



**Department of Decision Science
Faculty of Business
University of Moratuwa**

Semester 06

DA 3131 – Data Mining

Individual Assignment

Due Date of Submission

[27 / 10 / 2024]

Word Count

[2507]



**Department of Decision Science
Faculty of Business
University of Moratuwa
Semester 06**

**DA 3131 – Data Mining
Submission Sheet**

Name: Sandali Tharumini Wellehewa

Index No.:216151X

Mobile: 071-4395965

Email: wellehewast.21@uom.lk

Table of Contents

Executive Summary	3
Dataset Selection.....	4
Data Preprocessing	6
<i>Observations</i>	6
<i>Data Preprocessing Steps</i>	9
Exploratory Data Analysis (EDA).....	12
<i>Key Insights from the Descriptive Statistics:</i>	12
<i>Key Insights from the Visualizations:</i>	13
Model Building	20
<i>Model Selection</i>	20
<i>Feature Selection</i>	21
<i>Performance Evaluation</i>	23
<i>Overall Recommendation</i>	27
Model Optimization	28
<i>Tuning Process</i>	28
<i>Performance After Tuning</i>	30
<i>Before after comparison</i>	31
Insights and Business Recommendations	32
References	34
Appendix.....	35

Executive Summary

This report analyzes customer data from a superstore's recent marketing campaign that promoted a discounted gold membership. The aim was to understand customer behavior and improve future marketing strategies through predictive modeling.

Dataset Selection

A Dataset from a superstore's previous marketing campaign has been selected for this study, where customers were offered a gold membership at a discounted rate. The dataset includes customer demographics, purchasing behaviours, and whether they accepted the offer during the last campaign.

Data description is as follows;

- **Response** (target) - 1 if customer accepted the offer in the last campaign, 0 otherwise
- **ID** - Unique ID of each customer
- **Year_Birth** - Age of the customer
- **Complain** - 1 if the customer complained in the last 2 years
- **Dt_Customer** - date of customer's enrollment with the company
- **Education** - customer's level of education
- **Marital** - customer's marital status
- **Kidhome** - number of small children in customer's household
- **Teenhome** - number of teenagers in customer's household
- **Income** - customer's yearly household income
- **MntFishProducts** - the amount spent on fish products in the last 2 years
- **MntMeatProducts** - the amount spent on meat products in the last 2 years
- **MntFruits** - the amount spent on fruits products in the last 2 years
- **MntSweetProducts** - amount spent on sweet products in the last 2 years
- **MntWines** - the amount spent on wine products in the last 2 years
- **MntGoldProds** - the amount spent on gold products in the last 2 years
- **NumDealsPurchases** - number of purchases made with discount
- **NumCatalogPurchases** - number of purchases made using catalog (buying goods to be shipped through the mail)
- **NumStorePurchases** - number of purchases made directly in stores
- **NumWebPurchases** - number of purchases made through the company's website
- **NumWebVisitsMonth** - number of visits to company's website in the last month
- **Recency** - number of days since the last purchase

Shape of the Dataset (2240, 22) → 2240 rows & 22 columns

Id	int64
Year_Birth	int64
Education	object
Marital_Status	object
Income	float64
Kidhome	int64
Teenhome	int64
Dt_Customer	object
Recency	int64
MntWines	int64
MntFruits	int64
MntMeatProducts	int64
MntFishProducts	int64
MntSweetProducts	int64
MntGoldProds	int64
NumDealsPurchases	int64
NumWebPurchases	int64
NumCatalogPurchases	int64
NumStorePurchases	int64
NumWebVisitsMonth	int64
Response	int64
Complain	int64

This dataset is highly relevant to business as it helps the superstore reduce the cost of their upcoming campaign by identifying which customers are likely to purchase a membership. By using data mining techniques to predict customer responses, the store can focus their marketing efforts on high-potential customers, optimising resource allocation.

The problem at hand is a binary classification task, where the goal is to predict whether a customer will respond positively to the membership offer (target variable: Response = 1). This will enable the superstore to target the right audience and improve the efficiency of their marketing strategy.

Data Preprocessing

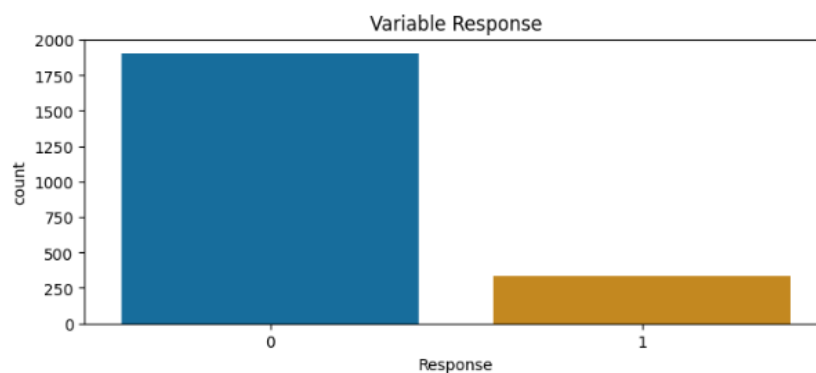
Observations

1. Missing Values

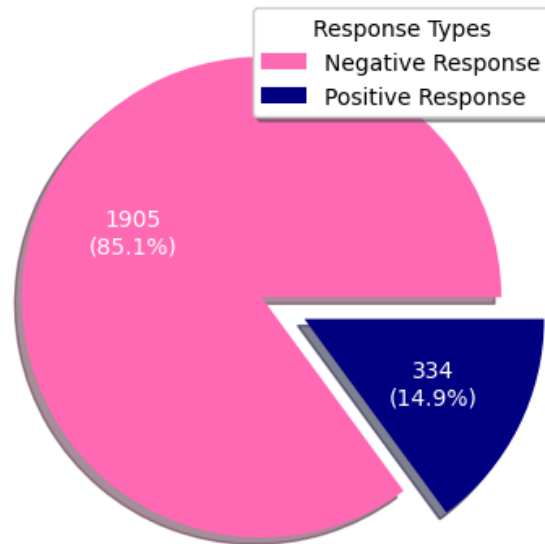
It was revealed that the dataset had 24 missing values in the Income.

