**Consumer Segmentation Analysis**

Karan Rajeshbhai Trivedi

M.Sc. Data Analytics

Webster University

CSDA 6010: Data Analytics Practicum

Prof. Dr. Ali Ovlia

**Abstract**

This project explores consumer segmentation and predictive modeling to support AXANTEUS, a market research agency, in understanding purchase behavior and brand loyalty among consumers. Leveraging a dataset of 600 consumer profiles, the study aimed at segment markets based on purchase behavior and the basis of purchase, identify value-conscious consumers, and predict brand loyalty.

The analysis involved data cleaning, normalization, and clustering using k-means. Optimal clusters were determined by using the elbow method and evaluated with silhouette scores. Predictive models, particularly Random Forest, classified consumers as value-conscious and forecasted brand runs. Feature importance highlighted critical predictors influencing consumer behavior.

Findings enable AXANTEUS’s clients to design cost-effective promotions and loyalty programs, aligning marketing strategies with distinct consumer segments. By integrating segmentation with predictive insights, this project contributes to more effective allocation of promotional budgets and increased consumer engagement.

**Executive Summary**

This project focuses on analyzing consumer behavior and purchase patterns to provide actionable insights for AXANTEUS, a market research agency. The primary goal is to segment consumers based on purchase behavior and on the basis of purchase, identify value-conscious consumers, and predict brand loyalty (measured by brand runs). The insights derived aim to help AXANTEUS’s clients, including consumer goods manufacturers and advertising agencies, design more targeted and cost-effective marketing strategies.

The analysis began with comprehensive data cleaning and preprocessing. Missing and zero values in critical variables such as education level, affluence index, and household size were systematically addressed using statistical imputation techniques. Key variables were normalized to prepare for clustering and predictive modeling.

Using k-means clustering, the dataset was segmented into distinct consumer groups based on purchase behavior, basis of purchase, and their combination. Optimal cluster numbers were determined using the elbow method, and silhouette scores evaluated the quality of the segmentation. Clusters were analyzed to identify value-conscious consumers, who are price-sensitive and responsive to discounts.

Predictive models were developed using Random Forest to classify consumers as value-conscious or not and to predict brand loyalty. Feature importance analysis revealed key drivers of consumer behavior, enabling a deeper understanding of factors influencing purchase decisions. Model performance was assessed using metrics such as accuracy, RMSE, and MAPE, ensuring robust and reliable predictions.

The results empower AXANTEUS and its clients to allocate promotional budgets more effectively, launch targeted campaigns, and enhance consumer loyalty through tailored reward programs. By linking demographic attributes to purchase patterns, the analysis provides a strategic framework for improving market segmentation and optimizing marketing investments.

**Table of Contents**

1.0 [Introduction 8](#_Toc166084884)

[2.0 Business Problems and Goals](#_Toc166084885) 8

[2.1 Business Problem:](#_Toc166084886) 8

[2.2 Analytics Goals:](#_Toc166084887) 8

[3.0 Data Preprocessing](#_Toc166084888) 9

[3.1 Attributes Definition:](#_Toc166084889) 9

[3.2 Data Preprocessing:](#_Toc166084890) 10

[3.2.1 Missing Values 1](#_Toc166084891)3

3.2.2 Checking for zeros 14

3.2.3 Data Transformation (Normalization) 15

3.3 Data Exploration 17

3.4 Correlation Analysis 24

**4.0 Predictor Analysis and Relevancy 25**

4.1 Variable Selection 25

4.2 Feature Importance Using Random Forest 26

**5.0 K-Means Clustering 27**

5.1 Cluster 1- Purchase Behavior 27

5.2 Cluster 2- Basis of Purchase 30

5.1 Cluster 3: Purchase Behavior and Basis of Purchase 31

**6.0 Data Partitioning Methods 33**

**7.0 Model Selection 34**

7.1 Classification Model 34

7.1.1 Logistic Regression Model 34

7.1.2 Decision Tree Model 36

7.1.3 Random Forest 37

7.2 Regression Model 39

7.2.1 Multiple Linear Regression 39

7.2.2 Regression Tree Model 40

7.2.3 Random Forest Model 41

**8.0 Best Model Selection 42**

**9.0 Report Model Performance 42**

9.1. Final Model Observation 42

9.2. Model Comparison - Strength & Weakness 43

9.3. Insights 43

**10.0 Model Evaluation 43**

**11.0 Observation & Conclusion 43**

11.1. Key Observations 43

11.2. Conclusion 44

**12.0 Recommendations** 45

**List of Figures**

[Figure 3.2.1: Attributes of Data 12](#_Toc166083488)

[Figure 3.2.1: First Six Rows and Dimension of DATA 13](#_Toc166083489)

[Figure 3.2.3: Checking Missing Values 14](#_Toc166083490)

[Figure 3.2.4:Checking Zero Values 15](#_Toc166083491)

[Figure 3.2.5: Normalised Data 1](#_Toc166083492)6

[Figure 3.3.1: Distribution of total volume 17](#_Toc166083493)

[Figure 3.3.2: Total volume vs. average price 18](#_Toc166083495)

[Figure 3.3.3: Eating habit distribution 19](#_Toc166083494)

[Figure 3.3.4: distribution of promotion volume 19](#_Toc166083495)

[Figure 3.3.5: Brand runs by socioeconomics 20](#_Toc166083496)

[Figure 3.3.6: Average volume, transactions and price by class 21](#_Toc166083497)

[Figure 3.3.7: relationship b/w affluence index and brand runs 22](#_Toc166083498)

[Figure 3.3.8: brand-wise percentage of volume purchased 23](#_Toc166083498)

[Figure 3.3.9: Percentage of volume purchased by price category 24](#_Toc166083498)

[Figure 3.4.1: Correlation matric - classification 24](#_Toc166083498)

[Figure 3.4.2: Correlation Matrix - regression 25](#_Toc166083498)

[Figure 4.1.1: random forest feature importance - classification 26](#_Toc166083498)

[Figure 4.1.2: Random forest feature importance - regression](#_Toc166083498) 27

[Figure 5.1.1: Silhoutte plot for Purchase behavior 28](#_Toc166083498)

[Figure 5.1.2: cluster centers – purchase behavior 29](#_Toc166083498)

[Figure 5.1.3 :cluster visualisation – purchase behavior 29](#_Toc166083498)

[Figure 5.1.4: silhoutte plot – basis of purchase 30](#_Toc166083488)

[Figure 5.1.5: cluster centers for basis of purchase 31](#_Toc166083489)

[Figure 5.1.6: cluster visualisation for basis of purchase 31](#_Toc166083490)

[Figure 5.1.7: silhoutte plot for purchase behavior & BOP 32](#_Toc166083491)

[Figure 5.1.8: cluster visualisation for purchase behavior & BOP 32](#_Toc166083492)

[Figure 5.1.9: cluster centers for purchase behavior & BOP 33](#_Toc166083493)

[Figure 6.1: data partitioning – classification model 33](#_Toc166083495)

[Figure 6.2: data partitioning – regression model](#_Toc166083494) 34

[Figure 7.1.1:logistic Regression model 34](#_Toc166083495)

[Figure 7.1.2: confusion matrix - logistic 35](#_Toc166083496)

[Figure 7.2.1: Decision tree model 36](#_Toc166083497)

[Figure 7.2.2: decision tree – confusion matrix 37](#_Toc166083497)

[Figure 7.3.1: random forest – confusion matrix 38](#_Toc166083497)

[Figure 7.3.2: tuned random forest – confusion matrix 38](#_Toc166083497)

[Figure 7.4: multiple linear regression model 39](#_Toc166083497)

[Figure 7.4.1: regression tree 40](#_Toc166083497)

**1.0 Introduction**

The goal of this project is to enable AXANTEUS, a leading market research agency, to analyze consumer purchasing behavior across various product categories. AXANTEUS specializes in tracking consumer purchases in both durable goods (e.g., autos and appliances) and nondurable goods (e.g., food and toiletries).

