Executive Summary

This report aims to address the question: *"Can the placement of products predict their sales volume? Moreover, does this relationship differ between in-store placements versus brochure placements or between product categories?* Three datasets including the causal, transaction, and product data were integrated, processed, and cleaned to answer the question. Initial exploratory data analysis showed that the peak shopping time exists, suggesting seasonality. Further evaluation of the top-selling products (drug-related items, groceries, and packaged meats) revealed distinct display preferences by department.

Regarding the model building, the predictors of sales volume selected were price, day, department, store size (cluster), sub-commodity description, commodity description, interaction (the combination of mailer and display), mailer (all categories), display (all categories), and coupons. These were used in an initial linear regression model. Recognising the potential for nonlinear relationships, a Random Forest (RF) algorithm was applied, resulting in an R^2 of 0.84, where 84% of the variance in sales volume can be explained by the predictors. Key predictors of sales volume were price, day, department, cluster, and sub-commodity description. While mailer and display were not in the top 5, their combined impact (interaction) remained notable. From the model results, the answers to the prior questions can be derived. The placement of products in both mailers and displays have limited predictive power for sales volume. Brochures displays have a stronger impact than in-store display, and the combination of both remains the most effective. Product categories contribute more to sales volume, with specific departments, such as drug-related products and packaged meats, benefiting from tailored display preferences. These products perform better when placed in their preferred display locations. Consequently, a few recommendations were made based on the feature of importance, including pricing strategies and placements on various mailers and in-store displays based on product categories. This analysis provides actionable insights to improve sales strategies, maximize promotional efficiency, and refine product placement approaches.

Table of Contents

1 Introduction

1.1 Introduction

Given the current economic climate, characterised by the ongoing cost of living crisis, high inflation, and shifting consumer lifestyles that make them rethink spending priorities, retailers face significant challenges in maintaining sales (Retail Insight Network, 2024). Additionally, the highly competitive retail market exacerbates these challenges, making it even harder for retailers to sustain their product sales whilst outperforming their competitors. Nonetheless, understanding the distinct factors influencing product sales is critical for businesses aiming to optimise their strategies and improve profitability. A key factor is product placement. Research has shown that three in four shoppers in supermarkets make a substantial number of buying decisions while in-store, underscoring that the location of a product can be regarded as an essential contributing factor in influencing purchasing decisions (Chandolia, 2024). This raises the question of whether product placement can predict sales volume and, if so, whether it can be leveraged as a strategic tool.

1.2 Problem statement

In recent years, research aimed at addressing predictive questions related to sales forecasting within the retail space has rapidly advanced. The development and implementation of machine learning tools have frequently been proposed as the solution, particularly due to their intelligent algorithms that can analyse vast amounts of time-series data and identify patterns (Arunraj and Ahrens, 2015). These tools have shown to have strong predictive capabilities, hence enabling businesses to make data-driven decisions that can be transformed into strategies that drive profitability.

Given the comprehensive dataset at hand, which includes household-level transactions over two years from 2500 frequent shoppers at the business, a simple descriptive analysis would be insufficient to answer the proposed business question of: “Can the placement of products predict their sales volume? Moreover, does this relationship differ between in-store placements versus brochure placements or between product categories?” Consequently, this report proposes a linear regression and the Random Forest (RF) algorithm as models to address the aforementioned question. Utilising the datasets provided, the data was cleaned, and a base model was created, developed, and tested. As such, this report outlines the entire process and presents the results. A literature review that justifies the use of RF will first be presented, followed by a section on the workflow, results, analysis, and evidence-based recommendations. Finally, the report concludes with a summary of the findings.

2 Literature Review

2.1. Machine Learning Algorithm – Linear Regression

Traditionally, both researchers and businesses have frequently used linear regression models to describe the relationship between demand and price, however analysing and predicting customer behaviour is often a complex endeavour. It could be argued that linear regression models might not be able to fully capture the aforementioned dynamic (Jonsson and Fredrikson, 2021). Conversely, the use of machine learning techniques enables processors to learn from large

datasets, uncover potentially hidden patterns, and automate complex processes. Additionally, machine learning techniques have consistently demonstrated superior effectiveness and accuracy compared to traditional statistical regressions. A comparative study on pricing optimisation in insurance markets found that machine learning models outperformed generalised linear models (GLMs) in all aspects, including handling complexity, large data volumes, and describing non-linear relationships (Spedicato et al, 2018). Machine learning models also demonstrated superior accuracy and discriminative power. These machine learning models are particularly advantageous for studying customer behaviour and sales as they effectively capture complex relationships that are often unsuitable for linear modelling. Furthermore, in forecasting future sales, time-series data often serve as the primary input, traditionally analysed using classical statistical methods like autoregressive moving averages or OLS. However, machine learning techniques have been argued to be more powerful and flexible as they allow the integration of additional external input variables beyond the specified time-series (Tsoumakas, 2018). This suggests that machine learning techniques might be more appropriate in the field of sales forecasting compared to their traditional counterparts.

That said, several concerns have been raised about comparing GLMs to machine learning techniques. Spedicato et al (2018) highlighted the “black box” nature of machine learning models, where parameters are more difficult compared to those in GLMs. In GLMs, the significance or the weight of the input variables can be easily interpreted by examining the model coefficients, a feature that is not readily available in machine learning models. This lack of interpretability can pose challenges when the target audience lacks technical expertise. Since sales predictions are often presented to managers and stakeholders who may not have a deep understanding of machine learning, misinterpretations could arise, making the insights less practical or even misleading if not communicated appropriately. Consequently, a feasible solution would be to choose simpler machine learning algorithms such as regression trees that are easier to interpret (Garre et al, 2020).

2.2. Machine Learning Algorithm – Random Forest

One machine learning algorithm that is often used in the context of sales forecasting is Random Forest (RF), introduced by Breiman (2001). RF addresses overfitting by constructing multiple decision trees and combining their results to improve model robustness and prediction accuracy. Its ability to process multi-dimensional data and capture non-linear relationships has led to a wide range of applications in retail, such as sales forecasting, customer behaviour analysis and demand forecasting. RF's effectiveness has garnered empirical support. For instance, Bauer and Kohavi (1999) showed that RF can effectively predict the sales of retail products. Moreover, its predictive performance surpasses traditional methods, particularly when factors such as promotional activities and seasonal changes are considered. However, the same researchers argue that RF might not be as efficient as simpler models when the data set is small (Amit & Geman, 1997). This suggests that the choice of whether to use RF should be tailored to the size of the data and business needs. Expanding on RF’s versatility, Rigatti (2017) further explored the use of RF in one of the various applications of retail data analytics, including customer segmentation and inventory optimisation, showing that RF can provide retailers with accurate market forecasts by analysing substantial amounts of transactional data and identifying the key factors influencing sales.

While RF can still fall under the risk of being considered a “black box model”, its strengths lie in handling intricate data patterns. RF excels in feature selection and modeling non-linear relationships, often outperforming simpler models in these areas. Louppe (2014) further noted

that RF is not only effective for classification but also serves a crucial role in feature selection and dimensionality reduction tasks, as it helps to identify the most predictive variables by assessing the significance of features. This is an especially important feature in determining important determinants of sales. Further stressing its ability to handle complex data Chandola et al. (2009).

3 Workflow

This section outlines the project's workflow for processing and handling the data. The initial subsections outlined the data description, preprocessing, and exploratory data analysis, while the latter section introduced the two models; linear regression and RF.

3.1. Data Description

To analyse the relationship between product features and sales volume, several supermarket datasets were used. The "transaction_data" file contained transaction history, while the "product" file provided information on product categories and descriptions. The "causal_data" file indicated product placement in brochures ("mailer") and on shelves ("display") at the time of sale. A detailed data description is illustrated in Appendix 1.

3.2. Data Preprocessing

In the data pre-processing step, the transaction data, causal data, and product data were integrated using three keys: PRODUCT_ID, WEEK_NO and STORE_ID. This led to a dataset of 20 columns and 2,596,590 records. A review revealed significant missing values in the mailer and display columns, suggesting potential missing campaigns or placement. Operating under the assumption that the products were not placed on special display or mailers under “promotion, the missing values were replaced with “0”. Moreover, descriptive analysis shows zero values in both quantity and sales, indicating voided transactions or free items. 18,919 rows were removed as a result. To ensure accurate analysis, the data was filtered to retain only transactions where both quantity and sales values were greater than zero (Hunter and Schmidt, 2004). A new feature was generated to indicate whether a transaction involved a redeemed coupon based on the "coupon_match_disc" attribute (Appendix 2-1).

