Consumer Behavior Analytics

Sai pranup Shobanaboina

Data Analytics, Webster University

Dr. Ali Ovlia

Table of Contents

[Executive Summary 4](#_Toc196758222)

[1. Project Introduction 4](#_Toc196758223)

[2. Business and Analytics Goals 5](#_Toc196758224)

[Business Problem 5](#_Toc196758225)

[Business Goal 5](#_Toc196758226)

[Business Objectives 5](#_Toc196758227)

[Analytics Approach 5](#_Toc196758228)

[3. Data Preprocessing 6](#_Toc196758229)

[Attributes Definition 6](#_Toc196758230)

[Data Exploration 7](#_Toc196758231)

[Summary Statistics for Numerical Variables 7](#_Toc196758232)

[Frequency Analysis of Categorical Variables 7](#_Toc196758233)

[Missing Values, Zero Values and NaNs Analysis 7](#_Toc196758234)

[Missing Value Imputation and Year\_Birth Correction 8](#_Toc196758235)

[Outlier Detection Using IQR Method 8](#_Toc196758236)

[Exploring Variable Relationships 9](#_Toc196758237)

[Correlation Heatmap 10](#_Toc196758238)

[Income vs. Total Spending 11](#_Toc196758239)

[Response Rate by Education Level 12](#_Toc196758240)

[Response Rate by Marital Status 13](#_Toc196758241)

[4. Predictor Analysis and Relevancy 14](#_Toc196758242)

[Multicollinearity Check for Numerical Variables 14](#_Toc196758243)

[Association Analysis for Binary Variables 14](#_Toc196758244)

[Relationship Between Numerical Variables and Response 14](#_Toc196758245)

[Key statistical findings 16](#_Toc196758246)

[5. Data Engineering and Transformation 17](#_Toc196758247)

[Log Transformation for Skewed Variables 17](#_Toc196758248)

[Ordinal Encoding of Education 18](#_Toc196758249)

[Dummy Encoding of Marital\_Status 18](#_Toc196758250)

[Winsorization of Income 18](#_Toc196758251)

[6. Dimension Reduction 19](#_Toc196758252)

[Redundancy and Relevance Filtering 19](#_Toc196758253)

[Avoidance of PCA 19](#_Toc196758254)

[Final Reduced Dataset 20](#_Toc196758255)

[7. Data Partitioning Methods 20](#_Toc196758256)

[For Classification 20](#_Toc196758257)

[For Regression 21](#_Toc196758258)

[8. Model Selection 21](#_Toc196758259)

[Segmentation - Unsupervised Learning 21](#_Toc196758260)

[Classification - Supervised Learning 21](#_Toc196758261)

[Regression - Supervised Learning 21](#_Toc196758262)

[9. Model fitting, validation accuracy and test accuracy 22](#_Toc196758263)

[Segmentation 22](#_Toc196758264)

[Classification Models 22](#_Toc196758265)

[Regression Models 22](#_Toc196758266)

[10. Report Models Performance 23](#_Toc196758267)

[Segmentation Performance – Silhouette Analysis 23](#_Toc196758268)

[Silhouette plot of K = 2 Clusters 23](#_Toc196758269)

[Silhouette plot of K = 3 Clusters 24](#_Toc196758270)

[Classification Models 25](#_Toc196758271)

[ROC Curves 25](#_Toc196758272)

[Regression Models 26](#_Toc196758273)

[Predictor Importance – Random Forest 26](#_Toc196758274)

[Random Forest Classification – Top Predictors: 27](#_Toc196758275)

[Random Forest Regression – Top Predictors: 28](#_Toc196758276)

[11. Model Evaluation 28](#_Toc196758277)

[Classification Models Evaluation 28](#_Toc196758278)

[Cost-Sensitive Learning Analysis 29](#_Toc196758279)

[Regression Models Evaluation 29](#_Toc196758280)

[Segmentation: K-Means Clustering Evaluation 30](#_Toc196758281)

[Interpretation of Cluster Profiles (K = 2) 30](#_Toc196758282)

[12. Observations and Conclusion 35](#_Toc196758283)

[Final Recommendations 35](#_Toc196758284)

[Conclusion and Future scope 36](#_Toc196758285)

[References 36](#_Toc196758286)

# Executive Summary

This analytics project was conducted to help Shop-Smart, a premium food and beverage retailer, enhance marketing **campaign effectiveness** by identifying which consumers are most likely to **respond** and spend more. Using historical transaction and demographic data, I have applied segmentation and predictive modeling techniques to better understand consumer behavior and guide personalized marketing. Key variables like Age, Income, and Total Spending were engineered, and data was cleaned and transformed to support high-performing models.

Based on the analysis, I recommend that Shop-Smart:

* **Target Cluster 2** - older, high-income, high-spending consumers who respond well to campaigns
* **Use logistic regression** to identify campaign responders with high accuracy and ROI
* Apply random forest regression to forecast spending and guide budget allocation
* Design **tiered promotional strategies** tailored to consumer segments and purchase channels (e.g., web, catalog)

These findings directly support Shop-Smart’s business goals of increasing engagement, improving ROI, and driving data-informed campaign decisions.

**Sai pranup Shobanaboina**

**Data Analyst | Shop-Smart**

**05/01/2025**

# Project Introduction

Understanding consumer behavior is essential for businesses aiming to personalize their marketing strategies and improve customer engagement. Shop-Smart, a retailer specializing in premium food and beverage products, has collected detailed consumer data including demographics, purchasing behavior, online interactions, and marketing campaign responses. This dataset offers an opportunity to uncover purchasing trends, segment customers based on their behavior and assess how various consumer groups respond to different types of marketing campaigns. The goal is to transform this raw consumer data into actionable insights that will guide targeted promotions and drive higher revenue.

This project applies both unsupervised and supervised modeling approaches to analyze consumer behavior at Shop-Smart. By identifying meaningful customer segments and modeling marketing responses, we aim to support strategic decisions such as which groups should be targeted with specific promotions. The study will also evaluate the impact of campaign efforts and determine factors influencing marketing engagement. Through this structured analytics approach, Shop-Smart will be better equipped to optimize its outreach strategy, improve conversion rates, and allocate marketing resources more effectively. Initial results from segmentation, classification, and regression models now provide early insights to inform these decisions.

# Business and Analytics Goals

## Business Problem

Shop-Smart lacks personalized targeting in its marketing campaigns, leading to low consumer engagement and poor ROI. Despite a large consumer database, the company struggles to identify which customers are most likely to respond to promotions.

## Business Goal

Leverage data analytics to improve campaign effectiveness and drive higher consumer engagement across marketing channels.

## Business Objectives

* Improve campaign response rates by analyzing consumer purchasing patterns and demographic traits over the next quarter.
* Increase marketing ROI by tailoring promotions to behavior-based consumer segments within the next two months.
* Identify consumers who are most likely to respond positively to future promotions, so marketing efforts can be better targeted.
* Forecast transaction-level marketing engagement using consumer and behavioral attributes to support quarterly revenue growth goals.

