Table of Contents

Table of Figures

Table of Tables

Executive Summary

This report analyses the strategic impact of product placement in retail sales, intending to provide actionable insights to retailers for optimising retail marketing and operational strategies. It delves into how the spatial layout of products in-store and in-brochures influences customer behaviour and impacts sales volume. Using data from 2500 frequent shoppers, the analysis leverages advanced predictive modelling techniques, such as XGBoost, to predict patterns and forecast sales based on product placement.

Statistical analyses, including Kruskal-Wallis and Dunn's tests, show that in-store and in-brochure product placement significantly affect sales volume. Predictive modeling of these effects attains an accuracy of 75.02% and an R^2 of 0.85 after optimising the hyperparameters and addressing overfitting challenges, establishing a robust performance. Feature importance analysis highlights that interactions between display and mailer as well as display positioning in high-traffic areas are critical drivers of sales volume. In terms of departments, grocery and packaged meat are the predominant contributors to sales, implying that targeted placement strategies for these departments yield substantial benefits.

However, the model also has limitations, including incomplete data, lack of temporal trends and demographic factors, which may affect prediction accuracy. This analysis highlights the critical role of data-driven product placement strategies, presenting actionable insights to optimise sales performance and elevate customer satisfaction, thereby enabling retailers to secure a competitive advantage in the market.

1 Introduction

In the competitive retail sector, product placement is one of the most influential tools to attract customers' attention and influence their purchase decision. In fact, effective store layouts can lead to substantial sales increase by exposing customers to high-margin product areas (Han, et al., 2022). While physical display can promote impulse purchases or maximize high-margin product purchases, brochure placements can encourage customers to make purchase intentions before entering the store. Whether in-store or in-brochure, product placement is shown to have a significant effect on consumer behaviour (Kacen, et al., 2012). Thus, to influence consumer behaviour and increase profits, it is crucial for retailers to focus on the spatial layout of products in-store and in brochures (Gul, et al., 2023). The effectiveness of product placement depends on several factors, including the type of product, customer demographics, and purchasing trends (Li, et al., 2022). Having a thorough understanding of these criteria can support retailers in strategically matching their product placement strategies with consumer preferences.

To deepen that understanding, this project examines how product placement can be used as a sales prediction tool, with implications for operational and marketing decisions. Based on the dataset on household level transactions provided by the retailer, we aim to identify patterns and develop a predictive model, providing the retailer with data-driven insights. In doing so, this analysis enables the retailer to enhance customer satisfaction and boost profitability.

2 Problem Statement

To support retailers in choosing effective marketing strategies and maximizing sales, we analyse the relationship between product placement and sales volume, focusing on whether product placement can predict sales volume. Specifically, we aim to determine if the effect of product placement differs between in-store and in-brochure placement or between product categories. Additionally, we will provide recommendations for further data that could improve the analysis, important to enhance future data collection and prediction accuracy.

To quantify the impact of product placement on sales, we build a predictive model based on data on transaction behaviour of 2500 frequent shopper households, provided by the retailer. In developing the model, we rely on the transaction data, providing information on the quantity of a product sold, as well as product data, containing detailed information on each product including groupings of products at different levels. Particularly relevant for our analysis is the information on product placement in the causal dataset. For each product, the data indicates whether and where it was displayed in-store or in-brochure. With ten detailed categories for in-store and in-brochure placements, the specific effect of display or mailer categories on sales volume can be identified.

The objective of this analysis is to provide actionable recommendations supporting the retailer in making informed, data-based decisions in terms of product placement. Our findings will suggest concrete actions regarding optimizing in-store layouts with improved product positioning, refining the use of promotional resources, and developing targeted marketing strategies depending on product categories. Ultimately, based on the provided data, our model will support the retailer in gaining a competitive advantage and maximizing revenue by understanding the effect of product placement on sales volume, and therefore customer behaviour.

3 Literature Review and Justification for Methodology

3.1 Literature Review

Product placement refers to the strategic placement of products within store shelves with the aim to increase product accessibility and influence customer purchase behaviour. Sorensen (2009) highlights that optimal product placement in high-traffic regions can boost visibility, which yields an increase in the purchase ratio. Similarly, Underhill (2009) delves into consumer psychology and shelf placement delves, exploring how product location, packaging, and visual appeal impact consumers. The study shows an increasing likelihood of purchase by redirecting consumer focus through eye-catching arrangements and conspicuous placements. These findings underline the importance of spatial layouts in improving purchase rates. In recent years, advanced technologies like machine learning have been increasingly used to analyse customer data, predict behavioural patterns, and simulate product placement to improve sales outcomes. Machine learning models, as discussed by Chen and Guestrin (2016), serve as powerful tools to enhance predictions and optimise shelf layouts. This approach aligns closely with the objective of our analysis, seeking to optimise product placement based on data-driven learning. In retail analytics, the XGBoost algorithm has proven particularly effective. For instance, the study by Turgut and Erdem (2022) predicts sales for fruits and vegetables based on this gradient boosting algorithm. This study demonstrates the model’s ability to handle complex, non-linear relationships in retail sales data, making it a valuable tool for predictive modelling.

In addition to in-store placement, in-brochure advertising is crucial in influencing consumer behaviour. The study by Gijsbrechts et al. (2003) provides valuable insights into brochure placement strategies in supermarkets. The research highlights that strategically positioning products in brochures can result in a substantial increase in sales volume for the promoted products. While the literature extensively explores the effects of product placement on sales performance, its focus predominantly is on consumer psychology, i.e., measuring the effect of product placement on consumer behaviour. Few studies adopt predictive modelling approaches, bridging the gap between theoretical research and practical applications in the retail sector. Thus, this analysis aims to offer a predictive framework which combines theoretical concepts with sales data to optimize product placement strategies for retailers.