Cardinality refers to the number of unique values within a column. Columns with low cardinality are ideal for one-hot encoding, while high cardinality columns, such as department and commodity descriptions, require ordinal encoding to reduce dimensionality. One-hot encoding was applied to the mailer and display with one column dropped to avoid redundancy. The "size of products" field was deemed uninformative and excluded from model training to ensure accuracy and validity (Moro et al., 2014). Moreover, to observe the influence of placement on sales volume, a column “INTERACTION” was generated as a combination of display and mailer (Appendix 2-2).

To address the influence of store size on sales volume, the stores were clustered by size for inclusion in the model (Appendix 2-3). Since specific store information is unavailable, stores were categorised based on derived patterns from the transaction data. K-means was applied to identify three store clusters: Cluster 0 (low transaction volume and few customers), Cluster 2 (medium transaction volume and customer traffic), and Cluster 1 (high transaction volume and customer traffic). These clusters were integrated into the dataset as predictors for the model.

Finally, to eliminate differences in the magnitude of numerical variables, the numerical columns were normalised. Using Standard Scaler, the variables were converted to have a mean of 0 and

a standard deviation of 1. The standardised columns included quantity purchased (QUANTITY), coded categorical variables (e.g. DEPARTMENT_en, etc.), and other numeric features (e.g. DAY and PRODUCT_PRICE) (Appendix 2-4).

3.3. Exploratory Data Analysis (EDA)

Following preprocessing, exploratory data analysis (EDA) was conducted to examine data distribution, uncover trends, patterns, and relationships, and identify seasonality in purchasing behaviour. The analysis also evaluated the effectiveness of display and mailer strategies. Detailed results are presented in the subsequent section.

4 Machine Learning

4.1. Multiple Linear Regression

As highlighted in the literature review, interpretability is a key consideration when presenting findings to business executives, as it allows for a clear understanding of the relationships we gained from correlation analysis (Appendix 4) without requiring technical expertise. For this reason, Multiple Linear Regression (MLR) was selected as the initial model. MLR extends the simple linear model by incorporating multiple predictors (Tsoumakas, 2018), including product price, product description, store details, day, display, and mailer information, to predict sales volume. In addition, columns related to coupons were utilised to reflect the usage of coupons, providing new valuable knowledge for improving model accuracy. Moreover, MLR can serve as a decider in the statistical significance of various variables in determining the sales volume.

To perform the modelling, the dataset was standardised (Appendix 3) and utilised, with the first 30% of the time-ordered data used as training data and the subsequent 10% to evaluate the model performance to preserve the time dependence of transactions (Sarvi, 2020). The predictors of sales volume selected were price, day, department, store size (cluster), sub-commodity description, commodity description, interaction (the combination of mailer and display), mailer (all categories), display (all categories), and coupons (Appendix 1). The Ordinary Least Squared (OLS) method was used to evaluate the model. R^2 score and Root Mean Squared Error (RMSE) were used as performance metrics. R^2 is the coefficient of determination that can be interpreted as the proportion of the information in the data that is explained by the model (Kuhn & Johnson, 2013).

4.2. Random Forest

Given the strengths of the RF model outlined in the literature review, it was selected to address the business question. RF makes decisions by building multiple decision trees during training and combining their outputs for the final prediction.

To perform the modelling, the original dataset that has not been standardised was utilised. This is due to the model's insensitivity to scaled numeric data (Kuhn & Johnson, 2013). A similar approach to the previous models was taken, where the first 30% of the time-ordered data was used as training data and the subsequent 10% to evaluate the model performance. The predictors of sales volume selected were those selected in the linear regression model. R^2 score and RMSE were again used as metrics of performance, serving as an additional tool to compare the effectiveness of the two models.

To optimise the model (Appendix 5), hyperparameter tuning using grid search was performed. This is where RF's parameters like the number of trees, maximum depth, and the splitting features are adjusted to find the optimal set of combinations of settings (Yang & Shami, 2020). 81 sets of combinations of parameters were tested. These settings impact the model's ability to learn patterns from the data and contribute to creating a more reliable and robust RF model.

Finally, to test for robustness, 70% of the data was used for training with a time-series split in line with Sarvi (2020), while the rest were reserved as out-of-sample data to test for performance. The R^2 score and RMSE were again used as metrics of performance.

5 Results

5.1. Exploratory Data Analysis

The time dependency of transaction data was analysed across three timeframes to understand customer behaviour shown in Figure 1. Between 10 AM and 8 PM, the highest transaction volumes occur, reflecting peak shopping hours. A seasonal pattern is evident, with certain days showing higher sales, possibly due to leisure shopping or promotions, though it is unclear if this aligns with weekdays or weekends. Weekly sales show consistent volumes with occasional outliers, likely driven by events, holidays, or campaigns.

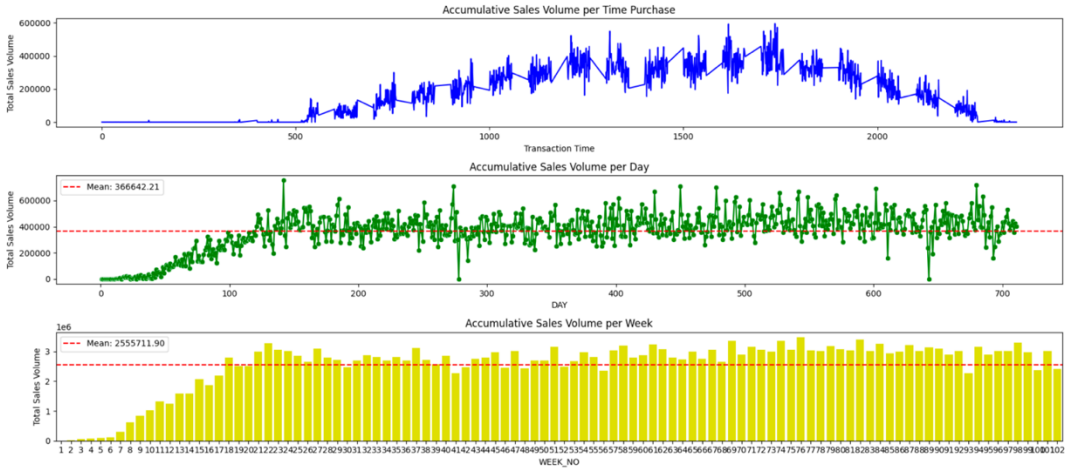

Figure 1 Transaction Seasonality

The causal data includes three distinct "strategies", including only display, mailer, and interaction (combination of the two). It is worth noting that the recorded data for each types of strategy is imbalanced. Regardless, during the EDA stage, the conversion of causal data (mailer, display, and interaction) into realised transactions was analysed to gauge strategy effectiveness. Conversion rates were calculated as the percentage of rows in transactional data (i.e., display) relative to corresponding rows in causal data. This process was repeated for mailer and interaction, allowing evaluation of each strategy's impact. Results show that 1.37% of mailer data were converted into actual transactions, hence, ceteris paribus, mailer's contribution to sales volume is 1.37%. Furthermore, the display's contribution was 1.29%, while the interaction (taking mailer and display into account) had the highest contribution at 2.92%. This suggests that customers are effectively influenced by brochures and product placements they prefer, hence underscoring the strategy's success in enhancing customer engagement and sales.

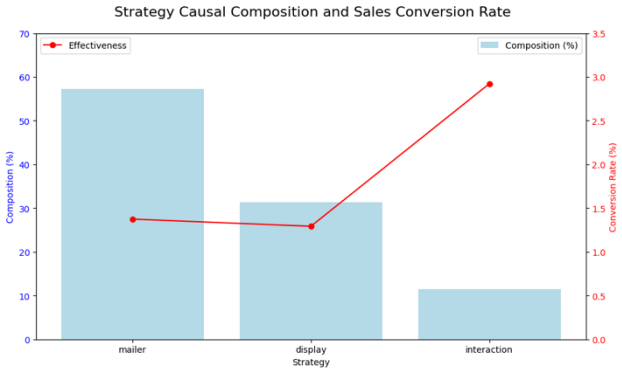

Figure 2 Strategy and Sales Conversion Rate

Building on the conversion rate analysis a deeper evaluation was conducted by breaking down the mailer, display, and interaction conversion rates for the top three best-selling products (department) determined by transaction frequency. The top three selling products were drug-related products, groceries, and packaged meat. Additionally, the model further breaks down each type of mailer and display (Figure 3). The results show that for display each product type (department) has their preferences of display. For drug-related products and packaged meats, the most common display choice is Display 2, typically located at the store rear, while groceries are predominantly placed in Display A, or on-shelf. In contrast, preferences for mailer positioning remain consistent across departments, with Mailer D, often featured on the front page, being the most favoured option.

