Enhancing Marketing Campaign Efficiency using Predictive Analytics

Agenda

Marketing Campaign Dataset

Problem Statement

- **Problem**: Artha Market faces a challenge despite consistent marketing efforts, revenue growth has stalled (from $4,225 in 2022 to $3,764 in 2023) and ROI is declining (from 4% in 2022 to -45% in 2023)

- I**ssue:** The marketing team struggles to identify customers likely to respond to future campaigns.

- I**mpact:** The company risks losing its competitive edge and wasting resources by targeting the wrong audience.

- Option were considered to solving the problem:
  1. Manual Segmentation
  2. Rule-Based Systems
  3. Machine Learning

Goal

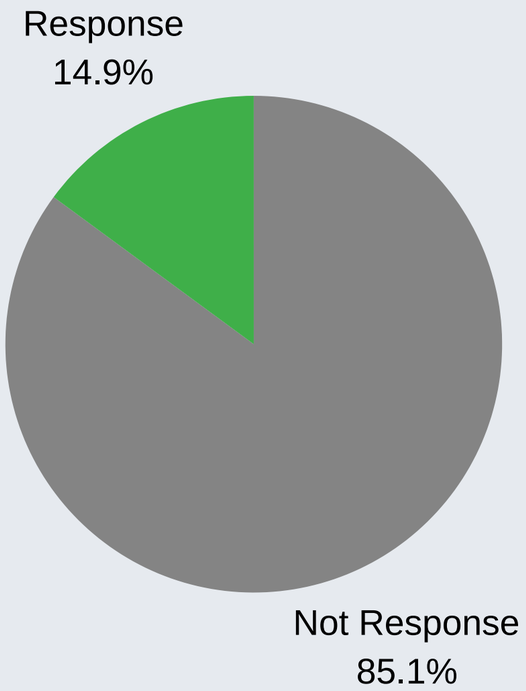
- **Why we choose Machine Learning?** Compared to manual and rule-based systems, machine learning offers greater adaptability and scalability to predict customer responses effectively.

- **Goal:** Enhance marketing efficiency by predicting who will respond using machine learning.

- This data-driven approach would enable:
  1. Targeting the Right Customers
  2. Personalizing Campaigns
  3. Optimizing Resource Allocation
  4. Measuring Campaign Effectiveness

Objective & Business Metrics

Leveraging ML models to drive better outcomes for marketing campaigns



25%

INCREASE RESPONSE RATE

Improve the response rate 25% by targeting the right customers

30%

REDUCE CAMPAIGN COST

Reduce campaign costs by 30% through reducing unnecessary spend aroun

100%

INCREASE ROI (RETURN OF INVESTMENT)

Improve ROI around 100% by focus campaign budgets on high-value customers

Dataset

29 columns, 2,240 row

Data Profile Customer	
ID	Unique identifier for each customer
Year_Birth	The year when the customer was born
Education	The level of education of the customer
Marital_Status	The marital status of the customer
Income	The annual income of the customer
Kidhome	The number of children in the customer's household
Teenhome	The number of teenagers in the customer's household
Dt_Customer	The date the customer was enrolled with the company
Recency	The number of days since the customer last made a purchase
Product Category	
MntWines	The amount spent on wine in the last 2 years
MntFruits	The amount spent on fruits in the last 2 years
MntMeatProducts	The amount spent on meat products in the last 2 years
MntFishProducts	The amount spent on fish products in the last 2 years
MntSweetProducts	The amount spent on sweet products in the last 2 years
MntGoldProds	The amount spent on gold products in the last 2 years

Amount Purchase by Purhasing Channel	
NumDealsPurchases	The number of purchases made with a discount
NumWebPurchases	The number of purchases made through the company’s website
NumCatalogPurchases	The number of purchases made using a catalog
NumStorePurchases	The number of purchases made directly in stores
NumWebVisitsMonth	The number of visits to the company’s website in the last month
Feedback from Customer	
AcceptedCmp3	1 if the customer accepted the offer in the 3rd campaign, 0 otherwise
AcceptedCmp4	1 if the customer accepted the offer in the 4th campaign, 0 otherwise
AcceptedCmp5	1 if the customer accepted the offer in the 5th campaign, 0 otherwise
AcceptedCmp1	1 if the customer accepted the offer in the 1st campaign, 0 otherwise
AcceptedCmp2	1 if the customer accepted the offer in the 2nd campaign, 0 otherwise
Complain	1 if the customer complained in the last 2 years, 0 otherwise
Z_CostContact	Cost per contact (fixed for all customers)
Z_Revenue	Revenue from the customer (fixed for all customers)
Response	1 if the customer accepted the offer in the last campaign, 0 otherwise