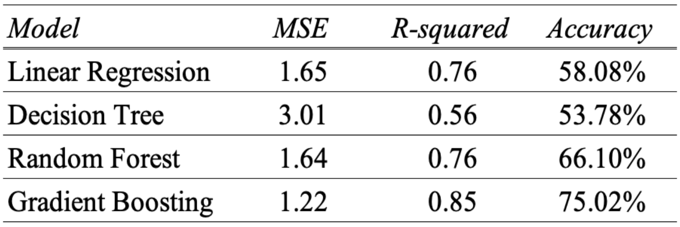
3.2 Pre-Processing of the Data

To ensure the dataset comprises only necessary information, we removed any redundant variables or observations. Specifically, the causal dataset contained approximately 30'000 duplicates, indicating different categories of display and mailer for the same product within the same store and week. While we assume that this is due to changes in product placement in that week, this assumption cannot be confirmed by the information provided. Therefore, the duplicates were dropped. To create a single dataset as the basis for the predictive model, the transaction, product, and causal data frames were merged. The number of variables was reduced to include only Product ID, Store ID, Week number, Quantity, Brand, Department, as well as display and mailer. Once merged, all rows with missing values for either display or mailer were removed, reducing the dataset from approximately 2.3 million observations to 482'705. Although a significant amount of data was lost, this step was necessary as display and mailer are essential for building the model. Furthermore, the data was aggregated at the product department level. This simplification was done to ensure the model's effectiveness as commodity descriptions and product-level data introduce noise and complexity beyond the scope of the modelling techniques. To further reduce the noise, we focused the analysis on the five departments with the highest quantity sold, grocery, drugs, packaged meat, produce and meat. Doing so allowed for a more meaningful analysis of departmental trends. To capture potential combined effects of departments and product placement, we introduced interaction terms. The interaction terms combined the departments with a binary variable defined for each display and mailer, indicating if the product was placed in-store or in-brochure or not. Finally, the target variable QUANTITY was log-transformed to reduce its right-skewness, variance and achieving a more normal distribution (Osborne, 2010).

3.3 Methodology

For an initial understanding of the relationship between product placement and sales volume, the categories for display and mailer were explored in terms of statistically significant differences in quantity sold. To select the appropriate statistical method, the data was assessed regarding normality and homogeneity of variances. The normality of the data distribution was tested using the Shapiro-Wilk test. The results showed a p-value of 0.00 for all categories in display and mailer, indicating a non-normal data distribution. The Q-Q Plots and distribution plots confirm this result, suggesting a right-skewed distribution of the data (see Appendix). Since none of the categories are normally distributed, non-parametric statistical methods seem more suitable. To check for homoscedasticity, we conducted Levene's test at the 5% significance level. For display, the p-value of 0.00 suggests unequal variance across the different categories. In contrast, for mailer, the p-value of 0.09 shows that the variances of the in-brochure categories are homogeneous. However, due to the non-normality of the data as well as the heteroscedasticity in the display variable, Kruskal-Wallis was deemed more appropriate to explore the relationship between product placement and sales volume as it is suitable for non-parametric data. The Kruskal-Wallis revealed statistically significant differences in the medians across the categories for both display and mailer, indicating that product placement affects sales volume. To determine which specific in-store or in-brochure placements significantly differ, we performed post-hoc pairwise comparison using Dunn's test.

Having confirmed a statistically significant effect of product placement on sales volume, a model was built to predict the sales volume based on specific product placement. The results of the initial data analysis suggest a non-linear, complex relationship between display and mailer, and quantity sold. Additionally, the data is mixed, including the categorical but not ordinal variables display and mailer. Thus, despite its interpretability, linear regression is not suitable to model the data. This is confirmed by the low accuracy result of 58.08% and the R^2 of 0.76 (see Table 1), implying that the predictive power of the model is not optimal.



Model	*MSE*	*R-squared*	*Accuracy*
Linear Regression	1.65	0.76	58.08%
Decision Tree	3.01	0.56	53.78%
Random Forest	1.64	0.76	66.10%
Gradient Boosting	1.22	0.85	75.02%

Table 1: Comparison of predictive model performances

While decision trees can handle the mixed data and capture non-linear relationships, the risk of overfitting is high, particularly with large data sets and few predictor variables (Kuhn & Johnson, 2013). With this sensitivity to data shape and its greedy approach, i.e., making locally optimal decisions at each split without considering the global optimum, decision trees cannot capture the complexity of the relationship in the data. This is represented in a low accuracy and R^2 as shown in Table 1. Nonetheless, a tree-based approach appears well-fitting due to its robustness, flexibility, and ability to identify non-linear patterns. Therefore, we attempted random forests. Random forests reduce model variance and prevent overfitting by only considering a random subset of predictors at each split in a tree, ensuring decorrelation between the trees. The method then combines multiple decision trees, each trained on bootstrapped samples of data, making the final prediction based on the majority vote across all trees (James, et al., 2021). However, while the model's performance increases to 0.76 for R^2 and 66.10% in accuracy (see Table 1), the predictions still are not very reliable or accurate. This limited predictive performance seems to be due to the highly complex data, which random forests fail to capture effectively. Additionally, the low number of features in the dataset limits the benefit of random feature selection, further limiting the model's effectiveness. To address these challenges and allow for more flexibility, gradient boosting was chosen as the final model. Instead of averaging predictions across multiple trees, gradient boosting grows trees sequentially (James, et al., 2021). Specifically, each tree is trained on the residual errors of the previous tree, allowing the model to focus on weak areas and slowly learning complex relationships within the data, gradually improving its predictions (Kuhn & Johnson, 2013). As we suspect highly complex, non-linear relationships within the dataset, this approach is well-suited. By tuning the parameters of the model via GridSearchCV and applying k-fold cross-validation, the XGBoost Regressor resulted in a well-performing and adaptive model with an R^2 of 0.85 and accuracy of 75.02%. Although the interpretability of the model itself is low, by running a feature importance analysis, the effect of different product placements in-store or in-brochure can be identified, providing valuable insights into product placement strategies.