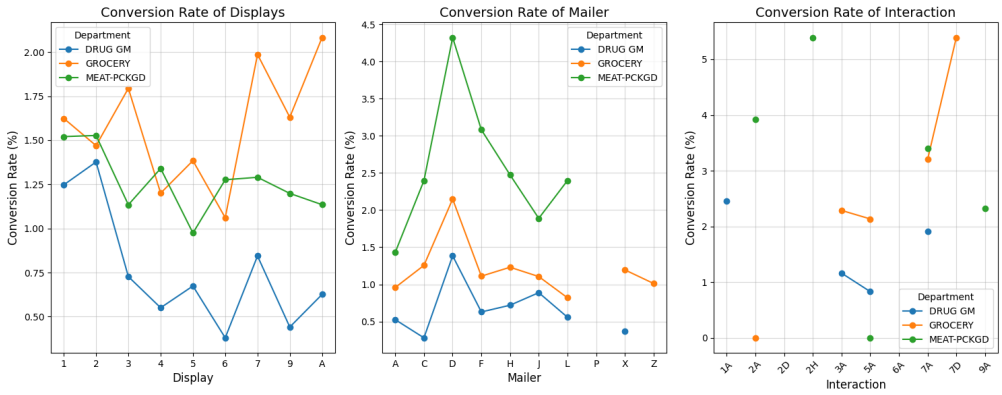

Figure 3 Sales Conversion Rate based on Department

5.2. Linear Regression

The OLS results of the model indicate an R^2 value of 0.02, meaning that only 2% of the variation in sales volume is explained by the independent variables (Appendix 6-1). This suggests that the model's performance is questionable in explaining the factors influencing sales volume. The model's inherent issue might be its inability to explain the non-linear relationship between variables.

5.3. Random Forest

The preliminary results from the RF model show that the initial model can explain 77% of the variance of sales volume, indicating a strong explanatory power of the model and the independent variables. However, based on the residual analysis (Appendix 6-2) and the 533.6 of RMSE

indicating that the model's predictions might deviate by an average of 533.6 units from the actual sales volume. This discrepancy is suspected to be due to the inclusion of fuel quantities (measured in liters), which are not comparable to the quantities of other commodities, such as meat products, typically purchased in discrete quantities. Nevertheless, as RF models rely on decision trees, splitting data at feature thresholds, they are not prone to numerical precisions or convergence that affect algorithms (Kuhn & Johnson, 2013). In this case, scaling quantities is not deemed necessary. In addition, a key motivation for developing this model is its ability to accommodate all products within the grocer's portfolio, ensuring broad applicability. To achieve this, all product types are included in the analysis. Following the grid search, the optimal parameters were selected and applied and the optimised model's R^2 when tested reached 0.84.

The feature of importance lists the contribution of input into the RF model. The top-ranked features are price, day, department, cluster, sub-commodity description, brand, commodity description, interaction effects, mailer categories (D, H, F, A), and display 5 (Appendix 6.2). Price emerges as the strongest predictor of sales volume, followed by product seasonality (day), department, and store size. These factors highlight predictive powers of factors beyond mailer or display placement on sales volume. Referring to the original question, the answer of "*Can placement of products predict sales volume?*" appears to lean more towards "no" to some extent as price, seasonality, and department (product categories) demonstrate stronger predictive and explanatory powers. However, mailer and display interactions might still have some contributions. While mailers and displays individually show limited impact, their combined effect appears more significant. For instance, customers may be more inclined to purchase an item when it is highlighted in a mailer and prominently displayed in the store. Moreover, to answer the second part of the question, mailers (particularly type D, H, F, and A) generally exert a greater influence on sales volume than display.

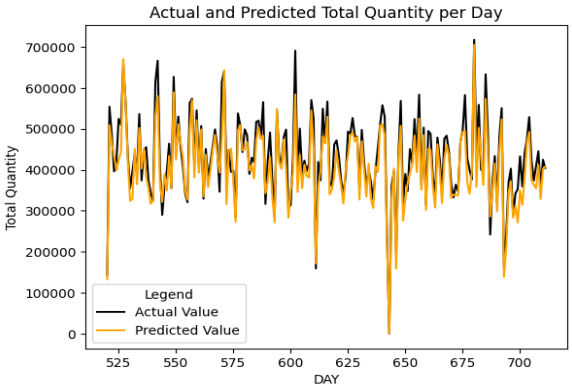

Figure 4 Random Forest Model Prediction Result

6 Recommendation

This section aims to provide recommendations based on the findings of this report. As such, the structure will be based on the RF's feature of importance, starting with the highest-ranked features. By prioritising these insights, the goal is to optimise promotional efforts and maximise sales impact.

1. Price ranked the highest in predicting sales volume, hence pricing strategies should be implemented to boost sales. Some examples could be bundle deals, or special discounts on products nearing expiration. Additionally, if the goal is to increase sales volume, volume-based pricing (lower prices for higher quantities purchased) could be adopted to stimulate demand.

2. Seasonality, particularly peak times, days, and hours proved to have a significant impact on sales volume. The findings show that across the grocery stores, 5-6 pm seems to be the peak time. To capitalise on this, strategies such as product demonstrations and offering samples could be introduced during this time to attract more customers, informing them of existing products and, in turn, making them more inclined to purchase them. Additionally, interactive displays could be especially effective during high-season months, ensuring customer engagement and sales are maximised.

3. To boost sales, mailer and display strategies should be tailored to each department's preferences. For instance, drug-related products perform well (translates to sales) in the store's rear (Display 2), so placing supplements and over-the-counter deals in this area can drive sales. Offer discounts on additional quantities, as this could further increase volume. Similarly, packaged meats should feature bulk items in display 2, encouraging larger purchases. This suggests that tailoring display location and promotion for each product type in both the mailer and display can optimise sales performance.

4. Given the differences in store sizes, promotional strategies can be tailored accordingly. Larger and smaller stores share the same top three departments: grocery, packaged meats, and drugs, while medium-sized stores have produce as the second best-selling category. Store executives should align their mailer and display promotions of specific product type according to the store size. Additionally, smaller stores, which generally experience lower traffic and transactions, could benefit from adopting additional strategies, such as targeted promotions or exclusive discounts, to boost customer engagement and drive sales.

7 Conclusion

This report revolves around answering the question *"Can the placement of products predict their sales volume? Moreover, does this relationship differ between in-store placements versus brochure placements or between product categories?".* Through building a base model and an RF algorithm afterwards, the model suggests that the placement of products cannot necessarily predict sales volume. This is due to other variables that the model deems to be more important, such as price, day, department, cluster, and sub-commodity description. Product category is the strongest predictor of sales volume, followed by in-store displays, which have more impact than brochures. However, using both displays and brochures together performs better.

Limitations within the data must be considered as they may influence result reliability. To improve future predictions and models, several additional data might be required:

1. The exact timing and date to better understand the seasonality of the transaction pattern regarding the month and week. This would allow for possible seasonal promotions such as Christmas deals.
2. Comprehensive placement information for mailer and display as the current missing value can be considered extensive. This will allow for more accurate model predictions.
3. In-depth store-specific information for all the grocery stores under the same company would give additional information such as location, size of the store, etc. This would in turn allow the model to take all this information into account and produce a better-tailored placement strategy for the store executives for each specific store.

8 References

Amit, Y., & Geman, D. (1997). “Shape quantization and recognition with randomized trees”. *Neural computation*, 9(7), 1545-1588. Available at: https://dl.acm.org/doi/abs/10.1162/neco.1997.9.7.1545

Arunraj, N. S., & Ahrens, D. (2015). “A hybrid seasonal autoregressive integrated moving average and quantile regression for daily food sales forecasting”. *International Journal of Production Economics*, *170*, 321-335. Available at: https://www.sciencedirect.com/science/article/pii/S0925527315003783

Bauer, E., & Kohavi, R. (1999). “An empirical comparison of voting classification algorithms: Bagging, boosting, and variants.” *Machine learning,* 36, 105-139. Available at: https://link.springer.com/article/10.1023/A:1007515423169

Breiman, L. (2001). “Random forests”. *Machine learning*, 45, 5-32. [online] Available at: https://link.springer.com/article/10.1023/a:1010933404324

Chandolia, H. (2020). *What data says about in-store product placement? - RSA America*. [online] RSA America. Available at: https://rsaamerica.com/what-data-says-about-in-store-product-placement/ [Accessed 15 Nov. 2024].

Chandola, V., Banerjee, A., & Kumar, V. (2009). “Anomaly detection: A survey.” *ACM computing surveys (CSUR)*, *41*(3), 1-58. Available at: https://dl.acm.org/doi/abs/10.1145/1541880.1541882

Garre, A., Ruiz, M. and Hontoria, E. (2020) ‘Application of machine learning to support production planning of a food industry in the context of waste generation under uncertainty’, *Operations Research Perspectives*. [Online] Available at:https://www.sciencedirect.com/science/article/pii/S2214716019301988 [Accessed 20 November 2024].

