Capstone Three - Predicting Customer Response to Marketing Campaigns

Executive Summary

The objective of this project is to predict customer response to a marketing campaign and identify key customer characteristics associated with successful engagement. Using historical customer, demographic, and transaction data, multiple machine learning models were developed and evaluated to determine the most effective approach for identifying likely responders.

The dataset consists of over 2,200 customers and includes demographic information, purchase behavior, prior campaign responses, and engagement metrics. Exploratory data analysis revealed that customers who responded to past campaigns tend to have higher overall spending, higher engagement with previous promotions, and fewer dependents at home.

Several classification models were evaluated, including Logistic Regression, Random Forest, and Gradient Boosting. Because only approximately 15% of customers responded to the campaign, model evaluation focused on recall and F1 score for the responder class rather than accuracy alone. A class-weighted Logistic Regression model was selected as the final model due to its superior ability to identify responders while maintaining reasonable overall performance.

Based on these findings, this project provides actionable recommendations for targeting future marketing campaigns toward customers most likely to respond, improving campaign efficiency and return on investment.

Problem Statement

Marketing campaigns are costly, and reaching customers who are unlikely to respond can significantly reduce return on investment. The goal of this project is to develop a predictive model that identifies customers who are most likely to respond to a marketing campaign, allowing marketing efforts to be better targeted and more effective.

Specifically, this project aims to answer the following questions:

- Which customer characteristics are most strongly associated with campaign response?
- Can a predictive model reliably identify likely responders despite class imbalance?
- How can these insights be used to inform future marketing strategy?

Data Overview

The dataset contains customer-level information for 2,237 individuals after data cleaning. Features include demographic attributes (age, education, marital status), household

composition, purchase behavior across multiple product categories, engagement metrics, and responses to previous marketing campaigns.

The target variable, Response, indicates whether a customer responded positively to the marketing campaign. The dataset exhibits class imbalance, with approximately 15% of customers responding. This imbalance informed both model selection and evaluation metrics.

Methodology

Data Wrangling and Feature Engineering

Missing values in income were imputed using the median. Unrealistic birth years were removed to ensure valid age calculations. Additional features such as customer age and recency in months were derived from existing variables.

Categorical variables were encoded using one-hot encoding, and numeric features were standardized to support models sensitive to feature magnitude.

Exploratory Data Analysis

Exploratory analysis focused on identifying relationships between customer behavior and campaign response. Visual analysis revealed that customers who responded tended to spend more across most product categories, particularly wine and meat products, and were more likely to have accepted previous campaigns.

Correlation analysis and comparative visualizations supported the selection of spending behavior and prior campaign engagement as key predictive features.

