# A Comprehensive Data Mining Approach Using SEMMA Methodology for Superstore Marketing Campaign Analysis

Rutuja Nemane

October 21, 2024

**Abstract**

This research paper presents a structured data mining approach using the SEMMA (Sample, Explore, Modify, Model, Assess) methodology applied to a real-world marketing campaign dataset from a retail Superstore. A decision tree classifier was utilized as the primary machine learning model, with a specific focus on its confusion matrix as an essential tool for evaluating model performance. The study aims to improve customer response predictions by revealing deeper insights into customer purchasing behavior and marketing effectiveness. Key data visualizations, such as income distribution, response rates, and recency of purchase, are included to complement the analysis and validate the model's performance.

## 1 Introduction

In the competitive landscape of retail marketing, understanding customer behavior is paramount for crafting effective campaigns. The Superstore marketing dataset provides an ideal environment to apply data mining techniques to uncover insights into customer responses and optimize marketing strategies. Given the complexity of customer behavior and diverse demographic factors, a robust data mining methodology is necessary.

The SEMMA methodology, developed by the SAS Institute, provides a structured approach for data mining tasks. By breaking the process into five distinct phases—Sample, Explore, Modify, Model, and Assess—SEMMA ensures a logical, step-by-step approach for predictive modeling. In this paper, we focus on the confusion matrix of the decision tree classifier, leveraging it as a vital metric for evaluating the performance of the predictive model. Additionally, we present several exploratory data visualizations to identify patterns and trends within the dataset.

## 2 Dataset Description

The dataset consists of 2,240 records and 22 attributes, providing a detailed view of customer demographics, purchasing history, and responses to marketing campaigns. The key features of the dataset include:

- **Year of Birth**: Customer's year of birth, providing insight into the age distribution.
- **Income**: Annual income in USD, representing the customer's financial status.
- **Marital Status**: Categorical feature indicating the customer's marital status (Married, Single, etc.).
- **Response**: Binary target variable indicating whether a customer responded to the marketing campaign (1 for responders, 0 for non-responders).

The dataset also includes various spending metrics (e.g., amounts spent on wines, fruits, meat products), which provide an understanding of customer preferences and purchasing behavior. The primary task is to predict whether a customer will respond to a marketing campaign based on these features.

## 3 SEMMA Methodology

The SEMMA methodology ensures a systematic data mining approach, which is outlined below:

## 3.1 Sample

The dataset, containing 2,240 records, was sampled down to 20% (448 records) to maintain computational efficiency while preserving data representativeness. Simple random sampling was applied to ensure unbiased selection. This sample size allowed for quicker iterations during model development without sacrificing accuracy. The primary objective of this step was to ensure that the sample was sufficient to generalize findings to the broader dataset while remaining manageable in size.

## 3.2 Explore

The exploration phase focused on understanding the dataset through statistical summaries and visualizations. Key insights into the dataset were uncovered, such as:

- **Income Distribution**: Income is skewed towards lower values, with a few high-income outliers.
- **Recency of Purchase**: Most customers had made recent purchases, indicating active engagement with the store.
- **Marketing Campaign Response**: A significant imbalance was observed, where only 6.7% of customers responded positively to the campaign.

The following visualizations were used to support the exploratory analysis:

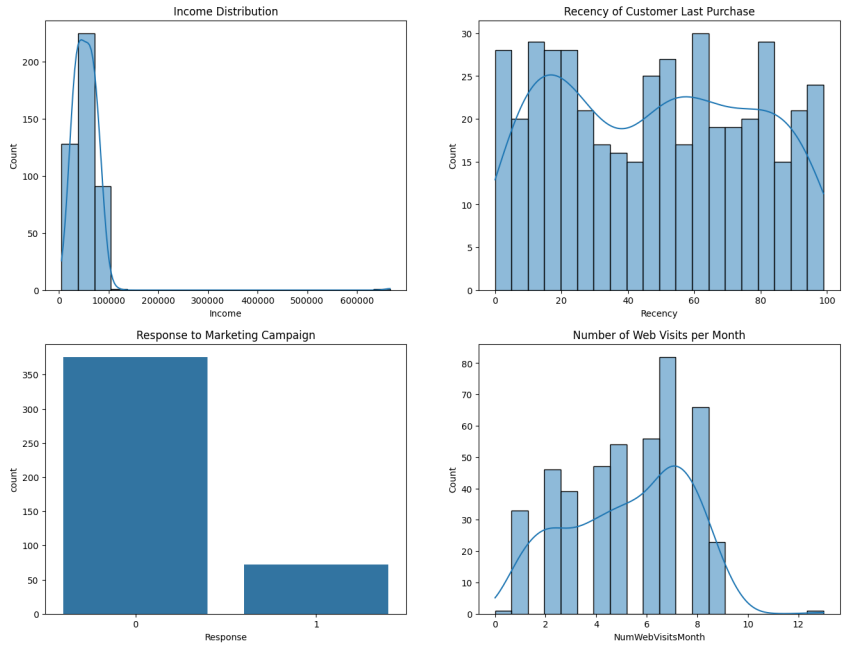


Figure 1:

These visualizations revealed important characteristics of the customer base and their purchasing behaviors, which guided the model-building phase.