Hunter, J.E. and Schmidt, F.L. (2004). *Methods of Meta-Analysis Corrected Error and Bias in Research Findings*. [online] ResearchGate. Available at: https://www.researchgate.net/publication/235726244_Methods_of_Meta-Analysis_Corrected_Error_and_Bias_in_Research_Findings

Jonsson, E. and Fredrikson, S. (2021) *An investigation of how well random forest regression can predict demand: Is random forest regression better at predicting the sell-through of close-to-date products at different discount levels than a basic linear model?* Dissertation. Available at: https://urn.kb.se/resolve?urn=urn:nbn:se:kth:diva-302025 [Accessed 20 November 2024].

Kuhn, M. and Johnson, K. (2013). “Applied Predictive Modeling.” *SpringerLink*. [online] doi:https://doi.org/10.1007-978-1-4614-6849-3.

Louppe, G. (2014). “Understanding random forests: From theory to practice”. arXiv preprint arXiv:1407.7502. Available at: https://arxiv.org/abs/1407.7502

Moro, S., Cortez, P. and Rita, P. (2014). “A data-driven approach to predict the success of bank telemarketing.” *Decision Support Systems*, [online] 62, pp.22–31. doi:https://doi.org/10.1016/j.dss.2014.03.001.

Retail Insight Network (2024). *2024: Another challenging year for retail*. [online] Retail Insight Network. Available at: https://www.retail-insight-network.com/analyst-comment/2024-challenging-year-retail/ [Accessed 15 Nov. 2024].

Rigatti, S. J. (2017). “Random forest.” *Journal of Insurance Medicine*, 47(1), 31-39. Available at : https://doi.org/10.17849/insm-47-01-31-39.1

Spedicato, G., Dutang, C. and Petrini, L. (2018) “ *Machine Learning Methods to Perform Pricing Optimization: A Comparison with Standard Generalized Linear Models*.” [online] Available at: https://www.casact.org/sites/default/files/2021-07/Machine-Learning-Methods-Spedicato-Dutang-Petrini.pdf.

Sarvi, T. (2020) “Predicting product sales in retail store chain.” Available at: https://aaltodoc.aalto.fi/server/api/core/bitstreams/2f656fd8-bc56-45b5-b6f7-9f3307380414/content.

Tsoumakas, G. (2018) “A survey of machine learning techniques for food sales prediction.” *Artificial Intelligence Review*, [online] 52(1), pp.441–447. doi: https://doi.org/10.1007/s10462-018-9637-z.

Yang, L. & Shami, A. (2020*) On hyperparameter optimization of machine learning algorithms: Theory and practice*. Neurocomputing (Amsterdam). [Online] 415295–316.

10 Appendix

APPENDIX 1 – Data Description

The “transaction_data” file was applied as a database since it contains basic transaction data, such as product ID, transaction date, quantity, store ID, etc. Based on the records of transactions, the files “product” and “causal_data” were utilised as look-up tables for gaining in-depth transaction insight. The “product” file was applied to gain the product description like category and commodity, and the “causal_data” was applied to gain the placement of the product when the transaction was being made. The table below shows all the columns we take into consideration in the raw data and if it is used in later machine learning part of the report.

Table 1 Original Dataset Description

Original Dataset	Column Title	used in machine learning	Description
transaction_data	HOUSEHOLD_KEY	NO	Uniquely identifies each household
	BASKET_ID	NO	Uniquely identifies a purchase occasion
	DAY	YES	Day when transaction occurred
	PRODUCT_ID	NO	Uniquely identifies each product
	QUANTITY	YES	Number of the products purchased during the trip
	SALES_VALUE	NO	Amount of dollars retailers receive from the sale
	STORE_ID	NO	Identifies unique stores
	COUPON_MATCH_DISC	NO	Discount applied due to retailer’s match of manufacturer coupon
	COUPON_DISC	NO	Discount applied due to manufacturer coupon
	RETAIL_DISC	NO	Discount applied due to retailer’s loyalty card programme
	TRANS_TIME	NO	Time of day when transaction occurred
	WEEK_NO	NO	Week of the transaction. Ranges 1 – 102
product	PRODUCT_ID	NO	Number that uniquely identifies each product
	DEPARTMENT	YES	Groups similar products together
	COMMODITY_DESC	YES	Groups similar products together at a lower level
	SUB_COMMODITY_DESC	YES	Groups similar products together at the lowest level
	MANUFACTURER	NO	Code that links products with same manufacturer together

	BRAND	YES	Indicates Private or National label brand
	CURR_SIZE_OF_PRODUCT	NO	Indicates package size (not available for all products)
causal_data	PRODUCT_ID	YES	Uniquely identifies each product
	STORE_ID	NO	Identifies unique stores
	WEEK_NO	NO	Week of the transaction
	DISPLAY	YES	Display location
	MAILER	YES	Mailer location

Table 2 Display and Mailer Description

field	contents
display	0 – Not on Display
	1 – Store Front
	2 – Store Rear
	3 – Front End Cap
	4 – Mid-Aisle End Cap
	5 – Rear End Cap
	6 – Side-Aisle End Cap
	7 – In-Aisle
	9 – Secondary Location Display
	A – In-Shelf
mailer	0 – Not on ad
	A – Interior page feature
	C – Interior page line item
	D – Front page feature
	F – Back page feature
	H – Wrap front feature
	J – Wrap interior coupon
	L – Wrap back feature
	P – Interior page coupon
	X – Free on interior page
	Z – Free on front page, back page or wrap

APPENDIX 2 – Derived Data

APPENDIX 2-1 – COUPON_FLAG

When consumers have coupons during their shopping, it may increase their desire to purchase specific products. To examine the impact of coupons on sales volume, a new column, "COUPON_FLAG", was generated to show if customers applied coupons when

the transaction was made and later used in model building. "COUPON_MATCH_DISC" in "transaction_data" was selected to check if the customers are applying coupons, if the "COUPON_MATCH_DISC" value is negative, the "COUPON_FLAG" value would be 1, indicating the existence of the discount.

APPENDIX 2-2 – INTERACTION

Column “INTERACTION” was generated to better understand the relationship between placement and sales volume. Columns “DISPLAY” and “MAILER” in the “causal_data” file were selected and combined to gain more in-depth information about their interaction with the sale volume in this column.

APPENDIX 2-3 - Cluster

Considering the traffic of every store might be different, we calculated the operating duration of the store, the number of transactions it had made, and the total products sold in the store and further added a variable “Cluster” in our model.

We aggregated and observed the relationships between various features through scatter plots, we considered factors such as transaction frequency, household, sold product categories and volume, and operating duration.