Id	0
Year_Birth	0
Education	0
Marital_Status	0
Income	24
Kidhome	0
Teenhome	0
Dt_Customer	0
Recency	0
MntWines	0
MntFruits	0
MntMeatProducts	0
MntFishProducts	0
MntSweetProducts	0
MntGoldProds	0
NumDealsPurchases	0
NumWebPurchases	0
NumCatalogPurchases	0
NumStorePurchases	0
NumWebVisitsMonth	0
Response	0
Complain	0

2. Class Imbalance in “Response” have identified

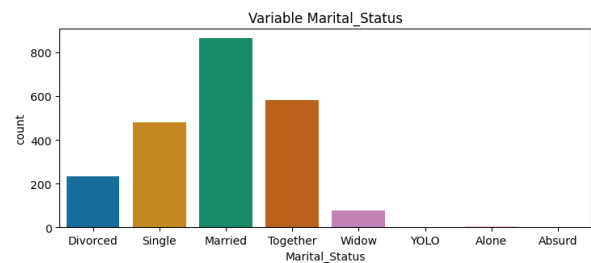
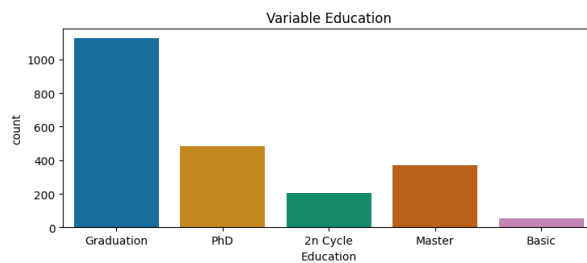


Percentage of Negative and Positive Response from the Marketing Campaign

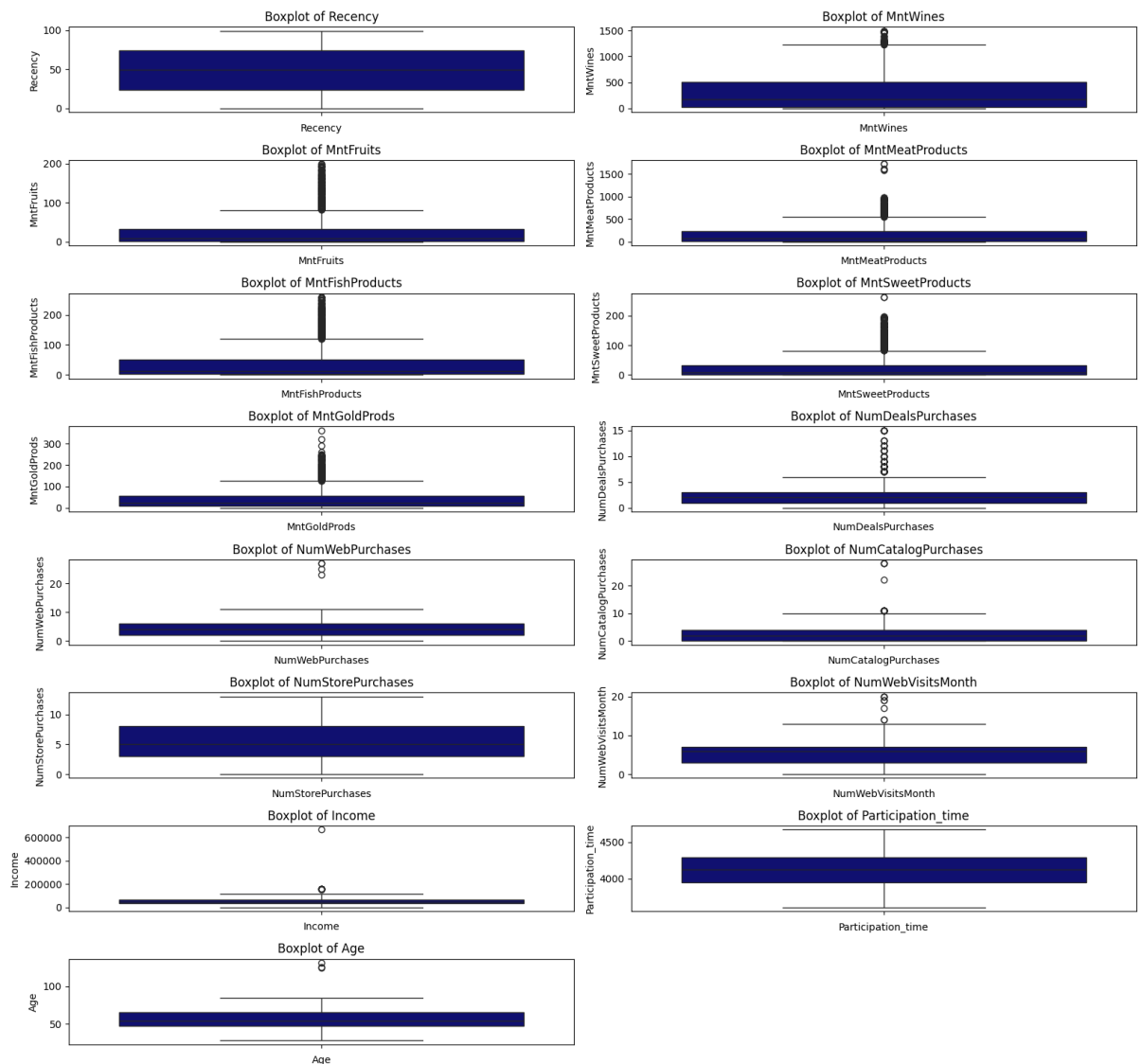


3. Erroneous Data in “Marital_Status” and “Education”

In Education, "2nd Cycle" equals a Master's. Marital status entries "Alone," "Absurd," and "YOLO" are erroneous.