4 Results, Limitations and Recommendations

4.1 Results

The initial analysis of the impact of product placement on sales volume showed a significant effect of the in-store display variable on sales. The Kruskal-Wallis test resulted in a p-value of 0.0004 and an H-statistic of 30.0582, indicating a statistically significant difference in the median quantity sold across categories of display, which underlines the importance of in-store product placement in driving sales. Similarly, for the mailer variable, the p-value of 0.0401 and H-statistic of 17.5976 suggest a statistically significant effect of in-brochure product placement on the quantity sold. However, comparing the H-statistic for display and mailer implies that the disparities between categories for display are more substantial than for mailer due to display's relatively high value. This was reflected in Dunn's test identifying the significant pairwise differences. For mailer, Dunn's test did not identify any significant disparities between categories, which is likely due to the small sample group size. Although prior research suggests that promotional mailers are one of the most effective advertisements (Gijsbrechts, et al., 2003), these results imply that the impact of in-brochure advertisement may be affected by unobserved factors.

To determine whether the product placement can predict sales volume, a gradient boosting approach was implemented using the XGBoost Regressor. The performance metrics of the first model included a Mean Squared Error (MSE) of 1.22 and an R^2 of 0.82 accounting for 82% of variance across the dataset. The accuracy revealed that the model correctly predicts 73% of the unknown data. While the model performance is good, it showed clear signs of overfitting when comparing the training with the validation error. The problem of overfitting was solved by adjusting the parameters of models via GridSearchCV and applying k-fold cross-validation. The final gradient boosting model achieved an MSE of 1.03, an R^2 explaining 85% of the variance and an accuracy score of 75.02%.

4.2 Feature Importance Analysis

To identify the main categories of in-store and in-brochure product placement driving sales volume, we conducted a feature importance analysis of the final XGBoost model. The results of the analysis as shown in Figure 1 are discussed below.

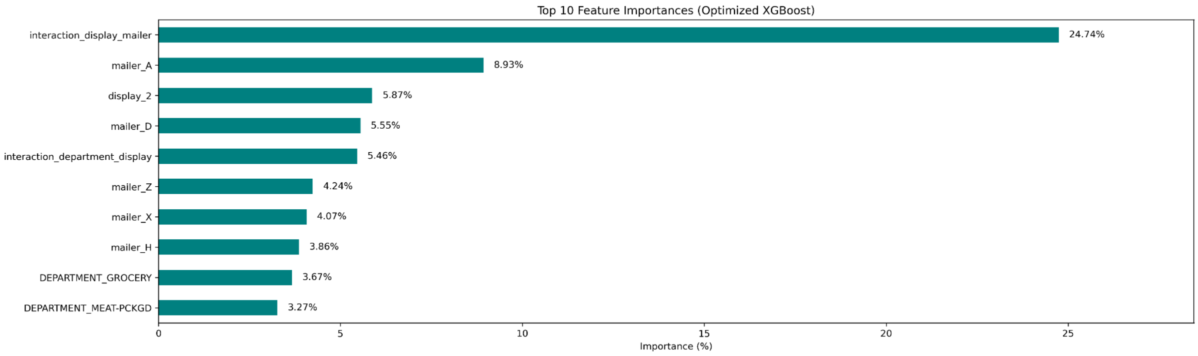


Figure 1: Feature importance analysis based on XGBoost

4.2.1 Interaction between Display, Mailer, and Department

The interaction between display and mailer was identified as the most important feature with a percentage of 24.743, suggesting that a strong in-store presence coupled with promotional mailing has a synergistic effect on consumers. Physical visibility and targeted marketing enhance product identification, increasing customers' likelihood to purchase them. The high importance score for this interaction emphasizes this its dominance in predicting sales volume. Additionally, the interaction between display and departments ranked amongst the most important features contributing 5.4567% to predictions. This indicates that the effectiveness of product placement in-store depends on the specific department. Thus, the positioning of display has a more pronounced effect on sales in departments generally attracting more customer attention.

4.2.2 In-Store Displays

For the display variable, surprisingly, the most important feature contributing to 5.8742% of predictions was the store rear. This might reflect high-demand or promotional products strategically being placed in the rear to draw customers into the store, thereby increasing their exposure to these displays. Additionally, generally the rear of stores is more structured and visually appealing compared to potentially cluttered and busy front areas, leading to higher interaction with displays in the rear. While store front is traditionally defined as the most important area for attracting customer attention, its significance in predicting sales volume is only at 1.5%. This could be due to customers being overexposed to displays at the front or the store fronts being cluttered with competing displays, reducing their individual effect on sales volume. Mid-aisle end caps, with a percentage of 2.57, likely gain importance due to the high visibility and traffic at this location. The remaining display categories only have limited predictive power and, therefore, are not analysed in detail.

4.2.3 Mailer Advertisements

The most important feature for mailer contributing to 8.93% of predictions, was the interior page feature with detailed promotions. The interior pages help customers make more informed purchasing decisions by providing more comprehensive product information. Moreover, by capturing attention first and deepening customer engagement with in-depth descriptions, interior features are likely to convert customer interest into sales. The second most impactful mailer category was the front page feature. These advertisements benefit from their prominent position, eliciting customers' initial attention when viewing the mailer. However, with a percentage of 5.55, the effect is less pronounced, probably because it only contains brief information, which might not be persuasive enough to induce purchasing intentions. Features related to free offers show a comparatively low effect but still contribute to sales predictions. Free promotions seem to have limited effectiveness on their own but could increase sales when combined with other incentives or promotional strategies, for instance, discounts for frequent customers. For more information on the feature importance of the remaining categories, refer to the Appendix.

4.2.4 Departments

To assess the effect of different product categories, i.e., departments, in predicting sales, grocery was among the ten most important features with a percentage of 3.67. Within this dataset, groceries proved to be more effective in increasing sales volume than other categories, possibly due to their regular consumption and stable repurchase rate. Therefore, they can be considered as a reliable contributor to overall sales turnover. The packaged meat department contributes to 3.27% of predictions and is one of the ten most important features. This importance reflects the year-round demand and convenience, implying that targeted promotions and strategic placement could further increase turnovers. In contrast, packaged meat and produce exhibit lower importance in sales predictions, likely due to their high-quality and premium-priced products which attracts a smaller customer segment and are purchased less frequently.