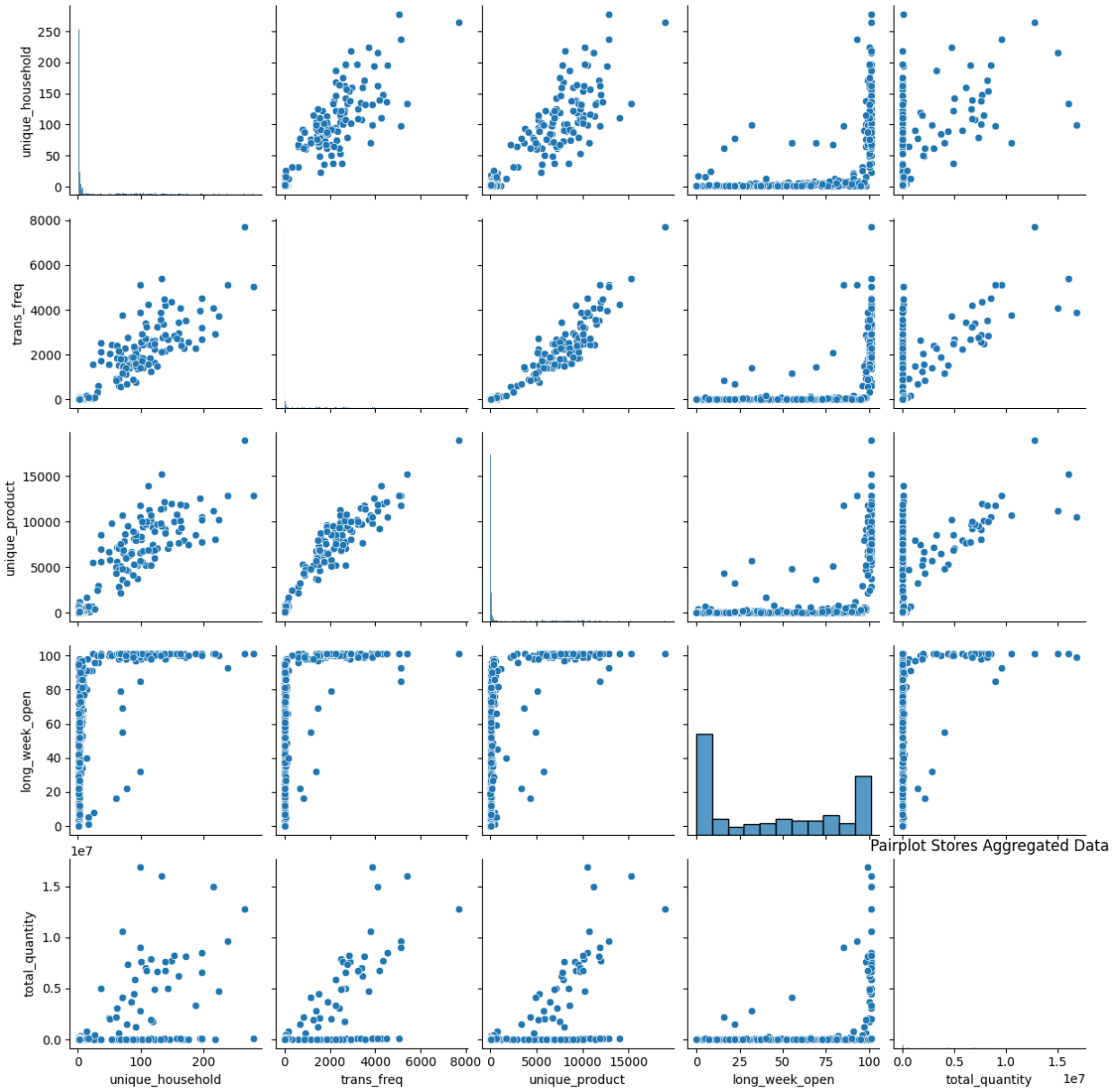

Figure 5 Pairplot of Dataset for Clustering

The Elbow Method was applied first to find the best classification number for later clustering, and the K-mean was utilised to separate the data, considering all the features above, into three clusters.

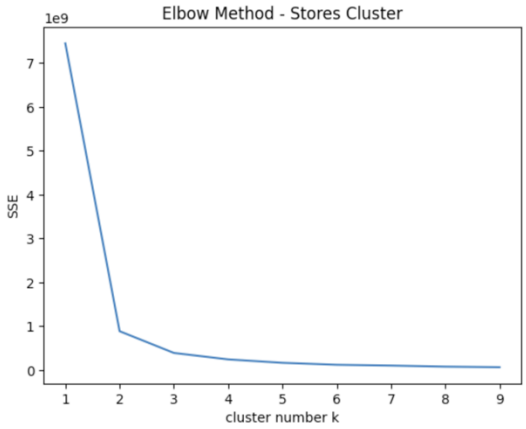

Figure 6 Elbow Method

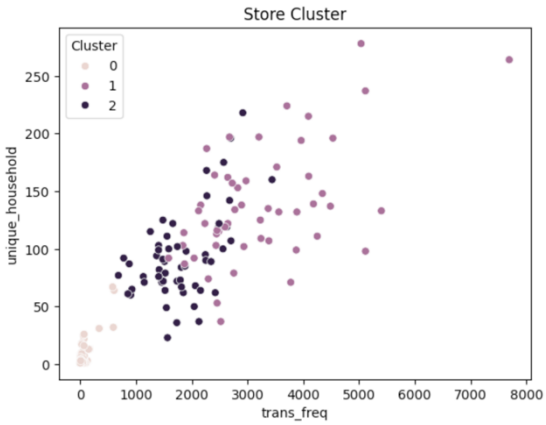

Figure 7 Clustering Result on Count of Unique Household and Transaction Frequency

APPENDIX 2-4 – PRODUCT_PRICE

Knowing that the nature of transactions would be highly related to the price of the product, four columns, which are "SALES_VALUE", "COUPON_MATCH_DISC", "COUPON_DISC", and "RETAIL_DISC", in "transaction_data" were selected to calculate a new column "REAL_SALE_VALUE", indicating the original sale price before customers applying discounts. Moreover, the column "REAL_SALE_VALUE" was divided by "QUANTITY" to demonstrate the real product price, and a new column "PRODUCT_PRICE" was generated subsequently and applied in later model building.

APPENDIX 3 - Data Distribution after Preprocessing

Descriptive analytics methods were utilised to explore variables and to give insight into the data's dispersion. We first demonstrated the distributions of day, quantity, price,

coupon, cluster, placement and category, then examined their relationships with our target feature, quantity.

Since these distributions appeared notably dispersed, the data shows arbitrary distribution due to more categorical data rather than numerical data. As for quantity and product price, left skewness is observed due to the low quantity of purchase and nature of the product price, making it not normally distributed, so we standardised these data to make the subsequent linear regression models fair.