**AXANTEUS clients include:**

Advertising agencies use the dataset monthly to help their clients modify advertising and promotional efforts.

Consumer goods manufacturers use AXANTEUS's database to track their brand's market share and client loyalty.

The study intends to go beyond typical demographic segmentation by including two key behavior-based factors: purchasing behavior (volume, frequency, and brand loyalty) and buy motivations (price sensitivity and promotional responsiveness). By studying these elements, AXANTEUS hopes to help its clients make better marketing and product positioning decisions.

**2.0 Business Problem and Goals**

**2.1 Business Problem:**

Traditional demographic segmentation methods are insufficient for understanding the complex purchasing behaviors of consumers. AXANTEUS's clients, such as consumer goods manufacturers and advertising agencies, face challenges in identifying and targeting high-value consumers effectively.

**2.2 Business Goal**:

The primary business goal of this project is to enable AXANTEUS's client, XRB, to design more cost-effective promotions targeted at appropriate market segments. By segmenting the market based on purchase behavior and the basis of purchase, AXANTEUS aims to gain insights into the demographic attributes associated with different purchase behaviors and degrees of brand loyalty. This will allow XRB to deploy promotion budgets more effectively and design consumer reward systems that increase brand loyalty.

**2.3 Analytical Goal**:

1. Segment consumers based on purchasing behavior and basis of purchase.

2. Identify value-conscious consumers who are sensitive to promotions and discounts.

3. Predict brand loyalty to help clients design targeted marketing strategies.

By achieving these goals, AXANTEUS can provide its clients with actionable insights to optimize promotional budgets, enhance consumer engagement, and improve brand loyalty. This project aims to bridge the gap between traditional segmentation methods and the need for more sophisticated, behavior-based segmentation.

**3.0 Data Preprocessing**

This section details the steps taken to clean and preprocess the dataset to ensure its suitability for analysis.

**3.1 Attributes Definition:**

The dataset consists of 600 observations and 46 variables, capturing various demographic and behavioral attributes of consumers. Key variables include:

* Demographic Information: SEC, FEH, MT, SEX, AGE, EDU, HS, CHILD, CS, Affluence.Index
* Purchase Behavior: No..of.Brands, Brand.Runs, Total.Volume, No..of..Trans, Value, Trans...Brand.Runs, Vol.Tran, Avg..Price
* Promotional and Brand-Specific Volumes: Pur.Vol.No.Promo...., Pur.Vol.Promo.6.., Pur.Vol.Other.Promo.., Br..Cd..57..144, Br..Cd..55, Br..Cd..272, Br..Cd..286, Br..Cd..24, Br..Cd..481, Br..Cd..352, Br..Cd..5, Others.999
* Product Category Preferences: Pr.Cat.1, Pr.Cat.2, Pr.Cat.3, Pr.Cat.4, PropCat.5, PropCat.6, PropCat.7, PropCat.8, PropCat.9, PropCat.10, PropCat.11, PropCat.12, PropCat.13, PropCat.14, PropCat.15

|  |  |  |  |
| --- | --- | --- | --- |
| **Attribute Category** | **Attribute** | **Variable Type** | **Description** |
| **Demographic** | Member ID | Integer | Unique identifier for each consumer. |
| SEC | Categorical (Ordinal) | Socioeconomic Class, where 1 = High and 5 = Low. |
| FEH | Categorical (Ordinal) | Eating Habit, with values: 1 = Vegetarian, 2 = Vegetarian but eats eggs, 3 = Non-vegetarian, 0 = Not specified. |
| MT | Integer | Native Language code ranging from 0 to 19. |
| SEX | Categorical | Gender of consumer: 1 = Male, 2 = Female. |
| AGE | Numeric | Age of the consumer. |
| EDU | Integer | Education level, with 1 as the minimum and 9 as the maximum education level. |
| HS | Integer | Household size, representing the number of members in the consumer's household. |
| CHILD | Integer | Number of children in the consumer's household. |
| CS | Categorical | Availability of television: 1 = Available, 2 = Not Available. |
| Affluence Index | Integer | Weighted index based on possession of durable goods, indicating the consumer’s level of affluence. |
| **Purchase Summary** | No. of Brands | Integer | Number of unique brands purchased by the consumer. |
| Brand Runs | Integer | Number of consecutive purchases of the same brand, indicating brand loyalty. |
| Total Volume | Numerical | Sum of the volume of products purchased by the consumer. |
| No. of Trans | Integer | Number of purchase transactions, with each brand purchased in a month counting as a separate transaction. |
| Value | Numerical | Total monetary value of the consumer's purchases. |
| Trans / Brand Runs | Numerical | Average number of transactions per brand run, indicating purchase frequency within the same brand. |
| Vol/Tran | Numerical | Average volume per transaction, indicating typical purchase quantity. |
| Avg. Price | Numerical | Average price per purchase transaction. |
| **Promotion-Based Purchase** | Pur Vol No Promo - % | Numerical | Percentage of purchases made without promotions. |
| Pur Vol Promo 6 % | Numerical | Percentage of purchases made under promotion code 6. |
| Pur Vol Other Promo % | Numerical | Percentage of purchases made under other promotions. |
| **Brand-wise Purchase** | Br. Cd. 57, 144, etc. | Numerical | Percentage of volume purchased under specific brand codes (e.g., Brand Code 57, 144, etc.), indicating brand preference. |
| **Price Category** | Pr Cat 1, 2, 3, 4 | Numerical | Price category classification (1-4) indicating the percentage of volume purchased under each price range. |
| **Selling Proposition** | PropCat 5, 6, ..., 15 | Numerical | Product proposition categories (5-15) indicating the percentage of volume purchased based on the selling proposition or reason for purchase (e.g., price, quality). |

**3.2 Data Preprocessing:**

The dataset, named consumer data, consists of 600 observations and 46 variables, with each row representing an individual consumer identified by a unique Member.id. The dataset includes a range of demographic and behavioural variables, capturing various aspects of consumer profiles and purchasing behaviours. Key variables include:

**Demographic Information:**

* SEC, FEH, MT: Variables likely representing socioeconomic classifications or demographic segmentation, coded as integer values.
* SEX: A binary variable indicating gender.
* AGE, EDU, HS: Variables representing age, education level, and housing status, respectively.
* CHILD: Count of children per household.
* CS and Affluence.Index: Variables indicating consumer affluence or socioeconomic status, potentially ordinal in nature.

**Purchase Behavior:**

* No..of.Brands and Brand.Runs: Indicators of brand diversity and frequency of brand engagement.
* Total.Volume, No..of..Trans, Value: Metrics representing the total volume of purchases, transaction frequency, and overall purchase value.
* Trans...Brand.Runs, Vol.Tran, and Avg..Price: Derived metrics providing insights into transactional volume, average transaction size, and pricing patterns.

**Promotional and Brand-Specific Volumes:**

Pur.Vol.No.Promo to Pur.Vol.Promo.6 : Indicate promotional and non-promotional purchase volumes, helping to analyze consumer response to promotions.

Various brand-specific volume metrics (e.g., Br..Cd...57..144, Br..Cd...55, Others.999) track purchases by brand codes, providing granular details on consumer brand preferences.

