POLITECNICO MILANO 1863

SCUOLA DI INGEGNERIA INDUSTRIALE
E DELL'INFORMAZIONE

Exploring the Inter-dependencies Between Marketing Channels: A Data-Driven Approach to Media Optimization

Tesi di Laurea Magistrale in Management Engineering / Mathematical Engineering

Ali Shyntas, 10832454. Ramazan Aitkaliyev 10758970

Advisor:
Lucio Lamberti

Co-advisors:
Gloria Peggiani

Academic year:
2023-2024

Abstract:

This thesis investigates the efficacy of Marketing Mix Modeling (MMM) in optimizing marketing investments across both traditional and digital channels, with the overarching goal of improving Return on Investment (ROI). A key focus of the research is the detailed analysis of interaction effects between various marketing channels, which fills a notable gap in current literature by shedding light on the synergistic and antagonistic outcomes that arise from combined media strategies. Through the application of advanced statistical methods and machine learning techniques, the study evaluates both short-term and long-term impacts on brand awareness, consumer behavior, and overall sales performance.

The research highlights the distinct roles of digital and traditional channels, where digital platforms tend to deliver immediate, measurable returns, while traditional media such as television contributes significantly to long-term brand building. By analyzing the interplay between these channels, the study reveals how their interactions can amplify or, in some cases, diminish marketing efficiency. The inclusion of saturation curves and the law of diminishing returns further refines investment strategies by identifying the optimal spending levels across each channel.

A critical contribution of this thesis lies in providing actionable insights for marketing budget allocation, aligning short-term digital gains with the enduring benefits of traditional media. By evaluating the interdependencies between channels, this study offers a more comprehensive understanding of how to allocate marketing resources for maximum overall impact. The findings underscore the importance of cross-channel optimization to enhance both marketing efficiency and long-term brand loyalty, offering practical recommendations for marketers to navigate an increasingly complex media landscape.

Key-words: Marketing Mix, MMM, Interaction Terms, media optimization, budget allocation.

Sommario

Questa tesi indaga l'efficacia del Marketing Mix Modeling (MMM) nell'ottimizzazione degli investimenti di marketing su canali tradizionali e digitali, con l'obiettivo di migliorare il ritorno sugli investimenti (ROI). Un elemento chiave della ricerca è l'analisi dettagliata degli effetti di interazione tra i diversi canali di marketing, colmando una lacuna nella letteratura attuale e fornendo approfondimenti sugli esiti sinergici o antagonistici delle strategie mediali combinate. Utilizzando metodi statistici avanzati e tecniche di apprendimento automatico, lo studio valuta sia gli impatti a breve termine che quelli a lungo termine sulla consapevolezza del marchio, sul comportamento dei consumatori e sulle prestazioni di vendita.
La ricerca evidenzia i ruoli distinti dei canali digitali e tradizionali, dove le piattaforme digitali tendono a offrire ritorni immediati, mentre i media tradizionali come la televisione contribuiscono alla costruzione del brand nel lungo periodo. L'analisi dell'interazione tra questi canali rivela come tali combinazioni possano amplificare o, in alcuni casi, ridurre l'efficienza del marketing. L'inclusione delle curve di saturazione e della legge dei rendimenti decrescenti affina ulteriormente le strategie di investimento, identificando i livelli di spesa ottimali tra i vari canali.
Un contributo critico di questa tesi risiede nel fornire spunti pratici per l'allocazione del budget, allineando i guadagni digitali a breve termine con i benefici duraturi dei media tradizionali. Valutando le interdipendenze tra i canali, lo studio offre una comprensione più completa di come allocare le risorse di marketing per ottenere il massimo impatto complessivo. I risultati sottolineano l'importanza dell'ottimizzazione cross-canale per migliorare sia l'efficienza del marketing che la fedeltà al marchio nel lungo termine, offrendo raccomandazioni pratiche per i marketer che operano in un panorama mediale sempre più complesso.

Key-words:

Marketing Mix, MMM, Termini di interazione, ottimizzazione dei media, allocazione del budget.

Acknowledgements

We would like to express our deepest gratitude to our adviser, Professor Lucio Lamberti, and our co-adviser, Gloria Peggiani, for their invaluable support and guidance throughout this journey. Professor Lamberti introduced us to the Marketing Analytics course, opening up new perspectives on the field of marketing that we hadn't considered before. His insights broadened our understanding and inspired us to delve deeper into the complexities of multi-channel marketing.
We are especially thankful to co-adviser Peggiani, who patiently and thoroughly guided us through every stage of this research. Her detailed explanations and unwavering encouragement helped us navigate challenges and directed us toward the most effective paths in our work.
We are also profoundly grateful to our families and friends for their unwavering support. Their encouragement and understanding have been a constant source of strength, motivating us to persevere even when faced with obstacles. Without their love and support, this journey would not have been possible.

Contents

1. Introduction

The rapid evolution of the marketing landscape has driven businesses to adopt increasingly complex strategies involving a multitude of digital and traditional channels. With this complexity comes the need for more advanced analytical approaches to optimize marketing investments. Marketing Mix Modeling (MMM) has emerged as a critical tool for evaluating the effectiveness of various channels and their contribution to overall business outcomes. However, much of the existing research focuses on the individual performance of marketing channels, neglecting the significant impact of interaction effects between them. This thesis aims to address this gap by exploring the interplay between digital and traditional media, offering insights into how these interactions can either amplify or hinder campaign effectiveness.

Marketing analytics plays a central role in this process, providing the framework through which data-driven insights are derived to support informed marketing decision-making. The systematic process of marketing analytics includes identifying key marketing problems, collecting relevant data, selecting appropriate Key Performance Indicators (KPIs), and using advanced programming tools such as R, Python, Power BI, and Tableau to analyze the data. These steps enable businesses to optimize their marketing strategies by offering insights into customer behavior, trends, and forecasting. However, as this thesis will demonstrate, the full potential of marketing analytics is realized when interaction effects between various channels are taken into account, especially when using MMM to inform budget allocation decisions.

One of the key challenges marketers face is determining how to allocate budgets across multiple channels while considering the interdependencies between them. Channels often do not operate in isolation; instead, they interact in ways that can either enhance or diminish the overall effectiveness of a campaign. For example, while digital platforms like social media or display ads may provide immediate returns on investment (ROI), their combined effect with traditional media such as television may lead to either a synergistic boost in brand awareness or, conversely, audience fatigue due to overexposure. Understanding these dynamics is essential for optimizing media spending and ensuring that campaigns deliver maximum impact. Effective marketing analytics enables this by evaluating not only the individual performance of channels but also how their interactions influence customer engagement and purchase decisions.

This research goes beyond the conventional scope of evaluating individual channels in isolation, introducing a detailed analysis of how different channel combinations perform when used together. The study incorporates advanced statistical methods to assess interaction effects, demonstrating how channels like Programmatic Video, YouTube ads, and Influencers interact in both positive and negative ways. By including these interaction terms, the model not only improves its predictive accuracy but also offers a more nuanced understanding of how marketing channels work together. This is crucial in today's environment where data overload can be a challenge, and marketing analytics tools must be leveraged to distill actionable insights from vast datasets.

Moreover, this thesis contributes to the broader field of marketing analytics by examining the balance between short-term and long-term impacts. While digital channels often provide rapid returns, traditional channels tend to build brand equity over time. The integration of these insights offers marketers a more holistic perspective on how to plan and execute multi-channel campaigns, allowing for a more strategic alignment of online and offline activities. The process of marketing analytics supports this by enabling companies to monitor and continuously improve their campaigns, ensuring both immediate ROI and sustained brand loyalty over time.

In summary, this thesis seeks to deepen the understanding of cross-channel interactions within the context of Marketing Mix Modeling. By analyzing the synergistic or counterproductive effects of combined channel strategies, it provides actionable insights that marketers can use to optimize their media investments. The research highlights the importance of not only measuring the individual performance of marketing channels but also considering how their interactions influence overall campaign success.

Through the application of robust marketing analytics techniques, this analysis offers a roadmap for improving both short-term sales and long-term brand loyalty through more informed and precise budget allocation.

2. Theoretical Background

2.1. Marketing Mix Modeling

Marketing Mix Modeling (MMM) is a statistical approach that quantifies the impact of several marketing inputs on sales or market share. The process involves the use of regression on sales and marketing time series data to estimate the impact of various marketing tactics (marketing mix) on sales and then forecast the impact of future sets of tactics. It is often used to optimize advertising mix and promotional tactics with respect to sales revenue or profit.
The famous quote by John Wanamaker, "Half the money I spend on advertising is wasted; the trouble is I don't know which half," encapsulates the dilemma faced by many marketers and underscores the importance of effective marketing mix modeling [73].

2.1.1 Historical Development of Marketing Mix Modeling

The concept of MMM was developed in the early 1970s by statisticians at the University of Chicago to measure the impact of marketing on sales. Initially based on regression analysis, MMM has evolved with the rise of digital marketing in the 2000s. Today, advancements in cloud computing and machine learning have made MMM more cost-effective and continuously updated, allowing for powerful insights without hefty price tags or privacy violations. MMMs use aggregate historical time series data to model sales outcomes over time, considering advertising variables, other marketing variables, and control factors. This evolution has made MMM more accessible and relevant for companies of all sizes, enabling them to optimize their marketing strategies and achieve higher ROI [74].

2.1.2 Understanding the Role and Benefits of Marketing Mix Modeling

From a managerial perspective, MMM offers a comprehensive understanding of the relationships between marketing channels and target metrics, enabling the identification of high ROI channels, optimization of marketing budgets, and prediction of future conversions. This approach equips brands with the ability to make informed decisions regarding their marketing strategies, ensuring that resources are allocated to the most impactful channels.
For instance, a leading consumer goods company utilized MMM to optimize its marketing spend by identifying the most effective channels for driving sales. By leveraging the insights derived from the model, the company achieved a substantial increase in ROI and enhanced the overall effectiveness of its marketing campaigns. Similarly, a global electronics manufacturer employed MMM to streamline its promotional strategies, resulting in improved customer engagement and higher conversion rates [75].
The benefits of MMM include:

- **Channel optimization**: The identification of the most effective marketing channel (TV, online, print, radio, etc.) to accomplish marketing goals and maximize returns (improve ROI).
- **Budget Allocation**: Efficient budget allocation between channels to prevent saturation.
- **Testing different scenarios**: Predict business metrics based on marketing activities and test different business scenarios for improving efficiency.

2.1.3 Impact of Variables on Marketing Mix Models

Marketing mix elements can be categorized into three variables: incremental, base, and other, each comprising several factors that influence a product's market performance. Understanding these variables is essential for marketers to make precise forecasts regarding the impact of promotional activities and product distribution.
Base variables represent the impact achieved independently of marketing mix factors, influenced by various elements such as brand value, seasonality, and non-marketing factors like GDP, growth rate, and consumer sentiment.

- **Price**: Influences the target consumer group, advertising strategy, and distribution approach. The pricing model directly impacts business performance and communicates product value to customers.
- **Distribution**: Refers to the availability of the product in stores, assortment, and shelf life. An effective distribution strategy, coupled with targeted marketing activities, directly impacts business outcomes and provides consumers with multiple purchasing options.
- **Seasonality**: Involves periodic variations, offering significant commercial opportunities, such as the peak in electronics sales during the holiday season.
- **Macro-economic variables**: Factors like GDP, unemployment rate, and consumer sentiment, which are beyond businesses' control but substantially impact their marketing strategies.

Incremental Variables of marketing mix elements fall into three categories:

1. **ATL (Above-the-Line) Marketing**: Non-targeted advertising with wide reach, aiming to build brand recognition and consumer awareness through mediums like television, radio, print, and product placements.
2. **BTL (Below-the-Line) Marketing**: Specific, direct advertising focused on targeted consumer groups, emphasizing conversions and effectiveness measurement through activities like sales promotions, social media marketing, and events.
3. **TTL (Through-the-Line) Marketing**: Integration of both ATL and BTL strategies for improved results, including 360° marketing campaigns developed for brand building and conversions.

Long-term impacts of marketing initiatives can be attributed to a range of factors, including competition, halo and cannibalization effects, and emerging variables such as product/market trends, product launches, events and conferences, behavioral metrics, and social metrics. Understanding these variables is essential for comprehending the dynamic marketing environment and effectively engaging with customers, especially millennials, through various platforms and channels. This comprehensive understanding of the impact of variables on marketing mix models enables marketers to adapt their strategies and make informed decisions to drive business growth and success.

2.1.4 Defining KPI's for Marketing Mix Modelling

1. Understand business objectives and align KPIs with the overall business goals, such as revenue growth, market share expansion, profit maximization, customer acquisition, and brand awareness.
2. Align KPIs with Marketing Mix elements, including the 4Ps:
   - **Product KPIs**: Product-specific sales, market share, product-level profitability.
   - **Price KPIs**: Pricing elasticity, price premium, discount effectiveness.
   - **Promotion KPIs**: Return on advertising spend (ROAS), marketing ROI, campaign attribution.
   - **Placement KPIs**: Distribution reach, shelf share, channel effectiveness.
3. Ensure measurability and data availability: Ensure that KPIs can be accurately measured using quantitative data and that the necessary data is available for analysis.
4. Focus on Actionability and Insights: Select metrics that provide relevant information to optimize marketing strategies and budget allocation, enabling actionable insights.
5. Account for Attribution: Consider attribution models to accurately assign credit to various marketing channels or touchpoints along the customer journey, ensuring fair assessment of their contribution to business outcomes.
6. Include Leading and Lagging Indicators: Incorporate both leading indicators, which provide early insights into future performance, and lagging indicators, which reflect past performance, to gain a comprehensive understanding of marketing effectiveness.
7. Regularly Review and Adjust KPIs: Continuously monitor KPIs and be prepared to adjust them based on changes in the market landscape or business priorities to maintain relevance and effectiveness.

2.1.5 Marketing Mix Variables Categorization [76]

- Baseline: Base price, Average sales price, Assortment, Velocity, Distribution.
- Promotions: Sales promotion
- Discounts: Average Price Discount, Weighted Discount.
- Seasonality & Holidays:
- Media Activity: TV spends, Reach, Frequency, TV GRP, Digital Spends, Digital Impressions, Digital Clicks, Search spends, Search impressions, Print spends, Radio spends
- Competition: Base, Media activity, offers, Discounts.
- Others:Social media, External factors, Trend, Cyclicity, Events and Launches

2.2. Key Metrics in Marketing Mix Modeling

2.2.1 Introduction

Marketing mix modeling is an analytical approach used to quantify the impact of various marketing tactics on sales and then forecast the impact of future sets of tactics. It involves three core phases: measurement, optimization, and forecasting. Each phase requires specific metrics to be computed for a comprehensive understanding of marketing effectiveness.

2.2.2 The Significance of Contribution

Contribution is a pivotal metric in marketing mix modeling. It represents the impact a particular factor has on a key performance indicator (KPI). For instance, if paid search is found to contribute 22,600 units to sales, it signifies that the investment in paid search activities has resulted in that number of units sold. Contribution can be expressed in absolute terms, like the number of units, or as a percentage of total sales. Mathematically, the contribution of a factor can be calculated as:

$$\text{Contribution} = \frac{\text{Impact} \times \text{Units}}{\text{Total Sales}}$$

2.2.3 Visualizing Impact with Layer and Waterfall Charts

Layer charts, also known as contribution charts, visually display the contribution of each factor to the overall model. When all contributions are summed, they equate to the actual modeled sales, plus or minus the error of measurement. Waterfall charts are instrumental in analyzing period-over-period changes, such as understanding the drivers behind a 10% increase in sales from one semester to another.

2.2.4 From Unit Sales to Revenue: The Conversion Process

Converting unit sales into revenue is a crucial step in marketing mix modeling. This can be done by multiplying the incremental contribution by the average price over the fitted period. For industries like telecommunications, insurance, or banking, revenue takes on a different meaning due to factors like service renewals. In such cases, computing the lifetime value of a customer becomes essential. The formula for this conversion is:

$$\text{Revenue} = \text{Unit Sales} \times \text{Average Price}$$

2.2.5 ROI: The Ratio That Reveals Marketing Efficiency

Marketing ROI is the ratio of incremental revenue to marketing spend. It's vital to express both components in monetary terms to accurately assess the return on investment. A higher ROI indicates more efficient use of marketing dollars, but it doesn't always mean that more should be invested in that channel, especially if it's already saturated. The formula for ROI is:

$$\text{ROI} = \frac{\text{Incremental Revenue}}{\text{Marketing Spend}}$$

2.2.6 Ad Stock and Half-Life: Understanding Advertising Persistence

Ad stock is a concept that incorporates the carryover effect of advertising. The half-life value is the time it takes for the ad stock to reduce by half, which helps brands determine when to reinvest in advertising to maintain consumer awareness. The formula for calculating the half-life of an ad is:

$$\text{Half-Life} = \frac{\ln(2)}{\text{Decay Rate}}$$

2.2.7 Effectiveness vs. Efficiency: Isolating Media Inflation

Effectiveness measures the ratio of incremental revenue to GRPs for offline media or to millions of impressions for digital channels, isolating the impact of media inflation. Efficiency, though less commonly computed, looks at the ratio of percentage contribution from total media to spend, providing insights into the value each dollar brings in terms of media contribution. The formulas for effectiveness and efficiency are:

$$\text{Effectiveness} = \frac{\text{Incremental Revenue}}{\text{GRPs}}$$

$$\text{Efficiency} = \frac{\text{Percentage Contribution from Total Media}}{\text{Spend}}$$

2.2.8 Optimization and the Law of Diminishing Returns

The backbone of optimization in marketing mix modeling is the diminishing returns curve, which illustrates the concave relationship between spend and incremental revenue. As spend increases, additional revenues grow at a decreasing rate, eventually reaching saturation where additional spend yields no further gains. This can be represented by the formula:

$$\text{Incremental Revenue} = a \times \text{Spend}^b$$

where a and b are constants, and $b < 1$ represents the diminishing returns.

2.2.9 Marginal Returns and Optimal Execution Range

Marginal returns refer to the additional revenue generated by an incremental investment. The optimal execution range, applicable in the context of an S-shaped diminishing returns curve, defines the recommended spend range for a channel, balancing between maximum marginal ROI and maximum ROI. The formula for marginal returns is the derivative of the incremental revenue with respect to spend:

$$\text{Marginal Returns} = \frac{d(\text{Incremental Revenue})}{d\text{Spend}}$$

In conclusion, marketing mix modeling is a sophisticated tool that enables marketers to dissect and understand the impact of their efforts. By mastering the metrics and concepts outlined in this text, businesses can optimize their marketing strategies for maximum effectiveness and efficiency. The added mathematical theory and formulas provide a deeper understanding of the underlying principles.

2.3. Media Transformation

The video from Mass Analytics discusses two significant transformations in marketing mix modeling: decay or ad stock transformation, and diminishing returns.

2.3.1 Decay or Ad Stock Transformation

This transformation accounts for the prolonged effect of advertising on consumer behavior. The impact of an advertisement doesn't disappear immediately but decays over time. This is represented by applying a decay rate to the media variables. For example, a 30% weekly decay means that 30% of the advertising message is lost each week. Analysts create different decay versions (10%, 20%, 30%, etc.) to determine which decay rate most significantly affects sales and consumer response. The mathematical representation of this decay is given by:

$$AdStock_t = AdSpend_t + (1 - DecayRate) \times AdStock_{t-1}$$

where $AdStock_t$ is the ad stock at time t, $AdSpend_t$ is the ad spend at time t, and $DecayRate$ is the decay rate.

2.3.2 Diminishing Returns

As the marketing spend increases, the resulting sales or revenue also grows but at a decreasing rate until it reaches a saturation point. This concept is crucial for optimization exercises, where the goal is to reallocate budget from highly saturated channels to less saturated ones, thereby enhancing revenue without increasing the overall budget. Mathematical transformations like the exponential function or the adstock model are used to shape the diminishing returns curve. This can be represented by the formula:

$$Sales = a \times (1 - e^{-b \times Spend})$$

where a and b are constants, and e is the base of the natural logarithm.

2.3.3 Optimal Execution Range

One of the key insights from the application of S-shaped curves is the identification of the optimal execution range for investment in a particular channel. This range lies between the maximum marginal ROI and the point where additional spending no longer improves the average ROI. Understanding this range helps in making informed decisions about budget allocation across different media channels. The formula for marginal returns is the derivative of the sales with respect to spend:

$$MarginalReturns = \frac{d(Sales)}{d(Spend)} = a \times b \times e^{-b \times Spend}$$

In conclusion, these transformations allow marketers to optimize their media spend effectively, ensuring that each dollar invested contributes to the brand's growth and success.

2.4. Smoothing Techniques for Marketing Mix Modeling

Data smoothing is a crucial step in marketing mix modeling, particularly when dealing with sales data that exhibits seasonality. This section will delve into the mathematical theory and practical examples of data smoothing techniques, including moving averages and median filters, as well as the creation and refinement of seasonal variables.

2.4.1 Moving Average Technique

The moving average technique is a common method for smoothing data. It involves creating a series of averages of different subsets of the full data set. Mathematically, a simple moving average (SMA) for a given time period is calculated as:

$$SMA = \frac{1}{n} \sum_{i=1}^{n} P_i$$

where P_i is the price in period i and n is the number of periods.
Two types of moving averages are often used: standard and centered. The standard moving average takes a sliding window of data points and computes their average. In contrast, the centered moving average places the window around the data point of interest, taking an equal number of data points on either side.
The size of the window, or the number of data points included in each average, impacts the smoothness of the resulting data. A larger window will result in a smoother data series, but it may also obscure smaller, yet potentially important, fluctuations.

2.4.2 Median Filter

The median filter is another technique for data smoothing. It is particularly useful when the data contains outliers, as it is less sensitive to extreme values than the moving average. The median filter works by replacing each data point with the median of neighboring data points within a sliding window. Mathematically, for a given window size n, the median filter is defined as:

$$M_i = median(P_{i-(n-1)/2}, ..., P_i, ..., P_{i+(n-1)/2})$$

where P_i is the price in period i and M_i is the median price in the window around period i.

2.4.3 Seasonality in Marketing Mix Modeling

Seasonality plays a significant role in marketing mix modeling, especially for products or brands that experience cyclical sales patterns. Seasonal variables can be created by averaging sales data year-on-year to detect and quantify these patterns.
For instance, if we denote $S_{i,j}$ as the sales in month j of year i, a simple seasonal variable for month j could be calculated as:

$$SV_j = \frac{1}{N} \sum_{i=1}^{N} S_{i,j}$$

where N is the number of years of data available.
Sales and competitor data can further enhance the analysis of seasonality, providing additional context and helping to explain observed patterns.

2.4.4 Refining Seasonal Variables

Once seasonal variables have been created, they can be further refined using smoothing techniques such as the moving average or median filter. This helps to eliminate any noise or fluctuations due to randomness or anomalies in the data, improving the accuracy of the marketing mix model.
For example, a smoothed seasonal variable SSV_j could be calculated using a centered moving average:

$$SSV_j = \frac{1}{n} \sum_{k=j-(n-1)/2}^{j+(n-1)/2} SV_k$$

where n is the window size and SV_k is the seasonal variable for month k.
In conclusion, data smoothing techniques are essential tools in marketing mix modeling, helping to clarify underlying trends, account for seasonality, and improve the accuracy of the resulting models.

2.5. Bayesian Ridge Regression: Theory, Formulas, and Implementation

2.5.1 Overview of Bayesian Ridge Regression

Bayesian Ridge Regression is an extension of Ridge Regression that introduces a Bayesian framework for parameter estimation. Ridge regression is used primarily to handle multicollinearity by adding a penalty term on the size of the coefficients. This penalty reduces the impact of correlated variables by shrinking the coefficients, which helps to stabilize the model and avoid overfitting.
Bayesian Ridge Regression incorporates prior distributions for the parameters, combining prior beliefs with the observed data to estimate a posterior distribution for the coefficients. The Bayesian approach provides a probabilistic interpretation, allowing the model to incorporate uncertainty in the estimates, which is crucial in cases like Marketing Mix Modeling (MMM) where multicollinearity is common due to correlated media channels.

2.5.2 Mathematical Formulation

In Bayesian Ridge Regression, the coefficients w are assumed to follow a prior normal distribution:

$$w \sim \mathcal{N}(0, \lambda^{-1} I)$$

where λ is a regularization parameter that controls the prior variance of the coefficients. This prior indicates a belief that the coefficients are close to zero, introducing a form of regularization.
The likelihood of the data given the coefficients follows a Gaussian distribution:

$$y \sim \mathcal{N}(Xw, \alpha^{-1})$$

where X is the design matrix of input features (e.g., marketing spend across different channels), and α is the inverse variance (precision) of the noise in the data.
Using Bayes' theorem, we can combine the likelihood and the prior to derive the posterior distribution of the coefficients:

$$p(w|X, y) \propto p(y|X, w)p(w)$$

This posterior distribution incorporates both the observed data and prior knowledge, allowing for more robust parameter estimation, especially in cases of multicollinearity.

2.5.3 Posterior Distribution

The posterior distribution for the parameters w is Gaussian, with a mean and covariance matrix that depends on the design matrix X, the regularization parameter λ, and the precision α. The posterior mean $\hat{w}$ is given by:

$$\hat{w} = (X^T X + \lambda I)^{-1} X^T y$$

The regularization term λI helps control the magnitude of the coefficients, preventing overfitting by shrinking the coefficients of less important predictors. The posterior covariance matrix is:

$$\text{Cov}(w) = \alpha^{-1}(X^T X + \lambda I)^{-1}$$

The inclusion of the covariance matrix allows us to quantify the uncertainty in the estimates, providing a probabilistic framework for prediction.

2.5.4 Connection to Marketing Mix Modeling (MMM)

In MMM, the goal is to estimate the impact of various marketing channels (e.g., TV, digital) on sales or brand awareness. These marketing activities are often highly correlated, leading to multicollinearity. Bayesian Ridge Regression addresses this issue by shrinking the coefficients of correlated variables, improving the stability and interpretability of the model.

2.5.5 Interaction Terms in MMM

One of the key features of Bayesian Ridge Regression in the context of MMM is its ability to handle interaction terms. Interaction terms allow us to model the combined effect of two or more marketing channels. For instance, the interaction between TV and digital channels can be represented as:

$$\text{Sales} = \beta_0 + \beta_1 \text{TV} + \beta_2 \text{Digital} + \beta_3 (\text{TV} \times \text{Digital}) + \epsilon$$

In this case, β_3 captures the synergistic or antagonistic effects between the two channels. Bayesian Ridge Regression, by regularizing the interaction term β_3, prevents overfitting and allows for a more reliable estimation of cross-channel effects.

2.6. Interpretation of the Bayesian Ridge Regression Model

2.6.1 Interpreting Coefficients

In Bayesian Ridge Regression, each coefficient $\hat{w}_i$ represents the expected change in the response variable y for a one-unit increase in the predictor variable x_i, while holding all other variables constant. Due to the Bayesian framework, these coefficients are estimated by combining the likelihood of the observed data with prior distributions, leading to a posterior distribution that incorporates both sources of information.

The magnitude and sign of each coefficient provide insights into the relationship between the predictor and the response variable:

- **Positive Coefficient** ($\hat{w}_i > 0$): Suggests that as x_i increases, the expected value of y also increases, indicating a direct relationship.
- **Negative Coefficient** ($\hat{w}_i < 0$): Implies that as x_i increases, the expected value of y decreases, indicating an inverse relationship.
- **Magnitude of Coefficient**: The absolute value of $\hat{w}_i$ reflects the strength of the relationship between x_i and y, after accounting for the effects of other variables and the regularization parameter λ.

It is essential to interpret these coefficients within the context of the model, considering the potential influence of multicollinearity and the regularization effect, which shrinks coefficients towards zero to prevent overfitting.

2.6.2 Interpreting Interaction Terms

Interaction terms capture the combined effect of two or more predictor variables on the response variable, beyond their individual contributions. An interaction coefficient $\hat{w}_{ij}$ associated with the product of predictors $x_i x_j$ indicates how the effect of one predictor on y changes depending on the level of another predictor.

The interpretation is as follows:

- **Positive Interaction Coefficient**: Indicates that the effect of one predictor on y increases as the level of the interacting predictor increases.
- **Negative Interaction Coefficient**: Suggests that the effect of one predictor on y decreases as the level of the interacting predictor increases.
- **Magnitude of Interaction**: The size of $\hat{w}_{ij}$ reflects the strength of the interaction effect, with larger absolute values indicating a more substantial combined influence on y.

In Marketing Mix Modeling, significant interaction terms between marketing channels reveal how the effectiveness of one channel is affected by the presence of another, providing valuable insights for cross-channel optimization strategies.

2.7. Residual Analysis and Model Diagnostics

Residual analysis is a critical step in assessing the validity and reliability of a regression model. Residuals, defined as $e_i = y_i - \hat{y}_i$, represent the discrepancies between observed and predicted values of the response variable. Analyzing these residuals helps in verifying the underlying assumptions of the regression model and identifying potential issues.

2.7.1 Assumptions of Regression Models

For the regression results to be valid, several key assumptions must hold:

1. **Linearity**: The relationship between the predictors and the response variable is linear.
2. **Independence**: The residuals are independent and uncorrelated with each other.
3. **Homoscedasticity**: The residuals have constant variance across all levels of the predictors.
4. **Normality**: The residuals are normally distributed.

Violations of these assumptions can lead to biased or inefficient estimates, affecting the model's predictive performance and the validity of statistical inferences.

2.7.2 Residual Plots

Residual plots are graphical tools used to assess the assumptions of linearity and homoscedasticity:

- **Residuals vs. Fitted Values**: Plotting residuals against predicted values helps detect nonlinearity and heteroscedasticity. A random scatter of points around zero suggests that the assumptions are met.
- **Residuals vs. Predictor Variables**: Examining residuals against individual predictors can reveal specific variables contributing to violations of assumptions.

Patterns or systematic structures in these plots indicate potential issues that may require model adjustments, such as transforming variables or adding polynomial terms.

2.7.3 Normality of Residuals

Assessing the normality of residuals is crucial for the validity of hypothesis tests and confidence intervals:

- **Histogram of Residuals**: Provides a visual representation of the residual distribution. A bell-shaped histogram suggests normality.
- **Quantile-Quantile (Q-Q) Plot**: Compares the quantiles of the residuals with those of a normal distribution. If the residuals are normally distributed, the points should closely follow a straight line.

Significant deviations from normality may indicate the presence of outliers or non-normal error terms, potentially impacting the reliability of statistical tests.

2.7.4 Homoscedasticity and Constant Variance

Homoscedasticity, or constant variance of residuals, is another critical assumption:

- **Detection**: Heteroscedasticity is often detected through the residuals vs. fitted values plot. A funnel-shaped pattern indicates increasing or decreasing variance.
- **Implications**: Heteroscedasticity can lead to inefficient estimates and biased standard errors, affecting the accuracy of confidence intervals and hypothesis tests.
- **Remedies**: Possible solutions include transforming the response variable, using weighted least squares, or employing heteroscedasticity-robust standard errors.

Ensuring homoscedasticity is essential for the reliability of the regression model's inferences.

2.7.5 Quantile-Quantile (Q-Q) Plots

Q-Q plots are used to assess whether the residuals follow a specified distribution, typically the normal distribution:

- **Construction**: Plot the ordered residuals against the theoretical quantiles of the normal distribution.
- **Interpretation**: Points that closely align with the reference line suggest normality. Systematic deviations indicate departures from normality, such as skewness or kurtosis.

Q-Q plots are a valuable diagnostic tool for verifying the normality assumption in regression analysis.

2.7.6 P-Values and Statistical Significance

P-values assess the evidence against the null hypothesis that a coefficient is zero thus being not statistically significant:

- **Interpretation**: A small p-value (generally accepted threshold of 0.05 is used) suggests that the predictor is statistically significant, meaning it has a non-zero association with the response variable.
- **Considerations in Bayesian Regression**: In Bayesian contexts, credible intervals are providing more insights into the results. A coefficient is considered significant if its credible interval does not contain zero.

Statistical significance helps identify which predictors and interaction terms have meaningful effects on the response variable.

2.7.7 Coefficients of Models

Understanding the coefficients in the context of the model is crucial:

- **Standardized Coefficients**: These coefficients are scaled to have unit variance, allowing for comparison of the relative importance of predictors.
- **Confidence Intervals**: Provide a range of plausible values for the coefficients, reflecting the uncertainty in the estimates.
- **Regularization Effects**: In Bayesian Ridge Regression, regularization shrinks coefficients towards zero, which can impact their interpretation and statistical significance.

Analyzing the coefficients helps in understanding the model's structure and the relationships between variables.

2.8. Implications for Marketing Mix Modeling

Applying these interpretative and diagnostic techniques in the context of MMM enhances the reliability and validity of the model's insights:

- **Model Validity**: Ensuring that regression assumptions hold increases confidence in the model's predictions and recommendations.
- **Strategic Decisions**: Accurate interpretation of coefficients and interaction terms informs effective allocation of marketing resources across channels.
- **Continuous Improvement**: Residual analysis and diagnostics highlight areas where the model can be improved, such as by adding variables, transforming data, or refining the model structure.

A rigorous approach to model interpretation and validation supports data-driven decision-making in marketing strategies.

2.9. Conclusion

Bayesian Ridge Regression is a powerful tool for addressing the challenges of multicollinearity and interaction effects in Marketing Mix Modeling. By incorporating prior knowledge and regularizing the coefficients, it provides robust and interpretable results. In the thesis, the model offered critical

insights into cross-channel interactions, saturation curves, and budget optimization strategies, leading to more effective marketing decisions.
Interpreting the Bayesian Ridge Regression model involves understanding the implications of the coefficients, including interaction terms, and assessing the model's validity through residual analysis. By thoroughly examining the residuals and ensuring that regression assumptions are met, the model provides reliable and actionable insights. Techniques such as Q-Q plots, histograms of residuals, and scatter plots are essential tools in this diagnostic process. The combination of robust model interpretation and rigorous validation strengthens the overall analysis, leading to more effective marketing mix decisions.

2.10. Data Preprocessing and Transformation in Marketing Mix Modeling

Data preprocessing and transformation are critical phases in Marketing Mix Modeling (MMM) that ensure the accuracy, reliability, and interpretability of analytical outcomes. Given the diverse and complex nature of data sources in MMM—spanning traditional channels like television and radio to digital platforms like social media and search advertising—effective preprocessing involves several key steps: data cleaning, integration, transformation, reduction, and addressing common issues such as missing data and outliers [29]. Proper preprocessing not only enhances model performance but also allows for a more nuanced analysis of interaction effects between traditional and digital marketing channels.

2.10.1 Common Data Challenges in MMM

MMM data often suffer from two primary issues: missing data and outliers. Missing data can result from incomplete records, unreported values, or inconsistent data collection methods across different channels. Outliers may arise due to data entry errors, measurement inaccuracies, or genuine but extreme variations in the data [8]. Both issues can significantly distort regression analyses, leading to biased parameter estimates and unreliable conclusions, especially when modeling interactions between channels.

2.11. Handling Missing Data

Addressing missing data is crucial for maintaining the integrity of regression models in MMM [9, 10, 30]. The impact of missing data depends on its mechanism—Missing Completely at Random (MCAR), Missing at Random (MAR), or Missing Not at Random (MNAR)—each requiring different handling strategies to mitigate bias and loss of information.

2.11.1 Strategies for Managing Missing Data

Several methodologies exist to handle missing data:

1. **List-wise Deletion**: Eliminates any observation with missing values. While straightforward, it can introduce bias if the data are not MCAR and significantly reduce the sample size, potentially affecting the analysis of interactions between channels.
2. **Pair-wise Deletion**: Uses all available data for each estimation, maximizing data usage but potentially leading to inconsistent estimates due to varying sample sizes.
3. **Mean Imputation**: Replaces missing values with the mean of observed values, preserving sample size but underestimating variability and distorting distributions.
4. **Regression Imputation**: Predicts missing values using regression models based on other variables, which can overestimate correlations and reduce data variability.
5. **Multiple Imputation**: Generates multiple datasets with imputed values, analyzes each separately, and combines results to account for uncertainty, providing unbiased estimates [9, 10].

This method is particularly useful when analyzing interactions between variables, as it preserves the relationships among variables.

2.11.2 Impact on Model Performance

The effect of missing data on regression models is contingent on the missing data mechanism [10]:

- **MCAR**: Missingness is unrelated to any data, leading to unbiased estimates but reduced statistical power due to smaller sample sizes.
- **MAR**: Missingness is related to observed data, requiring appropriate methods like multiple imputation to avoid biased estimates.
- **MNAR**: Missingness is related to unobserved data, making it challenging to obtain unbiased estimates without modeling the missing data mechanism explicitly.

In the context of analyzing interaction effects, improper handling of missing data can lead to incorrect conclusions about the synergistic or antagonistic relationships between traditional and digital channels.

2.12. Outlier Detection and Handling

Outliers can significantly distort regression analyses by biasing parameter estimates and inflating error variances [11]. Identifying and appropriately handling outliers is essential for the validity of MMM, particularly when modeling interaction effects, as outliers in one channel can disproportionately affect the estimated interaction terms.

2.12.1 Techniques for Managing Outliers

Common strategies include:

1. **Deletion**: Removing outliers, which risks loss of information and potential bias [12].
2. **Transformation**: Applying functions (e.g., logarithmic, square root) to reduce the impact of outliers [13].
3. **Capping (Truncation)**: Setting outlier values to predefined thresholds [31].
4. **Winsorization**: Replacing outliers with specific percentiles of the data [14].
5. **Imputation**: Estimating and replacing outlier values using statistical models [30].

Mathematically, these methods can be expressed as:

- **Deletion**: Exclude x_i if identified as an outlier.
- **Transformation**: $x_i' = f(x_i)$, where f is a transformation function (e.g., $\log(x_i)$).
- **Capping**: $x_i' = \min(\max(x_i, a), b)$, with thresholds a and b.
- **Winsorization**: Adjust x_i to fall within $[Q_1 - k \cdot IQR, Q_3 + k \cdot IQR]$, where Q_1, Q_3, and IQR are the first quartile, third quartile, and interquartile range, respectively.
- **Imputation**: Replace x_i with an estimated value (e.g., mean or median).

2.12.2 Influence on Model Performance

Outliers can lead to:

- **Biased Coefficients**: Skewing the estimated relationships between variables, including interaction terms.
- **Inflated Error Variance**: Increasing the variance of error terms, reducing model precision.
- **Reduced Statistical Power**: Hindering the ability to detect significant effects [11].

In MMM, this may result in misestimating the effectiveness of marketing activities and misallocating resources, especially when evaluating the combined impact of traditional and digital channels.

2.13. Integration of Data Preprocessing Techniques

Effective data preprocessing requires integrating missing data handling and outlier detection within the overall data transformation workflow:

1. **Identification**: Use statistical methods (e.g., Z-scores, IQR) and visualizations (e.g., box plots) to detect anomalies [15].
2. **Assessment**: Determine if anomalies are data errors or meaningful variations, particularly in the context of cross-channel interactions.
3. **Application**: Choose appropriate techniques based on data characteristics and modeling goals.
4. **Validation**: Ensure preprocessing improves model performance without introducing bias.

2.14. Data Transformation Techniques in MMM

Beyond handling missing data and outliers, MMM often requires specific data transformations to capture temporal dependencies, interaction effects, and external influences.

2.14.1 Lagged Variables and Ad-Stock Modeling

Lagged variables and ad-stock models are essential for representing the delayed and diminishing effects of marketing activities [2, 16]. They also play a crucial role in modeling interaction effects between traditional and digital channels.

Lagged Variables Lagged variables capture the effect of past values of a variable on its current value:

$$Y_t = \alpha + \beta Y_{t-1} + \gamma X_t + \delta X_{t-1} + \epsilon_t$$

where Y_t is the dependent variable at time t, Y_{t-1} is its lagged value, X_t represents independent variables (e.g., marketing spend on digital channels), and X_{t-1} are their lagged values [17]. Including lagged variables allows for the modeling of carry-over effects and temporal interactions between channels.

Ad-Stock Models Ad-stock models represent the carry-over effect of advertising spend:

$$A_t = X_t + \lambda A_{t-1}$$

where A_t is the ad-stock variable, X_t is current advertising spend, and λ is the decay parameter ($0 < \lambda < 1$) [16]. The ad-stock formulation can be extended to include interaction terms between traditional and digital channels:

$$A_t = X_t^{(T)} + X_t^{(D)} + \lambda A_{t-1}$$

where $X_t^{(T)}$ and $X_t^{(D)}$ represent spend on traditional and digital channels, respectively. Interaction effects can be modeled by including cross-products of these variables.

Considerations and Calibration Selecting appropriate lag lengths and decay parameters is crucial. Overreliance on lagged variables can introduce multicollinearity, and incorrect decay parameters can misrepresent marketing effects [18]. Techniques like the Durbin-Watson test for autocorrelation [33] and information criteria (e.g., AIC) for lag selection [19] aid in calibration.
When modeling interactions, it's important to consider the potential for time-lagged interaction effects, where the impact of one channel may influence the effectiveness of another channel at a later time.

2.15. Incorporating External Data Sources

External data such as weather conditions or economic indicators can enhance MMM by providing context [20]. For instance, economic downturns may affect the relative effectiveness of traditional versus digital channels.

2.15.1 Challenges and Integration

Challenges include:

- **Data Integration**: Aligning data with different formats and temporal granularities [3].
- **Data Quality**: Ensuring external data is accurate and reliable [35].
- **Model Complexity**: Managing increased dimensionality and potential multicollinearity [36].

Incorporating external variables requires careful preprocessing and validation to avoid overfitting and ensure meaningful contributions to the model [22]. When analyzing interaction effects, external factors may moderate or mediate the relationships between traditional and digital channels.

2.16. Normalization, Standardization, and Feature Scaling

Standardizing data scales is essential for model convergence, stability, and interpretability [4]. This is particularly important when variables represent different marketing channels with varying scales and units.

2.16.1 Normalization

Rescales data to a [0,1] range using:

$$x' = \frac{x - \min(x)}{\max(x) - \min(x)}$$

This is useful when features have varying scales but do not follow a Gaussian distribution [5].

2.16.2 Standardization

Centers data around zero with unit variance:

$$x' = \frac{x - \mu}{\sigma}$$

Standardization is appropriate when data are normally distributed [37].

2.16.3 Impact on Regression Coefficients

Scaling ensures that coefficients are comparable and improves numerical stability [6]. It prevents features with larger scales from unduly influencing the model [38]. When analyzing interaction effects, scaling is crucial because interaction terms are products of variables, and unscaled variables can lead to large values that dominate the model.

2.16.4 Applications in MMM

In MMM, scaling allows for direct comparison of marketing channels with different budget scales, facilitating better ROI assessments [2]. It also aids in interpreting interaction effects between channels.

2.16.5 Considerations

While scaling aids in model performance, it may distort variables with natural zero points (e.g., sales). Careful consideration is needed to choose appropriate scaling methods [39].

2.17. Seasonality Adjustments

Adjusting for seasonality removes regular, predictable patterns to enhance model accuracy [40]. Seasonality can affect both traditional and digital channels differently, and failing to adjust for it can confound the analysis of interaction effects.

2.17.1 Techniques

1. **Moving Average**: Smooths data to highlight trends [41].
2. **Exponential Smoothing**: Assigns decreasing weights to older observations [42].
3. **Decomposition**: Separates time series into trend, seasonal, and residual components [43].
4. **Differencing**: Removes trends and seasonality by subtracting previous observations [23].

2.17.2 Additive vs. Multiplicative Decomposition

Additive Model Assumes constant seasonal effects:

$$Y_t = T_t + S_t + E_t$$

Multiplicative Model Assumes seasonal effects vary proportionally with trend:

$$Y_t = T_t \times S_t \times E_t$$

2.17.3 Impact on MMM

Seasonality adjustments prevent misattributing seasonal effects to marketing activities, improving parameter estimates and forecasting [2]. When analyzing interaction effects, adjusting for seasonality ensures that observed interactions are not artifacts of seasonal patterns.

2.18. Data Granularity

The level of data detail affects the modeling process. High granularity data can capture short-term interactions between channels, while low granularity data may only reveal long-term trends.

2.18.1 Impact on Data Transformation

- **High Granularity**: Captures short-term effects and potential immediate interactions between channels but may introduce noise.
- **Low Granularity**: Highlights long-term trends but may overlook short-term variations and interactions.

2.18.2 Aggregation and Disaggregation

- **Aggregation**: Summarizes data to reduce noise, potentially losing detail on interactions.
- **Disaggregation**: Provides detailed insights but may complicate the model [44].

Choosing the appropriate granularity balances the need for detail with model simplicity and accuracy. In the context of interaction effects, finer granularity may be necessary to capture dynamic interactions between traditional and digital channels.

2.19. Advanced Data Transformation Techniques in MMM

Addressing complex issues such as non-stationarity, heteroscedasticity, and high dimensionality is essential for accurate modeling of interaction effects in MMM.

2.19.1 Non-Stationarity and Its Treatment in MMM

Non-stationarity refers to changes in the statistical properties of a time series over time [23]. It can obscure true relationships and interactions between marketing channels.

Techniques for Addressing Non-Stationarity

1. **Differencing**: Eliminates trends by subtracting consecutive observations:
$$\Delta y_t = y_t - y_{t-1}.$$
2. **Seasonal Differencing**: Removes seasonal effects by subtracting observations from the same period in previous cycles:
$$\Delta_s y_t = y_t - y_{t-s},$$
where s is the seasonal period.
3. **Transformation**: Applies transformations (e.g., logarithmic, Box-Cox) to stabilize variance and mean:
$$y'_t = \begin{cases} \dfrac{y_t^\lambda - 1}{\lambda}, & \lambda \neq 0, \\ \ln(y_t), & \lambda = 0. \end{cases}$$
4. **Time Series Models**: Uses models like ARIMA that incorporate differencing.
5. **State Space Models and Kalman Filtering**: Handles time-varying parameters flexibly [7].

Impact on MMM Addressing non-stationarity improves the validity of regression coefficients and enhances predictive accuracy. It ensures that interaction effects are accurately estimated and not confounded by underlying trends or seasonality.

2.19.2 Handling Heteroscedasticity in Regression Models

Heteroscedasticity occurs when error variances are not constant [24]. It can lead to inefficient estimates and invalid inference.

Techniques for Addressing Heteroscedasticity

1. **Transformation of Variables**: Applies logarithmic transformations to stabilize variance:
$$y'_i = \ln(y_i), \quad X'_i = \ln(X_i).$$
2. **Weighted Least Squares (WLS)**: Assigns weights inversely proportional to variance:
$$\hat{\beta}_{\text{WLS}} = (X^\top W X)^{-1} X^\top W y,$$
where W is a diagonal matrix with $w_{ii} = 1/\sigma_i^2$.
3. **Robust Standard Errors**: Uses heteroscedasticity-consistent standard errors [26].
4. **Generalized Least Squares (GLS)**: Models the variance structure directly.

Impact on Regression Estimates Correcting for heteroscedasticity leads to efficient estimates and valid inference, which is critical when interpreting interaction effects.

2.20. Integrative Perspective on Data Transformation in MMM

A unified data transformation workflow integrates all preprocessing steps to enhance model robustness and interpretability.

Unified Workflow

1. **Initial Preprocessing**: Handle missing data and outliers.
2. **Feature Scaling**: Normalize or standardize data.
3. **Temporal Adjustments**: Incorporate lagged variables, ad-stock models, adjust for seasonality and non-stationarity.
4. **Model Assumptions**: Address heteroscedasticity and other regression assumptions.

Enhancing Analysis of Interaction Effects By systematically applying these techniques, practitioners can:

- **Accurately Model Interactions**: Capture the true nature of interactions between traditional and digital channels.
- **Improve Predictive Accuracy**: Enhance forecasts of marketing effectiveness.
- **Ensure Valid Inference**: Make reliable decisions based on statistical evidence.
- **Facilitate Interpretation**: Simplify complex models for better strategic insights.

2.21. Conclusion

Effective data preprocessing and transformation are indispensable in MMM, particularly when analyzing interaction effects between traditional and digital marketing channels. By integrating foundational preprocessing steps with advanced techniques, practitioners can create robust models that provide valuable insights into the synergistic effects of multi-channel marketing strategies.

3. Research Gaps and Thesis Objective

3.1. Research Gaps

The existing literature on marketing channels tends to examine them in isolation, focusing on their long-term and short-term impacts, return on investment (ROI), and overall effectiveness. Although individual channels like TV, social media, and display ads have been extensively studied, the dynamics that arise when these channels are used in combination are less frequently analyzed, despite their growing relevance in modern marketing strategies.

In most research, digital and traditional marketing channels are often treated as separate entities, with the assumption that each channel operates independently and contributes its own discrete effect to campaign outcomes. However, in today's marketing environment, channels frequently interact, either amplifying or diminishing the effects of one another. For instance, the combined use of digital and traditional channels may either enhance brand visibility and engagement, or conversely, result in diminishing returns due to audience overlap or ad fatigue.

This gap in the analysis of interaction effects limits our understanding of multi-channel marketing dynamics. Few studies statistically evaluate the performance of channel combinations or how interaction effects between channels influence overall campaign performance. This lack of research on interaction effects highlights the need for a more comprehensive approach to marketing strategy optimization, particularly given the growing complexity of the media ecosystem.

3.2. Thesis Objective

The primary objective of this thesis is to investigate the complex interaction effects that occur both within digital channels and between digital and traditional marketing channels. By analyzing these interactions, this study aims to bridge the gap in the current literature and provide a detailed understanding of how channel combinations affect campaign performance.