EDA

Exploratory Data Analysis

Descriptive Statistics

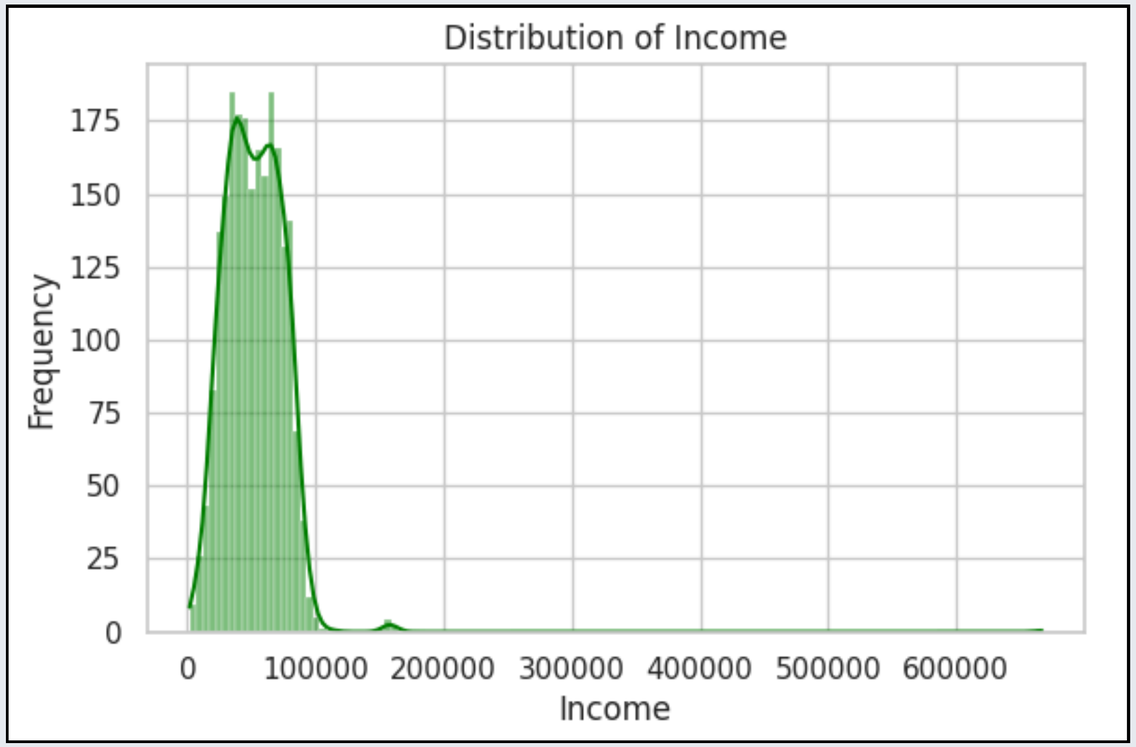
Univariate Analysis

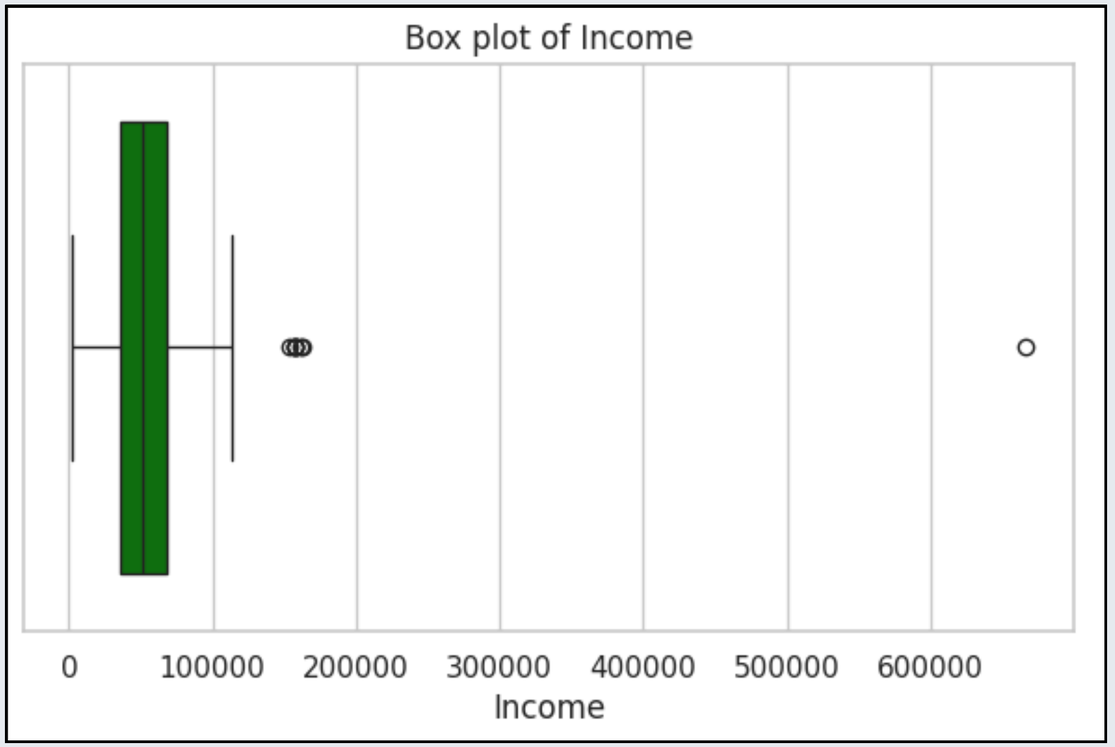
Bivariate Analysis

Data Distribution

Using histogram & boxplot

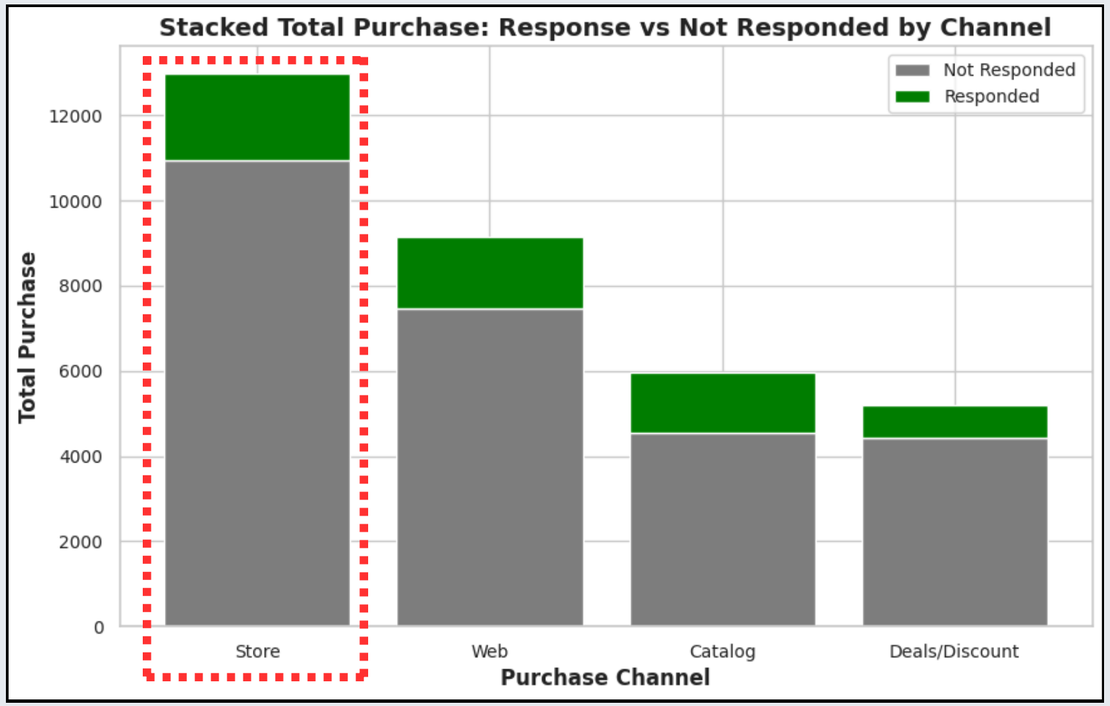
Income Distribution: Highly right-skewed distribution with most incomes clustered below 100,000 & A long tail extends to 600,000+, indicating some high-income outliers.