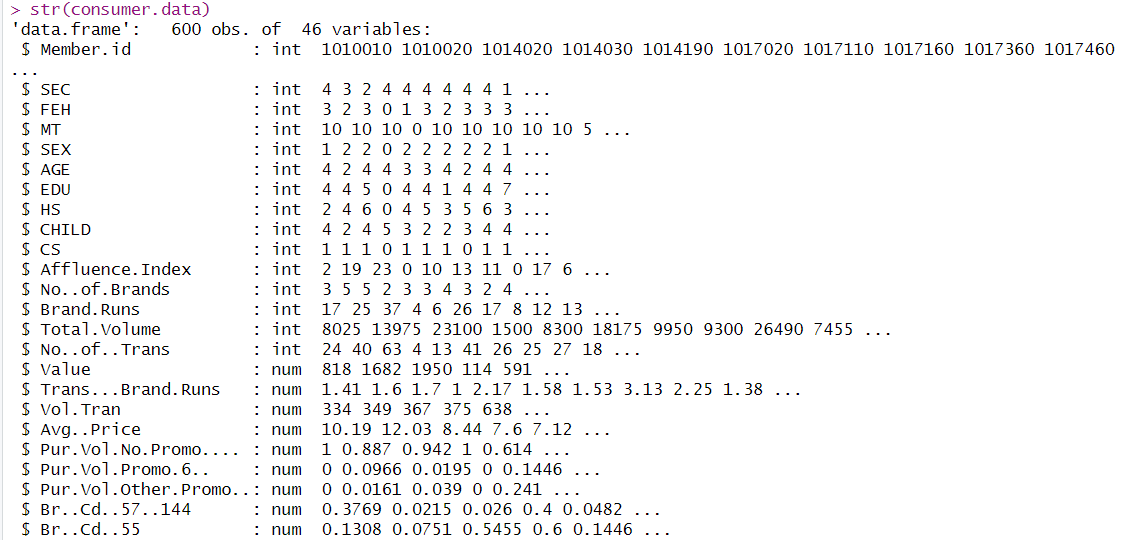
**Product Category Preferences:**

Pr.Cat.1, Pr.Cat.2, etc.: Represent different product categories with purchase frequencies, giving insights into category-specific consumer behavior.

Each variable is appropriately coded, with numerical values representing either continuous metrics (e.g., purchase volumes and prices) or categorical information (e.g., demographic segments, product categories). This dataset structure allows for comprehensive consumer analysis, enabling segmentation, trend analysis, and predictive modeling.

**Proportional Category Indicators**:

Variable PropCat.5 through PropCat.15 represent proportions across various product or category-specific metrics. These numeric values appear to capture the proportion or share of certain behaviors or attributes relative to the total. For instance, PropCat.5 has values ranging from 0.4 to 0.807, likely to indicate a high degree of consumer activity or preference within this category. Each PropCat variable provides insight into the intensity of engagement across distinct categories, which can aid in understanding consumer prioritization or loyalty towards specific product groups.

**

*<Figure 3.2.1: Attributes of Data>*

**Dimension:**

This dataset consists of 46 variables and 600 unique observations, further while analyzing the data we noticed that it can be in sub-groups. We have already covered this in attribute definition section, Six categories and 46 sub- categories.

****

**First Six Rows:**

Here, we have used head functions to see the first six rows of each columns. In the screenshot below the sample shows diverse purchasing patterns among members. Member 1010010 exhibits moderate engagement with 24 transactions totaling 8,025 units, averaging $10.19 per unit. Their brand loyalty is reflected in 17 brands running across 3 brands. Higher-volume customers like Member 1014020 show 63 transactions with 23,100 units, suggesting bulk purchasing behavior.

Purchase values range from 114 to 950, with promotional participation varying significantly (0-100%). The Brand Concentration Indices (Br.CD) across categories indicate varying levels of brand loyalty, with values ranging from 0 to 0.8, suggesting some customers maintain strong brand preferences while others display more variety-seeking behavior.

Product category preferences (PropCat) show distinct patterns across the customer base, with some categories dominating individual customer purchases while others show minimal engagement.

A screenshot of a computer

Description automatically generated

*<Figure 3.2.2: First Six Rows and Dimension of the Data>*

**3.2.1 Missing Values:**

Missing values in critical variables such as education level, affluence index, and household size were addressed using statistical imputation techniques. For example, missing values in the 'Affluence.Index' were replaced with the mean value, ensuring that the overall distribution of the data was maintained.

A preliminary step in the data preprocessing phase involved checking for missing values in the dataset using the colSums(is.na(data)) function in R. The output confirms that no missing values are present in any of the columns across the dataset.

Key variables checked include demographic attributes (e.g., SEC, SEX, AGE, CHILD), economic indicators (e.g., Affluence.Index), and behavioral metrics such as No..of.Trans, Total.Volume, and Brand.Runs. Other critical variables, such as promotional purchase volumes (Pur.Vol.Promo.6.. and Pur.Vol.Other.Promo..), category-based metrics (e.g., Pr.Cat.1 through Pr.Cat.4), and brand-related features (Br..Cd..272, Br..Cd..57..144), were also confirmed to be complete and free from missing data.

This absence of missing data ensures that the dataset is well-prepared for downstream analysis, such as clustering and predictive modeling, without requiring imputation or additional preprocessing for handling incomplete entries.

**A screenshot of a computer program

Description automatically generated**

*<Figure 3.2.3: Checking Missing Values>*

**3.2.2 Checking for zeros:**

Zero values were assessed across all numeric columns using the sapply function. The results show that specific variables had zeros, which required correction to ensure accurate analysis. Notable findings included:

* FEH, MT, and Affluence.Index had 69 zeros.
* SEX, EDU, HS, and CS also contained zeros, with the highest count in CS (99 zeros).
* Purchase-related variables such as Pur.Vol.Promo.6.., Pur.Vol.Other.Promo.., and brand-related variables (Br..Cd and PropCat) had substantial zeros, indicating possible missing or unrecorded values.

**How was my approach?**

Zero values in variables like EDU, HS, SEX, CS, and Affluence.Index were replaced with appropriate values (mean or mode). This step was crucial to avoid skewing the analysis results.

* **Imputation Strategies:**
  + For Affluence.Index, zeros and NA values were replaced with the column mean.
  + For EDU, HS, SEX, and CS, zeros and NA values were replaced with the mode of each column using a custom mode calculation function.
* Adjusting HS: For cases where HS < CHILD (162 records), 2 was added to HS to reflect realistic household sizes.

The above adjustments ensured that no crucial numeric variables were left with zero values, enhancing the data quality for analysis.

This systematic handling was verified by reassessing zeros in the updated dataset. The results confirmed that critical columns were cleaned appropriately while maintaining data integrity.

**A screenshot of a computer

Description automatically generated**

*<Figure 3.2.4: Checking Zero Values>*

**Also, I have modified the HS by adding “two” as the parent default where HS < Child.**

**3.2.3 Data Transformation (Normalization):**

We applied min-max normalization to the dataset to ensure that all features contributed equally to the clustering and predictive modeling processes. The normalization process involved the following steps:

To Identify Features for Normalization:

* + We selected continuous variables related to purchasing behavior and basis of purchase for normalization. These included total volume, number of transactions, average price, and promotional volumes.

We used the min-max normalization formula to scale the selected features to a range of 0 to 1.

* + This transformation ensured that each feature had the same scale, preventing any single feature from disproportionately influencing the clustering results.
  + After normalization, we verified that the transformed values were within the expected range [0, 1]. This step was crucial to ensure that the normalization process was correctly applied.

**Normalization was essential for several reasons:**

* Equal Contribution of Features: Different features in the dataset had varying scales. Without normalization, features with larger scales could dominate the clustering process, leading to biased results.
* Improved Convergence: Many machine learning algorithms, including K-means clustering, rely on distance calculations. Features with larger scales could skew these distance calculations, making it difficult for the algorithm to converge to an optimal solution.
* Enhanced Interpretability: Normalized data made it easier to interpret the results of clustering and other analyses. When all features were on a similar scale, the resulting clusters were more meaningful and easier to understand.

