid Search Results for RF models and its variations

Hyperparameters		RF		RF w Em		TRF		TRF w Em	
max_depth	n_estimators	RMSE	SD	RMSE	SD	RMSE	SD	RMSE	SD
5	100	1.33	0.0589	1.31	0.0663	1.33	0.0552	1.33	0.0542
5	200	1.33	0.0576	1.31	0.0633	1.33	0.0554	1.33	0.0547
7	100	1.31	0.0606	1.29	0.0721	1.32	0.0610	1.32	0.0599
7	200	1.31	0.0591	1.29	0.0717	1.32	0.0602	1.32	0.0597
9	100	1.30	0.0624	1.28	0.0755	1.31	0.0644	1.32	0.0646
9	200	1.30	0.0634	1.28	0.0732	1.31	0.0645	1.32	0.0652
11	100	1.30	0.0651	1.27	0.0725	1.31	0.0663	1.31	0.0672
11	200	1.30	0.0640	1.27	0.0733	1.31	0.0653	1.31	0.0668

Table B.2. Fit Time Results Across different RF Models (measured in seconds)

Hyperparameters		RF		RF w Em		TRF		TRF w Em	
max_depth	n_estimators	Mean	SD	Mean	SD	Mean	SD	Mean	SD
5	100	1179.35	8.60	1625.94	4.79	244.21	2.69	220.51	1.82
5	200	2285.63	27.76	3235.53	58.46	509.25	10.16	418.59	15.95
7	100	1800.85	6.00	2448.27	19.72	385.25	3.33	293.98	45.36
7	200	3504.37	13.94	4705.48	51.56	711.74	6.79	392.79	17.69
9	100	2405.61	12.89	3191.28	37.23	475.97	4.60	263.88	3.23
9	200	4739.36	11.74	6938.94	17.16	938.29	5.97	537.53	0.53
11	100	2907.48	215.86	4195.02	452.32	565.85	40.74	323.21	22.76
11	200	4378.36	8.62	5859.98	30.50	864.54	9.39	504.87	0.86

Table B.3. Grid Search Results for Gradient Boosting Regressor

Hyperparameters	Gradient Boosting Regressor	
	RMSE	**SD**
learning_rate = 0.01, max_depth = 5, max_iter = 100	1.81	0.0800
learning_rate = 0.01, max_depth = 5, max_iter = 200	1.38	0.0798
learning_rate = 0.01, max_depth = 7, max_iter = 100	1.80	0.0878
learning_rate = 0.01, max_depth = 7, max_iter = 200	1.38	0.0821
learning_rate = 0.01, max_depth = 9, max_iter = 100	1.80	0.0874
learning_rate = 0.01, max_depth = 9, max_iter = 200	1.38	0.0833
learning_rate = 0.01, max_depth = 11, max_iter = 100	1.80	0.0885
learning_rate = 0.01, max_depth = 11, max_iter = 200	1.38	0.0843
learning_rate = 0.05, max_depth = 5, max_iter = 100	1.32	0.0721
learning_rate = 0.05, max_depth = 5, max_iter = 200	1.32	0.0682
learning_rate = 0.05, max_depth = 7, max_iter = 100	1.31	0.0629
learning_rate = 0.05, max_depth = 7, max_iter = 200	1.31	0.0546
learning_rate = 0.05, max_depth = 9, max_iter = 100	1.31	0.0757
learning_rate = 0.05, max_depth = 9, max_iter = 200	1.30	0.0674
learning_rate = 0.05, max_depth = 11, max_iter = 100	1.31	0.0692
learning_rate = 0.05, max_depth = 11, max_iter = 200	1.30	0.0620
learning_rate = 0.1, max_depth = 5, max_iter = 100	1.33	0.0588
learning_rate = 0.1, max_depth = 5, max_iter = 200	1.33	0.0563
learning_rate = 0.1, max_depth = 7, max_iter = 100	1.32	0.0736
learning_rate = 0.1, max_depth = 7, max_iter = 200	1.31	0.0717
learning_rate = 0.1, max_depth = 9, max_iter = 100	1.32	0.0702
learning_rate = 0.1, max_depth = 9, max_iter = 200	1.30	0.052
learning_rate = 0.1, max_depth = 11, max_iter = 100	1.31	0.0643
learning_rate = 0.1, max_depth = 11, max_iter = 200	1.31	0.0639

B.1.1 Models Selected for Testing

Considering that all variations of the applied RF have different numbers of features, they were chosen individually. This is due to the fact that to capture, for example, the effect of embeddings, they may better fit with a higher max_depth.

- **Random Forest (RF)**
  - **Selected Parameters**: `max_depth = 7`, `n_estimators = 200`
  - **Reason**: Marginal improvement in RMSE beyond these values.
- **Random Forest with Embeddings (RF w Em)**
  - **Selected Parameters**: `max_depth = 11`, `n_estimators = 200`
  - **Reason**: Consistent improvement in RMSE with increasing complexity, and embeddings might require higher level of complexity
- **Targeted Random Forest (TRF)**
  - **Selected Parameters**: `max_depth = 9`, `n_estimators = 200`
  - **Reason**: Optimal performance reached with no substantial gains beyond these values.
- **Targeted Random Forest with Embeddings (TRF w Em)**
  - **Selected Parameters**: `max_depth = 9`, `n_estimators = 200`
  - **Reason**: Optimal performance reached with no substantial gains beyond these values.
- **Gradient Boosting Regressors (GBR)**
  - **Selected Parameters**: `learning_rate` = 0.05, `max_depth = 9`, `learning_rate` = 200
  - **Reason**: Very similar performance with learning_rate 0.1, but a lower learning rate may more slowly converge, and this slower adaption may be beneficial for stability of the model, thus chosen.