4. Outliers



Outliers in several variables were identified using box plots.

5. Absence of Duplicates

6. No features exhibit problematic multicollinearity

	Feature	VIF
0	Education	1.142404
1	Income	4.700078
2	Kidhome	1.946972
3	Teenhome	1.790667
4	Recency	1.062790
5	MntWines	3.262665
6	MntFruits	2.434451
7	MntMeatProducts	3.894551
8	MntFishProducts	2.690047
9	MntSweetProducts	2.530913
10	MntGoldProds	1.676023
11	NumDealsPurchases	1.868710
12	NumWebPurchases	2.415567
13	NumCatalogPurchases	3.346198
14	NumStorePurchases	2.552526
15	NumWebVisitsMonth	3.160993
16	Response	1.263373
17	Complain	1.008909
18	Participation_time	1.284100
19	Age	1.256884

Data Preprocessing Steps

1. Feature Engineering:

- **Age Calculation:** Derived Age by subtracting Year_Birth from the current year.
- **Participation Time:** Created Participation_time by calculating the number of days since enrollment using Dt_Customer.

2. Handling Marital Status:

- **Marital Status Adjustment:** Categories such as “Alone,” “Absurd,” and “YOLO” were reassigned based on household composition:
 - If the customer had children (Kidhome > 0 or Teenhome > 0) but was marked with an unconventional marital status, they were reclassified as "Married."
 - If there were no children (Kidhome == 0 and Teenhome == 0), they were reclassified as "Single."

Before Marital State Adjustment	After Marital State Adjustment
	

3. Education Adjustment:

- **Combining Similar Categories:** The “2n Cycle” category was merged with “Master”.

Before Education Adjustment		Education	After Education Adjustment	
Education			Education	
Graduation	1127		Graduation	1127
PhD	486		Master	573
Master	370		PhD	486
2n Cycle	203		Basic	54
Basic	54			

4. Missing Value Imputation:

- **Income:** Missing values in the Income column were imputed using the median income of each education level.(Since the income does not follows normal distribution according to QQ plot, Shapiro-Wilk Test, Kolmogorov-Smirnov Test moved to median Imputation)

5. Encoding:

- **Education:** Since Education is an ordinal variable, **label encoding** was applied (e.g., Basic = 0, Graduation = 1, etc.).
- **Marital Status (One-Hot Encoding):** Marital statuses were encoded using **one-hot encoding** as they are nominal variables.

6. Outlier Treatment:

- Outliers were identified using the IQR method and clipped accordingly.
- The IQR method identifies outliers based on data spread, clipping them to reduce their impact while preserving data structure, improving model performance without excessive data removal.

7. Duplicate Check and Cleaning:

- No duplicates were found, and irrelevant columns were removed to clean the dataset.

8. Class distribution has balanced using RandomOversampler.

```
New class distribution:  
Response  
1      1906  
0      1906
```

9. To prepare the dataset for modeling, the resampled data was split into training (70%) and testing (30%) sets.

```
Training set shape: (2668, 23)  
Testing set shape: (1144, 23)
```

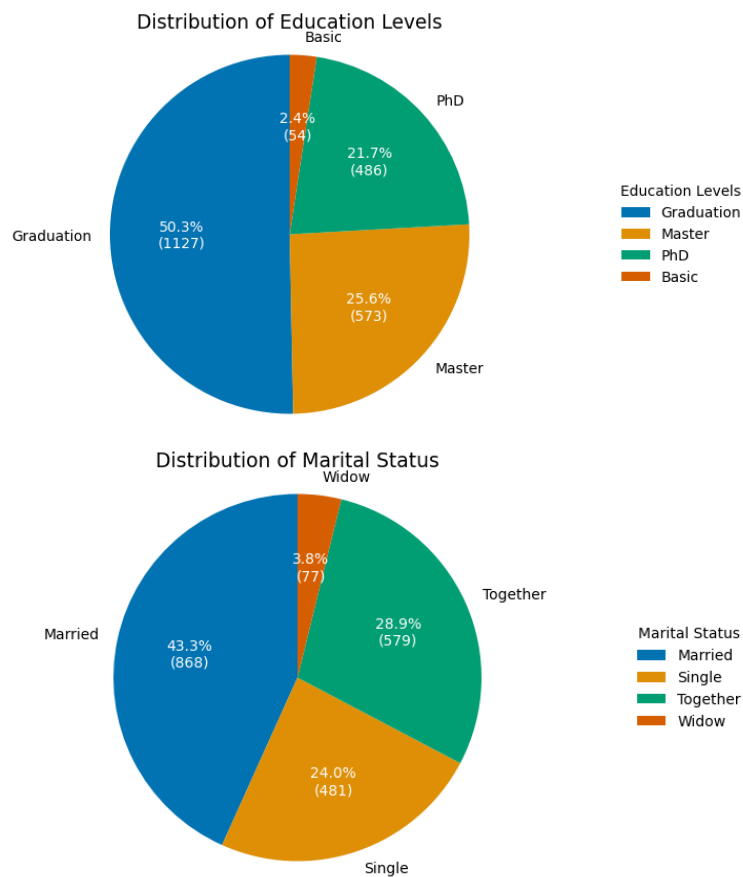
10. Standardization using StandardScaler, which was fitted on the training set and applied to transform sets, enhancing model performance and convergence.