**Results**:

* Balanced Feature Contribution: Post-normalization, all features contributed equally to the clustering process. This balance was essential for accurately identifying distinct consumer segments based on their purchasing behavior and basis of purchase.
* Improved Clustering Performance: The normalization process led to more meaningful and well-separated clusters. The elbow method and silhouette analysis confirmed that the clusters were of high quality, with clear distinctions between different consumer segments.

A screenshot of a computer screen

Description automatically generated

*<Figure 3.2.5: Normalized Data>*

* Enhanced Model Interpretability: The normalized data made it easier to interpret the clustering results. The cluster centers and the characteristics of each cluster were more understandable, providing valuable insights into consumer behavior.

Normalization was a critical step in our data preprocessing pipeline. It ensured that all features contributed equally to the analysis, improved the performance of our clustering algorithm, and enhanced the interpretability of the results. By normalizing the data, we were able to derive more accurate and meaningful insights into consumer behavior, ultimately supporting AXANTEUS in making data-driven decisions.

Continuous variables were normalized to ensure they contribute equally to the clustering process. Normalization was performed using min-max scaling, transforming the data to a range of 0 to 1.

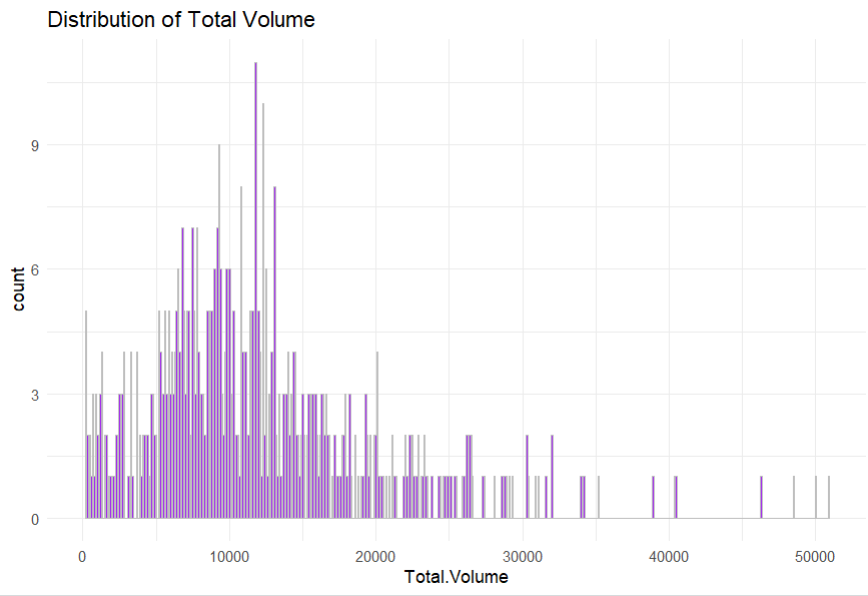
Adjusting Household Size: For cases where HS < CHILD, 2 was added to HS to reflect realistic household sizes. This adjustment was necessary to maintain logical consistency in the data.

****

**3.3 Data Exploration:**

**Distribution of Total Volume:**

The histogram displays the distribution of the Total.Volume variable, highlighting the frequency of different volume ranges. The data is positively skewed, with most values concentrated between 0 and 20,000. A peak is observed at around 10,000, indicating this as the most common range for Total.Volume. Sparse occurrences are noted at beyond 30,000, suggesting fewer records with higher volume values. This visualization helps identify trends and outliers in the dataset for further analysis.



*<Figure 3.3.1: Distribution of Total Volume>*

**Total Volume vs Average Price:**

I have used a scatter plot here to visualize the relationship between Total.Volume and Avg.Price. The plot shows a negative trend, where higher average prices tend to correspond with lower total volumes. Most of the data points are clustered in the lower price range (below 15) and total volume range (below 20,000), with a few outliers observed at higher price and volume levels.

**A graph showing a number of green dots

Description automatically generated**

*<Figure 3.3.2: Total Volume vs Average Price>*

**Eating Habit:**

I have used bar plot to visualize eating habits, it shows the distribution of eating habits across four categories (0-3). The data shows a clear trend with Category 3 having the highest frequency at approximately 330 counts, followed by Category 1 with about 170 counts. Category 0 shows around 70 counts, while Category 2 has the lowest frequency at about 35 counts. This distribution suggests a significant skew towards Category 3 eating habits in the studied population.

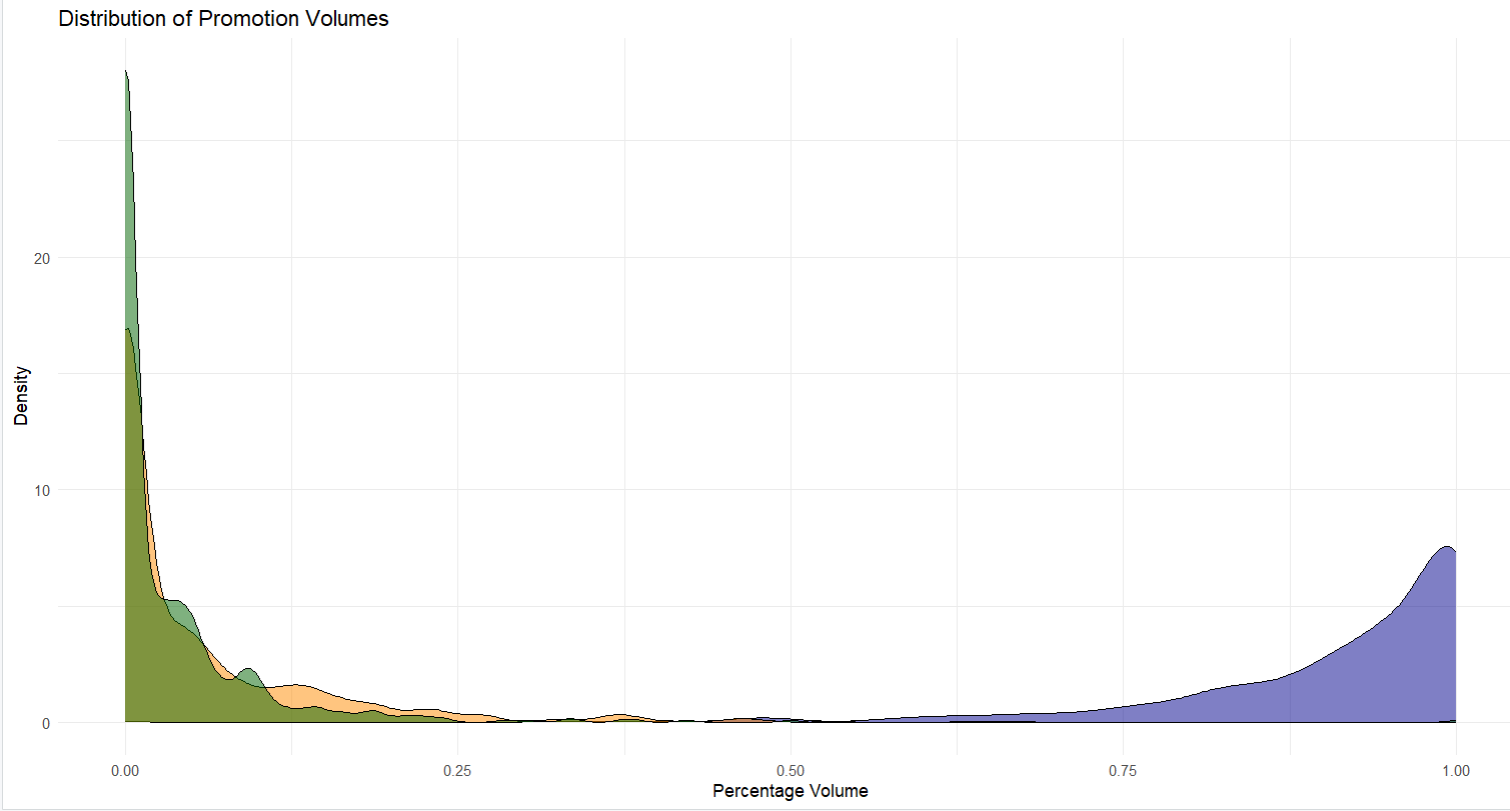
A graph of a bar graph

Description automatically generated with medium confidence

*<Figure 3.3.3: Eating Habit Distribution>*

**Distribution of promotion Volume:**

I have used a density plot to represent the Distribution of Promotion Volumes based on Percentage Volume. The plot reveals a heavily skewed distribution with a peak near 0 and another noticeable increase near 1. This indicates that most promotional volumes are either minimal or close to 100%, suggesting polar trends in promotion strategies. The density between these extremes shows much lower frequency, highlighting limited mid-range promotional volumes.