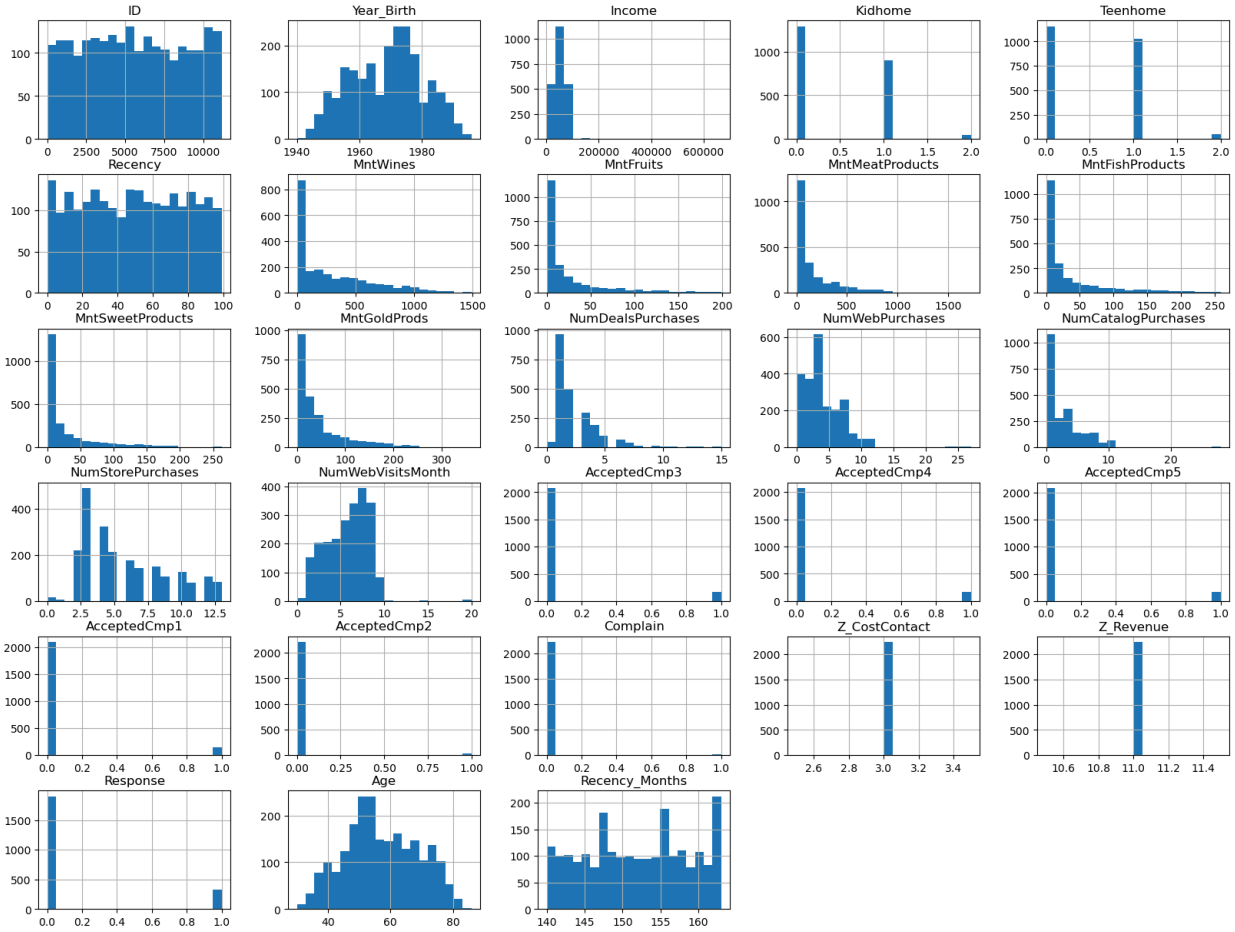

Figure 1: Numeric Feature Distributions

Histograms of all numeric variables in the dataset, showing their spread, central tendency, and potential outliers. These distributions provide insight into data variability and guide preprocessing decisions.