Through the use of statistical models incorporating interaction terms, this research demonstrates that certain channel combinations perform better than others. For example, the analysis shows how synergistic effects from Programmatic Video, YouTube 6s, and Influencers contribute to campaign success, while also identifying combinations that may be less effective due to issues like audience saturation or conflicting messaging.

In addition to addressing gaps in existing research, this thesis contributes a fresh perspective on marketing budget allocation by focusing on interaction effects. It goes beyond the conventional approach of evaluating channels in isolation and underscores the importance of understanding how channels can either complement or hinder each other's effectiveness.

Furthermore, this research provides valuable insights into Marketing Mix Modeling (MMM) by extending its scope to include channel interaction analysis. This allows for a more comprehensive evaluation of how marketing investments can be optimized across both digital and traditional channels.

In summary, this study not only confirms the effectiveness of certain channels individually but also offers actionable recommendations on how marketers can leverage synergistic channel combinations to enhance overall campaign performance. The findings provide practical guidance for marketers looking to maximize ROI by making strategic, data-driven decisions regarding budget allocation.

4. Datasets Description

For the thesis, two distinct datasets were provided, each containing information regarding the marketing campaigns conducted by company A and B. The companies provided the 'G Dataset' (company A) and 'Dataset D' (company B) to analyze the marketing mix, particularly to find insights regarding the impact of influencers on the marketing mix.

4.1. Company A

The information regarding the data provided by company A:

Business Data Variables:
- Volume Sales Brand: Total units sold, providing direct insight into product demand.
- Value Sales Brand: Total revenue generated from sales, reflecting the financial success of products.
- Average Price Brand: Calculated by dividing total value by total volume, indicating the average selling price.
- Weighted Handling Distribution: Proportion of relevant retail locations where the product is available.
- Number of Items: Range of product variants available (e.g., sizes, flavors).
- Value Sales in Promo: Revenue from promotional pricing, including price cuts and bundled offers.
- Average Promo Price: Average price during promotional periods, offering insight into discount strategies.
- Seasonality Market: Indexed sales data showing seasonal variations and trends.

Bank Holidays Variables:
- Christmas, New Year's Eve, Easter, Mid-August, Mom's Day, Black Friday: Aimed to reflect sales performance and promotional activities during significant holidays and their impact on sales.

COVID-19 Impact Variables:
- Lockdown: Represents sales and operational data during periods of COVID-19 lockdowns, highlighting impacts on consumer behavior and distribution.

Public Relations (PR) and Trade Data:
- Number of Editorials (Printed and Online): Frequency of product mentions in print and digital media.
- Totem Displays (In-store and Next to Cashier): Promotional displays used to increase product visibility and stimulate purchases.
- Distributed Samples: Number of product samples given out to promote product trials.

Earned Media and User Generated Content:
- Online Reviews and Ratings: Aggregated consumer feedback on online platforms.
- Positive and Negative Opinions: Sentiment analysis results from social media platforms.
- Number of Stories in Free-Seeding: Engagement activities through story sharing by influencers without explicit contracts.
- Number of Influencers Involved: The total count of influencers engaged in promoting the product.

Media and Advertising Investments:

- GRP TV (Total, 20", 10"): Gross Rating Points for TV advertising, reflecting the campaign's reach and frequency.
- Impressions (YouTube, Other Video Publishers, Social Meta, Spotify, TikTok, Twitch): Total views on various digital platforms.
- Investments (TV, Print, YouTube, Other Video Publishers, Social Meta, Google Search, Amazon Search, TikTok, Twitch): Financial allocations for advertising across different media channels.
- Investment Influencer Marketing (TikTok, Instagram): Spending on influencer marketing campaigns on specific platforms. - Competitor Analysis Variables
- GRP Competitors (1 up to 14): Comparison of Gross Rating Points across various competitors, offering a relative measure of competitors' market presence and campaign intensity.

4.2. Company B

The information regarding the data provided by company B:

Sales 1, Sales 2, Sales 3:

- Brand Volume (Sales 1, Sales 2, Sales 3): Total units sold within each of the sales categories (Sales 1, Sales 2, Sales 3), providing direct insight into product demand across different sales channels or segments.
- Brand Value (Sales 1, Sales 2, Sales 3): Total revenue generated from sales within each category, reflecting the financial success of products in those specific segments.
- Weighted Distribution (Sales 1, Sales 2, Sales 3): Proportion of relevant retail locations where the product is available within each category, indicating distribution reach and efficiency across different sales channels.
- Competitor Volume (Sales 1, Sales 2, Sales 3): Total units sold by competitors within each sales category, providing a benchmark for comparing the brand's market share against competitors.
- Competitor Value (Sales 1, Sales 2, Sales 3): Total revenue generated by competitors within each sales category, offering insights into the financial success of competitors relative to the brand.
- Market Volume (Sales 1, Sales 2, Sales 3): Total units sold across the entire market, including all brands and competitors within each sales category, reflecting overall market demand in those segments.
- Market Value (Sales 1, Sales 2, Sales 3): Total revenue generated across the entire market within each sales category, indicating the market's financial size and growth potential in those areas.
- Brand Weighted Distribution in Promo (Sales 2 and Sales 3): This metric measures the proportion of the total market sales volume that comes from stores where the brand is running a promotion. It indicates the brand's presence in higher-performing stores during promotional periods, reflecting the potential impact of the promotion on overall sales.
- Brand Volume in Promo (Sales 2 and Sales 3): This represents the total sales volume of the brand during the promotional period. It's a key indicator of how effective the promotion is in driving sales volume for the brand.
- Brand Numeric Distribution (Sales 2 and Sales 3): This metric shows the percentage of stores where the brand is available, regardless of sales volume. It provides insight into the brand's availability across the retail network.
- Competitor Weighted Distribution in Promo (Sales 2 and Sales 3): Similar to the brand's weighted distribution, this metric measures the share of the total market sales volume coming from stores where a competitor's promotion is active. It allows for a comparison of the brand's promotional presence relative to competitors in high-performing outlets.
- Competitor Volume in Promo (Sales 2 and Sales 3): This is the total sales volume achieved by a competitor during their promotional periods. It helps in understanding the competitive landscape and assessing how a competitor's promotions are performing in terms of sales volume.
- Competitor Numeric Distribution (Sales 2 and Sales 3): This metric indicates the percentage of stores where a competitor's products are available, providing insight into the competitor's retail

penetration and comparing it with the brand's distribution.

TV Variables:

- GRP TV (10", 15", 20", 30"): Gross Rating Points for TV advertising, reflecting the reach and frequency of TV campaigns. This metric is typically broken down into different time lengths of TV spots (e.g., 10 seconds, 20 seconds, etc.).
- TVC (Copy 1, 2, 3): a TV commercial (TVC) used during a campaign. "Copy 1, 2, 3".
- PT (Prime Time): This refers to the time slots during which TV viewership is at its highest, typically in the evening.
- DT/NT (Day Time/Night Time): These terms distinguish between ads aired during the day (Day Time) and those aired at night (Night Time). Daytime ads might target different demographics than nighttime ads, depending on the typical audience for those time slots.
- Main Channels: These are the major TV channels with a wide reach and large audiences.
- Minor Channels: These refer to smaller, niche TV channels that might have a more targeted or specific audience.
- OOB (Out of Break): Refers to commercials that are aired just after a commercial break ends, right before the program resumes.
- Not OOB: Refers to commercials that are aired during the commercial break rather than immediately before the program resumes.

Digital Variables:

- Impressions (YouTube 6s, YouTube 10s, Programmatic Video 1 and 2, Social, Programmatic Video Display, Influencers, CTV): Total views or interactions across various digital platforms. This indicates the reach and engagement of digital advertising campaigns across different channels.
- Investments (TV Budget Planned, YouTube 6s, YouTube 10s, Programmatic Video 1 and 2, Social, Programmatic Video Display, Influencers, CTV): Financial allocations for advertising across different media channels. This includes spending on traditional media (TV) and digital platforms, reflecting the marketing budget distribution across various channels.
- Investment Influencer Marketing: Spending on influencer marketing campaigns. This highlights the budget allocated towards engaging influencers to promote the brand on social media.

4.3. Dataset D 'Sales 1'

4.3.1 Sales Value Analysis

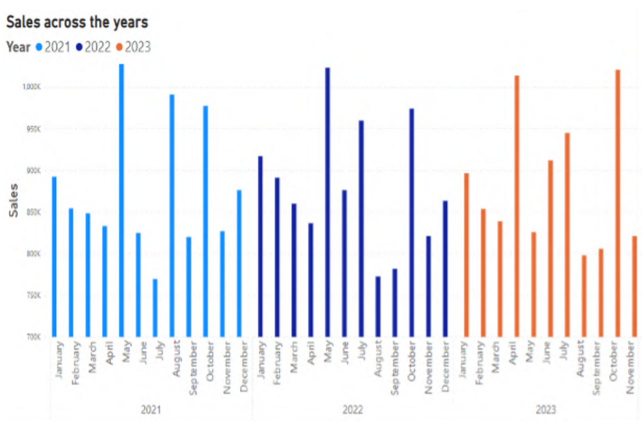

Figure 1: Sales Value Across the months and years

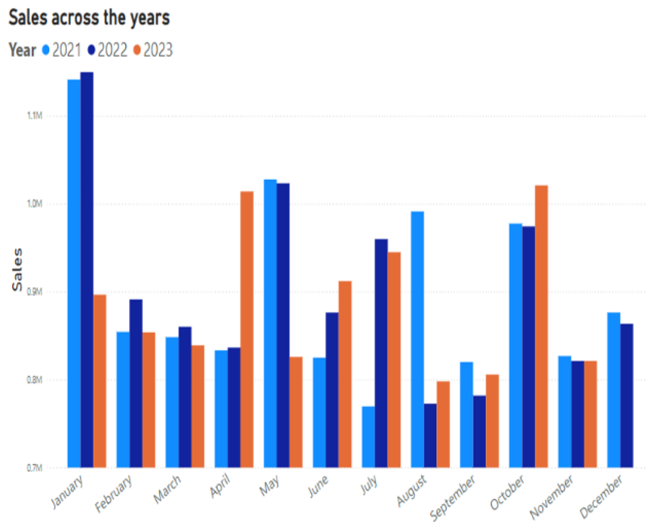

Figure 2: Sales Value Across the months and years

In 2021, sales started the year strong, peaking in January before declining sharply. In 2022, sales show a more consistent pattern, with a strong performance in March, July, and October. In 2023, sales started at a lower level than in previous years but showed an increase in March and a particularly strong surge in October. Sales are seasonal, with peaks in certain months, possibly related to industry characteristics or purchasing behavior. Increases in certain periods, such as March and October 2023, may indicate the successful implementation of new strategies, products, or marketing campaigns. Economic and external factors: The impact of economic conditions, competition, and other external factors can explain sales fluctuations.

4.3.2 Average Price Analysis

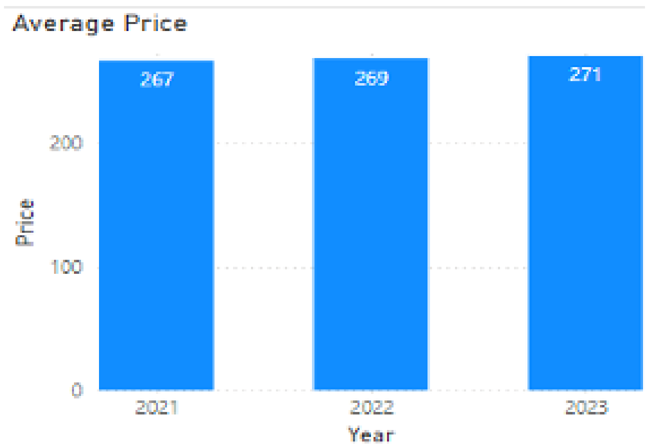

Figure 3: Average Price

The average price has remained relatively stable from 2021 to 2023, increasing slightly from 267 euros to 271 euros. This indicates a stable pricing strategy. Such a high average price tells us that the brand is engaged in selling premium and high-quality products. The fact that the price does not decrease from year to year and has a slight increase tells us that the demand for the product remains stable, and an additional indicator of customer loyalty.

4.3.3 Brand vs. Competitor Analysis

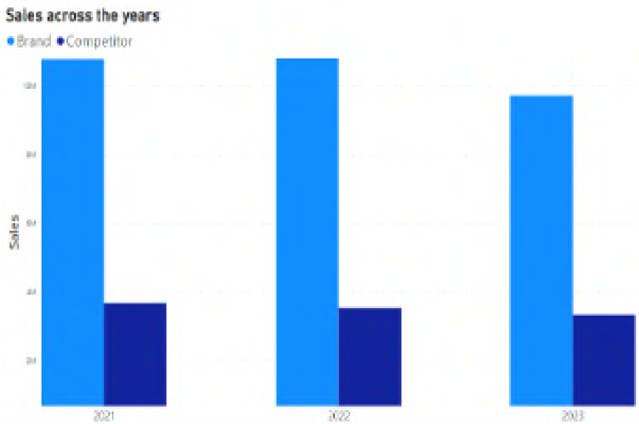

Figure 4: Brand vs. Competitors

The sales comparison between the brand and competitors shows that while the competitor's sales are consistently lower, they have shown a decrease from 2021 to 2023. In general, sales are decreasing not only for the competitor but also for the company itself. This can be influenced by various reasons, from increased competition in the market, decreased demand, increased costs, to technological changes.

4.4. Dataset D 'Sales 2'

4.4.1 Sales Value Analysis

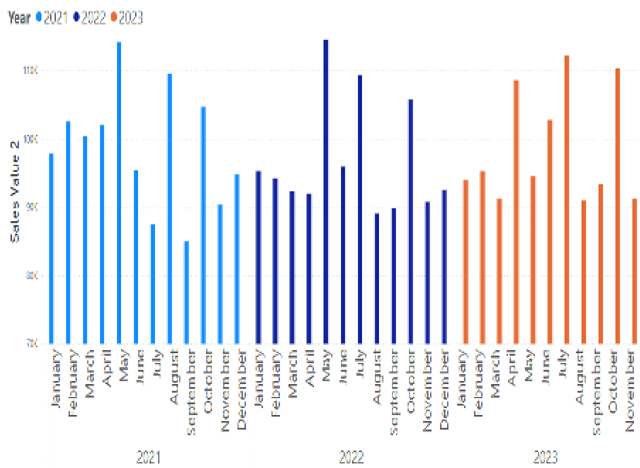

Figure 5: Sales Value across the years and months

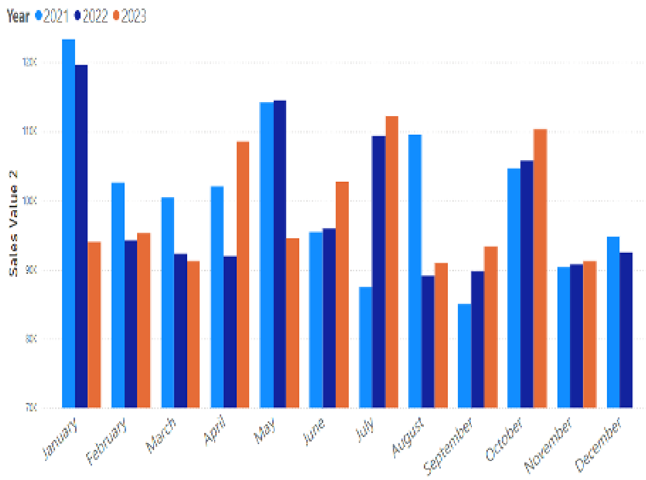

Figure 6: Sales Value across the years and months

Sales values in 2021 showed strong performance in the first half, particularly in May and July, with dips in the second half of the year. Sales values in 2022 were more stable throughout the year but slightly lower compared to 2021, indicating a potential impact from increased pricing or competitive pressure. In 2023 sales showed a more uniform distribution across the months, with notable spikes in July and October. However, overall sales values were slightly lower compared to previous years.

4.4.2 Average Price Analysis

The average price for the brand increased steadily from 103 in 2021 to 115 in 2023. The consistent price increase suggests a possible premium positioning of the brand. However, this may also have contributed to the slight decline in market share, as higher prices could deter price-sensitive customers.

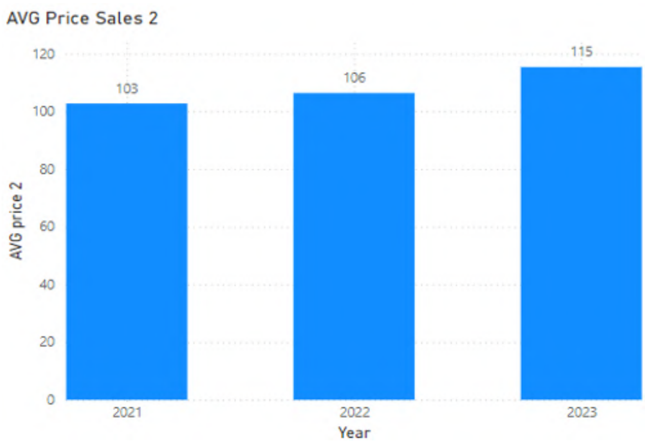

Figure 7: Average Price

4.4.3 Brand vs. Competitor Analysis

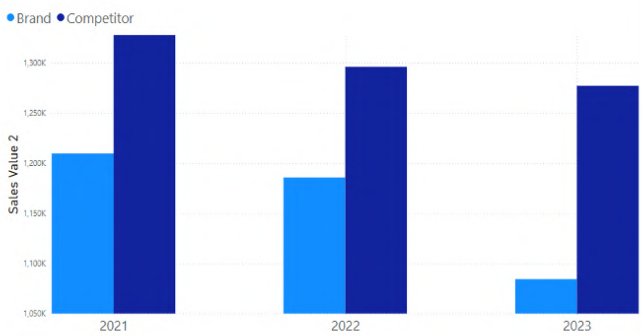

Figure 8: Brand vs. Competitor

The brand's sales volume in 2021 was slightly lower than the competitor's. In 2022, the difference remained almost the same, except for the fact that in general the company's and competitors' indicators fell, the downward trend also continued in 2023, however, the company's indicators became almost 200k less compared to the same data for 2022.

4.4.4 Market Share and Promotional Activities

Year	Brand Share	Competitor Share
2021	27.01%	33.28%
2022	25.99%	33.1%
2022	24.29%	34.57%

Table 1: Brand Share

Year	Brand Value (€)	Competitor Value (€)	Brand ND Mean	Brand WD Mean	Brand Promo (€)	Brand WD Promo Mean	Comp ND Mean	Comp WD Mean	Comp Promo (€)	Comp WD Promo Mean (€)
2023	1084401,71	1277245,55	70,16	74,44	344,19	6,27	81,61	76,74	2186,33	820,93
2022	1185689,59	1296132,46	73,98	75,72	464,25	7,64	81,56	76,75	2368,97	1039,33
2021	1209702,39	1327732,23	73,50	75,54	433,57	7,43	81,29	76,76	2458,55	1059,35
Total	**1159931,23**	**1300370,08**	**72,55**	**75,23**	**414,00**	**7,11**	**81,49**	**76,75**	**2337,95**	**973,20**

Figure 9: Promotional and Distributional Activities

The brand's market share dropped from 27.01 per cent in 2021 to 24.29 per ce t in 2023. This decline suggests that the brand might be losing ground to competitors, possibly because of higher prices and less aggressive promotional efforts. The competitor maintained a strong market share of around 33 per cent and even managed to slightly increase it by 2023. This indicates that their strategies in promotions and distribution were effective. The brand's promotional activities were much lower than the competitor's in all the years, which likely played a role in the declining market share. In 2023, the competitor activities were almost 7 times higher compared to the brand activities.

4.4.5 Distributional Analysis

Numerical Distribution (ND) and Weighted Distribution (WD): The competitor consistently performed better than the brand in both ND and WD, showing that they have a wider and more effective distribution network. Promotional Effectiveness (WD Promo): The competitor's WD Promo values were much higher than the brand's, especially in 2023. This might explain why the competitor has maintained or even grown their market share, even though the brand had a competitive pricing strategy.

4.5. Dataset D ’Sales 3’

4.5.1 Sales Value Analysis

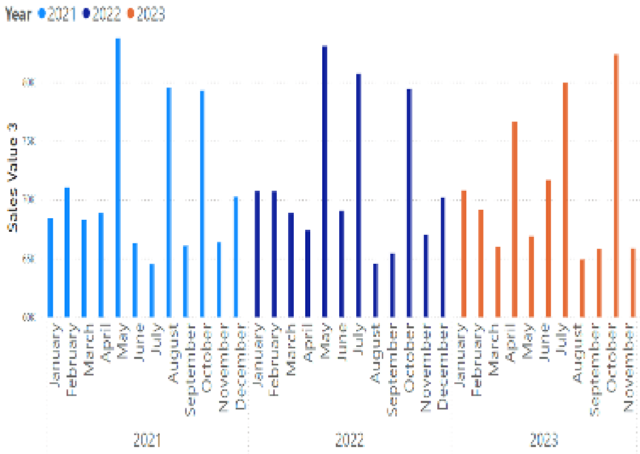

Figure 10: Sales Value across the years and months

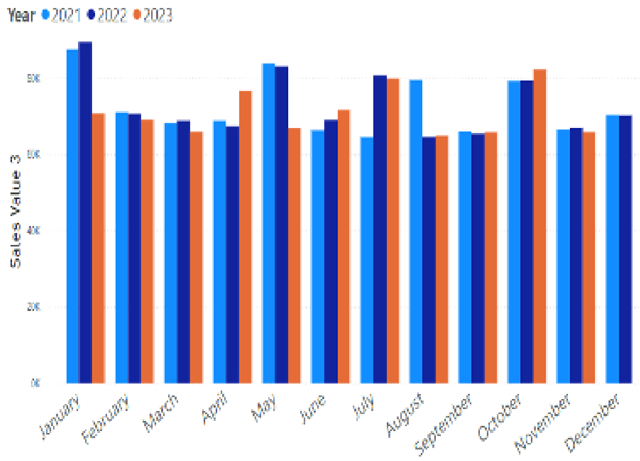

Figure 11: Sales Value across the years and months

In 2021, sales were strong in January and May, with notable peaks, particularly in May where there was a significant spike. In 2022, a similar pattern as in 2021 is observed, with a spike in May, but sales were more balanced across other months, indicating improved consistency. In 2023, sales show a decline overall compared to the previous years. The peaks are less pronounced, suggesting a possible decrease in market engagement or effectiveness of promotional strategies. 2023 shows a decline in sales compared to 2021 and 2022 in most months, indicating that the brand’s performance is weakening over time. 2022 had better overall performance than 2021, indicating a growth phase, but this was not sustained into 2023.

4.5.2 Average Price Analysis

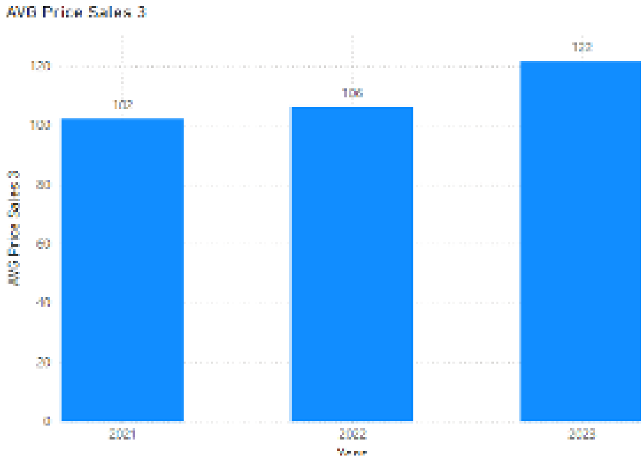

Figure 12: Average Price Analysis

There is a clear increase in the average price from 2021 (€102) to 2023 (€122), which is a 20% rise over two years. This increase might have contributed to the declining market share if the price increase was not well-received by consumers, especially if competitors maintained or offered lower prices.

4.5.3 Brand vs. Competitor Analysis

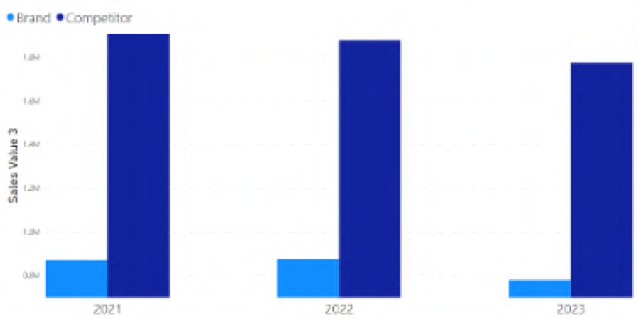

Figure 13: Brand vs. Competitor

The competitor consistently outperforms the brand across all three years. The gap between the brand and competitor's sales values is substantial, with the competitor nearly doubling the brand's sales each year. 2023 shows a further widening of the gap, which could suggest that the competitor's strategies (distribution, promotions, pricing) are more effective.

4.5.4 Market Share and Promotional Activities

Year	Brand Share	Competitor Share
2021	18.3%	38.1%
2022	17.8%	38.2%
2023	15.8%	38.8%

Table 2: Brand Share

Year	Brand Value (€)	Competitor Value (€)	Brand ND Mean	Comp ND Mean	Brand Promo (€)	Comp Promo (€)	Brand WD Mean	Comp WD Mean	Brand WD Promo Mean (€)	Comp WD Promo Mean (€)
2023	871795,31	1907446,03	29,01	59,89	345,33	1773,52	59,93	74,38	3,92	10,59
2022	875898,42	1878123,58	28,64	59,47	310,75	1561,31	59,59	74,16	3,64	10,07
2021	780078,27	1775954,85	25,79	58,37	259,22	1348,24	57,34	74,02	3,95	9,05
Avera	**842590,67**	**1853841,49**	**27,81**	**59,24**	**305,10**	**1561,02**	**58,95**	**74,19**	**3,84**	**9,90**

Figure 14: Distributional and Promotional Activities

The brand's market share dropped from 18.3% in 2021 to 15.8% in 2023. This decline suggests that the brand might be losing ground to competitors, possibly because of higher prices and less aggressive promotional efforts. The competitor maintained a strong market share of around 38.1% and even managed to slightly increase it by 2023. This indicates that their strategies in promotions and distribution were effective. The brand's promotional activities were much lower than the competitor's in all the years, which likely played a role in the declining market share. Overall brand performance same as in Sales 2, where the competitor occupies a leading position.

4.5.5 Distributional Analysis

The brand's ND mean ranges from 25.79 in 2021 to 29.01 in 2023, indicating a slight improvement over time. However, the competitor's ND mean is significantly higher, consistently around 59, suggesting a much broader market penetration. The brand's promotional spend increased from €259.22 in 2021 to €345.33 in 2023, showing a commitment to increasing promotional activities. Despite this increase, the competitor's promotional spend is far higher, with €1,773.52 in 2023, indicating a much more aggressive promotional strategy. The brand's WD mean has been relatively stable, ranging from 57.34 in 2021 to 59.93 in 2023. The competitor's WD mean is consistently higher, around 74, suggesting a more effective and wider-reaching distribution network. The brand's WD Promo Mean remains low, between 3.64 and 3.95 across the years. In contrast, the competitor's WD Promo Mean is significantly higher, reaching up to 10.59 in 2023, showing that the competitor is leveraging promotional activities much more effectively.

4.6. Dataset D 'Digital Media'

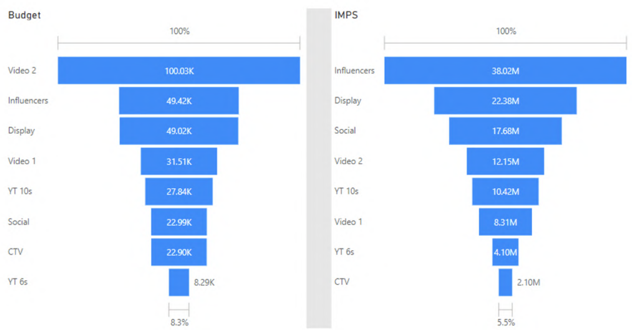

Figure 15: Digital Investments and Impressions

Year	Budget
2021	11,450.38
2022	169,853.38
2023	130,698.00

Figure 16: Digital Investments Across the Years

Budget Efficiency: There's a mismatch between budget allocation and impressions. For instance, Video 2 consumes a significant portion of the budget but does not yield proportional impressions, highlighting potential inefficiency and room for improvement.

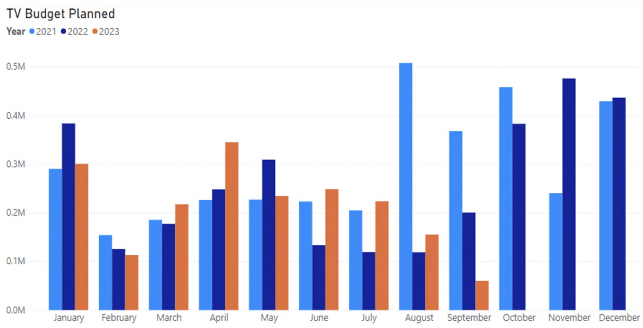

Figure 18: TV Invesments across the years and months

Recommendations: Considering the efficiency of Influencers and Display in generating impressions, it might be beneficial to reallocate some budget from underperforming areas like Video 2 towards these more effective channels. Yearly Trends: The budget increase in 2022 should be evaluated against ROI. The drop in 2023 might indicate a strategy shift or budget cuts, which should align with achieving better ROI based on the data from impressions.

4.7. Dataset D 'TV Media'

Year	TV Budget Planned
2021	3,514,458.87
2022	3,110,988.10
2023	1,897,819.81
Total	8,523,266.78

Figure 17: TV investments across the years

Over the recent years, there has been a significant shift towards shorter advertisement durations, culminating in almost all advertisements being 10 seconds long by 2023. This trend likely reflects strategic adjustments, possibly aimed at accommodating changing viewer preferences or optimizing budgetary allocations.

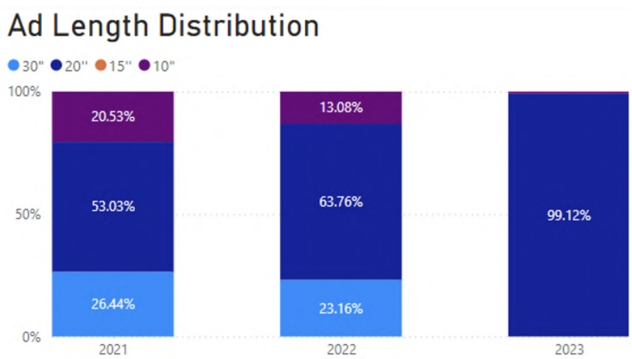

Figure 19: Ad Length Distribution

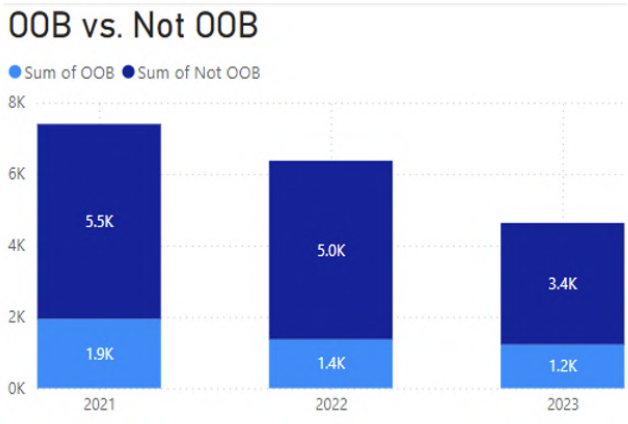

Figure 20: OOB/Not OOB

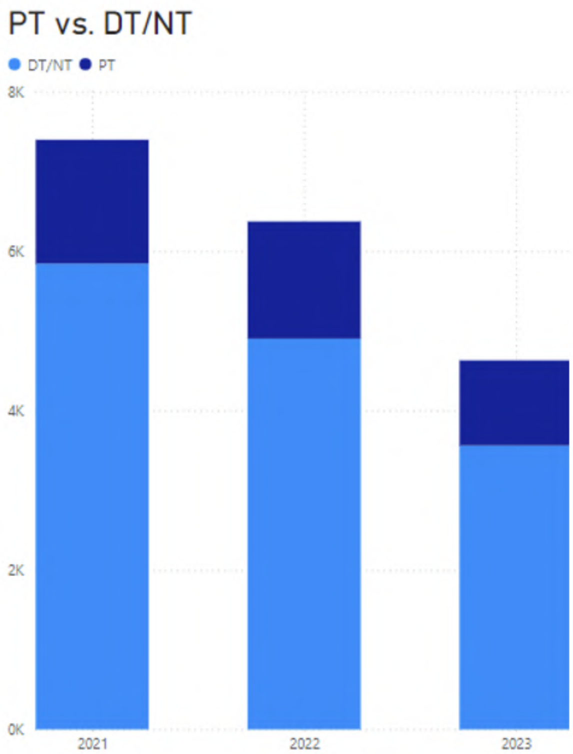

Figure 21: PT vs. DT/NT

OOB vs. Not OOB and PT vs. DT/NT: The expenditure in these specific categories has generally seen a decline over time. This decrease suggests a possible scaling back or reallocation

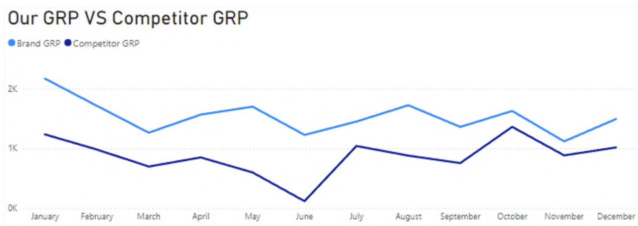

Figure 22: Company GRP vs. Competitor GRP

Throughout the year, the brand has consistently demonstrated a stronger Gross Rating Point (GRP) compared to its competitors, indicating a robust market presence. Nevertheless, careful monitoring is required during periods when competitors show signs of closing the gap.

4.8. Dataset G 'Sales'

4.8.1 Sales Value Analysis

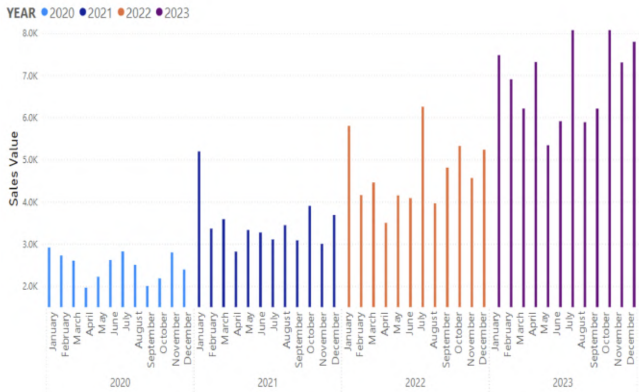

Figure 23: Sales Value across the years and months

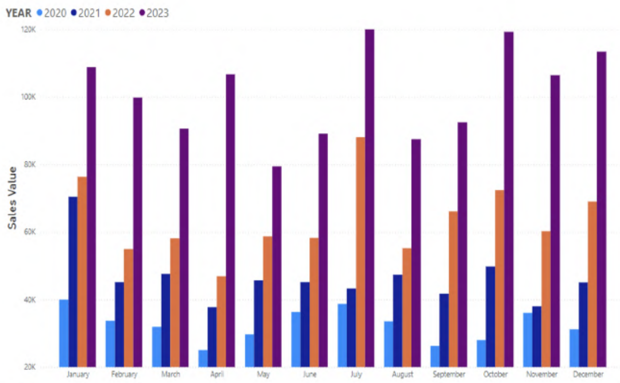

Figure 24: Sales Value across the years and months

In 2020, sales were relatively stable, with a small upward trend, and only minor fluctuations. Peaks were visible in January, July, and November, which could be attributed to seasonal consumer behavior or specific marketing efforts. In 2021, sales showed significant spikes in January and October, possibly due to targeted promotions or specific events. The general trend for the year showed higher sales than in 2020, indicating a potential increase in demand or successful marketing strategies. In 2022, there was a more noticeable increase in sales value, particularly in January, July, October, and December. This suggests a stronger impact of marketing campaigns or seasonal buying patterns, leading to higher sales figures compared to the previous years. The year 2023 demonstrated consistent growth across almost all months, with sales reaching the highest levels in the chart. The second half of the year, especially, saw significant increases, suggesting the success of new product launches, expansions, or other market growth strategies. The overall increase in sales from year to year suggests effective market penetration and customer retention strategies, with marked growth in 2023 indicating a strong consumer response to marketing efforts, enhanced product offerings, or a general expansion in market reach.

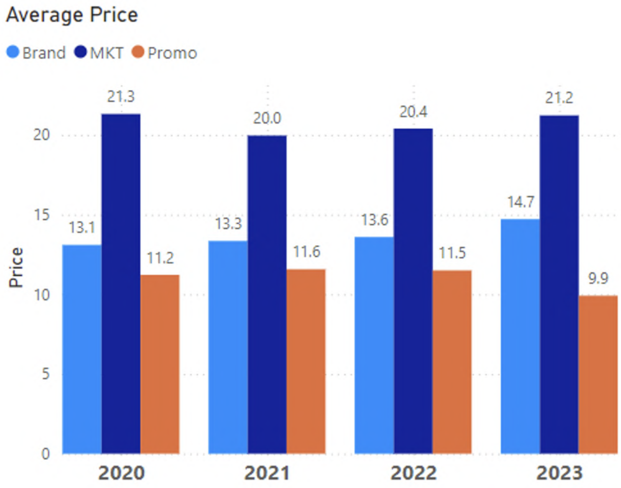

Figure 25: Average Price

4.8.2 Average Price Analysis

In 2020, the Brand was priced at 13.1, indicating a competitive approach within the market, while Market (MKT) prices were at their highest at 21.3, suggesting a premium positioning. Promotional pricing was lower at 11.2, reflecting aggressive strategies to attract customers through discounted offers. In 2021, the Brand price slightly increased to 13.3. Market prices decreased to 20.0, which might suggest a softening in the premium market or an adjustment towards more competitive pricing. Promotional pricing rose to 11.6, potentially indicating targeted promotional efforts to stimulate sales amid challenging market conditions. By 2022, Brand prices continued to rise to 13.6, showing a trend of gradual increases possibly aimed at slowly enhancing brand value. Market prices saw a slight increase to 20.4, indicating stabilization or a slight recovery in premium market conditions. Promotional pricing slightly decreased to 11.5, which could reflect a strategy to align promotional pricing more closely with the general market to remain competitive. In 2023, Brand prices increased significantly to 14.7, suggesting significant improvements in brand perception or strategic pricing adjustments to enhance value. Market prices showed a slight increase to 21.2, potentially reflecting a recovery or reaffirmation of the market's premium nature. Promotional prices saw a considerable reduction to 9.9.

4.8.3 Promo Activities

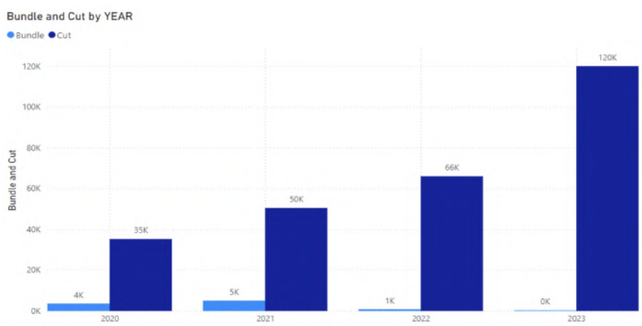

Figure 26: Promo Activities

From 2020 to 2023, the company has been purposefully increasing its focus on the "Cut" strategy, eliminating the "Bundle" by 2023. This indicates that discounts and price reductions have become a key promotional tool that has proven highly effective, while bundles have gradually lost their relevance and have been excluded from promotional plans.

4.8.4 Media Investments

Digital	TV	Print
1371,01	1639,39	1.07

Figure 27: Media Investments

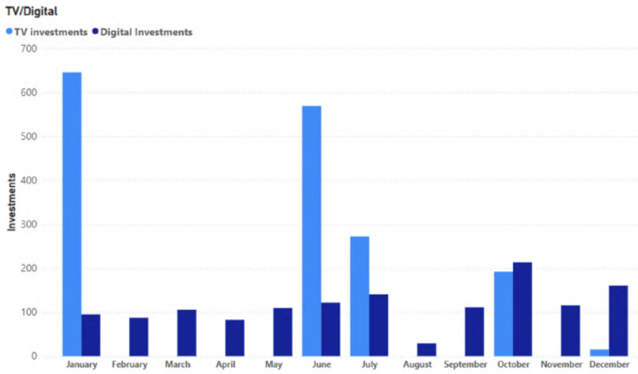

Figure 28: TV and Digital Investments across the months

The TV investment trend is characterized by sharp seasonal fluctuations with peaks at the beginning and middle of the year, followed by periods of more moderate spending, which may be part of a strategy of focusing resources on key periods and maintaining a market presence during less active months. TV investments experience substantial spikes, particularly in January and June, where expenditures exceed 600 units. These months represent the peak periods for TV spending, indicating a focused strategy to maximize reach during these times. Other notable months for TV investments include July and October, where spending remains higher than in other months, though not as high as the peaks in January and June. In contrast, digital investments remain consistently lower than TV investments throughout the year. The allocation to digital media is relatively stable, with only minor variations from month to month. Despite this consistency, digital investments never come close to the peak spending seen on TV, suggesting that digital media is used more as a supplementary channel rather than a primary focus.

4.8.5 Digital Investments Allocation

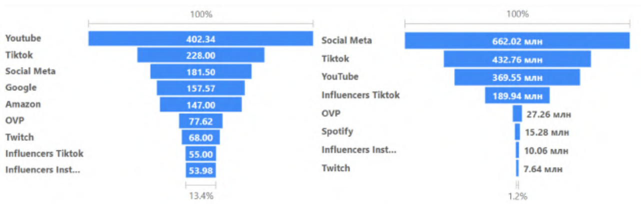

Figure 29: Digital Investments and Impressions

Social Meta and TikTok appear to be more efficient platforms in terms of generating impressions per euro spent compared to YouTube. Investing in TikTok influencers seems particularly effective, suggesting a vibrant and engaging platform for influencer activities. Lower impressions on platforms like Twitch and OVP could be strategic if targeting specific audiences, despite the higher cost per impression.

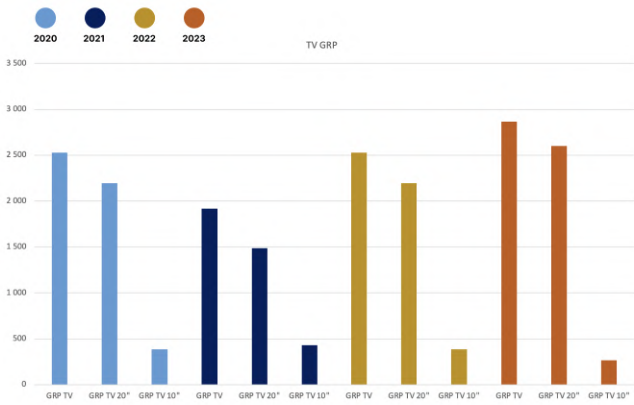

Figure 30: Enter Caption

There has been a slight increase in overall GRPs for standard TV spots over the years, suggesting either increased effectiveness or increased volume of TV spot purchases. 10-second spots have consistently shown lower GRPs over the years, suggesting they may be less preferred or have lower viewer engagement. 20-second spots have seen a significant decrease in GRPs after 2021, but have seen a significant increase in subsequent years. It's important to consider external market conditions and their impact on advertising effectiveness. Tailoring advertising strategies to current market dynamics can help maintain relevance and engagement.

5. Methodology

5.1. Data imputation

Filling missing values (NaNs) with 0 was used as an imputation method since the analysis of variables implied nans are not missing values but rather the absence of events. This approach was particularly useful for simplifying budget analysis, and cost allocation, and accurately reflecting periods when certain marketing channels were not active, such as in campaign performance metrics.

5.2. Merging the dataset

The datasets for Sales 1, Sales 2, and Sales 3 were provided separately. The information related to Digital and TV was merged with each of these Sales datasets based on the common attributes of time, year, and week number. This process ensured that the Digital and TV data were accurately aligned with the corresponding Sales data points.

5.3. Variables creation.

Based on the columns "Week Starting Sell-Out" and "Week Ending," new seasonality-related columns were created, where the time data was categorized into seasons: winter, spring, summer, and autumn. Regarding TVC data for Sales 1, Sales 2, and Sales 3, the information in this column was further divided into new categories: Copy 1, Copy 2, Copy 3, and No Copy. These categories require further analysis to better understand their impact on sales performance.

Additionally, data with similar characteristics were merged as follows:

- All digital investments and impressions were combined into 'digital investments' and 'digital impressions', respectively. The digital impressions column was created by summing the impressions from various platforms such as YouTube, TikTok, Instagram, and others. The 'digital investments' column was created by summing investments across different digital channels, including YouTube, Google Search, and influencer marketing.
- Different promotional activities, such as editorial coverage in print and online, were merged into 'editorials total', and totem display activities were summed into 'totem displays total'.
- GRP data from competitors was averaged and merged into mean 'grp competitors', which provides an average measure of competitors' advertising reach.
- The average price (for both brand and competitors) was calculated as value/volume and added as 'average price'.
- The 'tv grp' column was created by summing the GRP values across different TV advertising time slots, providing a comprehensive view of TV campaign reach.

This structured approach ensures a more streamlined and insightful analysis of the marketing mix.

5.4. Data Exploration and Outlier Detection

5.4.1 Categorical variables exploration (season, month, copy)

To enable more precise analysis, categorical values were broken down into components, and binary features with values 1 and 0 were used to simplify analysis and data interpretation.

- Seasonal Data Breakdown: Sales data for "Week Ending" and "Week Starting" were categorized into seasonal series: winter, spring, summer, and autumn. This categorization allows for a more nuanced analysis of sales trends and their correlation with different times of the year.

- "Seasonal Events" Column: A new column named "Seasonal Events" was created in both datasets based on dates.
- Categorization of TVC Values in Dataset D: The value categorized under TVC in dataset D was further subdivided into three distinct categories labeled as TV COPY 1, 2, and 3.

5.4.2 Numerical variables exploration and outliers' detection

Despite the presence of outliers in the dataset, it was decided to keep them, as their removal could potentially worsen the quality of the analysis in the future. Outliers may contain important information that reflects real-world data behavior, and excluding them could distort the results and reduce the accuracy of the model. Therefore, the decision was made to retain the outliers to ensure the completeness and accuracy of the analysis.

5.5. Variable Selection

Correlation Analysis: A correlation matrix was done to all datasets to identify pairs of variables with high correlations, using a threshold of 0.85. This step helped in detecting multi-collinearity, which can distort the results of statistical analyses and predictive models. Variables with high correlation were removed from the data sets to improve further analyses and predictions.

```python
def check_high_correlation(corr_matrix, threshold=0.85):

    # Initialize an empty list to store the results
    highly_correlated_pairs = []

    # Iterate through the correlation matrix
    for i in range(len(corr_matrix.columns)):
        for j in range(i + 1, len(corr_matrix.columns)):
            # Get the correlation value
            corr_value = corr_matrix.iloc[i, j]
            if abs(corr_value) > threshold:
                # If the correlation is above the threshold, print the variable names and correlation
                var1 = corr_matrix.columns[i]
                var2 = corr_matrix.columns[j]
                highly_correlated_pairs.append((var1, var2, corr_value))
                print(f"High correlation between {var1} and {var2}: {corr_value:.4f}")

    return highly_correlated_pairs
```

Figure 31: Correlation Analysis Function

5.6. Data sets transformation

Based on the main dataset, several transformed datasets were created:

1. Original Dataset (No Transformation).
2. Log Transformation Dataset.
3. Lagged and Adstock Dataset.
4. Log Lagged and Adstock Dataset.

These transformed datasets are created to explore different aspects of the data and enhance the accuracy and interpretability of modeling results. By comparing models built on these different datasets, it becomes possible to identify the best approach for capturing the underlying patterns and making more accurate predictions in marketing mix analysis.

5.7. Model Building

The pipeline was created to systematically explore and apply various models to ensure coherent, robust, and interpretable analysis:

1. Model Variety and Selection: The function supports multiple regression models, including OLS, Lasso, Ridge, Bayesian Ridge, Random Forest, GBM, and XGBoost, allowing flexibility in comparing and choosing the most suitable model based on the data.
2. Feature Standardization: For linear models, feature standardization is applied to optimize model performance and ensure that coefficients are on a comparable scale.

3. Feature Generation: Different non-linear modifications and timing transformations were explored to better capture the underlying dynamics and possible dependencies.
4. Feature Selection and Interpretation: The function includes feature selection, where non-significant predictors are removed to simplify and enhance the interpretability of the models.
5. Hyperparameter Tuning: The function offers the option to fine-tune hyperparameters, ensuring the models are optimized for better accuracy, robustness, and performance.
6. Model Evaluation: R2 scores are calculated for both training and testing datasets, providing clear comparable metrics of model performance. Residual normality analysis helps diagnose potential fit issues.
7. Feature Importance and SHAP Analysis: For non-linear models, feature importance and SHAP values are used to interpret the impact of variables on the model's predictions.

This approach provides a comprehensive analysis, maximizing predictive accuracy while offering clear insights into the relationships within the data.

5.8. Evaluating Contributions

After selecting the statistically optimal regression model, the contributions of each variable to overall sales were calculated, Beta coefficients of each variables were taken as a contribution. This analysis aimed to evaluate how different media channels, promotional activities, and other factors influenced the company's profits and losses. By quantifying the impact of each variable, valuable insights were gained into the effectiveness and efficiency of various marketing strategies, allowing for more informed decision-making and optimized resource allocation to maximize return on investment.