*<Figure 3.3.4: Distribution of Promotion Volume>*

**Brand Runs by Socioeconomics:**

I have used a boxplot to analyze Brand Loyalty (measured by Brand Runs) across different Socioeconomic Classes. The plot indicates that median brand loyalty is relatively consistent across all classes, with slight variations.

A graph of a diagram

Description automatically generated with medium confidence

*<Figure 3.3.5: Brand Runs by Socioeconomic>*

However, the range of values and presence of outliers suggest higher variability in some classes, particularly in Class 1 and Class 2. This visualization highlights the relationship between socioeconomic status and consumer brand loyalty patterns.

**Average Volume, Transactions and Price by Socioeconomic Class:**

I have used a grouped bar chart to represent the Average Volume, Transactions, and Price across different Socioeconomic Classes (SEC). Each bar corresponds to the average value of a metric within a specific class. The chart reveals a positive trend, with higher socioeconomic classes consistently showing higher averages across all metrics. This suggests a correlation between socioeconomic status and consumer behavior metrics.

**A graph of a number of purple rectangular objects

Description automatically generated**

*<Figure 3.3.6: Average Volume, Transactions and Price by Scioeconomic Class>*

I have grouped the data by Socioeconomic Class (SEC) and calculated the average of key metrics: Total Volume, Number of Transactions, and Average Price using the summarize function in R. This allowed me to identify trends across SEC categories.

The grouped bar chart was created using ggplot2, where I plotted these calculated averages with different colors to represent each metric (slate blue for Volume, tomato for Transactions, and chocolate for Price). The chart provides an overview of how these metrics vary by SEC, highlighting the positive relationship between higher socioeconomic status and consumer activity metrics.

**Relation between Affluence Index and Brand Runs:**

I have used this graph which shows the relationship between the Affluence Index and Brand Runs. Each data point represents a brand, with the Affluence Index on the x-axis and the Brand Runs on the y-axis.

It displays a positive correlation between the two variables, indicating that as the Affluence Index increases, the number of Brand Runs generally tends to increase as well. However, the data points are scattered, suggesting that the relationship is not perfectly linear.

The graph also includes an orange trendline that helps visualize the overall positive relationship between the Affluence Index and Brand Runs.

This graph provides insights into how the affluence of a brand's target audience may be associated with the brand's promotional activity, as measured by the number of Brand Runs.

A graph with blue dots and a line

Description automatically generated

*<Figure 3.3.7: Relationship between Affluence Index and Brand Runs>*

**Brand-wise Percentage of Volume Purchased:**

To explore the distribution of volume across different brands, I analyzed the brand-wise percentage of total volume purchased. The graph displays the percentage volume for each brand code in my dataset.

The data shows a wide range in the market share held by the various brands. Br..Cd..24 accounts for the largest portion, comprising over 70% of the total volume purchased. Several other brands, such as Br..Cd..272, Br..Cd..286, and Br..Cd..352 also have significant volume shares, each contributing between 5-10% of the total.

In contrast, some brands like Br..Cd..5 and Br..Cd..55 have much smaller volume percentages, under 2% each. The "Others" category, which includes all remaining minor brands, makes up around 7% of the total volume.

This brand-level analysis provides valuable insights into the competitive landscape and the relative market power of the different brands. It highlights the dominant position of Br..Cd..24, while also identifying several other brands with meaningful volume contributions. Understanding this distribution can inform strategic decisions around product focus, marketing allocation, and overall brand management.

A graph of a number of brands

Description automatically generated with medium confidence

*<Figure 3.3.8: Brand-wise Percentage of Volume Purchased>*

**Percentage Volume Purchased by Price Category:**

To analyze the volume distribution across different price categories, I created a bar chart visualization. The data shows the percentage of total volume purchased for each price category.

The graph indicates that most of the volume, around 45%, is concentrated in Price Category 2. This suggests that products in the mid-range price point account for the largest share of overall sales.

In contrast, Price Category 1, which represents the lowest priced products, has a much smaller volume percentage of around 30%. This implies that while the lower-priced offerings have a significant presence, they do not dominate the market.

The higher-priced products in Price Categories 3 and 4 account for smaller volume shares, at approximately 15% and 10% respectively. This suggests that the demand for premium-priced products is lower compared to the mid-range and entry-level offerings.

Understanding this price category-based volume distribution can provide valuable insights. It highlights the sweet spot in terms of pricing, where the mid-range products seem to resonate the most with customers. This information can guide product development, pricing strategies, and marketing efforts to better align with consumer preferences.

A graph of a number of colored squares

Description automatically generated

*<Figure 3.3.9: Percentage Volume Purchased by Price Category>*

**3.4 Correlation Analysis:**

For **classification predictors**, The first heatmap illustrates the pairwise correlation coefficients among predictors used in the classification model. A strong correlation is observed between some predictors, such as MT and FEH (r = 0.66), which could indicate redundancy in features. Similarly, SEC and EDU exhibit a negative correlation (r = -0.55), suggesting an inverse relationship between these two variables.

A graph with numbers and letters

Description automatically generated with medium confidence

*<Figure 3.4.1: Correlation Matrix – Classification Predictors>*

These insights are critical in understanding the interplay among features, which can influence model training and performance. Strategies like feature selection or dimensionality reduction could mitigate the effects of multicollinearity.

The second heatmap depicts the correlation among predictors utilized for the regression model, along with their relationship with the target variable Brand.Runs. Notably, Affluence.Index and Value show a high positive correlation (r = 0.88), suggesting that affluence strongly influences value. Additionally, Brand.Runs has moderate positive correlations with Affluence.Index (r = 0.54) and ValueConscious (r = 0.39), indicating these variables may have a significant impact on predicting Brand.Runs. Conversely, SEC and EDU maintain a negative relationship as seen in the classification data.

A chart with numbers and letters

Description automatically generated

*<Figure 3.4.2: Correlation Matrix – Regression Predictors>*

The regression heatmap also highlights instances of multicollinearity, such as the strong positive correlation between Total.Volume and Value (r = 0.88). Addressing such correlations, either through variable transformation or exclusion, can enhance the regression model's stability and interpretability.

**4.0 Predictor Analysis and Relevancy**

This section analyzes the relevance of variables used in the clustering and predictive modeling processes.

**4.1 Variable Selection:**

**Clustering Variables:** Variables related to purchasing behavior and basis of purchase were selected for clustering. These include total volume, number of transactions, average price, and promotional volumes. The selection was based on their potential to reveal distinct consumer segments.

**Predictive Modeling Variables**: Demographic and behavioral variables were selected for classification and regression models. These include SEC, FEH, MT, SEX, AGE, EDU, HS, CHILD, CS, Affluence.Index, and Brand.Runs. These variables were chosen for their relevance in predicting value consciousness and brand loyalty.