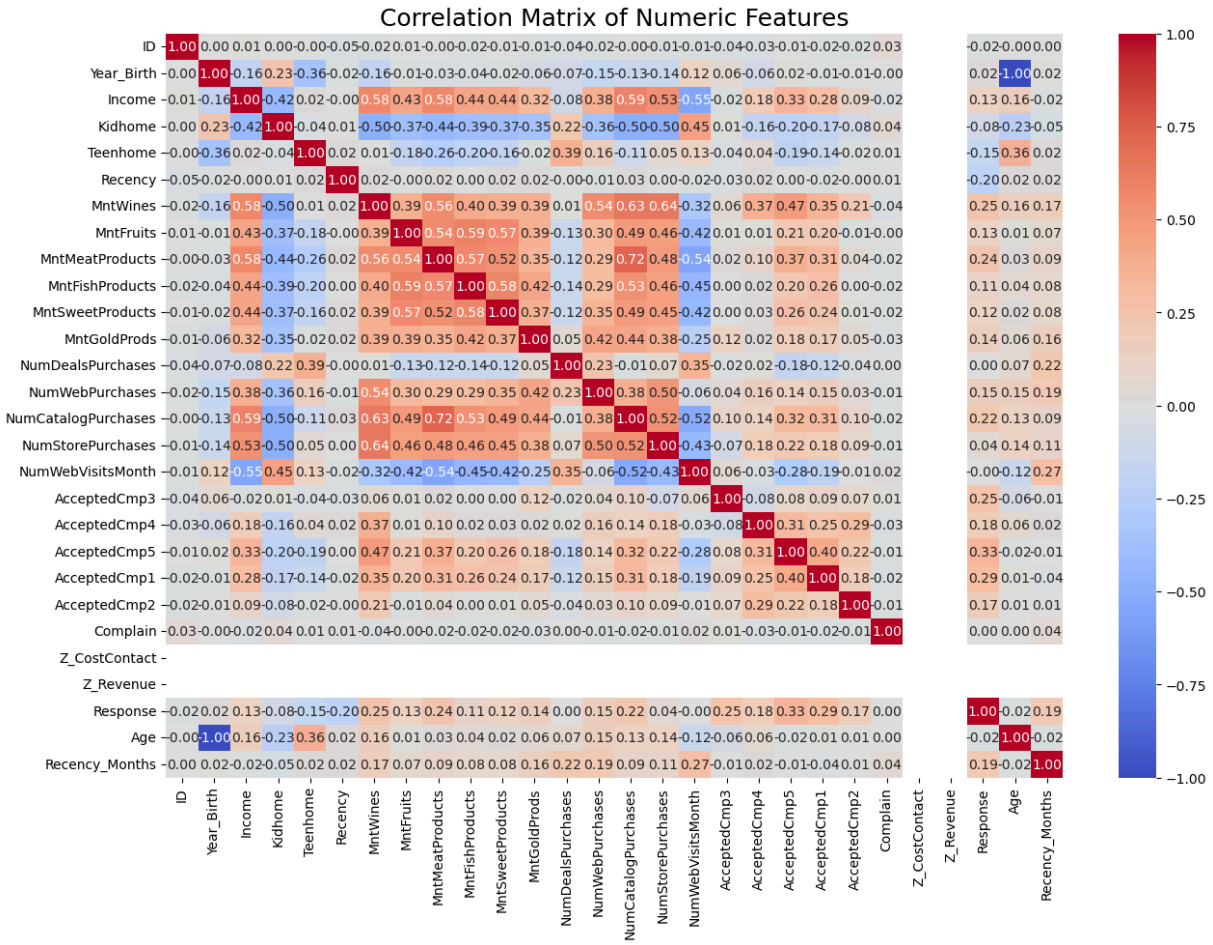

Figure 2: Correlation Matrix of Numeric Features

This heatmap displays the Pearson correlation coefficients between numeric variables in the dataset, highlighting key relationships relevant to predicting campaign response. Notably, the target variable Response shows positive correlations with customer spending features such as MntWines, MntFruits, and prior campaign acceptances (e.g., AcceptedCmp3, AcceptedCmp5). These insights guided feature selection and model development by emphasizing spending and prior engagement as important predictors.