4.3 Limitations and Data Improvements

The analysis faced several limitations that should be addressed to enhance robustness and prediction accuracy. First, in pre-processing the data, due to missing information about product placement, 79.12% of the dataset was dropped. This likely resulted in a decrease in the model sensitivity to product placement effects. Additionally, 30'000 duplicates were removed which possibly corresponded to changes in the placement of specific products in a particular week, potentially introducing valuable information into the model. Secondly, by aggregating the data on department level, the analysis focused on a very high level of product categories, not taking into account more specific commodity descriptions. Moreover, the dataset omitted seasonal influences, which are key factors in the retail sector. Some products experience seasonal peaks in terms of sales volume. Ignoring these trends distorts the results. As such, including seasonal information could substantially improve prediction accuracy.
Furthermore, the dataset did not account for consumers' responses to specific promotional campaigns or the campaigns' interaction with product positioning strategies. Including campaign-specific data would also help refine the results obtained from a coordinated marketing campaign. Lastly, our predictive model did not include any demographics such as age, income, and buying patterns as components, thus not assessing how product placements impact different customer segments. Furthermore, other store-level characteristics such as

regional consumption preferences or the number of customers entering a store, were ignored. These factors could provide a more in-depth understanding of the differences in effectiveness of placements. For instance, store-front displays may have different effectiveness according to the store location.

4.4 Recommendations

Based on the analysis, actionable recommendations are provided to the retailer with the objective of optimizing the product placement strategies to increase sales volume. As shown in the feature importance analysis, combining product placement in-store and in-brochure substantially influences sales volume. Therefore, retailers should implement coordinated marketing campaigns, aligning mailer advertisement with effective in-store displays featuring the same products in both channels to increase their visibility and increase the quantity sold.

The feature importance analysis for display provides valuable insights into how store layouts can strategically be optimized to increase sales. As the store rear was shown to be the most effective location for display, high-demand or promotional items should be placed in the rear, drawing customers deeper into the store, and making use of the impact of that display on sales. As the second most impactful location, mid-aisle end caps should be used for positioning high-margin products, increasing their visibility in this high-traffic area. While the store front is somewhat impactful, simplifying the displays in the store front can enhance the impact by focusing on fewer top-performing products to avoid overwhelming customers.

In-brochure placements play a significant role in increasing sales volume. Specifically, interior pages have shown to be effective in influencing customer behaviour. Thus, detailed promotions should be featured on interior pages. To maximize the effect of these promotions, the front page should include teasers to capture customers' attention and guide them to more detailed promotions inside. Additionally, to increase the influence of free promotions, they could be combined with other promotional strategies such as loyalty discounts for better sales rates.

To achieve the most significant impact for product placement on sales performance, campaigns should be tailored to the high-performing departments, grocery and packaged meat. Specifically, high-performing grocery items and meat packages should be placed at store rears or mid-aisle end caps and be advertised on pages inside the brochure to maximize sales. To improve the sales performance of the unpackaged meat and produce departments, the high-quality, premium-priced products should be paired with complementary products in bundled discounts to attract price-sensitive customers.

For future data collection and research, the data integrity and sample size should be improved by collecting more observations and details on product placement categories. Moreover, considering seasonal trends, regional differences, customer demographics, and campaign-specific data can add further important features for predicting sales volume. Most importantly, analysing low-level product categories can lead to more detailed insights into the effects of product placement on sales volume.

5 Conclusions

This analysis report covers the critical role of data-driven strategies in optimising product placements in retail stores to refine sales rates and customer satisfaction. Through statistical evaluation and predictive modelling techniques, the significant influence of in-store displays on sales volumes, particularly in store rear and mid-aisle end caps, is highlighted. Mailer campaigns offer opportunities when targeted with detailed interior pages and promotional features.

The predictive model employing XGBoost demonstrated high accuracy and R^2, highlighting the effectiveness of machine learning techniques in uncovering complex relationships between product placement and sales. However, its performance is constrained by limitations such as incomplete data, the absence of temporal trends, and the lack of demographic variables, underscoring the need for future refinements.

Retailers are strategically encouraged to align in-store layouts and promotional efforts with high-performing departments to maximize sales. Merging displays with targeted mailer campaigns can amplify effectiveness. This study provides actionable admonitions, empowering retailers with strategies to gain a competitive edge through optimised and impactful product placement.

Bibliography

Chen, T. & Guestrin, C., 2016. Xgboost: A scalable tree boosting system. *Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining,* pp. 785-794.

Curhan, R. C., 1974. The effects of merchandising and temporary promotional activities on the sales of fresh fruits and vegetables in supermarkets. *Journal of Marketing Research,* 11(3), pp. 286-294.

Gijsbrechts, E., Campo, K. & Goossens, T., 2003. The impact of store flyers on store traffic and store sales: a geo-marketing approach. *Journal of retailing,* 79(1), pp. 1-16.

Gul, E., Lim, A. & Xu, J., 2023. Retail store layout optimization for maximum product visibility. *Journal of the Operational Research Society,* 74(4), pp. 1079-1091.

Han, Y., Chandukala, S. R. & Li, S., 2022. Impact of different types of in-store displays on consumer purchase behavior. *Journal of Retailing,* 98(3), pp. 432-452.

James, G., Witten, D., Hastie, T. & Tibshirani, R., 2021. *An Introduction to Statistical Learning: with Applications in R.* 2nd ed. New York: Springer US.

Kacen, J. J., Hess, J. D. & Walker, D., 2012. Spontaneous selection: The influence of product and retailing factors on consumer impulse purchases. *Journal of retailing and consumer services,* 19(6), pp. 578-588.

Kuhn, M. & Johnson, K., 2013. *Applied Predictive Modeling.* New York: Springer Nature.