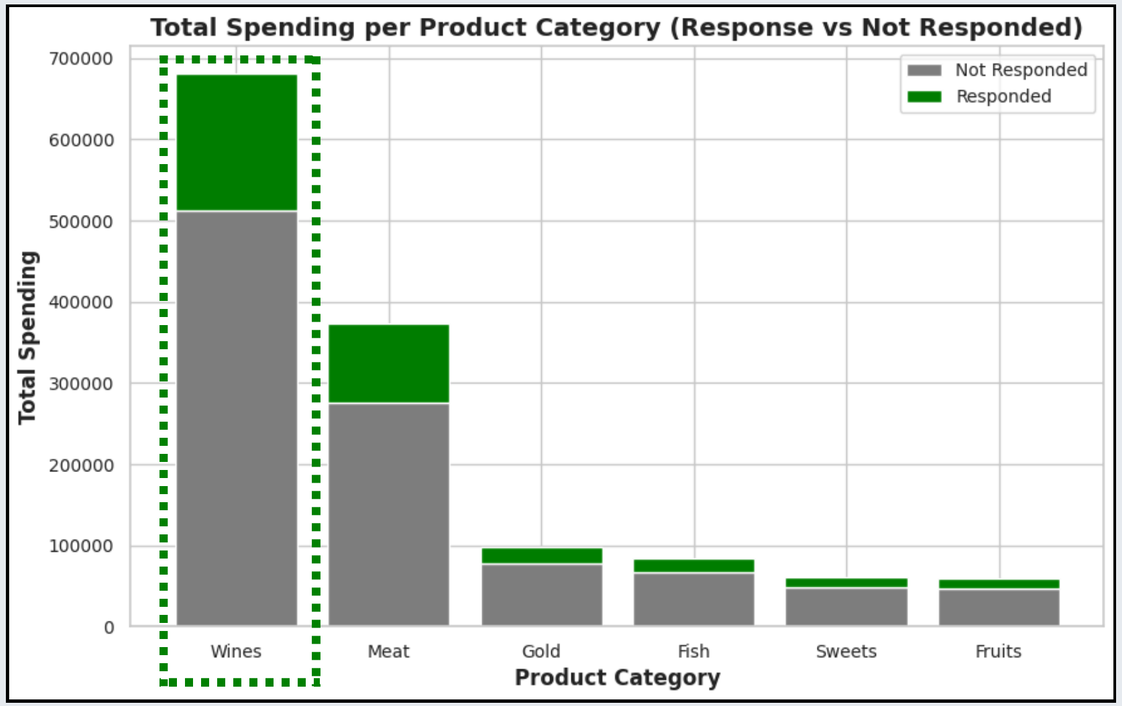




Segmentation

- Channel: The **Store channel is the most popular** among customers, while Web and Catalog have higher response rates. Deals/Discount has the lowest response rate.
- Product: **Wines is the most popular and profitable product category**, with Wines and Meat having higher response rates. Gold and Sweets have the lowest response rates.
- Deals/Discount and Gold/Sweets categories require attention to improve customer engagement and response rates.





Accepted Campaign Analysis

- The Campaign 1 and 3 have the highest overall number of deals purchases, the effectiveness of each campaign in driving customer engagement and purchases varies.
- Campaign 1 and 5 demonstrate strong response rates, suggesting they are more successful in converting potential customers into actual buyers.
- Most campaign Cmp3-5 is the majority of the campaign that delivers to deals purchases with more than 73% of deals purchases.

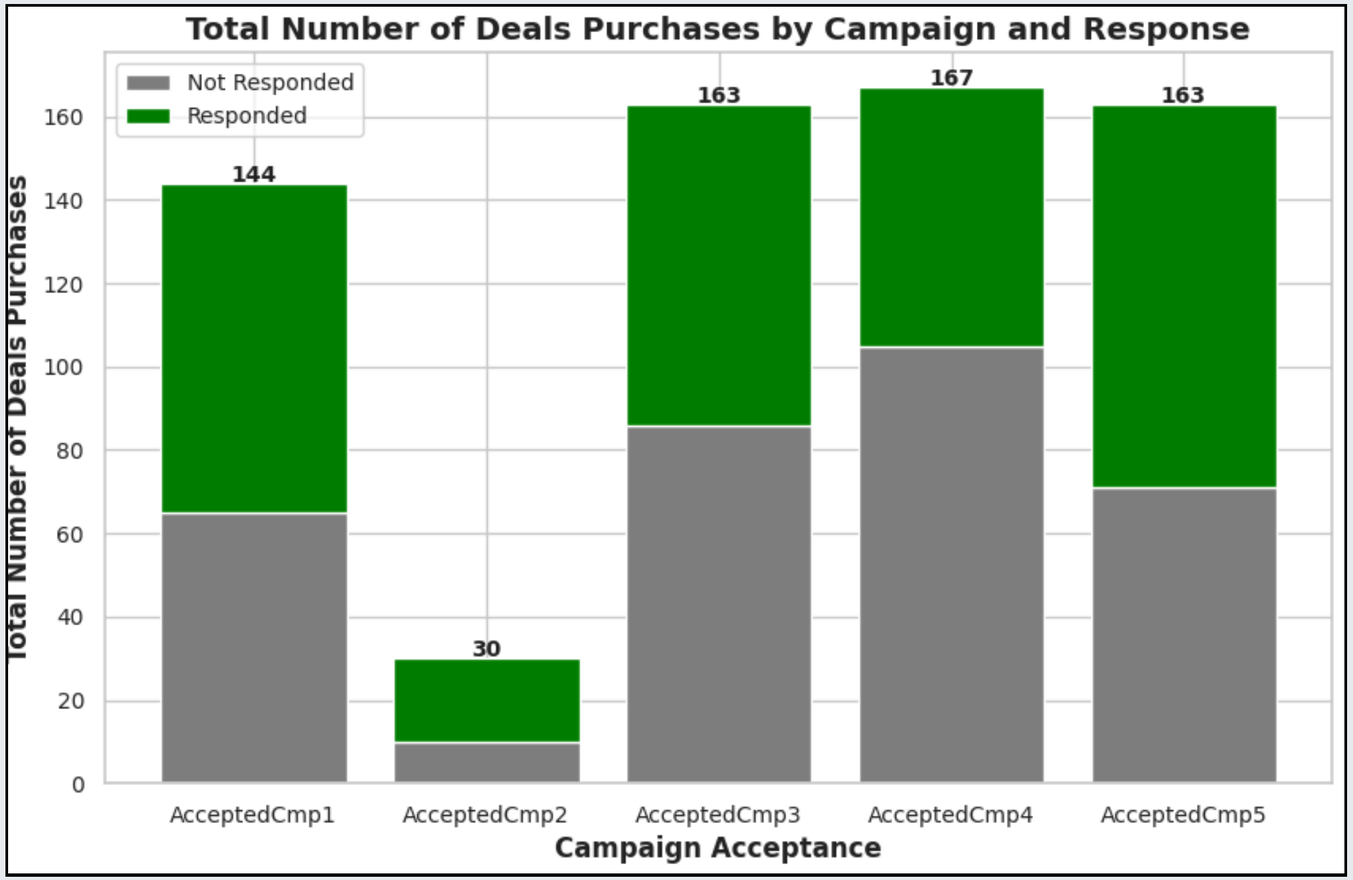
Overall Response: The majority of customers do not respond to campaigns, indicating a need for improvement in campaign effectiveness.

85.1%

Not Response

14.9%

Response



Data Pre-Processing

Mapping Values

Missing & Duplicated Values

Outliers

Mapping Values

Data Transformation for **Marital Status** & **Education** Feature

Marital Status	Mapped Marital Status	Education	Mapped Education
Single	Single	Basic	Basic
Alone	Single	Graduation	Bachelor
Married	Married	2n Cycle	Master
Together	Married	Master	Master
Divorced	Divorced	PhD	PhD
Widow	Divorced		
YOLO	Removed		
Absurd	Removed		

Mapping values is to **standardize** the data by converting inconsistent or ambiguous values into more consistent categories.

Since the value 'Absurd', 'YOLO' has an unclear meaning and only has a small number (**4 rows**), we will **remove** this data row in the next process.