B.2 Question 1 Supporting Material

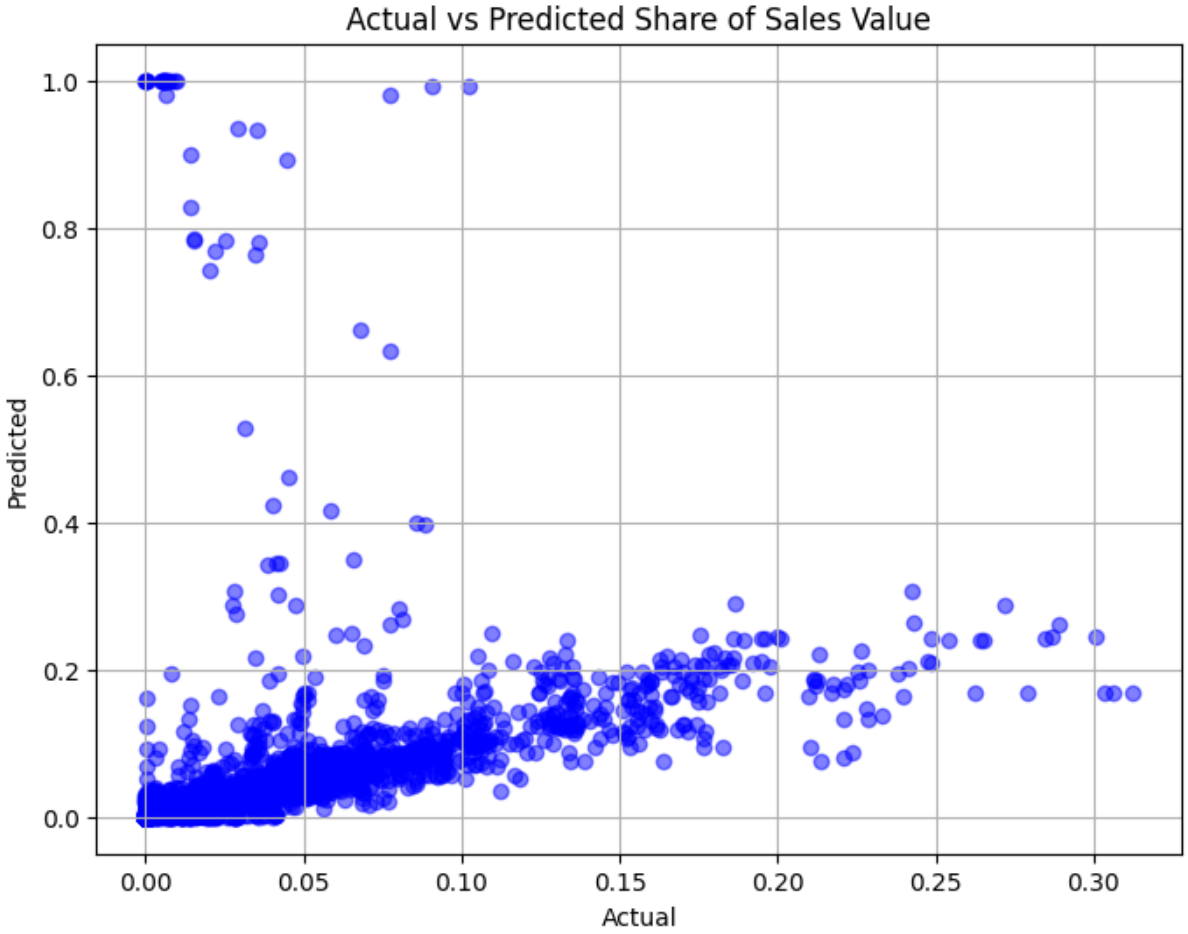
Table B.4. Model's Performance on the Hold-out-of-Sample Test Set (with Log Transformed Prophet)

Subcategory	**Metric**	**Statistical Method**	**Machine Learning Methods**	
		Prophet	**Random Forest**	**Gradient Boosting**
Ketchup	RMSE	2.45	2.08	2.06
	MAE	1.16	0.91	0.90
	sMAPE	54.12	49.55	49.84
Mayonnaise	RMSE	9.65	1.16	1.11
	MAE	2.08	0.46	0.46
	sMAPE	46.03	39.34	40.77
Meat & Fish Sauces	RMSE	7.27	0.40	0.36
	MAE	0.76	0.22	0.20
	sMAPE	44.89	50.49	51.33
Mustard	RMSE	9.32	1.94	1.70
	MAE	2.59	1.14	1.04
	sMAPE	39.83	35.98	35.88
Other Dressings	RMSE	1.74	1.09	1.03
	MAE	1.03	0.69	0.67
	sMAPE	36.25	32.39	33.75
Salad Dressings	RMSE	1.56	1.06	1.01
	MAE	0.83	0.59	0.59
	sMAPE	40.44	35.27	36.31
Total	RMSE	7.03	1.08	1.03
	MAE	1.19	0.46	0.44
	sMAPE	44.69	44.60	45.51

***NOTE: RMSE and MAE are both on a scale of 0-100% for better interpretation purposes.**

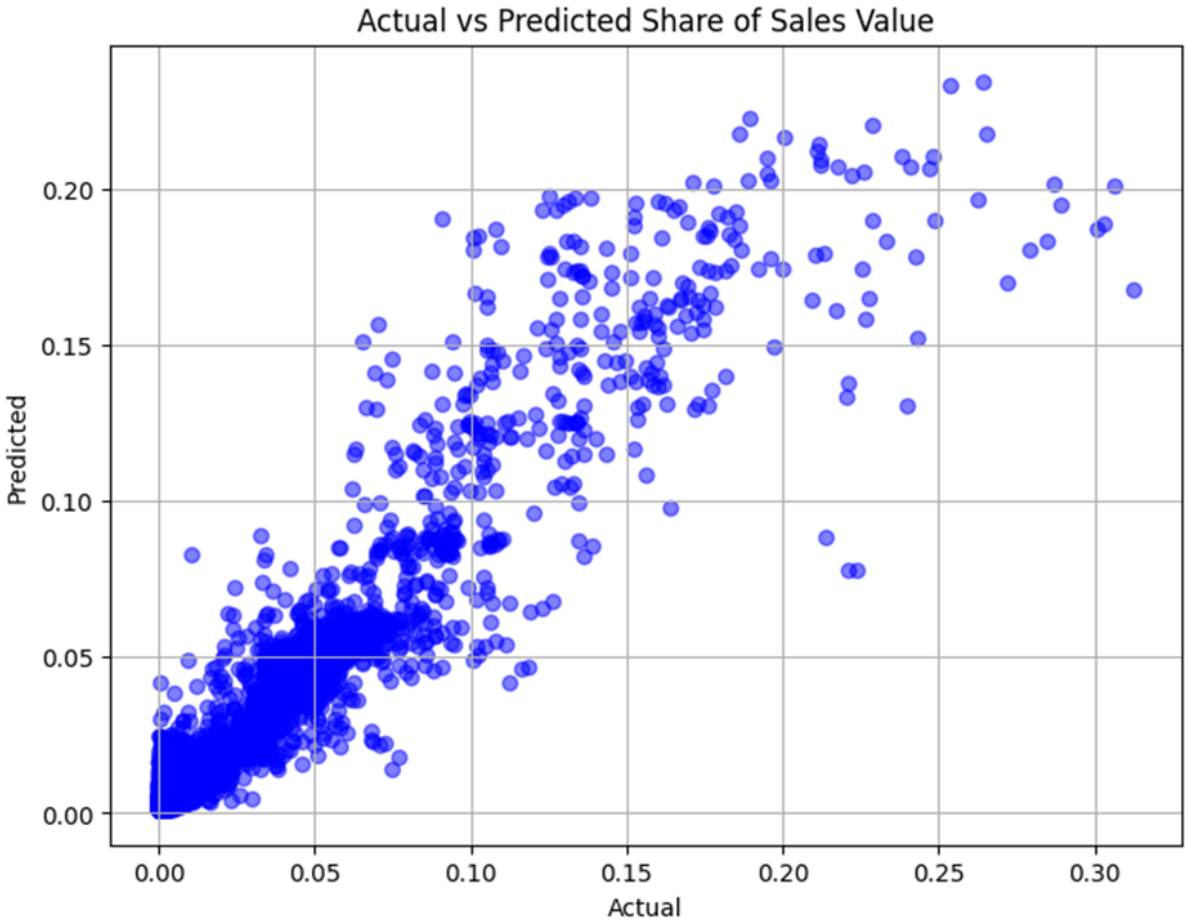
Return to the reading spot by clicking the number, Section 5.1