## Analytics Approach

The analytics process follows a structured methodology to extract actionable insights:

1. **Define the Problem**

how consumer-level data can support personalized marketing and targeted outreach?

1. **Collect Data**

Use the historical consumer database, including demographics, transaction history, marketing exposure, and campaign responses.

1. **Data Cleaning and Engineering**

Identify and handle missing values, standardize attributes, correct anomalies, transform skewed variables, engineer features and create a well-structured dataset for modeling.

1. **Data Analysis**

Explore purchasing trends, online activity, and campaign engagement using EDA and statistical summaries.

1. **Extract Knowledge**

* Unsupervised Learning: Apply clustering to segment consumers by behavior and preferences.
* Supervised Learning: Build predictive models to classify responders and estimate spending engagement.
* Interpretation: Derive actionable insights to guide segmentation and campaign personalization.

1. **Insights Presentation:**

Use visualizations and summaries to communicate consumer insights clearly and concisely.

1. **Decision-Making Support:**

Recommend data-driven targeting strategies to enhance future campaign performance.

# Data Preprocessing

## Attributes Definition

The dataset contains 2,240 observations and 29 variables, capturing consumer demographics, product purchase behavior, online activity, and marketing campaign responses. Each attribute is categorized as shown in the below Table 1, based on its data type and modeling relevance.

|  |  |  |  |
| --- | --- | --- | --- |
| ****Column Name**** | ****Description**** | ****Data Type**** | ****Category**** |
| ID | Unique consumer identifier | Categorical | Nominal |
| Year\_Birth | Birth year of the consumer | Temporal (int) | Temporal |
| Education | Consumer’s education level | Categorical | Ordinal |
| Marital\_Status | Marital status of the consumer | Categorical | Nominal |
| Income | Consumer’s yearly household income | Numerical (float) | Continuous |
| Kidhome | Number of children in the household | Numerical (int) | Discrete |
| Teenhome | Number of teenagers in the household | Numerical (int) | Discrete |
| Dt\_Consumer | Date of consumer’s enrollment with the company | Date | Date |
| Recency | Number of days since last purchase | Numerical (int) | Continuous |
| MntWines | Amount spent on wine | Numerical (int) | Continuous |
| MntFruits | Amount spent on fruits | Numerical (int) | Continuous |
| MntMeatProducts | Amount spent on meat products | Numerical (int) | Continuous |
| MntFishProducts | Amount spent on fish products | Numerical (int) | Continuous |
| MntSweetProducts | Amount spent on sweets | Numerical (int) | Continuous |
| MntGoldProds | Amount spent on gold products | Numerical (int) | Continuous |
| NumDealsPurchases | Number of purchases with a discount deal | Numerical (int) | Discrete |
| NumWebPurchases | Number of purchases through website | Numerical (int) | Discrete |
| NumCatalogPurchases | Number of catalog-based purchases | Numerical (int) | Discrete |
| NumStorePurchases | Number of in-store purchases | Numerical (int) | Discrete |
| NumWebVisitsMonth | Number of website visits in the last month | Numerical (int) | Discrete |
| AcceptedCmp1 | Campaign 1 accepted (1 = Yes, 0 = No) | Categorical (binary) | Binary |
| AcceptedCmp2 | Campaign 2 accepted (1 = Yes, 0 = No) | Categorical (binary) | Binary |
| AcceptedCmp3 | Campaign 3 accepted (1 = Yes, 0 = No) | Categorical (binary) | Binary |
| AcceptedCmp4 | Campaign 4 accepted (1 = Yes, 0 = No) | Categorical (binary) | Binary |
| AcceptedCmp5 | Campaign 5 accepted (1 = Yes, 0 = No) | Categorical (binary) | Binary |
| Complain | Whether the consumer has filed a complaint (1 = Yes, 0 = No) | Categorical (binary) | Binary |
| Z\_CostContact | Internal contact cost variable (constant) | Numerical (int) | Constant |
| Z\_Revenue | Internal revenue variable (constant) | Numerical (int) | Constant |
| Response | Final campaign response (1 = Yes, 0 = No) | Categorical (binary) | Binary |

Table 1: Attributes Definition

## Data Exploration

### Summary Statistics for Numerical Variables

To understand the overall structure and distribution of the dataset, summary statistics were computed for all numerical variables. The dataset consists of 2,240 consumers with a diverse range of purchasing and demographic behaviors. The Income variable showed a widespread, ranging from 1,730 to 666,666, indicating potential outliers or high-income anomalies, and it contains 24 missing values. Most purchase-related variables (e.g., MntWines, MntFruits, NumWebPurchases) have minimum values of zero, suggesting some consumers have not engaged in certain purchasing categories. The Year\_Birth ranges from 1893 to 1996, revealing possible data entry errors or exceptional age values. Although Year\_Birth is stored as an integer, it represents a time-based attribute and is therefore classified as a temporal variable rather than continuous. Since raw year values are not easily interpretable in business contexts and could lead to misleading calculations (e.g., summing two years), a derived variable “Age” was created for downstream analysis. Variables like Recency, NumDealsPurchases, and NumCatalogPurchases also displayed skewed distributions. These observations provide early insight into consumer diversity and will guide further analysis such as missing value treatment, outlier handling, and transformation if necessary.

### Frequency Analysis of Categorical Variables

To better understand the distribution of consumer profiles, frequency tables were generated for all categorical variables. The variables Education and Marital\_Status are multi-level categorical variables, while variables like AcceptedCmp1 to AcceptedCmp5, Complain, and Response are binary categorical.

In the Education variable, the majority of consumers had completed Graduation (1,127), followed by Master's (370) and PhD (486). Smaller groups included “2n Cycle” (203) and “Basic” education (54). These categories are preserved as-is, as each reflects distinct education levels that could influence spending patterns and campaign responsiveness.

In the Marital\_Status variable, the largest groups were “Married” (864), “Together” (580), and “Single” (480), with “Divorced” (232) and “Widow” (77) also showing notable representation. However, rare or inconsistent entries such as “YOLO” (2), “Absurd” (2), and “Alone” (3) were grouped under “Single” to reduce noise and improve modeling clarity. The “Widow” category was retained separately, as it may reflect a distinct consumer segment in terms of age, lifestyle, and marketing response behavior.

The binary categorical variables reflect response indicators for previous marketing campaigns. Most consumers did not accept offers (0) in campaigns AcceptedCmp1 to AcceptedCmp5, but a small proportion (e.g., 144 accepted Campaign 1, 163 accepted Campaign 3 and 5) did respond positively. The final Response variable showed a success rate of around 15%, indicating the class imbalance that will need to be addressed in the modeling stage.