Missing & Duplicated Values

Because the missing values contained are only **1.25% (28 row)**, less than 5%, we can handle it by **deleting rows** with missing values.

By checking duplicate data, we can ensure that our dataset remains clean and reliable. This step is crucial for maintaining data integrity, as duplicates can lead to incorrect analyses and insights.

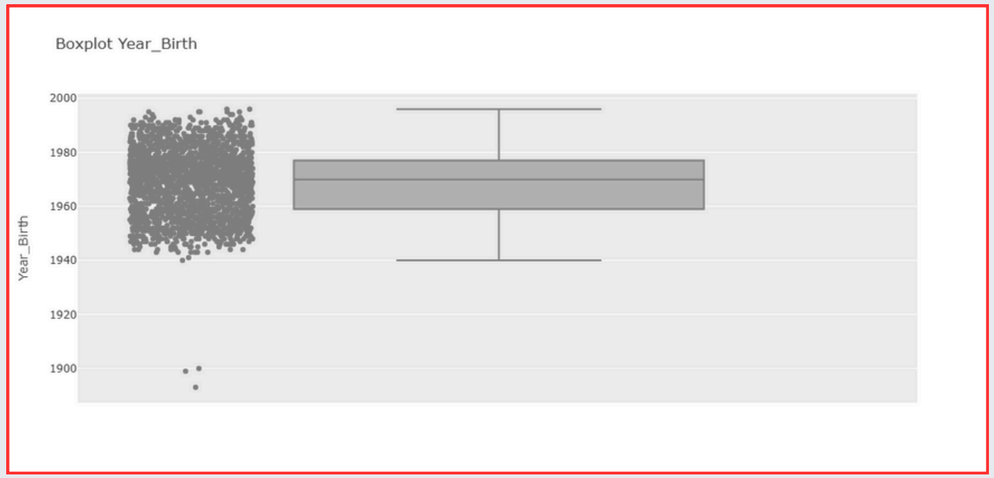
Feature	Missing Count	Missing Percentage
Marital Status	4	0.18%
Income	24	1.07%

No Duplicate Rows

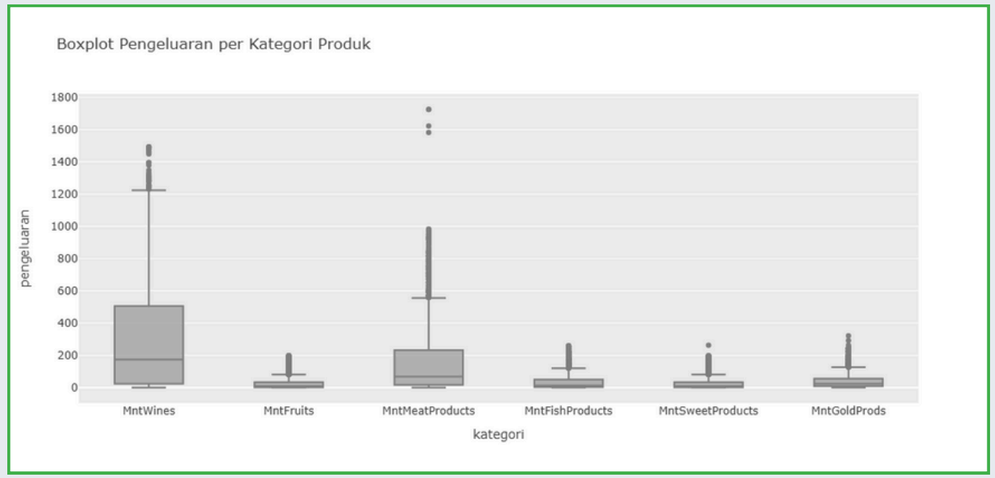
Outliers

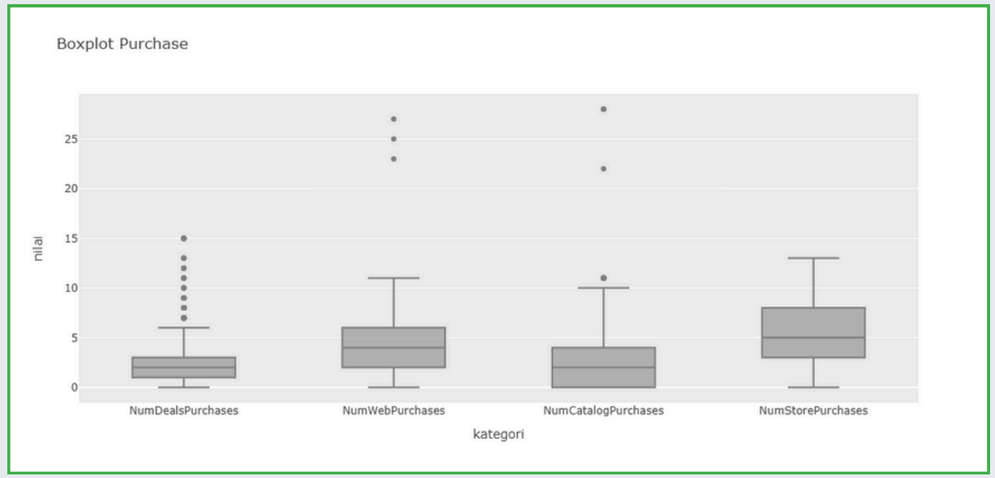
The number of outliers contained in the entire data is 1216, which is more than 50% of the existing data, so special attention is needed, outliers in the ‘income’ and ‘year_birth’ columns are **removed**.Then the other outliers is handling by **scaling**, because handle outliers so as not to damage the existing data.