Figure B.1.
Prophet Actuals vs Predictions



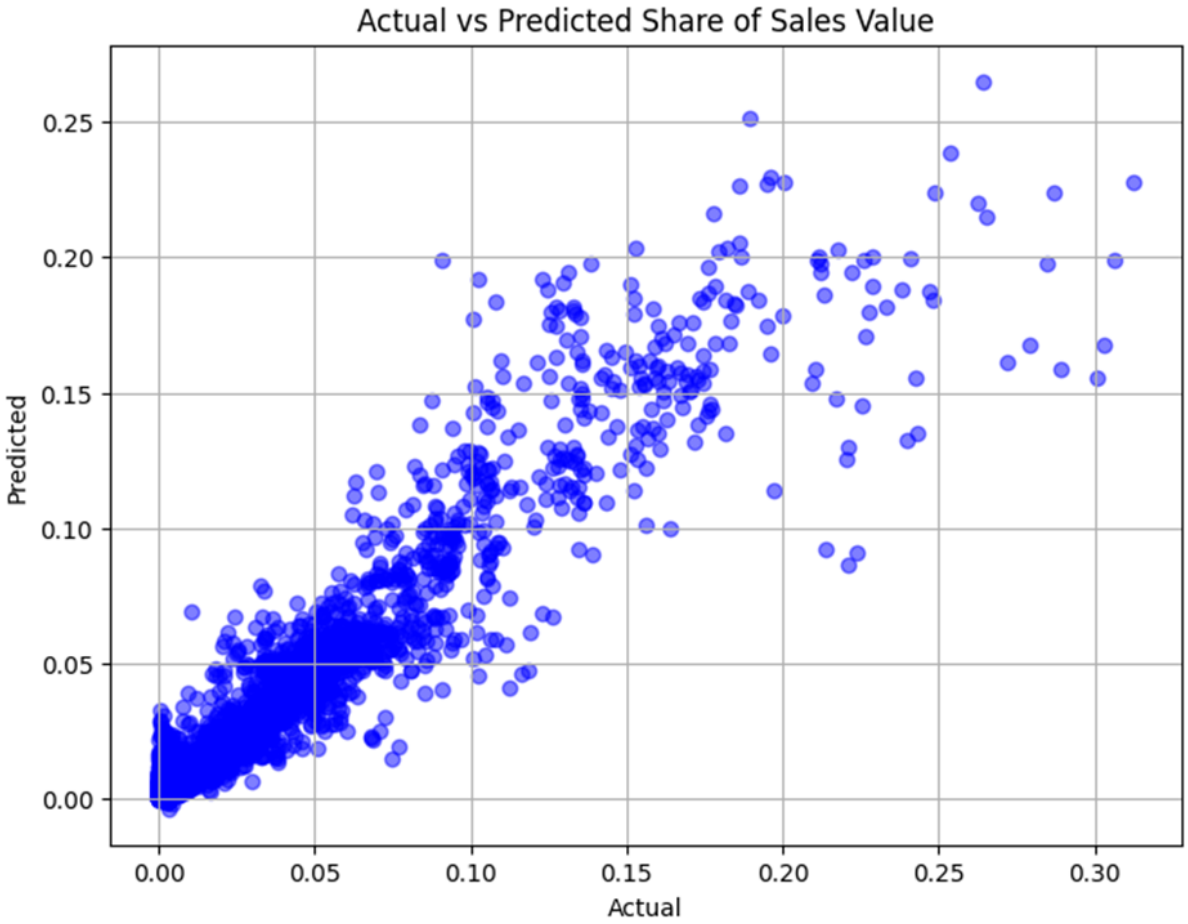
Return to the reading spot by clicking the number, Section 5.1

Figure B.2.
RF Actuals vs Predictions



Return to the reading spot by clicking the number, Section 5.1

Figure B.3.
GBR Actuals vs Predictions



Return to the reading spot by clicking the number, Section 5.1

Table B.5.
ANOVA and Post-Hoc Comparisons using Tukey HSD

Model	Mean RMSE	Std. Dev.	Std. Error
Gradient Boosting (GBR)	1.05	0.04	0.0013
Prophet	6.99	0.55	0.0175
Random Forest (RF)	1.10	0.04	0.0013

Source	F Value	Num DF	Den DF	p-value
Model	114097.78	2	1998	0.000

Comparison	Mean Difference	Std. Error	p-adj	Lower CI	Upper CI
GBR vs. Prophet	5.94	3.03	0.000***	5.91	5.98
GBR vs. RF	0.05	0.03	0.002**	0.02	0.08
Prophet vs. RF	-5.89	3.01	0.000***	-5.93	-5.86

Note: The significance level is ($p < .05$) *Return to the reading spot by clicking the number, Section* 5.1