Exploratory Data Analysis (EDA)

Key Insights from the Descriptive Statistics:

1. **Income Distribution:**
 - Wide disparities among individuals.
2. **Age:**
 - An older demographic.
3. **Household Composition:**
 - Many households have no children (teens and kids).
4. **Purchase Behavior:**
 - Highest spending is on wine, followed by meat products.
 - Physical stores are the primary shopping channel, with fewer online and catalog purchases.
5. **Recency:**
 - Average recency is 49 days, with a range from 0 to 99 days, showing varied engagement frequency.
6. **Marketing Response:**
 - Only 15% of customers respond positively to marketing campaigns, indicating potential for campaign improvement.
7. **Complaints:**
 - Very few complaints, which could indicate either customer satisfaction or limited service interaction.
8. **Participation in Loyalty Programs:**
 - Average participation time is approximately 11.3 years, showing strong customer loyalty.

Key Insights from the Visualizations:

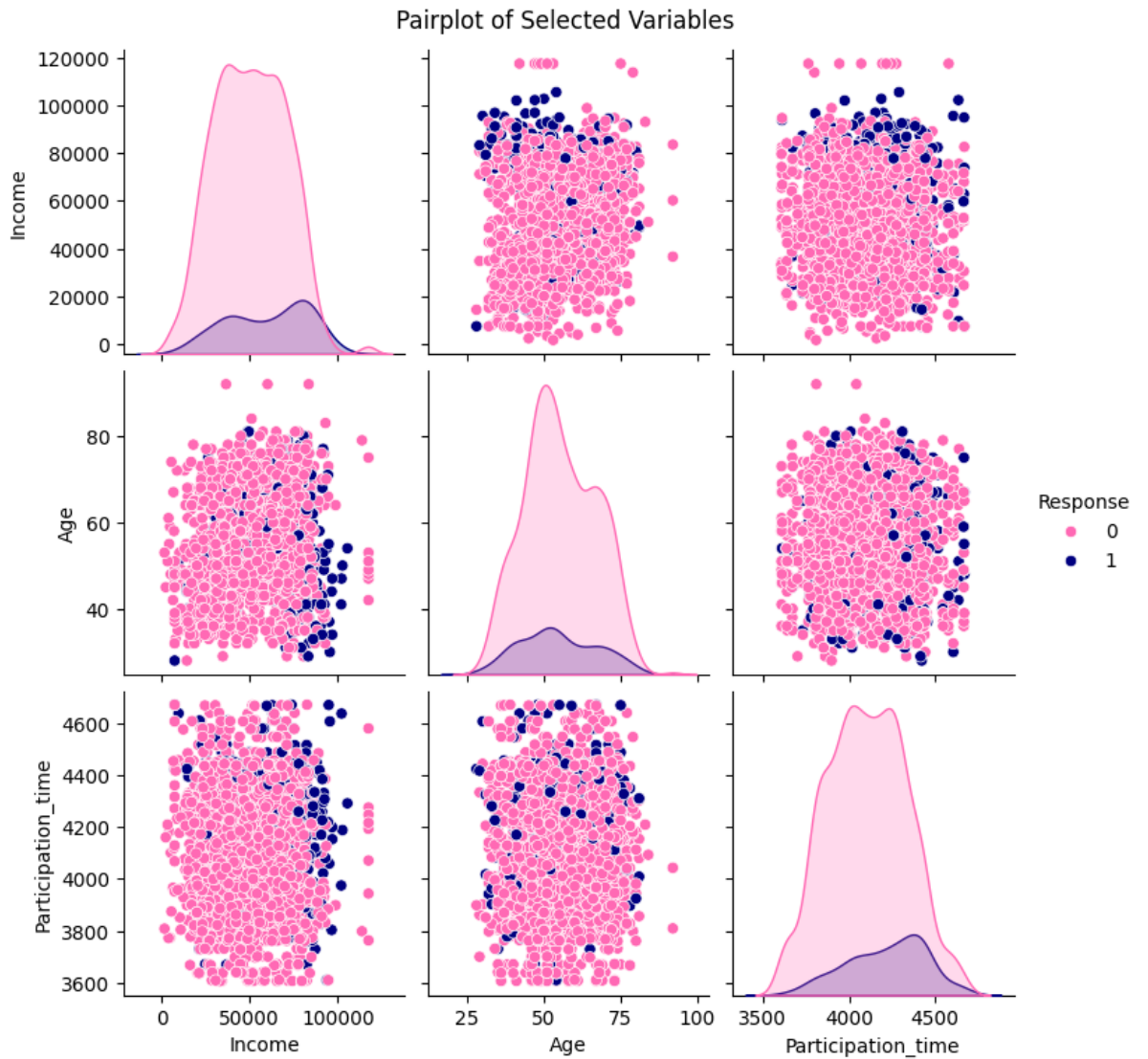


- The customer base is predominantly educated, with the majority holding a graduate degree. Additionally, a significant portion of customers have advanced degrees (Master's and PhD), highlighting a well-educated consumer demographic.
- The majority of customers are married, with a small minority who are widowed. There is also a notable presence of individuals who are single or in a partnership, reflecting a diverse range of marital statuses within the customer base.