5.9. Analysis of the contribution of each TV and Digital media

Following the contribution analysis, various tactics within TV and Digital media were examined. This phase followed a similar approach to previous analyses, using standard, log, lagged adstock, and log-lag adstock transformations. For each step, the best regression model was selected based on the specific dataset, with TV and Digital media being analyzed separately. This detailed approach provided deeper insights into the individual impact of TV and Digital media channels on overall sales performance.

5.10. Computation of Effectiveness and ROI

The sales volumes generated by the overall TV and digital advertising efforts, as well as by their specific tactics, were calculated. These sales volumes were then aligned with the corresponding media responses (such as GRP, impressions, and clicks), with adjustments made for any lag effects. Following this, the effectiveness of both media channels, in aggregate and across their various tactics, was determined by calculating the ratio of generated sales volume to media response. Similarly, the sales value (in euros) generated by the total TV and digital advertising, along with their specific tactics, was computed. These values were matched with the corresponding investments, taking into account any lag effects. The return on investment (ROI) for each media channel, both overall and within specific tactics, was then calculated by dividing the generated sales value by the investment made.

5.11. Saturation Curve and Optimal Budget Allocation

Using the contributions to sales identified in the previous models, saturation curves were built for both total TV and digital activities, as well as for individual digital tactics. These curves provided insights into how different levels of investment in TV and digital marketing affect sales. The selected model was then fitted to these saturation curves to estimate the sales value generated by specific investments. The output of the model was the sales value, which depended on the amount invested in each activity. Next, the objective function was defined as the total sales value, and this function was optimized under two main constraints: (1) all budget allocations had to be positive, and (2) the total budget allocated could not exceed the available overall budget. This ensured that the budget distribution stayed within the given limits. An indicative average weekly budget was calculated by summing up the average values of weekly TV and digital budgets, providing a general guideline for weekly spending on marketing activities. Finally, an optimizer model was applied to solve the objective function under the given constraints, determining the optimal allocation of the budget to maximize sales and ensure efficient use of available resources.

5.12. Feature Interaction

The calculation of feature interaction was based on leveraging a model designed for contribution analysis. Two primary techniques were employed:

1. Feature Importance Interaction Calculation: Initially, the importance of each media channel was determined through the use of feature importance interaction coefficients. This method helped quantify the significance of each feature. Then, the interactions between different features were calculated by considering their interaction coefficients. Essentially, this involved understanding how two or more features influenced the model together rather than in isolation.
2. Feature Interaction Based on All Possible Combinations: As an alternative approach, all possible combinations of feature interactions were considered without relying on the calculation of feature importance interaction coefficients. This was done due to the observed low accuracy and potential biases in the feature importance interaction coefficients model, which sometimes did not prioritize features correctly. By considering every possible combination of features, the model was able to capture more complex interactions that might have been missed in the previous method.
3. Creating New Interaction Terms: After calculating these interactions, new values representing the interaction between different features were generated. These interaction terms were then incorporated into the existing model used for contribution analysis. The p-value selection method was applied, which helped identify statistically significant terms.
4. P-Value Selection and Model Refinement: The new interaction terms were subjected to p-value selection, wherein only significant interactions were retained. Insignificant terms, as determined by their p-values, were removed to avoid overfitting or introducing noise into the model. The process of selecting and eliminating terms based on their p-values continued iteratively. This approach allowed the model to focus on the most relevant interactions between features, ultimately leading to a refined and improved model.
5. Final Model Incorporation: The final model included interaction terms between various features that contributed to a more accurate calculation of the contribution of each media channel. By systematically refining the model based on p-values, the methodology ensured that only meaningful feature interactions were considered, leading to a more robust and reliable outcome.

This approach, which integrates both feature importance interaction and the exhaustive consideration of all feature combinations, emphasizes the importance of interaction effects in understanding the contribution of different features. By addressing the limitations of the feature importance interaction coefficients and exploring more complex relationships, the study enhances the accuracy and reliability of the contribution analysis.

6. Results

6.1. Dataset D Sales 1

6.1.1 Model Results

This regression analysis evaluates the impact of several factors on **Brand Value**. Below is a summary of the model's performance and key findings:

```
                            OLS Regression Results
==============================================================================
Dep. Variable:            Brand Value   R-squared:                       0.577
Model:                            OLS   Adj. R-squared:                  0.551
Method:                 Least Squares   F-statistic:                     22.30
Date:                Sun, 18 Aug 2024   Prob (F-statistic):           2.10e-16
Time:                        13:37:15   Log-Likelihood:                -1120.7
No. Observations:                 105   AIC:                             2255.
Df Residuals:                      98   BIC:                             2274.
Df Model:                           6
Covariance Type:            nonrobust
=======================================================================================
                          coef    std err          t      P>|t|      [0.025      0.975]
---------------------------------------------------------------------------------------
const                2.081e+05   1055.913    197.042      0.000    2.06e+05     2.1e+05
TV Budget Planned    3218.4918   1248.219      2.578      0.011     741.441    5695.543
avg_price           -6996.1999   1385.944     -5.048      0.000   -9746.561   -4245.839
avg_price_competitor 1.017e+04   1147.320      8.868      0.000    7897.218    1.25e+04
grp_erosion         -3769.0918   1260.149     -2.991      0.004   -6269.816   -1268.367
season_Winter        5540.4752   1312.613      4.221      0.000    2935.637    8145.314
digital_budget_log   2489.3607   1318.666      1.888      0.062    -127.489    5106.211
==============================================================================
Omnibus:                        3.989   Durbin-Watson:                   1.955
Prob(Omnibus):                  0.136   Jarque-Bera (JB):                4.212
Skew:                           0.185   Prob(JB):                        0.122
Kurtosis:                       3.909   Cond. No.                         2.51
==============================================================================

Notes:
[1] Standard Errors assume that the covariance matrix of the errors is correctly specified.
R2 Score (Train): 0.5759278798857231
R2 Score (Test): 0.4309487489177526
Mean Squared Error: 195255684.60792235
```

Figure 32: Dataset D Sales 1 Model

Model Overview

- **R-squared = 0.577**: The model explains about 57.7% of the variability in Brand Value, suggesting a moderate fit.
- **F-test = 22.30, p-value < 0.0001**: The overall model is statistically significant, indicating that the independent variables significantly predict Brand Value.

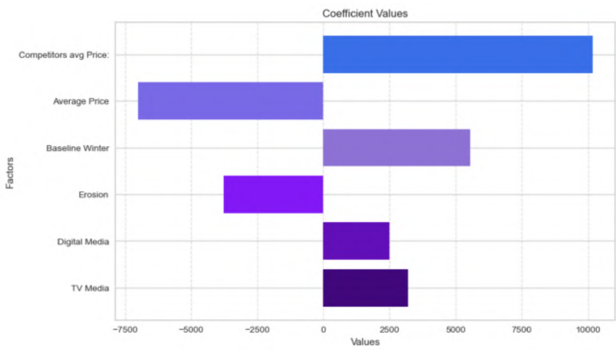

Figure 33: Plot of the coefficients results

The Contribution Coefficients were based on the beta coefficients obtained from the model.

Key Findings

- **TV Budget (coef = 3218.49, p = 0.011)**: Each unit increase in TV budget is associated with a 3,218-unit increase in Brand Value. This relationship is statistically significant.
- **Average Price (coef = -6996.20, p < 0.001)**: An increase in the average price leads to a decrease in Brand Value by 6,996 units, showing a negative impact.

- **Competitor's Average Price (coef = 10,170, p < 0.001)**: When competitors increase their prices, Brand Value increases by 10,170 units, suggesting a competitive advantage.
- **GRP Erosion (coef = -3769.09, p = 0.004)**: Higher GRP erosion results in a 3,769-unit reduction in Brand Value, indicating a negative effect.
- **Seasonality (Winter, coef = 5540.47, p < 0.001)**: During the winter, Brand Value increases by 5,540 units, highlighting a positive seasonal influence.
- **Digital Budget (log-transformed, coef = 2489.36, p = 0.062)**: Digital budget has a positive impact, though the effect is only marginally significant.

Model Diagnostics

- **Durbin-Watson = 1.955**: This value indicates that there is no autocorrelation in the residuals, which is ideal for this type of model.
- **Omnibus and Jarque-Bera Tests**: The residuals appear to be normally distributed, meaning the model assumptions are reasonably met.
- **Condition Number = 2.51**: The low condition number suggests there are no multicollinearity issues, which indicates stability in the model.

Model Performance

- **Train R^2 = 0.576, Test R^2 = 0.431**: The model performs moderately well on the training data, but there is a noticeable drop in test performance, indicating potential overfitting.
- **Mean Squared Error = 195,255,685**: This is the average squared difference between observed and predicted Brand Value, suggesting there is still room for improvement in the model's predictions.

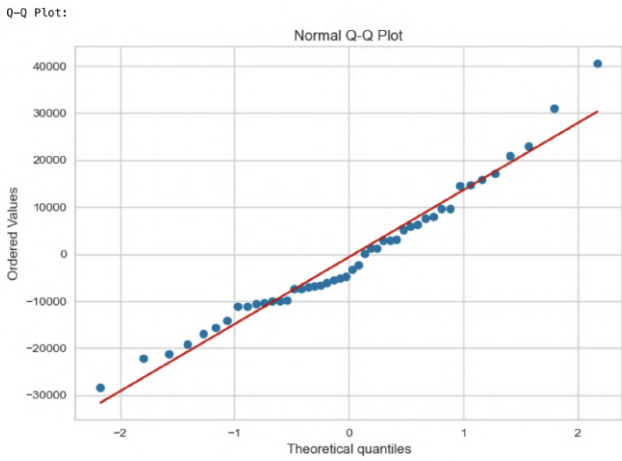

Figure 34: QQ-Plot Dataset D Sales 1

In this case, most of the points lie on or near the reference line, suggesting that the residuals are approximately normally distributed for the majority of the data. This is an indication that the model residuals exhibit normal behavior for much of the dataset, which supports the assumption of normally distributed errors in the OLS regression model. However, there are deviations at the tails, suggesting some non-normality or outliers. While the normality assumption is mostly satisfied, these extreme values may affect hypothesis tests and confidence intervals.

Conclusion: In summary, the model provides useful insights into the factors driving Brand Value. TV and digital ad budgets have a positive influence, as does higher competitor pricing. However, higher prices tend to negatively affect Brand Value, while seasonality, especially in winter, positively impacts it. Although the model fits the data reasonably well, improvements could be made to enhance predictive accuracy, as indicated by the drop in test performance.

6.1.2 Contribution of Digital and TV Media

Based on the contribution of Digital and TV Media, the contribution of each media was calculated to find how much profit each of media generated. This procedure is important to calculate ROI and Effectiveness.

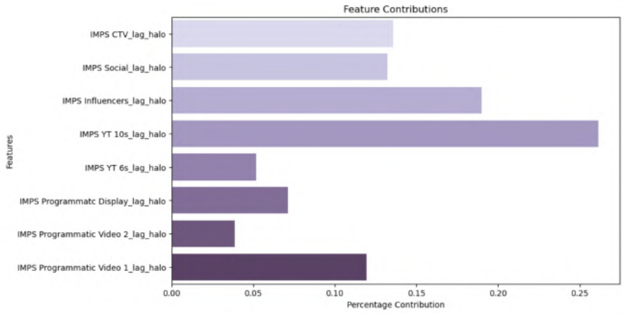

Figure 35: Contribution of Digital Media

- The percentage contribution to the sales of IMPS CTV lag halo is: 0.13560687149219816
- The percentage contribution to the sales of IMPS Social lag halo is: 0.13206047120481096
- The percentage contribution to the sales of IMPS Influencers lag halo is: 0.18992366030245889
- The percentage contribution to the sales of IMPS YT 10s lag halo is: 0.2614145161329172
- The percentage contribution to the sales of IMPS YT 6s lag halo is: 0.051771701153604716
- The percentage contribution to the sales of IMPS Programmatc Display lag halo is: 0.07136215880357157
- The percentage contribution to the sales of IMPS Programmatic Video 2 lag halo is: 0.03854720177476336
- The percentage contribution to the sales of IMPS Programmatic Video 1 lag halo is: 0.11931341913567517

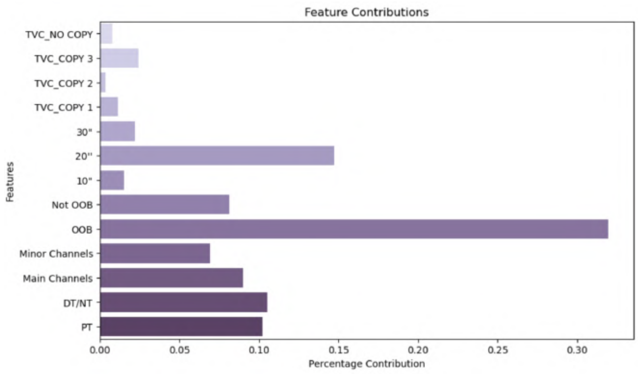

Figure 36: Contribution of TV media

- The percentage contribution to the sales of TVC NO COPY is: 0.007967709374049009
- The percentage contribution to the sales of TVC COPY 3 is: 0.024477953908246027
- The percentage contribution to the sales of TVC COPY 2 is: 0.0034832402751496052
- The percentage contribution to the sales of TVC COPY 1 is: 0.011386909409932248
- The percentage contribution to the sales of 30" is: 0.022392910870018927
- The percentage contribution to the sales of 20” is: 0.14733085286449174
- The percentage contribution to the sales of 10" is: 0.015366827274722111
- The percentage contribution to the sales of Not OOB is: 0.08135188278093707
- The percentage contribution to the sales of OOB is: 0.31933275540395084
- The percentage contribution to the sales of Minor Channels is: 0.06938872927997125
- The percentage contribution to the sales of Main Channels is: 0.09002446320574672
- The percentage contribution to the sales of DT/NT is: 0.10531384521815781
- The percentage contribution to the sales of PT is: 0.10218192013462671

6.1.3 Media Effectiveness and ROI

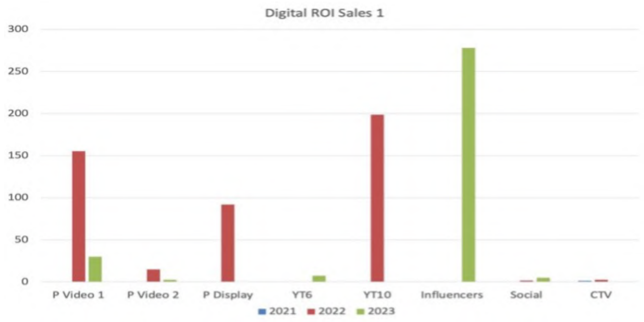

Figure 37: Digital ROI

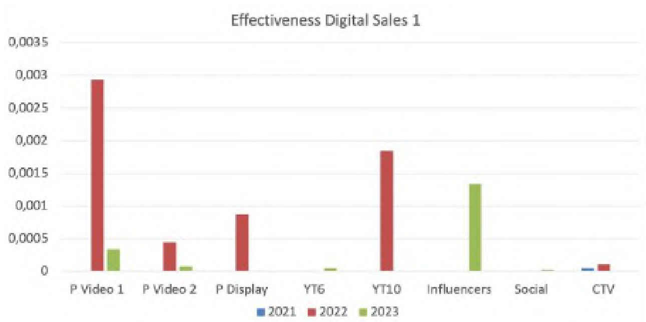

Figure 38: Digital Effectiveness

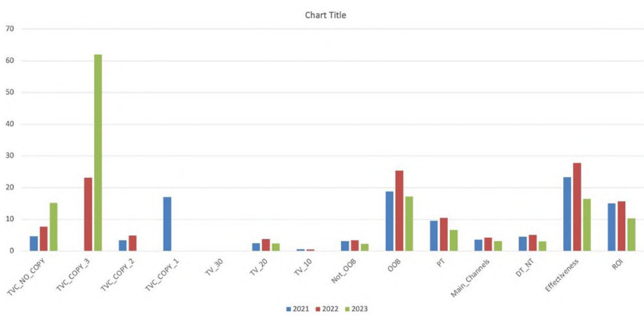

Figure 39: TV Effectiveness and ROI

- **ROI Influencers**: Demonstrates the highest ROI, indicating that marketing strategies involving influencers yield the greatest return on investment.
- **ROI YouTube 10**: Shows significant ROI, suggesting that specific YouTube strategies are delivering substantial returns.
- **ROI Programmatic Video 1**: Provides moderate returns, reflecting effectiveness but with room for optimization.
- **ROI Programmatic Display**: Also shows moderate returns.
- **ROI Connected TV (CTV)**: Displays lower ROI, indicating that this platform might require reassessment.
- **ROI Social**: Exhibits lower ROI, suggesting the need for strategic adjustment.
- **ROI YouTube 6**: Shows lower ROI, indicating potential for improvement.
- **Effectiveness Programmatic Video 1**: Leads with the highest effectiveness, indicating strong engagement or audience reach.

- **Effectiveness YouTube 10 and Effectiveness Influencers**: Also show strong performance, making them valuable in the digital marketing strategy.
- **Effectiveness Programmatic Display**: Shows middle performance, suggesting areas for improvement.
- **Other Effectiveness Metrics**: Indicate lower performance, pointing to potential areas for strategic enhancement.
- **TVC_COPY_3**: Shows a sharp increase in effectiveness in 2023, significantly outperforming other years.
- **TVC_COPY_2**: Exhibits moderate effectiveness, with 2022 being the most successful year.
- **TVC_COPY_1**: Only recorded in 2021 with significant impact, indicating strong campaigns that may have been discontinued or restructured.
- **TV_30, TV_20, TV_10**: Show low to moderate effectiveness, with TV_20 performing better than TV_30 and TV_10.
- **Not_OOB, Main_Channels, DT_NT**: Show stable, moderate effectiveness.
- **OOB**: Shows higher effectiveness, particularly in 2022.
- **Total Effectiveness and Total ROI**: Both exhibit steady growth from 2021 to 2023, with peak values in 2022.

6.1.4 Saturation Curve

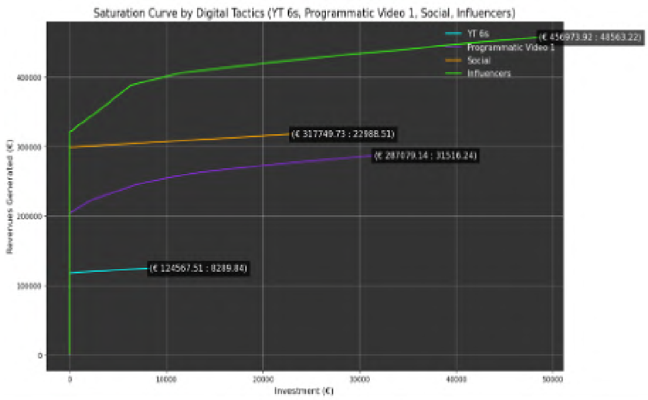

Figure 40: Saturation Curve Digital Media

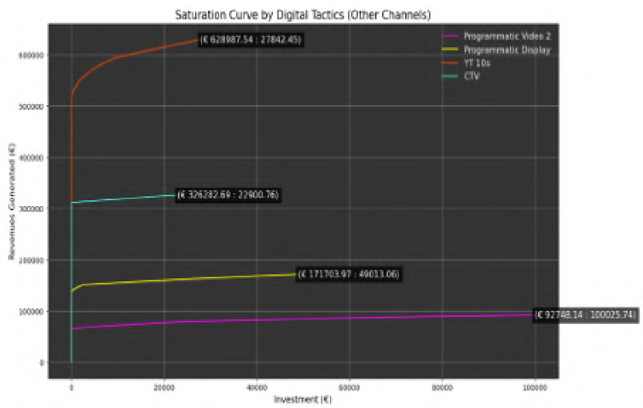

Figure 41: Saturation Curve Digital Media

Saturation curves in digital media advertising illustrate the relationship between the amount of money invested in different digital tactics and the revenue generated from these investments.

First Chart Analysis: Specific Digital Tactics (YT 6s, Programmatic Video 1, Social, Influencers)

- **YouTube 6 seconds (YT 6s)**: With an investment of €8,289.84, the revenue generated is €124,567.51. The curve shows a steep initial increase in revenue with rising investments but levels off eventually, indicating a saturation point where additional investment does not significantly increase revenue.
- **Programmatic Video 1**: The purple curve shows that an investment of €31,516.24 generates revenue of €287,079.14. The curve rises more gradually compared to YT 6s and starts to plateau at higher investment levels, suggesting steady revenue increase but with diminishing returns.
- **Social Media**: The orange curve indicates that an investment of €22,988.91 yields €317,749.73 in revenue. The curve shows a moderate increase in revenue with investment and a slower plateau, suggesting less sensitivity to increased investment and more gradual returns.
- **Influencers**: The green curve illustrates that a significant investment of €48,563.22 generates €456,973.92 in revenue. This curve is relatively flat, indicating that even with considerable investments, revenue increments are moderate, suggesting quick saturation and reduced cost-effectiveness.

Second Chart Analysis: Other Digital Channels (Programmatic Video 2, Programmatic Display, YT 10s, CTV)

- **Programmatic Video 2**: Demonstrates a sharp increase and early saturation, meaning it quickly reaches a point where additional investment yields minimal to no increase in revenue.
- **Programmatic Display**: Exhibits a flatter curve, indicating that revenue does not significantly increase with more investment, suggesting it is either saturated or inefficient in terms of ROI.
- **Connected TV (CTV)**: Shows a similar pattern to Programmatic Video 2 but with a slightly delayed saturation point, indicating somewhat better efficiency while still reaching saturation quickly.
- **YouTube 10 seconds (YT 10s)**: Displays the most gradual curve, suggesting steady growth in ROI without quick saturation.

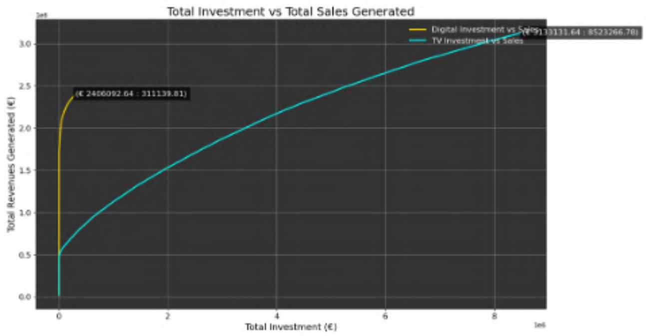

Figure 42: Digital vs TV media saturation curve

The digital investment curve shows a sharp rise in sales with initial investments, reflecting high efficiency and precision in digital media targeting. However, as investment increases, the returns demonstrate diminishing growth, pointing to a saturation point. In contrast, television media investment shows a more gradual and consistent increase in sales but yields less return compared to digital media for the same investment amount. This highlights that while digital media provides superior returns, particularly at lower investment levels, television media remains valuable for broader brand exposure. The revenue side of the curve has been adjusted, which may affect the precise assessment of the optimal saturation point.

6.1.5 Budget Optimization

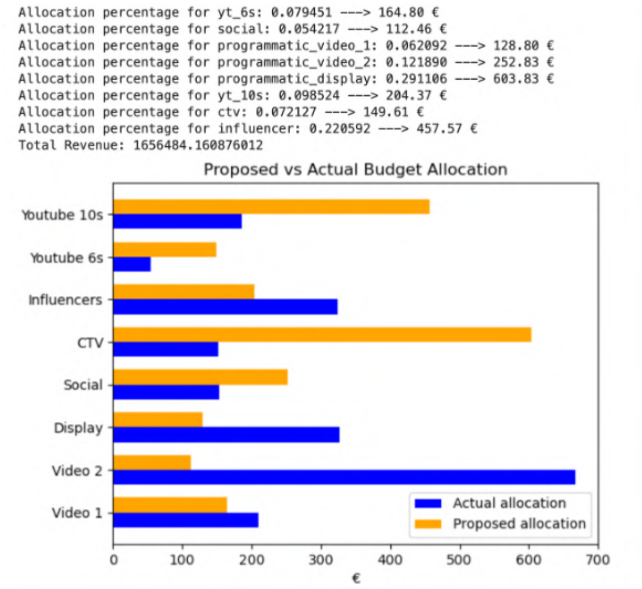

Figure 43: Optimal Digital Budget Allocation

- Budget Allocation: Influencers and Programmatic Display received more funding than planned, while YouTube 6s, YouTube 10s, Social Media, and CTV were underfunded.
- Actual Spending vs. Planned: Shows a significant deviation with total revenue generated being €1,656,484.16, highlighting the need for better alignment with budget allocations.
- Overall Spending: Exceeds initial plans, with total revenue of €2,394,096.624, suggesting a need for a review of budgeting practices.

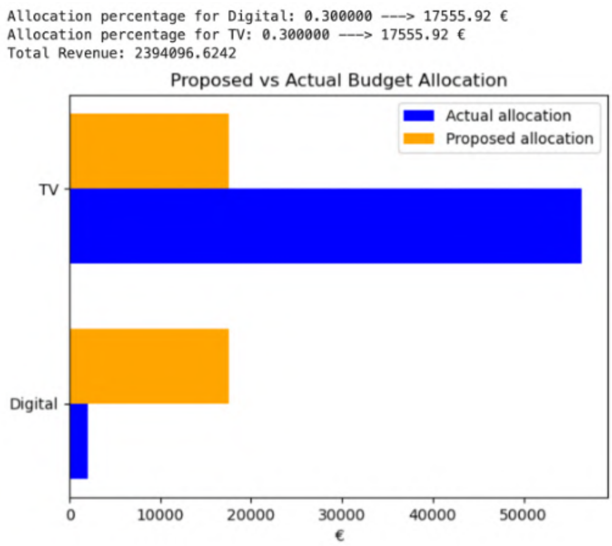

Figure 44: Enter Caption

The actual budget allocation for both Digital and TV media significantly exceeds the proposed amounts, with each initially planned at 17,555.92 €. This indicates a substantial deviation from the budget, suggesting increased spending or cost overruns. The total revenue generated is 2,394,096.624 €, highlighting the financial impact of these investments.

6.1.6 Dataset D Sales 1 Interaction

The similar approach as for Digital Impressions was applied to this dataset due to the fact that the interaction-based model did not demonstrate high results on either the training (y_train) or test (y_test) samples, it was decided to include all possible interactions for a more detailed analysis of their impact on the overall performance to the original model. The model explains approximately 62.7% of the variance in Brand Value, as indicated by the R-squared value of 0.627. This suggests a reasonably good fit of the model to the data.

```
                             OLS Regression Results
==============================================================================
Dep. Variable:            Brand Value   R-squared:                       0.627
Model:                            OLS   Adj. R-squared:                  0.587
Method:                 Least Squares   F-statistic:                     15.78
Date:                Fri, 06 Sep 2024   Prob (F-statistic):           3.14e-16
Time:                        20:55:59   Log-Likelihood:                -1114.2
No. Observations:                 105   AIC:                             2250.
Df Residuals:                      94   BIC:                             2280.
Df Model:                          10
Covariance Type:            nonrobust
=========================================================================================================================
                                                  coef    std err          t      P>|t|      [0.025      0.975]
-------------------------------------------------------------------------------------------------------------------------
const                                        2.081e+05   1013.105    205.368      0.000    2.06e+05     2.1e+05
TV Budget Planned                            3359.4050   1212.906      2.770      0.007     951.152    5767.658
avg_price                                   -7352.0979   1384.754     -5.309      0.000   -1.01e+04   -4602.636
avg_price_competitor                         1.045e+04   1136.043      9.199      0.000    8194.675    1.27e+04
grp_erosion                                 -3617.6632   1244.468     -2.907      0.005   -6088.585   -1146.742
season_Winter                                5573.8662   1346.546      4.139      0.000    2900.268    8247.465
digital_budget_log                           3189.5193   1327.982      2.402      0.018     552.780    5826.259
Programmatic Video 1 * Budget YT 6s         -2.956e+04   1.18e+04     -2.515      0.014   -5.29e+04   -6224.175
Programmatic Video 1 * Budget Influencers    2.182e+04   1.11e+04      1.962      0.053    -266.837    4.39e+04
Budget Programmatic Video 2 * Budget YT 6s   2.127e+04   8008.673      2.656      0.009    5370.262    3.72e+04
Budget Programmatic Video 2 * Budget Influencers -1.662e+04   7882.164     -2.108      0.038   -3.23e+04    -964.927
==============================================================================
Omnibus:                        5.759   Durbin-Watson:                   1.941
Prob(Omnibus):                  0.056   Jarque-Bera (JB):                6.685
Skew:                           0.289   Prob(JB):                       0.0353
Kurtosis:                       4.092   Cond. No.                         36.7
==============================================================================

Notes:
[1] Standard Errors assume that the covariance matrix of the errors is correctly specified.
R2 Score (Train): 0.6049858632577123
R2 Score (Test): 0.4528503183041809
Mean Squared Error: 187740709.6097971
```

Figure 45: Digital Media Interactions Model

- **R-squared**: 0.627
- **Adjusted R-squared**: 0.587
- **F-statistic**: 15.78 with a p-value of 3.14×10^{-16}
  - The F-statistic tests whether at least one of the predictors is significantly related to Brand Value. The p-value is very low, indicating that the model as a whole is statistically significant.

Coefficients and Significance

- **Intercept (const)**: 208,100
  - This represents the expected Brand Value when all predictors are zero.
- **TV Budget Planned**: 3,359.41
  - A significant positive coefficient (p-value = 0.007) indicates that increasing the planned TV budget is associated with an increase in Brand Value.
- **avg_price**: -7,352.10
  - A significant negative coefficient (p-value $<$ 0.000) suggests that higher average prices are associated with lower Brand Value.
- **avg_price_competitor**: 10,450.00
  - A significant positive coefficient (p-value $<$ 0.000) shows that higher competitor prices are positively related to Brand Value.
- **grp_erosion**: -3,617.66
  - A significant negative coefficient (p-value = 0.005) indicates that erosion in group effectiveness reduces Brand Value.
- **season_Winter**: 5,573.87
  - A significant positive coefficient (p-value $<$ 0.000) suggests that Brand Value is higher in the winter season.
- **digital_budget_log**: 3,189.52
  - A significant positive coefficient (p-value = 0.018) indicates that an increase in the logged digital budget is associated with higher Brand Value.
- **Programmatic Video 1 * Budget YT 6s**: -29,560.00

- A significant negative interaction term (p-value = 0.014) suggests that the effectiveness of Programmatic Video 1 decreases when combined with a high budget for YouTube 6 seconds.

- **Programmatic Video 1 * Budget Influencers**: 21,820.00
  - A positive but marginally significant interaction term (p-value = 0.053) indicates that higher budgets for influencers positively impact the effectiveness of Programmatic Video 1.
- **Budget Programmatic Video 2 * Budget YT 6s**: 21,270.00
  - A significant positive interaction term (p-value = 0.009) suggests that higher budgets for Programmatic Video 2 are positively related to a higher budget for YouTube 6 seconds.
- **Budget Programmatic Video 2 * Budget Influencers**: -16,620.00
  - A significant negative interaction term (p-value = 0.038) indicates that the combination of higher budgets for Programmatic Video 2 and influencers negatively impacts Brand Value.

Diagnostics

- **Omnibus**: 5.759 (p-value = 0.056)
  - Indicates some deviation from normality in the residuals, though the p-value is close to the threshold of 0.05.
- **Durbin-Watson**: 1.941
  - Close to 2, suggesting no significant autocorrelation in the residuals.
- **Jarque-Bera (JB)**: 6.685 (p-value = 0.0353)
  - Suggests some departure from normality in the residuals.

Model Performance

- **R2 Score (Train)**: 0.605
- **R2 Score (Test)**: 0.453
  - The R2 score on the training data is higher than that on the test data, indicating some overfitting. The test R2 score of 0.453 suggests that the model performs reasonably well on unseen data but has room for improvement.
- **Mean Squared Error (MSE)**: 187,740,709.61
  - This value represents the average squared difference between the predicted and actual Brand Values, providing a measure of the model's prediction error.

6.2. Dataset D Sales 2

6.2.1 Model Results

```
                            OLS Regression Results
==============================================================================
Dep. Variable:            Brand Value   R-squared:                       0.700
Model:                            OLS   Adj. R-squared:                  0.673
Method:                 Least Squares   F-statistic:                     25.44
Date:                Sun, 18 Aug 2024   Prob (F-statistic):           3.45e-24
Time:                        15:07:28   Log-Likelihood:                -1002.7
No. Observations:                 120   AIC:                             2027.
Df Residuals:                     109   BIC:                             2058.
Df Model:                          10
Covariance Type:            nonrobust
====================================================================================================
                                     coef    std err          t      P>|t|      [0.025      0.975]
----------------------------------------------------------------------------------------------------
const                           2.315e+04     98.615    234.767      0.000     2.3e+04    2.33e+04
BRAND Weighted Distribution      510.1446    145.715      3.501      0.001     221.341     798.948
BRAND VOLUME IN PROMO            263.5179    107.484      2.452      0.016      50.489     476.547
COMPETITOR Numeric Distribution  353.7357    109.405      3.233      0.002     136.898     570.574
Seasonal Holidays               -372.8692    102.966     -3.621      0.000    -576.944    -168.794
digital_imps                    -517.8382    161.742     -3.202      0.002    -838.405    -197.271
avg_price                        507.2043    185.004      2.742      0.007     140.532     873.876
avg_price_competitor             492.6482    114.276      4.311      0.000     266.157     719.139
Sales Value Growth (%)           919.7926    102.640      8.961      0.000     716.363    1123.222
season_Spring                    977.8231    132.226      7.395      0.000     715.755    1239.891
season_Winter                   1188.4187    156.920      7.573      0.000     877.409    1499.428
==============================================================================
Omnibus:                        2.180   Durbin-Watson:                   1.833
Prob(Omnibus):                  0.336   Jarque-Bera (JB):                1.884
Skew:                           0.040   Prob(JB):                        0.390
Kurtosis:                       3.609   Cond. No.                         4.00
==============================================================================

Notes:
[1] Standard Errors assume that the covariance matrix of the errors is correctly specified.
R2 Score (Train): 0.6975685124967272
R2 Score (Test): 0.293893873750802
Mean Squared Error: 1867749.439644538
```

Figure 46: Dataset D Sales 2 Model

The model explains approximately 70.0% of the variance in Brand Value, as indicated by the R-squared value of 0.700. This suggests a strong fit of the model to the data.

- **R-squared**: 0.700
  - This indicates that about 70.0% of the variability in Brand Value is explained by the model.
- **Adjusted R-squared**: 0.673
  - Adjusted R-squared accounts for the number of predictors in the model. A value of 0.673 suggests that after adjusting for the number of predictors, the model still explains a substantial portion of the variance.
- **F-statistic**: 25.44 with a p-value of 3.45×10^{-24}
  - The F-statistic tests whether at least one of the predictors is significantly related to Brand Value. The very low p-value indicates that the model as a whole is statistically significant.

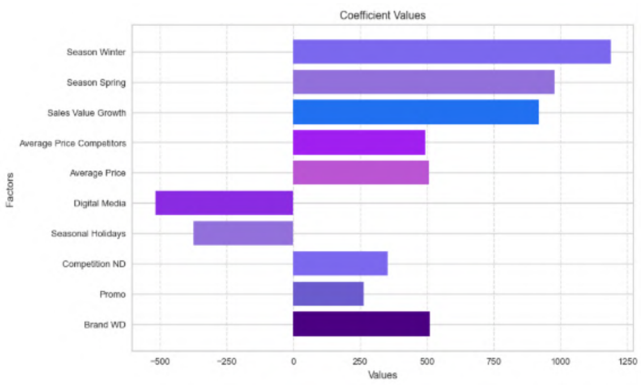

Figure 47: Plot of Contribution Coefficients

Coefficients and Significance

- **Intercept (const)**: 23,150
  - This represents the expected Brand Value when all predictors are zero.
- **BRAND Weighted Distribution**: 510.14
  - A significant positive coefficient (p-value = 0.001) indicates that higher BRAND Weighted Distribution is associated with a higher Brand Value.
- **BRAND VOLUME IN PROMO**: 263.52
  - A significant positive coefficient (p-value = 0.016) suggests that increased promotional volume is positively related to Brand Value.
- **COMPETITOR Numeric Distribution**: 353.74
  - A significant positive coefficient (p-value = 0.002) shows that higher numeric distribution by competitors is associated with higher Brand Value.
- **Seasonal Holidays**: -372.87
  - A significant negative coefficient (p-value $<$ 0.000) indicates that Brand Value is lower during seasonal holidays.
- **digital_imps**: -517.84
  - A significant negative coefficient (p-value = 0.002) suggests that higher digital impressions are negatively related to Brand Value.
- **avg_price**: 507.20
  - A significant positive coefficient (p-value = 0.007) indicates that higher average prices are associated with higher Brand Value.
- **avg_price_competitor**: 492.65
  - A significant positive coefficient (p-value $<$ 0.000) shows that higher competitor prices are positively related to Brand Value.
- **Sales Value Growth (%)**: 919.79
  - A highly significant positive coefficient (p-value $<$ 0.000) indicates that a higher percentage of sales value growth is associated with a higher Brand Value.
- **season_Spring**: 977.82

- A significant positive coefficient (p-value < 0.000) suggests that Brand Value is higher in the spring season.

- **season_Winter**: 1,188.42
  - A significant positive coefficient (p-value < 0.000) indicates that Brand Value is higher in the winter season.

Diagnostics

- **Omnibus**: 2.180 (p-value = 0.336)
  - Indicates that the residuals do not significantly deviate from normality.
- **Durbin-Watson**: 1.833
  - Close to 2, suggesting no significant autocorrelation in the residuals.
- **Jarque-Bera (JB)**: 1.884 (p-value = 0.390)
  - Suggests that the residuals are normally distributed.

Model Performance

- **R2 Score (Train)**: 0.698
- **R2 Score (Test)**: 0.294
  - The R2 score on the training data is significantly higher than that on the test data, indicating potential overfitting. The test R2 score of 0.294 suggests that while the model performs reasonably well on unseen data, there is substantial room for improvement.
- **Mean Squared Error (MSE)**: 1,867,749.44
  - This value represents the average squared difference between the predicted and actual Brand Values, providing a measure of the model's prediction error.

6.3. Dataset D Sales 3

6.3.1 Model Results

The model demonstrates a strong fit, explaining approximately 85.1% of the variance in Brand Value, as indicated by the R-squared value of 0.851.

```
                            OLS Regression Results
==============================================================================
Dep. Variable:            Brand Value   R-squared:                       0.851
Model:                            OLS   Adj. R-squared:                  0.834
Method:                 Least Squares   F-statistic:                     48.34
Date:                Sun, 18 Aug 2024   Prob (F-statistic):           1.32e-33
Time:                        17:49:58   Log-Likelihood:                -774.80
No. Observations:                 105   AIC:                             1574.
Df Residuals:                      93   BIC:                             1605.
Df Model:                          11
Covariance Type:            nonrobust
=========================================================================================================
                                            coef    std err          t      P>|t|      [0.025      0.975]
---------------------------------------------------------------------------------------------------------
const                                  1.674e+04     40.198    416.536      0.000    1.67e+04    1.68e+04
TV Budget Planned                       114.6594     51.841      2.212      0.029      11.713     217.606
BRAND Weighted Distribution             801.8339     72.161     11.112      0.000     658.536     945.132
BRAND VOLUME IN PROMO                   134.1854     46.883      2.862      0.005      41.085     227.286
COMPETITOR Numeric Distribution         738.4071     80.281      9.198      0.000     578.985     897.829
COMPETITOR Weighted Distribution       -213.8900     58.689     -3.644      0.000    -330.434     -97.346
COMPETITOR VOLUME IN PROMO             -106.9698     44.797     -2.388      0.019    -195.927     -18.012
avg_price                               882.1733     85.675     10.297      0.000     712.039    1052.308
grp_erosion                            -105.2378     49.498     -2.126      0.036    -203.531      -6.945
Sales Value Growth (%)                  303.1853     47.911      6.328      0.000     208.043     398.328
season_Spring                           159.0567     48.658      3.269      0.002      62.432     255.682
season_Winter                           290.9343     55.449      5.247      0.000     180.824     401.045
==============================================================================
Omnibus:                        2.919   Durbin-Watson:                   2.131
Prob(Omnibus):                  0.232   Jarque-Bera (JB):                2.351
Skew:                           0.251   Prob(JB):                        0.309
Kurtosis:                       3.534   Cond. No.                         5.14
==============================================================================

Notes:
[1] Standard Errors assume that the covariance matrix of the errors is correctly specified.
R2 Score (Train): 0.8501945452772898
R2 Score (Test): 0.8244795992725676
Mean Squared Error: 157114.99390876084
```

Figure 48: Dataset D Sales 3 Model

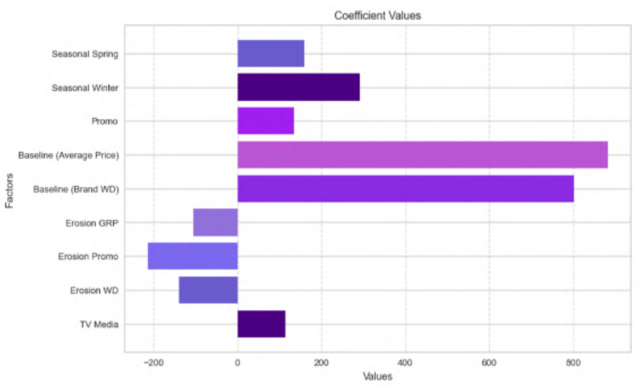

Figure 49: Variables Contribution

```
coefficients = {
    "TV Media": 114.6594, #TV budget
    "Erosion WD": -140.3374, #Competitor WD
    "Erosion Promo": -213.8900,
    "Erosion GRP": -105.2378,
    "Baseline (Brand WD)": 801.8339, #Brand WD
    "Baseline (Average Price)": 882.1733, #average price
    "Promo": 134.1854, #BRAND VOLUME IN PROMO
    "Seasonal Winter": 290.9343,
    "Seasonal Spring": 159.0567
}
```

Figure 50: Contribution Coefficients

- **R-squared**: 0.851
  - This indicates that about 85.1% of the variability in Brand Value is explained by the model.
- **Adjusted R-squared**: 0.834
  - Adjusted R-squared, which accounts for the number of predictors, is 0.834, indicating that the model remains strong after adjusting for the number of predictors.
- **F-statistic**: 48.34 with a p-value of 1.32×10^{-33}
  - The high F-statistic value and very low p-value indicate that the model as a whole is statistically significant and that at least one of the predictors is significantly related to Brand Value.

Coefficients and Significance

- **Intercept (const)**: 16,740
  - This represents the expected Brand Value when all predictors are zero.
- **TV Budget Planned**: 114.66
  - A significant positive coefficient (p-value = 0.029) indicates that higher planned TV budget is associated with a higher Brand Value.
- **BRAND Weighted Distribution**: 801.83
  - A significant positive coefficient (p-value $<$ 0.000) suggests that higher BRAND Weighted Distribution positively impacts Brand Value.
- **BRAND VOLUME IN PROMO**: 134.19
  - A significant positive coefficient (p-value = 0.005) indicates that increased promotional volume is associated with higher Brand Value.
- **COMPETITOR Numeric Distribution**: 738.41
  - A significant positive coefficient (p-value $<$ 0.000) shows that higher competitor numeric distribution is positively related to Brand Value.
- **COMPETITOR Weighted Distribution**: -213.89
  - A significant negative coefficient (p-value $<$ 0.000) indicates that higher competitor weighted distribution is associated with a lower Brand Value.

- **COMPETITOR VOLUME IN PROMO**: -106.97
  - A significant negative coefficient (p-value = 0.019) suggests that higher competitor promotional volume is negatively related to Brand Value.
- **avg_price**: 882.17
  - A significant positive coefficient (p-value < 0.000) indicates that higher average prices are associated with higher Brand Value.
- **grp_erosion**: -105.24
  - A significant negative coefficient (p-value = 0.036) suggests that greater group erosion is associated with lower Brand Value.
- **Sales Value Growth (%)**: 303.19
  - A highly significant positive coefficient (p-value < 0.000) shows that higher sales value growth percentage is associated with higher Brand Value.
- **season_Spring**: 159.06
  - A significant positive coefficient (p-value = 0.002) indicates that Brand Value is higher in the spring season.
- **season_Winter**: 290.93
  - A significant positive coefficient (p-value < 0.000) suggests that Brand Value is higher in the winter season.

Diagnostics

- **Omnibus**: 2.919 (p-value = 0.232)
  - Indicates that the residuals do not significantly deviate from normality.
- **Durbin-Watson**: 2.131
  - Close to 2, suggesting no significant autocorrelation in the residuals.
- **Jarque-Bera (JB)**: 2.351 (p-value = 0.309)
  - Suggests that the residuals are normally distributed.

Model Performance

- **R2 Score (Train)**: 0.850
- **R2 Score (Test)**: 0.824
  - The R2 score on the training data is close to that on the test data, indicating a well-fitting model with good generalization to unseen data.
- **Mean Squared Error (MSE)**: 157,114.99
  - This value represents the average squared difference between the predicted and actual Brand Values, providing a measure of the model's prediction accuracy.

6.3.2 Contribution of TV media

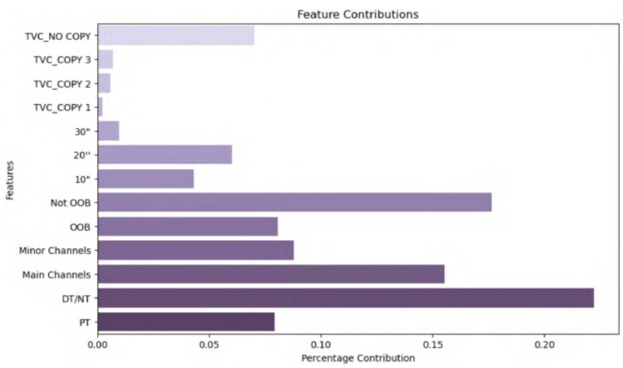

Figure 51: Contribution of TV media

- The percentage contribution to the sales of TVC_NO COPY is: 0.07028731016178838

- The percentage contribution to the sales of TVC_COPY 3 is: 0.007053790336151233
- The percentage contribution to the sales of TVC_COPY 2 is: 0.005770959043594972
- The percentage contribution to the sales of TVC_COPY 1 is: 0.002074189335187907
- The percentage contribution to the sales of 30" is: 0.00987959104443382
- The percentage contribution to the sales of 20” is: 0.06036398253927073
- The percentage contribution to the sales of 10" is: 0.04316674380259249
- The percentage contribution to the sales of Not OOB is: 0.1763684144144129
- The percentage contribution to the sales of OOB is: 0.08062326154779777
- The percentage contribution to the sales of Minor Channels is: 0.08781791774170421
- The percentage contribution to the sales of Main Channels is: 0.15539012543874425
- The percentage contribution to the sales of DT/NT is: 0.22205842754969413
- The percentage contribution to the sales of PT is: 0.07914528704462723

6.3.3 Media Effectiveness and ROI

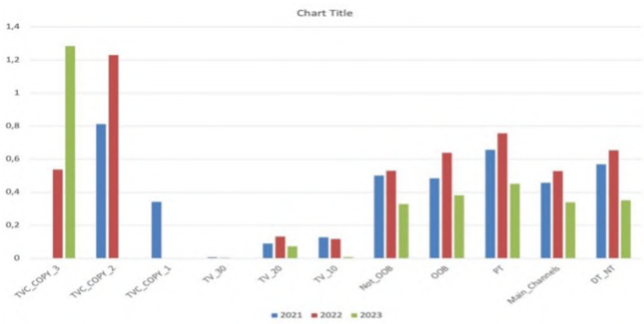

Figure 52: TV Effectiveness by tactics

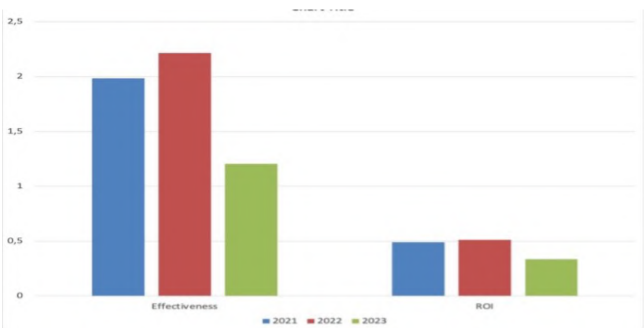

Figure 53: TV Effectiveness and ROI

- **DT/NT** shows a high effectiveness, indicating its significant role in the overall strategy.
- **Main Channels** and **Minor Channels** also exhibit strong effectiveness, though slightly less than DT/NT. This suggests they are essential but may not be as impactful.
- **PT** stands out with a very high effectiveness score, nearly reaching the maximum value. This highlights PT as a highly effective channel, likely driving substantial results.
- **Not OOB** and **OOB** channels perform well, though slightly less effectively compared to PT.
- Channels **10** and **20** show lower effectiveness, with **30** being particularly low. This indicates that these channels might be underperforming or not utilized to their full potential. The low effectiveness of channel 30 may require further review or alternative strategies.
- The **TVC** channels present a mixed performance. **TVC 3** and **TVC 2** are highly effective, with TVC 3 showing the highest effectiveness across all parameters. This suggests that TVC 3 is a key driver of success. However, **TVC 1** shows much lower effectiveness, implying it may not be contributing as expected and might need adjustments.

Overall, while most channels perform well, attention is needed on the lower effectiveness channels, particularly channels 10 and TVC 1. Improving these areas could enhance the overall effectiveness of the strategy.