B.3 Question 2 Supporting Visualizations

Table B.6. Model's Performance on the Hold-out-of-Sample Test Set

Subcategory	Metric	RF Baseline	RF with Em	TRF	TRF with Em
Mayonnaise	RMSE	1.16	1.09	1.11	1.11
	MAE	0.46	0.46	0.45	0.45
	sMAPE	39.34%	39.53%	38.84%	38.90%
Meat & Fish Sauces	RMSE	0.40	0.35	0.40	0.40
	MAE	0.22	0.20	0.22	0.22
	sMAPE	50.49%	48.53%	48.97%	49.03%
Salad Dressings	RMSE	1.06	1.08	1.10	1.10
	MAE	0.59	0.60	0.60	0.60
	sMAPE	35.27%	35.95%	35.56%	35.59%
Ketchup	RMSE	2.08	2.07	2.13	2.12
	MAE	0.91	0.91	0.90	0.91
	sMAPE	49.55%	49.50%	48.35%	48.27%
Mustard	RMSE	1.94	2.09	1.67	1.67
	MAE	1.14	1.24	1.04	1.04
	sMAPE	35.98%	36.67%	35.49%	35.48%
Other Dressings	RMSE	1.09	1.09	1.11	1.11
	MAE	0.69	0.67	0.70	0.71
	sMAPE	32.39%	32.38%	32.57%	32.62%
Total	RMSE	1.08	1.07	**1.06**	1.06
	MAE	0.46	0.45	**0.45**	0.45
	sMAPE	44.60%	43.80%	**43.67%**	43.71%

Return to the reading spot by clicking the number, Section 5.2.1

Figure B.4.
Baseline Random Forest Feature Importance

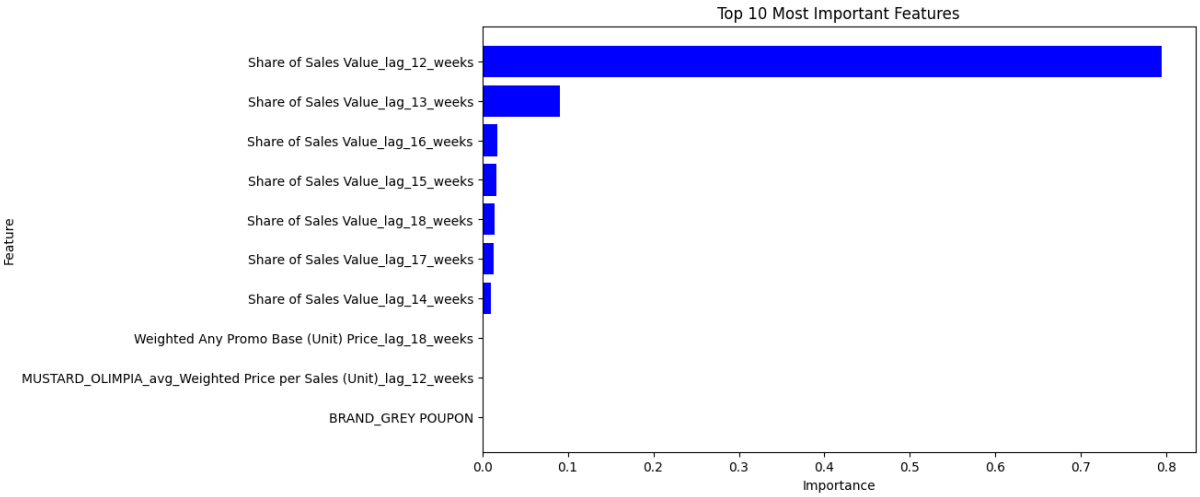
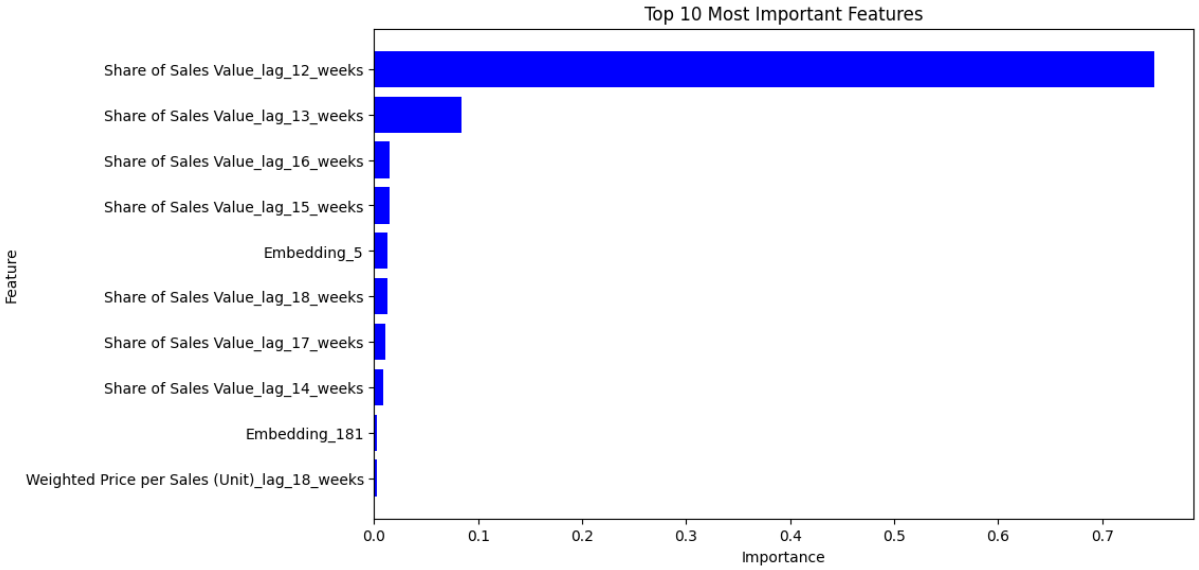