These distributions provide a baseline understanding of the consumer population and offer direction for encoding strategies, segment formation, and classification modeling in later phases of the analysis.

### Missing Values, Zero Values and NaNs Analysis

The dataset contains a small number of missing values, specifically 24 entries in the Income variable. This minor level of missingness can be handled using simple imputation techniques without significantly impacting analysis integrity. Additionally, several numerical attributes contain zero values, which may carry analytical meaning or indicate lack of activity. For example, Kidhome and Teenhome contain 1,293 and 1,158 zero entries respectively, indicating many consumers do not have children or teenagers in the household. Similarly, attributes such as MntFruits (400 zeros), MntFishProducts (384 zeros), and MntSweetProducts (419 zeros) suggest that a significant portion of consumers made no purchases in these categories. Transaction-related variables like NumCatalogPurchases (586 zeros) and NumWebPurchases (49 zeros) also show many users did not engage through those channels. These zero values will be retained for further analysis, as they may provide valuable insights into consumer preferences, engagement behaviors, and potential segmentation patterns.

### Missing Value Imputation and Year\_Birth Correction

The dataset contained missing values in the Income variable as mentioned earlier. As the Income distribution appeared approximately symmetrical (based summary statistics), median imputation was chosen. Median imputation is widely recommended for numerical variables with low levels of missingness, especially when the variable may contain outliers, as the median is robust to skewness and extreme values (Soley-Bori, 2013).

In the Year\_Birth variable, two outlier values were detected: 1893 and 1900. A manual check confirmed that each appeared only once. As birth years before 1900 are implausible for contemporary consumer behavior and may reflect data entry errors, both values were imputed with the median year (1970). This approach is preferred over deletion to preserve data integrity and sample size (Allison, 2001).

### Outlier Detection Using IQR Method

A group of graphs showing the different types of income

Description automatically generated with medium confidence

Figure 1: Outliers Detection for Numerical Variables

A group of graphs with numbers

Description automatically generated with medium confidence

Figure 2: Continued - Outliers Detection for Numerical Variables

Outliers in the dataset were analyzed using boxplots as seen in the above Figure 1 and Figure 2 and statistically flagged using the Interquartile Range (IQR) method. This method was chosen over the Z-score approach because several numerical variables-such as MntFruits, MntMeatProducts, and MntSweetProducts-were observed to have right-skewed distributions during the summary statistics phase. The IQR method does not assume normality and is therefore better suited for detecting outliers in skewed or heavy-tailed data (Leys et al., 2013).

This approach flagged a substantial number of high-end outliers in spending-related variables, further confirming the variability highlighted earlier:

* MntFruits: 227 outliers
* MntSweetProducts: 248 outliers
* MntMeatProducts: 175 outliers
* MntGoldProds: 207 outliers

These outliers align with prior observations of wide value ranges and skewed distributions, suggesting the presence of high-spending customers rather than data errors. As a result, these values will be retained in the dataset, since removing them may obscure valuable insights into premium customer segments. In business analytics, preserving such extreme behaviors is often critical for capturing consumer diversity and supporting targeted marketing strategies (Aggarwal, 2015).

## Exploring Variable Relationships

To deepen understanding of consumer behavior patterns, visualizations were used to explore relationships between key variables. This helps guide feature selection, transformation, and model development in future stages.

### Correlation Heatmap

A correlation heatmap of numerical variables was generated to identify linear relationships. The heatmap revealed:

A graph of a graph

Description automatically generated with medium confidence

Figure 3: Correlation Heatmap for Numerical Variables

In the summary statistics, many purchase-related variables (e.g., MntWines, MntMeatProducts, MntGoldProds) had high ranges and extreme values, suggesting a few customers spend significantly more. In the outlier detection, those same variables had the highest number of flagged outliers, indicating a skewed pattern of high spenders.

The heatmap as illustrated above (Figure 3) then confirmed those behaviors - strong positive correlations among the same spending variables - reinforcing that heavy spenders tend to spend across multiple categories.

These patterns will help identify redundant features and refine segmentation strategies in the upcoming modeling phase.

### Income vs. Total Spending

A graph with a red line

Description automatically generated

Figure 4: Income Vs. Total Spending

The above (Figure 4) scatter plot of Income vs. Total Spending was used to explore how consumer earnings relate to their overall expenditure across product categories. Consistent with earlier summary statistics and outlier detection, a widespread in income and total spending was observed, with a few consumers showing disproportionately high values. A positive trend exists - higher-income consumers generally spend more - but several exceptions were evident. Some high-income individuals had low spending and vice versa, reflecting diverse consumer profiles. These patterns align with earlier correlation analysis and suggest that income alone does not fully explain spending behavior, reinforcing the need for segmentation and multivariate modeling in later stages.

**Note\***

The plot appears compressed near the origin due to extreme outliers in both Income and Total\_Spending, including one consumer with an income of 666,666. While these cases offer insight into premium consumers, such extremes can skew cluster centroids in segmentation models. To ensure balanced groupings, winsorization may be considered in later phases, limiting the effect of extreme values without discarding important consumer profiles (Wilcox, 2011). This allows the clustering algorithm to capture broader behavioral trends while maintaining robustness.

Additionally, a new variable **Total\_Spending** was engineered by summing up all product-related monetary attributes: MntWines, MntFruits, MntMeatProducts, MntFishProducts, MntSweetProducts, and MntGoldProds. This aggregated measure captures each consumer’s overall purchasing power and is used in subsequent analysis to evaluate spending behavior in relation to income and marketing response.

### Response Rate by Education Level

A graph of a graph with a number of blue bars

Description automatically generated with medium confidence

Figure 5: Response rate by Education Level

The above (Figure 5) proportional bar chart was used to explore how marketing response rates vary by education level. The chart suggests that consumers with PhD and Master’s degrees had relatively higher proportions of positive responses. However, frequency analysis shows that the majority of respondents came from the Graduation group, due to its large base size. This distinction highlights the importance of combining proportional response rates with actual counts to identify impactful segments. Future segmentation should consider both engagement rate and segment size to optimize targeting.

### Response Rate by Marital Status

A graph of a number of gray rectangular objects

Description automatically generated with medium confidence

Figure 6:Response Rate by Marital Status

The above (Figure 6) proportional bar chart was created to examine how marketing campaign response rates vary across different marital status groups. The chart reveals that Widow and Divorced consumers had relatively higher response rates compared to other groups, while Married and Together segments showed the lowest proportions of positive responses. Interestingly, Single consumers also showed a slightly above-average engagement level.

These trends provide early insight into how lifestyle or household structure may influence consumer responsiveness to marketing campaigns. While further analysis is needed to validate their predictive power, this visualization highlights Marital\_Status as a potentially valuable predictor for future modeling and segmentation efforts. It complements earlier demographic insights and supports the inclusion of marital factors in marketing personalization strategies.

