# An In-Depth Data Mining Approach Using SEMMA for Retail Marketing Campaign Analysis

**Chayan Shah**

**25th October 2024**


## Abstract

This study demonstrates a structured approach to data mining using the SEMMA (Sample, Explore, Modify, Model, Assess) methodology, applied to a real-world retail Superstore marketing campaign dataset. A decision tree classifier was employed as the primary machine learning model, with the confusion matrix serving as a key performance metric. The analysis aims to enhance customer response predictions by uncovering insights into customer purchasing behaviours and the overall effectiveness of marketing strategies. Visualizations, including income distribution, response rates, and recency of purchases, provide additional context to validate model performance and highlight key patterns within the customer data.


## 1. Introduction

In today's competitive retail environment, understanding customer preferences and behaviour is essential for designing effective marketing campaigns. The Superstore dataset presents a unique opportunity to apply data mining methods to examine customer responses and improve campaign strategies. Given the complex interplay of demographics, purchasing habits, and campaign engagement, a methodical approach is required.

The SEMMA methodology, developed by the SAS Institute, provides a clear and logical framework for handling data mining tasks. Divided into five phases—Sample, Explore, Modify, Model, and Assess—SEMMA facilitates a systematic approach to predictive modelling. In this study, we focus on applying a decision tree classifier and evaluating its effectiveness through the confusion matrix, with supporting exploratory visualizations to interpret patterns within the customer data.

## 2. Dataset Description

The dataset used in this study comprises 2,240 records and 22 attributes, each detailing various aspects of customer demographics, purchasing behaviour, and marketing campaign responses. Key features include:

- **Year of Birth**: Represents customer age, helping analyze the age distribution across responders.

- **Income**: Annual income (USD) captures customers' financial status, which may correlate with response likelihood.
- **Marital Status**: Categorical indicator of whether a customer is Married, Single, etc., aiding in demographic analysis.
- **Response**: Binary target variable marking customers who responded to the campaign (`1` for responders, `0` for non-responders).

The dataset further includes spending metrics (e.g., the amount spent on wines, fruits, and meats), which are valuable for understanding customer preferences. The main objective is to predict a customer's likelihood of responding to marketing efforts based on these features.

## 3. SEMMA Methodology

The SEMMA framework guides the data mining process, ensuring a structured approach for predictive modelling. Below, we outline each phase:

### 3.1 Sample

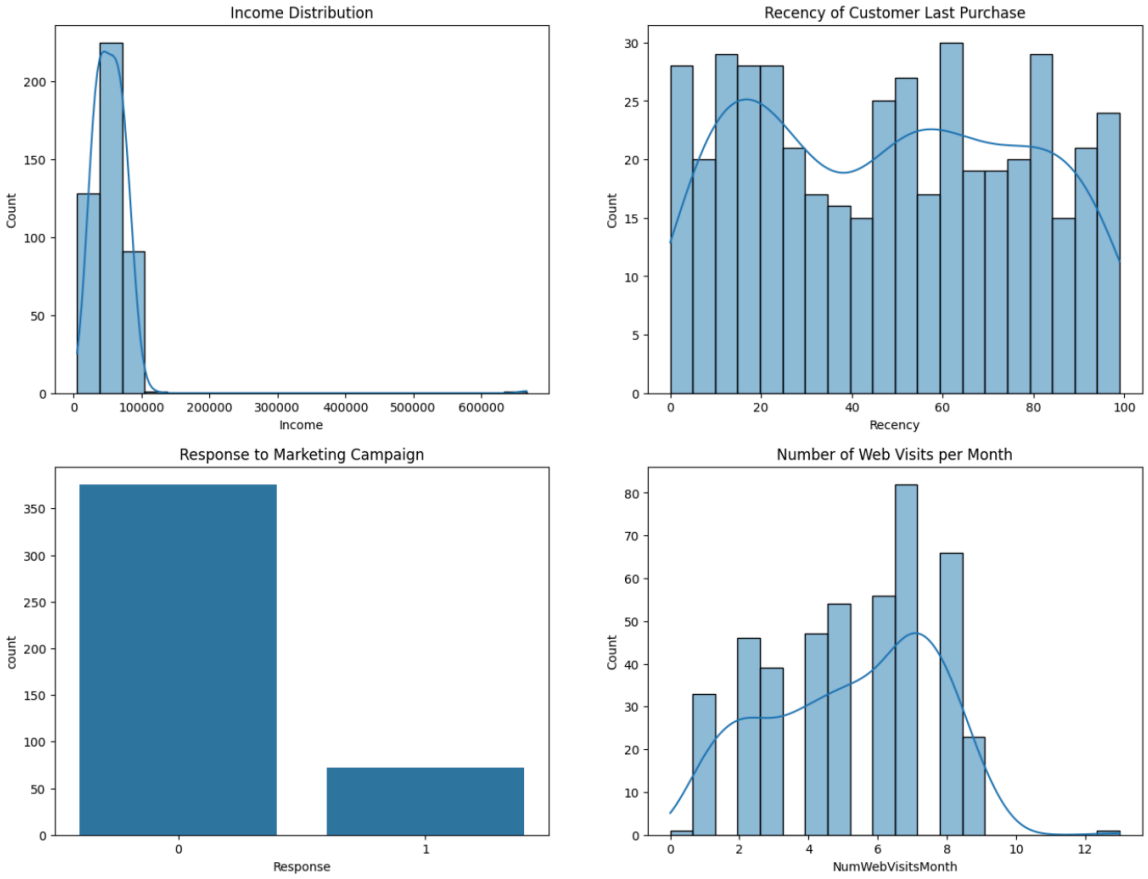
To enhance computational efficiency without losing data representativeness, a 20% sample (448 records) was extracted from the full dataset of 2,240 records using simple random sampling. This sample enabled faster iterations during model development, balancing resource efficiency and model accuracy. Sampling ensured a dataset size that allowed generalizable findings to the larger dataset.