Figure B.5.
Random Forest with Embeddings Feature Importance



Return to the reading spot by clicking the number, Section 5.2.1

Figure B.6.
Embedding_5 Items Descending Order

ITEM	Embedding_5
UNILEVER / UNILEVER BESTFOODS HELLMANNS LIGHT MAYONNAISE MAYONNAISE WITH LOW CALORIE CLAIM JAR GLASS 600G X1	-0.114440918
UNILEVER / UNILEVER BESTFOODS HELLMANNS LIGHTER THAN LIGHT MAYONNAISE MAYONNAISE WITH LOW CALORIE CLAIM JAR GLASS 400G X1	-0.114074707
UNILEVER / UNILEVER BESTFOODS HELLMANNS LIGHT MAYONNAISE MAYONNAISE WITH LOW CALORIE CLAIM JAR GLASS 400G X1	-0.1137084961
UNILEVER / UNILEVER BESTFOODS HELLMANNS LIGHT MAYONNAISE MAYONNAISE WITH LOW CALORIE CLAIM JAR GLASS 800G X1	-0.1137084961
KRAFT HEINZ COMPANY HEINZ TOMATO KETCHUP 50% LESS SUGAR & SALT KETCHUP STYLE SAUCE BOTTLE PLASTIC 500ML X1	-0.0991210938
KRAFT HEINZ COMPANY HEINZ MAYORACHA MAYO SRIRACHA SAUCE MAYONNAISE BOTTLE PLASTIC 400ML X1	-0.0982666016
KRAFT HEINZ COMPANY HEINZ SERIOUSLY GOOD MAYONNAISE MAYONNAISE JAR GLASS 710ML X1	-0.095703125
KRAFT HEINZ COMPANY HEINZ TOMATO KETCHUP 50% LESS SUGAR & SALT KETCHUP STYLE SAUCE BOTTLE PLASTIC 235G X1	-0.0956420898
KRAFT HEINZ COMPANY HEINZ ORGANIC TOMATO KETCHUP KETCHUP STYLE SAUCE WITH ORGANIC CLAIM BOTTLE PLASTIC 580G X1	-0.0919189453
UNILEVER / UNILEVER BESTFOODS HELLMANNS LIGHT MAYONNAISE MAYONNAISE WITH LOW CALORIE CLAIM BOTTLE PLASTIC 250ML X1	-0.0908203125
KRAFT HEINZ COMPANY HEINZ SERIOUSLY GOOD LIGHT MAYONNAISE MAYONNAISE BOTTLE PLASTIC 220ML X1	-0.0891113281
UNILEVER / UNILEVER BESTFOODS HELLMANNS LIGHT MAYONNAISE MAYONNAISE WITH LOW CALORIE CLAIM BOTTLE PLASTIC 750ML X1	-0.0885620117
KRAFT HEINZ COMPANY HEINZ SERIOUSLY GOOD LIGHT MAYONNAISE MAYONNAISE BOTTLE PLASTIC 800ML X1	-0.0874633789
KRAFT HEINZ COMPANY HEINZ TOMATO KETCHUP 50% LESS SUGAR & SALT KETCHUP STYLE SAUCE BOTTLE PLASTIC 435G X1	-0.0860595703
UNILEVER / UNILEVER BESTFOODS HELLMANNS ORGANIC MAYONNAISE MAYONNAISE JAR GLASS 270G X1	-0.0857543945

Return to the reading spot by clicking the number, Section 5.2.1

Table B.7.
Total Time for Validation and Grid Search for Different Models

Model	Total Time (minutes)	Total Time (seconds)
Random Forest (RF)	286	17160
Random Forest with Embeddings (RF w Em)	393	23580
Targeted Random Forest (TRF)	57	3420
Targeted Random Forest with Embeddings (TRF w Em)	36	2160

Return to the reading spot by clicking the number, Section 5.2.1

Figure B.7.
LASSO without Embeddings Feature Importance

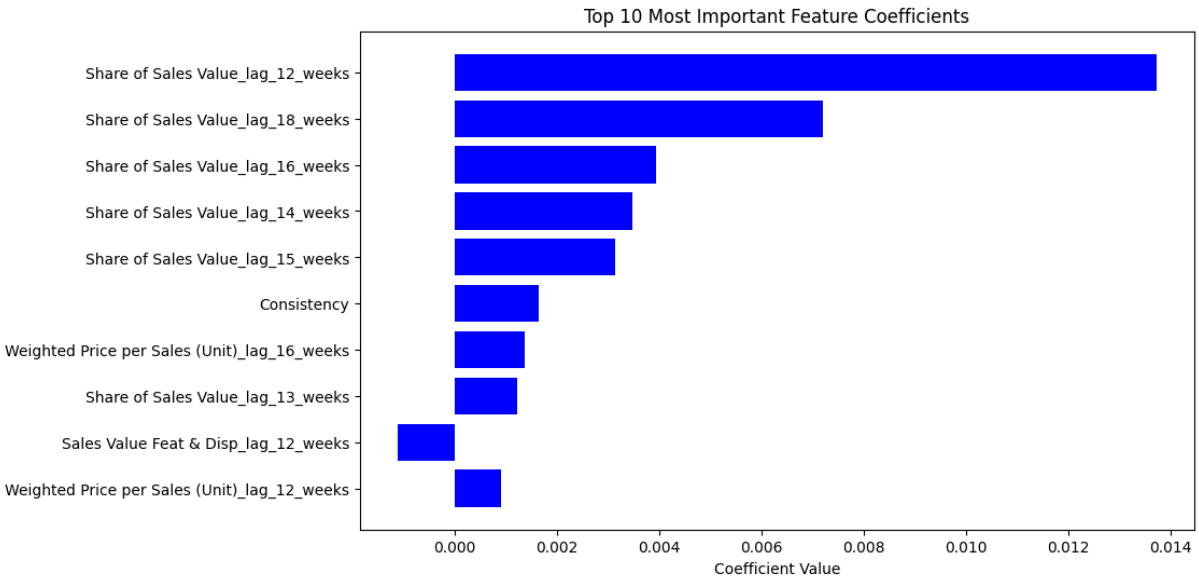
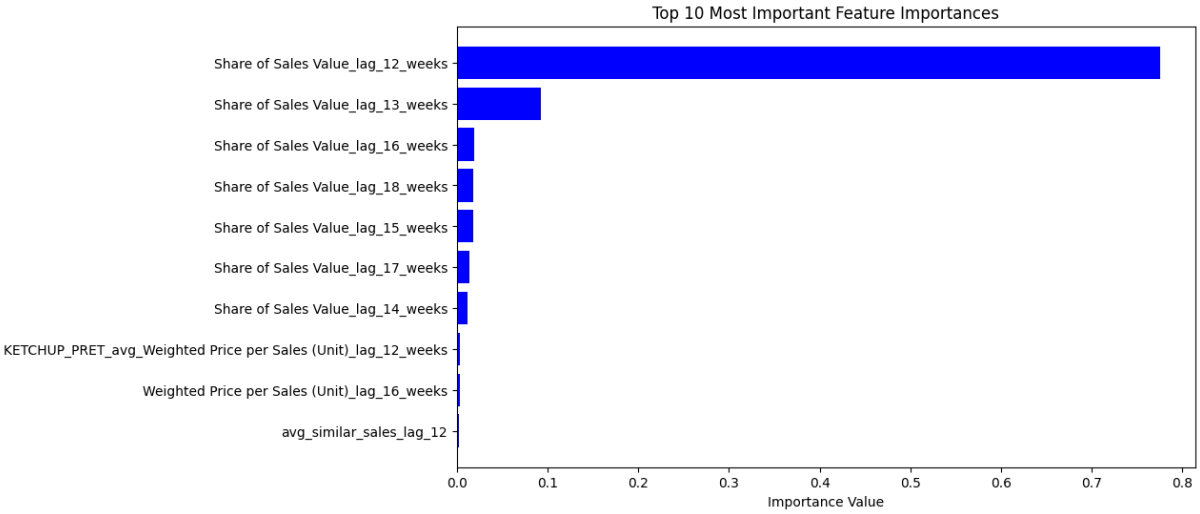
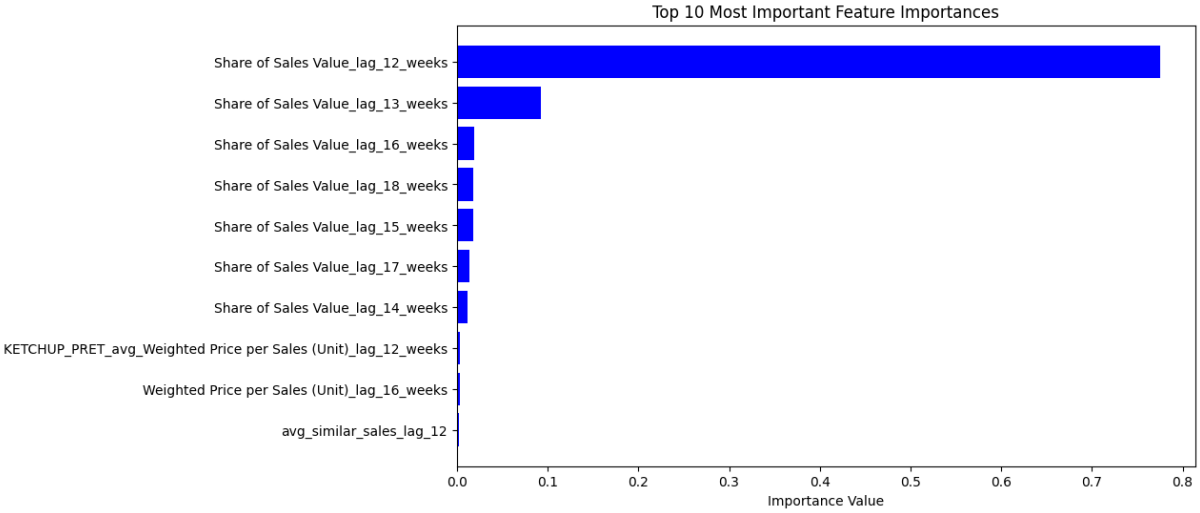