Li, J., Guo, F., Xu, J. & Yu, Z., 2022. What influences consumers' intention to purchase innovative products: Evidence from China. *Frontiers in Psychology,* Volume 13, 838244.

Osborne, J., 2010. Improving your data transformations: Applying the Box-Cox transformation. *Practical Assessment, Research, and Evaluation,* 15(1).

Sørensen, E., 2009. *The materiality of learning: Technology and knowledge in educational practice.* s.l.:Cambridge University Press.

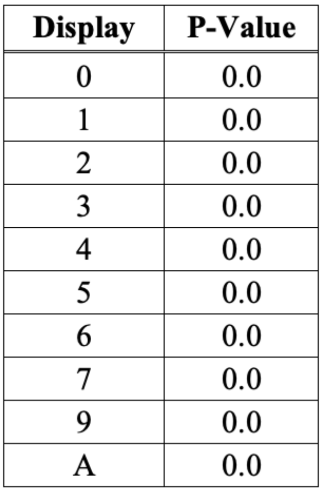
Turgut, Y. & Erdem, M., 2022. Forecasting of retail produce sales based on XGBoost algorithm. In: *Industrial Engineering in the Internet-of-Things World: Selected Papers from the Virtual Global Joint Conference on Industrial Engineering and Its Application Areas, GJCIE 2020, August 14–15, 2020.* s.l.:Springer International Publishing, pp. 27-43.

Underhill, P., 2009. *Why we buy: The science of shopping.* Updated and revised ed. New York: Simon & Schuster.

Appendix

Display – Shapiro-Wilk Normality Test

Table 2 below shows the p-values for the Shapiro-Wilk normality test for display. With a significance level of 5%, all p-values are below, indicating that the data is not normally distributed for all groups.

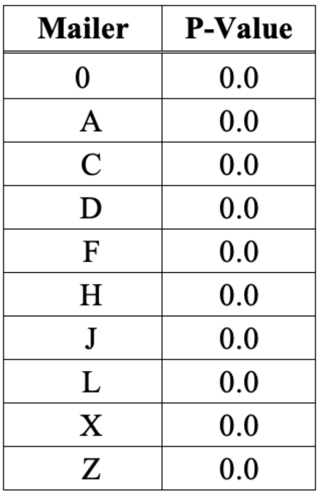


Display	P-Value
0	0.0
1	0.0
2	0.0
3	0.0
4	0.0
5	0.0
6	0.0
7	0.0
9	0.0
A	0.0

Table 2: p-values of Shapiro-Wilk Test for Display

Mailer – Shapiro-Wilk Normality Test

Table 3 below shows the p-values for the Shapiro-Wilk normality test for display. With a significance level of 5%, all p-values are below, indicating that the data is not normally distributed for all groups.



Mailer	P-Value
0	0.0
A	0.0
C	0.0
D	0.0
F	0.0
H	0.0
J	0.0
L	0.0
X	0.0
Z	0.0

Table 3: p-values of Shapiro-Wilk Test for Mailer

Display – Q-Q Plots and Distribution Plots

The Figure 2 below visualizes Q-Q plots and distribution plots for each category, providing insights into data distribution for the QUANTITY variable across groups. These plots are comparing the quantiles of the observed data with the theoretical quantiles of a normal distribution. The Q-Q plots reveal that the observed data across all groups deviate significantly from normality represented by the red diagonal line. This indicates that the data in all groups do not follow a normal distribution. The histograms with overlaid KDE curves show the frequency distribution of the quantity sold for each group . Most groups exhibit highly right-skewed distribution, with a larger concentration of data near lower values and long tails extending toward higher values.