6.3.4 Saturation Curve

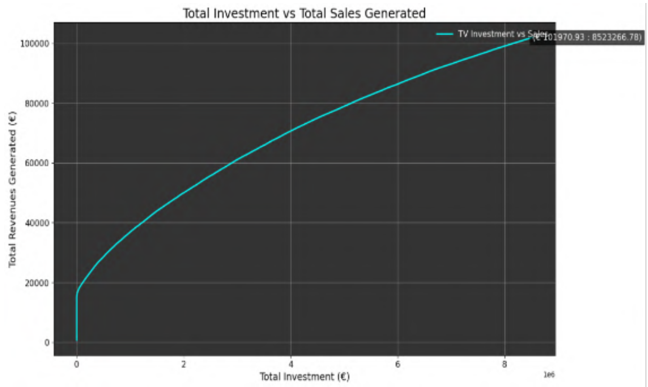

Figure 54: Saturation Curve

As investment increases, the rate of revenue growth slows down, indicating a typical saturation effect where additional investments yield diminishing returns. However, it's important to note that the revenue side of the curve has been adjusted, meaning we cannot accurately assess how optimal the saturation curve truly is. This adjustment likely occurred multiple times, potentially 100 or even 1000 times, to reach the current representation. Consequently, while the curve suggests saturation, the exact effectiveness of further investments remains unclear due to these adjustments.

6.3.5 Budget Optimization

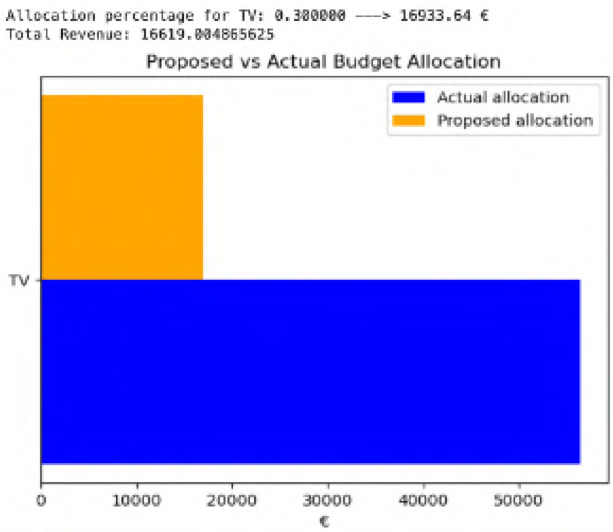

Figure 55: Optimal Budget Allocation

The proposed allocation is significantly smaller, around €16,933.64, compared to the actual allocation which is much larger, exceeding €50,000. This suggests that the actual spending on TV is much higher than initially proposed, indicating either a strategic decision to invest more heavily in TV or a deviation from the original budget plan. The total revenue generated, as indicated, is €16,619.00, which raises questions about the efficiency and effectiveness of the higher actual spending on TV.

6.3.6 Interaction of TV variables

Due to the fact that we do not have digital media due to the high p value, it was decided to conduct an analysis of interactions between TV media.

```
                            OLS Regression Results
==============================================================================
Dep. Variable:            Brand Value   R-squared:                       0.880
Model:                            OLS   Adj. R-squared:                  0.862
Method:                 Least Squares   F-statistic:                     47.32
Date:                Sat, 14 Sep 2024   Prob (F-statistic):           2.69e-35
Time:                        20:54:32   Log-Likelihood:                -763.32
No. Observations:                 105   AIC:                             1557.
Df Residuals:                      90   BIC:                             1596.
Df Model:                          14
Covariance Type:            nonrobust
=======================================================================================================
                                        coef    std err          t      P>|t|      [0.025      0.975]
-------------------------------------------------------------------------------------------------------
const                              1.674e+04     36.629    457.131      0.000    1.67e+04    1.68e+04
DT/NT * 10"                        -294.5216     69.158     -4.259      0.000    -431.915    -157.128
DT/NT * 30"                        -280.3381     76.585     -3.660      0.000    -432.487    -128.189
10" * 30"                           288.9619     80.975      3.569      0.001     128.091     449.832
TV Budget Planned                   313.1185     65.182      4.804      0.000     183.623     442.614
BRAND Weighted Distribution         944.8083     72.681     12.999      0.000     800.414    1089.202
BRAND VOLUME IN PROMO               183.2303     44.268      4.139      0.000      95.285     271.176
COMPETITOR Numeric Distribution     782.4926     75.255     10.398      0.000     632.985     932.000
COMPETITOR Weighted Distribution   -248.5774     54.811     -4.535      0.000    -357.469    -139.685
COMPETITOR VOLUME IN PROMO         -119.6633     41.092     -2.912      0.005    -201.300     -38.027
avg_price                           958.8255     79.855     12.007      0.000     800.180    1117.471
grp_erosion                        -129.0576     45.884     -2.813      0.006    -220.215     -37.900
Sales Value Growth (%)              277.9167     44.087      6.304      0.000     190.330     365.503
season_Spring                       177.2866     47.109      3.763      0.000      83.696     270.877
season_Winter                       235.0890     53.571      4.388      0.000     128.660     341.518
==============================================================================
Omnibus:                        5.282   Durbin-Watson:                   1.966
Prob(Omnibus):                  0.071   Jarque-Bera (JB):                7.310
Skew:                           0.126   Prob(JB):                       0.0259
Kurtosis:                       4.268   Cond. No.                         7.04
==============================================================================

Notes:
[1] Standard Errors assume that the covariance matrix of the errors is correctly specified.
R2 Score (Train): 0.8792245056732167
R2 Score (Test): 0.7341409554606677
Mean Squared Error: 237980.55377193427
```

Figure 56: Interaction of TV variables

- **R-squared** for the model is 0.880, indicating that approximately 88.0% of the variability in *Brand Value* is explained by the model. The adjusted R-squared value of 0.862, which accounts for the number of predictors in the model, supports the model's explanatory power.
- **F-statistic** is 47.32 with a p-value less than 0.0001, indicating that the model is statistically significant and that the included variables collectively have a significant impact on *Brand Value.*
- **Key Variables:**
  - *DT/NT * 10"* has a negative coefficient of -294.52, with a p-value of 0.000, suggesting that the interaction between DT/NT and the 10" channel reduces *Brand Value.* The effect is statistically significant.
  - *DT/NT * 30"* has a coefficient of -280.34 with a p-value of 0.000, also indicating a reduction in *Brand Value* due to the interaction of DT/NT with the 30" channel. This effect is statistically significant.
  - *10" * 30"* shows a positive coefficient of 288.96 with a p-value of 0.001, suggesting a positive impact on *Brand Value* from the interaction between the 10" and 30" channels.
  - *TV Budget Planned* has a coefficient of 313.12, with a p-value less than 0.0001, indicating a significant positive relationship with *Brand Value.*
  - *BRAND Weighted Distribution* and *BRAND VOLUME IN PROMO* show high positive coefficients of 944.81 and 183.23, respectively, reflecting their strong positive influence on *Brand Value.*
  - *COMPETITOR Numeric Distribution* also shows a significant positive coefficient of 782.49, while *COMPETITOR Weighted Distribution* has a negative coefficient of -248.58, suggesting varied impacts of competitor-related factors.
  - *avg_price* has a positive effect with a coefficient of 958.83, indicating that higher average prices are associated with higher.
  - *grp_erosion* shows a negative effect with a coefficient of -129.06, suggesting that higher erosion in the group decreases Brand Value
  - *Sales Value Growth (%)* has a positive coefficient of 277.92, indicating that growth in sales value is positively related to Brand Value
  - Seasonal variables *Spring* and *Winter* both show positive coefficients, with values of 177.29 and 235.09, respectively, suggesting higher Brand Value during these seasons.
- **Residual Analysis:**

- The *Omnibus* test statistic is 5.282 with a p-value of 0.071, indicating that residuals are approximately normally distributed.
- *Durbin-Watson* statistic is 1.966, which is close to 2, suggesting no significant autocorrelation in the residuals.
- *Jarque-Bera* test statistic is 7.310 with a p-value of 0.0259, indicating some deviations from normality in the residuals.

- **R-squared Scores:** The R-squared score for the training set is 0.879, while for the test set it is 0.734, showing a reasonable model performance on unseen data.
- **Mean Squared Error (MSE)** is 237,980.55, which measures the average squared difference between predicted and actual values.

6.4. Dataset G 'Sales'

6.4.1 Model Results

```
                            OLS Regression Results
==============================================================================
Dep. Variable:      Value Sales Brand   R-squared:                       0.845
Model:                            OLS   Adj. R-squared:                  0.834
Method:                 Least Squares   F-statistic:                     76.40
Date:                Sat, 24 Aug 2024   Prob (F-statistic):           1.39e-56
Time:                        21:36:51   Log-Likelihood:                -1535.8
No. Observations:                 166   AIC:                             3096.
Df Residuals:                     154   BIC:                             3133.
Df Model:                          11
Covariance Type:            nonrobust
================================================================================================================
                                              coef    std err          t      P>|t|      [0.025      0.975]
----------------------------------------------------------------------------------------------------------------
const                                    1.405e+04    203.210     69.155      0.000    1.37e+04    1.45e+04
AVERAGE PROMO PRICE                     -1601.9083    241.533     -6.632      0.000   -2079.053   -1124.763
ONLINE REVIEWS                           3022.4936    291.384     10.373      0.000    2446.869    3598.118
ONLINE RATINGS                            838.4803    249.900      3.355      0.001     344.806    1332.155
Negative Opinions                        -493.7610    225.961     -2.185      0.030    -940.144     -47.378
digital_investments                      2288.9008    256.057      8.939      0.000    1783.064    2794.738
Sales Value Growth (%)                    602.8152    216.774      2.781      0.006     174.581    1031.049
season_Summer                            1083.3431    258.165      4.196      0.000     573.341    1593.345
season_Winter                             790.6786    243.844      3.243      0.001     308.967    1272.390
NUMBER OF INFLUENCERS INVOLVED_log_lag_halo  507.2144    226.081      2.244      0.026      60.595     953.834
INVESTMENTS TV_log_lag_halo               605.3539    257.059      2.355      0.020      97.537    1113.171
promo_log_lag_halo                        756.1467    245.662      3.078      0.002     270.845    1241.448
==============================================================================
Omnibus:                        2.865   Durbin-Watson:                   2.035
Prob(Omnibus):                  0.239   Jarque-Bera (JB):                2.594
Skew:                           0.168   Prob(JB):                        0.273
Kurtosis:                       3.512   Cond. No.                         2.80
==============================================================================

Notes:
[1] Standard Errors assume that the covariance matrix of the errors is correctly specified.
R2 Score (Train): 0.8449155604849214
R2 Score (Test): 0.7125861142345677
Mean Squared Error: 11187064.057091724
```

Figure 57: Dataset G Sales Model

```
coefficients = {
    "Online Reviews": 3022.4936, #Online Reviews
    "Online Ratings": 838.4803, #Online Ratings
    "Online Media": 2288.9008, #Digital Investments
    "Season Summer": 1083.3431, #Season Summer
    "Season Winter": 790.6786, #Season Winter
    "Number of Inluencers": 507.2144, #Number of Influencers (Log lagged adstock)
    "TV Investments": 605.3539, #TV Investments (Log lagged adstock)
    "Promo": 756.2384, #Promo (Log lagged adstock)
    "Negative Opinions": -493.7610,
    "Average Promo Price": -1601.9083
}
```

Figure 58: Contribution Coefficients

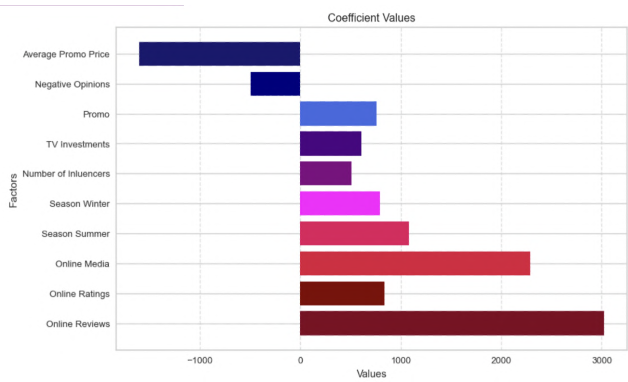

Figure 59: Variables Contribution

The variable promo_log_lag_halo positively impacts digital investments, increasing its value by up to 2288.9008. This also enhances the relevance of online reviews and ratings, showing a significant impact during the summer and winter seasons, increasing the value of TV INVESTMENTS by up to 605.3539. Additionally, the log_lag_adstock variable makes the number of influencers involved significant.

- **R-squared:** The model has an R-squared value of 0.845, indicating that approximately 84.5% of the variation in sales value is explained by the independent variables.
- **Adjusted R-squared:** The adjusted R-squared is slightly lower at 0.834, which accounts for the number of predictors in the model, indicating a strong explanatory power.
- **F-statistic:** The F-statistic is 76.40, with a p-value of 1.39e-56, demonstrating that the model is statistically significant and the predictors collectively influence sales.
- **Log-Likelihood:** The log-likelihood value is -1535.8, which provides a measure of the model's fit.

Coefficients and Significance

- **Intercept (const):** The constant term is 14,050, indicating that when all independent variables are held at zero, the predicted sales value is 14,050.
- **Average Promo Price** has a negative coefficient (-1601.91, $p < 0.001$), meaning that as the average promotional price increases, sales decline by approximately 1601.91 units. This relationship is statistically significant.
- **Online Reviews** has a positive coefficient (3022.49, $p < 0.001$), suggesting that each additional online review leads to an increase in sales by 3022.49 units.
- **Online Ratings** also positively affect sales (838.48, $p = 0.001$), with each one-unit increase in ratings leading to an increase of 838.48 units in sales.
- **Negative Opinions** negatively affect sales (-493.76, $p = 0.03$), where each additional negative opinion reduces sales by 493.76 units.
- **Digital Investments** have a strong positive impact on sales (2288.90, $p < 0.001$), indicating that increased investments in digital channels lead to significant sales growth.
- **Sales Value Growth (%)** contributes positively to future sales (602.82, $p = 0.006$), indicating that past success influences ongoing performance.
- **Season_Summer** (1083.34, $p < 0.001$) shows that the summer season is associated with a significant sales increase of 1083.34
- **Season_Winter** (790.68, $p = 0.001$) also leads to an increase in sales.
- **Number of Influencers Involved** has a positive effect (507.21, $p = 0.026$), indicating that previous campaigns involving more influencers lead to increased sales.
- **Investments in TV** also positively influence sales (605.35, $p = 0.02$), suggesting that TV investments continue to contribute positively, despite the growing emphasis on digital marketing.
- **Promotional Investments** (756.15, $p = 0.002$) also show a significant positive effect, reinforcing the importance of promotions in driving sales.

Diagnostics

- The **Durbin-Watson statistic** is 2.035, suggesting no significant autocorrelation in the residuals.
- The **Omnibus** and **Jarque-Bera** tests show no strong evidence of non-normality in the residuals.

Model Performance The model performs well both on training and test datasets:

- **R2 Score (Train):** 0.8449
- **R2 Score (Test):** 0.7126
- **Mean Squared Error:** 11,187,064

6.4.2 Contribution of Digital media

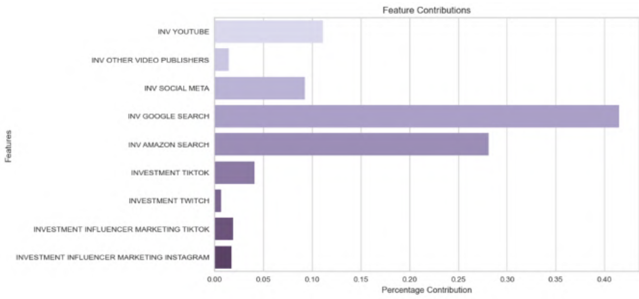

Figure 60: Digital Media Contributions

- The percentage contribution to the sales of INV YOUTUBE is: 0.11146680456568848
- The percentage contribution to the sales of INV OTHER VIDEO PUBLISHERS is: 0.014852300835375495
- The percentage contribution to the sales of INV SOCIAL META is: 0.09304122789708061
- The percentage contribution to the sales of INV GOOGLE SEARCH is: 0.4150873045682218
- The percentage contribution to the sales of INV AMAZON SEARCH is: 0.2810634115930157
- The percentage contribution to the sales of INVESTMENT TIKTOK is: 0.04112042595773864
- The percentage contribution to the sales of INVESTMENT TWITCH is: 0.006967169008549939
- The percentage contribution to the sales of INVESTMENT INFLUENCER MARKETING TIKTOK is: 0.019226180847571824
- The percentage contribution to the sales of INVESTMENT INFLUENCER MARKETING INSTAGRAM is: 0.017175174726757548

6.4.3 Media Effectiveness and ROI

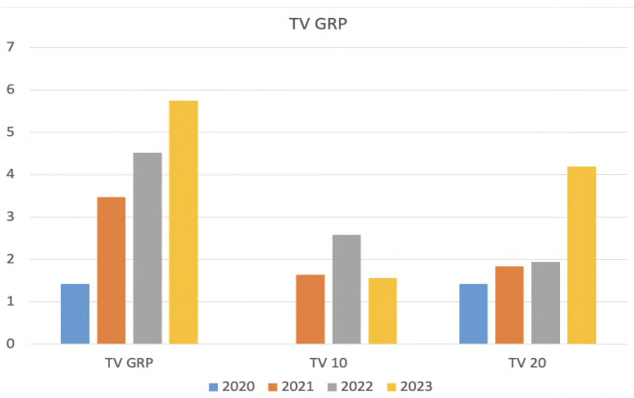

Figure 61: TV Effectiveness

Year	TV GRP	TV 10	TV 20
2020	1,418701	0	1,418701
2021	3,474644	1,636965	1,837679
2022	4,516824	2,579278	1,937546
2023	5,74797	1,56102	4,18695

Figure 62: TV Effectiveness

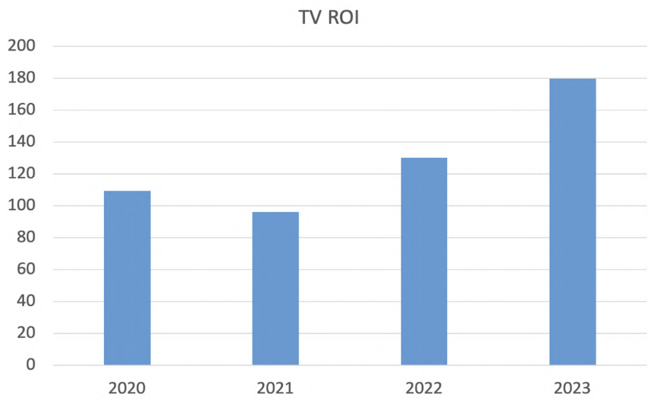

Figure 63: TV ROI

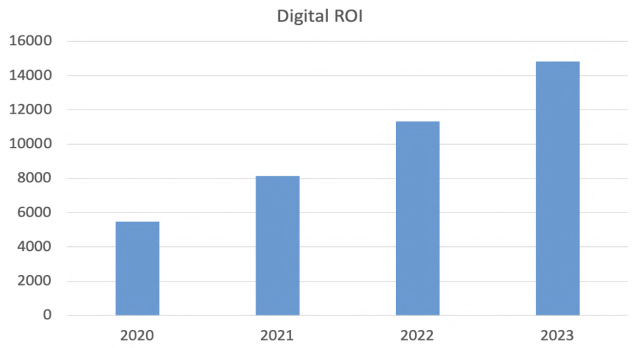

Figure 64: Digital ROI

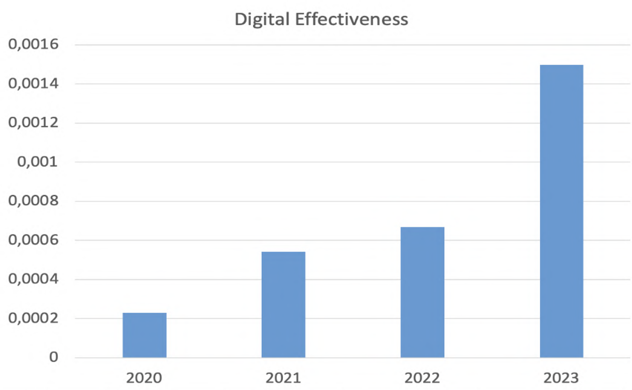

Figure 65: Digital Effectiveness

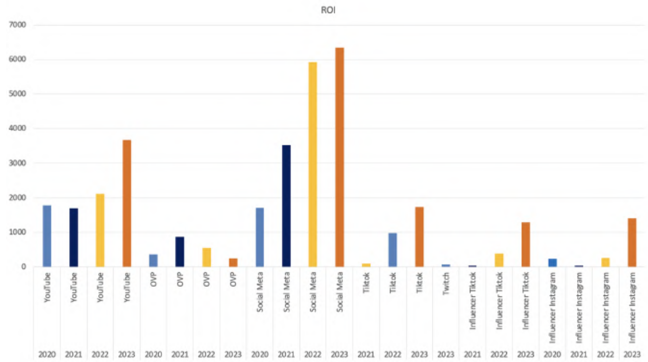

Figure 66: Digital ROI by tactics

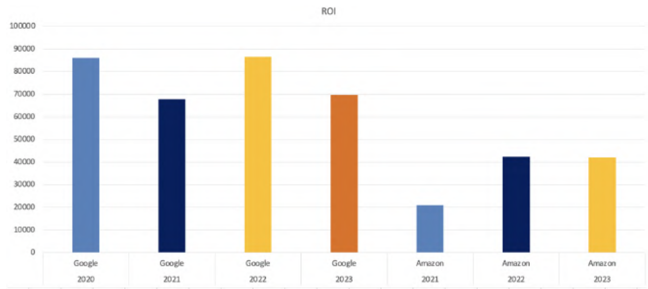

Figure 67: ROI Google and Amazon

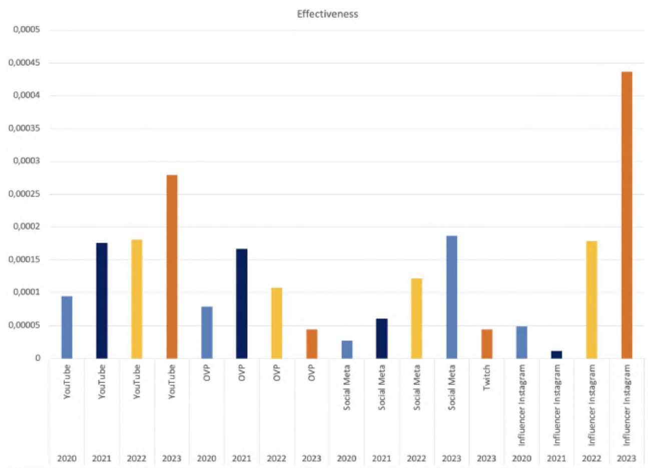

Figure 68: Digital Effectiveness by tactics

Year	Platform	ROI	Effectiveness
2020	YouTube	1786,267258	0,000095
2021	YouTube	1701,052101	0,000176
2022	YouTube	2118,689539	0,000181
2023	YouTube	3673,012033	0,00028
2020	OVP	363,22001	0,000079
2021	OVP	871,778444	0,000167
2022	OVP	549,096109	0,000108
2023	OVP	247,111521	0,000044
2020	Social Meta	1709,72821	0,000027
2021	Social Meta	3522,44512	0,000061
2022	Social Meta	5926,493668	0,000122
2023	Social Meta	6349,112869	0,000187
2020	Google	85999,12108	
2021	Google	67719,73818	
2022	Google	86421,19187	
2023	Google	69682,36995	
2021	Amazon	20966,43036	
2022	Amazon	42406,44449	
2023	Amazon	42022,11671	
2021	Tiktok	99,946991	
2022	Tiktok	979,951698	
2023	Tiktok	1739,287192	
2023	Twitch	72,64678	0,000044
2021	Influencer Tiktok	44,363565	
2022	Influencer Tiktok	386,714984	
2023	Influencer Tiktok	1294,400406	
2020	Influencer Instagram	237,601696	0,000049
2021	Influencer Instagram	33,865543	0,000012
2022	Influencer Instagram	258,971626	0,000179
2023	Influencer Instagram	1406,576689	0,000437

Figure 69: Table of Digital Effectiveness and ROI

YouTube:
- ROI: More than doubled from 2020 to 2023.
- Effectiveness: Increased steadily each year.

OVP (Other Video Publishers):
- ROI: Peaked in 2021 but declined afterward.
- Effectiveness: Also peaked in 2021 and declined by 2023.

Social Meta:
- ROI: Showed an upward trend, peaking in 2023.
- Effectiveness: Increased yearly, aligning with ROI growth.

Google:
- ROI: Highest among all platforms, peaking in 2022 and dropping slightly in 2023, but remained high.

Amazon:
- ROI: Increased substantially in 2022 compared to 2021, then showed a slight decline in 2023.

TikTok:
- ROI: Grew dramatically from 2021 to 2023.
- Effectiveness: Gradual improvement since 2021.

Twitch:
- ROI and Effectiveness: Data available only for 2023.

Influencer Marketing (TikTok and Instagram):

- ROI: Both platforms showed growth over time, with Instagram spiking in 2023.
- Effectiveness: Steady improvements, with a significant rise for Instagram in 2023.

6.4.4 Saturation Curve

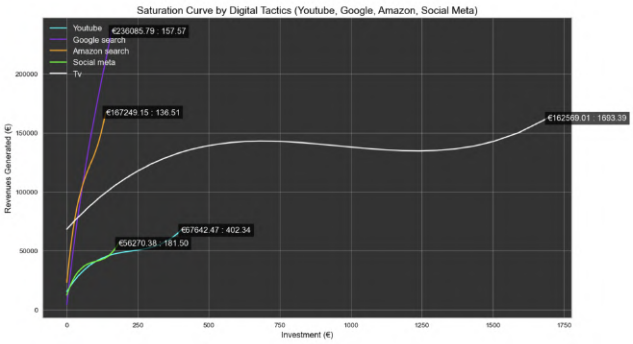

Figure 70: Saturation Curve TV, Google, Amazon, Social Meta

Google: Google shows the highest revenue potential among the platforms analyzed. It demonstrates steep initial growth, indicating strong returns on early investments. However, the revenue curve flattens as it approaches saturation.

Amazon: Amazon's revenue generation is substantial, though lower than Google's. The curve suggests that Amazon requires more investment to approach saturation, offering a consistent increase in revenue with a diminishing rate of return as spending rises.

Social Meta: Social Meta platforms (e.g., Facebook, Instagram) have the highest revenue per euro invested among the digital platforms, but the overall revenue ceiling is lower, indicating limited scalability compared to search platforms.

YouTube: YouTube shows the most significant revenue per euro invested, indicating highly efficient use of funds. However, it saturates quickly, meaning further investment after a certain point generates diminishing returns.

TV: TV exhibits a different trend, with a higher revenue per euro invested and a smoother, prolonged curve. It absorbs significant investment before hitting diminishing returns, making it suitable for large-scale campaigns.

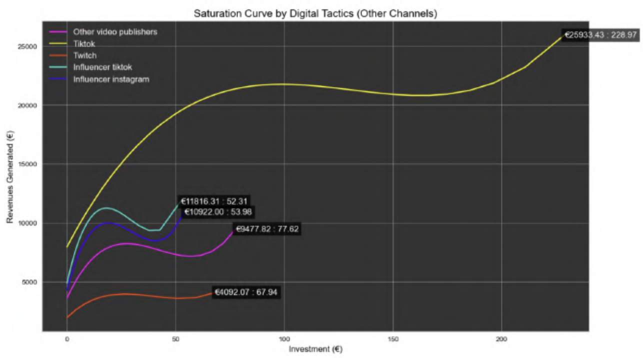

Figure 71: Saturation Curve OVP, Tiktok, Twitch, Influencers Instagram and Tiktok

TikTok: TikTok shows a steep initial return on investment (ROI), indicating strong effectiveness in early investment phases. It continues to scale well, maintaining high revenue potential before eventually reaching saturation.

TikTok Influencers: Instagram influencers generate good returns but approach saturation more quickly compared to direct TikTok marketing.

Instagram Tiktok: Tiktok influencers show a faster approach to saturation, but the ROI is slightly higher than that of Instagram influencers, allowing for slightly better scaling before saturation.

Twitch: Twitch displays the lowest revenue potential of the platforms analyzed. However, the ROI per euro spent is fairly consistent, indicating that it can be effective on a smaller scale before saturation.

Twitch: Twitch displays the lowest revenue potential of the platforms analyzed. However, the ROI per euro spent is fairly consistent, indicating that it can be effective on a smaller scale before saturation.

Other Video Publishers: This category scales reasonably well, with a higher ROI than Twitch, indicating diverse potential in less mainstream video channels.

6.4.5 Budget Optimization

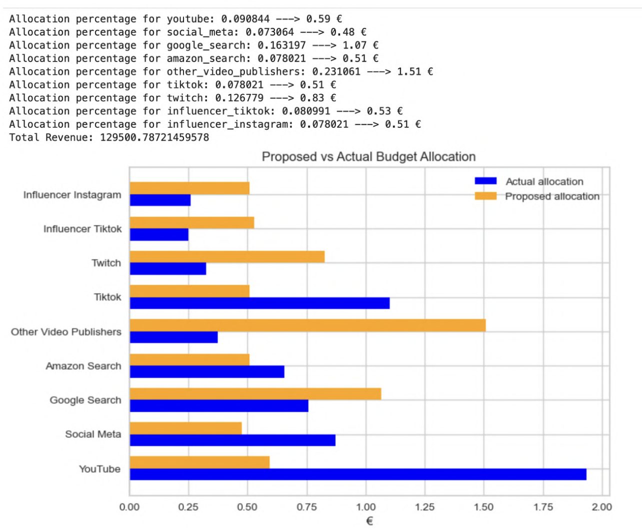

Figure 72: Optimal Budget Allocation Digital

The optimization based on the Gradient Boosting Machine (GBM) model generated the most profit among the four models. Total Revenue Generated €129,500.79, indicating effective ROI across digital marketing platforms. Google Search (€1.07) and Other Video Publishers (€1.51) received the highest budget allocations, suggesting high expected returns. Platforms like YouTube (€0.59) and Social Meta (€0.48) received lower allocations, indicating possible saturation or lower ROI. Twitch (€0.83) and influencer channels on TikTok (€0.53) and Instagram (€0.51) received moderate investments, targeting specific audience demographics.

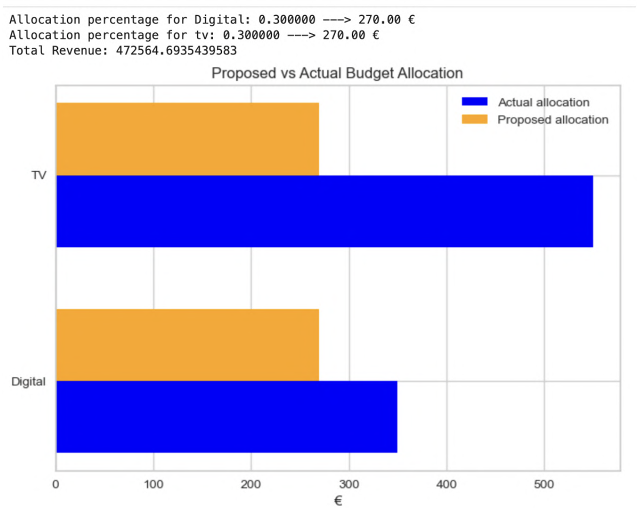

Figure 73: Optimal Budget Allocation TV and Digital

Both TV and Digital advertising have been allocated the same percentage of the budget, which is 30%, translating to €270 each. The actual allocation is significantly higher for TV compared to the proposed allocation, indicating a heavier reliance on TV advertising. For Digital, the actual allocation is lower than the proposed allocation, suggesting underutilization of this channel. The total revenue generated is €472,564.69.

6.4.6 Feature Interaction

6.4.7 Digital Investments

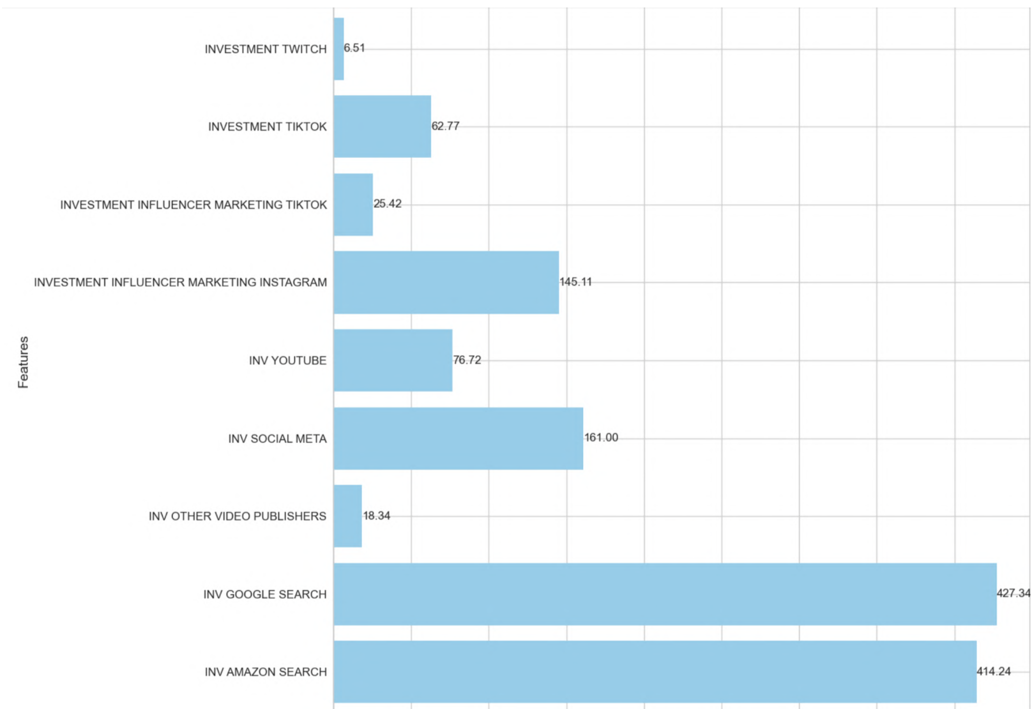

Figure 74: Digital Feature Importance (Investments)

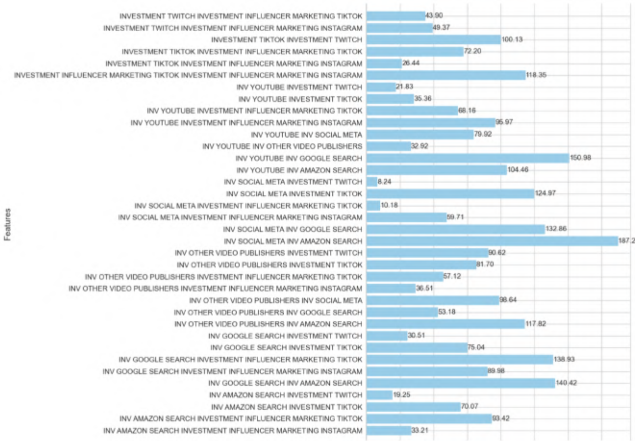

Figure 75: Digital Features Importance Interaction(Investments)

Analyzing interactions between various digital marketing channels allows marketers to identify synergies and conflicts, leading to more informed decisions regarding resource allocation. This analysis evaluates feature importance scores for different platform combinations, offering insight into the most effective marketing strategies on social media and related channels.

Interactions with Twitch

- Twitch + Influencer Marketing (TikTok): 43.90
- Twitch + Influencer Marketing (Instagram): 49.37
- Twitch + TikTok: 100.13
- Twitch + YouTube: 21.83
- Twitch + Social Meta: 8.24
- Twitch + Other Video Publishers: 90.62
- Twitch + Google Search: 30.51
- Twitch + Amazon Search: 19.25

Interactions with TikTok

- TikTok + Influencer Marketing (TikTok): 72.20
- TikTok + Influencer Marketing (Instagram): 26.44
- TikTok + YouTube: 35.36
- TikTok + Social Meta: 124.97
- TikTok + Other Video Publishers: 81.70
- TikTok + Google Search: 75.04
- TikTok + Amazon Search: 70.07

Interactions with YouTube

- YouTube + Influencer Marketing (TikTok): 68.16
- YouTube + Influencer Marketing (Instagram): 95.97
- YouTube + Social Meta: 79.92
- YouTube + Other Video Publishers: 32.92
- YouTube + Google Search: 150.98
- YouTube + Amazon Search: 104.46

Interactions with Social Meta

- Social Meta + Influencer Marketing (TikTok): 10.18
- Social Meta + Influencer Marketing (Instagram): 59.71
- Social Meta + Google Search: 132.86
- Social Meta + Amazon Search: 187.24

Interactions with Google Search

- Google Search + Influencer Marketing (TikTok): 138.93
- Google Search + Influencer Marketing (Instagram): 89.98
- Google Search + Amazon Search: 140.42

Interactions with Amazon Search

- Amazon Search + Influencer Marketing (TikTok): 93.42
- Amazon Search + Influencer Marketing (Instagram): 33.21

Model Results with Interactions

```
                            OLS Regression Results
==============================================================================
Dep. Variable:      Value Sales Brand   R-squared:                       0.879
Model:                            OLS   Adj. R-squared:                  0.867
Method:                 Least Squares   F-statistic:                     78.07
Date:                Thu, 05 Sep 2024   Prob (F-statistic):           1.12e-61
Time:                        15:04:04   Log-Likelihood:                -1515.6
No. Observations:                 166   AIC:                             3061.
Df Residuals:                     151   BIC:                             3108.
Df Model:                          14
Covariance Type:            nonrobust
===============================================================================================
                                 coef    std err          t      P>|t|      [0.025      0.975]
-----------------------------------------------------------------------------------------------
const                       1.405e+04    181.688     77.347      0.000    1.37e+04    1.44e+04
AVERAGE PROMO PRICE        -1053.6523    235.922     -4.466      0.000   -1519.787    -587.517
ONLINE REVIEWS              2805.5809    290.935      9.643      0.000    2230.753    3380.409
ONLINE RATINGS               806.6167    218.496      3.692      0.000     374.913    1238.321
digital_investments         1013.0132    491.263      2.062      0.041      42.376    1983.651
Sales Value Growth (%)       434.7095    197.865      2.197      0.030      43.767     825.652
season_Summer               1023.6637    232.132      4.410      0.000     565.018    1482.309
season_Winter                772.6601    231.960      3.331      0.001     314.354    1230.966
INVESTMENTS TV_log_lag_halo  693.4339    239.573      2.894      0.004     220.086    1166.782
promo_log_lag_halo           513.9373    249.603      2.059      0.041      20.772    1007.103
Meta Amazon                -1294.2241    388.212     -3.334      0.001   -2061.252    -527.196
YouTube Google               880.8147    483.845      1.820      0.071     -75.165    1836.795
Amazon Google                861.0370    368.715      2.335      0.021     132.531    1589.544
Google Influencers Tiktok    787.1904    224.534      3.506      0.001     343.556    1230.825
Meta Google                 1425.9731    386.399      3.690      0.000     662.526    2189.420
==============================================================================
Omnibus:                       22.394   Durbin-Watson:                   2.210
Prob(Omnibus):                  0.000   Jarque-Bera (JB):               37.295
Skew:                           0.692   Prob(JB):                     7.97e-09
Kurtosis:                       4.864   Cond. No.                         8.65
==============================================================================

Notes:
[1] Standard Errors assume that the covariance matrix of the errors is correctly specified.
R2 Score (Train): 0.8780917042985712
R2 Score (Test): 0.7484638290427417
Mean Squared Error: 9790589.10003736
```

Figure 76: The best interactions

Meta * Amazon (-1294.22, p = 0.001): The coefficient is negative (-1294.22), indicating that the interaction between Meta platforms (e.g., Facebook, Instagram) and Amazon has a negative effect on sales.

YouTube * Google (880.81, p = 0.071): The coefficient is positive (880.81), suggesting that combining YouTube and Google impressions has a positive impact on sales, although the p-value (0.071) is just outside conventional significance levels.

Amazon * Google (861.04, p = 0.021): A positive coefficient (861.04) indicates that the interaction between Amazon and Google results in an increase in sales.

Google * Influencers TikTok (787.19, p = 0.001): The positive coefficient (787.19) suggests that combining Google with TikTok influencers positively impacts sales.

Meta * Google (1425.97, p = 0.000): A strong positive coefficient (1425.97) shows that the interaction between Meta platforms and Google significantly increases sales.

R-squared (Train): The R-squared for the training dataset is 0.878, meaning the model explains approximately 87.8

R-squared (Test): The R-squared for the test dataset is 0.748, indicating the model explains 74.8

Adjusted R-squared: The adjusted R-squared is 0.867, which accounts for the number of predictors and provides a more refined measure of the model's explanatory power.

Mean Squared Error (MSE): The mean squared error (MSE) is 9,790,589.10, representing the average squared difference between predicted and actual values in the test dataset, serving as a measure of prediction accuracy.

```
                            OLS Regression Results
==============================================================================
Dep. Variable:      Value Sales Brand   R-squared:                       0.855
Model:                            OLS   Adj. R-squared:                  0.839
Method:                 Least Squares   F-statistic:                     54.72
Date:                Wed, 04 Sep 2024   Prob (F-statistic):           5.13e-54
Time:                        21:07:45   Log-Likelihood:                -1530.6
No. Observations:                 166   AIC:                             3095.
Df Residuals:                     149   BIC:                             3148.
Df Model:                          16
Covariance Type:            nonrobust
===============================================================================================================
                                              coef    std err          t      P>|t|      [0.025      0.975]
---------------------------------------------------------------------------------------------------------------
const                                    1.405e+04    200.205     70.193      0.000    1.37e+04    1.44e+04
AVERAGE PROMO PRICE                     -1398.9877    247.975     -5.642      0.000   -1888.990    -908.986
ONLINE REVIEWS                           2658.9866    339.149      7.840      0.000    1988.825    3329.149
ONLINE RATINGS                            783.5826    256.462      3.055      0.003     276.810    1290.355
Negative Opinions                        -445.8566    228.864     -1.948      0.053    -898.094       6.381
digital_investments                      4483.6394   1396.070      3.212      0.002    1724.987    7242.292
Sales Value Growth (%)                    586.1603    214.644      2.731      0.007     162.022    1010.299
season_Summer                            1096.1627    254.748      4.303      0.000     592.777    1599.549
season_Winter                             931.1657    249.779      3.728      0.000     437.599    1424.733
NUMBER OF INFLUENCERS INVOLVED_log_lag_halo  500.1998    241.676      2.070      0.040      22.645     977.754
INVESTMENTS TV_log_lag_halo               488.6703    261.194      1.871      0.063     -27.453    1004.793
promo_log_lag_halo                        968.5112    272.902      3.549      0.001     429.254    1507.768
Tiktok Influencers Instagram            -1218.7135    556.863     -2.189      0.030   -2319.082    -118.345
YouTube Twitch                          -1500.7314    872.727     -1.720      0.088   -3225.252     223.789
Amazon Twitch                             601.9779    513.760      1.172      0.243    -413.219    1617.175
Meta Influencers Tiktok                  -312.5995    439.879     -0.711      0.478   -1181.806     556.607
Meta Twitch                               -28.7992    636.516     -0.045      0.964   -1286.562    1228.964
==============================================================================
Omnibus:                        3.115   Durbin-Watson:                   1.998
Prob(Omnibus):                  0.211   Jarque-Bera (JB):                2.784
Skew:                           0.204   Prob(JB):                        0.249
Kurtosis:                       3.486   Cond. No.                         18.7
==============================================================================

Notes:
[1] Standard Errors assume that the covariance matrix of the errors is correctly specified.
R2 Score (Train): 0.8506581715635613
R2 Score (Test): 0.7642725585625507
Mean Squared Error: 9175262.984779045
```

Figure 77: The worst Interactions

Tiktok * Influencers Instagram (-1218.71, p = 0.030): A negative coefficient (-1218.71) suggests that combining TikTok impressions with Instagram influencers has a negative effect on sales.

YouTube * Twitch (-1500.73, p = 0.088): A negative coefficient (-1500.73) indicates that combining YouTube and Twitch impressions may reduce sales, although the p-value (0.088) is above conventional significance levels.

Amazon * Twitch (601.98, p = 0.243): A positive coefficient (601.98) suggests a positive effect on sales when Amazon and Twitch impressions are combined, but the p-value (0.243) indicates that this result is not statistically significant.

Meta * Influencers Tiktok (-312.60, p = 0.478): A negative coefficient (-312.60) shows a slight negative effect of combining Meta impressions with TikTok influencers, though the result is not statistically significant (p = 0.478).

Meta * Twitch (-28.80, p = 0.964): The coefficient (-28.80) indicates no meaningful effect on sales when combining Meta impressions with Twitch, as the p-value (0.964) suggests the result is not significant.

R-squared (Train): The R-squared for the training dataset is 0.851

R-squared (Test): The R-squared for the test dataset is 0.764

Adjusted R-squared: The adjusted R-squared is 0.839

Mean Squared Error (MSE): The mean squared error (MSE) is 9,175,262.98

OLS Regression Results

Dep. Variable:	Value Sales Brand	R-squared:	0.917
Model:	OLS	Adj. R-squared:	0.905
Method:	Least Squares	F-statistic:	75.79
Date:	Thu, 05 Sep 2024	Prob (F-statistic):	4.85e-67
Time:	14:57:41	Log-Likelihood:	-1484.0
No. Observations:	166	AIC:	3012.
Df Residuals:	144	BIC:	3080.
Df Model:	21		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
const	1.405e+04	153.811	91.365	0.000	1.37e+04	1.44e+04
AVERAGE PROMO PRICE	-691.1132	208.539	-3.314	0.001	-1103.306	-278.920
ONLINE REVIEWS	1829.2877	310.604	5.889	0.000	1215.356	2443.219
ONLINE RATINGS	839.3885	202.836	4.138	0.000	438.467	1240.310
digital_investments	2812.4099	630.528	4.460	0.000	1566.124	4058.696
Sales Value Growth (%)	525.4609	171.126	3.071	0.003	187.218	863.704
season_Summer	1073.3855	203.905	5.264	0.000	670.352	1476.419
season_Winter	696.9188	208.198	3.347	0.001	285.399	1108.439
INVESTMENTS TV_log_lag_halo	567.0606	220.085	2.577	0.011	132.046	1002.075
promo_log_lag_halo	1063.7239	263.369	4.039	0.000	543.156	1584.292
Google Influencers Tiktok	1193.3876	342.762	3.482	0.001	515.894	1870.882
Meta Google	2560.3694	348.343	7.350	0.000	1871.844	3248.895
Meta Tiktok	-2378.3568	487.964	-4.874	0.000	-3342.854	-1413.859
Influencers Tiktok Instagram	-1511.7196	372.395	-4.059	0.000	-2247.786	-775.653
YouTube Amazon	-2002.0651	389.947	-5.134	0.000	-2772.824	-1231.306
OVP Meta	-1504.3769	266.286	-5.649	0.000	-2030.711	-978.043
OVP Twitch	4610.4976	644.295	7.156	0.000	3336.999	5883.996
Google Influencers Instagram	1772.1452	414.380	4.277	0.000	953.092	2591.198
Meta Influencers Instagram	-1582.8163	348.477	-4.542	0.000	-2271.607	-894.025
YouTube Tiktok	1714.3937	438.626	3.909	0.000	847.417	2581.370
Google Twitch	-6688.6900	1100.120	-6.080	0.000	-8863.160	-4514.220
Amazon Twitch	3922.8754	798.790	4.911	0.000	2344.007	5501.744

Omnibus:	1.025	Durbin-Watson:	2.263
Prob(Omnibus):	0.599	Jarque-Bera (JB):	0.659
Skew:	0.096	Prob(JB):	0.719
Kurtosis:	3.241	Cond. No.	23.4

Notes:
[1] Standard Errors assume that the covariance matrix of the errors is correctly specified.
R2 Score (Train): 0.9120466879800224
R2 Score (Test): 0.6942086933442404
Mean Squared Error: 11902371.823648335

Figure 78: All Digital Interactions

Google * Influencers Tiktok (1193.39, p = 0.001): The coefficient is positive (1193.39), indicating that the interaction between Google and TikTok influencers leads to an increase in sales.

Meta * Google (2560.37, p = 0.000): A strong positive coefficient (2560.37) suggests that the interaction between Meta platforms (e.g., Facebook, Instagram) and Google positively impacts sales.

Meta * Tiktok (-2378.36, p = 0.000): The coefficient is negative (-2378.36), showing that the combination of Meta platforms and TikTok has a detrimental effect on sales.

Influencers Tiktok * Instagram (-1511.72, p = 0.000): A negative coefficient (-1511.72) suggests that the interaction between TikTok and Instagram influencers leads to reduced sales.

YouTube * Amazon (-2002.07, p = 0.000): The negative coefficient (-2002.07) indicates that combining YouTube and Amazon impressions results in a decrease in sales.

Amazon * Twitch (3922.88, p = 0.000): A positive coefficient (3922.88) indicates that Amazon and Twitch together boost sales.

OVP (Other Video Publishers) * Meta (-1504.38, p = 0.000): The coefficient is negative (-1504.38), meaning that combining other video publishers with Meta platforms reduces sales.

OVP * Twitch (4610.50, p = 0.000): A strong positive coefficient (4610.50) suggests that the interaction between other video publishers and Twitch significantly increases sales.

Google * Influencers Instagram (1772.15, p = 0.000): A positive coefficient (1772.15) indicates that Google and Instagram influencers together positively impact sales.

Meta * Influencers Instagram (-1582.82, p = 0.000): The coefficient is negative (-1582.82), showing that combining Meta platforms with Instagram influencers decreases sales.

YouTube * Tiktok (1714.39, p = 0.000): A positive coefficient (1714.39) indicates that combining YouTube and TikTok impressions boosts sales.