These visual relationships reinforce the patterns observed during summary statistics and outlier detection, offering additional context on consumer behavior and response trends. They provide an early indication of which attributes may be most relevant for segmentation, classification, and campaign prediction models. In the next phase, these predictors will be formally analyzed, reduced if necessary, and transformed to support more effective modeling and insight generation.

# Predictor Analysis and Relevancy

To identify the most relevant variables for future modeling, this section evaluates the associations and potential predictive power of numerical and binary categorical variables with respect to the campaign Response variable. The goal is to ensure that only meaningful, non-redundant variables are carried forward for model development and segmentation.

## Multicollinearity Check for Numerical Variables

Variance Inflation Factor (VIF) analysis was performed on all numerical variables to detect multicollinearity. All VIF values were well below the commonly accepted threshold of 5 (maximum ~2.95 for NumCatalogPurchases), indicating no severe multicollinearity among the numerical variables. These results are consistent with the earlier correlation heatmap, where variables like MntWines, MntMeatProducts, and MntGoldProds displayed moderate positive correlations but not to a degree that justifies removal.

Hence, all numerical variables are retained at this stage for further modeling, as none appear statistically redundant or collinear (James et al., 2013).

## Association Analysis for Binary Variables

Chi-square tests of independence were used to evaluate associations among binary categorical variables (AcceptedCmp1–5, Complain, and Response). The results revealed that the current campaign response (Response) was strongly and significantly associated with previous campaign acceptances. This supports both business intuition and earlier frequency observations that consumers who responded positively to past offers are more likely to engage again.

However, the variable Complain showed no significant association with any campaign response, including the current one. This aligns with prior frequency analysis, which showed the variable was heavily imbalanced (only 21 out of 2,240 consumers filed complaints). While this variable may be important from a service quality or customer support lens, it appears to lack predictive power in the context of marketing engagement and may be removed or deprioritized during later stages of dimensionality reduction.

## Relationship Between Numerical Variables and Response

To explore how individual numerical variables relate to marketing response, boxplots were first used to visualize distribution differences between responders and non-responders. This step complements earlier outlier analysis and summary statistics by offering a comparative look at how each variable behaves across response groups.

A group of graphs showing different sizes and colors

Description automatically generated with medium confidenceA group of graphs with text

Description automatically generated with medium confidence

Figure 7: Relationship Between Numerical Variables and Response

**Key visual findings from the above Figure 7 include:**

* Product Spending Variables: Consumers who responded positively to the campaign (Response = 1) generally had higher spending on MntWines, MntMeatProducts, MntGoldProds, and MntFruits. These variables also had a large number of flagged outliers earlier, reinforcing that high-spending consumers are more engaged.
* Recency: A visibly lower median Recency was observed for responders, suggesting that recent shoppers are more likely to engage with campaigns. This aligns with earlier correlation observations.
* Income: Responders tend to have slightly higher income levels, although the visual gap was subtle. This suggests income may contribute to response behavior, which will be verified with statistical testing.
* NumWebPurchases and NumCatalogPurchases: These variables also showed slightly higher median values among responders, indicating greater online engagement among the converted group.
* Non-distinct Variables: Year\_Birth, NumWebVisitsMonth, and NumDealsPurchases displayed minimal or no visual differences across response groups, suggesting weaker individual relationships.

#### Wilcoxon Rank-Sum Test for Statistical Validation

To statistically validate the visual patterns, the Wilcoxon rank-sum test was applied to each numerical variable. This non-parametric test was selected due to earlier findings of skewed distributions in variables such as MntSweetProducts, Income, and Recency, where normality assumptions for t-tests would not hold (Leys et al., 2013).

## Key statistical findings

Significant Differences: Most purchasing variables (e.g., MntWines, MntMeatProducts, MntFruits, NumWebPurchases, NumCatalogPurchases) had highly significant p-values, confirming that higher spending and purchase activity is associated with increased campaign responsiveness.

Recency and Income: Both showed statistically significant differences across response groups, reinforcing earlier visual and business assumptions.

Non-significant Variables: Year\_Birth, NumWebVisitsMonth, and NumDealsPurchases did not show significant group differences, indicating these may be less informative for prediction.

**Summary**

This analysis provides a crucial bridge between exploratory insights and formal variable evaluation. It confirms that variables reflecting high spending, recent activity, and past campaign engagement hold strong predictive relevance and should be prioritized for modeling. Conversely, variables like Complain, Year\_Birth, and NumWebVisitsMonth appear less useful and can be considered for removal or transformation during feature engineering or dimensionality reduction.

By grounding these conclusions in both visual and statistical evidence and aligning them with earlier data profiling, this step ensures the modeling phase is built on robust, interpretable, and business-relevant inputs. In the next section, appropriate transformations will be made such as log scaling, encoding, and feature derivation to prepare the refined variables for modeling.

# Data Engineering and Transformation

To prepare the dataset for modeling and improve interpretability and performance, several transformations were performed based on prior exploratory analysis and predictor evaluation.

## Log Transformation for Skewed Variables

As identified in earlier summary statistics and outlier detection, several spending-related variables such as MntWines, MntMeatProducts, and Total\_Spending, were heavily right-skewed with extreme outliers. This skewness posed challenges for modeling due to the influence of a small number of high-spending consumers. To address this, a log transformation using log1p(), (log(1 + x)) was applied to normalize distributions while preserving zero values, as recommended in prior research (Osborne, 2002).

A graph of different sizes and colors

Description automatically generated with medium confidence

Figure 8: Original Vs. Log-Transformed Spending Variables

The above (Figure 8) faceted histogram comparison illustrates the transformation’s effectiveness. In the right-hand panels, the original versions show extreme right-skew, with most values clustered at the low end and long tails stretching toward high values. In contrast, the left-hand panels (log-transformed) exhibit more symmetric, compact distributions with reduced influence from outliers. This improves model interpretability, enhances numerical stability, and prevents high-spending consumers from dominating learning algorithms.

These normalized variables (e.g., log\_MntWines, log\_Total\_Spending) will be prioritized during model development to support more accurate, balanced, and generalizable predictive performance.

**Derived Variable: Age**

While Year\_Birth was initially included in the dataset, it is not directly interpretable in a marketing context. Therefore, an Age variable was created using the formula:

Age = 2015 - Year\_Birth,

Year 2015 is the reference year based on consumer enrollment dates. This transformation enables a clearer understanding of how consumer life stage relates to spending and campaign responsiveness.

### Ordinal Encoding of Education

The Education variable was converted into an ordinal numeric scale to reflect increasing levels of qualification. Based on business logic and marketing relevance, the categories were ordered as: "Basic" < "2n Cycle" < "Graduation" < "Master" < "PhD".

This transformation allows modeling algorithms to understand the progression of education levels while preserving ordinal relationships.