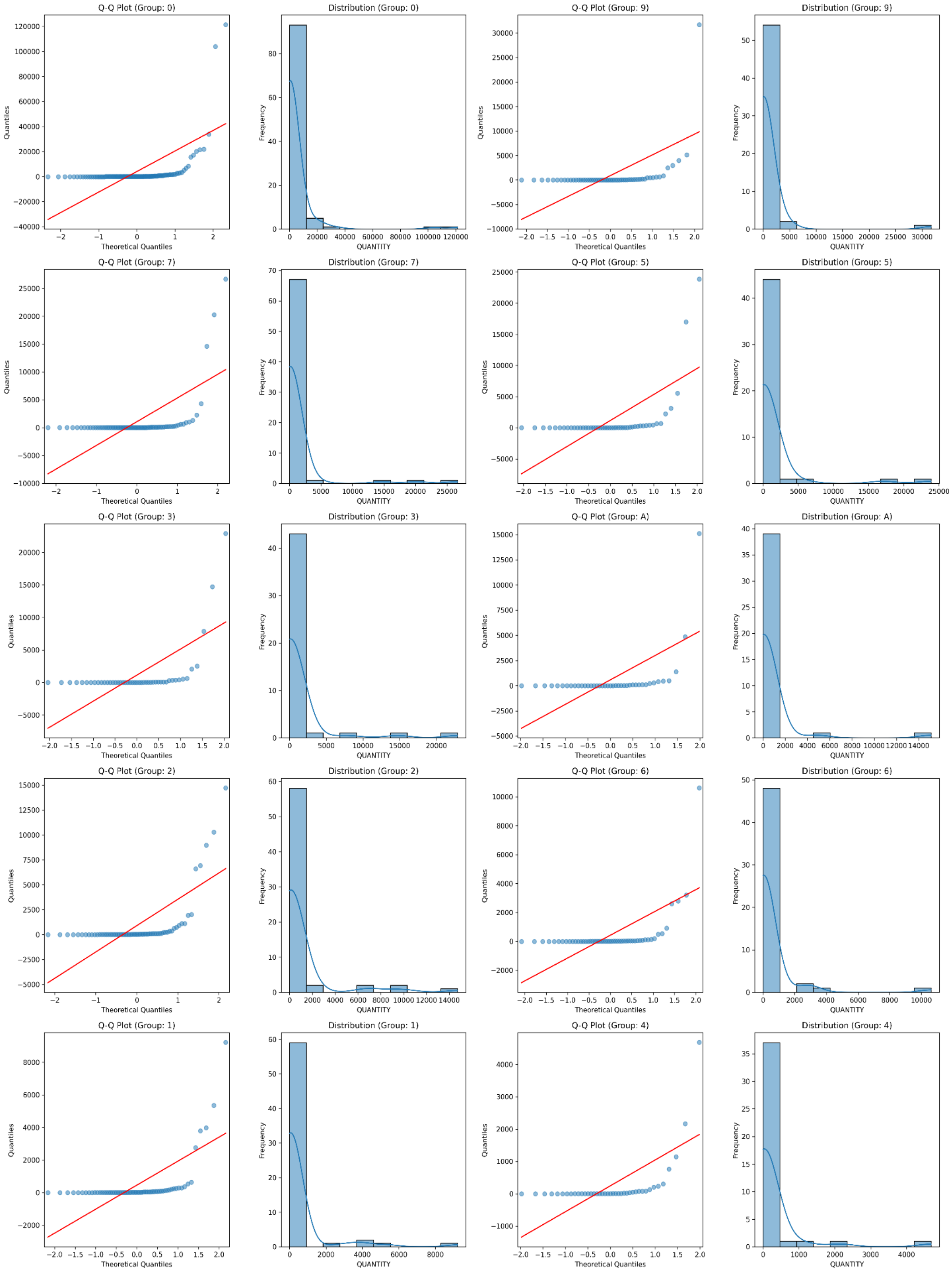


Figure 2: Q-Q Plots and Distribution Plots for Display Categories

Mailer – Q-Q Plots and Distribution Plots

Figure 3 shows Q-Q plots and distribution plots for QUANTITY across the different mailer categories. The visualizations highlight the data distribution and deviations from normality. The Q-Q plots suggest that the data for all mailer groups substantially deviates from a normal distribution. The Distribution Plots represented by the histograms paired with the KDE curve exhibit strong right-skewness across all mailer groups.

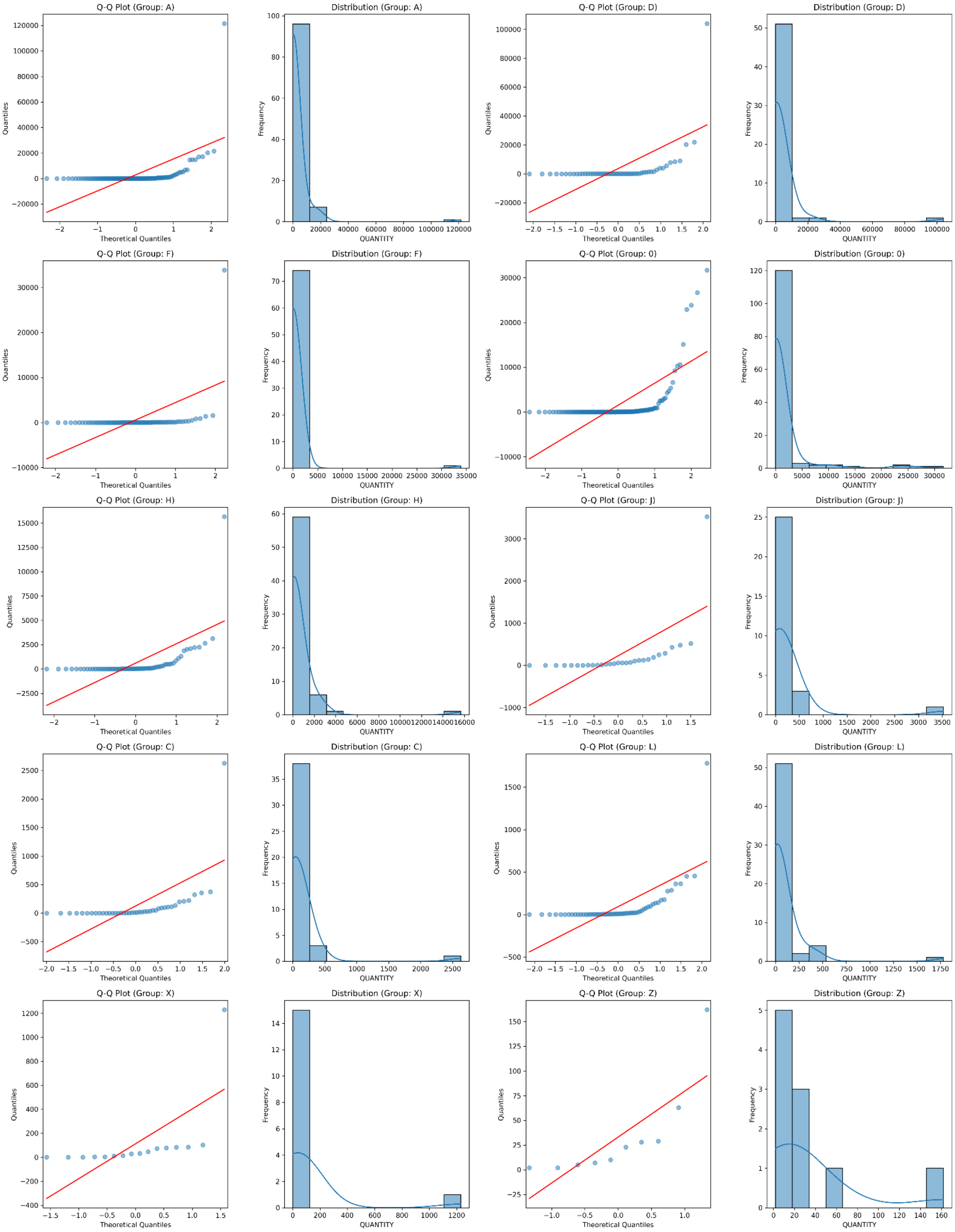
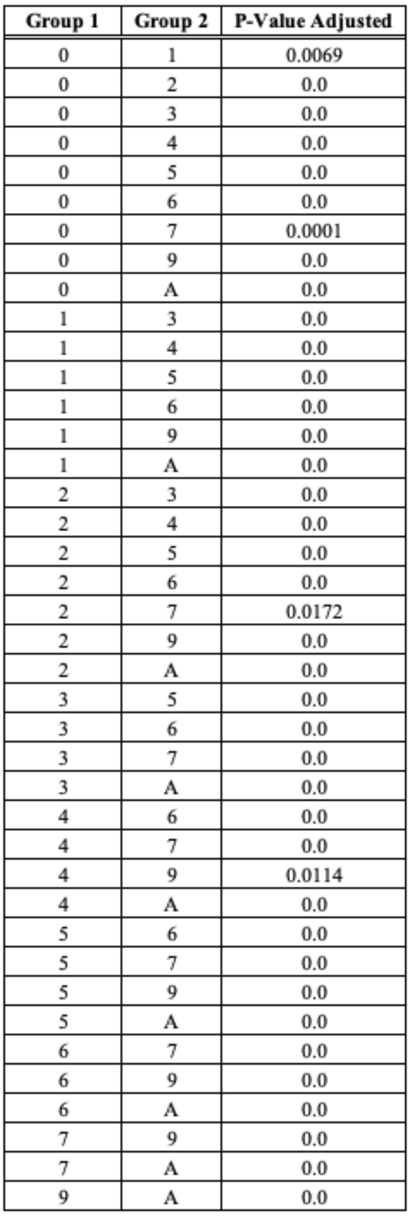