Remove Outliers



Scaling





Feature Engineering

Feature Selection

Feature Extraction

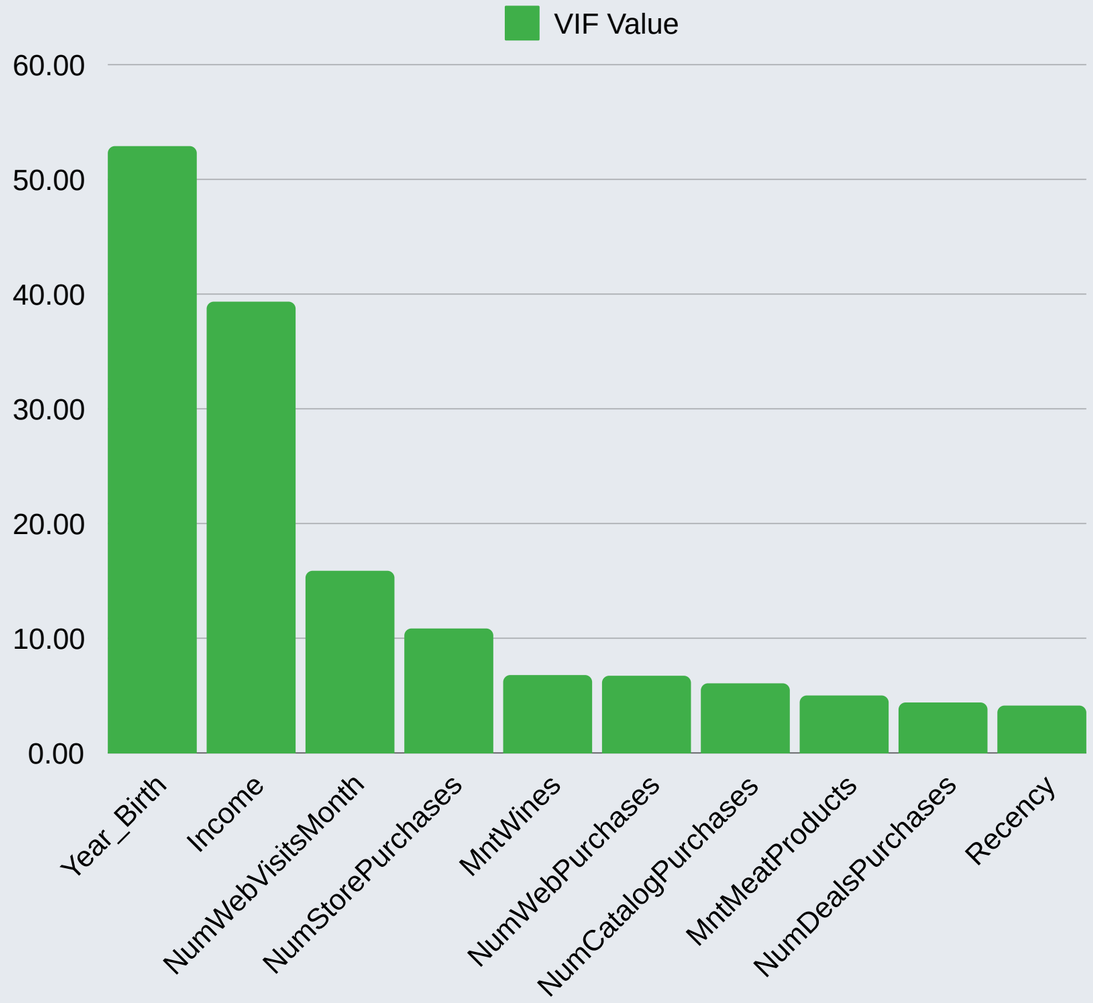
Additional Feature

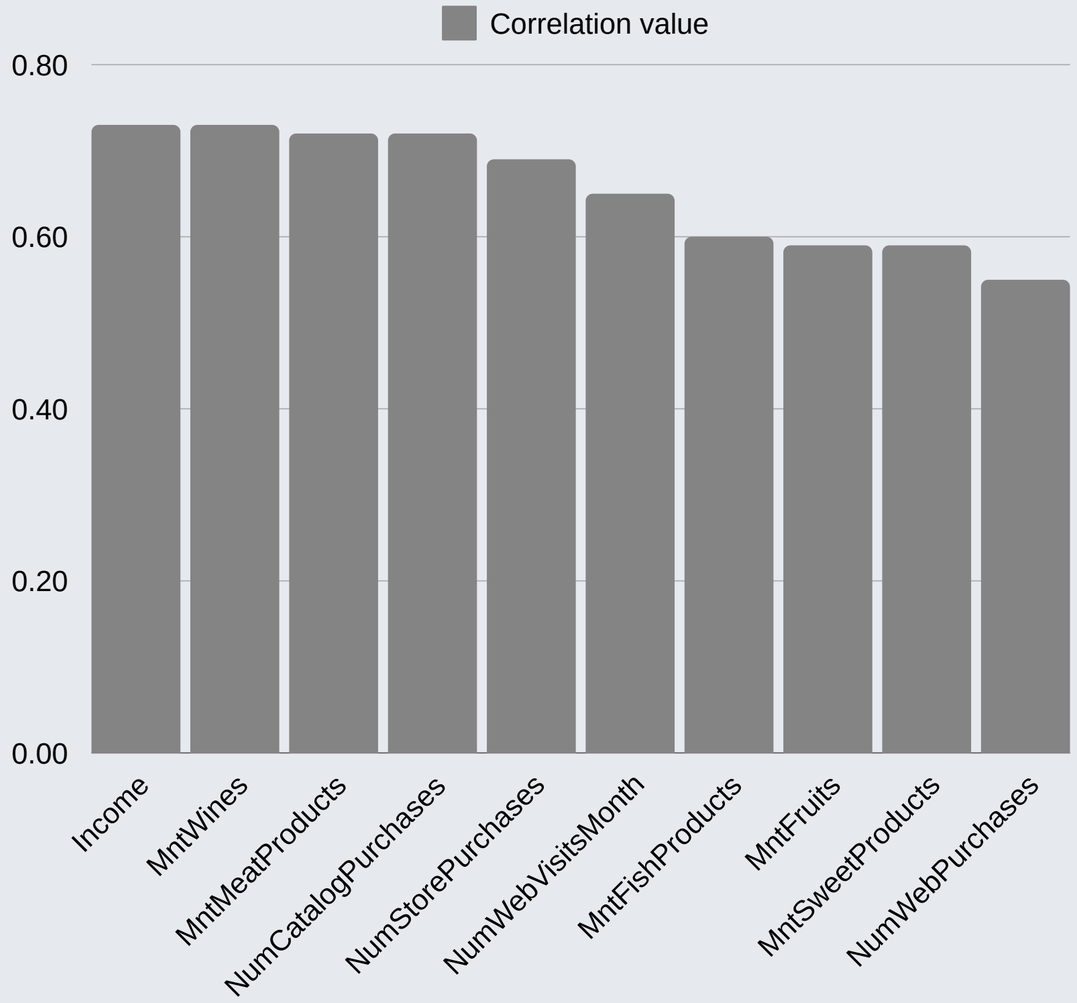
Feature Selection

VIF Analysis and Correlation Values

Removing feature with High VIF Value : ‘Year_Birth’, ‘Income’ and NumWebVisitMonth’

Why Remove Features with High VIF (>10)? Multicollinearity Reduces Model Interpretability, Increases Variance of Coefficients and Degrades Model Performance





Feature Extraction & Encoding

Extracting feature and transforming Categorical Data

Feature Category	Features	Description
Income-Based	**Income Per Capita**	Normalizes income by family size to understand spending behavior.
Spending Patterns	**Total Spending, Category Spending Ratios, Average Monthly Spending**	Provides overall spending metrics and spending preferences.
Purchase Behavior	**Total Purchases, Web vs. Store Purchases Ratio, Deal Purchase Ratio**	Indicates purchase frequency, channel preference, and deal responsiveness.
Campaign Response	**Campaign Acceptance Count, Response Rate**	Measures customer engagement with marketing campaigns.

Why One-Hot Encoding for Marital Status?

- No Order: Marital status categories have no specific order.
- Independence: Each category is separate.

Why Label Encoding for Education?

- Order: Education levels have a clear order.
- Simplicity: Easy to use.