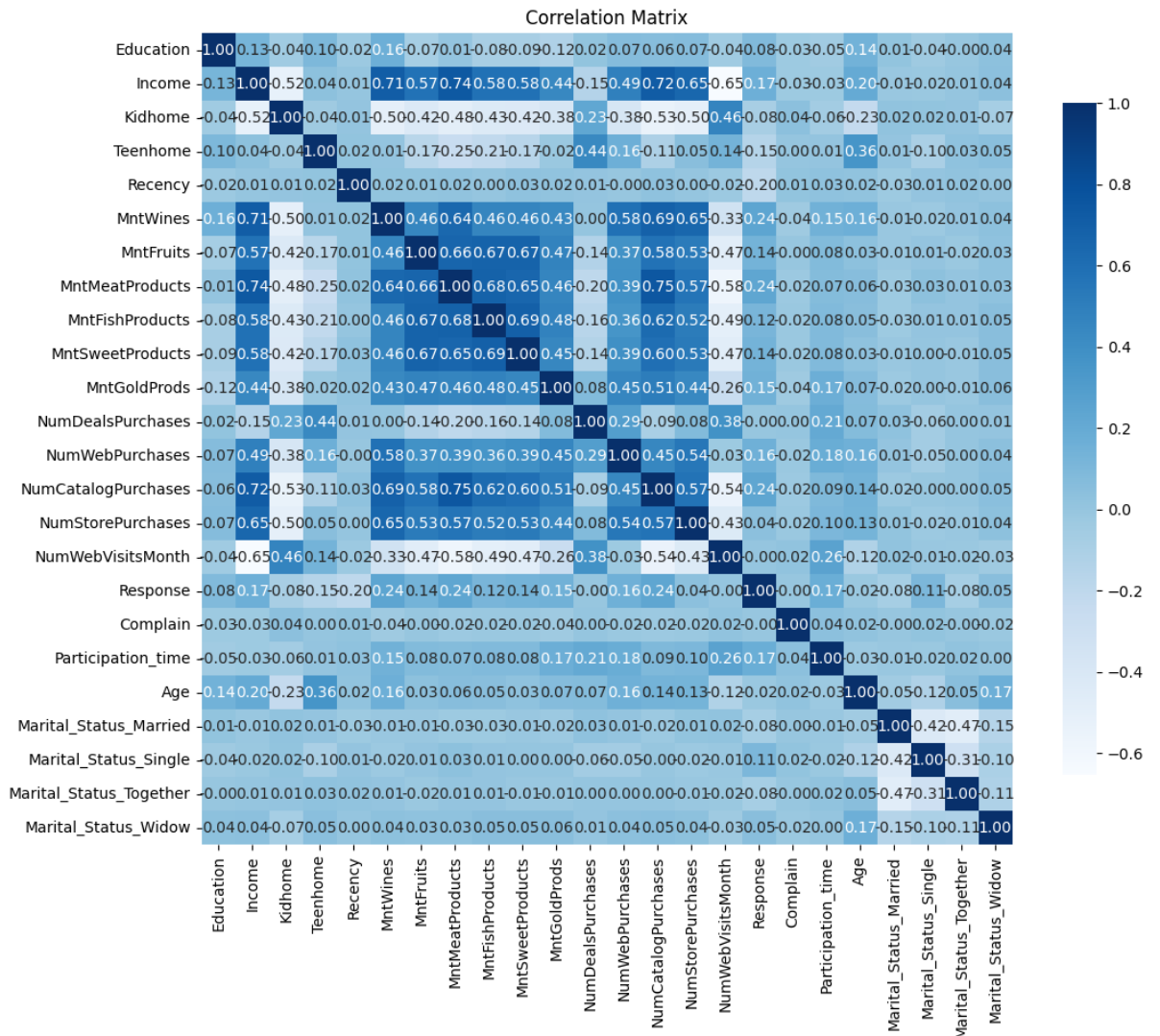
Analysis Of Variable Response



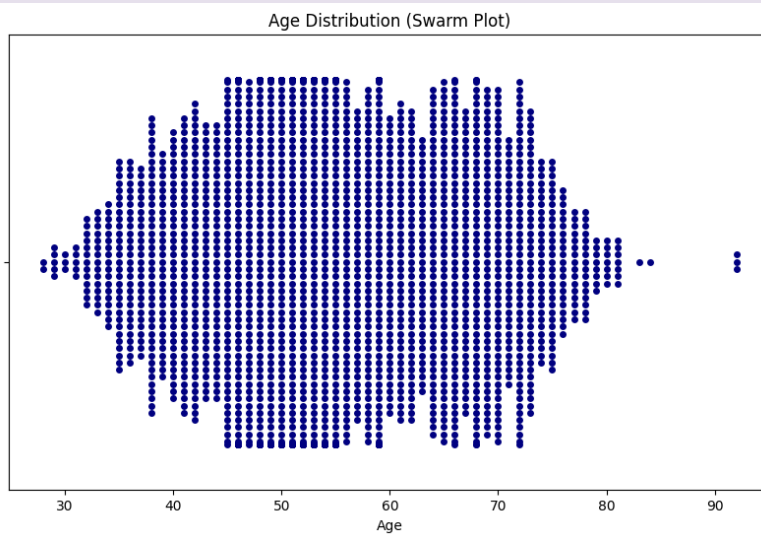
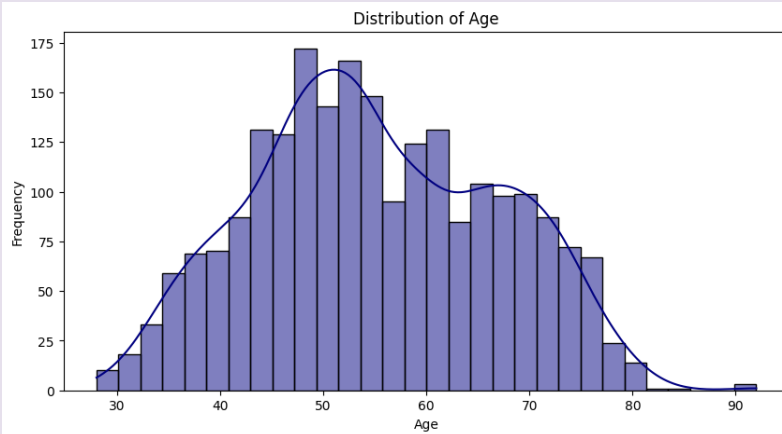
- **Education Level:** Basic-educated individuals are less likely to respond positively to marketing campaigns.
- **Household Composition:** Customers without teens or kids at home are more likely to respond positively to campaigns.
- **Complaints:** Customers with no complaints show a higher likelihood of responding positively.