Figure B.8.
Targeted Random Forest Feature Importances



Return to the reading spot by clicking the number, Section 5.2.1

Figure B.9.
Targeted Random Forest with Embeddings Important Features



Return to the reading spot by clicking the number, Section 5.2.1

Table B.8.
ANOVA and Post-Hoc Comparisons using Tukey HSD

Model	Mean RMSE	Std. Dev.	Std. Error
RF_EM_RMSE	1.07	0.04	0.0012
RF_RMSE	1.08	0.04	0.0013
TRF_EM_RMSE	1.06	0.04	0.0013
TRF_RMSE	1.06	0.04	0.0013

Source	F Value	Num DF	Den DF	p-value
Model	654.59	3	2997	0.000

Comparison	Mean Difference	Std. Error	p-adj	Lower CI	Upper CI
RF_EM_RMSE vs. RF_RMSE	0.0060	0.0031	0.004	0.0014	0.0106
RF_EM_RMSE vs. TRF_EM_RMSE	-0.0129	0.0066	0.000	-0.0175	-0.0083
RF_EM_RMSE vs. TRF_RMSE	-0.0120	0.0061	0.000	-0.0167	-0.0074
RF_RMSE vs. TRF_EM_RMSE	-0.0189	0.0096	0.000	-0.0235	-0.0143
RF_RMSE vs. TRF_RMSE	-0.0180	0.0092	0.000	-0.0227	-0.0134
TRF_EM_RMSE vs. TRF_RMSE	0.0009	0.0005	0.964	-0.0038	0.0055

Note: The significance level is ($p < .05$)
Return to the reading spot by clicking the number, Section 5.2.2