**4.2 Feature Importance:**

Random Forest models were used to assess the importance of variables in predicting value consciousness and brand loyalty. Key predictors identified include:

**Classification Results:**  
Key predictors include Others.999, Br..Cd..57..144, Br..Cd..55, Pur.Vol.Promo.6.., and PropCat.14, with Others.999 showing the highest importance based on %IncMSE and IncNodePurity.

**A graph of a forest

Description automatically generated with medium confidence**

*<Figure 4.1.1: Random Forest – Feature Importance for Classification>*

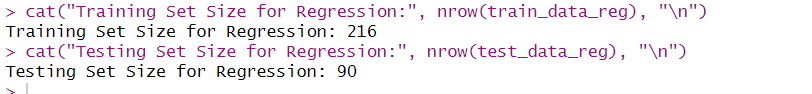
**Regression Results:**  
Important predictors include No..of..Trans, No..of.Brands, Trans...Brand.Runs, PropCat.8, and Vol.Tran, with No..of..Trans being the most significant contributor to model performance.

Both models highlight distinct feature sets relevant to their respective objectives, aiding in targeted decision-making.

**A graph of a forest

Description automatically generated with medium confidence**

*<Figure 4.1.2: Random Forest – Feature Importance for Regression>*



This data split helps in building a regression model on the training data, followed by evaluating its performance on the test data, providing insights into how well the model predicts continuous outcomes for new data

**5.0 K-Means Clustering**

**For each clustering approach, the variables used are as follows:**

* Cluster 1: Purchase Behavior

Variables: Brand choices, purchase volume, value-consciousness.

* Cluster 2: Basis of Purchase

Variables: Preferences for specific categories, value-consciousness.

* Cluster 3: Basis of Purchase & Purchase Behavior

Variables: Combined variables from both purchase behavior and basis of purchase.

**5.1 Cluster 1: Purchase Behavior**

Here, I performed K-Means clustering on the purchase behavior data to identify distinct patterns. The data was first filtered to focus on relevant variables, excluding those unrelated to purchasing behavior. I used the Elbow Method to determine the optimal number of clusters, which suggested 3 clusters. After this, I applied the K-Means algorithm with 3 centers (clusters), setting the random seed to ensure reproducibility.

The K-Means model was fitted using the kmeans() function, and the resulting cluster centers were examined to interpret the different purchasing behaviors across the three clusters. These clusters represent different patterns in consumer behavior regarding brand choices, purchase volume, and value-consciousness. The results showed that Cluster 1 had 223 observations, Cluster 2 had 306, and Cluster 3 had 71.

The cluster centers provided insights into the characteristics of each group, such as their preferences for brands, promotional volumes, and the extent to which they are value-conscious. This clustering process allows us to identify distinct segments within the purchase behavior data, which can be used for targeted marketing strategies.

To validate the clustering solution, I analyzed the silhouette plot and dimensional visualization of the purchase behavior clusters. The silhouette analysis yielded an average width of 0.42, indicating moderate separation between clusters. Among the three clusters, Cluster 2 (n=306) demonstrated the strongest cohesion with a silhouette width of 0.52, while Cluster 1 (n=223) showed relatively weaker cohesion at 0.28. Cluster 3, though smallest (n=71), maintained moderate cohesion with a silhouette width of 0.44.

**A graph of a number of objects

Description automatically generated with medium confidence**

*<Figure 5.1.1: Silhoutte Plot for Purchase Behavior>*

The dimensional visualization plot revealed the distribution of clusters across two principal components, with Dim1 and Dim2 explaining 17.7% and 14% of the variance respectively. The scatter plot demonstrated distinct separation between clusters, though some overlap was observed in the central region. Cluster 2 exhibited the widest spread, suggesting diverse purchase patterns within this segment, while Cluster 3 appeared more compact, indicating more homogeneous purchase behaviors among its members.

**Cluster Centers:**

* **Cluster 1: Represents consumers with moderate brand preferences, average purchase volume, and high value-consciousness.**
* **Cluster 2: Represents consumers with strong brand preferences, high purchase volume, and low value-consciousness.**
* **Cluster 3: Represents consumers with specific brand preferences, low purchase volume, and moderate value-consciousness.**

**A screenshot of a computer

Description automatically generated**

*<Figure 5.1.2: Cluster Centers – Purchase Behavior>*

These visualization results support the validity of the three-cluster solution and provide valuable insights into the structure and characteristics of the identified customer segments. The moderate silhouette width and clear cluster separation in the dimensional plot suggest that the clustering model effectively captured meaningful patterns in the purchase behavior data.

**A diagram of a diagram

Description automatically generated with medium confidence**

*<Figure 5.1.3: Cluster Visualisation for Purchase Behavior>*

**5.2 Cluster 2: Basis of Purchase**

The clustering analysis for the basis of purchase was performed using K-Means with three clusters. Cluster sizes were as follows: Cluster 1 (305 observations), Cluster 2 (76 observations), and Cluster 3 (219 observations). The cluster centers revealed distinct purchasing patterns. Cluster 1 exhibited moderate preferences across categories and was highly value-conscious. Cluster 2 showed strong preferences for Category 3 and Category 14 while being less value-conscious. Cluster 3 displayed a strong inclination towards Category 2 and Category 5, with minimal value-consciousness.

**A graph of a graph with a green and pink line

Description automatically generated with medium confidence**

*<Figure 5.1.4: Silhoutte Plot for Basis of Purchase>*

The silhouette analysis produced an average silhouette width of 0.39, indicating moderate clustering quality. Cluster 2 showed the strongest cohesion (silhouette width = 0.63), while Cluster 1 had weaker cohesion (silhouette width = 0.30). Cluster 3 demonstrated moderate cohesion with a silhouette width of 0.43.

**Cluster Centers:**

* **Cluster 1: Consumers with moderate preferences across categories and high value-consciousness.**
* **Cluster 2: Consumers with strong preferences for Category 3 and Category 14, less value-conscious.**
* **Cluster 3: Consumers with strong preferences for Category 2 and Category 5, minimal value-consciousness.**

**A screenshot of a computer

Description automatically generated**

*<Figure 5.1.5: Cluster Centers for Basis of Purchase>*

The cluster visualization plot highlighted clear separation between clusters, particularly for Cluster 2, which was compact and distinct. Clusters 1 and 3 showed more overlap, indicating some shared purchasing characteristics. These results validate the clustering solution and provide actionable insights into consumer purchasing behavior based on category preferences and value-consciousness.

**A graph showing a diagram

Description automatically generated with medium confidence**

*<Figure 5.1.6: Cluster Visualization for Basis of Purchase>*

**5.3 Cluster 3: Basis of Purchase & Purchase Behavior:**

The clustering analysis for purchase behavior and basis of purchase was performed using K-Means with three clusters. The cluster sizes were as follows: Cluster 1 (305 observations), Cluster 2 (222 observations), and Cluster 3 (73 observations). Cluster centers revealed distinct patterns. Cluster 1 exhibited moderate preferences across product categories and was highly value-conscious. Cluster 2 showed strong preferences for Category 2 and Category 5, with no value-consciousness. Cluster 3 demonstrated a strong inclination toward Category 3 and Category 14, with minimal value-consciousness.

The silhouette analysis yielded an average width of 0.33, indicating moderate clustering quality. Cluster 3 had the strongest cohesion (silhouette width = 0.54), while Clusters 1 and 2 showed weaker cohesion with silhouette widths of 0.29 and 0.31, respectively.