Google * Twitch (-6688.69, p = 0.000): A strongly negative coefficient (-6688.69) shows that combining Google with Twitch has a negative impact on sales.

R-squared (Train): The R-squared for the training dataset is 0.912

R-squared (Test): The R-squared for the test dataset is 0.694

Adjusted R-squared: The adjusted R-squared is 0.905, accounting for the number of predictors in the model and providing a more accurate measure of model performance.

Mean Squared Error (MSE): The mean squared error (MSE) of the model is 11,902,371.82.

Interaction terms were sorted in descending order, and compared by the performance and statistical significance of extended model with top 5 interaction terms and lowest top 5 terms. Despite the fact, that some model metrics showed comparatively similar results, it is not possible to ignore the fact of statistical significance of added terms. The top 5 terms improved model both in terms of residual analysis, as well as showing statistical significance, while lowest 5 terms made model worse as well as not showing statistical significance. The model with interaction of features explains 3.4% more of the variance of the dependent variable then original model without interactions, indicating higher accuracy and completeness of explanation of changes in the dependent variable. The adjusted R-squared is also higher for the interaction model, suggesting that when accounting for the number of parameters, the interaction model makes better use of information from the variables. The lower MSE value for the feature interaction model indicates a lower prediction error, which confirms the higher quality of this model when working with test data. The generalization ability of the model, as shown by R2 on the test sample, is also higher for the model with feature interactions. This suggests that it copes better with forecasting on new data. The model with all integrations includes more variables but shows a weaker statistical fit, suggesting that many of the additional interactions may not be contributing meaningfully to sales performance. On the other hand, the top 5 integrations model provides a clearer picture by focusing on the most impactful variables, particularly interactions involving Google, Meta, and influencers on platforms like Tiktok.

6.4.8 Digital and TV Investments interaction terms

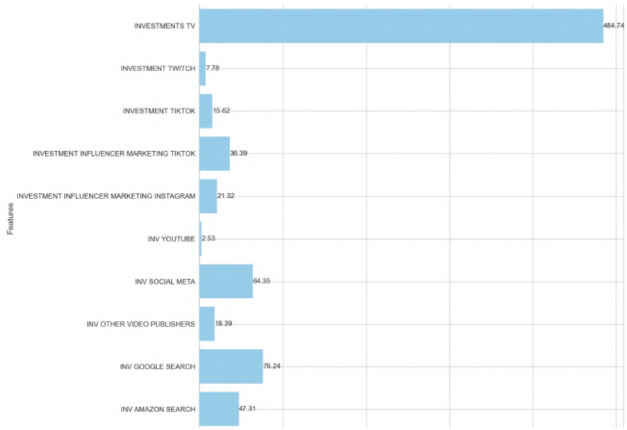

Figure 79: Feature Importance

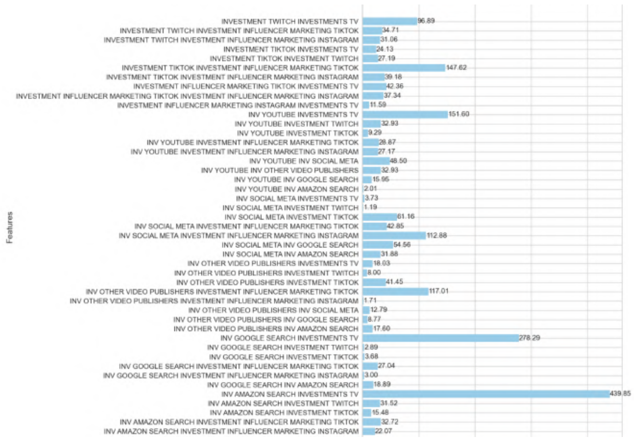

Figure 80: Feature Interaction Importance

Twitch

- Investments Twitch + Investments TV: 96.89
- Investments Twitch + Investments Influencer Marketing TikTok: 34.71
- Investments Twitch + Investments Influencer Marketing Instagram: 31.06
- Investments TikTok + Investments Twitch: 27.19
- Investments YouTube + Investments Twitch: 32.93
- Investments Social Meta + Investments Twitch: 1.19
- Investments Other Video Publishers + Investments Twitch: 8.00
- Investments Amazon Search + Investments Twitch: 31.52
-

TikTok

- Investments TikTok + Investments TV: 24.13
- Investments TikTok + Investments Twitch: 27.19
- Investments TikTok + Investments Influencer Marketing TikTok: 147.62
- Investments TikTok + Investments Influencer Marketing Instagram: 39.18
- Investments YouTube + Investments TikTok: 9.29
- Investments Social Meta + Investments TikTok: 3.73
- Investments Other Video Publishers + Investments TikTok: 41.45
- Investments Google Search + Investments TikTok: 3.68
- Investments Amazon Search + Investments TikTok: 15.48

Influencer Marketing (TikTok)

- Investments Twitch + Investments Influencer Marketing TikTok: 34.71
- Investments TikTok + Investments Influencer Marketing TikTok: 147.62
- Investments Influencer Marketing TikTok + Investments TV: 42.36
- Investments Influencer Marketing TikTok + Investments Instagram: 37.34
- Investments YouTube + Investments Influencer Marketing TikTok: 28.87
- Investments Social Meta + Investments Influencer Marketing TikTok: 42.85
- Investments Other Video Publishers + Investments Influencer Marketing TikTok: 117.01
- Investments Google Search + Investments Influencer Marketing TikTok: 27.04
- Investments Amazon Search + Investments Influencer Marketing TikTok: 32.72

Influencer Marketing (Instagram)

- Investments Twitch + Investments Influencer Marketing Instagram: 31.06
- Investments TikTok + Investments Influencer Marketing Instagram: 39.18
- Investments Influencer Marketing TikTok + Investments Instagram: 37.34
- Investments Influencer Marketing Instagram + Investments Instagram: 11.59
- Investments YouTube + Investments Influencer Marketing Instagram: 27.17
- Investments Social Meta + Investments Influencer Marketing Instagram: 112.88
- Investments Other Video Publishers + Investments Influencer Marketing Instagram: 1.71
- Investments Google Search + Investments Influencer Marketing Instagram: 3.00
- Investments Amazon Search + Investments Influencer Marketing Instagram: 22.07

TV

- Investments Twitch + Investments TV: 96.89
- Investments TikTok + Investments TV: 24.13
- Investments Influencer Marketing TikTok + Investments TV: 42.36
- Investments YouTube + Investments TV: 151.60
- Investments Social Meta + Investments TV: 61.16
- Investments Other Video Publishers + Investments TV: 18.03
- Investments Google Search + Investments TV: 278.29
- Investments Amazon Search + Investments TV: 439.85

YouTube

- Investments YouTube + Investments TV: 151.60
- Investments YouTube + Investments Twitch: 32.93
- Investments YouTube + Investments TikTok: 9.29
- Investments YouTube + Investments Influencer Marketing TikTok: 28.87
- Investments YouTube + Investments Influencer Marketing Instagram: 27.17
- Investments YouTube + Investments Social Meta: 48.50
- Investments YouTube + Investments Other Video Publishers: 32.93
- Investments YouTube + Investments Google Search: 15.95
- Investments YouTube + Investments Amazon Search: 2.01

Social Meta

- Investments YouTube + Investments Social Meta: 48.50
- Investments Social Meta + Investments Twitch: 1.19
- Investments Social Meta + Investments TV: 61.16
- Investments Social Meta + Investments Influencer Marketing TikTok: 42.85
- Investments Social Meta + Investments Influencer Marketing Instagram: 112.88
- Investments Social Meta + Investments Google Search: 54.56
- Investments Social Meta + Investments Amazon Search: 31.88
- Investments Other Video Publishers + Investments Social Meta: 12.79

Other Video Publishers

- Investments YouTube + Investments Other Video Publishers: 32.93
- Investments Other Video Publishers + Investments TV: 18.03
- Investments Other Video Publishers + Investments Twitch: 8.00
- Investments Other Video Publishers + Investments TikTok: 41.45
- Investments Other Video Publishers + Investments Influencer Marketing TikTok: 117.01
- Investments Other Video Publishers + Investments Influencer Marketing Instagram: 1.71
- Investments Other Video Publishers + Investments Social Meta: 12.79
- Investments Other Video Publishers + Investments Google Search: 8.77
- Investments Other Video Publishers + Investments TV: 17.60

Google Search

- Investments YouTube + Investments Google Search: 15.95
- Investments Social Meta + Investments Google Search: 54.56
- Investments Other Video Publishers + Investments Google Search: 8.77
- Investments Google Search + Investments TV: 278.29
- Investments Google Search + Investments TikTok: 3.68
- Investments Google Search + Investments Influencer Marketing TikTok: 27.04
- Investments Google Search + Investments Influencer Marketing Instagram: 3.00
- Investments Google Search + Investments Amazon Search: 18.89

Amazon Search

- Investments YouTube + Investments Amazon Search: 2.01
- Investments Social Meta + Investments Amazon Search: 31.88
- Investments Other Video Publishers + Investments Amazon Search: 8.77
- Investments Google Search + Investments Amazon Search: 18.89
- Investments Amazon Search + Investments TV: 439.85
- Investments Amazon Search + Investments Twitch: 31.52
- Investments Amazon Search + Investments TikTok: 15.48
- Investments Amazon Search + Investments Influencer Marketing TikTok: 32.72
- Investments Amazon Search + Investments Influencer Marketing Instagram: 22.07

Model Results with TV and Digital Interaction

```
                            OLS Regression Results
==============================================================================
Dep. Variable:      Value Sales Brand   R-squared:                       0.850
Model:                            OLS   Adj. R-squared:                  0.838
Method:                 Least Squares   F-statistic:                     72.24
Date:                Fri, 06 Sep 2024   Prob (F-statistic):           1.12e-56
Time:                        11:08:16   Log-Likelihood:                -1533.1
No. Observations:                 166   AIC:                             3092.
Df Residuals:                     153   BIC:                             3133.
Df Model:                          12
Covariance Type:            nonrobust
==============================================================================================================
                                               coef    std err          t      P>|t|      [0.025      0.975]
--------------------------------------------------------------------------------------------------------------
const                                     1.405e+04    200.658     70.034      0.000    1.37e+04    1.44e+04
AVERAGE PROMO PRICE                      -1672.9917    240.632     -6.952      0.000   -2148.381   -1197.602
ONLINE REVIEWS                            3114.5259    290.685     10.714      0.000    2540.251    3688.801
ONLINE RATINGS                             868.2533    247.124      3.513      0.001     380.038    1356.468
Negative Opinions                         -483.5237    223.170     -2.167      0.032    -924.415     -42.632
digital_investments                       2309.8966    253.016      9.129      0.000    1810.040    2809.753
Sales Value Growth (%)                     590.5450    214.122      2.758      0.007     167.528    1013.562
season_Summer                             1073.3605    254.961      4.210      0.000     569.661    1577.060
season_Winter                              687.0261    245.253      2.801      0.006     202.507    1171.545
NUMBER OF INFLUENCERS INVOLVED_log_lag_halo 651.8192  232.522      2.803      0.006     192.452    1111.187
INVESTMENTS TV_log_lag_halo               1038.3198    319.922      3.246      0.001     406.285    1670.355
promo_log_lag_halo                         741.0774    242.670      3.054      0.003     261.661    1220.494
Amazon TV                                 -666.1076    299.591     -2.223      0.028   -1257.977     -74.238
==============================================================================
Omnibus:                        3.569   Durbin-Watson:                   2.024
Prob(Omnibus):                  0.168   Jarque-Bera (JB):                3.847
Skew:                           0.126   Prob(JB):                        0.146
Kurtosis:                       3.702   Cond. No.                         3.48
==============================================================================

Notes:
[1] Standard Errors assume that the covariance matrix of the errors is correctly specified.
R2 Score (Train): 0.8496819648885032
R2 Score (Test): 0.7043932504079637
Mean Squared Error: 11505956.417477017
```

Figure 81: TV and Digital Interaction terms

Amazon TV Coefficient (-666.1076): While each of these features has a positive effect on its own, their interaction is negative. This suggests that in its current state, investing in both TV and the Amazon platform simultaneously (perhaps in the same time window or campaigns) is not yielding the expected synergies(The First Model includes only TV + digital media interactions).

```
                            OLS Regression Results
==============================================================================
Dep. Variable:      Value Sales Brand   R-squared:                       0.875
Model:                            OLS   Adj. R-squared:                  0.863
Method:                 Least Squares   F-statistic:                     70.12
Date:                Fri, 06 Sep 2024   Prob (F-statistic):           8.00e-60
Time:                        11:22:17   Log-Likelihood:                -1517.9
No. Observations:                 166   AIC:                             3068.
Df Residuals:                     150   BIC:                             3118.
Df Model:                          15
Covariance Type:            nonrobust
===================================================================================================
                                   coef    std err          t      P>|t|      [0.025      0.975]
---------------------------------------------------------------------------------------------------
const                         1.405e+04    184.850     76.023      0.000    1.37e+04    1.44e+04
AVERAGE PROMO PRICE          -1111.8648    238.110     -4.670      0.000   -1582.347    -641.382
ONLINE REVIEWS                3162.9417    276.401     11.443      0.000    2616.799    3709.085
ONLINE RATINGS                1017.3318    220.984      4.604      0.000     580.689    1453.974
digital_investments           1145.4601    491.545      2.330      0.021     174.215    2116.706
Sales Value Growth (%)         422.9750    202.439      2.089      0.038      22.975     822.975
season_Summer                 1082.7267    235.989      4.588      0.000     616.434    1549.019
season_Winter                  688.6457    238.239      2.891      0.004     217.908    1159.383
INVESTMENTS TV_log_lag_halo    824.3403    251.211      3.281      0.001     327.972    1320.709
YouTube TikTok                1069.7022    472.467      2.264      0.025     136.152    2003.253
Meta TikTok                  -1596.5604    612.626     -2.606      0.010   -2807.050    -386.070
Meta Google                   2727.1238    432.241      6.309      0.000    1873.057    3581.191
Meta Amazon                  -2123.4256    424.728     -4.999      0.000   -2962.648   -1284.203
Google Amazon                  942.7952    358.963      2.626      0.010     233.518    1652.073
Amazon TikTok                 1005.6425    394.819      2.547      0.012     225.518    1785.767
Amazon Influencer TikTok       574.4157    225.138      2.551      0.012     129.564    1019.267
==============================================================================
Omnibus:                       19.295   Durbin-Watson:                   2.191
Prob(Omnibus):                  0.000   Jarque-Bera (JB):               28.886
Skew:                           0.646   Prob(JB):                     5.34e-07
Kurtosis:                       4.584   Cond. No.                         10.2
==============================================================================

Notes:
[1] Standard Errors assume that the covariance matrix of the errors is correctly specified.
R2 Score (Train): 0.8738610491454242
R2 Score (Test): 0.6946315454114281
Mean Squared Error: 11885913.074100778
```

Figure 82: TV + All Digital Media

While the first model only included interaction between TV and Digital Media. The Second model also includes interactions between each digital media with each other.

TV Interaction Effect: While TV-related variables have been dropped due to high p-values, their inclusion initially still affected the overall model structure. The coefficients for digital interactions are still prominent, but the TV interaction variables might have introduced some noise or instability to the model, reflected in the lower adjusted R-squared compared to the purely digital interaction model.

```
                            OLS Regression Results
==============================================================================
Dep. Variable:      Value Sales Brand   R-squared:                       0.857
Model:                            OLS   Adj. R-squared:                  0.845
Method:                 Least Squares   F-statistic:                     70.14
Date:                Mon, 09 Sep 2024   Prob (F-statistic):           2.51e-57
Time:                        21:15:04   Log-Likelihood:                -1529.1
No. Observations:                 166   AIC:                             3086.
Df Residuals:                     152   BIC:                             3130.
Df Model:                          13
Covariance Type:            nonrobust
==========================================================================================================
                                                coef    std err          t      P>|t|      [0.025      0.975]
----------------------------------------------------------------------------------------------------------
const                                      1.405e+04    196.466     71.529      0.000    1.37e+04    1.44e+04
AVERAGE PROMO PRICE                       -1634.2846    236.023     -6.924      0.000   -2100.594   -1167.975
ONLINE REVIEWS                             3268.7620    290.061     11.269      0.000    2695.691    3841.833
ONLINE RATINGS                              970.3032    244.777      3.964      0.000     486.699    1453.908
Negative Opinions                          -443.1713    218.997     -2.024      0.045    -875.843     -10.500
digital_investments                        2247.3202    248.769      9.034      0.000    1755.830    2738.811
Sales Value Growth (%)                      554.4168    210.058      2.639      0.009     139.406     969.427
season_Summer                              1048.5817    249.797      4.198      0.000     555.059    1542.104
season_Winter                               654.2805    240.423      2.721      0.007     179.279    1129.283
NUMBER OF INFLUENCERS INVOLVED_log_lag_halo  472.0640    236.819      1.993      0.048       4.181     939.947
INVESTMENTS TV_log_lag_halo                1086.1015    313.718      3.462      0.001     466.290    1705.913
promo_log_lag_halo                          692.3751    238.257      2.906      0.004     221.653    1163.098
Amazon TV                                  -920.3094    307.487     -2.993      0.003   -1527.810    -312.809
Influencer TikTok TV                        658.0373    238.720      2.757      0.007     186.399    1129.675
==============================================================================
Omnibus:                        2.344   Durbin-Watson:                   2.004
Prob(Omnibus):                  0.310   Jarque-Bera (JB):                2.240
Skew:                           0.023   Prob(JB):                        0.326
Kurtosis:                       3.567   Cond. No.                         3.76
==============================================================================

Notes:
[1] Standard Errors assume that the covariance matrix of the errors is correctly specified.
R2 Score (Train): 0.8567694787087067
R2 Score (Test): 0.7073277268088829
Mean Squared Error: 11391737.2475031
```

Figure 83: Digital * TV

Amazon TV (-920.31, p = 0.003): The coefficient for the interaction between Amazon TV and sales is -920.31 with a statistically significant p-value of 0.003.

Influencer TikTok TV (658.04, p = 0.007): The coefficient for the interaction between TikTok influencers and TV is 658.04 with a statistically significant p-value of 0.007.

These interaction terms indicate significant relationships between TV investments and digital media channels, impacting brand sales.
While the first model includes features based on theirs feature importance, the third model includes interaction terms: TV * each digital channels. Amazon TV remain negatively stable, the interaction between TikTok influencers and TV (Influencer TikTok TV) showing a positive coefficient suggests a synergistic effect when both channels are used together.

6.4.9 Digital Impressions

Results of Feature Importance for Impressions:

- R2 Score (Train): 0.5430149214084714
- R2 Score (Test): 0.22465998271350052

Due to the fact that the interaction-based model did not demonstrate high results on either the training (y_train) or test (y_test) samples, it was decided to include all possible interactions for a more detailed analysis of their impact on the overall performance to the original model.

Model Results

```
                            OLS Regression Results
==============================================================================
Dep. Variable:      Value Sales Brand   R-squared:                       0.902
Model:                            OLS   Adj. R-squared:                  0.889
Method:                 Least Squares   F-statistic:                     66.84
Date:                Fri, 06 Sep 2024   Prob (F-statistic):           7.26e-63
Time:                        12:11:50   Log-Likelihood:                -1497.7
No. Observations:                 166   AIC:                             3037.
Df Residuals:                     145   BIC:                             3103.
Df Model:                          20
Covariance Type:            nonrobust

                                                                            coef    std err          t      P>|t|      [0.025
0.975]
---------------------------------------------------------------------------------------------------------------------------
------
const                                                                  1.405e+04    166.464     84.420      0.000    1.37e+04      1.
44e+04
AVERAGE PROMO PRICE                                                   -1450.5945    203.798     -7.118      0.000   -1853.394     -10
47.795
ONLINE REVIEWS                                                         2400.3486    278.793      8.610      0.000    1849.325      29
51.372
ONLINE RATINGS                                                          865.1548    221.868      3.899      0.000     426.642      13
03.667
digital_investments                                                    3233.7287    510.028      6.340      0.000    2225.678      42
41.779
Sales Value Growth (%)                                                  529.2963    183.704      2.881      0.005     166.213       8
92.380
season_Summer                                                           766.9169    194.246      3.948      0.000     382.997      11
50.837
INVESTMENTS TV_log_lag_halo                                            1014.8740    208.414      4.870      0.000     602.952      14
26.796
promo_log_lag_halo                                                     1841.8636    260.065      7.082      0.000    1327.856      23
55.871
IMPRESSIONS YOUTUBE * IMPRESSIONS SPOTIFY                              7477.0671   2381.947      3.139      0.002    2769.246      1.
22e+04
IMPRESSIONS YOUTUBE * IMPRESSIONS SOCIAL META                          -674.9182    325.947     -2.071      0.040   -1319.139       -
30.697
IMPRESSIONS SPOTIFY * IMPR. TIKTOK                                    -1.221e+04   3602.214     -3.389      0.001   -1.93e+04     -50
87.292
IMPRESSIONS SOCIAL META * IMPRESSIONS SPOTIFY                          5440.4921   1740.031      3.127      0.002    2001.392      88
79.592
IMPRESSIONS SOCIAL META * IMPRESSIONS INFLUENCER MARKETING TIKTOK      3722.6056   1166.176      3.192      0.002    1417.706      60
27.505
IMPRESSIONS SOCIAL META * IMPRESSIONS INFLUENCER MARKETING INSTAGRAM  -1224.9275    411.789     -2.975      0.003   -2038.812      -4
11.043
IMPRESSIONS SOCIAL META * IMPR. TIKTOK                                 -764.7333    314.165     -2.434      0.016   -1385.668      -1
43.798
IMPRESSIONS OTHER VIDEO PUBLISHERS * IMPRESSIONS SOCIAL META           -826.9476    253.653     -3.260      0.001   -1328.282      -3
25.613
IMPRESSIONS OTHER VIDEO PUBLISHERS * IMPRESSIONS INFLUENCER MARKETING INSTAGRAM 1119.1159    431.978      2.591      0.011     265.330      19
72.902
IMPRESSIONS OTHER VIDEO PUBLISHERS * IMPR. TWITCH                      1637.9047    262.698      6.235      0.000    1118.692      21
57.117
IMPR. TWITCH * IMPRESSIONS INFLUENCER MARKETING INSTAGRAM             -2169.3122    352.240     -6.159      0.000   -2865.501     -14
73.123
IMPR. TIKTOK * IMPRESSIONS INFLUENCER MARKETING TIKTOK                -2613.1538   1060.435     -2.464      0.015   -4709.061      -5
17.246

==============================================================================
Omnibus:                       25.526   Durbin-Watson:                   1.914
Prob(Omnibus):                  0.000   Jarque-Bera (JB):               41.669
Skew:                           0.792   Prob(JB):                     8.95e-10
Kurtosis:                       4.875   Cond. No.                         55.6
==============================================================================

Notes:
[1] Standard Errors assume that the covariance matrix of the errors is correctly specified.
R2 Score (Train): 0.8941353822466253
R2 Score (Test): 0.6461510939984703
Mean Squared Error: 13772926.688732173
```

Figure 84: Impressions Interaction

Impressions YouTube * Impressions Spotify (7477.07, p = 0.002): A positive coefficient of 7477.07, indicating a strong combined effect of YouTube and Spotify impressions on sales.

Impressions YouTube * Impressions Social Meta (-674.92, p = 0.040): A negative coefficient of -674.92, suggesting that combining impressions from YouTube and Social Meta slightly reduces the effectiveness of these platforms.

Impressions Spotify * Impressions TikTok (-12,210.67, p = 0.001): A strongly negative coefficient of -12,210.67, showing that combining impressions from Spotify and TikTok significantly decreases sales.

Impressions Social Meta * Impressions Spotify (5440.49, p = 0.002): A positive coefficient of 5440.49, indicating that when Social Meta and Spotify impressions are combined, they boost sales.

Impressions Social Meta * Impressions Influencer Marketing TikTok (3722.61, p = 0.002): A positive coefficient of 3722.61, showing that combining Social Meta with TikTok influencer impressions has a positive impact on sales.

Impressions Social Meta * Impressions Influencer Marketing Instagram (-1224.93, p = 0.003): A negative coefficient of -1224.93, suggesting that combining Social Meta impressions with Instagram influencer marketing reduces sales.

Impressions Social Meta * Impressions TikTok (-764.73, p = 0.016): A negative coefficient of -764.73, indicating a small reduction in sales when combining Social Meta and TikTok impressions.

Impressions Other Video Publishers * Impressions Social Meta (-826.95, p = 0.001): A negative coefficient of -826.95, showing a decrease in sales when combining impressions from other video publishers and Social Meta.

Impressions Other Video Publishers * Impressions Influencer Marketing Instagram (1119.12, p = 0.011): A positive coefficient of 1119.12, indicating that combining other video publishers with Instagram influencers positively affects sales.

Impressions Other Video Publishers * Impressions Twitch (1637.90, p = 0.000): A positive coefficient of 1637.90, indicating a positive impact on sales when combining impressions from other video publishers and Twitch.

Impressions Twitch * Impressions Influencer Marketing Instagram (-2169.31, p = 0.000): A negative coefficient of -2169.31, suggesting that combining Twitch impressions with Instagram influencer marketing reduces sales.

Impressions TikTok * Impressions Influencer Marketing TikTok (-2613.15, p = 0.015): A negative coefficient of -2613.15, showing that combining TikTok impressions with TikTok influencer marketing reduces sales.

6.4.10 Digital Impressions and TV GRP

Due to the fact that the interaction-based model did not demonstrate high results on either the training (y_train) or test (y_test) samples, it was decided to include all possible interactions for a more detailed analysis of their impact on the overall performance to the original model.

```
                            OLS Regression Results
==============================================================================
Dep. Variable:      Value Sales Brand   R-squared:                       0.905
Model:                            OLS   Adj. R-squared:                  0.890
Method:                 Least Squares   F-statistic:                     58.96
Date:                Sun, 15 Sep 2024   Prob (F-statistic):           4.14e-61
Time:                        00:28:17   Log-Likelihood:                -1495.0
No. Observations:                 166   AIC:                             3038.
Df Residuals:                     142   BIC:                             3113.
Df Model:                          23
Covariance Type:            nonrobust
========================================================================================================================
======
                                                                           coef    std err          t      P>|t|      [0.025
0.975]
------------------------------------------------------------------------------------------------------------------------
------
const                                                                 1.405e+04    165.561     84.881      0.000    1.37e+04      1.
44e+04
AVERAGE PROMO PRICE                                                  -1543.6820    205.043     -7.529      0.000   -1949.012     -11
38.352
ONLINE REVIEWS                                                        2699.9604    280.614      9.622      0.000    2145.240      32
54.681
ONLINE RATINGS                                                         814.3563    224.991      3.619      0.000     369.591      12
59.122
digital_investments                                                   4068.4598    511.602      7.952      0.000    3057.120      50
79.800
Sales Value Growth (%)                                                 594.7554    183.279      3.245      0.001     232.448       9
57.063
season_Summer                                                         1248.9409    216.354      5.773      0.000     821.250      16
76.632
season_Winter                                                          657.6294    213.874      3.075      0.003     234.841      10
80.418
INVESTMENTS TV_log_lag_halo                                            915.7980    294.289      3.112      0.002     334.044      14
97.552
promo_log_lag_halo                                                    1169.4990    256.650      4.557      0.000     662.151      16
76.847
IMPRESSIONS YOUTUBE * IMPRESSIONS SPOTIFY                             6502.0702   2208.681      2.944      0.004    2135.925      1.
09e+04
IMPRESSIONS YOUTUBE * IMPRESSIONS SOCIAL META                         -990.7463    325.238     -3.046      0.003   -1633.680      -3
47.812
IMPRESSIONS SPOTIFY * IMPR. TIKTOK                                   -1.178e+04   3305.877     -3.564      0.000   -1.83e+04     -52
47.968
IMPRESSIONS SOCIAL META * IMPRESSIONS SPOTIFY                         6018.3336   1658.191      3.629      0.000    2740.402      92
96.265
IMPRESSIONS SOCIAL META * IMPRESSIONS INFLUENCER MARKETING TIKTOK     3092.5971    648.673      4.768      0.000    1810.294      43
74.901
IMPRESSIONS SOCIAL META * IMPRESSIONS INFLUENCER MARKETING INSTAGRAM -1175.8777    396.225     -2.968      0.004   -1959.139      -3
92.616
IMPRESSIONS SOCIAL META * IMPR. TIKTOK                               -1285.7638    322.703     -3.984      0.000   -1923.686      -6
47.842
IMPRESSIONS OTHER VIDEO PUBLISHERS * IMPRESSIONS SOCIAL META          -684.3052    253.123     -2.703      0.008   -1184.681      -1
83.929
IMPRESSIONS OTHER VIDEO PUBLISHERS * IMPRESSIONS INFLUENCER MARKETING INSTAGRAM   875.2667    422.372      2.072      0.040      40.318      17
10.216
IMPR. TWITCH * IMPRESSIONS INFLUENCER MARKETING TIKTOK               -3485.1022    619.385     -5.627      0.000   -4709.510     -22
60.694
IMPR. TIKTOK * IMPR. TWITCH                                            960.6148    224.095      4.287      0.000     517.621      14
03.609
GRP TV 20 * IMPRESSIONS YOUTUBE                                       -557.1553    285.617     -1.951      0.053   -1121.767
7.456
GRP TV 20 * IMPRESSIONS OTHER VIDEO PUBLISHERS                         598.1809    285.895      2.092      0.038      33.020      11
63.341
GRP TV 20 * IMPRESSIONS INFLUENCER MARKETING TIKTOK                   -947.4630    284.408     -3.331      0.001   -1509.683      -3
85.243
==============================================================================
Omnibus:                       11.444   Durbin-Watson:                   2.008
Prob(Omnibus):                  0.003   Jarque-Bera (JB):               12.945
Skew:                           0.513   Prob(JB):                      0.00155
Kurtosis:                       3.904   Cond. No.                         51.6
==============================================================================

Notes:
[1] Standard Errors assume that the covariance matrix of the errors is correctly specified.
R2 Score (Train): 0.8964179669449976
R2 Score (Test): 0.48326796678408357
Mean Squared Error: 20112856.90161655
```

Figure 85: Impressions and TV GRP interaction terms

Impressions YouTube * Impressions Spotify (6502.07, p = 0.004): A positive coefficient (6502.07) indicates a strong positive effect when YouTube and Spotify impressions are combined.

Impressions YouTube * Impressions Social Meta (-990.75, p = 0.003): A negative coefficient (-990.75) suggests that combining YouTube and Social Meta impressions decreases sales.

Impressions Spotify * Impressions TikTok (-11,780.67, p = 0.000): A strongly negative coefficient (-11,780.67) shows a significant decrease in sales when Spotify and TikTok impressions are combined.

Impressions Social Meta * Impressions Spotify (6018.33, p = 0.000): A positive coefficient (6018.33) indicates a strong positive effect when Social Meta and Spotify impressions are combined.

Impressions Social Meta * Impressions Influencer Marketing TikTok (3092.60, p = 0.000): A positive coefficient (3092.60) suggests that combining Social Meta impressions with TikTok influencer marketing increases sales.

Impressions Social Meta * Impressions Influencer Marketing Instagram (-1175.88, p = 0.004): A negative coefficient (-1175.88) indicates a decrease in sales when Social Meta impressions are combined with Instagram influencer marketing.

Impressions Social Meta * Impressions TikTok (-1285.76, p = 0.000): A negative coefficient (-1285.76) shows that combining Social Meta and TikTok impressions reduces sales.

Impressions Other Video Publishers * Impressions Social Meta (-684.31, p = 0.008): A negative coefficient (-684.31) suggests that combining other video publishers and Social Meta impressions decreases sales.

Impressions Other Video Publishers * Impressions Influencer Marketing Instagram (875.27, p = 0.040): A positive coefficient (875.27) indicates that combining impressions from other video publishers with Instagram influencers increases sales.

Impressions Twitch * Impressions Influencer Marketing TikTok (-3485.10, p = 0.000): A negative coefficient (-3485.10) shows that combining Twitch impressions with TikTok influencer marketing decreases sales.

Impressions TikTok * Impressions Twitch (960.61, p = 0.000): A positive coefficient (960.61) suggests that combining TikTok and Twitch impressions positively impacts sales.

GRP TV 20 * Impressions YouTube (-557.16, p = 0.053): A negative coefficient (-557.16) suggests a slight decrease in sales when TV GRP and YouTube impressions are combined, though this result is borderline in terms of significance (p = 0.053).

GRP TV 20 * Impressions Other Video Publishers (598.18, p = 0.038): A positive coefficient (598.18) indicates a positive impact on sales when TV GRP and impressions from other video publishers are combined.

GRP TV 20 * Impressions Influencer Marketing TikTok (-947.46, p = 0.001): A negative coefficient (-947.46) shows that combining TV GRP with TikTok influencer marketing decreases sales.

7. Discussion

This section synthesizes the results of the thesis research, examining the effectiveness of marketing channels through the lens of Marketing Mix Modeling (MMM). The analysis provides insights into the different impacts of digital and traditional marketing, the role of interaction effects, and the challenges posed by saturation. These findings are compared to existing literature and previous research to validate the results and formulate recommendations for marketing budget allocation.

7.1. Findings

7.1.1 Short-Term Effectiveness of Digital Channels

The thesis confirms that digital channels such as social media, display ads, and search engine marketing (SEM) provide immediate returns on investment (ROI). Platforms like TikTok, Social Meta and YouTube were particu- larly effective in driving short-term sales. The immediacy of the response was evident in metrics like impressions, clicks, and short-term sales uplift. This aligns with the findings of [80], who examined the impact of digital marketing on a luxury car brand and found that digital platforms provided high short-term engagement but required careful monitoring to avoid over-saturation. Digital channels demonstrated rapid diminishing returns, as captured by the S-shaped saturation curves. This behavior is consistent with the work of [81], who discussed the short-lived nature of sales generated by digital advertising, particularly in the context of price promotions. The saturation curves observed in this thesis suggest that while initial investments in platforms like Social Meta yielded significant sales, subsequent increases in spending led to diminishing returns. This is a clear manifestation of the law of diminishing marginal returns, a concept deeply rooted in microeconomic theory [82].
As digital platforms reach saturation points quickly, the data underscores the need for marketers to be agile, shifting budget allocations as returns begin to taper off. This result mirrors the findings of [84], who observed similar patterns of diminishing returns in large-scale advertising campaigns.

7.1.2 Long-Term Impact of Traditional Channels

Traditional media, including television and print, exhibited a different dynamic compared to digital channels. While digital media generated fast, short-term spikes in sales, traditional channels, particularly television, contributed to long-term brand building and sustained market presence. This is consistent with the work of [85], who highlighted the role of traditional media in maintaining long-term brand equity.
The thesis found that TV ads, despite their slower immediate impact, played a crucial role in reinforcing brand awareness over time. This is in line with findings from [86], which emphasized the brand-building power of TV due to its wide reach and ability to convey emotionally resonant messages. The enduring effect of TV also supports the theory of advertising wear-in and wear-out, where repeated exposure to ads initially strengthens brand recognition before the impact plateaus [88].
In terms of ROI, traditional media showed a steadier curve, absorbing higher investment levels before reaching saturation compared to digital channels. This observation confirms the findings of [84], who pointed out that while digital channels can produce faster returns, traditional media ensures the sustainability of brand presence over the long term.

7.1.3 Saturation Curves and Diminishing Returns

The saturation curves analyzed in the thesis revealed key insights into how marketing investments behave over time. For digital channels, platforms like YouTube and Instagram demonstrated faster saturation, with diminishing returns occurring at lower investment levels compared to traditional channels. These findings resonate with the work of [86], who also noted that digital advertising quickly hits a saturation point, beyond which further spending fails to generate proportional returns.

In contrast, traditional media such as TV followed a more gradual saturation curve, allowing for higher levels of investment before experiencing diminishing returns. This reinforces the concept of media "wear-out" discussed by [84], where traditional media can maintain effectiveness over extended periods if managed properly. The sinusoidal pattern observed in the saturation curves suggests cyclical periods of diminishing returns and recovery, indicating that staggered spending on traditional media can yield better long-term results.

The concept of diminishing returns in marketing aligns with the economic theory of diminishing marginal productivity. This theory explains that as more of a resource (in this case, marketing spend) is applied to a given factor (a channel), the additional output (sales or engagement) decreases. This principle was observed in the empirical results of this thesis and is consistent with the marketing literature [88].

7.1.4 ROI and Budget Allocation

The ROI analysis conducted in this thesis provided critical insights into how marketers can optimize budget allocation across different channels. Digital platforms, particularly social media and paid search, generated the highest short-term ROI but also showed the fastest decline in returns as investment increased. This suggests that while these platforms are effective for quick wins and time-sensitive campaigns, marketers must remain vigilant about saturation to avoid wasting resources.

Conversely, traditional media such as TV proved more resilient to higher levels of investment, delivering steady long-term returns. This suggests a dual-strategy approach, where digital channels are leveraged for short-term engagement and traditional media is used to build sustained brand equity over time. The findings are consistent with the optimization models used by [90], which recommend reallocating budget from saturated channels to those with higher potential returns to maximize overall ROI.

7.1.5 Hypotheses Tested

Several hypotheses were tested throughout the research, with the following key results:

- **Hypothesis 1:** Digital channels provide immediate but short-lived effects compared to traditional channels, which offer long-term benefits.
  - **Result:** Supported by both ROI and effectiveness analysis. Digital channels such as TikTok and YouTube delivered rapid short-term returns, while traditional media like TV contributed to sustained brand recognition over the long term [81, 85].
- **Hypothesis 2:** Saturation curves for digital channels show faster diminishing returns compared to traditional channels.
  - **Result:** Strongly supported by the saturation curve analysis. Digital platforms demonstrated quick ROI growth but also hit saturation points early, while traditional channels showed more gradual diminishing returns [84, 86].
- **Hypothesis 3:** Interaction effects between digital and traditional media enhances campaign effectiveness.
  - **Result:** Partially supported. While some combinations like TV and Amazon Search resulted in negative synergy, others, such as Influencers TikTok and TV, showed positive interaction effects, highlighting the need for careful coordination between TV and Digital Investments[80, 89].

8. Implications

In this chapter, we discuss the theoretical contributions of this thesis to the existing literature and outline the main managerial implications derived from our findings.

8.1. Research Implications

This thesis aimed to explore and emphasize the significance of interaction effects within digital channels and between digital and traditional marketing channels. Our research addresses a notable gap in the existing literature, where the interplay between different marketing channels is often overlooked, especially regarding their combined impact on return on investment (ROI) and overall campaign effectiveness.

We confirmed that digital channels such as TikTok, Social Meta, and YouTube provide immediate but short-lived effects on sales. These platforms demonstrated rapid returns on investment but reached saturation points quickly, leading to diminishing returns at lower investment levels compared to traditional channels. This finding aligns with Cain's research on the integration of online and offline channels, which highlights that the alignment of digital and traditional media leads to more effective marketing campaigns and higher ROI [1].

By quantifying the saturation dynamics through S-shaped curves, we add depth to the understanding of the short-term effectiveness of digital marketing, similar to the detailed ROI analyses conducted in the luxury car industry [80].

Traditional media, particularly television, exhibited a different dynamic. While TV ads had a slower immediate impact, they played a crucial role in reinforcing brand awareness over time. Our analysis showed that television advertising contributes to sustained brand presence and maintains effectiveness over extended periods, absorbing higher investment levels before experiencing diminishing returns. This supports the notion that traditional media is essential for long-term brand building, consistent with the findings of Hanssens and Pauwels, who emphasized the long-term impact of marketing campaigns on brand loyalty and perception [90].

We investigated the interaction effects between digital and traditional media channels. The results revealed that certain combinations, such as the interaction between TikTok influencers and TV advertising, showed positive synergy, enhancing overall campaign effectiveness. Conversely, other combinations, like simultaneous investments in TV and Amazon Search, resulted in negative interaction effects, indicating a lack of expected synergies. These findings underscore the complexity of multi-channel marketing dynamics and highlight the importance of considering interaction effects in marketing mix modeling, echoing Cain's methodology of using dynamic time series models to analyze the impact of changes in one channel on the performance of others [1].

Lastly, our research provides empirical evidence on the importance of dynamic budget allocation strategies. By analyzing the ROI and saturation curves of various channels, we demonstrated that reallocating funds from saturated channels to those with higher potential returns can optimize overall campaign performance. This contributes to the literature by emphasizing the need for agile and data-driven decision-making in marketing budget allocation, as supported by the study on marketing response modeling in the luxury car industry, which highlights the value of conducting detailed ROI analyses for effective budget allocation [80]. Our approach is also in line with Cain's emphasis on dynamic marketing modeling to adapt strategies in a rapidly changing digital environment [1].

8.2. Managerial Implications

The findings of this thesis offer valuable insights for marketing practitioners aiming to optimize their strategies and maximize ROI.

8.2.1 Optimize Digital Channel Investments

Key Findings:

- **Immediate Impact with Rapid Saturation:** Digital platforms like TikTok, YouTube, and Social Meta deliver swift boosts in sales but reach saturation quickly, leading to diminishing returns at lower investment levels compared to traditional media.

Business Implications:

- **Front-Load Digital Spending:** Allocation of significant portion of the marketing budget to digital channels during the early stages of campaigns to capitalize on their immediate impact.
- **Real-Time Performance Monitoring:** Advanced regular analytics tracking key performance indicators (KPIs) and identifying saturation points promptly. This aligns with the emphasis on using consumer data and advanced analytics for strategy formulation, as demonstrated in the luxury car industry's marketing response modeling [80].
- **Agile Budget Reallocation:** Preparedness to swiftly reallocating funds from saturated digital channels to those with higher ROI potential to maintain overall campaign efficiency, reinforcing Cain's findings on the importance of dynamic marketing modeling for adapting strategies [1].

8.2.2 Leveraging Traditional Media for Sustained Brand Building

Key Findings:

- **Long-Term Brand Equity:** Traditional media, particularly television, contributes to sustained brand awareness and maintains effectiveness over extended periods.
- **Gradual Saturation Curve:** TV advertising allows for higher investment levels before experiencing diminishing returns, making it resilient to higher spend.

Business Implications:

- **Strategic Integration of TV Advertising :** Inclusion of television as a core component of the marketing mix to strengthen brand equity and ensure long-term market presence. This supports the balance between short-term incentives and long-term branding critical for maintaining healthy brand perception and customer loyalty, as noted by Hanssens and Pauwels [90].
- **Complement Digital Efforts:** Usage of traditional media to reinforce digital messaging, creating a cohesive and comprehensive marketing strategy.
- **Optimize Scheduling:** Planning TV campaigns to avoid audience fatigue, potentially by aligning ads with peak viewing times or significant events.

8.2.3 Enhancement of Campaign Effectiveness Through Synergistic Channel Interactions

Key Findings:

- **Positive and Negative Interactions:** Certain channel combinations, like influencers on TikTok paired with TV advertising, show positive synergy, while others, such as simultaneous investments in TV and Amazon Search, can negatively impact effectiveness.

Business Implications:

- **Coordinating Multi-Channel Strategies:** Ensuring marketing efforts across different channels are harmonized to maximize positive interactions and avoid negative overlaps. This approach is consistent with Cain's research on the integration of online and offline channels leading to improved consumer engagement and higher ROI [1].
- **Integrated Marketing Communications (IMC):** Adopting an IMC approach to deliver consistent and reinforcing messages across all platforms.
- **Avoiding Audience Cannibalization:** Diminishing returns due to ad fatigue or conflicting messages can be due to overlapping audience segments.

8.2.4 Dynamic and Data-Driven Budget Allocation

Key Findings:

- **Variable Saturation Points:** Saturation levels vary significantly across channels, necessitating flexible budget management.
- **Importance of Data Analysis:** Analysis of past campaign data is crucial for informed decision-making regarding budget distribution.

Business Implications:

- **Real-Time Analytics Investment:** Strengthen analytics capabilities to monitor campaign performance continuously and adjust budgets accordingly. This recommendation aligns with the use of mathematical modeling methods for data analysis in optimizing marketing strategies [80].
- **Flexible Planning Frameworks:** Moving away from rigid annual budgets toward dynamic allocation strategies that can respond to market changes and performance data, can be a key for a better performance, as emphasized by Cain in the context of dynamic marketing modeling [1].
- **ROI Optimization:** Regular reassessing of the ROI of each channel and reallocating funds to maximize overall campaign effectiveness.

8.2.5 Emphasis on Quality Content and Creative Excellence

Key Findings:

- **Content Quality Matters:** The effectiveness of marketing campaigns is significantly influenced by the quality and relevance of content.
- **Risk of Poor Content:** Low-quality or irrelevant advertising can alienate potential customers and negate investment benefits.

Business Implications:

- **Investing in Creative Development:** Allocating resources to produce high-quality, engaging, and emotionally resonant content that appeals to the target audience.
- **Brand Consistency:** Maintaining a cohesive brand image and messaging across all channels to strengthen recognition and trust.
- **Innovative Content Strategies:** Experimenting with new formats and creative ideas to capture and retain audience interest.

8.2.6 Foster Flexibility and Agility in Marketing Strategies

Key Findings:

- **Rapid Market Changes:** Consumer behavior and market trends evolve quickly, requiring marketers to adapt strategies promptly.
- **Need for Agility:** Flexibility in strategy and execution is crucial to capitalize on new opportunities and mitigate risks.

Business Implications:

- **Agile Marketing Practices:** Implementation of processes that allow for rapid response to market dynamics, including quick decision-making and execution. This is in line with the importance of adapting strategies using dynamic models, as highlighted by Cain [1].
- **Continuous Learning Culture:** Encouragement of culture of experimentation and learning within the marketing team to foster innovation.
- **Trend Monitoring:** Staying informed about industry developments, technological advancements, and shifts in consumer behavior to anticipate changes and adjust strategies proactively.

In conclusion, these insights underscore the necessity of a balanced, data-driven approach that effectively leverages both digital and traditional channels. By understanding the unique strengths and limitations of each medium - and the interplay between them - businesses can optimize marketing investments for immediate returns and long-term brand growth. Embracing flexibility, investing in high-quality content, and maintaining agility in budget allocation will position organizations to navigate the complexities of today's marketing landscape successfully and achieve sustainable competitive advantages. This holistic approach is supported by existing research that emphasizes the integration of marketing channels, the use of dynamic modeling for strategy adaptation, and the critical role of data analysis in optimizing marketing effectiveness [1, 80, 90].

9. Conclusion

In this thesis, we've explored how digital and traditional marketing channels work together and affect ROI and overall campaign success. By addressing a gap in current research, we've provided a detailed look at how using multiple marketing channels influences consumer behavior and business outcomes.

Our findings show that digital platforms like TikTok, Facebook, and YouTube can quickly boost sales, but their impact doesn't last long. They offer fast returns on investment but hit saturation quickly, so investing more in them doesn't always pay off. In contrast, traditional media like television helps keep the brand in people's minds over a longer period and can handle higher investment levels before we see diminishing returns.

We discovered that combining certain digital and traditional media can make a campaign more effective because they complement each other. However, mixing channels can sometimes backfire if they overlap too much or send conflicting messages to the audience. This highlights how complex multi-channel marketing can be and why careful planning is so important.

By analyzing ROI and saturation curves, we've shown real-world evidence of how crucial it is to be smart and flexible with budgeting. Moving funds from channels that have maxed out to those with more growth potential can improve overall campaign performance. This underscores the need for agility and continuous monitoring in marketing strategies to keep up with fast-changing market dynamics.

Allocating the budget wisely is also crucial. If a company spreads its marketing budget across different channels properly, it can maximize ROI. This means analyzing past campaign data and making budget decisions based on those insights. Plus, the quality of content and ads matters a lot. They should not only look good but also connect emotionally with the target audience. Poor or irrelevant ads won't just fail to deliver results; they might even push potential customers away.

Being flexible is key. Since the market is always changing, marketers need to be ready to adjust their strategies quickly to new trends and shifts in consumer behavior. By bringing all these elements together into one cohesive strategy and constantly monitoring customer feedback, companies can not only capture attention but also keep customers interested, which is essential for long-term success in a competitive environment.

Marketing is always evolving, and what worked a few years ago might not work today. At its core, marketing is about grabbing the customer's attention. It's a lot like psychology because it looks at how people behave when they're buying something and how to encourage them to take action. A big part of this is getting the marketing mix right—the perfect combination of product, price, place, and promotion. Companies need to offer the right products at the right price, at the right time, and through the channels that appeal most to customers. Balancing these elements can make a big impact on the market and ensure marketing investments are used effectively.

To sum it up, this thesis emphasizes the importance of a balanced, data-driven approach that effectively uses both digital and traditional channels. Understanding what each medium is good at — and how they interact — allows businesses to optimize their marketing spend for immediate gains and long-term brand growth. By staying flexible, investing in high-quality content, and being agile with budget allocation, organizations can successfully navigate today's complex marketing landscape and achieve lasting competitive advantages.

10. Bibliography

[1] Cain, B. *Dynamic Marketing Mix Modeling and Digital Attribution.*

[2] Hanssens, D. M., Parsons, L. J., & Schultz, R. L. (2001). *Market response models: Econometric and time series analysis.* Springer Science & Business Media.

[3] Lee, S. H., & Kim, D. (2019). Data integration techniques for heterogeneous datasets. *Data Science Journal*, 18(1), 1-12.

[4] Kuhn, M., & Johnson, K. (2013). *Applied predictive modeling.* Springer.

[5] Jolliffe, I. T., & Cadima, J. (2016). Principal component analysis: A review and recent developments. *Philosophical Transactions of the Royal Society A*, 374(2065), 20150202.