ONE HOT ENCODING

‘Marital Status’ Feature
(Divorced, Married, Single)

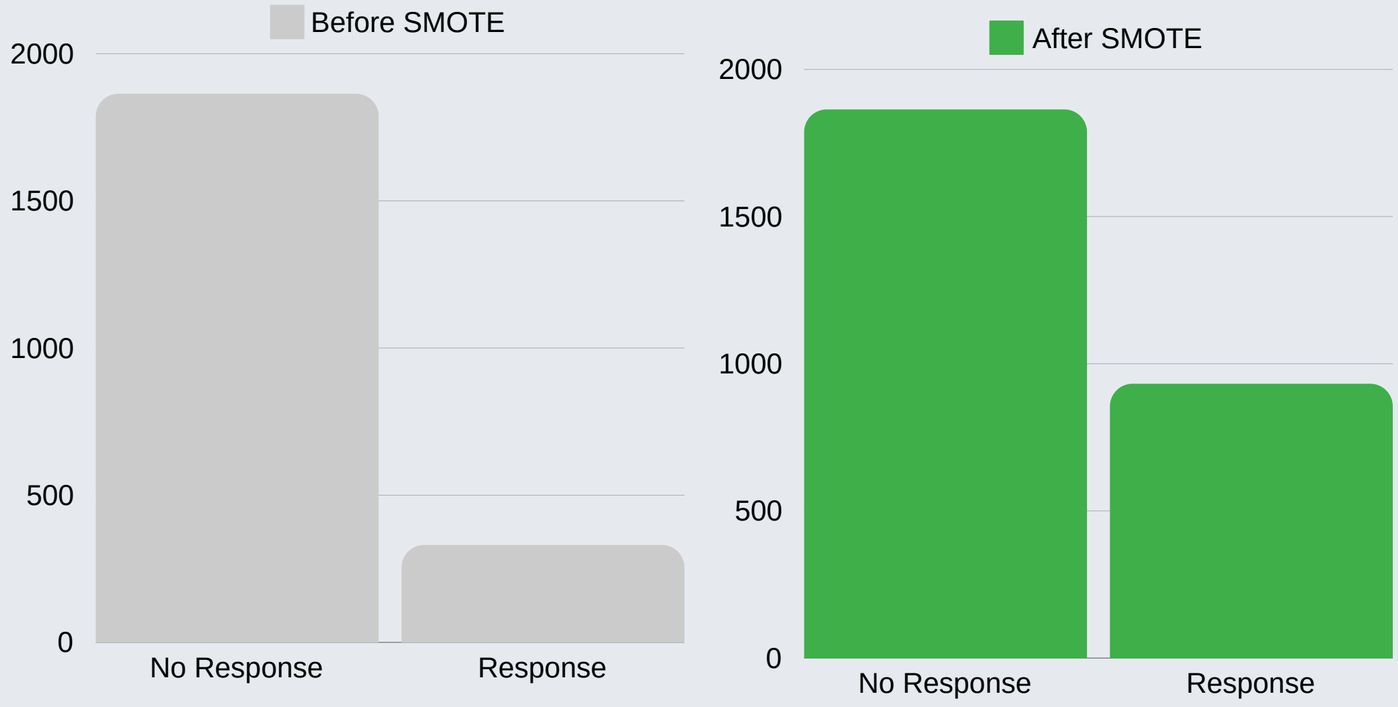
LABEL ENCODING

‘Education’ Feature
(Basic, Bachelor, Master, PhD)

Handle Class Imbalance

SMOTE (Oversampling)

Before SMOTE	After SMOTE
1864	1864
331	932

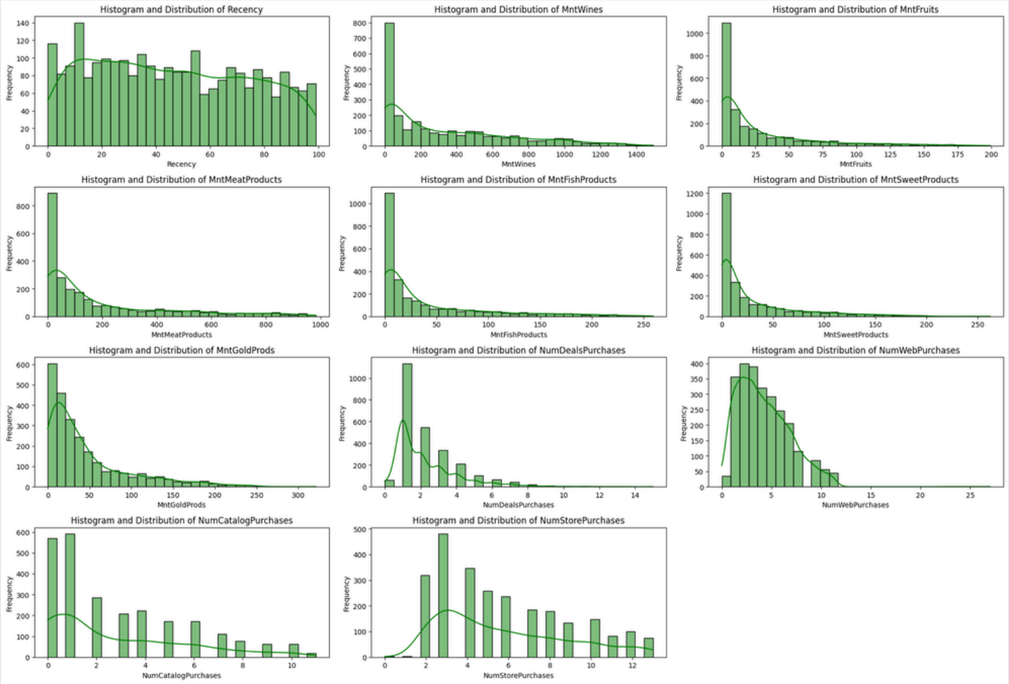


The most recommended handling is oversampling for the minority class (Response). These techniques can increase the number of examples for the minority class by generating **synthetic examples**, thereby increasing the balance in the dataset without reducing the amount of data from the majority class. **Using 1:2 avoid a lot of synthetic data because it causes bias.**