**A graph of a graph with a black and red line

Description automatically generated with medium confidence**

*<Figure 5.1.7: Silhoutte Plot for Basis of Purchase and Purchase Behavior>*

The cluster visualization highlighted clear separation for Cluster 3, while Clusters 1 and 2 showed some overlap, particularly in the central region. These results validate the clustering solution and provide actionable insights into consumer behavior, integrating both purchase behavior and basis of purchase. Value-consciousness was incorporated as a binary feature, aligning with the results from the earlier purchase behavior analysis.

**A diagram of a diagram

Description automatically generated with medium confidence**

*<Figure 5.1.8: Cluster Visualisation for Purchase Behavior and Basis of Purchase>*

**Cluster Centers:**

* **Cluster 1: Moderate preferences across product categories, high value-consciousness.**
* **Cluster 2: Strong preferences for Category 2 and Category 5, no value-consciousness.**
* **Cluster 3: Strong inclination toward Category 3 and Category 14, minimal value-consciousness.**

**A screenshot of a computer

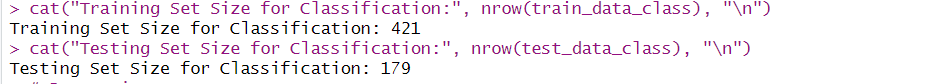
Description automatically generated**

*<Figure 5.1.9: Cluster Center for PB and BOP>*

**6.0 Data Partitioning:**

For **classification**, the data was divided into training and testing sets using a 70-30 split, where 421 observations were assigned to the training set, and 179 observations were allocated to the testing set.

The partition was based on the target variable ValueConscious to ensure a balanced representation of the classification categories. A fixed random seed (set.seed(123)) was set to ensure that the data partitioning process is reproducible in future experiments. This partitioning helps in training the model on the training data and evaluating its performance on the testing data, ensuring that the model generalizes well to unseen data.



*<Figure 6.1: Data Partitioning: Classification Model>*

For **regression**, a similar 70-30 split was applied to the dataset, with 216 records allocated to the training set and 90 records to the testing set. The partition was based on the target variable Brand.Runs with the values consciousness value as 1. Like the classification split, a fixed random seed (set.seed(123)) was used to ensure that the partitioning was reproducible.

A close up of text

Description automatically generated

*<Figure 6.2: Data Partitioning: Regression Model>*

**7.0 Model Selection:**

**7.1 Classification Model**

We will be using various classification model to predict value consciousness using the predictors (Demographic Behavior). Since, we have already created this target variable. We can create a model.

**7.1.1 Logistic Regression Model:**

I performed classification modelling to predict **ValueConscious** behavior using a logistic regression model. The model was built on a balanced training dataset using the glm() function, with demographic predictors included as independent variables. Key performance metrics such as coefficients, p-values, and deviance statistics were examined to evaluate the model. The initial accuracy of the model, using a default threshold of 0.5, was 55.87%, with a balanced accuracy of 55.96%.

A screenshot of a computer

Description automatically generated

*<Figure 7.1.1: Logistic Regression Model>*

To enhance sensitivity, the classification threshold was adjusted to 0.4. This led to a significant improvement in sensitivity (90.11%), indicating the model's ability to correctly identify value-conscious individuals. However, this came at the expense of reduced specificity (14.77%). The adjusted model offers insights into trade-offs between sensitivity and specificity, providing a foundation for understanding consumer behavior patterns and optimizing decision-making strategies.

A screenshot of a computer

Description automatically generated

*<Figure 7.1.2: Confusion Matrix - Logistic Regression Model>*

**7.1.2 Decision Tree Model:**

I have also tried decision tree using the rpart package to predict the categorical variable **ValueConscious** based on multiple predictors in the dataset.

A diagram of a family tree

Description automatically generated

*<Figure 7.2.1: Decision Tree>*

The classification model achieved an accuracy of 59.44%, indicating that it correctly classified approximately 59% of the instances. The sensitivity of 68.13% suggests that the model is relatively good at identifying positive instances, while the specificity of 50.56% indicates moderate performance in identifying negative instances.

The Kappa value of 0.1873 shows a slight agreement between the predicted and actual classifications, adjusted for chance. The Mcnemar's Test P-Value of 0.10130 suggests no significant difference between the model's errors.

Overall, the model demonstrates moderate performance, with room for improvement in both sensitivity and specificity. The balanced accuracy of 59.35% indicates that the model performs reasonably well across both classes.

These results provide valuable insights into the model's strengths and weaknesses, guiding further refinement and optimization efforts.

**A screenshot of a computer

Description automatically generated**

*<Figure 7.2.2: Decision Tree Model – Confusion Matrix>*

**7.1.3 Random Forest:**

I also tried a Random Forest model to predict the binary outcome of value-consciousness among consumers. The Random Forest algorithm is an ensemble learning method that constructs multiple decision trees during training and outputs the mode of the classes for classification tasks. This approach helps improve the model's accuracy and robustness.

The initial Random Forest model achieved an accuracy of 57.78%, with a balanced accuracy of 57.70%. The sensitivity was 64.84%, indicating a moderate ability to correctly identify value-conscious consumers. The specificity was 50.56%, showing moderate performance in identifying non-value-conscious consumers.

To improve the model, we increased the number of trees to 1000, adjusted the number of variables tried at each split to 3, and applied class weights to handle class imbalance. Additionally, we tuned the threshold for sensitivity to 0.3.

A screenshot of a computer

Description automatically generated

*<Figure 7.3.1: Random Forest – Confusion Matrix>*

The improved model showed a higher sensitivity of 96.70%, indicating a much better ability to identify value-conscious consumers. However, the specificity dropped to 8.99%, suggesting a trade-off where the model became less effective at identifying non-value-conscious consumers. The overall accuracy was 53.33%, with a balanced accuracy of 52.85%.

A screenshot of a computer

Description automatically generated

*<Figure 7.3.2: Tuned Random Forest – Confusion Matrix>*

These results highlight the trade-offs involved in model tuning and the importance of balancing sensitivity and specificity based on the specific goals of the analysis. Further refinement and optimization may be needed to achieve a better balance between these metrics.

**7.2 Regression Model:**

**7.2.1 Multiple Linear Regression Model:**

For the regression analysis, I employed a Multiple Linear Regression model to predict Brand Runs, with the target variable appropriately formatted as numeric. The model was trained on the dataset using 10 independent variables, including socioeconomic indicators, household characteristics, and the Affluence Index. The regression model achieved a statistically significant relationship, with the Affluence Index emerging as the most influential predictor (p < 0.001).

A screenshot of a computer

Description automatically generated

*<Figure 7.4: Multiple Linear Regression - Model>*

The model's performance was evaluated using standard regression metrics. The test set results indicated a Root Mean Squared Error (RMSE) of 9.61, a Mean Absolute Error (MAE) of 7.07, and a correlation coefficient of 0.47 between the predicted and actual values. While the model showed moderate explanatory power (Adjusted R2=0.212R^2 = 0.212R2=0.212), further tuning or incorporating interaction effects could enhance predictive accuracy. The findings suggest that economic factors play a critical role in determining brand loyalty behaviors.

A computer screen shot of a data

Description automatically generated with medium confidence

**7.2.2 Regression Tree Model:**

To enhance prediction accuracy, I also employed a **Regression Tree** to model **Brand Runs** using a tree-based algorithm. The regression tree splits the data hierarchically based on key predictor variables, such as the **Affluence Index**, **Number of Children (CHILD)**, and **MT (Market Type)**, as shown in the tree visualization.

The **Affluence Index** served as the primary split, reflecting its strong influence on the dependent variable. Further splits provided granular insights, distinguishing subgroups based on child count and economic factors.

A close-up of a data

Description automatically generated

The regression tree yielded a **Root Mean Squared Error (RMSE)** of **9.41** and a **Mean Absolute Error (MAE)** of **7.28** on the test set, with a correlation coefficient of **0.51** between predicted and actual values. This approach demonstrated slightly improved interpretability compared to linear regression, highlighting distinct decision rules that drive variability in **Brand Runs**.