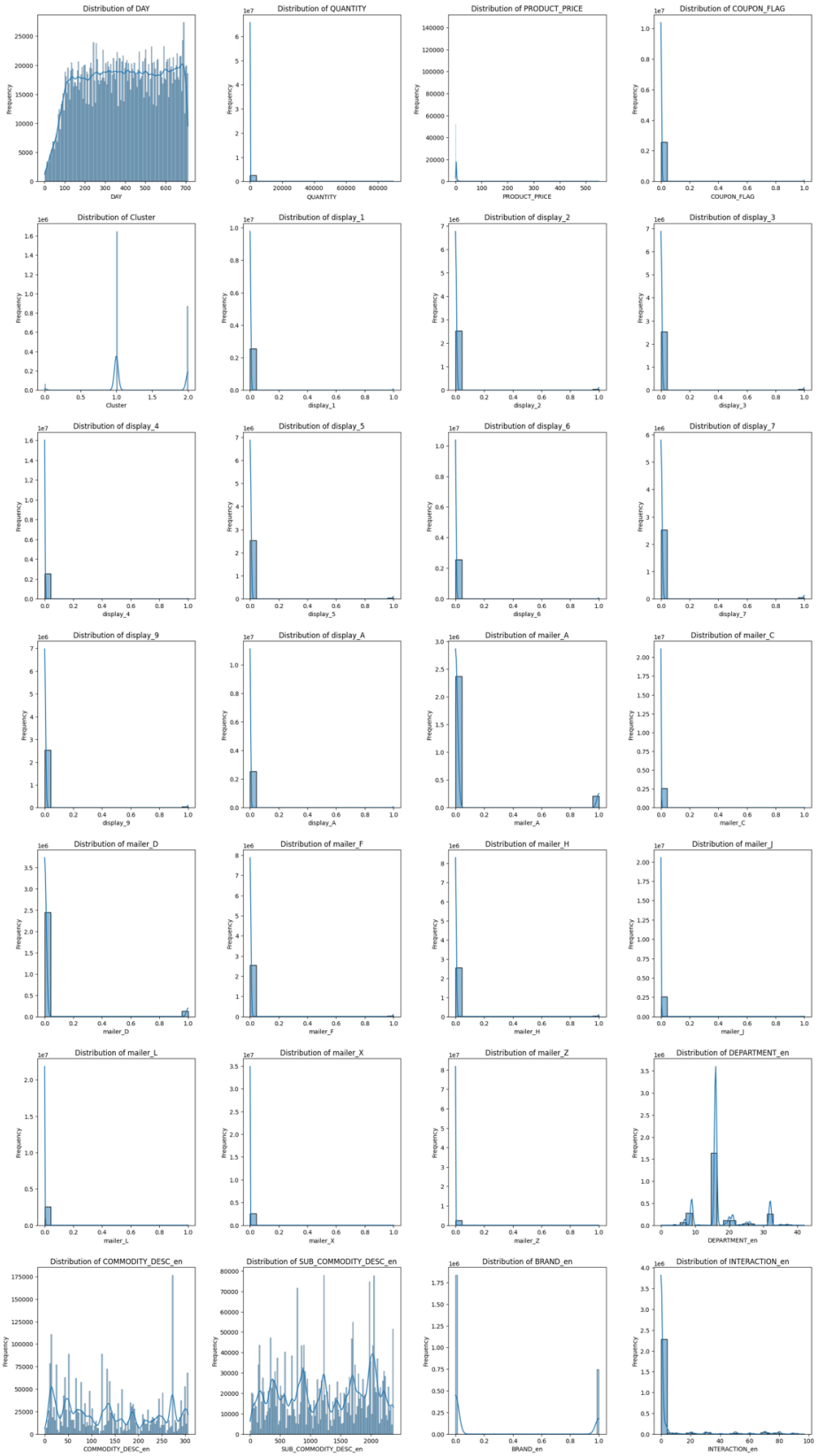

Figure 8 Data Distribution Before Standardisation

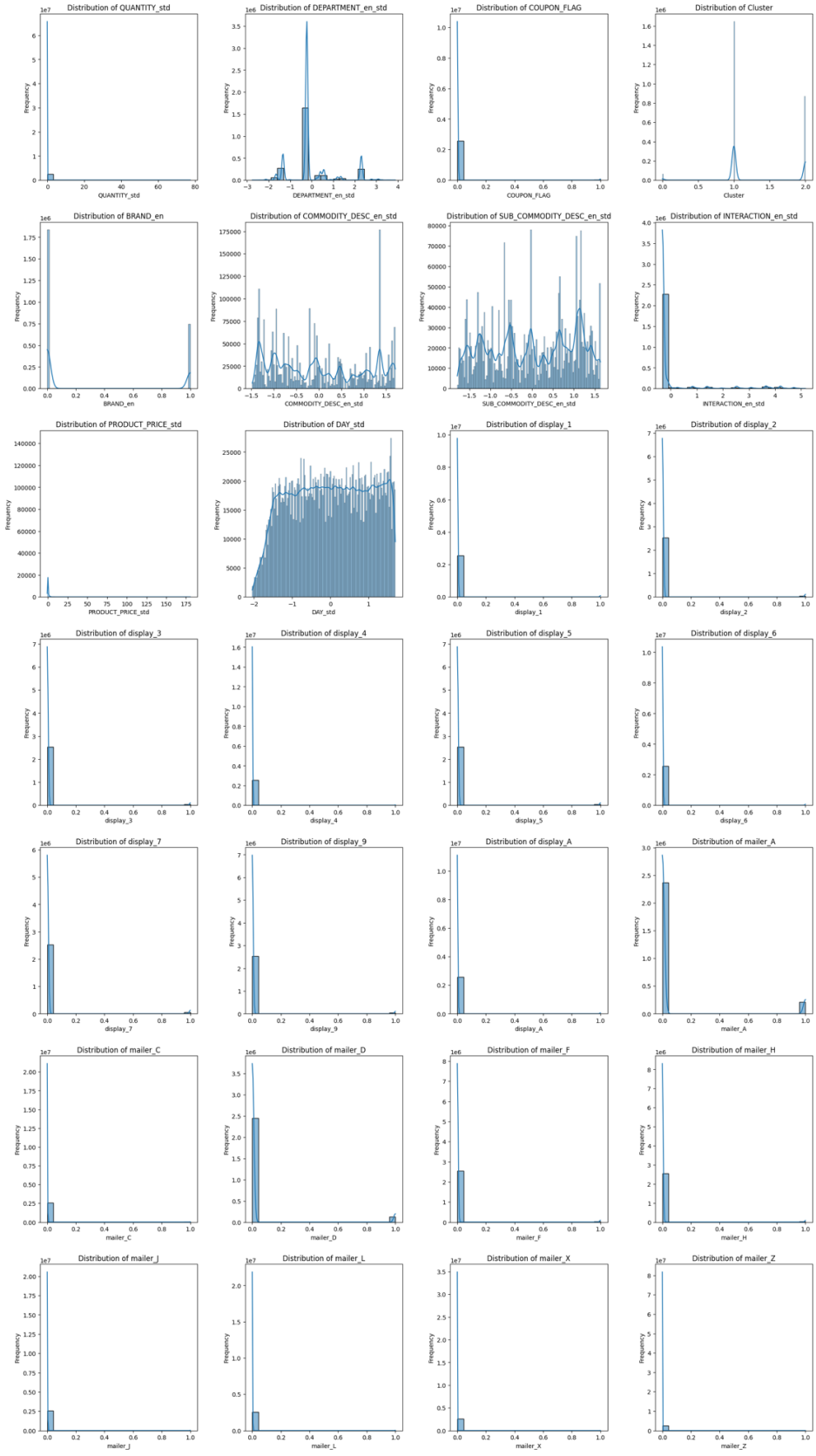

Figure 9 Data Distribution after Standardisation

APPENDIX 4 - Correlation Analysis

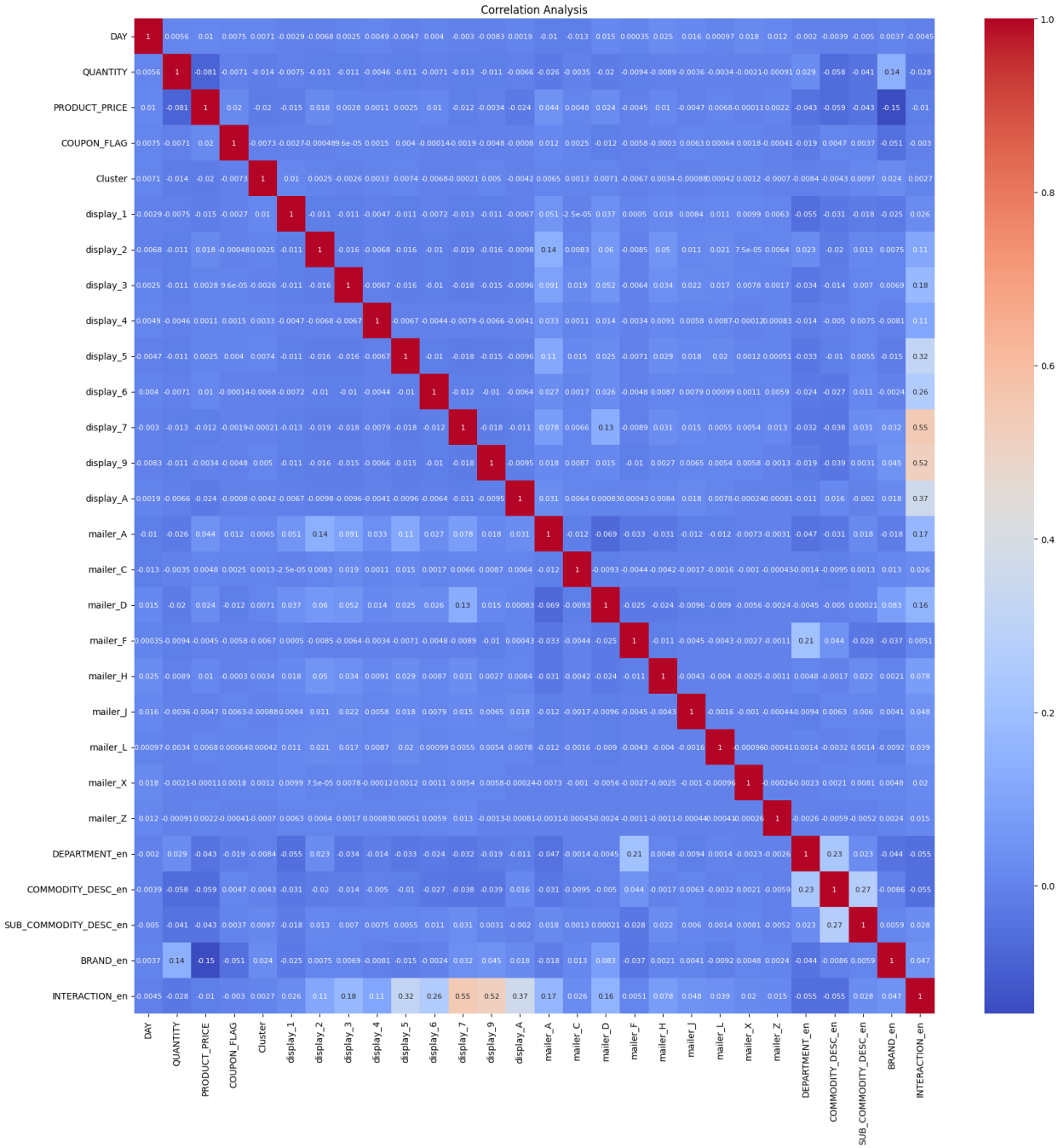

Figure 910 Correlation Analysis Before Standardisation

To understand how the variables correlate before and after the standardisation, two correlation matrices were developed for better comparison. The value did not change after the standardisation since the distribution of the data would not be impacted by standardisation.

A Correlation value higher than 0.7 would be considered highly correlated and, therefore, dropped to avoid multicollinearity. After studying, there is no observation of highly correlated data with the predictors. The only slightly higher correlation value is display/mailer to interaction since the column “INTERACTION” is a derived variable generated for implying the product is on shelves and brochures simultaneously.

APPENDIX 5 – Model Optimization

To optimize the model's performance, hyperparameter tuning was conducted using a grid search approach. Grid search systematically evaluates all possible combinations of hyperparameters to identify the best set of parameters for the model. The hyperparameters tested, their values, and the importance of testing them are summarized in the table below:

Table 3 Parameter Testing using Grid Search

Parameter	**Value Tested**	**Implications**
n_estimators	[50,100,200]	Controls the number of trees in the ensemble. Higher values may improve performance but increase training time.
max_depth	[10,20,None]	Sets the maximum depth of each tree. Limiting depth can reduce overfitting, while None allows unlimited depth.
min_samples_split	[2,5,10]	The minimum number of samples required to split an internal node. Higher values can prevent overfitting.
Min_samples_leaf	[1,2,4]	The minimum number of samples required to be at a leaf node. Higher values make the model more robust.