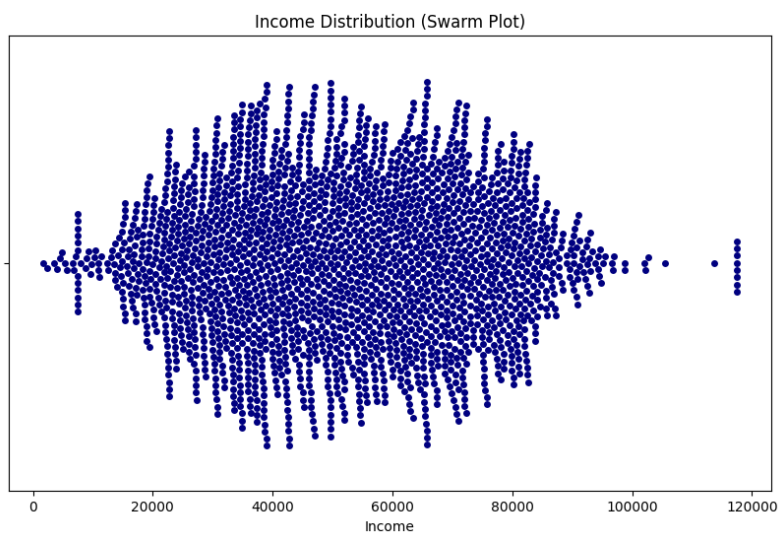
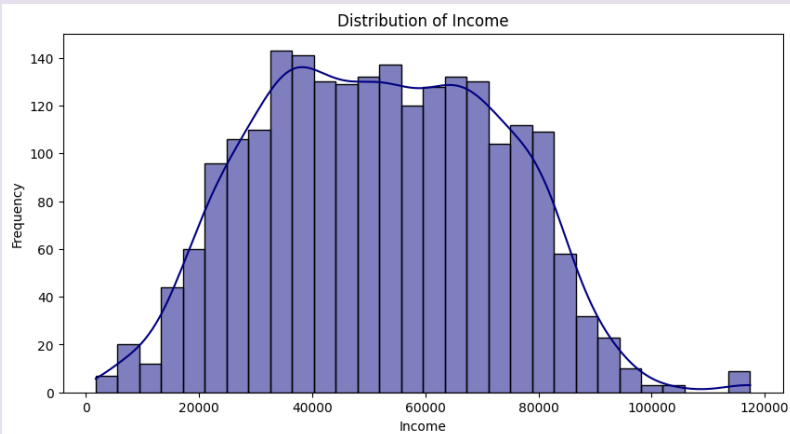
The variables **Income**, **Age**, and **Participation_time** show no strong correlations or distinct distributions between response classes.



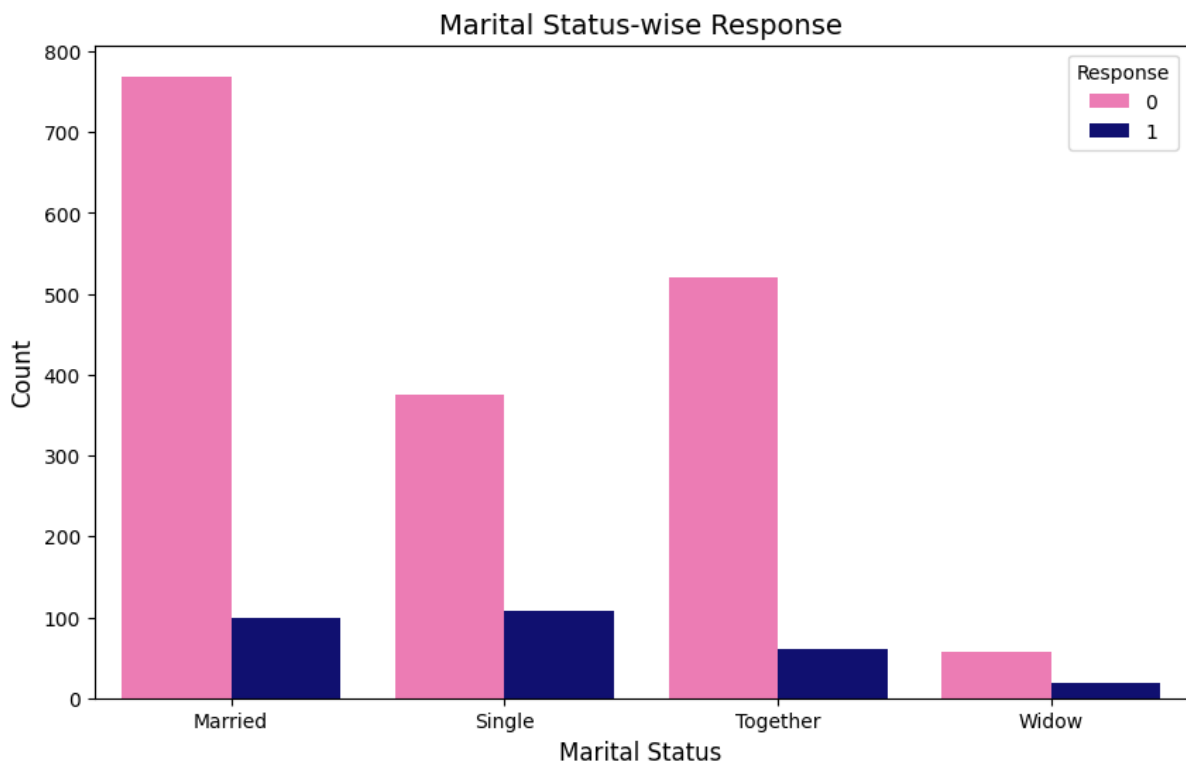
Absence of correlations exceeding 0.8 suggests these variables are not highly collinear.