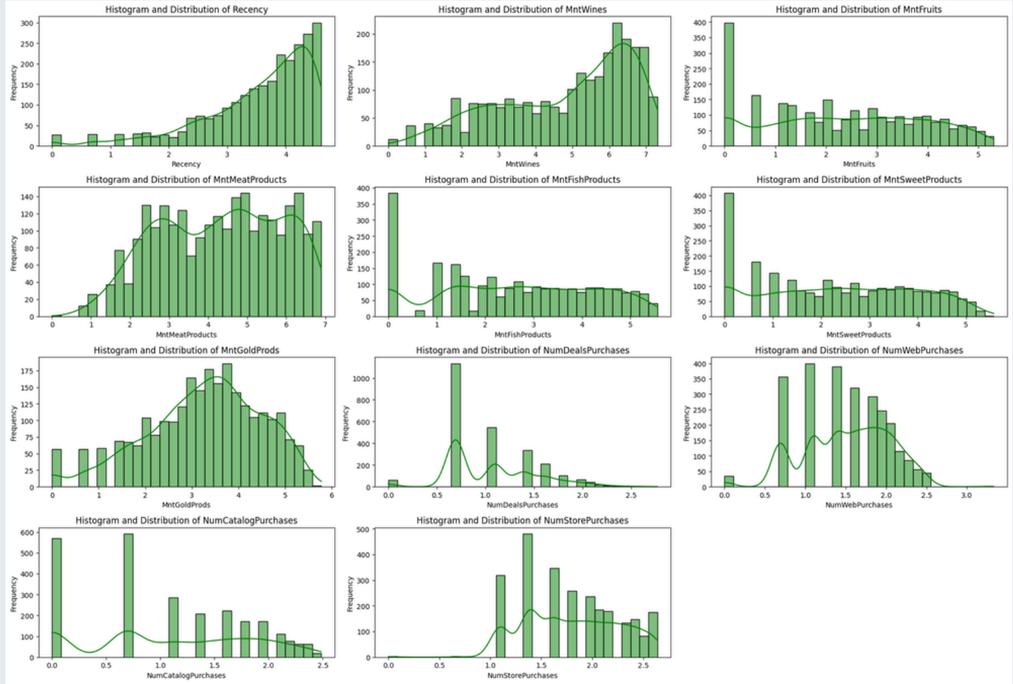
Feature Scaling

Using log transformation

This code applies a log transformation to **reduce the skewness** in numerical data. It works by converting each value using the formula **`log(1 + x)`**, which helps to spread out highly skewed data. These technique is useful for making **data more normally distributed, and usual more used on data with many outliers.**



Log Transformation



Modeling

Split Data Train

Modeling

Evaluation (metrics & underfit/overfit)

Hyperparameter

Feature Importance

Metrics Modeling

Precision : key performance metric in classification models, It focuses on how many of the predicted positive responses were actually correct and focusing minimizing false positives.

$$\text{Precision} = \frac{\text{True Positives}}{\text{True Positives} + \text{False Positives}}$$

- **Minimizing false positive**s: In marketing, a false positive means targeting someone who will not actually respond to the offer. This leads to wasted resources, as marketing efforts (emails, ads, or promotions) are spent on uninterested individuals.
- **Efficiency of marketing spend**: By focusing on precision, the campaign can better allocate its budget and efforts toward those more likely to respond, reducing unnecessary costs.
- **Customer experience**: Sending offers to uninterested people may lead to negative perceptions, potentially harming brand reputation.

		Predicted	
		Not Response	Response
Actual	Not Response	547	27
	Response	65	212

Modeling Comparison

Comparing 7 modeling algorithm to find the best **precision**

Logistic Regression is chosen as the best-fit modeling algorithm because it has the highest precision and correct positive predictions, indicating its superior performance in terms of accuracy and reliability

No	Model	Precision (Train)	Precision (Test)	Status (> 10% overfit, <-10% underfit)
0	**Logistic Regression**	**0.803**	**0.797**	**Best fit**
1	KNN	0.752	0.585	Overfit
2	Decision Tree	1.000	0.736	Overfit
3	Random Forest	0.997	0.852	Overfit
4	AdaBoost	0.892	0.835	Overfit
5	Gradient Boosting	0.981	0.877	Overfit
6	XGBoost	0.998	0.840	Overfit

Hyperparameter Tuning

Logistic Regression Algorithm

Model	Precision Before	Precision After	**Increase**
Train Data	0.803	0.898	**11.83 %**
Test Data	0.797	0.887	**11.29 %**

Logistic Regression is chosen as the best-fit modeling algorithm because it has the highest precision and correct positive predictions, indicating its superior performance in terms of accuracy and reliability.

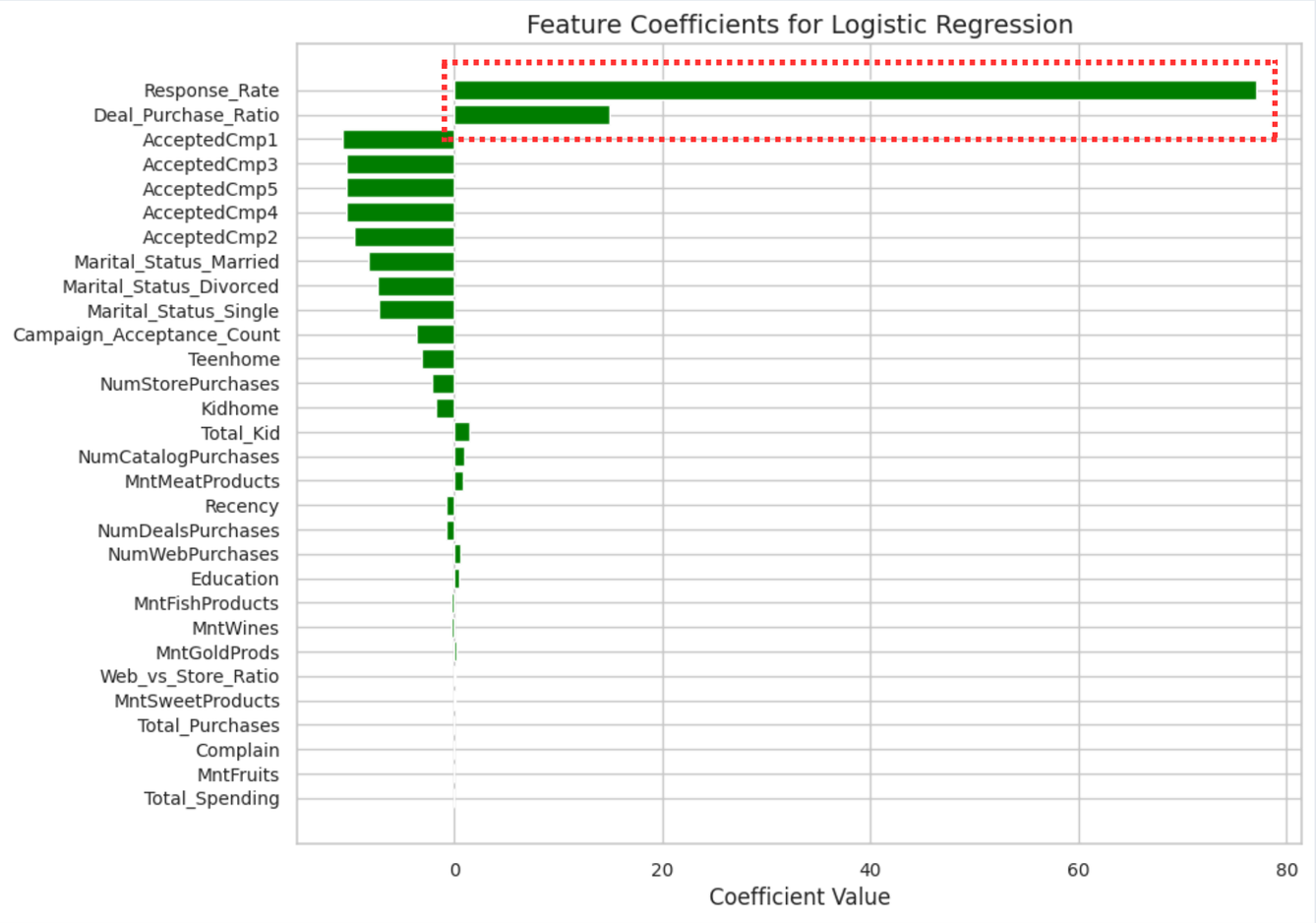
- Data Splitting: The data was divided into 5 parts (folds) for cross-validation to prevent overfitting.
- Hyperparameter Tuning: 120 different combinations of hyperparameters were tested to find the best configuration.
- Model Evaluation: The model's performance was measured using precision on both training and test data.