### Dummy Encoding of Marital\_Status

The Marital\_Status variable, being nominal, was dummy encoded to avoid imposing an artificial order. Using one-hot encoding (excluding the intercept), binary flags were created for each marital category. This allows flexible incorporation of marital dynamics into modeling without violating assumptions of linearity or order.

### Winsorization of Income

A single consumer was found with an exceptionally high income (666,666), well beyond typical values. To reduce its undue influence on both clustering and regression, the Income variable was Winsorized at the 95th percentile, replacing all extreme values above this threshold with the capped limit (Ghosh & Vogt, 2012). This approach preserves the customer for analysis while preventing distortion in models sensitive to scale or outliers. The winsorized version of Income is used consistently across classification, regression, and clustering tasks.

A graph showing a red line and a dotted line

Description automatically generated

Figure 9: Winsorized Income VS. Total Spending

To support earlier insights and maintain consistency, the above (Figure 9) updated scatter plot visualizes the relationship after winsorizing extreme income values. Compared to the original, the chart now presents a clearer concentration of consumer patterns across the mid-range income levels, while still preserving the general upward trend between earnings and expenditure. This reinforces the initial observation that income is positively associated with spending, though not deterministically, while addressing prior distortion caused by outliers.

These transformations preserve the interpretability of consumer attributes while optimizing the dataset for downstream modeling tasks such as classification and clustering. The inclusion of business-aware features like Age, log-transformed spend metrics and winsorized income ensures the models will be both accurate and actionable.

# Dimension Reduction

To ensure modeling efficiency and interpretability, this section focuses on reducing the number of input variables by removing redundant, low-impact, or overlapping attributes. The goal is to retain only those variables that contribute meaningful variance and align with business objectives, as established in earlier analysis.

## Redundancy and Relevance Filtering

Based on the outcomes of the multicollinearity check (Variance Inflation Factor analysis), Wilcoxon rank-sum tests, and chi-square tests performed in the previous section, the following variables were identified for exclusion:

* Year\_Birth: Replaced with the more interpretable Age variable. As discussed earlier, age provides clearer insights into consumer lifecycle stages than raw birth year, and Year\_Birth showed no significant relationship with campaign response.
* NumWebVisitsMonth and NumDealsPurchases: These variables did not show statistically significant differences between responders and non-responders in Wilcoxon tests and offered minimal business interpretability. Their exclusion reduces dimensional noise without sacrificing predictive value.
* Complain: Found to be highly imbalanced and statistically insignificant in both chi-square tests and frequency analysis. It was deemed irrelevant for predicting campaign responsiveness.
* Z\_CostContact and Z\_Revenue: Both are constant across all observations and therefore provide no variance or modeling value. These were safely discarded.

MntWines, MntFruits, MntMeatProducts, MntFishProducts, MntSweetProducts, MntGoldProds, Total\_Spending: Although valuable in earlier analysis, these original spending variables were replaced by their log-transformed versions to address skewness and outliers more effectively (Osborne, 2002). Their transformed versions (log\_\*) are retained instead for better model performance and interpretability.

These decisions are fully aligned with findings from the predictor analysis, which showed that high-spending behaviors (especially in wine, meat, and gold products) were strong indicators of campaign response. However, transformation was required to account for extreme skewness and variability.

## Avoidance of PCA

While Principal Component Analysis (PCA) is a common dimensionality reduction technique, it was deliberately avoided in this project. PCA transforms original variables into abstract principal components, which often sacrifices interpretability. Since the business goal is not just predictive accuracy but also actionable insights into consumer behavior, retaining transparent and explainable variables is crucial. This ensures that the marketing team can understand and act on model outputs.

## Final Reduced Dataset

The final dataset contains 29 variables including:

* Log-transformed monetary variables (7)
* Key demographic and behavioral variables (e.g., Income, Recency, Age)
* Encoded categorical variables: Education\_Ord and dummies for Marital\_Status
* Previous campaign indicators (AcceptedCmp1 to 5)
* Response variable

The retained variables were selected not only based on statistical evidence, but also for their business interpretability and modeling utility. This refined dataset will be used in subsequent stages of supervised learning, ensuring that the model is both efficient and meaningful.

# Data Partitioning Methods

To evaluate predictive performance and generalizability, the dataset was split into training and testing subsets. Given the dual nature of the project - classification and regression - two separate partitioning strategies were adopted. For the Segmentation, whole dataset is used without any need of data partitioning.

For Classification  
To support Shop-Smart’s objective of improving future campaign targeting, the Response variable was selected as the target for classification modeling. Unlike AcceptedCmp1 through AcceptedCmp5, which reflect past campaign engagement, Response captures whether a consumer responded to the final or most recent campaign. This variable best aligns with the business question: “Which consumers are most likely to respond to an upcoming campaign?” Modeling this outcome allows the marketing team to optimize future outreach and improve return on investment.

Meanwhile, the AcceptedCmp1-5 variables serve as predictors in the model, representing a consumer’s historical responsiveness and engagement, key behavioral traits used to forecast future intent. This modeling structure respects temporal causality, which is vital for valid prediction tasks (Kuhn & Johnson, 2013). Moreover, this approach aligns with best practices in campaign modeling, where prior behavior informs future response likelihood (Wedel & Kamakura, 2000).

While alternative approaches such as combining past campaigns into a composite outcome were considered, such transformations would shift the objective toward modeling cumulative behavior, not next-action prediction, which was clearly emphasized in both the business and analytics goals. Therefore, retaining Response as the classification target supports a forward-looking, decision-oriented model, ensuring that insights are both interpretable and actionable for real-time marketing interventions (Fawcett, 2006).  
The Response variable, indicating campaign acceptance, was found to be highly imbalanced, with approximately 85% non-responders (0) and 15% responders (1). In such cases, traditional random splitting may distort class proportions, leading to biased model training and evaluation.  
To preserve class distribution across splits, stratified partitioning was used to split the data into:

* 80% training set
* 20% testing set

This ensures the minority class is proportionally represented in both datasets, improving the reliability of performance metrics like precision, recall, and AUC.

Given the imbalance, resampling strategies such as oversampling the minority class or using class weighting will also be considered during model building to further address this issue (He & Garcia, 2009).

For Regression  
The regression task aims to predict total consumer spending behavior (log-transformed to address skewness). Since the target variable is continuous and reasonably distributed, random sampling with 80/20 train-test split was sufficient. No stratification was necessary in this case.

These partitioning strategies ensure a balanced and fair foundation for building classification and regression models, while acknowledging the underlying structure and distribution of each target variable.

# Model Selection

To prepare for predictive and segmentation analysis, suitable models were selected for each of the three analytics tasks: classification, regression, and clustering. The goal of this section is to establish the modeling direction and indicate which methods will be evaluated in subsequent phases.

## Segmentation - Unsupervised Learning

To identify natural consumer groupings based on behavior and demographics, k-means clustering was selected. Since all variables used were numeric or binary encoded, k-means was preferred over k-prototypes. Two configurations-k = 2 and k = 3-were chosen for comparison and will be evaluated using silhouette scores in the next section.