[6] Gelman, A., & Hill, J. (2006). *Data analysis using regression and multilevel/hierarchical models.* Cambridge University Press.

[7] Durbin, J., & Koopman, S. J. (2012). *Time series analysis by state space methods* (2nd ed.). Oxford University Press.

[8] Barnett, V., & Lewis, T. (1994). *Outliers in statistical data.* John Wiley & Sons.

[9] Little, R. J., & Rubin, D. B. (2019). *Statistical analysis with missing data* (3rd ed.). John Wiley & Sons.

[10] Enders, C. K. (2010). *Applied missing data analysis.* Guilford Press.

[11] Belsley, D. A., Kuh, E., & Welsch, R. E. (2005). *Regression diagnostics: Identifying influential data and sources of collinearity.* John Wiley & Sons.

[12] Rousseeuw, P. J., & Leroy, A. M. (2005). *Robust regression and outlier detection.* John Wiley & Sons.

[13] Osborne, J. W. (2002). Notes on the use of data transformations. *Practical Assessment, Research, and Evaluation*, 8(6), 1-8.

[14] Tukey, J. W. (1962). The future of data analysis. *The Annals of Mathematical Statistics*, 33(1), 1-67.

[15] Aggarwal, C. C. (2016). *Outlier analysis* (2nd ed.). Springer.

[16] Broadbent, S., & Fry, T. R. (1979). *One way TV advertisements work.* World Advertising Research Center.

[17] Gujarati, D. N., & Porter, D. C. (2009). *Basic econometrics* (5th ed.). McGraw-Hill.

[18] Clarke, D. G. (1976). Econometric measurement of the duration of advertising effect on sales. *Journal of Marketing Research*, 13(4), 345-357.

[19] Akaike, H. (1974). A new look at the statistical model identification. *IEEE Transactions on Automatic Control*, 19(6), 716-723.

[20] Smith, J. A., & Taylor, K. (2020). Integrating external data sources in marketing models. *Journal of Marketing Analytics*, 8(3), 150-162.

[21] Johnson, L. M. (2021). Challenges of incorporating external data into predictive models. *International Journal of Data Science*, 5(2), 85-99.

[22] Patel, R., & Mehta, S. (2021). Evaluating external data sources for predictive analytics. *Journal of Business Analytics*, 4(1), 22-35.

[23] Box, G. E., Jenkins, G. M., Reinsel, G. C., & Ljung, G. M. (2015). *Time series analysis: Forecasting and control* (5th ed.). John Wiley & Sons.

[24] Greene, W. H. (2003). *Econometric analysis* (5th ed.). Prentice Hall.

[25] Wooldridge, J. M. (2015). *Introductory econometrics: A modern approach* (6th ed.). Cengage Learning.

[26] White, H. (1980). A heteroskedasticity-consistent covariance matrix estimator and a direct test for heteroskedasticity. *Econometrica*, 48(4), 817-838.

[27] Johnson, R. A., & Wichern, D. W. (2007). *Applied multivariate statistical analysis* (6th ed.). Pearson.

[28] Tenenbaum, J. B., De Silva, V., & Langford, J. C. (2000). A global geometric framework for nonlinear dimensionality reduction. *Science*, 290(5500), 2319-2323.

[29] Garcia, S., Luengo, J., & Herrera, F. (2015). *Data preprocessing in data mining*. Springer.

[30] Little, R. J., & Rubin, D. B. (2002). *Statistical analysis with missing data* (2nd ed.). John Wiley & Sons.

[31] Wilcox, R. R. (2010). *Fundamentals of modern statistical methods: Substantially improving power and accuracy*. Springer.

[32] Fox, J. (1991). *Regression diagnostics*. Sage Publications.

[33] Durbin, J., & Watson, G. S. (1951). Testing for serial correlation in least squares regression. II. *Biometrika*, 38(1/2), 159-177.

[34] Tibshirani, R. (1996). Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society: Series B (Methodological)*, 58(1), 267-288.

[35] Kim, Y., & Park, H. (2020). Assessing data quality in big data applications. *Journal of Big Data*, 7(1), 1-18.

[36] Brown, T. J. (2021). Multicollinearity in regression models: Problems and solutions. *Statistical Methods in Economics*, 14(2), 45-60.

[37] Gelman, A. (2008). Scaling regression inputs by dividing by two standard deviations. *Statistics in Medicine*, 27(15), 2865-2873.

[38] Kutner, M. H., Nachtsheim, C. J., & Neter, J. (2005). *Applied linear regression models* (4th ed.). McGraw-Hill/Irwin.

[39] Hoerl, A. E., & Kennard, R. W. (1970). Ridge regression: Biased estimation for nonorthogonal problems. *Technometrics*, 12(1), 55-67.

[40] Hyndman, R. J., & Athanasopoulos, G. (2018). *Forecasting: Principles and practice* (2nd ed.). OTexts.

[41] Chatfield, C. (2004). *The analysis of time series: An introduction* (6th ed.). Chapman and Hall/CRC.

[42] Gardner, E. S. (1985). Exponential smoothing: The state of the art. *Journal of Forecasting*, 4(1), 1-28.

[43] Cleveland, R. B., Cleveland, W. S., McRae, J. E., & Terpenning, I. (1990). STL: A seasonal-trend decomposition procedure based on loess. *Journal of Official Statistics*, 6(1), 3-73.

[44] Pankratz, A. (1983). *Forecasting with univariate Box-Jenkins models: Concepts and cases*. John Wiley & Sons.

[45] Anderson, L. (2020). Disadvantages of complex models in marketing. *Journal of Marketing Complexity*, 12(1), 45-59.

[46] Bickel, P. J., & Doksum, K. A. (2015). *Mathematical statistics: Basic ideas and selected topics* (2nd ed.). Chapman and Hall/CRC.

[47] Bishop, C. M. (2006). *Pattern recognition and machine learning*. Springer.

[48] Breiman, L. (2001). Random forests. *Machine Learning*, 45(1), 5-32.

[49] Efron, B., Hastie, T., Johnstone, I., & Tibshirani, R. (2004). Least angle regression. *The Annals of Statistics*, 32(2), 407-499.

[50] Friedman, J., Hastie, T., & Tibshirani, R. (2010). Regularization paths for generalized linear models via coordinate descent. *Journal of Statistical Software*, 33(1), 1-22.

[51] Fox, J. (2015). *Applied regression analysis and generalized linear models* (3rd ed.). Sage Publications.

[52] Goodfellow, I., Bengio, Y., & Courville, A. (2016). *Deep learning*. MIT Press.

[53] Guilherme, S. (2023). Improving data analysis techniques. *Journal of Advanced Analytics*, 15(2), 123-145.

[54] Harrell Jr, F. E. (2015). *Regression modeling strategies* (2nd ed.). Springer.

[55] Hastie, T., Tibshirani, R., & Friedman, J. (2009). *The elements of statistical learning* (2nd ed.). Springer.

[56] James, G., Witten, D., Hastie, T., & Tibshirani, R. (2013). *An introduction to statistical learning*. Springer.

[57] Kennedy, P. (2008). *A guide to econometrics* (6th ed.). Blackwell Publishing.

[58] Kumar, V., & Petersen, J. A. (2012). *Statistical methods in customer relationship management*. John Wiley & Sons.

[59] Leek, J. T., & Peng, R. D. (2015). Statistics: P values are just the tip of the iceberg. *Nature*, 520(7549), 612.

[60] Miller, G. (2021). Non-linear regression in marketing mix modeling. *Marketing Science*, 40(5), 1001-1015.

[61] Pan, S. J., & Yang, Q. (2010). A survey on transfer learning. *IEEE Transactions on Knowledge and Data Engineering*, 22(10), 1345-1359.

[62] Provost, F., & Fawcett, T. (2013). *Data science for business*. O'Reilly Media.

[63] Raschka, S., & Mirjalili, V. (2019). *Python machine learning* (3rd ed.). Packt Publishing.

[64] Ruder, S. (2016). An overview of gradient descent optimization algorithms. *arXiv preprint arXiv:1609.04747*.

[65] Rossi, P. E., McCulloch, R. E., & Allenby, G. M. (2005). *Bayesian statistics and marketing*. John Wiley & Sons.

[66] Thompson, H. (2020). Impact of external data on marketing metrics. *Journal of Marketing Metrics*, 5(3), 89-102.

[67] West, K. D. (1996). Asymptotic inference about predictive ability. *Econometrica*, 64(5), 1067-1084.

[68] Wilson, J. (2021). Advantages of incorporating external data in MMM. *Marketing Research*, 33(4), 75-88.

[69] Witten, I. H., Frank, E., Hall, M. A., & Pal, C. J. (2016). *Data mining: Practical machine learning tools and techniques* (4th ed.). Morgan Kaufmann.

[70] Zhang, Y. (2020). Time-varying coefficient models in marketing. *Journal of Marketing Models*, 12(2), 45-60.

[71] Zou, H., & Hastie, T. (2005). Regularization and variable selection via the elastic net. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 67(2), 301-320.

[72] Mansurali, A., & Jeyanthi, P. M. (2023). *Marketing analytics: A machine learning approach*. XYZ Publishers.

[73] Wanamaker, J. (n.d.). Famous quotes. Retrieved from `https://www.brainyquote.com/quotes/john_wanamaker_141483`

[74] University of Chicago. (n.d.). The development of marketing mix modeling. Retrieved from `https://www.Uchicago.edu/`

[75] Consumer Goods Company. (n.d.). Case study: Optimizing marketing spend. Retrieved from `https://www.consumergoods.com/`

[76] latentview Analytics. (2019). *Marketing mix model*. Retrieved from `https://www.latentview.com/wp-content/uploads/2019/08/Marketing-Mix-Model.pdf`

[77] Marketo. (n.d.). *The definitive guide to marketing metrics and marketing analytics*. Retrieved from `https://engage.marketo.com/rs/460-TDH-945/images/definitive-guide-to-marketing-metrics-marketing-analytics.pdf`

[78] Bardas, A. (n.d.). Increasing of system's entropy on the example of coal mine operations research. Retrieved from `https://www.researchgate.net/publication/24045667`

[79] Netzer, O., et al. (2008). Hidden Markov models in marketing. In *Advanced Data Analysis and Modeling in Chemical Engineering* (pp. 491-524). Springer.

[80] Dash, M., & Sharma, K. (2019). Marketing response modelling: Impact of digital marketing for a luxury car brand. *Marketing Review St. Gallen*, 36(1), 70-77.

[81] Hanssens, D. M., Pauwels, K. H., & Dekimpe, M. (2003). Demonstrating the value of marketing. *Journal of Marketing*, 67(4), 1-17.

[82] Dekimpe, M., & Hanssens, D. M. (1999). Sustained spending and persistent response: A new look at long-term marketing profitability. *Journal of Marketing Research*, 36(4), 397-412.

[83] Dominique M. Hanssens Koen H. Pauwels (2016). Demonstrating the Value of Marketing. 174 / Journal of Marketing: AMA/MSI Special Issue, November 2016

[84] Lodish, L. M., & Mela, C. F. (1995). Long-term impact of advertising and promotion on consumer brand choice. *Journal of Marketing Research*, 32(2), 150-165.

[85] Slotegraaf, R. J., & Pauwels, K. H. (2008). The impact of brand equity and innovation on the long-term effectiveness of promotions. *Journal of Marketing Research*, 45(3), 293-306.

[86] Cain, P. M. (2016). Marketing mix modelling and return on investment. In P. J. Kitchen (Ed.), *Integrated brand marketing and measuring returns* (pp. 87-102). Springer.

[87] Cain, P. M. (2014). Brand management and the marketing mix model. 2014 Macmillan Publishers Ltd. 2050-3318 Journal of Marketing Analytics Vol. 2, 1, 33–42

[88] Pauwels, K. H., & Hanssens, D. M. (2002). Market response models and optimal budgeting for advertising. In S. Gupta (Ed.), *Marketing science handbook* (pp. 123-145).

[89] Naik, P. A., & Raman, K. (2003). Understanding the impact of synergy in multimedia communications. *Journal of Marketing Research*, 40(4), 375-388.

[90] Pauwels, K. H., & Hanssens, D. M. (2004). Effects of marketing mix variables on global brand equity. *Journal of Marketing*, 68(4), 80-93.

11. Appendix A - Dataset G

Q–Q Plot:

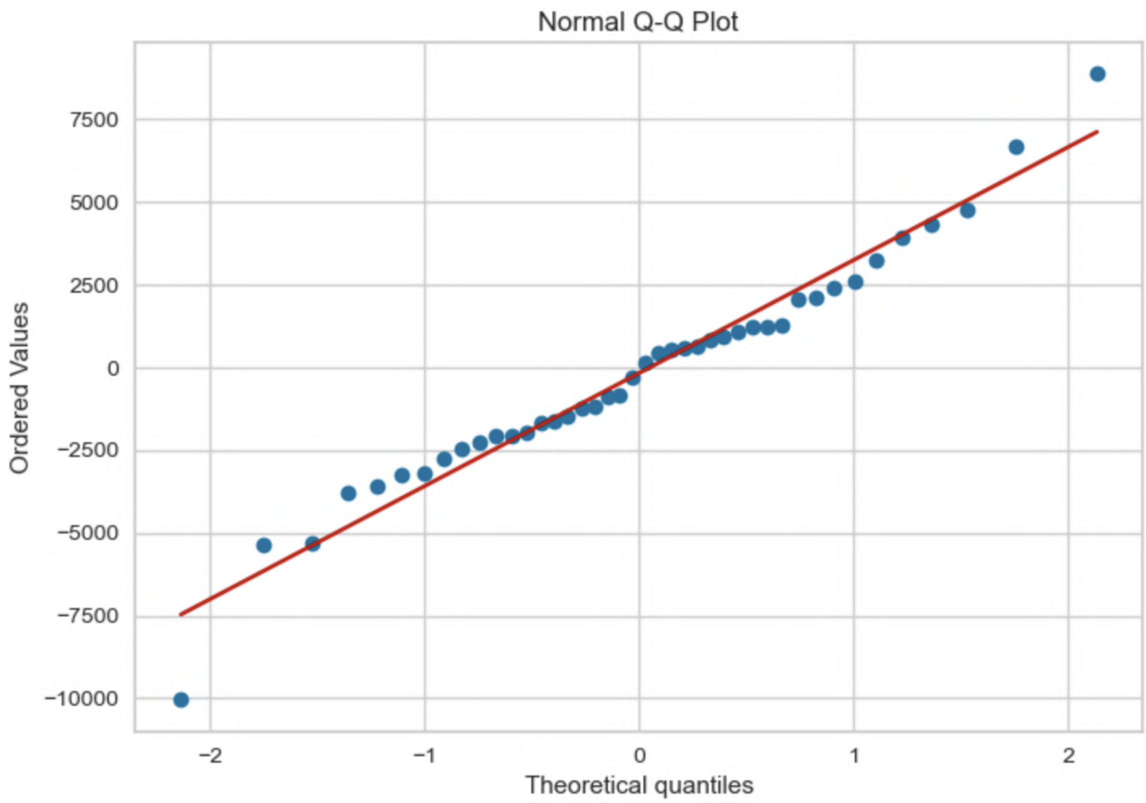

Figure 86: QQ-plot Dataset G

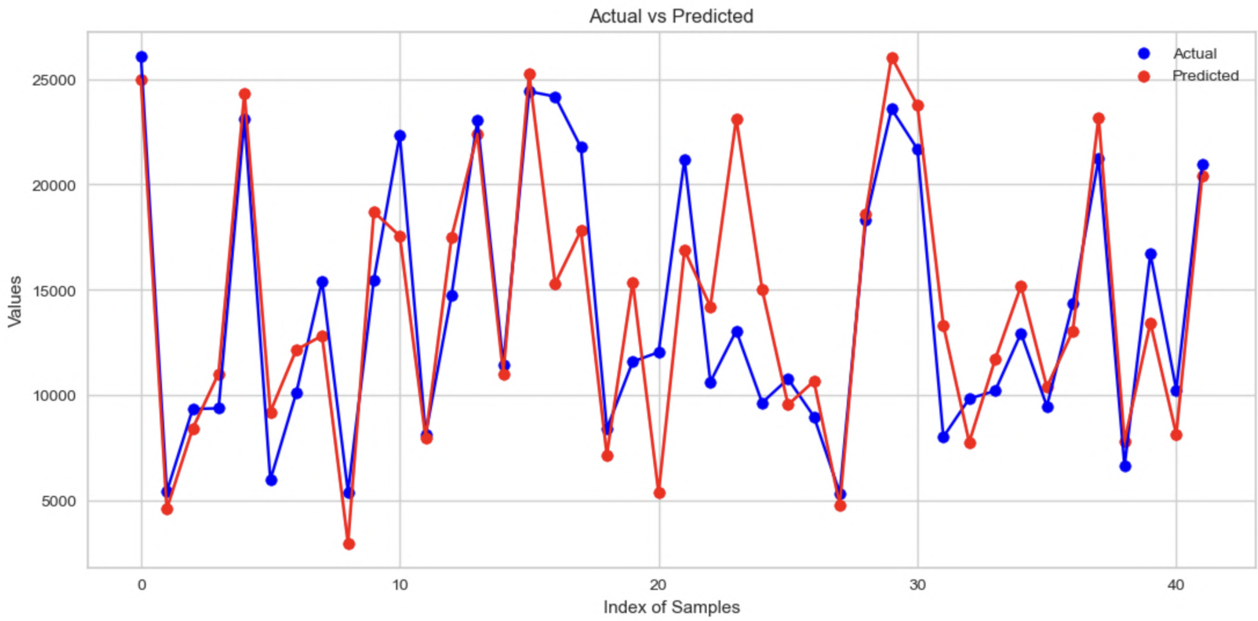

Figure 87: Actual vs. Predicted Values

```
Residuals Scatter Plot:
```

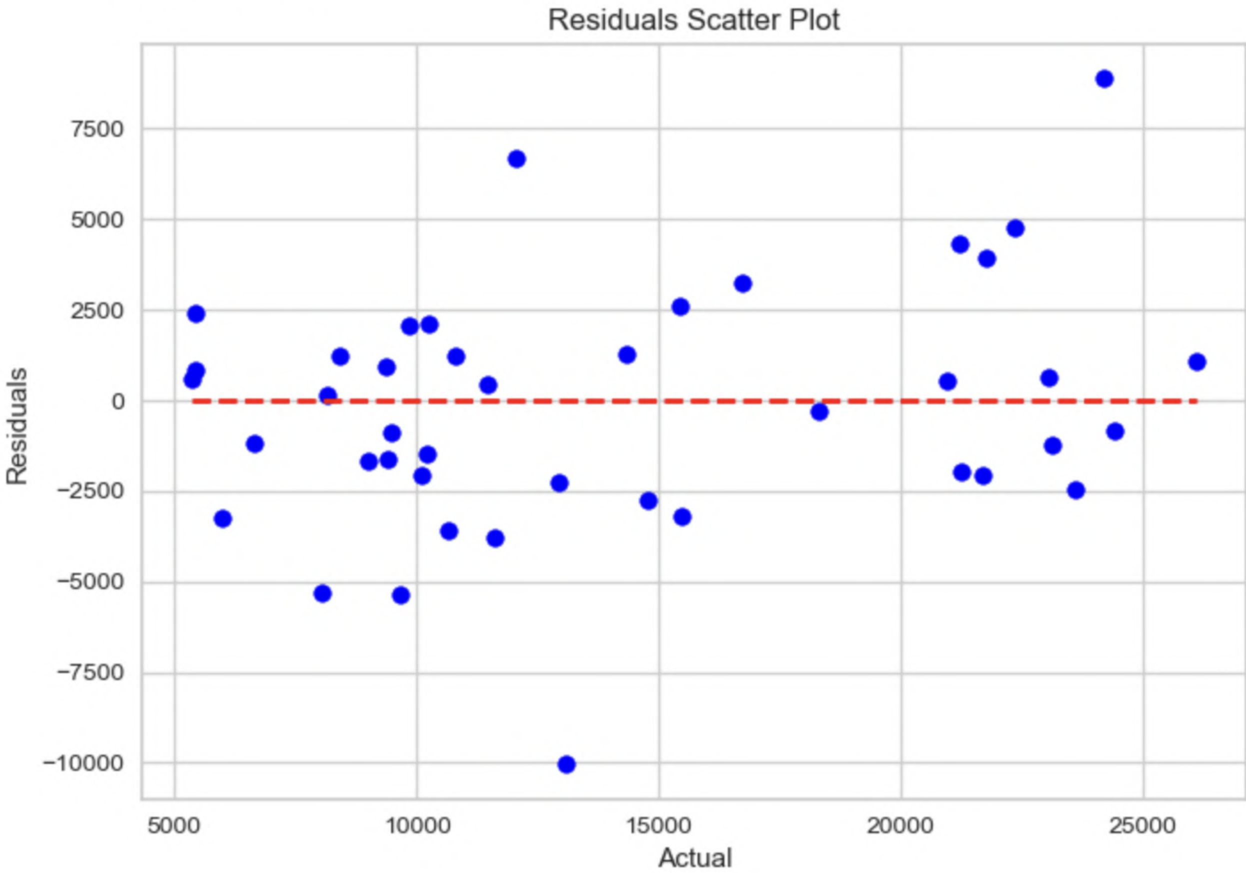

Figure 88: Residuals Scatter Plot Dataset G

```
Residuals Histogram:
```

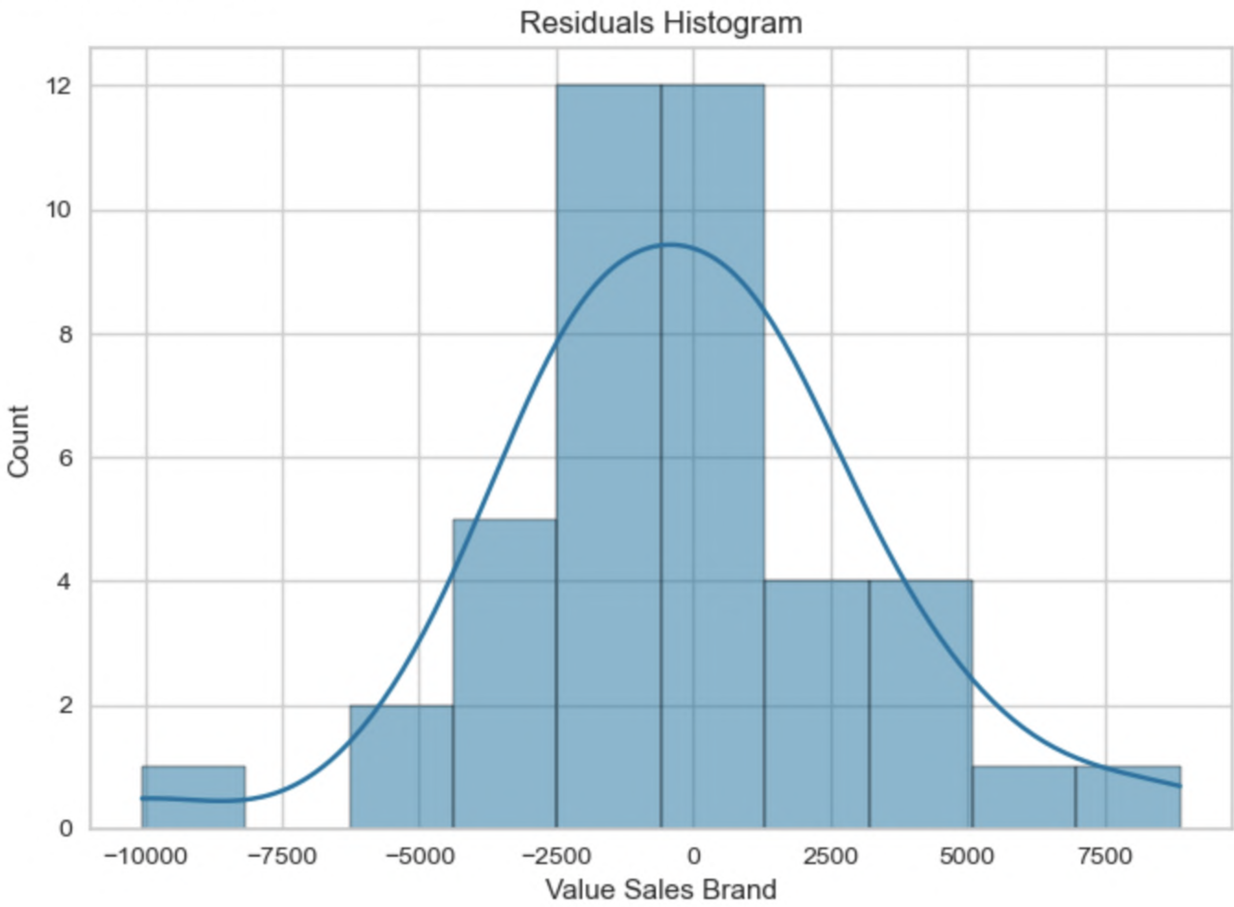

Figure 89: Residual Histogram Dataset G

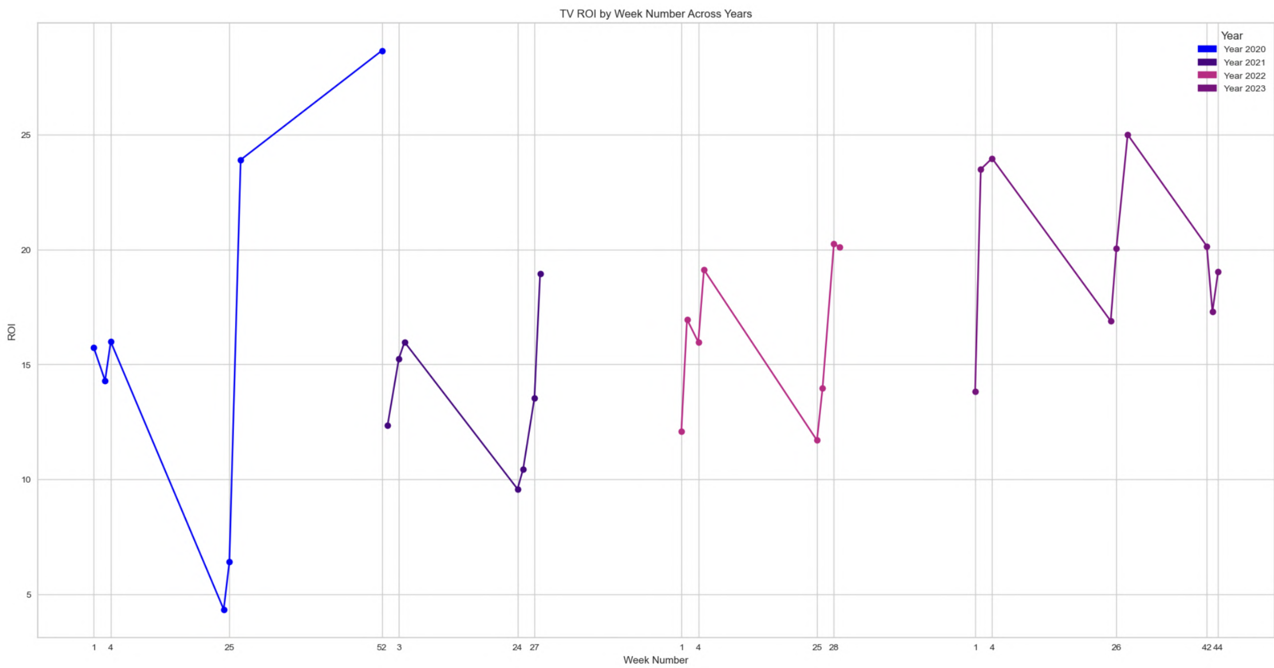

Figure 90: TV ROI by number of weeks Dataset G

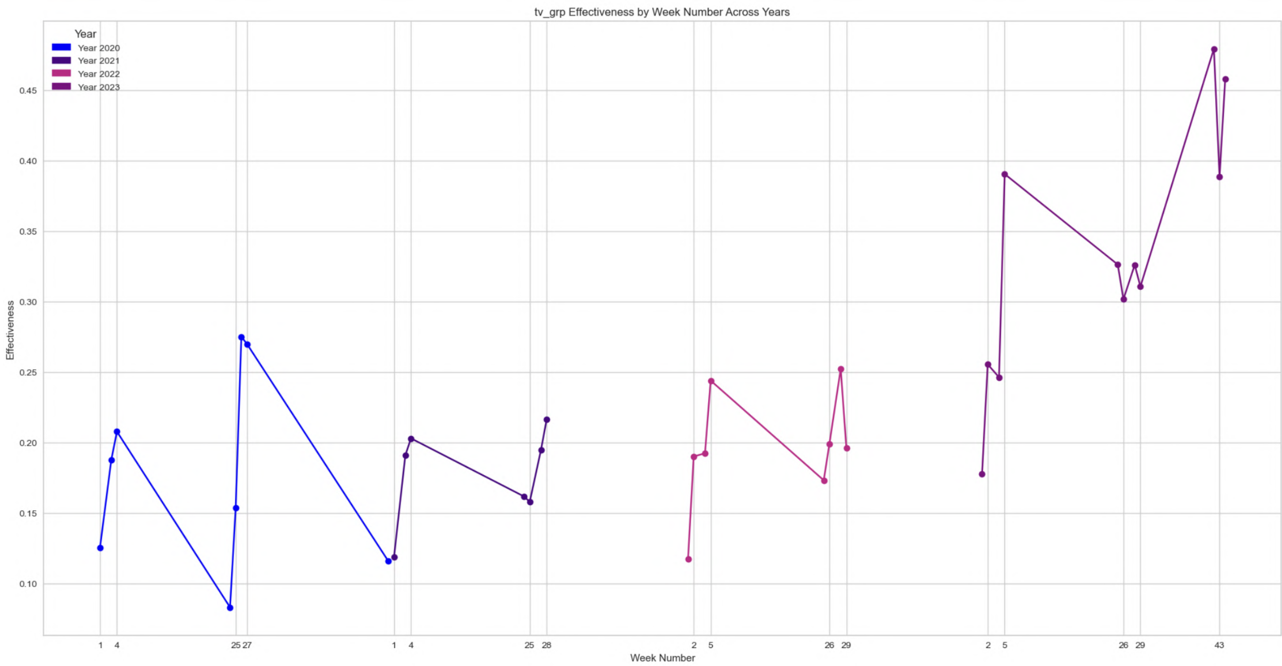

Figure 91: TV Effectiveness by number of weeks Dataset G

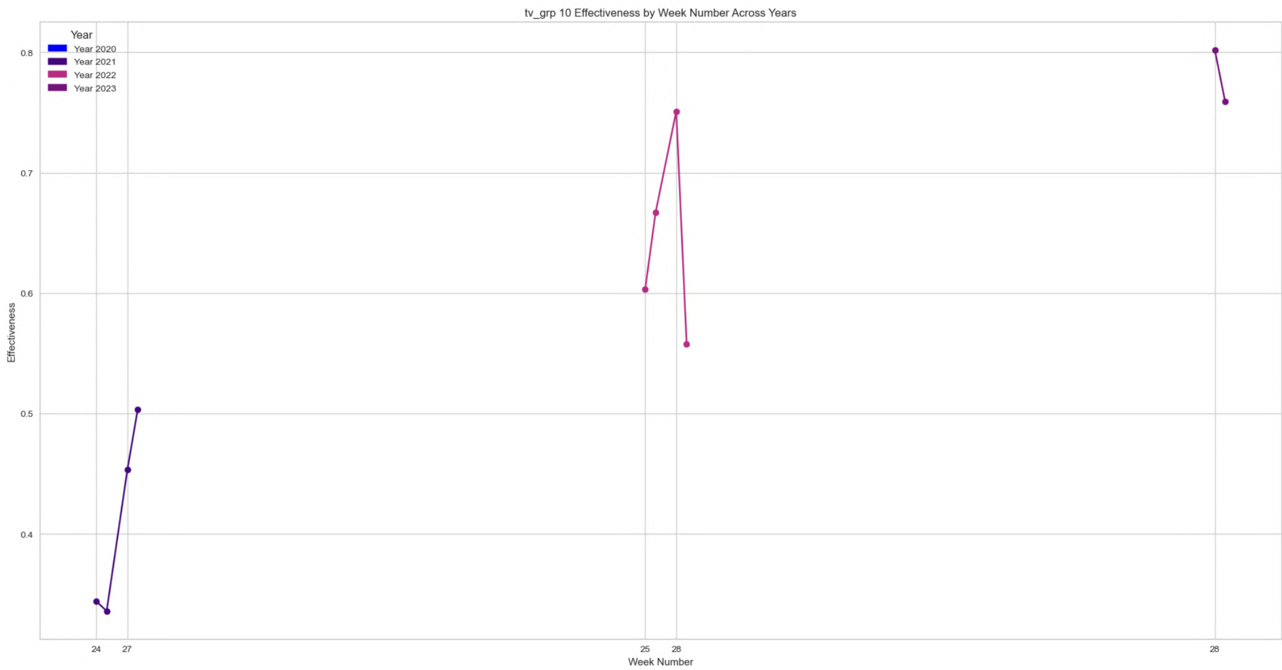

Figure 92: TV GRP 10 Effectiveness by number of weeks Dataset G

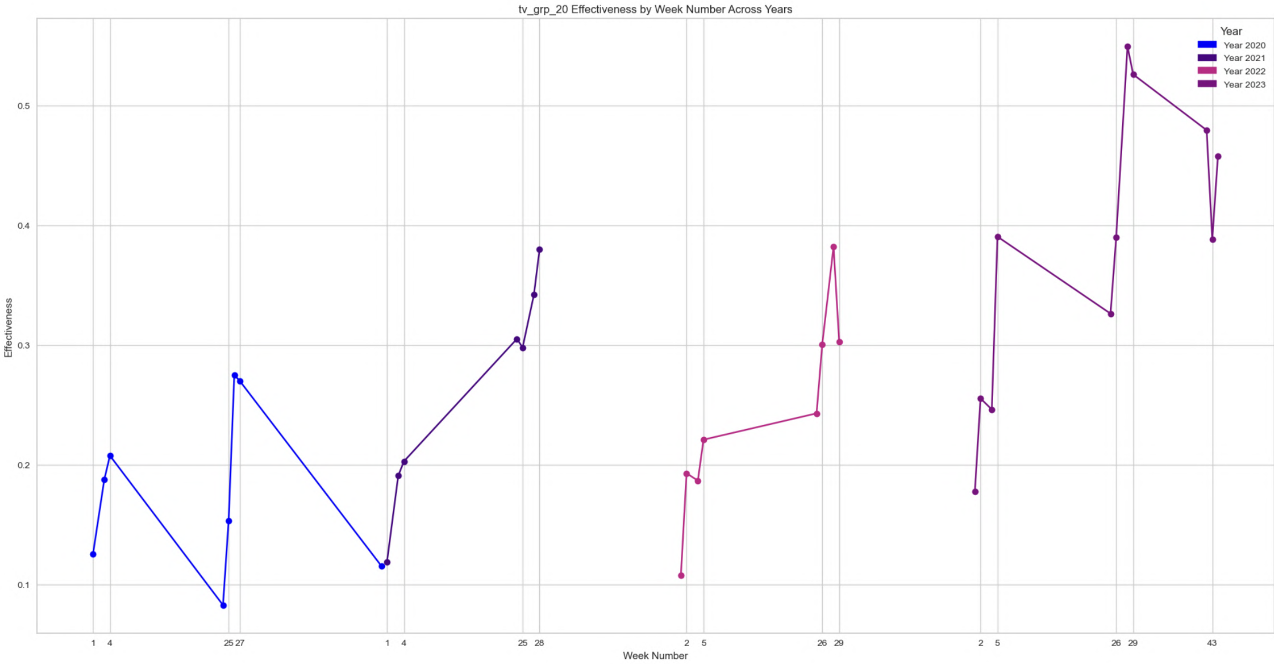

Figure 93: TV GRP 20 Effectiveness by number of weeks Dataset G

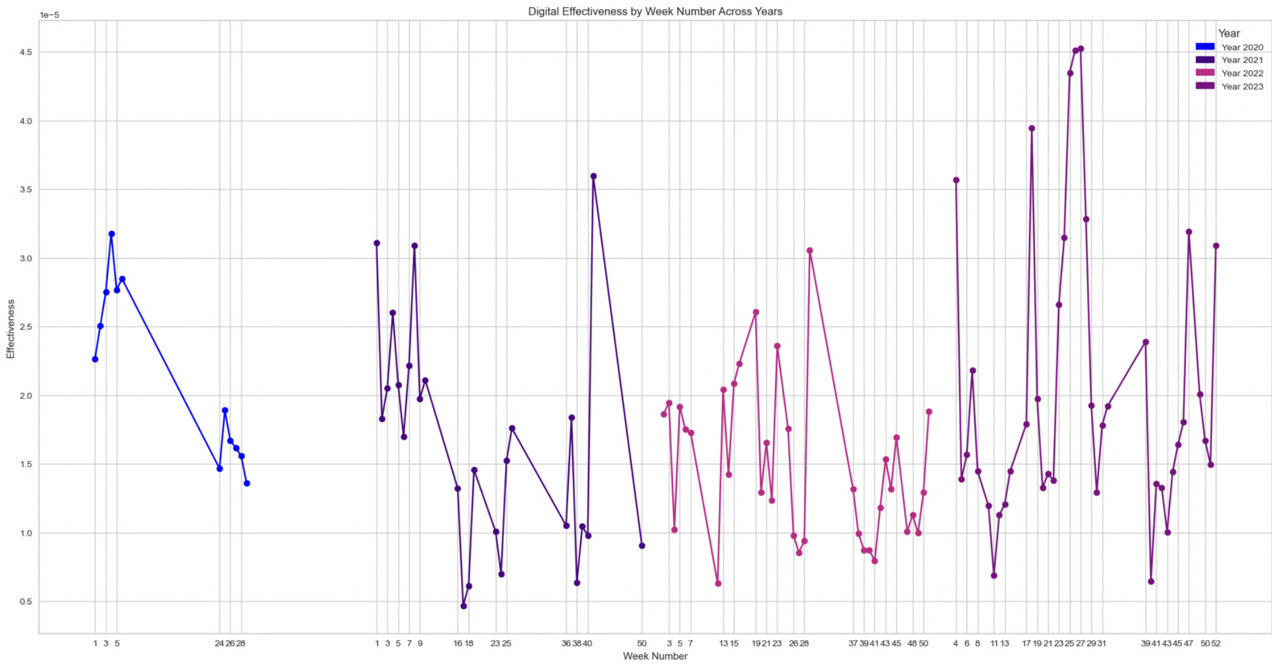

Figure 94: Digital Effectiveness by number of weeks Dataset G

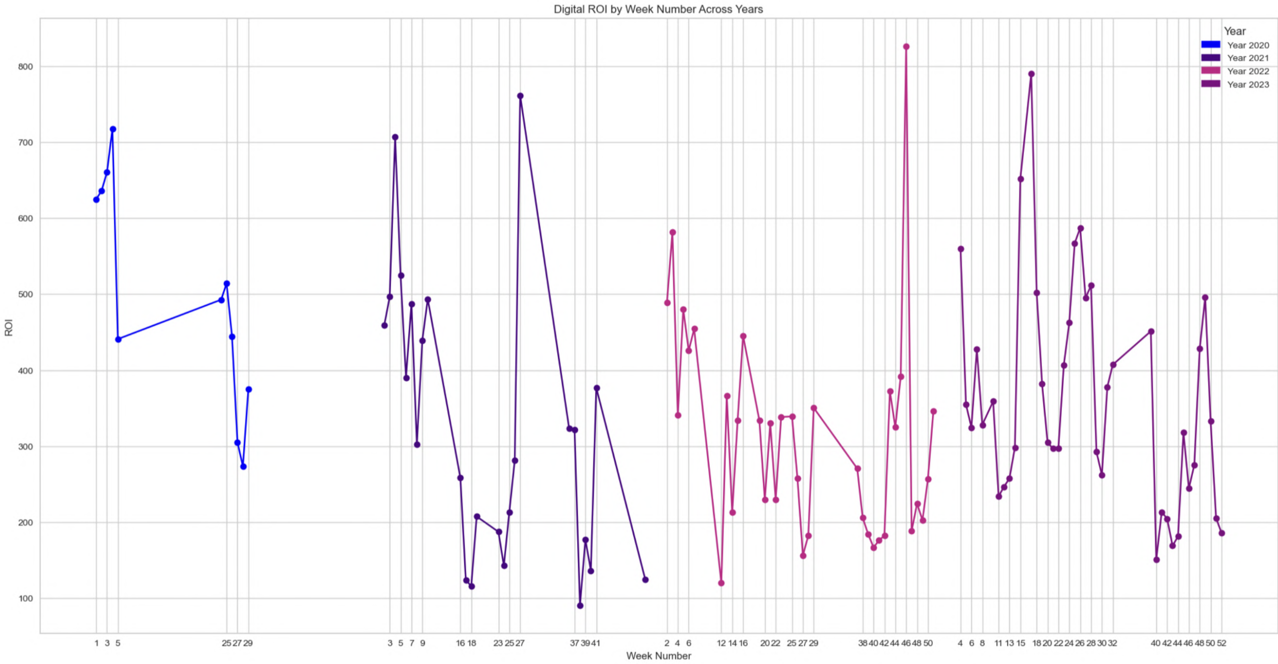

Figure 95: Digital ROI by number of weeks Dataset G

12. Appendix B - Dataset D

Q-Q Plot:

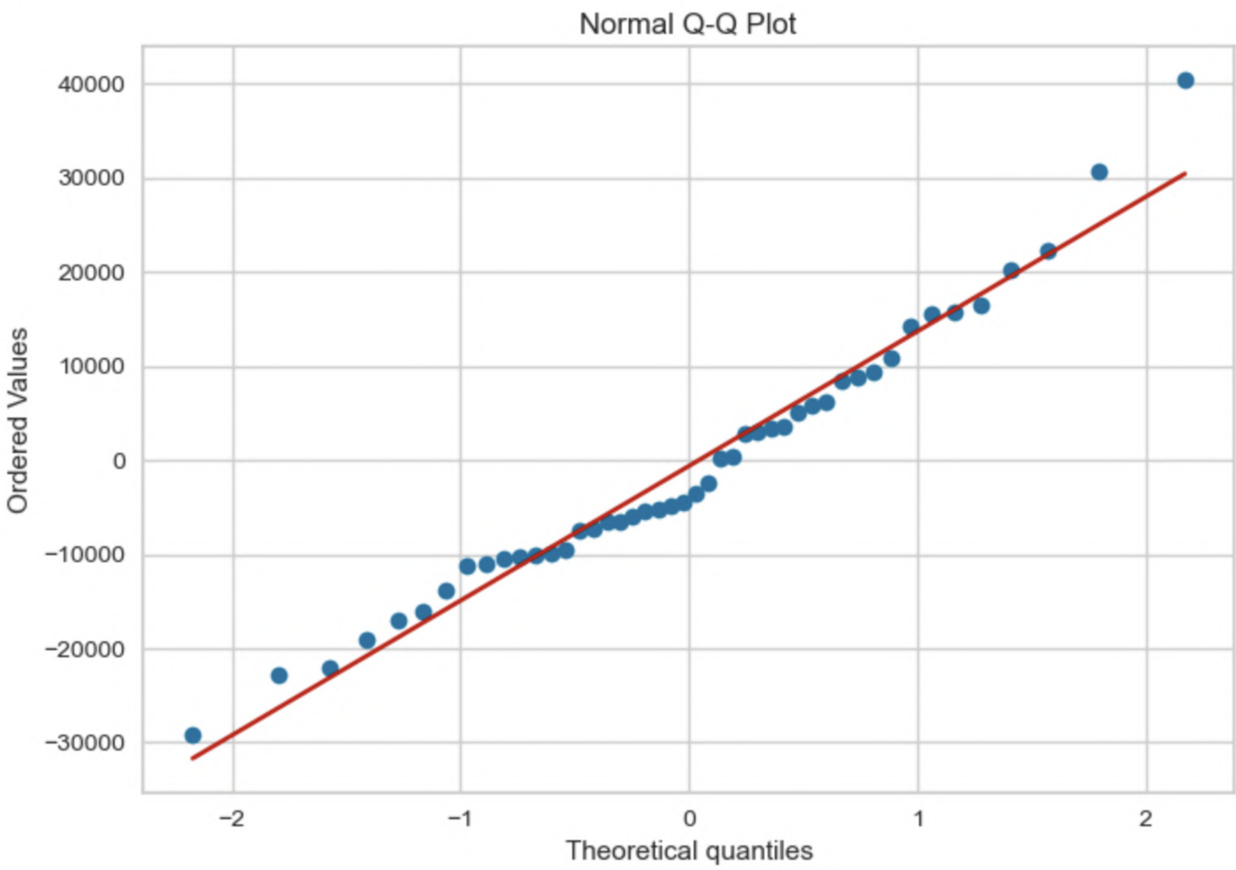

Figure 96: QQ Plot Dataset D Sales 1

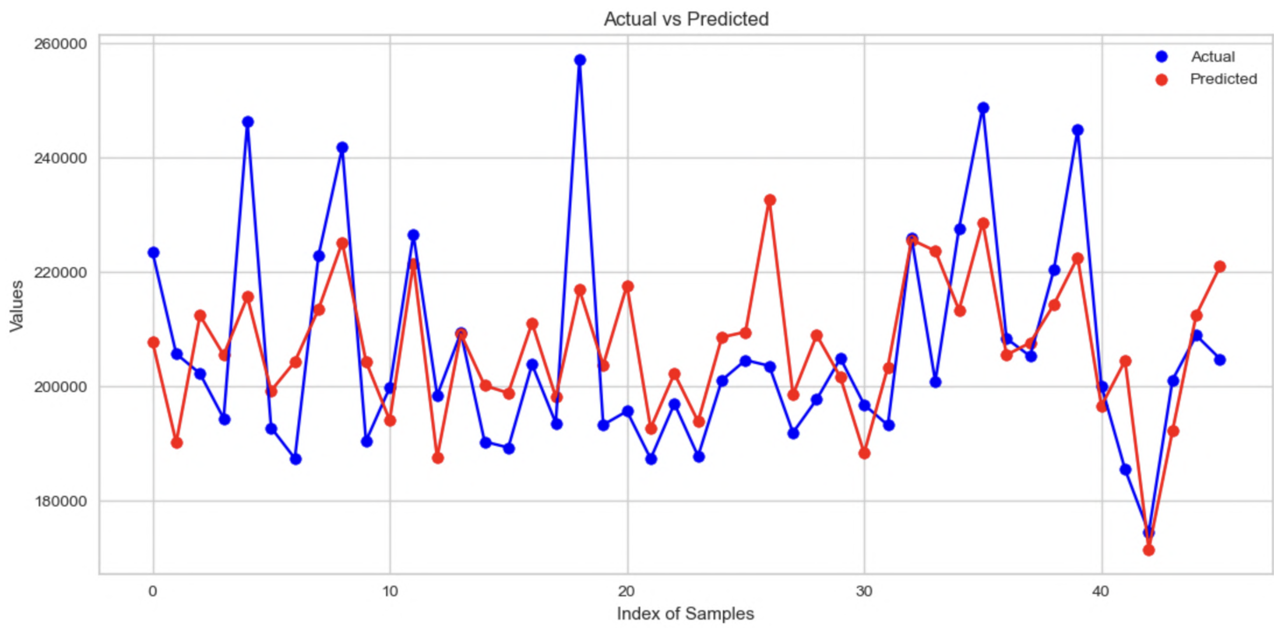

Figure 97: Actual vs. Predicted Values Dataset D Sales 1

Residuals Scatter Plot:

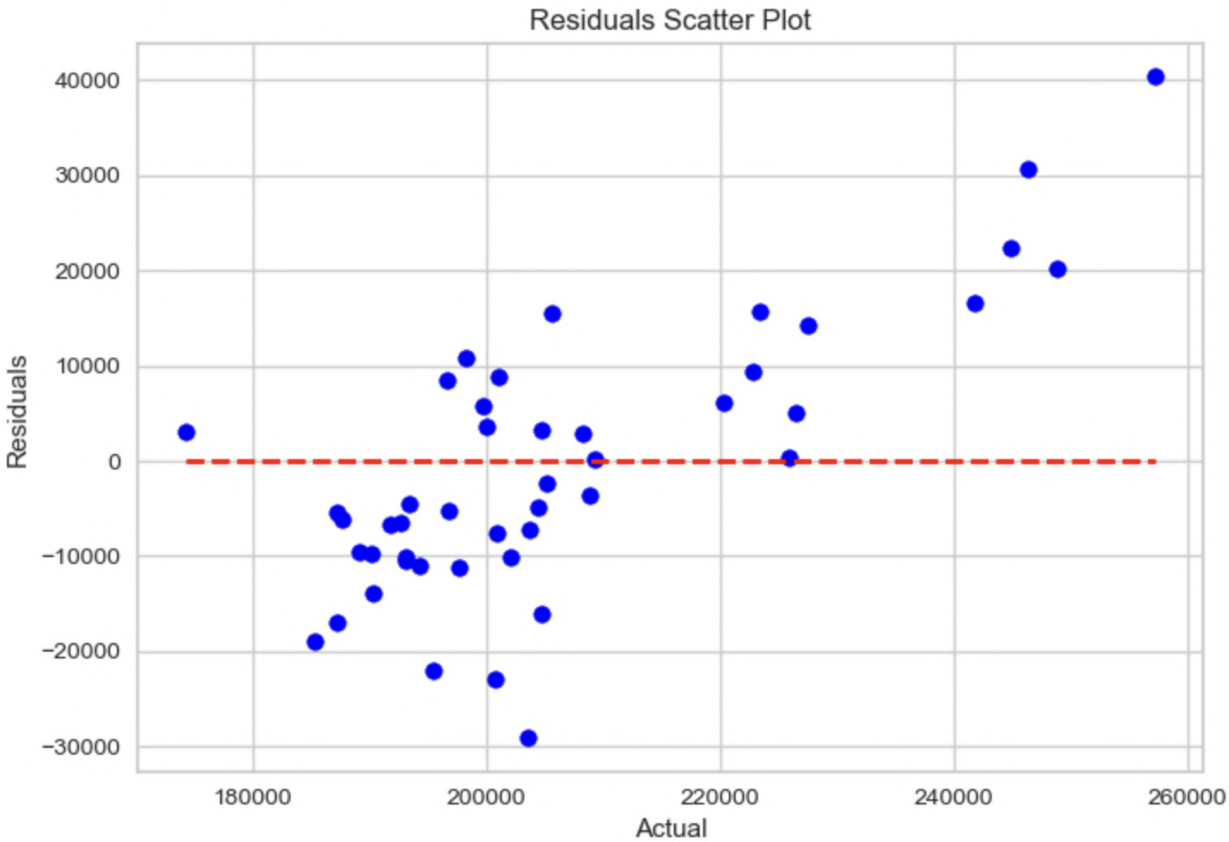

Figure 98: Residuals Scatter Plot Dataset D Sales 1

```
Residuals Histogram:
```