B.4 Question 3 Supporting Visualizations

Figure B.10.
Feature Importances for all Subcategories

Figure B.11. Mayonnaise

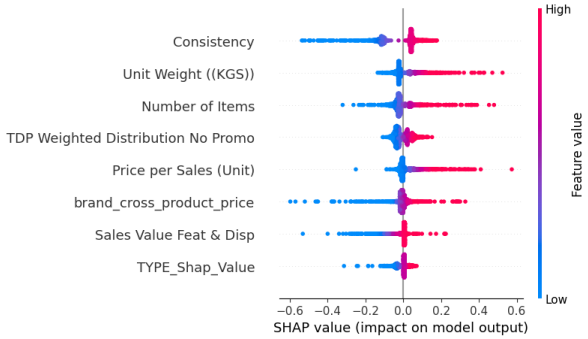


Figure B.12. Meat & Fish Sauces

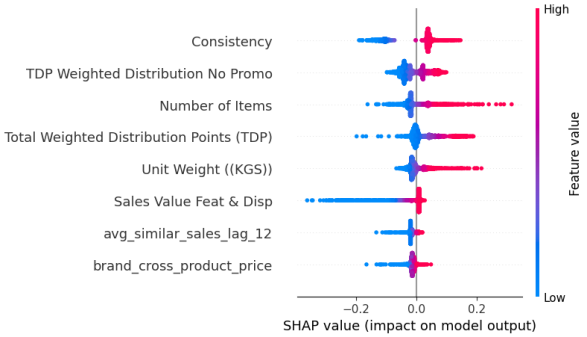


Figure B.13. Salad Dressings

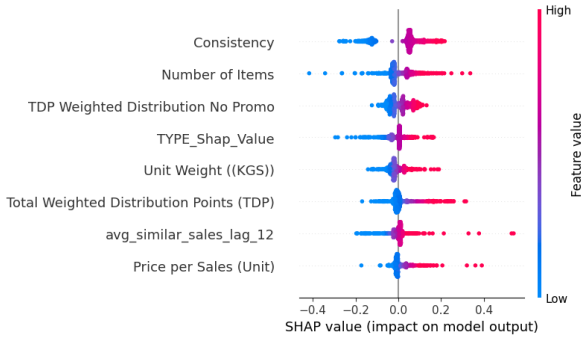


Figure B.14. Ketchup

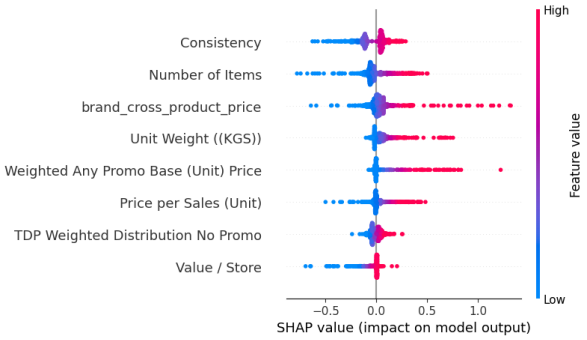


Figure B.15. Mustard

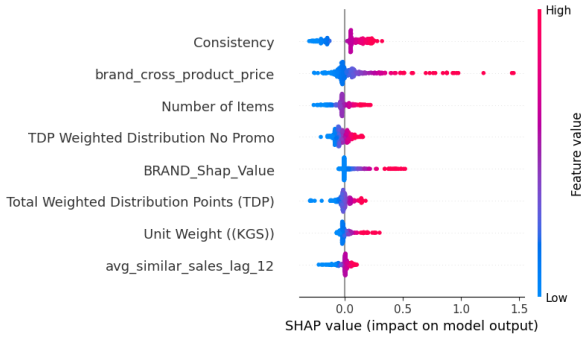
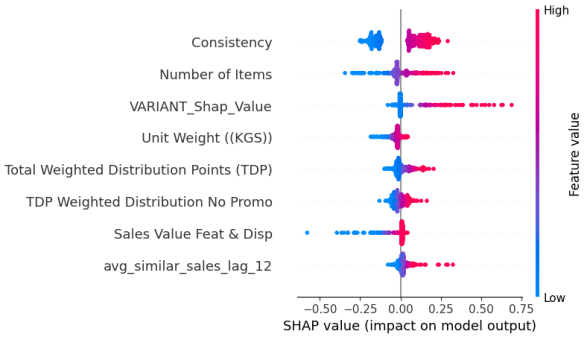


Figure B.16. Other Dressings



Return to the reading spot by clicking the number, Section 5.3.1

B.4.1 Local Feature Importances

Return to the reading spot by clicking the number, Section 5.3.2

Figure B.17.
Waterfall Plot of Feature Importance for the Highest Share SKU within Ketchup Subcategory

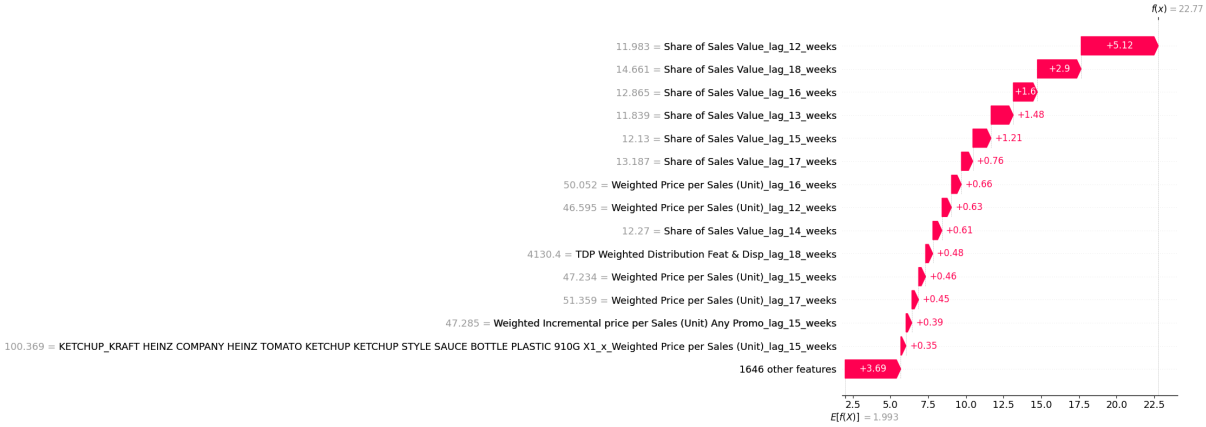


Figure B.18.
Waterfall Plot of Feature Importance for the Highest Share SKU within Meat & Fish Sauces Subcategory