Best Model Parameters

- Regularization: C=100 (balance between fitting the data and avoiding overfitting)
- Iterations: max_iter=100 (maximum number of iterations during training)
- Regularization Type: penalty='l1' (sparse model, many coefficients are zero)
- Solver: solver='liblinear' (efficient optimization algorithm)

Feature Importance

After Hyperparameter Tuning



1. Response Rate: The "Response Rate" feature itself has a strong positive coefficient, suggesting that customers who have responded to previous campaigns are more likely to respond again.
2. Deal_Purchase_Ratio: the second most influential feature, also with a strong positive impact. A higher deal purchase ratio (more frequent purchases on deals) is positively correlated with customer response related to the marketing campaign.

Business Metrics Evaluation

	Precision	Recall	F1 Score	AUC-ROC
Train Data	0.8994	0.7787	0.8347	-
Test Data	0.887	0.7653	0.8217	0.942

167%

INCREASING RESPONSE RATE

From 14.91% to 39.81% response rate after using ML.

↓ 72%

REDUCE MARKETING COST

From $6,720 to $1,884 for marketing cost. Bcs just sending campaigns into 628 potential customer from 2012 overall.

↑ 201%

INCREASING ROI RETURN OF INVESMENT

From -45.33% to 45.97% with predicted revenue $5,760 for once marketing campaign.

Business Recommendation

	Focus on Boosting Initial Engagement and Optimize Retargeting	**Strengthen Repeat Purchase Programs**	**Data-Driven and Personalized Approach**
	Most campaigns have low response rates (Mode 0.00 and average 0.06), but when customers do respond, they tend to make a purchase. Many customers make only one purchase (Mode Number of Purchases 1)	Repeat deals purchases are still low, with an average of 2.31 and a mode of 1. Most customers do not return to buy again after their first interaction	Logistic regression analysis shows some campaigns have a negative impact. Personalized campaigns based on customer data could yield better results.
	1. Optimize Customer Segmentation: Use predictive analytics to identify high-conversion potential customer groups. Focus marketing efforts on these segments to increase initial campaign responses. A/B testing for various campaign formats is also recommended to identify the most effective types of campaigns 2. Behavior-Based Retargeting: Implement more targeted retargeting strategies to reach customers who have interacted but haven't made a repeat purchase, offering discounts or special promotions.	1. Upselling Strategies with Product Bundling or Cross Selling: Provide additional incentives, such as discounts or exclusive offers, to first-time buyers to encourage repeat purchases. Also, offer product bundles to motivate customers to buy multiple items, like combining wine and meats	1. Regularly update predictive models with the latest customer data to maintain accuracy and adapt to behavioral changes. Use these data-driven insights to personalize marketing campaigns, focusing on factors that increase response rates and sales

Implementation

“Customer Respond Predictor” by Streamlit

Key Features:

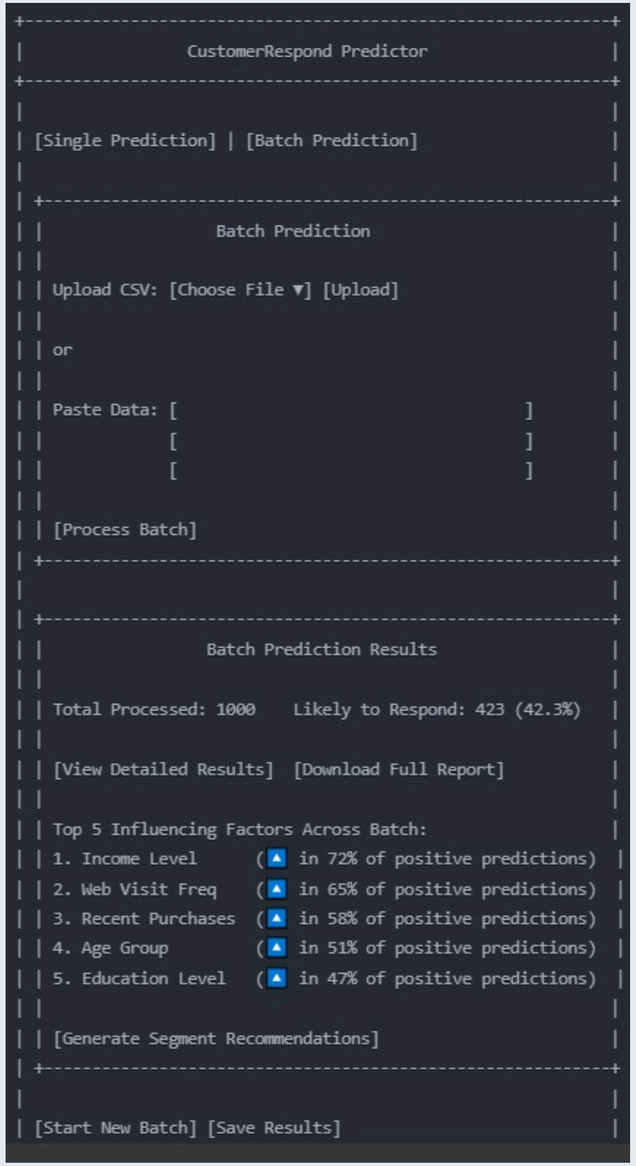
- Individual and Batch Prediction: Analyzes both single customer data and large datasets.
- Accurate Predictions: Identifies customers most likely to respond to marketing campaigns.
- Actionable Insights: Provides recommendations for targeted marketing strategies.

How it Works:

1. Data Input: Upload or paste customer data.
2. Prediction: Model analyzes data to predict response likelihood.
3. Insights: Reveals key factors influencing predictions and suggests actions.

Benefits:

- Improved Targeting: Focus marketing efforts on high-potential customers.
- Increased ROI: Maximize campaign effectiveness and reduce costs.
- Data-Driven Decisions: Make informed choices based on predictive analytics.



Required Feature to Improve Analysis

more understanding of customers & enhancing predictive models

Metrics	Required Feature
1. Customer Lifetime Value (CLV) Prediction	- Historical purchase data (transaction dates, amounts) - Purchase frequency (number of transactions per customer) - Average order value (total amount spent divided by the number of transactions)
2. Social Media Engagement Score	- Social media interaction data (likes, shares, comments, posts) - Customer IDs linked to social media accounts
3. Product Affinity Scores	- Purchase history (product categories, purchase amounts) - Customer IDs
4. Sentiment Analysis of Customer Reviews	- Customer reviews or feedback text - Sentiment analysis tool or library (e.g., VADER, TextBlob, or custom NLP model)