Figure 3: Q-Q Plots and Distribution Plots for Display Categories

Display – Dunn’s Test

Dunn’s test for post-hoc pairwise comparison checks for significant differences in the effect on sales volume between different groups. Table 4 shows all adjusted p-values below the 5% significance level, indicating a statistically significant difference for the groups listed.

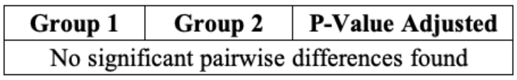


Group 1	Group 2	P-Value Adjusted
0	1	0.0069
0	2	0.0
0	3	0.0
0	4	0.0
0	5	0.0
0	6	0.0
0	7	0.0001
0	9	0.0
0	A	0.0
1	3	0.0
1	4	0.0
1	5	0.0
1	6	0.0
1	9	0.0
1	A	0.0
2	3	0.0
2	4	0.0
2	5	0.0
2	6	0.0
2	7	0.0172
2	9	0.0
2	A	0.0
3	5	0.0
3	6	0.0
3	7	0.0
3	A	0.0
4	6	0.0
4	7	0.0
4	9	0.0114
4	A	0.0
5	6	0.0
5	7	0.0
5	9	0.0
5	A	0.0
6	7	0.0
6	9	0.0
6	A	0.0
7	9	0.0
7	A	0.0
9	A	0.0

Table 4: Adjusted p-values for Dunn's Test for Display

Mailer – Dunn’s Test

Dunn’s test for post-hoc pairwise comparison checks for significant differences in the effect on sales volume between different groups. As Table 5 indicates no statistically significant differences have been found between the mailer categories.



Group 1	Group 2	P-Value Adjusted
No significant pairwise differences found		

Table 5: Adjusted p-values for Dunn's Test for Mailer

Performance Analysis for Optimised XGBoosting Model

The figure provides a visual analysis for the optimized XGBoost model performance, representing insights into residual patterns, prediction accuracy, and model learning behavior.

Residual Distribution

Figure 4 indicates the residuals are centered around zero and roughly follow a normal distribution. This aligns with the expectation of a well-fitted model.

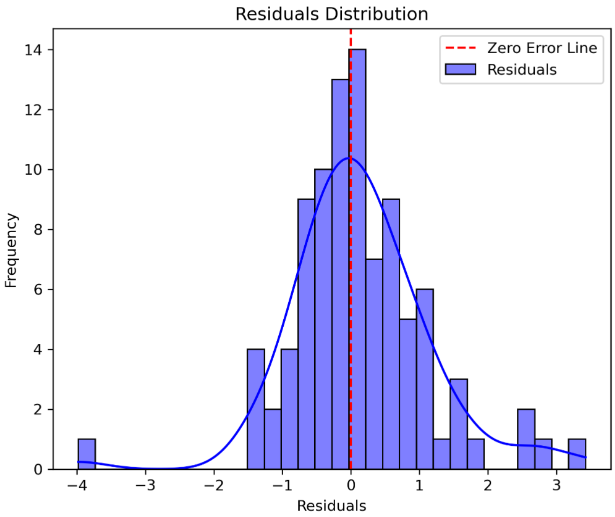


Figure 4: Residual Distribution

Residuals vs. Predicted Value

The scatterplot in Figure 5 visualizes the relationship between predicted and residual values. The residuals exhibit a relatively even spread without a systematic pattern, indicating that the model does not suffer from significant bias.

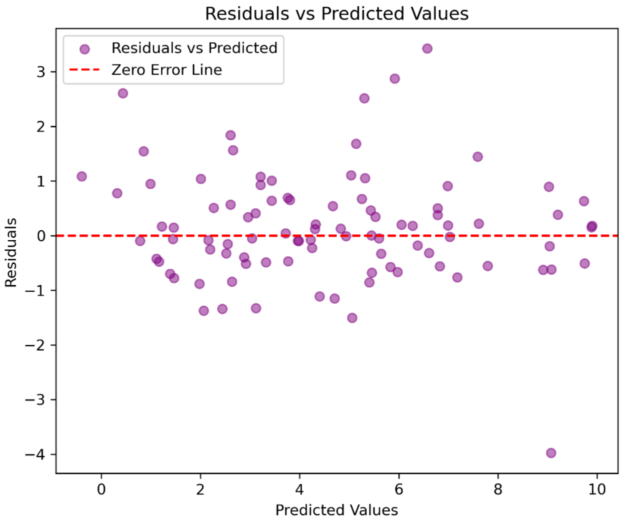


Figure 5: Residuals vs. Predicted Value

Predicted vs. Actual Values

Figure 6 compares the predicted values against actual values. The points in the plot are closely aligned along the diagonal, showing the high accuracy of the model.