### 3.2 Explore

The exploration phase involved analyzing statistical summaries and creating visualizations to gain insight into customer characteristics:

- **Income Distribution**: Most customers had lower incomes, with a few high-income outliers.
- **Recency of Purchase**: A large proportion of customers had recently interacted with the store, indicating strong customer engagement.
- **Response to Marketing Campaigns**: The dataset showed an imbalance, with only 6.7% of customers responding to campaigns.

These visualizations and insights were critical for guiding the model-building phase, ensuring that key patterns and imbalances were understood and addressed.




### 3.3 Modify

The modification phase involved preparing the data for modeling through several preprocessing steps:

- **Imputation of Missing Values**: Missing values in `Income` were imputed using the median, chosen due to the skewed income distribution, which would be disproportionately affected by mean imputation.
- **Normalization of Features**: Numerical features (`Income`, `Recency`, spending variables) were standardized using Z-scores, aligning their scales for model compatibility.
- **One-Hot Encoding**: Categorical features such as `Marital Status` and `Education` were converted to dummy variables, allowing models to interpret these categories without introducing bias.
- **Class Balancing**: To address the imbalance between responders and non-responders, SMOTE (Synthetic Minority Over-sampling Technique) was applied to oversample the minority class, improving model sensitivity to the positive class.

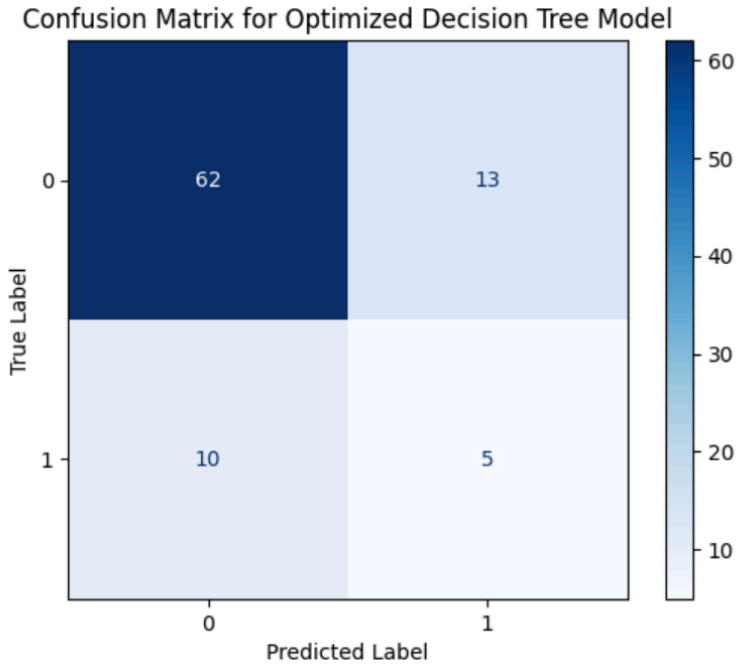
### 3.4 Model

The modeling phase aimed to predict customer responses using a decision tree classifier. Initially, a simple decision tree achieved an accuracy of 86%. However,

due to the class imbalance, the model underperformed in predicting responders, resulting in a low F1-score for the minority class. To refine the model:

- **SMOTE**: As part of the preprocessing, SMOTE was applied to generate synthetic samples for the minority class, balancing the dataset.
- **Hyperparameter Tuning**: A grid search was conducted to optimize parameters, such as `max_depth`, `min_samples_split`, and `min_samples_leaf`, enhancing model performance.

The optimized decision tree confusion matrix is shown in Figure 2, highlighting true positives, true negatives, false positives, and false negatives, offering a deeper understanding of classification accuracy.




### 3.5 Assess

The final phase assessed the model's performance using metrics derived from the confusion matrix, such as accuracy, precision, recall, and F1-score. The confusion matrix allowed us to analyze specific model errors:

- **False Positives**: Non-responders incorrectly classified as responders.
- **False Negatives**: Responders incorrectly classified as non-responders.

The F1-score served as a balanced metric for precision and recall, providing a more complete assessment of model performance with imbalanced classes.

## 4. Conclusion

This study emphasizes the value of the confusion matrix in evaluating machine learning models, particularly in imbalanced datasets. Applying the SEMMA methodology allowed us to build a predictive model capable of classifying customer

responses to marketing campaigns. While the decision tree model accurately classified non-responders, future improvements could focus on enhancing the model's predictive accuracy for responders, possibly by exploring advanced models like Gradient Boosting or XGBoost.

The SEMMA framework provided a methodical approach, enabling comprehensive data exploration, preparation, modeling, and assessment. This structure is especially useful for retail marketing analysis, as it aids in understanding customer behavior and optimizing campaign effectiveness.