The results of the hyperparameter tuning process revealed that the optimal set of parameters for the model were {'max_depth': 10, 'min_samples_leaf': 4, 'min_samples_split': 10, 'n_estimators': 50}. Using this configuration, the model achieved a best R^2 score of 0.8426, indicating strong predictive performance. The feature importance of the model indicates

Using the best parameter, we trained the model again using 70% of the data making it into 5 iterations time-series split to assess the robustness of the model, the R^2 score and RMSE were captured for each iteration of the training process, and the results are visualized in the figure below, highlighting the consistency and stability of the model across different range of dataset tested.

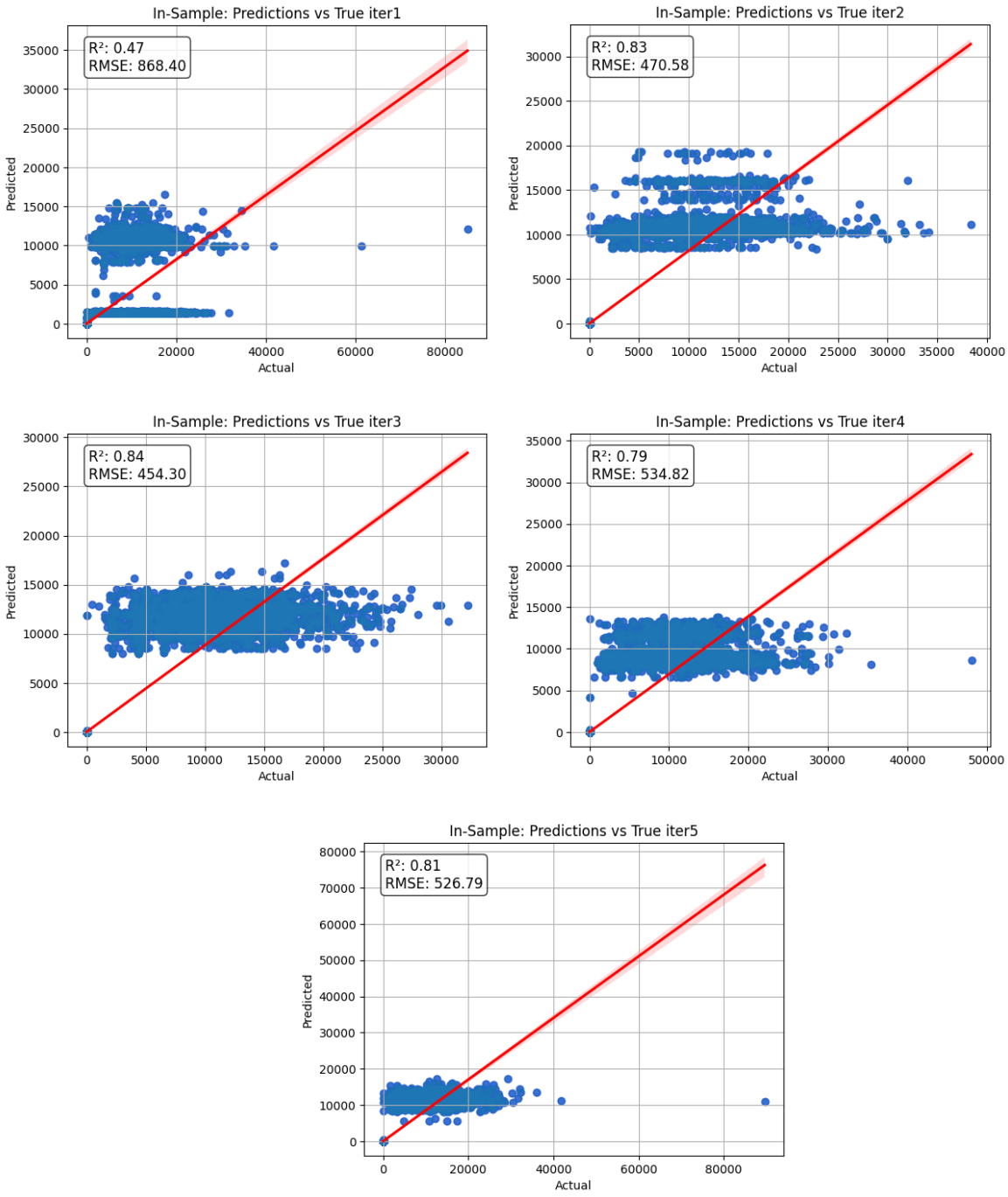

Figure 11 Scatterplot of Actual vs Prediction in each iteration

To validate the results, the optimized model was tested on out-of-sample data, demonstrating strong performance. The model successfully explained 84% of the variance, indicating its robustness and reliability in predicting unseen data. This result highlights the model's ability to generalize effectively beyond the training set, confirming its validity and optimized performance. The result also captures the trends when it is compared with the actual value.