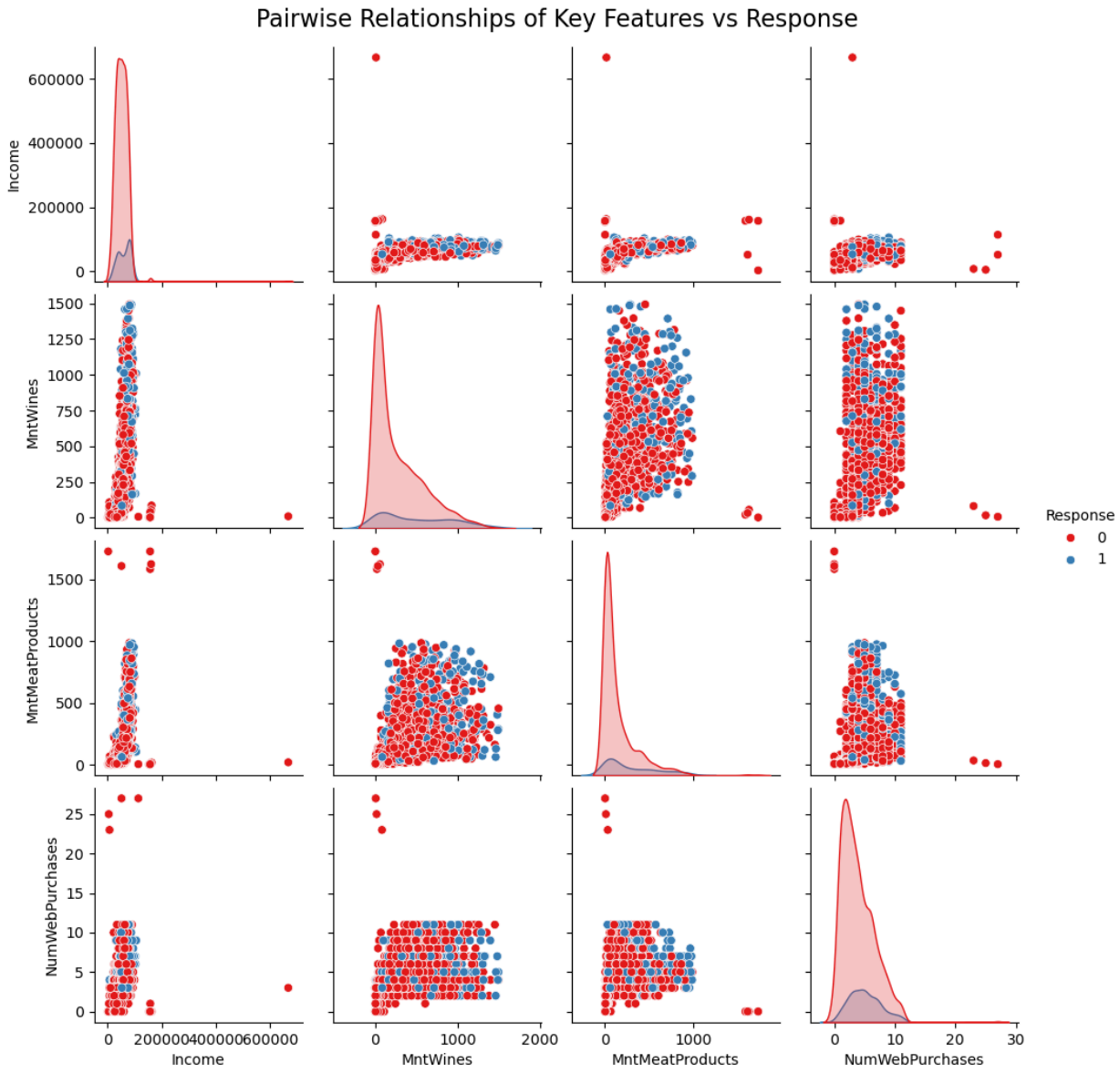

Figure 3: Pairwise Relationships Of Key Features vs Response
Income, MntWines, MntMeatProducts, and NumWebPurchases; grouped by customer response status. Scatterplots illustrate feature interactions and clustering patterns for responders (blue) and non-responders (red), while diagonal density plots show the distribution differences within each feature by response class. This visualization helps identify potential predictors and relationships relevant to modeling customer behavior.

Modeling Approach

The dataset was split into 80/20 training and testing subsets. Multiple classification models were evaluated, including Logistic Regression, Random Forest, and Gradient Boosting.

Due to class imbalance, model performance was evaluated using recall and F1 score for the responder class in addition to overall accuracy. Hyperparameter tuning was performed for the Random Forest model using cross-validation.

Results and Model Performance

While tree-based models achieved higher overall accuracy, they struggled to correctly identify responders. The class-weighted Logistic Regression model achieved the highest recall and F1 score for the responder class, making it the most suitable choice for this business problem.

Table 1: Model Performance Comparison on Test Set

Model	Accuracy	Class 1 Recall	Class 1 F1
Logistic Regression (Balanced)	0.815	0.75	0.57
Random Forest	0.877	0.36	0.49
Gradient Boosting	0.884	0.43	0.54
Random Forest (Tuned)	0.873	0.32	0.45

Comparison of four machine learning models for predicting marketing campaign response. While Random Forest and Gradient Boosting achieved higher overall accuracy, class-weighted Logistic Regression captured the largest proportion of positive responders, which was the primary objective of the campaign analysis.

Final Model Selection

Given the business objective of identifying as many likely responders as possible, the class-weighted Logistic Regression model was selected as the final model. Although its overall accuracy was lower than some alternatives, it significantly outperformed other models in identifying responders, which is critical for targeted marketing campaigns.

Recommendations

- Prioritize customers with high historical spending and prior campaign engagement, as these features were the strongest predictors of response.
- Use the predictive model to target a smaller, higher-quality audience, improving campaign efficiency and reducing marketing costs.

- Continuously retrain the model using new campaign data to adapt to changing customer behavior and improve performance over time.

Limitations and Future Work

This analysis relies on historical campaign data, which may not fully capture changing customer preferences. Additionally, the model does not incorporate real-time behavioral data or external factors such as seasonality.

Future work could explore additional feature engineering, alternative imbalance handling techniques, or more advanced models such as XGBoost. Incorporating real-time customer interactions could further improve predictive accuracy.