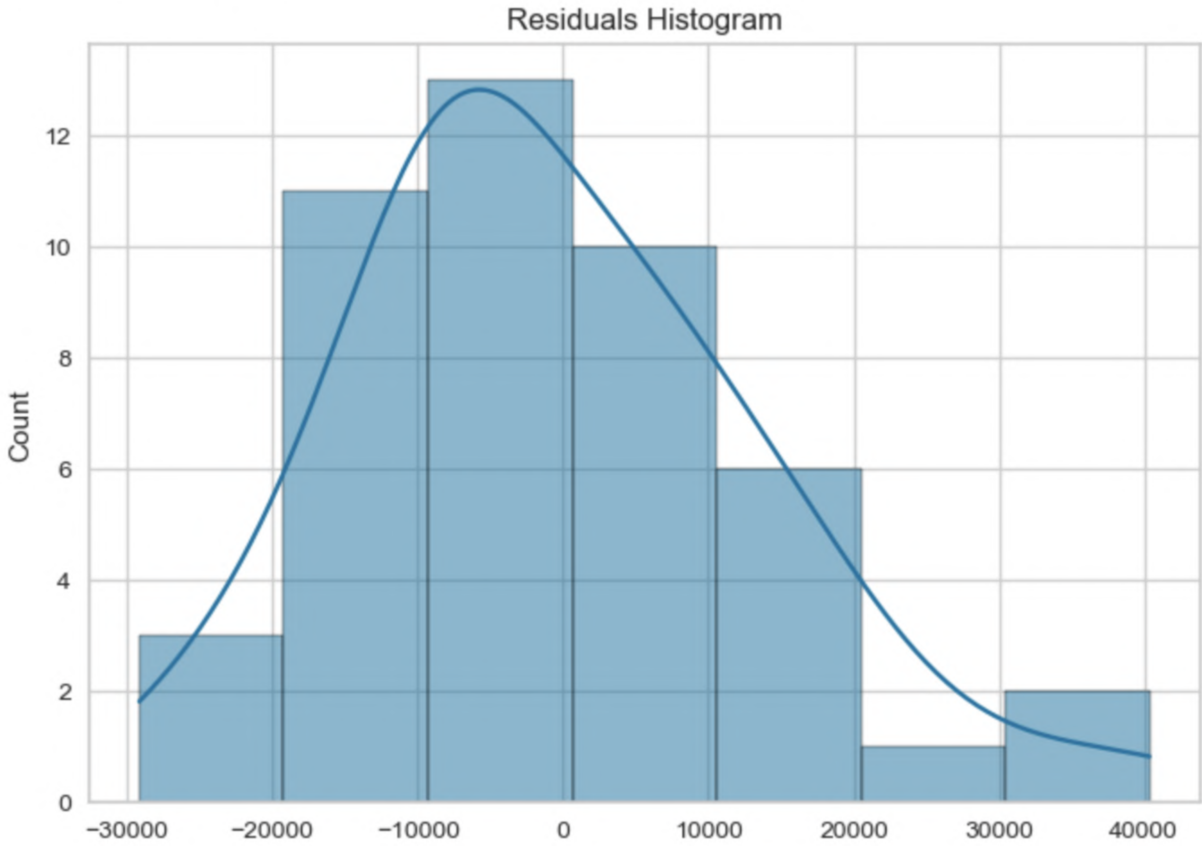

Figure 99: Residuals Histogram Dataset D Sales 1

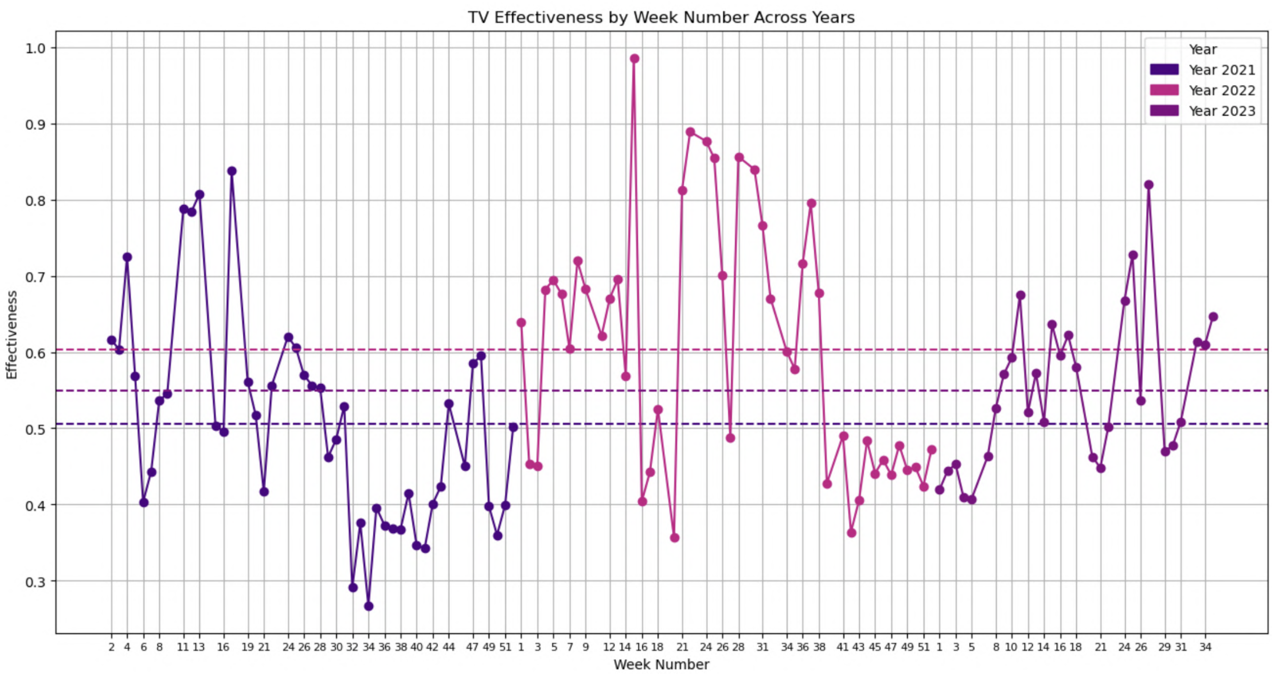

Figure 100: TV Effectiveness Dataset D Sales 1

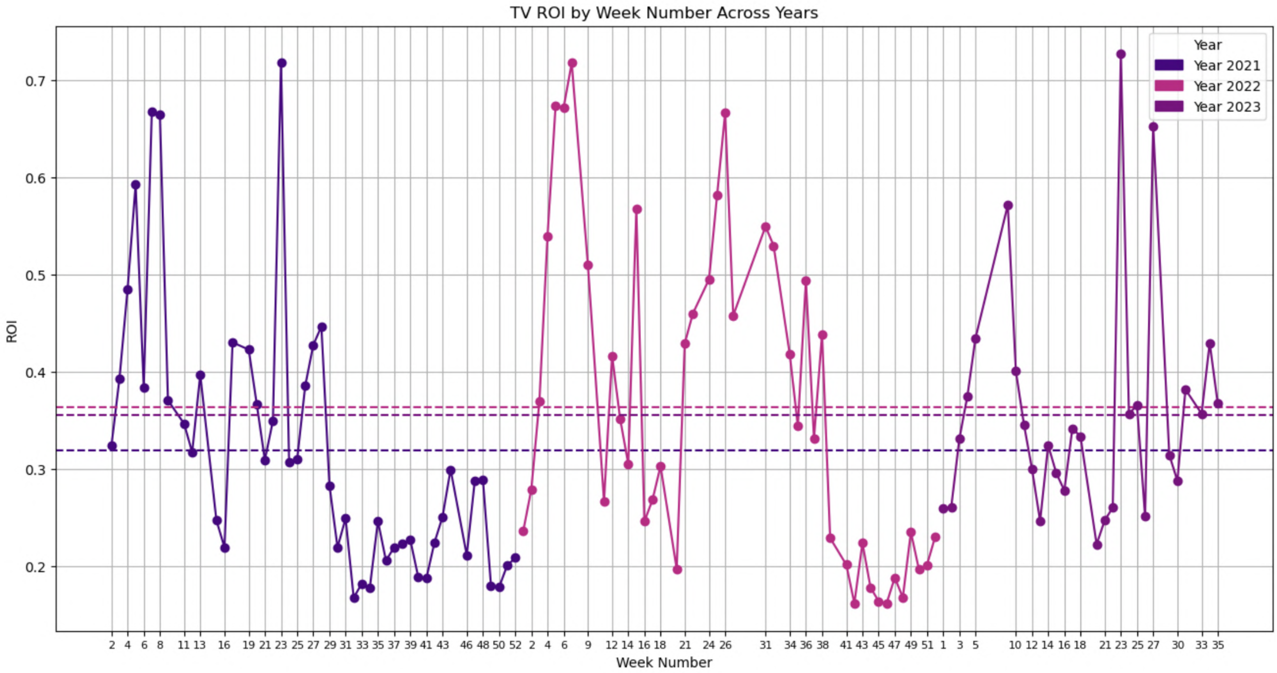

Figure 101: TV ROI Dataset D Sales 1

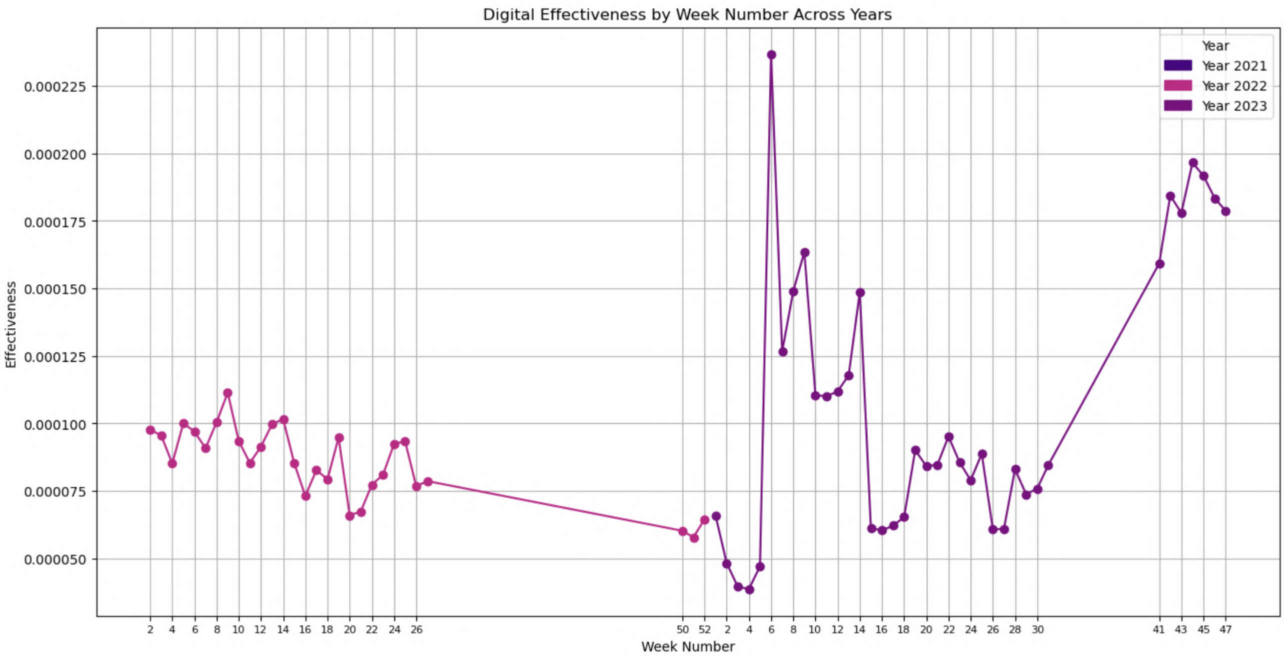

Figure 102: Digital Effectiveness Dataset D Sales 1

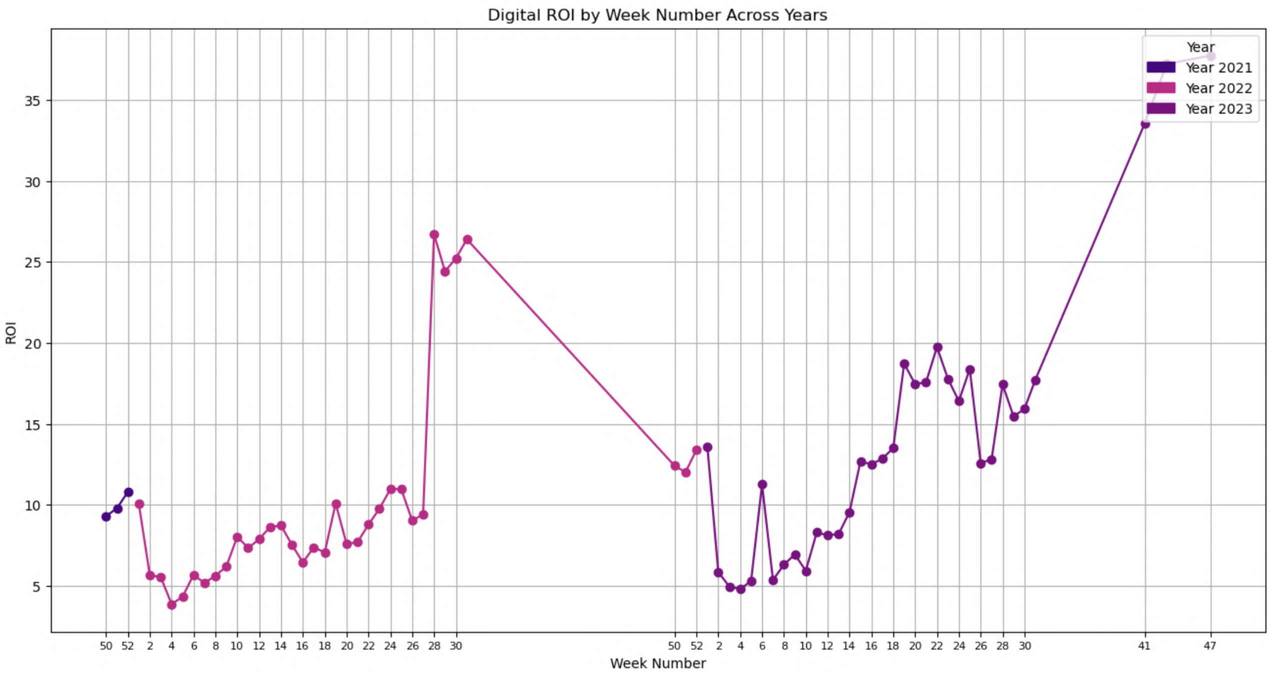

Figure 103: Digital ROI Dataset D Sales 1

```
The percentage contribution to the sales of IMPS CTV_lag_halo is: 0.13560687149219816
The percentage contribution to the sales of IMPS Social_lag_halo is: 0.13206047120481096
The percentage contribution to the sales of IMPS Influencers_lag_halo is: 0.18992366030245889
The percentage contribution to the sales of IMPS YT 10s_lag_halo is: 0.2614145161329172
The percentage contribution to the sales of IMPS YT 6s_lag_halo is: 0.051771701153604716
The percentage contribution to the sales of IMPS Programmatc Display_lag_halo is: 0.07136215880357157
The percentage contribution to the sales of IMPS Programmatic Video 2_lag_halo is: 0.03854720177476336
The percentage contribution to the sales of IMPS Programmatic Video 1_lag_halo is: 0.11931341913567517
```

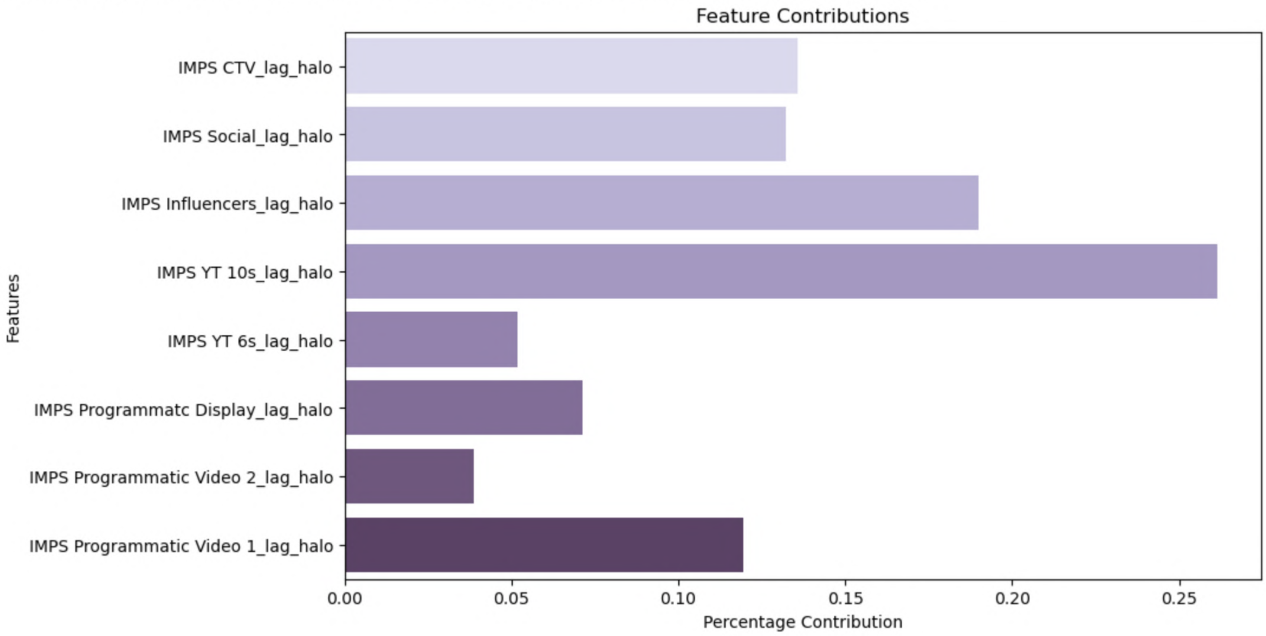

Figure 104: Digital Features Importance Coefficients Dataset D Sales 1

```
The percentage contribution to the sales of TVC_NO COPY is: 0.007967709374049009
The percentage contribution to the sales of TVC_COPY 3 is: 0.024477953908246027
The percentage contribution to the sales of TVC_COPY 2 is: 0.0034832402751496052
The percentage contribution to the sales of TVC_COPY 1 is: 0.011386909409932248
The percentage contribution to the sales of 30" is: 0.022392910870018927
The percentage contribution to the sales of 20'' is: 0.14733085286449174
The percentage contribution to the sales of 10" is: 0.01536682727472211
The percentage contribution to the sales of Not OOB is: 0.08135188278093707
The percentage contribution to the sales of OOB is: 0.31933275540395084
The percentage contribution to the sales of Minor Channels is: 0.06938872927997125
The percentage contribution to the sales of Main Channels is: 0.09002446320574672
The percentage contribution to the sales of DT/NT is: 0.10531384521815781
The percentage contribution to the sales of PT is: 0.10218192013462671
```

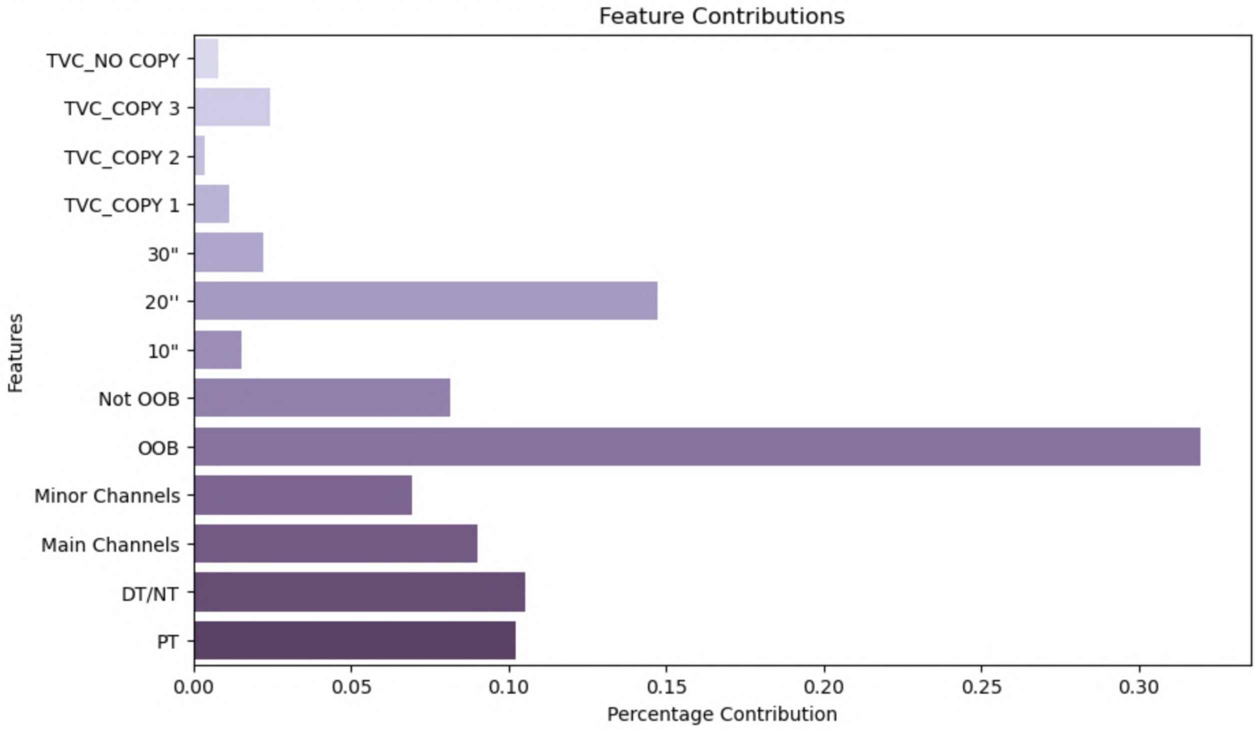

Figure 105: TV Features Importance Coefficients Dataset D Sales 1

Q-Q Plot:

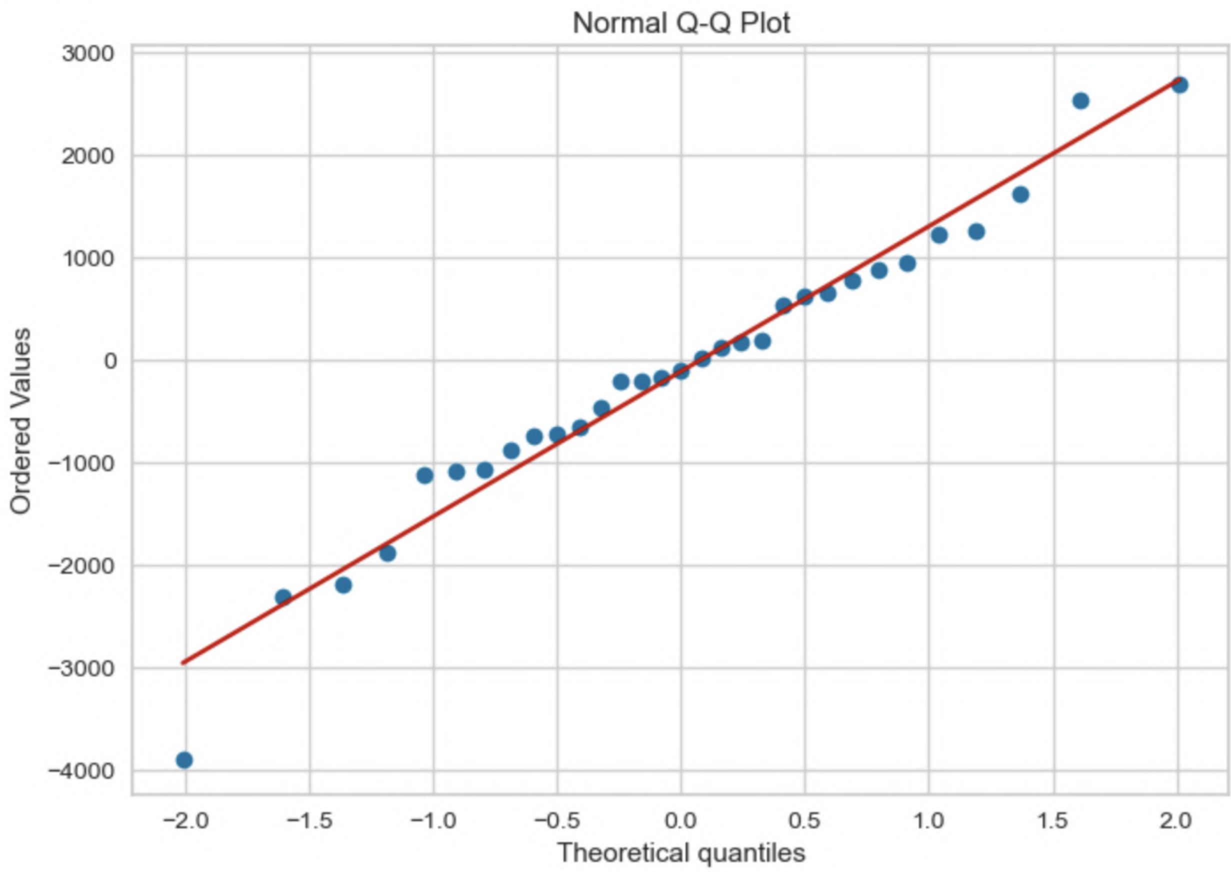

Figure 106: QQ-plot Dataset D Sales 2

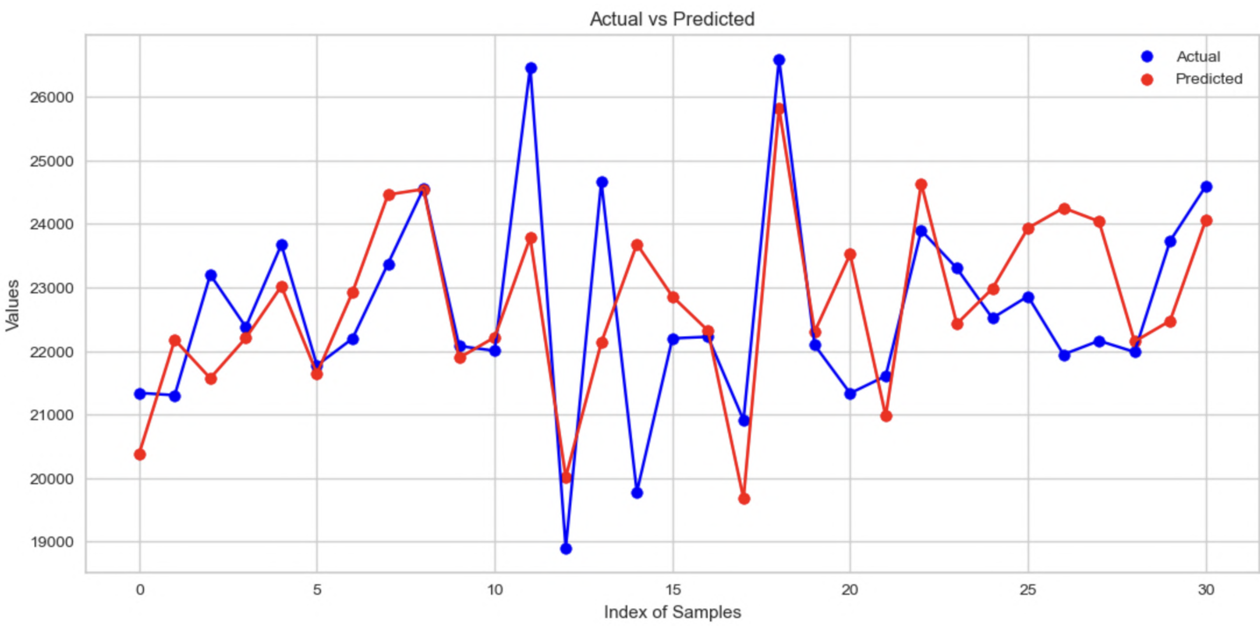

Figure 107: Actual vs. Predicted Values Dataset D Sales 2

Residuals Scatter Plot:

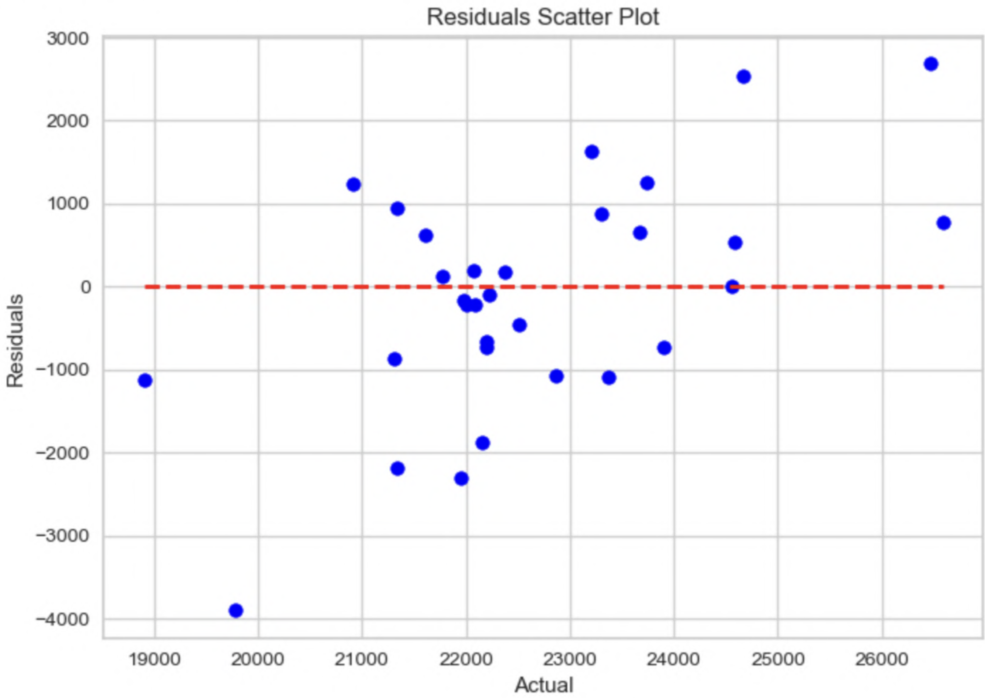

Figure 108: Residuals Scatter Plot Dataset D Sales 2

Residuals Histogram:

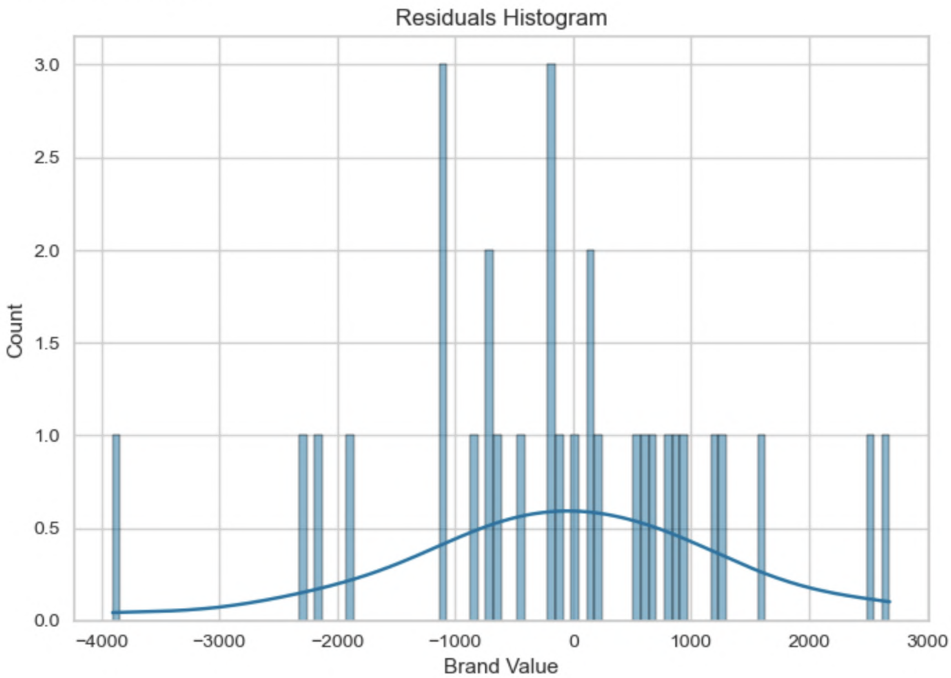

Figure 109: Residuals Histogram Dataset D Sales 2

```
Q-Q Plot:
```

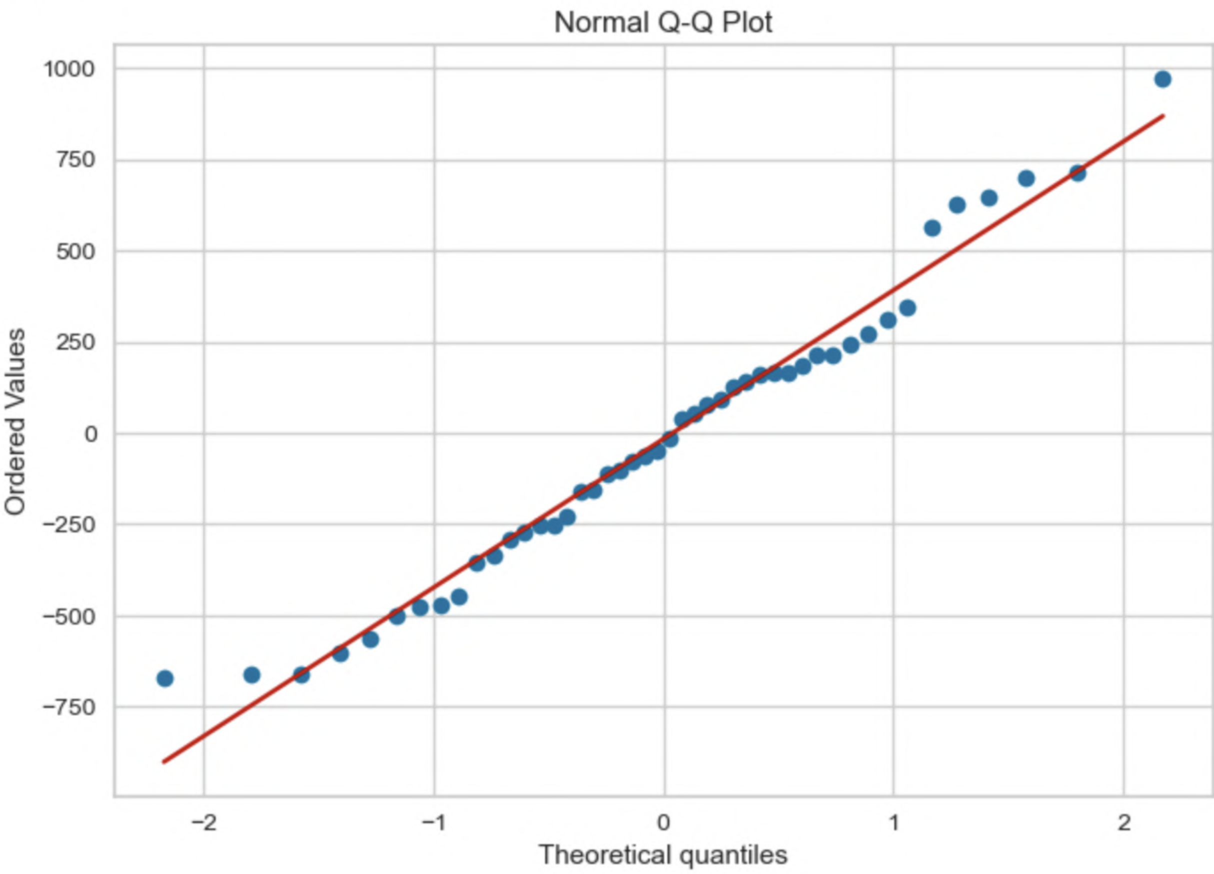

Figure 110: QQ-plot Dataset D Sales 3

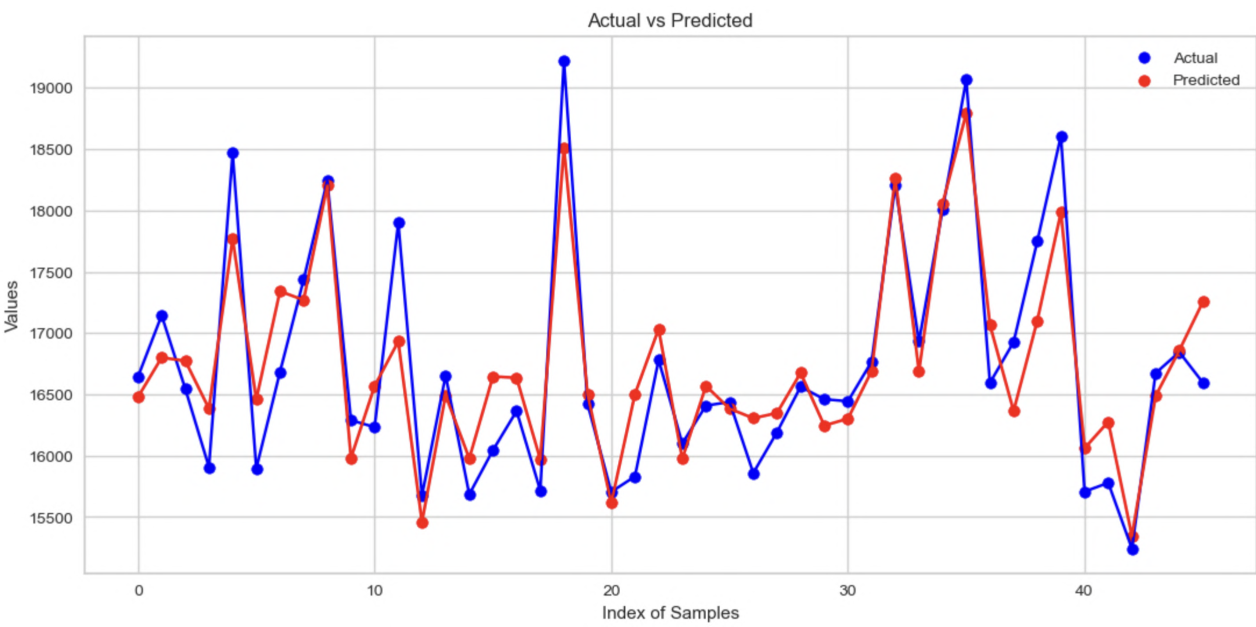

Figure 111: Actual vs. Predicted Values Dataset D Sales 3

```
Residuals Scatter Plot:
```

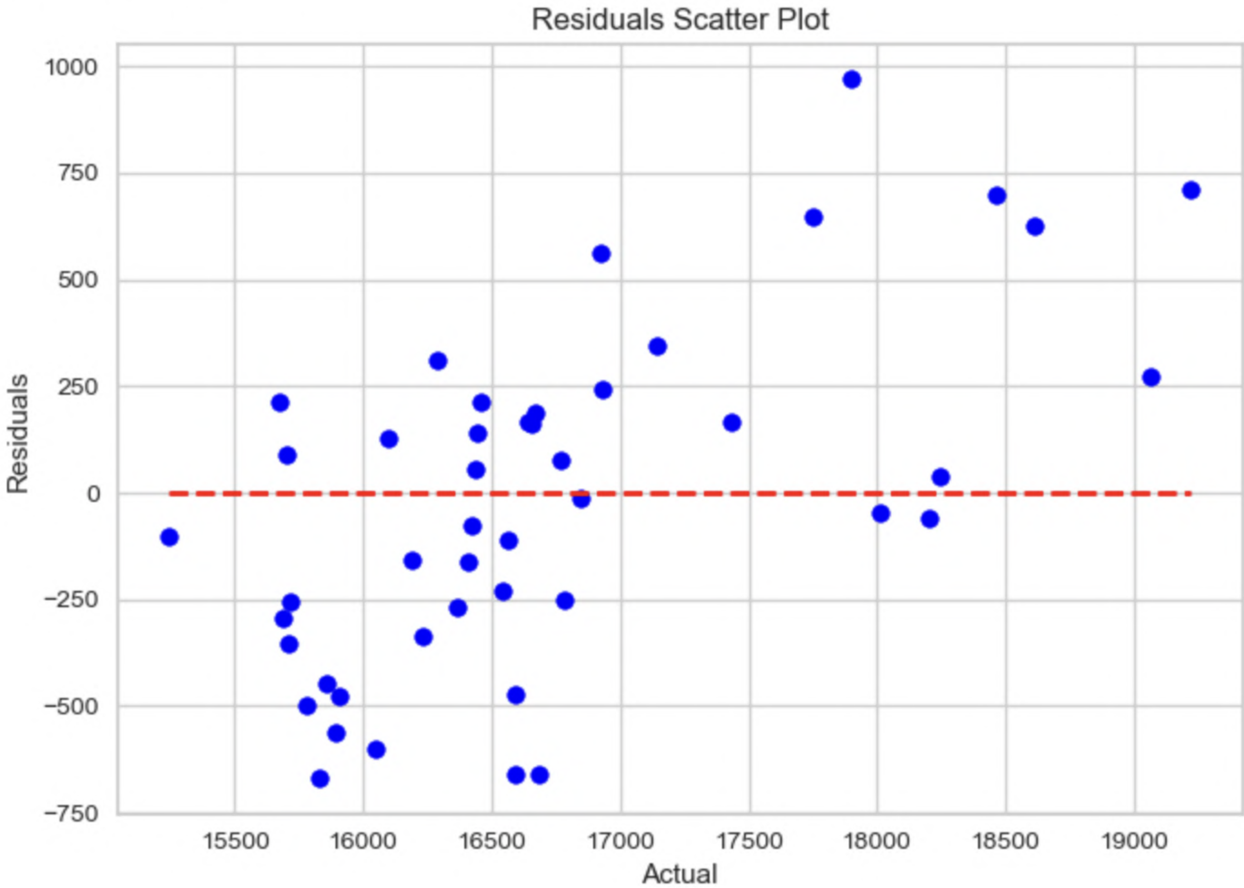

Figure 112: Residual Scatter Plot Dataset D Sales 3

```
Residuals Histogram:
```

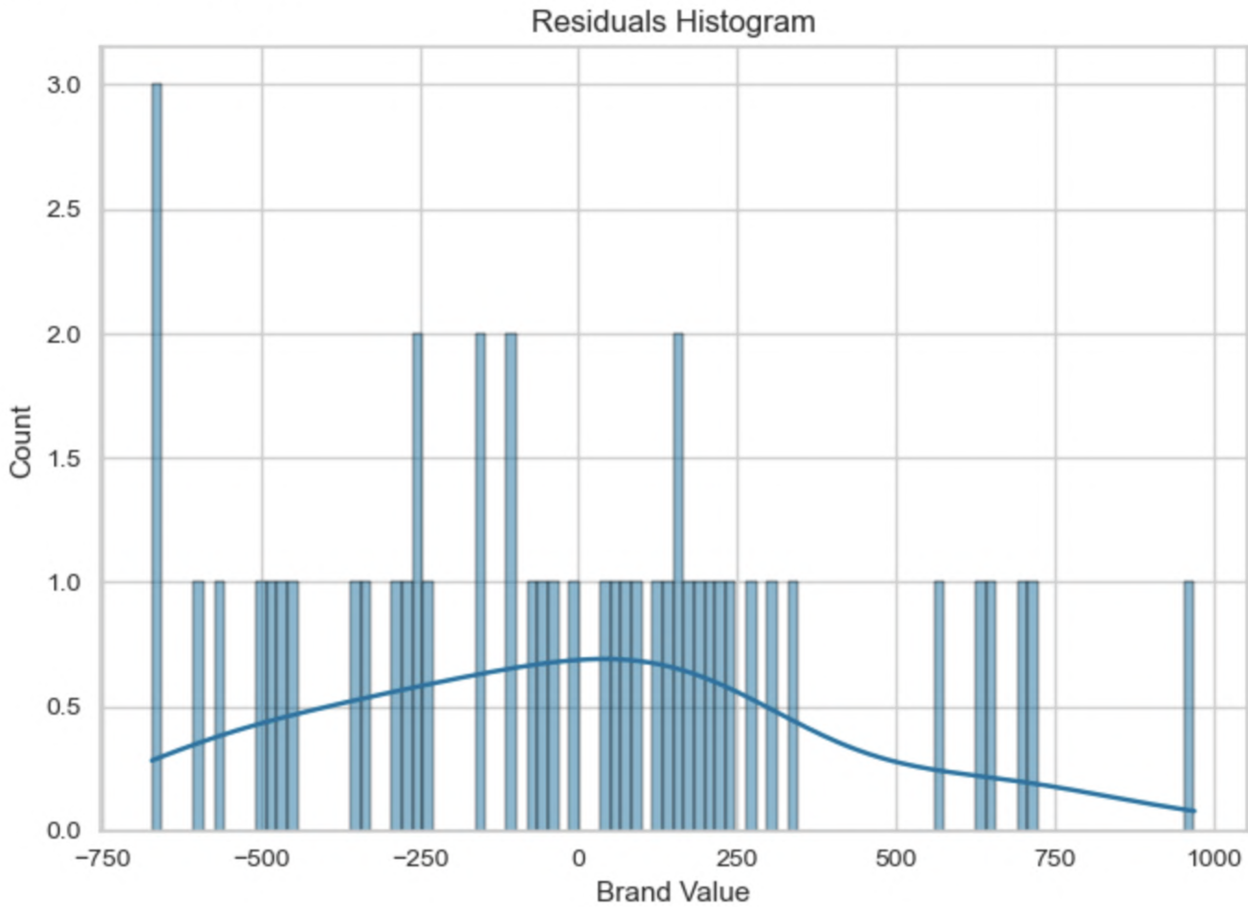

Figure 113: Residual Histogram Dataset D Sales 3

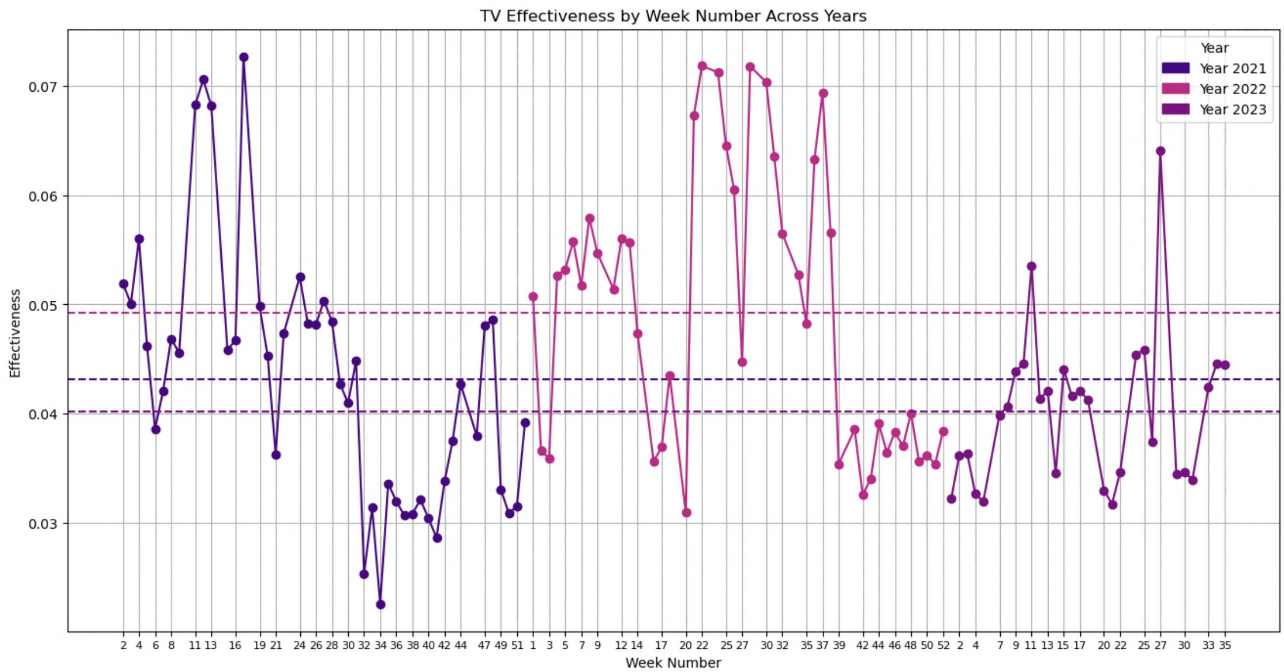

Figure 114: TV Effectiveness Dataset D Sales 3

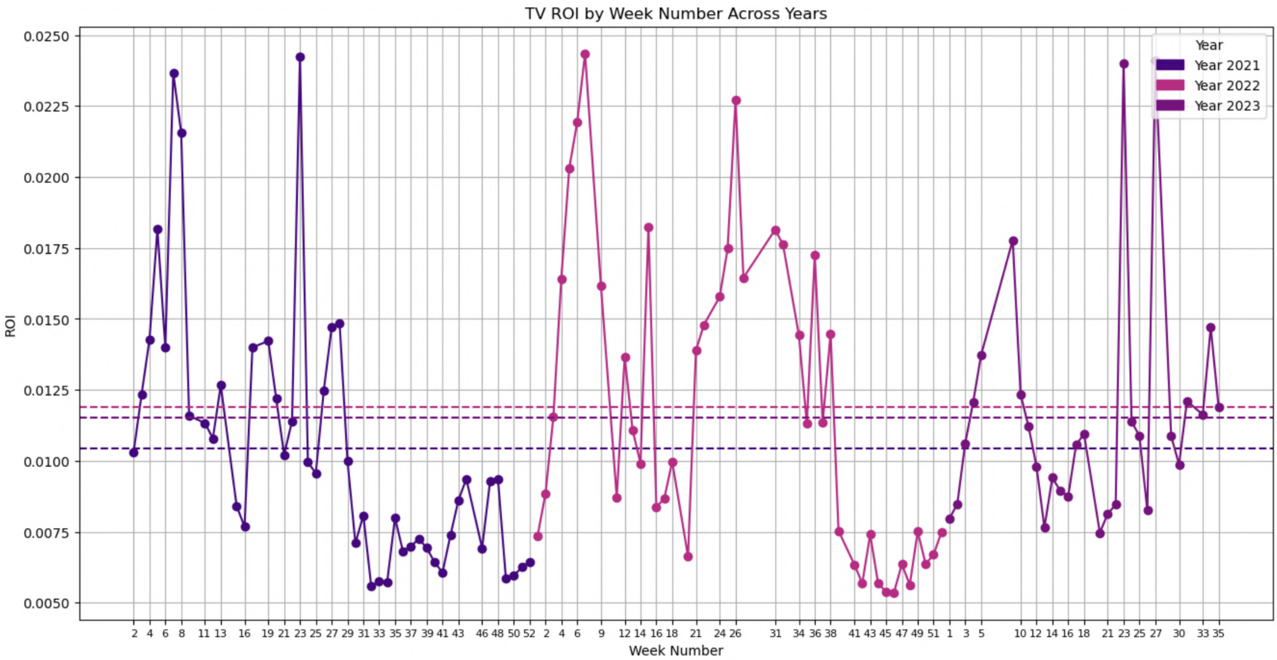

Figure 115: TV ROI Dataset D Sales 3

```
The percentage contribution to the sales of TVC_NO COPY is: 0.07028731016178838
The percentage contribution to the sales of TVC_COPY 3 is: 0.007053790336151233
The percentage contribution to the sales of TVC_COPY 2 is: 0.005770959043594972
The percentage contribution to the sales of TVC_COPY 1 is: 0.002074189335187907
The percentage contribution to the sales of 30" is: 0.00987959104443382
The percentage contribution to the sales of 20'' is: 0.06036398253927073
The percentage contribution to the sales of 10" is: 0.04316674380259249
The percentage contribution to the sales of Not OOB is: 0.1763684144144129
The percentage contribution to the sales of OOB is: 0.08062326154779777
The percentage contribution to the sales of Minor Channels is: 0.08781791774170421
The percentage contribution to the sales of Main Channels is: 0.15539012543874425
The percentage contribution to the sales of DT/NT is: 0.22205842754969413
The percentage contribution to the sales of PT is: 0.07914528704462723
```