Most customers fall between 47 and 65 years old, indicating an older demographic.



Broad income range with a majority of income of \$35,000 to \$70,000



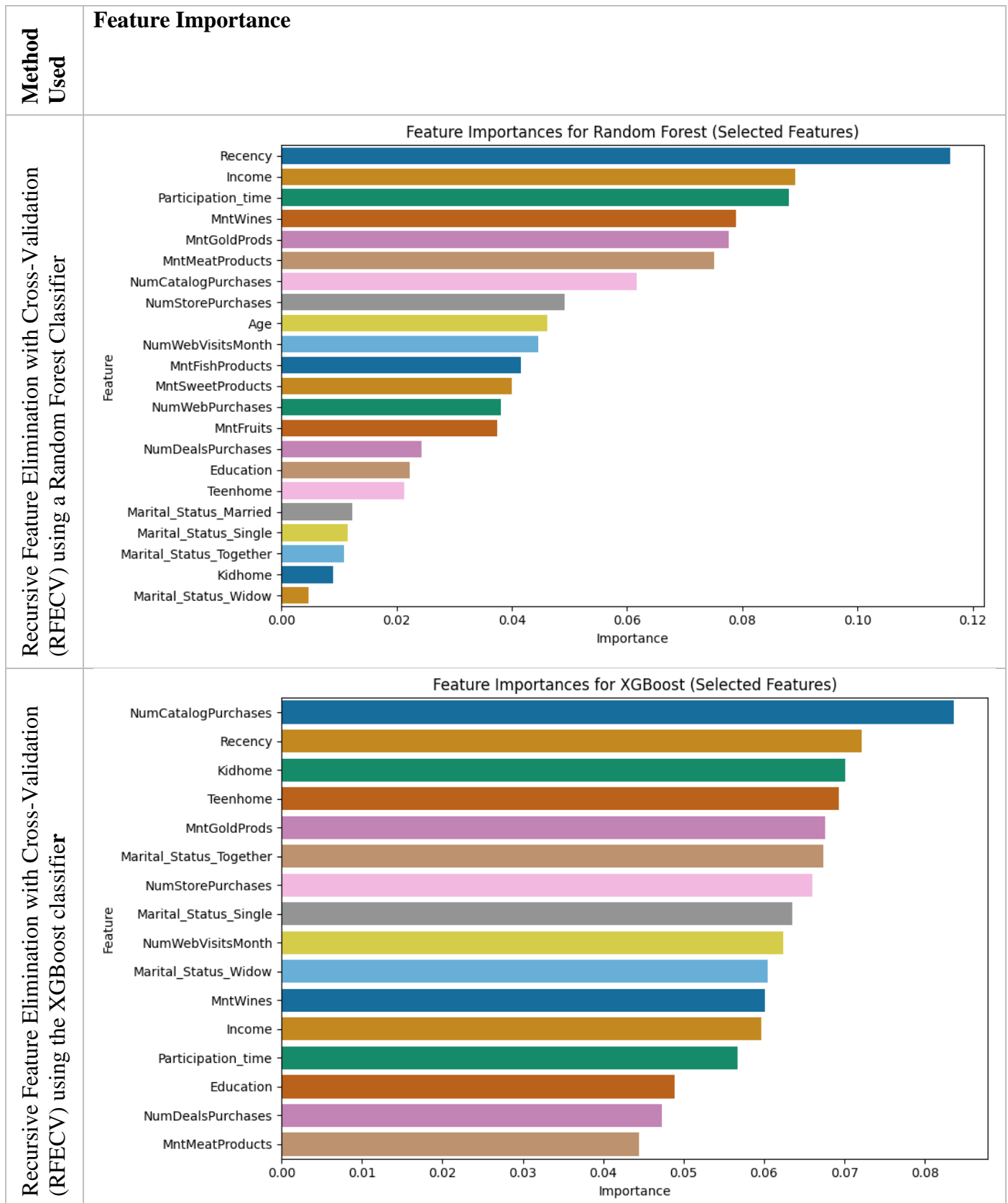
No Significant Marital Status Impact: Doesn't have a significant difference in acceptance rates between different marital statuses. All groups show similar patterns of low acceptance.

Model Building

Model Selection

Model	Justification
Random Forest	Random Forest is ideal for this dataset as it handles a large number of features well, avoids overfitting by averaging multiple decision trees, and captures complex relationships. Its ability to rank feature importance helps identify key factors influencing customer responses.
Decision Tree	Decision Tree is a simple and interpretable model, making it useful for understanding the decision-making process behind customer behavior. It works well with both categorical and numerical variables and can easily handle the imbalanced nature of the dataset.
XG Boost	XGBoost is chosen for its ability to deliver high performance, especially with structured data like this. It excels at handling imbalanced data, reducing bias and variance, and improving prediction accuracy by efficiently learning from the dataset's features through boosting techniques.