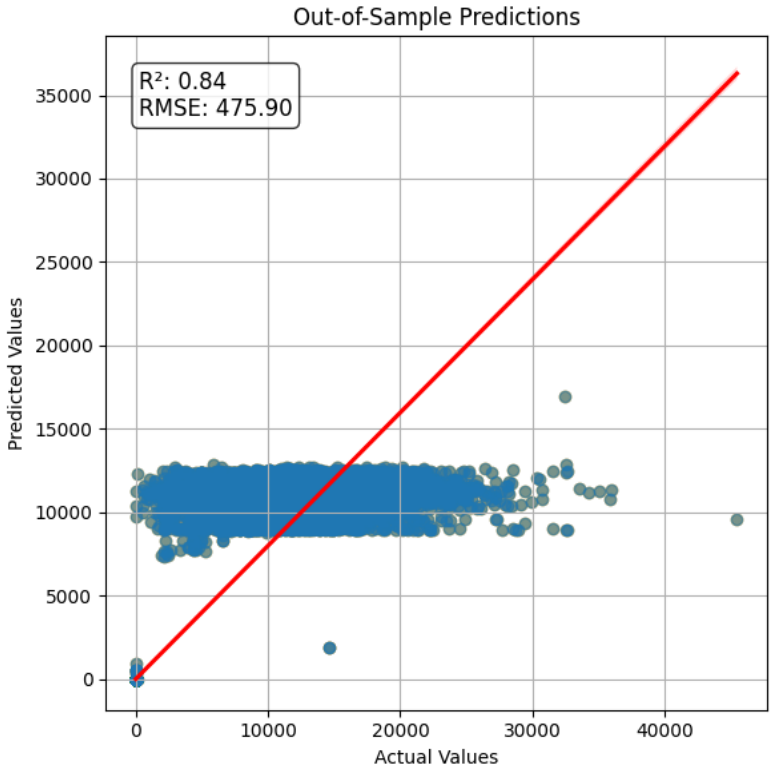

Figure 12 Scatter Plot Actual vs Prediction in Out-of-Sample

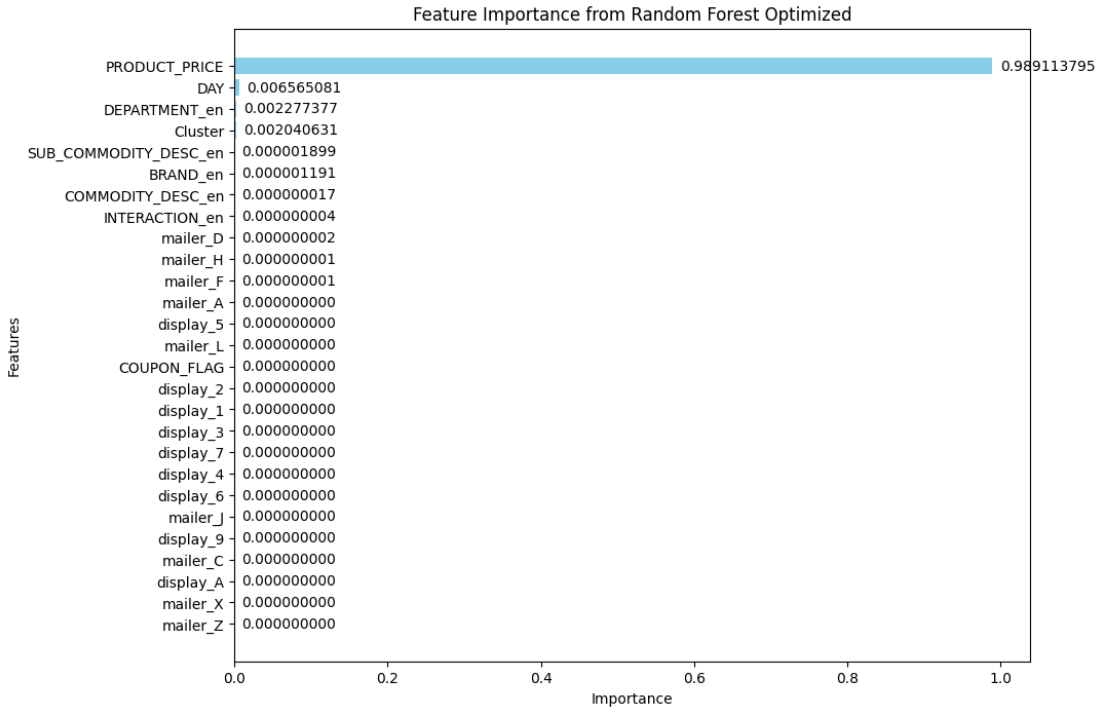

Figure 13 Feature Importance from Optimized Random Forest

APPENDIX 6 – Regression Results and Residual Analysis

APPENDIX 6-1 – Linear Regression Model

Table 2: OLS Regression Results

Dep. Variable: QUANTITY_std.	R- squared: 0.030
Model: OLS	Adj. R-squared: 0.030
Method: Leat Squares	F-statistic: 972.0
Durbin-watson: 1.861	Prob(F-statistic):0.00

	coef	std err	t	P>|t|	[0.025	0.975]
const	-8.80E+05	1.26E+09	-0.001	0.999	-2.47E+09	2.46E+09
DEPARTMENT_en_std	0.0455	0.001	39.292	0	0.043	0.048
COUPON_FLAG	0.0154	0.014	1.13	0.258	-0.011	0.042
Cluster	0.0491	0.002	25.057	0	0.045	0.053
BRAND_en	0.2701	0.002	111.126	0	0.265	0.275
COMMODITY_DESC_en_std	-0.0633	0.001	-54.561	0	-0.066	-0.061
SUB_COMMODITY_DESC_en_std	-0.024	0.001	-21.332	0	-0.026	-0.022
INTERACTION_en_std	-2.73E+06	3.90E+09	-0.001	0.999	-7.65E+09	7.64E+09
PRODUCT_PRICE_std	-0.061	0.001	-54.373	0	-0.063	-0.059
DAY_std	0.0398	0.003	12.413	0	0.033	0.046
display_1	1.54E+06	2.21E+09	0.001	0.999	-4.32E+09	4.33E+09
display_2	3.09E+06	4.41E+09	0.001	0.999	-8.64E+09	8.65E+09
display_3	4.63E+06	6.62E+09	0.001	0.999	-1.30E+10	1.30E+10
display_4	6.17E+06	8.82E+09	0.001	0.999	-1.73E+10	1.73E+10
display_5	7.71E+06	1.10E+10	0.001	0.999	-2.16E+10	2.16E+10
display_6	9.26E+06	1.32E+10	0.001	0.999	-2.59E+10	2.60E+10
display_7	1.08E+07	1.54E+10	0.001	0.999	-3.03E+10	3.03E+10
display_9	1.23E+07	1.76E+10	0.001	0.999	-3.46E+10	3.46E+10
display_A	1.37E+07	1.96E+10	0.001	0.999	-3.85E+10	3.85E+10
mailer_A	1.54E+05	2.21E+08	0.001	0.999	-4.32E+08	4.33E+08
mailer_C	3.09E+05	4.41E+08	0.001	0.999	-8.64E+08	8.65E+08
mailer_D	4.63E+05	6.62E+08	0.001	0.999	-1.30E+09	1.30E+09
mailer_F	6.17E+05	8.82E+08	0.001	0.999	-1.73E+09	1.73E+09
mailer_H	7.71E+05	1.10E+09	0.001	0.999	-2.16E+09	2.16E+09
mailer_J	9.26E+05	1.32E+09	0.001	0.999	-2.59E+09	2.60E+09
mailer_L	1.08E+06	1.54E+09	0.001	0.999	-3.03E+09	3.03E+09
mailer_X	0	0	nan	nan	0	0
mailer_Z	0	0	nan	nan	0	0

Note: coefficients with (*) are significant at 5% significance level.

To validate the model outcomes, residual analysis was conducted in both models.

For the linear regression model, the "Actual vs. Predicted" plot shows a significant clustering of predicted values around zero, regardless of the actual values, indicating potential issues with the model's ability to reflect the actual relationship. The line deviating significantly from the clustered points implies a lack of strong linear relationships in the dataset. If the residuals also deviate systematically from the line, it might indicate the model is underfitting or there is non-linearity. Hence, a residual plot was developed.

In the “Residuals Plot”, residuals are imbalanced and concentrated on one end. This indicates heteroscedasticity (non-constant variance of residuals), which violates a key assumption of linear regression.

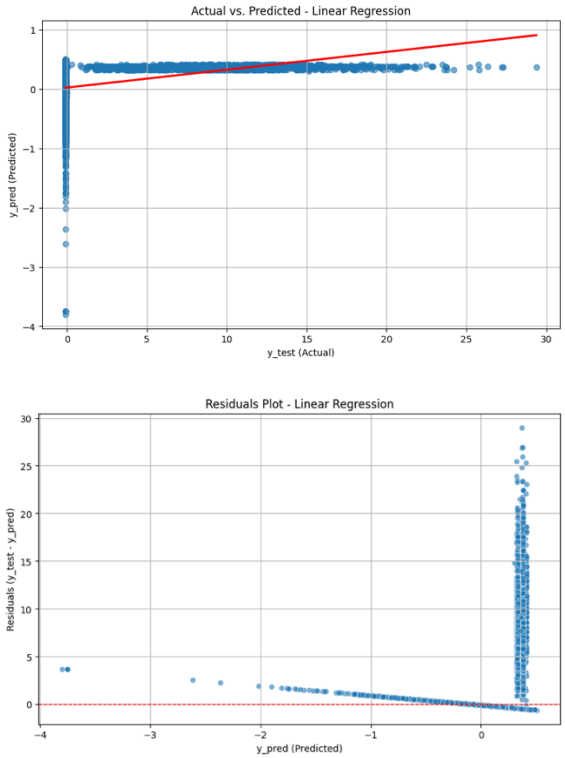

Figure 14 Actual vs Predicted Value (top) and Residual Analysis (bottom)

APPENDIX 6-2 – Random Forest Model

As for the random forest model, the "Actual vs. Predicted" plot shows that the predicted value scattering around the red line represents perfect predictions. For small values, the predictions are denser and closer to the red line. As for larger values, the predictions tend to spread more, showing the model struggles with extreme cases. In the “Residuals Plot”, ideally, residuals should be randomly distributed around zero without patterns. The residuals show clustering near zero for

mid-range predictions, indicating the model is accurate for this range. As for the smaller and larger values, the residuals skewed and spread more, especially in the larger cases. All in all, the RF here is effective at predicting most cases but struggles with outliers or extreme values, needed more attention.

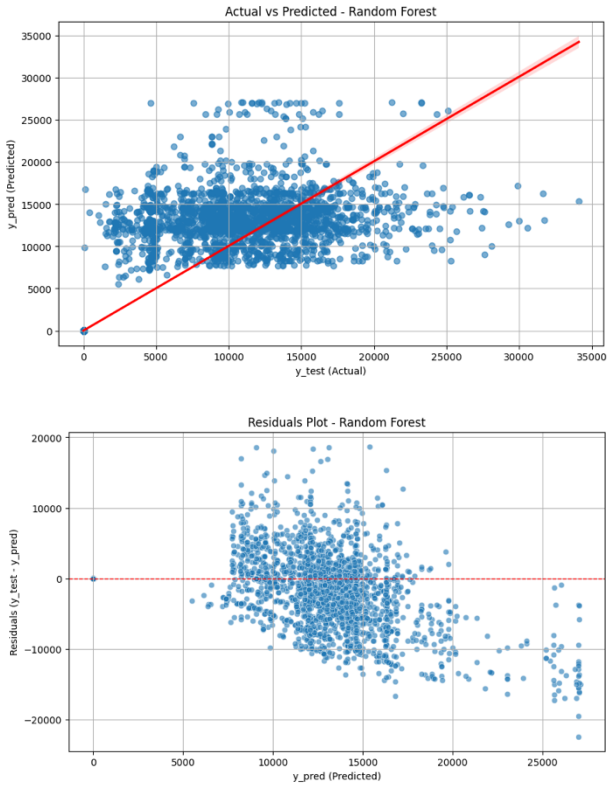

Figure 15 Actual vs Predicted Random Forest (top) and Residual Plot (bottom)