## 3.3 Modify

The modification phase focused on preparing the dataset for modeling. The following steps were performed:

- **Imputation of Missing Values**: Missing values in the ‘Income‘ column were imputed using the median. This approach was chosen due to the skewness in the income distribution, which would have otherwise distorted results with mean imputation.
- **Normalization of Features**: Numerical features, including ‘Income‘, ‘Recency‘, and spending variables, were normalized using Z-scores to standardize the scales of the features.

- **One-Hot Encoding**: Categorical features such as ‘Marital Status‘ and ‘Education‘ were converted to dummy variables using one-hot encoding.
- **Class Balancing**: Since the dataset exhibited class imbalance, with significantly fewer responders, SMOTE (Synthetic Minority Over-sampling Technique) was applied to oversample the minority class (responders).

### 3.4 Model

The modeling phase aimed to predict customer responses to the marketing campaign using a decision tree classifier. Initially, a basic decision tree model was built, achieving an accuracy of 86%. However, due to the class imbalance, the model performed poorly in predicting responders, leading to a suboptimal F1-score for the minority class.

To address this, we employed the following strategies:

- **SMOTE**: As part of the preprocessing step, SMOTE was used to generate synthetic samples for the minority class (responders), thereby balancing the dataset.
- **Hyperparameter Tuning**: A grid search was conducted to optimize hyperparameters such as ‘$\max_depth$‘, ‘$min_samples_s$

The confusion matrix for the optimized decision tree model is shown in Figure 2. It provides a breakdown of the true positives, true negatives, false positives, and false negatives, highlighting the model's classification errors.

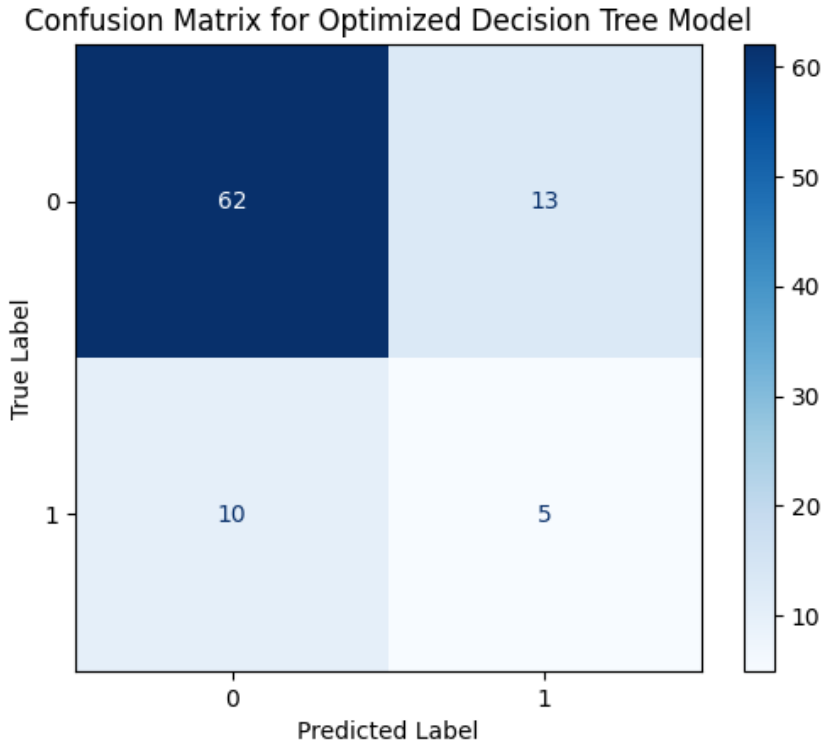


Figure 2: Confusion Matrix for Optimized Decision Tree Model

The confusion matrix indicates that the model effectively identifies non-responders, though there is room for improvement in predicting responders, with some false positives and false negatives remaining.

### 3.5 Assess

The final phase involved assessing the model's performance based on key metrics such as accuracy, precision, recall, and F1-score. These metrics, derived from the confusion matrix, provided insight into the model's ability to classify responders and non-responders. Specifically, the confusion matrix helped identify where the model made errors:

- **False Positives**: Instances where non-responders were incorrectly classified as responders.
- **False Negatives**: Instances where responders were incorrectly classified as non-responders.

The F1-score was calculated to balance the precision and recall, providing a more comprehensive measure of the model's performance in handling the imbalanced dataset.

# 4 Conclusion

This study highlights the importance of the confusion matrix as a critical tool for evaluating machine learning models, particularly when dealing with class imbalance. By applying the SEMMA methodology, we developed a predictive model capable of identifying customer responses to marketing campaigns. While the model performed well in identifying non-responders, future improvements could focus on enhancing the model's ability to predict responders accurately, perhaps by exploring more advanced models such as Gradient Boosting or XGBoost.

The SEMMA framework provided a structured approach that ensured systematic data exploration, modification, modeling, and assessment. In the context of retail marketing analytics, this approach is valuable for understanding customer behavior and optimizing campaign strategies.