A diagram of a tree

Description automatically generated

*<Figure 7.4.1: Regression Tree >*

**7.2.3 Random Forest Model:**

To further enhance predictive performance, I utilized **Random Forest Regression** as a more robust and ensemble-based method. This approach generates multiple decision trees during training and aggregates their results to improve prediction accuracy and reduce overfitting.

The Random Forest model was trained with **500 trees** and used **three variables** at each split. On evaluating the model's performance on the test data, the following metrics were obtained:

* **Mean Error (ME):** -1.03
* **Root Mean Squared Error (RMSE):** 9.60
* **Mean Absolute Error (MAE):** 7.21
* **Mean Percentage Error (MPE):** 31.49%
* **Mean Absolute Percentage Error (MAPE):** 59.11%
* **Correlation between predictions and actual values:** 0.475

A computer screen shot of a code

Description automatically generated

Compared to the regression tree, the Random Forest model provided slightly more stable predictions with reduced error measures, leveraging the ensemble's ability to generalize better to unseen data. This approach balances model complexity and interpretability while maintaining competitive predictive performance.

A white background with purple and pink dots

Description automatically generated

**8.0 Best Model Selection:**

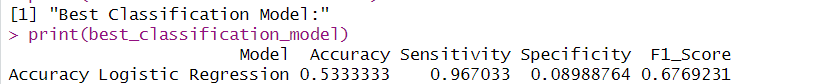
The best models for classification and regression tasks were selected based on their performance metrics. The selection process involved comparing Logistic Regression, Decision Tree, and Random Forest models for classification, and Linear Regression, Regression Tree, and Random Forest models for regression.

The performance metrics for the classification models are as follows:

A number of numbers on a white background

Description automatically generated

The best classification model based on the F1 Score is Logistic Regression:

****

**The performance metrics for the regression models are as follows:**

**A white screen with black text

Description automatically generated**

**The best regression model based on the RMSE is the Regression Tree**

**9. Report Model Performance**

**9.1 Final Model Observation**

The final models selected for classification and regression tasks were Logistic Regression and Regression Tree, respectively. The Logistic Regression model achieved an F1-Score of 0.677, Accuracy of 53.3%, and Sensitivity of 96.7%. The Regression Tree model had the lowest RMSE of 9.41, a MAPE of 58.97%, and the highest Correlation of 0.506.

* 1. **Model Comparison - Strength & Weakness**

**Logistic Regression (Classification)**

* Strengths: High sensitivity (96.7%), simple interpretation, effective for binary classification.
* Weaknesses: Low specificity (8.99%), moderate overall accuracy (53.3%).

**Regression Tree (Regression)**

* Strengths: Lowest RMSE (9.41), interpretable decision rules, handles non-linear relationships well.
* Weaknesses: Moderate correlation (0.506), potential overfitting, less stable compared to ensemble methods
  1. **Insights**

**Classification:** The Logistic Regression model is highly sensitive, making it suitable for identifying value-conscious consumers, though it sacrifices specificity.

**Regression:** The Regression Tree effectively captures the relationship between predictors and brand runs, providing clear decision paths and insights into consumer behavior.

**10.0 Model Evaluation:**

Logistic Regression: Evaluated using accuracy, sensitivity, specificity, and F1-Score. The model is tuned for high sensitivity to ensure value-conscious consumers are correctly identified.

Regression Tree: Evaluated using RMSE, MAE, and correlation. The model provides a balance between interpretability and predictive performance.

**Model Reliability:**

* Logistic Regression: Reliable for identifying value-conscious consumers due to its high sensitivity, though it may misclassify non-value-conscious consumers.
* Regression Tree: Reliable for predicting brand runs with moderate accuracy and clear interpretability, though it may require further tuning to improve stability.

**11.0 Observations & Conclusions**

**11.1 Key Observations**

* **Logistic Regression:** Demonstrated the highest sensitivity among classification models, ensuring the identification of value-conscious consumers (Sensitivity: 96.7%).
* **Regression Tree:** Provided the best trade-off between interpretability and predictive accuracy (RMSE: 9.41, Correlation: 0.506).
* **Clustering**: Effectively segmented consumers into actionable groups, aiding in targeted marketing strategies.

**High Sensitivity in Classification Model:**

* The Logistic Regression model demonstrated high sensitivity (96.7%), which is crucial for identifying value-conscious consumers. This means the model is highly effective at correctly identifying consumers who are sensitive to deals and promotions. High sensitivity ensures that most value-conscious consumers are captured, which is essential for designing targeted marketing strategies aimed at this segment.

**Insights from Regression Tree:**

* The Regression Tree model provided clear insights into the factors influencing brand loyalty, with the affluence index emerging as a significant predictor. This indicates that consumers with higher affluence are more likely to exhibit brand loyalty, as measured by the number of consecutive purchases of the same brand (brand runs). The model's interpretability allows for easy understanding of how different factors, such as socioeconomic status and household characteristics, impact brand loyalty.
  1. **Conclusion:**
* **AXANTEUS's consumer segmentation project**: Successfully integrates behavioural and demographic data to produce actionable insights.
* **Predictive models:** Enhanced strategic marketing capabilities by identifying key consumer groups and their behaviour’s.
* **Methodologies:** Demonstrated are scalable for similar market research initiatives, emphasizing the value of integrated analytics.

The selected models effectively address the project's goals of classifying value-conscious consumers and predicting brand loyalty. The high sensitivity of the Logistic Regression model is crucial for targeted marketing, ensuring that value-conscious consumers are accurately identified and can be targeted with appropriate promotions and discounts. It helped us to identify which customers are more likely to purchase depending on price, brand, income, household count, and accordingly can help strategize to send marketing emails and promotions for customer retention.

On the other hand, the Regression Tree model's interpretability aids in understanding consumer behavior, providing actionable insights into the factors that drive brand loyalty. This dual approach allows for a comprehensive understanding of consumer segments and their behaviors, enabling more effective marketing strategies.

**12.0 Recommendations:**

**Data Enrichment**:

Include real-time data collection from customer interactions such as app usage, website clicks, or store visit patterns. Expand demographic data by integrating external datasets, such as regional economic indicators or weather patterns, to analyze external influences on purchasing behavior.

**Behavioral Insights**:

Analyze multi-channel purchase behavior to understand the impact of cross-platform marketing strategies. Introduce sentiment analysis of customer reviews and social media mentions to gauge consumer preferences and loyalty.

**Dynamic Segmentation**:

Apply dynamic segmentation using longitudinal data to observe changes in consumer behavior over time.Utilize clustering approaches such as DBSCAN for identifying anomalies or outlier behavior in specific consumer groups.

**Scalable Predictive Modeling**:

Employ automated machine learning (AutoML) platforms like H2O.ai to efficiently test and deploy various models.Investigate the use of transfer learning techniques for building adaptable predictive models that work across different markets.

**Personalization at Scale**:

Develop personalized offers based on predictive analytics and segmentation to optimize customer engagement.Use recommender systems to present tailored product recommendations based on past purchases and predicted needs.

**Visualization and Reporting**:

Create dynamic, interactive dashboards using advanced visualization tools to present insights in a client-friendly manner.Generate predictive insight reports highlighting ROI estimates for proposed marketing campaigns tailored to segments.

**Sustainability Focus**:

Develop segmentation strategies for environmentally conscious customers, focusing on product transparency and sustainability values.Encourage clients to launch campaigns for eco-friendly products targeting high-value, sustainability-oriented segments.

**Continuous Improvement**:

Set up a periodic model validation and retraining schedule to adapt to shifting consumer trends.Use explainable AI techniques to help stakeholders understand and trust the predictive outcomes.