**Note:**

A derived variable Family\_Size was engineered specifically for segmentation purposes. It combines household structure using marital status (coded as 1 for single/divorced/widow, 2 for married/together) with Kidhome and Teenhome. This feature provides clearer segmentation patterns and does not replace or remove the original household composition variables.

## Classification - Supervised Learning

To predict whether a consumer will respond to a marketing campaign (Response). The following classification models were selected:

* **Logistic Regression:** for interpretability and baseline comparison
* **Random Forest Classification:** for capturing non-linear relationships and interactions

Stratified sampling and oversampling will be used to address class imbalance during model training and validation.

## Regression - Supervised Learning

To estimate overall consumer engagement in spending (log\_Total\_Spending), the following regression models were selected:

* **Linear Regression:** to establish a baseline understanding of linear relationships
* **Random Forest Regression:** to capture complex spending behaviors and improve prediction accuracy

These models will be implemented and evaluated in the following section. The next phase will focus on model fitting, validation accuracy, and performance comparison across both supervised and unsupervised tasks.

# Model fitting, validation accuracy and test accuracy

To validate the feasibility and predictive strength of the selected models, supervised and unsupervised learning techniques were applied to the preprocessed dataset.

## Segmentation

For the clustering task, k-means was applied using scaled numeric variables (including winsorized income, log-transformed spending, and engineered features like family size). Cluster memberships were assigned for:

* k = 2, yielding clusters of 1,071 and 1,169 consumers.
* k = 3, yielding clusters of 783, 853, and 604 consumers.

These cluster configurations will be interpreted in detail in the following section to support consumer profiling and marketing strategy development.

**Clustering Design Justification:**

To ensure unbiased and interpretable segmentation, only intrinsic consumer attributes and behavioral variables were included in clustering such as income, age, education, channel engagement, and product purchases. Downstream outcome variables (e.g., campaign responses or complaint behavior) were intentionally excluded to prevent data leakage and allow the discovery of natural customer groupings. These outcome variables were later used to profile and evaluate the clusters post hoc. This approach follows best practices in customer segmentation, where clustering is performed on stable traits and followed by outcome-based interpretation (Dolnicar, 2004).

## Classification Models

As mentioned earlier about the class imbalance, oversampling on the minority class has been performed only for the classification data.

Two classification models were trained to predict consumer campaign response:

* Logistic Regression: Accuracy on test set: 77.9%
* Random Forest Classifier: Accuracy on test set: 86.2%

The random forest classifier demonstrated stronger predictive performance, likely due to its ability to capture nonlinear interactions and variable importance.

## Regression Models

To ensure interpretability, inverse log transformation was performed on the model predictions and originals before calculating RMSE. This allows error metrics to be reported on the original spending scale, aligning with business understanding. This transformation will also be applied during evaluation and future prediction phases.

To predict consumer spending behavior, two regression models were fitted using log\_Total\_Spending as the response variable:

* Linear Regression: RMSE (original scale): 202.13
* Random Forest Regression: RMSE (original scale): 115.82

Random forest regression yielded significantly lower RMSE, suggesting stronger performance in capturing the nonlinear relationship between predictors and total spending.

In summary, the initial model results suggest that random forest models offer superior performance in both classification and regression tasks. In the next section, we will conduct a deeper evaluation of each model, reporting confusion matrices, ROC curves, and further performance diagnostics.

# Report Models Performance

## Segmentation Performance – Silhouette Analysis

To assess cluster quality, silhouette plots were used to visualize how well each consumer fits within their assigned segment across two clustering solutions:

### Silhouette plot of K = 2 Clusters

Figure 10: Silhouette plot of 2 clusters

A graph of a number of objects

Description automatically generated with medium confidence

### Silhouette plot of K = 3 Clusters

A graph of a number of colored triangles

Description automatically generated with medium confidence

Figure 11: Silhouette plot of 3 clusters

|  |  |  |  |
| --- | --- | --- | --- |
| Clustering Solution | Avg. Silhouette Width | Cluster Sizes | Interpretation |
| ****K = 2**** | **0.26** | Cluster 1: 1071 (Avg. Si = 0.30)  Cluster 2: 1169 (Avg. Si = 0.22) | Moderate structure with well-separated clusters |
| ****K = 3**** | 0.16 | Cluster 1: 783  Cluster 2: 853  Cluster 3: 604  (Si ≈ 0.09–0.22) | Weak separation: third cluster poorly formed |

Table 2: Clusters Silhouette Summary

The K = 2 solution showed a stronger clustering structure, as reflected by the higher average silhouette score and more distinct separation between groups as illustrated by above Figure 10 and Figure 11. These results will guide final cluster interpretation in the next section.

## Classification Models

Both models were evaluated on the test set using confusion matrices, ROC AUC, and other standard classification metrics. The positive class was set to 1, corresponding to campaign responders.

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Model | Accuracy | Sensitivity | Specificity | Balanced Accuracy | AUC |
| Logistic Regression | 0.779 | 0.742 | 0.785 | 0.764 | 0.8366 |
| Random Forest | 0.862 | 0.439 | 0.935 | 0.687 | 0.8308 |

Table 3: Performance Metrics of Classification models

While Random Forest achieved higher accuracy overall as shown in the Table 3, Logistic Regression showed better sensitivity (true positive rate), which may be critical depending on the campaign goal. These tradeoffs will be discussed in the evaluation section.

### ROC Curves

To visualize classification performance, ROC curves were plotted for both models. These curves illustrate the trade-off between sensitivity and specificity across different thresholds. Both models show curves rising well above the diagonal, indicating better-than-random performance.

#### Logistic Regression ROC Curve

AUC: 0.8366

A graph of a logistic regression

Description automatically generated

Figure 12: Logistic Regression ROC Curve

#### Random Forest ROC Curve

AUC: 0.8308

A green line graph with text

Description automatically generated

Figure 13: Random Forest Regression ROC Curve

Both models achieved similar AUC values as seen from the above Figure 12 and Figure 13, supporting their reliability in classifying responders, though with different tradeoffs in sensitivity vs. specificity.

## Regression Models

Regression models were evaluated on the original scale by applying an inverse log transformation to the predicted values.

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| Model | MAE | MSE | RMSE | MAPE | R² | Adj R² |
| Linear Regression | 102.25 | 40857.17 | 202.13 | 17.55% | 0.9057 | 0.8999 |
| Random Forest Regression | 50.35 | 13414.27 | 115.82 | 7.90% | 0.9711 | 0.9693 |

Table 4: Regression Models Evaluation Metrics

The Random Forest Regression clearly outperformed Linear Regression across all error metrics. Notably, it achieved more than 50% lower RMSE and a substantially better MAPE with a higher adjusted R2, indicating stronger predictive precision.