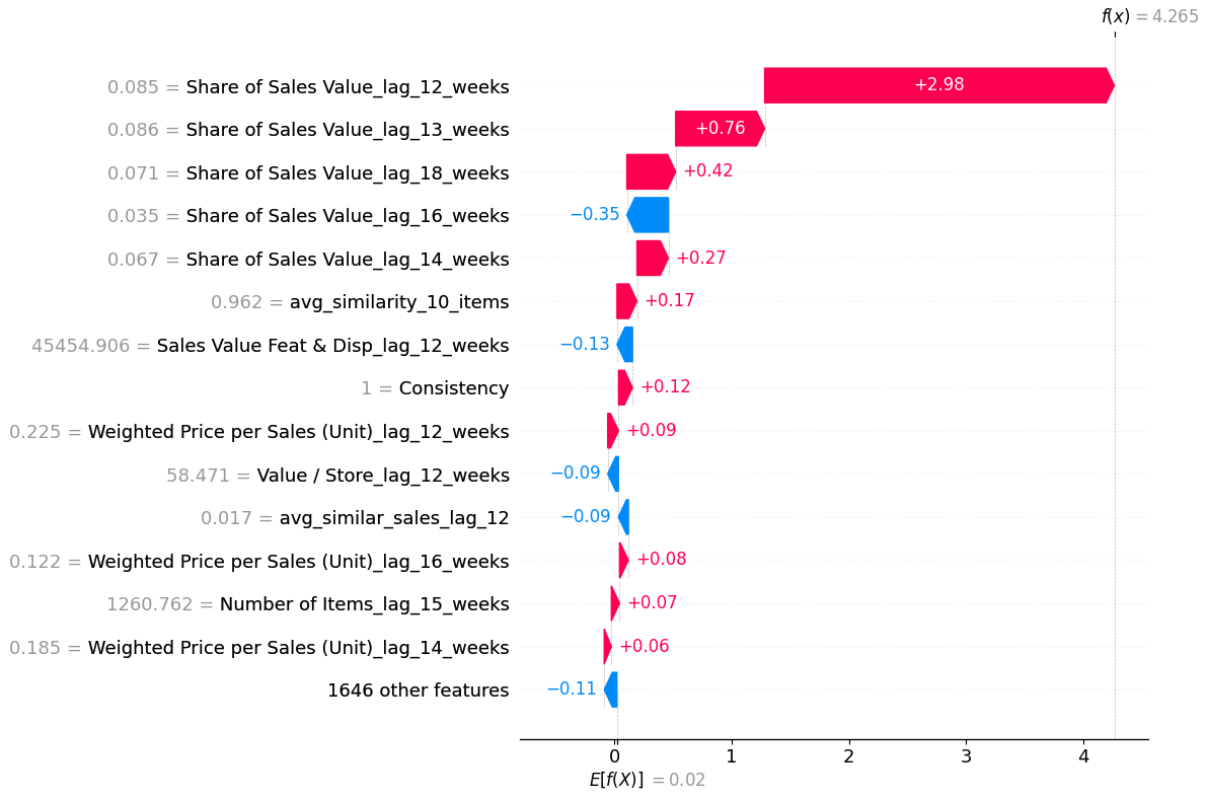


Figure B.19.
Waterfall Plot of Feature Importance for the Highest Share SKU within Mayonnaise Subcategory

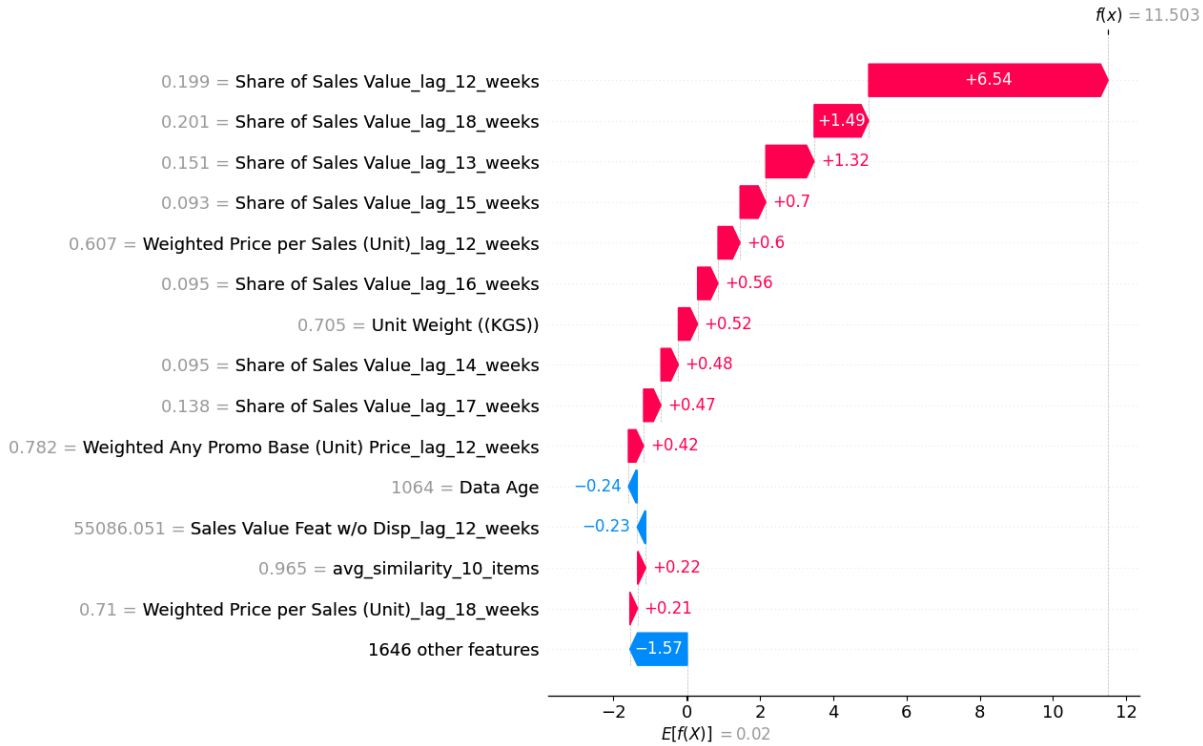


Figure B.20.
Waterfall Plot of Feature Importance for the Highest Share SKU within Other Dressings Subcategory

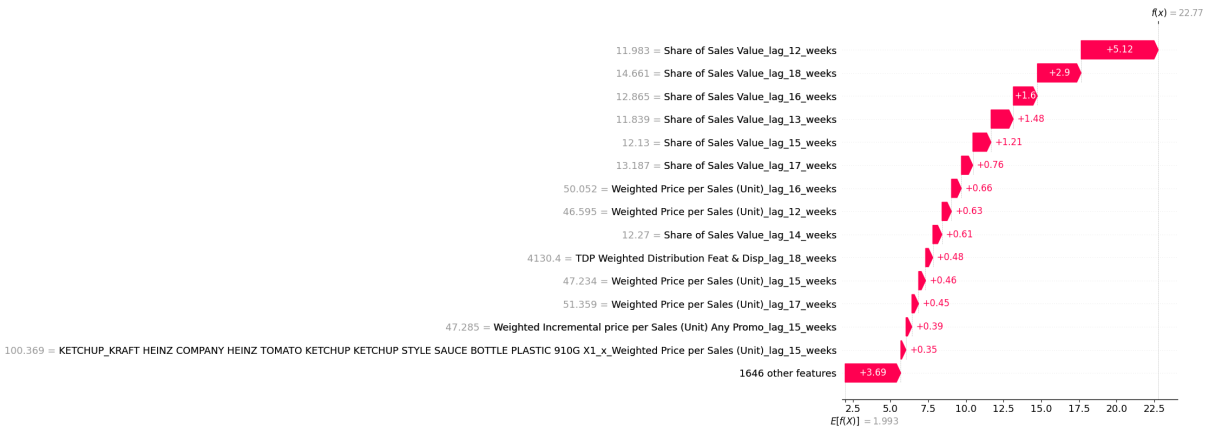


Figure B.21.
Waterfall Plot of Feature Importance for the Highest Share SKU within Salad Dressings Subcategory