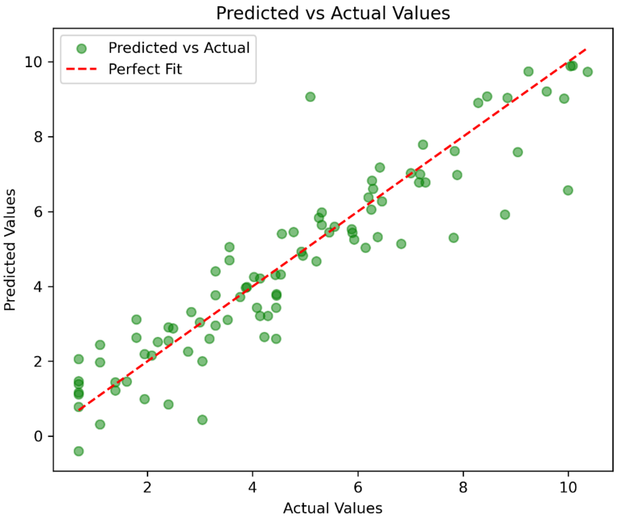


Figure 6: Predicted vs. Actual Values

Learning Curve

The training error as shown in Figure 7 is consistently lower than the validation error, which indicates a well-fitted model. A key observation in the plot is that as the training size increases, the validation error decreases, stabilizing as more data is utilized. This indicated that the model benefits from the additional training data.

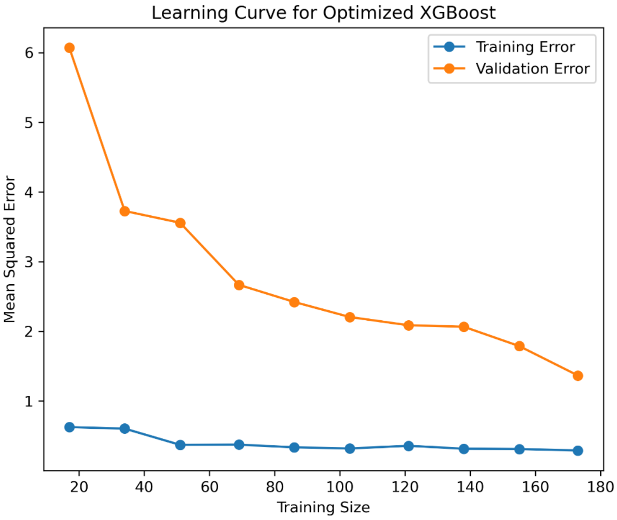


Figure 7: Learning Curve for Optimized XGBoost

Important Features for Predictions

Feature Importance for Display and Mailer

Figure 8 shows the feature importances for display and mailer, respectively.

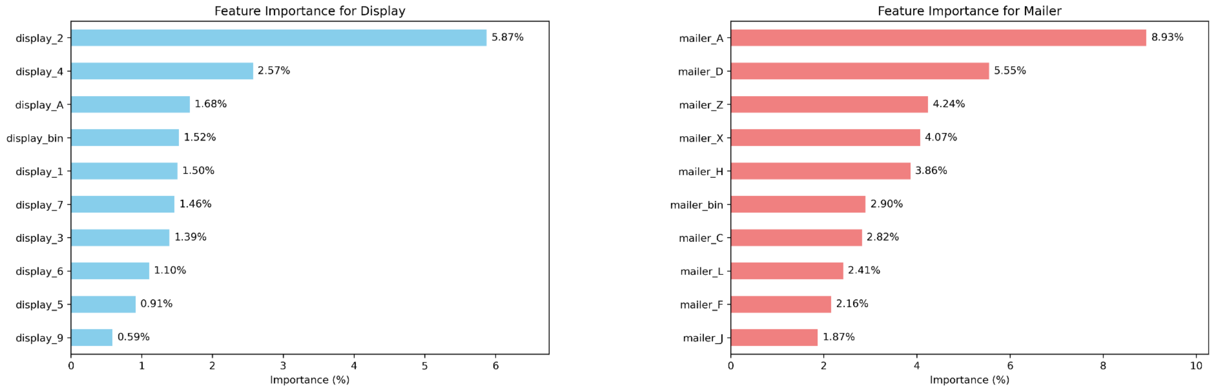


Figure 8: Feature Importances for Display and Mailer

Feature Importance for Department and Interaction Terms

Figure 9 shows the feature importances for departments and interactions, respectively.

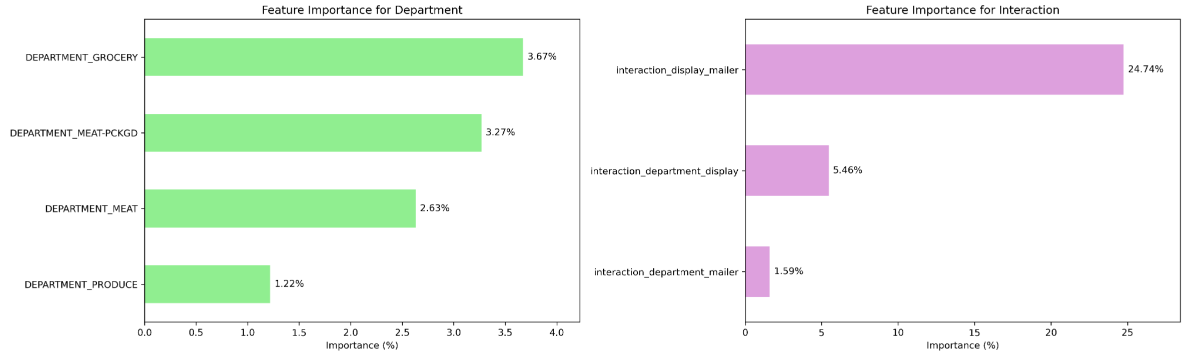


Figure 9: Feature Importance for Departments and Interaction Terms