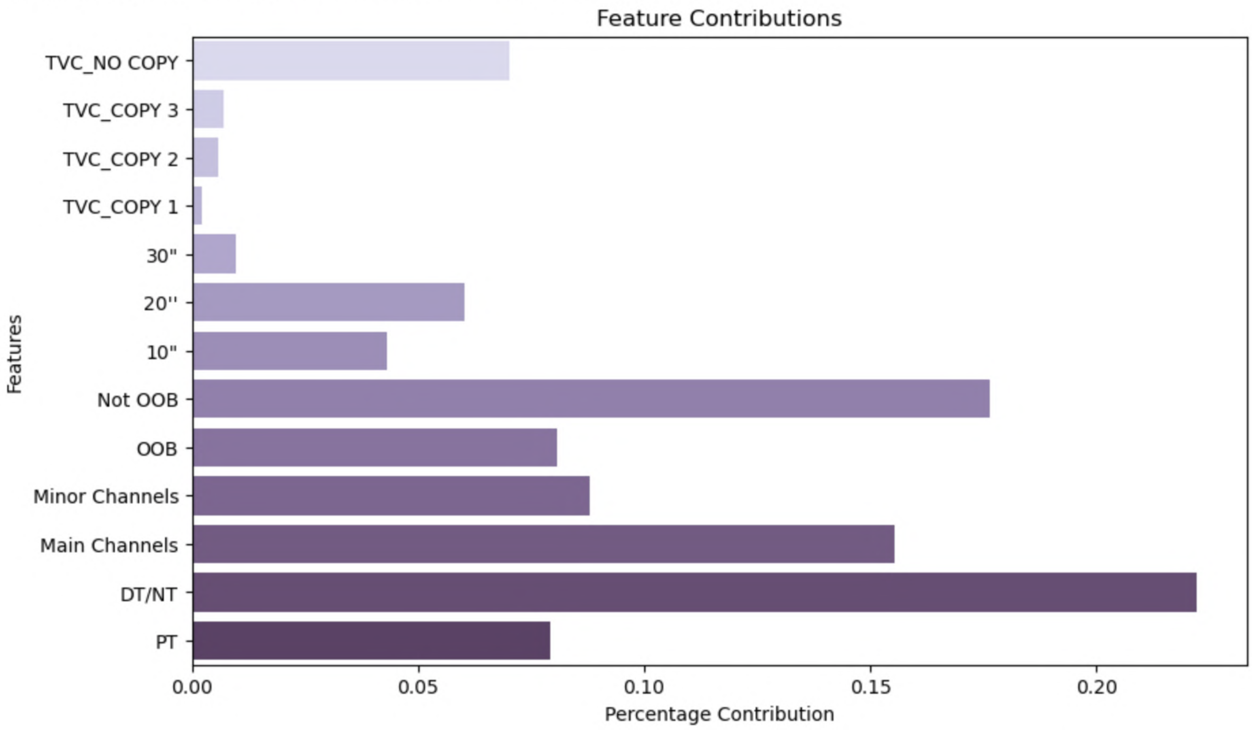

Figure 116: TV Features Importance Coefficients Dataset D Sales 3

13. Appendix C - Models Building

```
import pandas as pd
import numpy as np
from scipy.stats import skew, kurtosis

def analyze_and_transform_features(df):
    # Create a copy of the original DataFrame
    df_log = df.copy()

    # Iterate over each feature in the DataFrame
    features = df_log.columns
    columns_to_drop = []

    for feature in features:
        # Calculate skewness and kurtosis for the feature
        feature_skewness = skew(df_log[feature])
        feature_kurtosis = kurtosis(df_log[feature])

        # Check if the feature needs transformation
        needs_transformation = abs(feature_skewness) > 1 or feature_kurtosis > 3

        if needs_transformation:
            print(f"Feature: {feature}")
            print(f"Skewness: {feature_skewness}")
            print(f"Kurtosis: {feature_kurtosis}")
            print(f"Feature '{feature}' needs transformation due to skewness or kurtosis.")

            # Apply log transformation
            log_feature = np.log(df_log[feature] + 1)  # Add 1 to avoid log(0)
            new_feature_name = f"{feature}_log"
            df_log[new_feature_name] = log_feature

            # Mark the original feature for dropping
            columns_to_drop.append(feature)

    # Drop the original columns that were transformed
    df_log.drop(columns=columns_to_drop, inplace=True)

    # Return the transformed DataFrame
    return df_log
```

Figure 117: Skewness and Kurtosis Code

```python
import pandas as pd

# Assume 'df_lag_ads' is your DataFrame

# Lagging the variables by 1 week starting from the second column
lag = 1
for col in df_lag_ads.columns[1:]:  # Start from the second column
    df_lag_ads[col + '_lag'] = df_lag_ads[col].shift(lag)

# Keeping the first column and dropping the original columns except the first one
df_lag_ads = df_lag_ads[[df_lag_ads.columns[0]] + [col for col in df_lag_ads.columns if col.endswith('_lag')]]
df_lag_ads = df_lag_ads.dropna().reset_index(drop=True)

# Implementing the adstock effect
window_adstock = 10
rate_adstock = 0.5

for var in df_lag_ads.columns[1:]:  # Apply adstock starting from the second column
    df_lag_ads[var + '_halo'] = df_lag_ads[var]
    for i in range(1, window_adstock + 1):
        df_lag_ads[var + '_halo'] += df_lag_ads[var].shift(i).fillna(0) * rate_adstock ** i

# Keeping the first column and dropping the intermediate columns after applying adstock
df_lag_ads = df_lag_ads[[df_lag_ads.columns[0]] + [col for col in df_lag_ads.columns if col.endswith('_halo')]]
df_lag_ads = df_lag_ads.dropna().reset_index(drop=True)

df_lag_ads
```

Figure 118: Lag and Adstock Code

```python
def calculate_mse(y_test, y_pred):
    """Calculate and return the mean squared error."""
    error = mean_squared_error(y_test, y_pred)
    return error

def calculate_residuals(y_test, y_pred):
    """Calculate and return the residuals."""
    residuals = y_test - y_pred
    return residuals

def plot_qq(residuals):
    """Plot the Q-Q plot of the residuals."""
    stats.probplot(residuals, dist="norm", plot=plt)
    plt.title("Normal Q-Q Plot")
    plt.show()

def plot_actual_vs_predicted_scatter(y_test, y_pred):
    """Plot the actual vs predicted values."""
    plt.scatter(y_test, y_pred)
    plt.plot([y_test.min(), y_test.max()], [y_test.min(), y_test.max()], color='red', linestyle='--', lw=2)
    plt.xlabel("Actual")
    plt.ylabel("Predicted")
    plt.title("Actual vs Predicted")
    plt.show()

def plot_residuals_histogram(residuals):
    """Plot the residuals histogram."""
    sns.histplot(residuals, kde=True)
    plt.title("Residuals Histogram")
    plt.show()
```

Figure 119: Residual Plot Codes 1

```
def plot_actual_vs_predicted(y_test, y_pred):
    """Plot the actual vs predicted values."""
    plt.figure(figsize=(13, 6))
    plt.scatter(range(len(y_test)), y_test, color='blue', label='Actual')
    plt.scatter(range(len(y_pred)), y_pred, color='red', label='Predicted')
    plt.title("Actual vs Predicted")
    plt.xlabel("Index of Samples")
    plt.ylabel("Values")
    plt.legend()
    plt.show()

    plt.figure(figsize=(13, 6))
    plt.plot(range(len(y_test)), y_test, color='blue', label='Actual')
    plt.plot(range(len(y_pred)), y_pred, color='red', label='Predicted')
    plt.title("Actual vs Predicted")
    plt.xlabel("Index of Samples")
    plt.ylabel("Values")
    plt.legend()
    plt.show()

    plt.figure(figsize=(13, 6))
    plt.plot(range(len(y_test)), y_test, color='blue')
    plt.plot(range(len(y_pred)), y_pred, color='red')
    plt.scatter(range(len(y_test)), y_test, color='blue', label='Actual')
    plt.scatter(range(len(y_pred)), y_pred, color='red', label='Predicted')
    plt.title("Actual vs Predicted")
    plt.xlabel("Index of Samples")
    plt.ylabel("Values")
    plt.legend()
    plt.show()
```

Figure 120: Residual Plot Codes 2

```
def plot_residuals_scatter(y_test, residuals):
    """Plot the residuals scatter plot."""
    plt.scatter(y_test, residuals, color='blue')
    plt.plot([y_test.min(), y_test.max()], [0, 0], color='red', linestyle='--', lw=2)
    plt.title("Residuals Scatter Plot")
    plt.xlabel("Actual")
    plt.ylabel("Residuals")
    plt.show()

def analysis_of_residuals(y_test, y_pred):
    """Perform analysis of residuals."""
    # Calculate residuals
    residuals = y_test - y_pred

    # Calculate mean squared error
    mse = mean_squared_error(y_test, y_pred)
    print(f"Mean Squared Error: {mse}")

    # Plot Q-Q plot
    print("\nQ-Q Plot:")
    plot_qq(residuals)

    # Plot actual vs predicted
    print("\nActual vs Predicted:")
    plot_actual_vs_predicted(y_test, y_pred)

   # print("\nActual vs Predicted Scatter:")
   # plot_actual_vs_predicted_scatter(y_test, y_pred)

    # Plot residuals scatter plot
    print("\nResiduals Scatter Plot:")
    plot_residuals_scatter(y_test, residuals)

    # Plot residuals histogram
    print("\nResiduals Histogram:")
    plot_residuals_histogram(residuals)
```

Figure 121: Residual Plot Codes 3

```python
# Define the regression function
def perform_regression(X_train: pd.DataFrame, X_test: pd.DataFrame, y_train: pd.Series,
                       model_type: str = 'OLS', simple: bool = True, alpha: float = 1.0,
                       tune_hyperparameters: bool = False, cv: int = 5, feature_selection: bool = True):

    model = None
    y_pred = None
    y_train_pred = None

    # Standardize features if needed
    if model_type in ['BayesianRidge', 'Lasso', 'Ridge']:
        scaler = StandardScaler()
        X_train_scaled = scaler.fit_transform(X_train)
        X_test_scaled = scaler.transform(X_test)
        X_train = pd.DataFrame(X_train_scaled, index=X_train.index, columns=X_train.columns)
        X_test = pd.DataFrame(X_test_scaled, index=X_test.index, columns=X_test.columns)
```

Figure 122: Standartization Dataset G

```python
if model_type == 'OLS':
    # Add a constant to the features for OLS regression
    X_train_const = sm.add_constant(X_train)
    X_test_const = sm.add_constant(X_test)

    # Initialize and fit the model
    model = sm.OLS(y_train, X_train_const)
    results = model.fit()
    y_pred = results.predict(X_test_const)
    y_train_pred = results.predict(X_train_const)

    # Print the summary
    if simple:
        print(results.summary())
    else:
        # Get the p-values of the features
        p_values = results.pvalues

        # Get the features with p-value less than 0.05
        significant_features = p_values[p_values < 0.05].index

        # Perform regression again with the significant features
        X_train_significant = X_train_const[significant_features]
        X_test_significant = X_test_const[significant_features]

        # Initialize and fit the model
        model = sm.OLS(y_train, X_train_significant)
        results = model.fit()

        # Print the summary
        print(results.summary())

        # Make predictions
        y_pred = results.predict(X_test_significant)
        y_train_pred = results.predict(X_train_significant)
```

Figure 123: OLS Dataset G

```
elif model_type == 'RandomForest':
    model = RandomForestRegressor(n_estimators=1000, random_state=42)
    if tune_hyperparameters:
        param_grid = {
            'n_estimators': [100, 500, 1000],
            'max_features': ['auto', 'sqrt', 'log2'],
            'max_depth': [None, 10, 20, 30],
            'min_samples_split': [2, 5, 10],
            'min_samples_leaf': [1, 2, 4],
        }
        grid = RandomizedSearchCV(model, param_grid, cv=cv, n_iter=10, random_state=42)
        grid.fit(X_train, y_train)
        model = grid.best_estimator_
    else:
        model.fit(X_train, y_train)

    y_pred = model.predict(X_test)
    y_train_pred = model.predict(X_train)

    # Feature importance
    importance = model.feature_importances_
    sorted_idx = np.argsort(importance)
    plt.barh(X_train.columns[sorted_idx], importance[sorted_idx])
    plt.xlabel("Random Forest Feature Importance")
    plt.show()
```

Figure 124: RandomForest Dataset G

```
elif model_type == 'GBM':
    model = GradientBoostingRegressor()
    if tune_hyperparameters:
        param_grid = {
            'n_estimators': [100, 500, 1000],
            'learning_rate': [0.01, 0.1, 0.05],
            'max_depth': [3, 5, 7],
        }
        grid = GridSearchCV(model, param_grid, cv=cv)
        grid.fit(X_train, y_train)
        model = grid.best_estimator_
    else:
        model.fit(X_train, y_train)

    y_pred = model.predict(X_test)
    y_train_pred = model.predict(X_train)

    # Feature importance
    importance = model.feature_importances_
    sorted_idx = np.argsort(importance)
    plt.barh(X_train.columns[sorted_idx], importance[sorted_idx])
    plt.xlabel("GBM Feature Importance")
    plt.show()
```

Figure 125: GradientBoostingRegressor Dataset G

```
elif model_type == 'XGBoost':
    model = XGBRegressor()
    if tune_hyperparameters:
        param_grid = {
            'n_estimators': [100, 500, 1000],
            'learning_rate': [0.01, 0.1, 0.2],
            'max_depth': [3, 5, 7],
        }
        grid = RandomizedSearchCV(model, param_grid, cv=cv, n_iter=10, random_state=42)
        grid.fit(X_train, y_train)
        model = grid.best_estimator_
    else:
        model.fit(X_train, y_train)

    y_pred = model.predict(X_test)
    y_train_pred = model.predict(X_train)

    # Feature importance with SHAP
    explainer = shap.Explainer(model, X_train)
    shap_values = explainer(X_train)
    shap.summary_plot(shap_values, X_train, plot_type="bar")
```

Figure 126: XGBoost Dataset G

```
elif model_type == 'Lasso':
    lasso = Lasso(alpha=alpha, max_iter=10000)
    if tune_hyperparameters:
        param_grid = {'alpha': np.logspace(-5, 5, 100)}
        grid = GridSearchCV(lasso, param_grid, cv=cv, scoring='r2')
        grid.fit(X_train, y_train)
        lasso = grid.best_estimator_
    else:
        lasso.fit(X_train, y_train)

    y_pred = lasso.predict(X_test)
    y_train_pred = lasso.predict(X_train)

    # Select features with non-zero coefficients
    selected_features = X_train.columns[lasso.coef_ != 0]
    print("Selected features:", selected_features)

    # Use OLS for interpretability
    X_train_selected = sm.add_constant(X_train[selected_features])
    X_test_selected = sm.add_constant(X_test[selected_features])

    ols_model = sm.OLS(y_train, X_train_selected)
    ols_results = ols_model.fit()
    print(ols_results.summary())
    y_pred = ols_results.predict(X_test_selected)
    y_train_pred = ols_results.predict(X_train_selected)
```

Figure 127: Lasso Dataset G

```python
    elif model_type == 'BayesianRidge':
        model = BayesianRidge()
        if tune_hyperparameters:
            param_grid = {'alpha_1': [1e-6, 1e-4, 1e-2, 1e0], 'alpha_2': [1e-6, 1e-4, 1e-2, 1e0]}
            grid = GridSearchCV(model, param_grid, cv=cv)
            grid.fit(X_train, y_train)
            model = grid.best_estimator_
        else:
            model.fit(X_train, y_train)
        y_pred = model.predict(X_test)
        y_train_pred = model.predict(X_train)

        # Coefficients interpretation
        coef = model.coef_
        selected_features = X_train.columns[np.abs(coef) > 1e-5]
        X_train_selected = sm.add_constant(X_train[selected_features])
        results = sm.OLS(y_train, X_train_selected).fit()
        print(results.summary())
```

Figure 128: BayesianRidge Dataset G

```python
elif model_type == 'Ridge':
    ridge = Ridge(alpha=alpha)
    if tune_hyperparameters:
        param_grid = {'alpha': np.logspace(-5, 5, 100)}
        grid = GridSearchCV(ridge, param_grid, cv=cv, scoring='r2')
        grid.fit(X_train, y_train)
        ridge = grid.best_estimator_
    else:
        ridge.fit(X_train, y_train)

    y_pred = ridge.predict(X_test)
    y_train_pred = ridge.predict(X_train)

    # Select features with coefficients above a certain threshold
    threshold = 1e-5  # Define a threshold for significance
    selected_features = X_train.columns[np.abs(ridge.coef_) > threshold]
    print("Selected features for Ridge:", selected_features)

    # Use OLS for interpretability
    X_train_selected = sm.add_constant(X_train[selected_features])
    X_test_selected = sm.add_constant(X_test[selected_features])

    ols_model = sm.OLS(y_train, X_train_selected)
    ols_results = ols_model.fit()
    print(ols_results.summary())
    y_pred = ols_results.predict(X_test_selected)
    y_train_pred = ols_results.predict(X_train_selected)

else:
    raise ValueError(f"Unsupported model type: {model_type}")

# Cross-Validation (Optional)
if model_type not in ['OLS'] and not simple:
    cv_scores = cross_val_score(model, X_train, y_train, cv=cv, scoring='r2')
    print(f"Cross-Validation R2 Score: {np.mean(cv_scores)} ± {np.std(cv_scores)}")

# Calculate and print R2 scores
r2_train = r2_score(y_train, y_train_pred)
r2_test = r2_score(y_test, y_pred)
print(f"R2 Score (Train): {r2_train}")
print(f"R2 Score (Test): {r2_test}")

return model, y_pred
```

Figure 129: Ridge Dataset G

```python
def perform_regression(X_train: pd.DataFrame, X_test: pd.DataFrame, y_train: pd.Series,
                       model_type: str = 'BayesianRidge', alpha: float = 1.0,
                       tune_hyperparameters: bool = False, cv: int = 5,
                       include_interactions: bool = False, degree: int = 2):
    """
    Perform regression with the specified model type, optionally perform hyperparameter tuning,
    and return the final model, predictions, R2 scores, and feature importance.
    """

    model = None
    y_pred = None
    y_train_pred = None
    feature_importance = None
    original_feature_importance = None
    interaction_feature_importance = None

    if include_interactions:
        # Initialize the PolynomialFeatures transformer for interaction terms only
        poly = PolynomialFeatures(degree=degree, interaction_only=True, include_bias=False)

        # Generate interaction features
        X_train_interactions = pd.DataFrame(
            poly.fit_transform(X_train),
            columns=poly.get_feature_names_out(X_train.columns),
            index=X_train.index
        )
        X_test_interactions = pd.DataFrame(
            poly.transform(X_test),
            columns=poly.get_feature_names_out(X_test.columns),
            index=X_test.index
        )

        # Separate original features and interaction features
        original_feature_names = X_train.columns
        interaction_feature_names = [name for name in X_train_interactions.columns if name not in original_feature_names]

        X_train_combined = pd.concat([X_train, X_train_interactions[interaction_feature_names]], axis=1)
        X_test_combined = pd.concat([X_test, X_test_interactions[interaction_feature_names]], axis=1)

        X_train = X_train_combined
        X_test = X_test_combined
```

Figure 130: Feature Importance and Interaction model Dataset G

```
# Standardize features if needed
if model_type in ['BayesianRidge', 'Lasso', 'Ridge', 'SVR', 'AdaBoost']:
    scaler = StandardScaler()
    X_train_scaled = scaler.fit_transform(X_train)
    X_test_scaled = scaler.transform(X_test)
    X_train = pd.DataFrame(X_train_scaled, index=X_train.index, columns=X_train.columns)
    X_test = pd.DataFrame(X_test_scaled, index=X_test.index, columns=X_test.columns)

elif model_type == 'BayesianRidge':
    model = BayesianRidge()
    if tune_hyperparameters:
        param_grid = {'alpha_1': [1e-6, 1e-4, 1e-2, 1e0], 'alpha_2': [1e-6, 1e-4, 1e-2, 1e0]}
        grid = GridSearchCV(model, param_grid, cv=cv)
        grid.fit(X_train, y_train)
        model = grid.best_estimator_
    else:
        model.fit(X_train, y_train)
    y_pred = model.predict(X_test)
    y_train_pred = model.predict(X_train)
    feature_importance = np.abs(model.coef_)
else:
    raise ValueError(f"Unsupported model type: {model_type}")

# Calculate and print R2 scores
r2_train = r2_score(y_train, y_train_pred)
r2_test = r2_score(y_test, y_pred)
print(f"R2 Score (Train): {r2_train}")
print(f"R2 Score (Test): {r2_test}")

# Separate feature importances into original and interaction features
if include_interactions:
    original_feature_importance = feature_importance[:len(original_feature_names)]
    interaction_feature_importance = feature_importance[len(original_feature_names):]

    # Plot feature importance for original features
    plot_feature_importance(original_feature_importance, original_feature_names, "Feature Importance for Original Features")

    # Plot feature importance for interaction features
    plot_feature_importance(interaction_feature_importance, interaction_feature_names, "Feature Importance for Interaction Features")
else:
    # Plot all features together if no interactions are used
    plot_feature_importance(feature_importance, X_train.columns, f"Feature Importance for {model_type}")

return model, y_pred, r2_train, r2_test, feature_importance
```

Figure 131: BayesianRidge Model for Interaction and Importance Dataset G

```python
def perform_regression(X_train: pd.DataFrame, X_test: pd.DataFrame, y_train: pd.Series,
                       model_type: str = 'OLS', simple: bool = True, alpha: float = 1.0,
                       tune_hyperparameters: bool = False, cv: int = 5, feature_selection: bool = True):
    model = None
    y_pred = None
    y_train_pred = None

    if model_type in ['BayesianRidge', 'Lasso', 'Ridge']:
        scaler = StandardScaler()
        X_train_scaled = scaler.fit_transform(X_train)
        X_test_scaled = scaler.transform(X_test)
        X_train = pd.DataFrame(X_train_scaled, index=X_train.index, columns=X_train.columns)
        X_test = pd.DataFrame(X_test_scaled, index=X_test.index, columns=X_test.columns)
    if model_type == 'OLS':
        # Add a constant to the features for OLS regression
        X_train_const = sm.add_constant(X_train)
        X_test_const = sm.add_constant(X_test)

        # Initialize and fit the model
        model = sm.OLS(y_train, X_train_const)
        results = model.fit()
        y_pred = results.predict(X_test_const)
        y_train_pred = results.predict(X_train_const)
        # Print the summary
        if simple:
            print(results.summary())
        else:
            # Get the p-values of the features
            p_values = results.pvalues
            # Get the features with p-value less than 0.05
            significant_features = p_values[p_values < 0.05].index
            # Perform regression again with the significant features
            X_train_significant = X_train_const[significant_features]
            X_test_significant = X_test_const[significant_features]
            # Initialize and fit the model
            model = sm.OLS(y_train, X_train_significant)
            results = model.fit()
            # Print the summary
            print(results.summary())
            # Make predictions
            y_pred = results.predict(X_test_significant)
            y_train_pred = results.predict(X_train_significant)
```

Figure 132: Model for finding contributions Dataset D sales 1

```python
elif model_type == 'BayesianRidge':
    model = BayesianRidge()
    if tune_hyperparameters:
        param_grid = {'alpha_1': [1e-6, 1e-4, 1e-2, 1e0], 'alpha_2': [1e-6, 1e-4, 1e-2, 1e0]}
        grid = GridSearchCV(model, param_grid, cv=cv)
        grid.fit(X_train, y_train)
        model = grid.best_estimator_
    else:
        model.fit(X_train, y_train)
    y_pred = model.predict(X_test)
    y_train_pred = model.predict(X_train)

    # Coefficients interpretation
    coef = model.coef_
    selected_features = X_train.columns[np.abs(coef) > 1e-5]
    X_train_selected = sm.add_constant(X_train[selected_features])
    results = sm.OLS(y_train, X_train_selected).fit()
    print(results.summary())
```

Figure 133: Model for finding contributions Dataset D sales 1

```
elif model_type == 'RandomForest':
    model = RandomForestRegressor(n_estimators=1000, random_state=42)
    if tune_hyperparameters:
        param_grid = {
            'n_estimators': [100, 500, 1000],
            'max_features': ['auto', 'sqrt', 'log2'],
            'max_depth': [None, 10, 20, 30],
            'min_samples_split': [2, 5, 10],
            'min_samples_leaf': [1, 2, 4],
        }
        grid = RandomizedSearchCV(model, param_grid, cv=cv, n_iter=10, random_state=42)
        grid.fit(X_train, y_train)
        model = grid.best_estimator_
    else:
        model.fit(X_train, y_train)

    y_pred = model.predict(X_test)
    y_train_pred = model.predict(X_train)

    # Feature importance
    importance = model.feature_importances_
    sorted_idx = np.argsort(importance)
    plt.barh(X_train.columns[sorted_idx], importance[sorted_idx])
    plt.xlabel("Random Forest Feature Importance")
    plt.show()
```

Figure 134: Model for finding contributions Dataset D sales 1

```
elif model_type == 'GBM':
    model = GradientBoostingRegressor()
    if tune_hyperparameters:
        param_grid = {
            'n_estimators': [100, 500, 1000],
            'learning_rate': [0.01, 0.1, 0.05],
            'max_depth': [3, 5, 7],
        }
        grid = GridSearchCV(model, param_grid, cv=cv)
        grid.fit(X_train, y_train)
        model = grid.best_estimator_
    else:
        model.fit(X_train, y_train)

    y_pred = model.predict(X_test)
    y_train_pred = model.predict(X_train)

    # Feature importance
    importance = model.feature_importances_
    sorted_idx = np.argsort(importance)
    plt.barh(X_train.columns[sorted_idx], importance[sorted_idx])
    plt.xlabel("GBM Feature Importance")
    plt.show()
```

Figure 135: Model for finding contributions Dataset D sales 1

```python
# Define the regression function
def perform_regression(X_train: pd.DataFrame, X_test: pd.DataFrame, y_train: pd.Series,
                       model_type: str = 'OLS', simple: bool = True, alpha: float = 1.0):
    model = None
    y_pred = None
    y_train_pred = None

    # Scaling for certain models
    if model_type in ['BayesianRidge', 'SVR', 'AdaBoost']:
        scaler = StandardScaler()
        X_train_scaled = scaler.fit_transform(X_train)
        X_test_scaled = scaler.transform(X_test)
        X_train = pd.DataFrame(X_train_scaled, index=X_train.index, columns=X_train.columns)
        X_test = pd.DataFrame(X_test_scaled, index=X_test.index, columns=X_test.columns)
    if model_type == 'OLS':
        # Add a constant to the features for OLS regression
        X_train_const = sm.add_constant(X_train)
        X_test_const = sm.add_constant(X_test)

        # Initialize and fit the model
        model = sm.OLS(y_train, X_train_const)
        results = model.fit()
        y_pred = results.predict(X_test_const)
        y_train_pred = results.predict(X_train_const)
    elif model_type == 'BayesianRidge':
        model = BayesianRidge()
        model.fit(X_train, y_train)
        y_pred = model.predict(X_test)
        y_train_pred = model.predict(X_train)
    elif model_type == 'RandomForest':
        model = RandomForestRegressor(n_estimators=1000, random_state=42)
        model.fit(X_train, y_train)
        y_pred = model.predict(X_test)
        y_train_pred = model.predict(X_train)
    elif model_type == 'GBM':
        model = GradientBoostingRegressor()
        model.fit(X_train, y_train)
        y_pred = model.predict(X_test)
        y_train_pred = model.predict(X_train)
```

Figure 136: Model for finding feature contributions Dataset D sales 1

```python
elif model_type == 'XGBoost':
    model = XGBRegressor()
    model.fit(X_train, y_train)
    y_pred = model.predict(X_test)
    y_train_pred = model.predict(X_train)
elif model_type == 'Lasso':
    # Use fixed alpha value for Lasso and increase max_iter
    lasso = Lasso(alpha=alpha, max_iter=10000)
    lasso.fit(X_train, y_train)
    y_pred = lasso.predict(X_test)
    y_train_pred = lasso.predict(X_train)
    # Select features with non-zero coefficients
    selected_features = X_train.columns[lasso.coef_ != 0]
    print("Selected features:", selected_features)
elif model_type == 'Ridge':
    # Use fixed alpha value for Ridge
    ridge = Ridge(alpha=alpha)
    ridge.fit(X_train, y_train)
    y_pred = ridge.predict(X_test)
    y_train_pred = ridge.predict(X_train)
    # Select features with coefficients above a certain threshold
    threshold = 1e-5  # Define a threshold for significance
    selected_features = X_train.columns[np.abs(ridge.coef_) > threshold]
    print("Selected features for Ridge:", selected_features)
elif model_type == 'AdaBoost':
    model = AdaBoostRegressor(n_estimators=1000, random_state=42)
    model.fit(X_train, y_train)
    y_pred = model.predict(X_test)
    y_train_pred = model.predict(X_train)
elif model_type == 'SVR':
    model = SVR(kernel='linear')
    model.fit(X_train, y_train)
    y_pred = model.predict(X_test)
    y_train_pred = model.predict(X_train)

else:
    raise ValueError(f"Unsupported model type: {model_type}")
r2_train = r2_score(y_train, y_train_pred)
r2_test = r2_score(y_test, y_pred)
print(f"R2 Score (Train): {r2_train}")
print(f"R2 Score (Test): {r2_test}")
return model, y_pred
```

Figure 137: Model for finding feature contributions Dataset D sales 1

```python
def perform_regression(X_train: pd.DataFrame, X_test: pd.DataFrame, y_train: pd.Series,
                       model_type: str = 'OLS', simple: bool = True, alpha: float = 1.0,
                       tune_hyperparameters: bool = False, cv: int = 5, feature_selection: bool = True):
    model = None
    y_pred = None
    y_train_pred = None
    # Standardize features if needed
    if model_type in ['BayesianRidge', 'Lasso', 'Ridge']:
        scaler = StandardScaler()
        X_train_scaled = scaler.fit_transform(X_train)
        X_test_scaled = scaler.transform(X_test)
        X_train = pd.DataFrame(X_train_scaled, index=X_train.index, columns=X_train.columns)
        X_test = pd.DataFrame(X_test_scaled, index=X_test.index, columns=X_test.columns)
    if model_type == 'OLS':
        # Add a constant to the features for OLS regression
        X_train_const = sm.add_constant(X_train)
        X_test_const = sm.add_constant(X_test)
        # Initialize and fit the model
        model = sm.OLS(y_train, X_train_const)
        results = model.fit()
        y_pred = results.predict(X_test_const)
        y_train_pred = results.predict(X_train_const)
        # Print the summary
        if simple:
            print(results.summary())
        else:
            # Get the p-values of the features
            p_values = results.pvalues

            # Get the features with p-value less than 0.05
            significant_features = p_values[p_values < 0.05].index
            # Perform regression again with the significant features
            X_train_significant = X_train_const[significant_features]
            X_test_significant = X_test_const[significant_features]
            # Initialize and fit the model
            model = sm.OLS(y_train, X_train_significant)
            results = model.fit()
            # Print the summary
            print(results.summary())
            # Make predictions
            y_pred = results.predict(X_test_significant)
            y_train_pred = results.predict(X_train_significant)
```

Figure 138: Model for finding contributions Dataset D Sales 2

```python
elif model_type == 'BayesianRidge':
    model = BayesianRidge()
    if tune_hyperparameters:
        param_grid = {'alpha_1': [1e-6, 1e-4, 1e-2, 1e0], 'alpha_2': [1e-6, 1e-4, 1e-2, 1e0]}
        grid = GridSearchCV(model, param_grid, cv=cv)
        grid.fit(X_train, y_train)
        model = grid.best_estimator_
    else:
        model.fit(X_train, y_train)
    y_pred = model.predict(X_test)
    y_train_pred = model.predict(X_train)
    # Coefficients interpretation
    coef = model.coef_
    selected_features = X_train.columns[np.abs(coef) > 1e-5]
    X_train_selected = sm.add_constant(X_train[selected_features])
    results = sm.OLS(y_train, X_train_selected).fit()
    print(results.summary())
elif model_type == 'RandomForest':
    model = RandomForestRegressor(n_estimators=1000, random_state=42)
    if tune_hyperparameters:
        param_grid = {
            'n_estimators': [100, 500, 1000],
            'max_features': ['auto', 'sqrt', 'log2'],
            'max_depth': [None, 10, 20, 30],
            'min_samples_split': [2, 5, 10],
            'min_samples_leaf': [1, 2, 4],
        }
        grid = RandomizedSearchCV(model, param_grid, cv=cv, n_iter=10, random_state=42)
        grid.fit(X_train, y_train)
        model = grid.best_estimator_
    else:
        model.fit(X_train, y_train)

    y_pred = model.predict(X_test)
    y_train_pred = model.predict(X_train)
    # Feature importance
    importance = model.feature_importances_
    sorted_idx = np.argsort(importance)
    plt.barh(X_train.columns[sorted_idx], importance[sorted_idx])
    plt.xlabel("Random Forest Feature Importance")
    plt.show()
```

Figure 139: Model for finding contributions Dataset D Sales 2

```python
elif model_type == 'GBM':
    model = GradientBoostingRegressor()
    if tune_hyperparameters:
        param_grid = {
            'n_estimators': [100, 500, 1000],
            'learning_rate': [0.01, 0.1, 0.05],
            'max_depth': [3, 5, 7],
        }
        grid = GridSearchCV(model, param_grid, cv=cv)
        grid.fit(X_train, y_train)
        model = grid.best_estimator_
    else:
        model.fit(X_train, y_train)

    y_pred = model.predict(X_test)
    y_train_pred = model.predict(X_train)

    # Feature importance
    importance = model.feature_importances_
    sorted_idx = np.argsort(importance)
    plt.barh(X_train.columns[sorted_idx], importance[sorted_idx])
    plt.xlabel("GBM Feature Importance")
    plt.show()
```

Figure 140: Model for finding contributions Dataset D Sales 2

```
elif model_type == 'XGBoost':
    model = XGBRegressor()
    if tune_hyperparameters:
        param_grid = {
            'n_estimators': [100, 500, 1000],
            'learning_rate': [0.01, 0.1, 0.2],
            'max_depth': [3, 5, 7],
        }
        grid = RandomizedSearchCV(model, param_grid, cv=cv, n_iter=10, random_state=42)
        grid.fit(X_train, y_train)
        model = grid.best_estimator_
    else:
        model.fit(X_train, y_train)

    y_pred = model.predict(X_test)
    y_train_pred = model.predict(X_train)

    # Feature importance with SHAP
    explainer = shap.Explainer(model, X_train)
    shap_values = explainer(X_train)
    shap.summary_plot(shap_values, X_train, plot_type="bar")
```

Figure 141: Model for finding contributions Dataset D Sales 2

```
elif model_type == 'Lasso':
    lasso = Lasso(alpha=alpha, max_iter=10000)
    if tune_hyperparameters:
        param_grid = {'alpha': np.logspace(-5, 5, 100)}
        grid = GridSearchCV(lasso, param_grid, cv=cv, scoring='r2')
        grid.fit(X_train, y_train)
        lasso = grid.best_estimator_
    else:
        lasso.fit(X_train, y_train)

    y_pred = lasso.predict(X_test)
    y_train_pred = lasso.predict(X_train)

    # Select features with non-zero coefficients
    selected_features = X_train.columns[lasso.coef_ != 0]
    print("Selected features:", selected_features)

    # Use OLS for interpretability
    X_train_selected = sm.add_constant(X_train[selected_features])
    X_test_selected = sm.add_constant(X_test[selected_features])

    ols_model = sm.OLS(y_train, X_train_selected)
    ols_results = ols_model.fit()
    print(ols_results.summary())
    y_pred = ols_results.predict(X_test_selected)
    y_train_pred = ols_results.predict(X_train_selected)
```

Figure 142: Model for finding contributions Dataset D Sales 2

```
elif model_type == 'Ridge':
    ridge = Ridge(alpha=alpha)
    if tune_hyperparameters:
        param_grid = {'alpha': np.logspace(-5, 5, 100)}
        grid = GridSearchCV(ridge, param_grid, cv=cv, scoring='r2')
        grid.fit(X_train, y_train)
        ridge = grid.best_estimator_
    else:
        ridge.fit(X_train, y_train)

    y_pred = ridge.predict(X_test)
    y_train_pred = ridge.predict(X_train)

    # Select features with coefficients above a certain threshold
    threshold = 1e-5  # Define a threshold for significance
    selected_features = X_train.columns[np.abs(ridge.coef_) > threshold]
    print("Selected features for Ridge:", selected_features)

    # Use OLS for interpretability
    X_train_selected = sm.add_constant(X_train[selected_features])
    X_test_selected = sm.add_constant(X_test[selected_features])

    ols_model = sm.OLS(y_train, X_train_selected)
    ols_results = ols_model.fit()
    print(ols_results.summary())
    y_pred = ols_results.predict(X_test_selected)
    y_train_pred = ols_results.predict(X_train_selected)

else:
    raise ValueError(f"Unsupported model type: {model_type}")

# Cross-Validation (Optional)
if model_type not in ['OLS'] and not simple:
    cv_scores = cross_val_score(model, X_train, y_train, cv=cv, scoring='r2')
    print(f"Cross-Validation R2 Score: {np.mean(cv_scores)} ± {np.std(cv_scores)}")

# Calculate and print R2 scores
r2_train = r2_score(y_train, y_train_pred)
r2_test = r2_score(y_test, y_pred)
print(f"R2 Score (Train): {r2_train}")
print(f"R2 Score (Test): {r2_test}")

return model, y_pred
```

Figure 143: Model for finding contributions Dataset D Sales 2

```
# Define the regression function
def perform_regression(X_train: pd.DataFrame, X_test: pd.DataFrame, y_train: pd.Series,
                       model_type: str = 'OLS', simple: bool = True, alpha: float = 1.0,
                       tune_hyperparameters: bool = False, cv: int = 5, feature_selection: bool = True):
    model = None
    y_pred = None
    y_train_pred = None
    # Standardize features if needed
    if model_type in ['BayesianRidge', 'Lasso', 'Ridge']:
        scaler = StandardScaler()
        X_train_scaled = scaler.fit_transform(X_train)
        X_test_scaled = scaler.transform(X_test)
        X_train = pd.DataFrame(X_train_scaled, index=X_train.index, columns=X_train.columns)
        X_test = pd.DataFrame(X_test_scaled, index=X_test.index, columns=X_test.columns)
```

Figure 144: Model for finding Contribution Dataset D Sales 3

```python
if model_type == 'OLS':
    # Add a constant to the features for OLS regression
    X_train_const = sm.add_constant(X_train)
    X_test_const = sm.add_constant(X_test)
    # Initialize and fit the model
    model = sm.OLS(y_train, X_train_const)
    results = model.fit()
    y_pred = results.predict(X_test_const)
    y_train_pred = results.predict(X_train_const)
    # Print the summary
    if simple:
        print(results.summary())
    else:
        # Get the p-values of the features
        p_values = results.pvalues
        # Get the features with p-value less than 0.05
        significant_features = p_values[p_values < 0.05].index
        # Perform regression again with the significant features
        X_train_significant = X_train_const[significant_features]
        X_test_significant = X_test_const[significant_features]
        # Initialize and fit the model
        model = sm.OLS(y_train, X_train_significant)
        results = model.fit()
         # Print the summary
        print(results.summary())

        # Make predictions
        y_pred = results.predict(X_test_significant)
        y_train_pred = results.predict(X_train_significant)
```

Figure 145: Model for finding Contribution Dataset D Sales 3

```python
elif model_type == 'BayesianRidge':
    model = BayesianRidge()
    if tune_hyperparameters:
        param_grid = {'alpha_1': [1e-6, 1e-4, 1e-2, 1e0], 'alpha_2': [1e-6, 1e-4, 1e-2, 1e0]}
        grid = GridSearchCV(model, param_grid, cv=cv)
        grid.fit(X_train, y_train)
        model = grid.best_estimator_
    else:
        model.fit(X_train, y_train)
    y_pred = model.predict(X_test)
    y_train_pred = model.predict(X_train)

    # Coefficients interpretation
    coef = model.coef_
    selected_features = X_train.columns[np.abs(coef) > 1e-5]
    X_train_selected = sm.add_constant(X_train[selected_features])
    results = sm.OLS(y_train, X_train_selected).fit()
    print(results.summary())

elif model_type == 'RandomForest':
    model = RandomForestRegressor(n_estimators=1000, random_state=42)
    if tune_hyperparameters:
        param_grid = {
            'n_estimators': [100, 500, 1000],
            'max_features': ['auto', 'sqrt', 'log2'],
            'max_depth': [None, 10, 20, 30],
            'min_samples_split': [2, 5, 10],
            'min_samples_leaf': [1, 2, 4],
        }
        grid = RandomizedSearchCV(model, param_grid, cv=cv, n_iter=10, random_state=42)
        grid.fit(X_train, y_train)
        model = grid.best_estimator_
    else:
        model.fit(X_train, y_train)

    y_pred = model.predict(X_test)
    y_train_pred = model.predict(X_train)
    # Feature importance
    importance = model.feature_importances_
    sorted_idx = np.argsort(importance)
    plt.barh(X_train.columns[sorted_idx], importance[sorted_idx])
    plt.xlabel("Random Forest Feature Importance")
    plt.show()
```

Figure 146: Model for finding Contribution Dataset D Sales 3

```python
# Define the regression function
def perform_regression(X_train: pd.DataFrame, X_test: pd.DataFrame, y_train: pd.Series,
                       model_type: str = 'OLS', simple: bool = True, alpha: float = 1.0):
    model = None
    y_pred = None
    y_train_pred = None

    # Scaling for certain models
    if model_type in ['BayesianRidge', 'SVR', 'AdaBoost']:
        scaler = StandardScaler()
        X_train_scaled = scaler.fit_transform(X_train)
        X_test_scaled = scaler.transform(X_test)
        X_train = pd.DataFrame(X_train_scaled, index=X_train.index, columns=X_train.columns)
        X_test = pd.DataFrame(X_test_scaled, index=X_test.index, columns=X_test.columns)

    if model_type == 'OLS':
        # Use HuberRegressor as a robust alternative to OLS
        model = HuberRegressor()
        model.fit(X_train, y_train)
        y_pred = model.predict(X_test)
        y_train_pred = model.predict(X_train)

    elif model_type == 'BayesianRidge':
        model = BayesianRidge()
        model.fit(X_train, y_train)
        y_pred = model.predict(X_test)
        y_train_pred = model.predict(X_train)

    elif model_type == 'RandomForest':
        model = RandomForestRegressor(n_estimators=1000, random_state=42, max_features='sqrt')
        model.fit(X_train, y_train)
        y_pred = model.predict(X_test)
        y_train_pred = model.predict(X_train)

    elif model_type == 'GBM':
        # Use Huber loss for robustness in Gradient Boosting
        model = GradientBoostingRegressor(loss='huber', n_estimators=1000, random_state=42)
        model.fit(X_train, y_train)
        y_pred = model.predict(X_test)
        y_train_pred = model.predict(X_train)
```

Figure 147: Model for finding feature contribution Dataset D Sales 3

```python
elif model_type == 'XGBoost':
    model = XGBRegressor()
    model.fit(X_train, y_train)
    y_pred = model.predict(X_test)
    y_train_pred = model.predict(X_train)

elif model_type == 'Lasso':
    # Use HuberRegressor as a robust alternative to Lasso
    model = HuberRegressor()
    model.fit(X_train, y_train)
    y_pred = model.predict(X_test)
    y_train_pred = model.predict(X_train)

elif model_type == 'Ridge':
    # Use HuberRegressor as a robust alternative to Ridge
    model = HuberRegressor()
    model.fit(X_train, y_train)
    y_pred = model.predict(X_test)
    y_train_pred = model.predict(X_train)

elif model_type == 'AdaBoost':
    model = AdaBoostRegressor(n_estimators=1000, random_state=42)
    model.fit(X_train, y_train)
    y_pred = model.predict(X_test)
    y_train_pred = model.predict(X_train)

elif model_type == 'SVR':
    # Replacing SVR with HuberRegressor for robustness
    model = HuberRegressor()
    model.fit(X_train, y_train)
    y_pred = model.predict(X_test)
    y_train_pred = model.predict(X_train)

else:
    raise ValueError(f"Unsupported model type: {model_type}")

# Calculate and print R2 scores
r2_train = r2_score(y_train, y_train_pred)
r2_test = r2_score(y_test, y_pred)
print(f"R2 Score : {r2_train}")

return model, y_pred
```

Figure 148: Model for finding feature contribution Dataset D Sales 3

```python
# Calculate SHAP values for the GBM model
explainer = shap.Explainer(model, X_train)
shap_values = explainer(X_test)

# Initialize lists for contributors and their SHAP values
contributors = X_train.columns.tolist()
shap_value = np.zeros(len(contributors))

# Compute the mean absolute SHAP value for each feature
for i in range(len(contributors)):
    shap_value[i] = np.abs(shap_values[:, i].values).mean()

# Compute the percentage contributions
perc_contributions = np.zeros(len(contributors))
for i in range(len(contributors)):
    perc_contributions[i] = shap_value[i] / shap_value.sum()
    print(f'The percentage contribution to the sales of {contributors[i]} is: {perc_contributions[i]}')

# Create a custom color palette from dark red to dark purple
colors = sns.color_palette("ch:3.5,-.2,dark=.3", n_colors=len(contributors))

# Plot the percentage contributions with the gradient colors
plt.figure(figsize=(10, 6))
sns.barplot(x=perc_contributions, y=contributors, palette=colors)
plt.title("Feature Contributions")
plt.xlabel("Percentage Contribution")
plt.ylabel("Features")
plt.show()
```

Figure 149: Feature Contribution Code Dataset D Sales 3