## Predictor Importance – Random Forest

To understand key predictors driving model decisions, variable importance plots were generated from the Random Forest models.

### Random Forest Classification – Top Predictors:

A graph with dots on it

Description automatically generated

Figure 14: Predictor Importance - Random Forest Classification

Top predictors included Recency, log\_Total\_Spending, log\_MntMeatProducts, and NumCatalogPurchases as seen from the above Figure 14.

### Random Forest Regression – Top Predictors:

A graph with numbers and lines

Description automatically generated

Figure 15: Predictor Importance - Random Forest Regression

Key spending predictors such as log\_MntWines, log\_MntMeatProducts, and NumCatalogPurchases showed the highest influence as seen from the above Figure 15.

These visuals set the stage for detailed interpretation in the upcoming evaluation section.

# Model Evaluation

This section evaluates the performance and relevance of the selected models for classification, regression, and segmentation. Beyond raw metrics, it emphasizes interpretability, practical utility, and alignment with business objectives.

## Classification Models Evaluation

Two models were trained to predict whether a consumer would respond positively to a marketing campaign: Logistic Regression and Random Forest Classification. Given the class imbalance (positive response rate ≈ 15%), oversampling via the ovun.sample() method was employed on the training data to enhance the minority class representation without loss of information.

* Logistic Regression showed better sensitivity, suggesting it was more effective in detecting actual responders, which is crucial when campaign reach is limited and targeting efficiency is essential.
* Random Forest, while offering higher overall accuracy and specificity, underperformed in sensitivity. This suggests it favored the majority class (non-responders) even after oversampling, a known limitation in tree-based models without additional class weighting (Kuhn & Johnson, 2013).

ROC Curve Insights: Both models demonstrated reasonable discriminatory ability with AUC scores above 0.83. Logistic regression achieved marginally better class separation, reinforcing its value as a reliable baseline model for imbalanced binary classification problems (Fawcett, 2006).

Variable Importance (Random Forest): The Random Forest’s variable importance plot revealed Recency, log\_Total\_Spending, and log\_MntMeatProducts as top predictors, aligning with domain expectations. Consumers who engaged recently and exhibited high spending were more likely to respond. This supports targeting strategies that prioritize recently active, high-value customers.

### Cost-Sensitive Learning Analysis

To complement standard performance metrics, a Cost-Sensitive Learning (CSL) framework was applied to the classification models. In this approach, the business impact of misclassification is explicitly quantified. The cost of incorrectly predicting a responder as a non-responder (False Negative) was set at $13.00 (missed donation), and the cost of incorrectly predicting a non-responder as a responder (False Positive) was set at $0.68 (cost of mailing without donation).

Cost-Sensitive Learning Formula:

Where,

Cost(FN) = $13.00

Cost(FP) = $0.68

**Logistic Regression:**

False Negatives (FN) = 17

False Positives (FP) = 82

Total Cost (Logistic Regression) = (13 \* 17) + (0.68 \* 82) = 276.76

Total Cost for Logistic Regression = $276.76

**Random Forest:**

False Negatives (FN) = 37

False Positives (FP) = 25

Total Cost (Random Forest) = (13 \* 37) + (0.68 \* 25) = 498.00

Total Cost for Random Forest = $498.00

**Interpretation:**

Since the goal of cost-sensitive learning is to minimize total misclassification cost, Logistic Regression clearly outperforms Random Forest. Logistic Regression leads to a lower total misclassification cost ($276.76) compared to Random Forest’s higher cost ($498.00).

This cost-based evaluation provides stronger business justification: Logistic Regression is more effective in preserving Shop-Smart’s marketing budget and maximizing the return from targeted campaigns by minimizing both mailing costs and lost responder opportunities.

## Regression Models Evaluation

The goal was to estimate a consumer’s Total Spending based on demographic and behavioral attributes. Since spending distributions were heavily skewed, log transformation was applied to stabilize variance. Post-modeling, inverse log transformation was used to recover predictions on the original scale for interpretable evaluation.

* Random Forest Regression significantly outperformed the linear model across all metrics, especially in MAPE and MAE, making it a better fit for capturing non-linear and interaction effects within consumer behavior.
* The improvement in R² (0.97) indicates that the model explains a large proportion of the variance in spending, while a low MAPE suggests predictions are on average within 8% of actual values, a strong result in marketing analytics.

Variable Importance (Random Forest): Key drivers of spending included log\_MntWines, log\_MntMeatProducts, and catalog/store/web purchases, which aligns with segmentation hypotheses from earlier clustering steps. These insights can guide cross-selling strategies and personalized marketing offers.

## Segmentation: K-Means Clustering Evaluation

K-Means clustering was applied on scaled variables including consumer demographics, purchase patterns, and key encoded categorical attributes (e.g., marital status dummies). The Response variable was excluded to ensure unbiased segmentation based on natural consumer behavior rather than campaign outcomes.

* The K = 2 solution was selected based on its higher average silhouette score (0.258) compared to K = 3 (0.158), indicating better cluster cohesion and separation (Rousseeuw, 1987).
* The silhouette plot for K = 2 showed two reasonably distinct clusters, while the third cluster in K = 3 showed poor structure, reinforcing the selection of K = 2.

### Interpretation of Cluster Profiles (K = 2)

To derive actionable business insights, the clusters were profiled based on key numeric and categorical variables. Below charts show how spending behavior, demographics, and campaign engagement differ by cluster.

#### Total Spending by Cluster

**A diagram of a cluster

Description automatically generated with medium confidence**

Figure 16: Total Spending by Cluster

To provide a business-level interpretation of consumer value, Total Spending was back-transformed from its log scale and visualized in the original dollar scale. The distribution as illustrated above (Figure 16) clearly indicates that Cluster 2 exhibits a substantially higher median and overall range of spending compared to Cluster 1. This finding reinforces the value of prioritizing Cluster 2 in high-return campaigns.

#### Income by Cluster

A diagram of a cluster

Description automatically generated with medium confidence

Figure 17: Income by Cluster

#### Age by Cluster

A diagram of a cluster

Description automatically generated with medium confidence

Figure 18: Age by Cluster

**Key Observations from the Figure 17 and 18:**

* Cluster 2 comprises consumers with higher income, higher total spending, and a slightly older age profile.
* Cluster 1 represents younger consumers with moderate-to-low income and lower overall spending.

#### Response and AcceptedCmp1–5 by Cluster

A group of blue and pink squares

Description automatically generated

Figure 19: Response and AcceptedCmp1–5 by Cluster

#### Marital Status Dummies by Cluster

A screenshot of a graph

Description automatically generated

Figure 20: Marital Status by Cluster

**Key Observations from the Figure 19 and 20:**

* Cluster 2 had a significantly higher positive response rate to past campaigns and greater participation across campaigns (Cmp1–Cmp5).
* Marital status proportions were fairly balanced across clusters, but Cluster 2 showed slightly higher proportions in married and together segments, potentially indicating stable household units with stronger purchasing power.

**These segmentation results reveal two distinct consumer profiles:**

* Cluster 1: Younger, budget-conscious consumers with limited response to marketing.
* Cluster 2: Older, high-spending, high-response consumers who present ideal targets for personalized, ROI-driven campaigns.

These behavioral clusters form the basis for data-driven targeting strategies discussed in the next section.

# Observations and Conclusion

This project was initiated to address Shop-Smart’s business need for improved consumer engagement through personalized marketing strategies. By leveraging structured data analytics, the project transformed historical consumer data into insights that support smarter segmentation, targeted outreach, and more effective resource allocation, directly aligning with the business and analytics goals.

**Key Observations:**

* **Consumer Segmentation:**  
  K-means clustering revealed two distinct consumer groups.
  + Cluster 1 represents younger, lower-income consumers with modest spending and limited marketing responsiveness.
  + Cluster 2 includes older, higher-income consumers with significantly higher spending and stronger engagement across multiple campaigns. This segmentation directly supports the goal of behavior-based targeting and lays the foundation for personalized marketing strategies.
* **Predictive Modeling for Engagement and Spend Forecasting:**  
  Supervised models confirmed that consumer behaviors such as recency of purchase, prior campaign responses, and overall spending patterns are strong indicators of future engagement.  
  Classification models enabled identification of likely responders, while regression models supported transaction-level forecasting, aligning with the objectives of improving response rates and forecasting revenue growth.
* **Addressing Analytical Challenges:**  
  Early concerns around data skewness and class imbalance were successfully handled through appropriate transformations and resampling techniques. This strengthened the models' reliability and ensured that findings were both statistically sound and business relevant.

Final Recommendations  
Based on the findings from segmentation, predictive modeling, and profit-driven analysis, the following recommendations are proposed to optimize marketing effectiveness:

* **Prioritize High-Value Consumers:** Focus targeted campaigns toward Cluster 2 consumers who demonstrate higher spending and greater responsiveness.
* **Implement Personalized Campaigns:** Design differentiated offers and messaging strategies based on consumer demographics and past purchasing behaviors.
* **Utilize Sensitive Targeting Models:** Deploy logistic regression models to maximize responder identification and campaign profitability.
* **Forecast Engagement Dynamically:** Use transaction-level engagement predictions to better allocate marketing budgets and set realistic revenue expectations.
* **Scale Analytics Across Marketing:** Extend segmentation and prediction frameworks across future campaigns to refine personalization and drive continuous improvement.

These recommendations lay the groundwork for Shop-Smart's future strategies, ensuring that marketing initiatives are data-driven, targeted, and results-oriented.

Overall, the analytics insights generated from this project support data-driven decision-making by identifying actionable consumer segments, optimizing campaign targeting, and enabling more accurate revenue forecasting, meeting both the analytics and business objectives established at the outset.

# Conclusion and Future scope

This project set out to support Shop-Smart’s overarching business goal of improving marketing campaign effectiveness and boosting consumer engagement by leveraging consumer-level analytics. Through a structured and methodical analytics approach, from data preprocessing and transformation to segmentation and predictive modeling, the project delivered meaningful insights that directly aligned with the company's business and analytics objectives.

Despite challenges such as skewed spending distributions, class imbalance, and the need for careful variable selection, the analysis successfully identified distinct consumer segments and built reliable predictive models. Segmentation revealed two clear market groups, allowing for behavior-based targeting, while classification and regression modeling provided actionable intelligence on consumer responsiveness and spending engagement. These findings confirmed early assumptions about the value of consumer behavior analytics and demonstrated the ability of data science to transform raw consumer records into strategic marketing decisions.

Looking ahead, the outcomes of this project lay the groundwork for even broader analytics integration across Shop-Smart’s marketing initiatives. Future efforts could involve developing unified consumer scoring models that combine multi-campaign responsiveness, implementing profit-based optimization frameworks to refine marketing ROI further, and adopting real-time scoring systems to personalize engagement dynamically. With this scalable, adaptable analytics framework, Shop-Smart is well-positioned to evolve its marketing strategies and capitalize on emerging consumer behaviors in a rapidly changing marketplace.

# References

* Allison, P. D. (2001). Missing data. Sage Publications. <https://methods.sagepub.com/book/mono/missing-data/toc>
* Soley-Bori, M. (2013). Dealing with missing data: Key assumptions and methods for applied analysis [Technical report]. Boston University.  
  <https://www.researchgate.net/publication/323266125_Dealing_with_missing_data_key_assumptions_and_methods_for_applied_analysis>
* Leys, C., Ley, C., Klein, O., Bernard, P., & Licata, L. (2013). Detecting outliers: Do not use standard deviation around the mean, use absolute deviation around the median. Journal of Experimental Social Psychology, 49(4), 764–766. <https://doi.org/10.1016/j.jesp.2013.03.013>
* Aggarwal, C. C. (2015). Outlier analysis (2nd ed.). Springer.

<https://doi.org/10.1007/978-3-319-47578-3>

* Osborne, J. W. (2002). Notes on the use of data transformations. Practical Assessment, Research, and Evaluation, 8(6), 1–7. <https://doi.org/10.7275/2m2h-c250>
* He, H., & Garcia, E. A. (2009). Learning from imbalanced data. IEEE Transactions on Knowledge and Data Engineering, 21(9), 1263–1284. <https://doi.org/10.1109/TKDE.2008.239>
* Ghosh, D., & Vogt, A. (2012). Outliers: An evaluation of methodologies. In Proceedings of the Joint Statistical Meetings. American Statistical Association. <https://www.researchgate.net/publication/258994085_Outliers_An_Evaluation_of_Methodologies>
* Wilcox, R. R. (2011). Modern statistics for the social and behavioral sciences: A practical introduction. CRC Press. <https://doi.org/10.1201/b10960>
* Dolnicar, S. (2004). Beyond "commonsense segmentation": A systematics of segmentation approaches in tourism. Journal of Travel Research, 42(3), 244–250. <https://doi.org/10.1177/0047287503258830>
* Kuhn, M., & Johnson, K. (2013). Applied predictive modeling. Springer. <https://doi.org/10.1007/978-1-4614-6849-3>
* Shmueli, G., Bruce, P. C., Gedeck, P., & Patel, N. R. (2020). Data mining for business analytics: Concepts, techniques, and applications in R (2nd ed.). Wiley. <https://www.dataminingbook.com/editions/r-2nd-edition>
* Fawcett, T. (2006). An introduction to ROC analysis. Pattern Recognition Letters, 27(8), 861–874. <https://doi.org/10.1016/j.patrec.2005.10.010>
* Wedel, M., & Kamakura, W. A. (2000). Market segmentation: Conceptual and methodological foundations (2nd ed.). Springer. <https://link.springer.com/book/10.1007/978-1-4615-4651-1>