Feature Selection



To improve model consistency, we then selected the common variables from both sets to run the Decision Tree (DT), Random Forest (RF), and XGBoost models.

```
# Common selected features
common_features = ['Education', 'Income', 'Kidhome', 'Teenhome', 'Recency', 'MntWines', 'MntMeatProducts', 'MntGoldProds',
                   'NumDealsPurchases', 'NumCatalogPurchases', 'NumStorePurchases',
                   'NumWebVisitsMonth', 'Participation_time', 'Marital_Status_Single',
                   'Marital_Status_Together', 'Marital_Status_Widow']
```

Performance Evaluation

Comparing Accuracies

	Training Accuracy	Testing Accuracy	Interpretation
Decision Tree	0.9959	0.9126	High accuracy on both the training and testing sets. However, the testing accuracy of 91.26% suggests that the model may be slightly overfitted, as indicated by the near-perfect training accuracy.
Random Forest	0.9959	0.9432	Similar training accuracy to the Decision Tree but has higher testing accuracy, suggesting that it generalizes better. With an accuracy of 94.32%, it outperforms the Decision Tree in classification performance.
XG Boost	0.9959	0.9318	Performed well, with a testing accuracy of 93.18%. While slightly lower than Random Forest, it still shows strong generalization capabilities.

Comparing Classification Reports

1. Decision Tree

Excellent recall for Class 1 (98%), but lower recall for Class 0 (85%), indicating it performs better at predicting positive responders while **misclassifying more negative responders**.

Decision Tree Classification Report:				
	precision	recall	f1-score	support
0	0.98	0.85	0.91	584
1	0.86	0.98	0.92	560
accuracy			0.91	1144
macro avg	0.92	0.91	0.91	1144
weighted avg	0.92	0.91	0.91	1144

2. Random Forest

Balanced and high precision and recall for both classes (91-98%), making it the best performer overall with 94.32% accuracy.

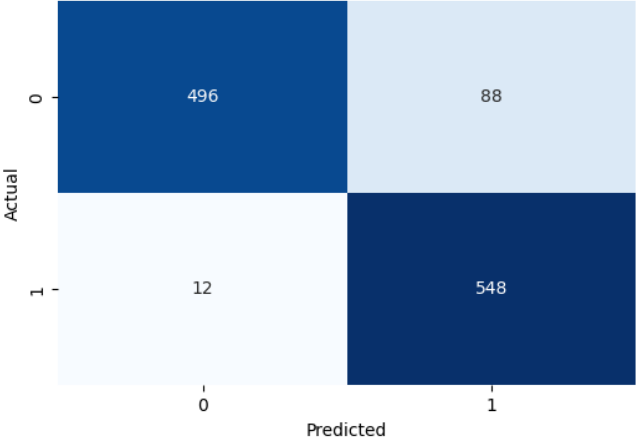
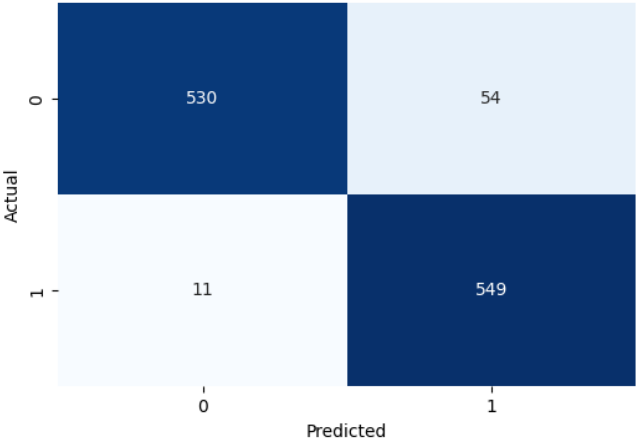
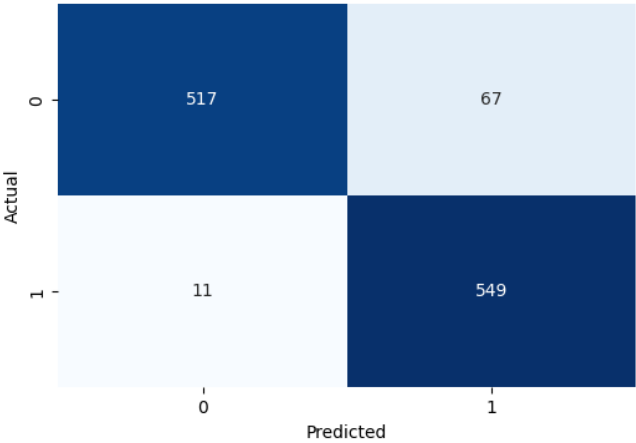
Random Forest Classification Report:				
	precision	recall	f1-score	support
0	0.98	0.91	0.94	584
1	0.91	0.98	0.94	560
accuracy			0.94	1144
macro avg	0.95	0.94	0.94	1144
weighted avg	0.95	0.94	0.94	1144

3. XGBoost

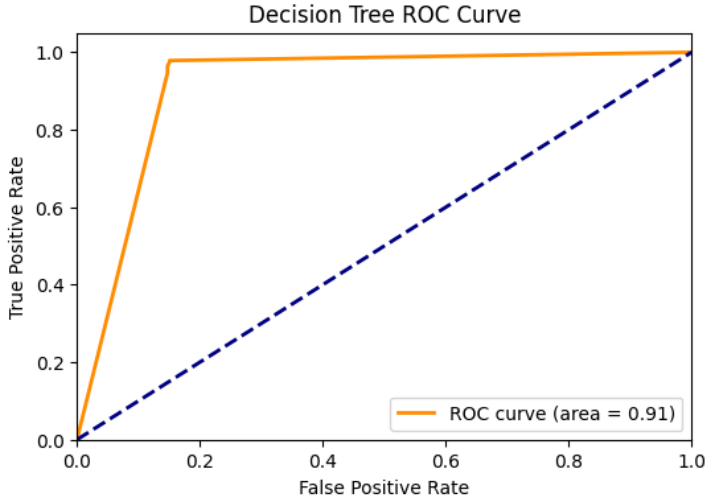
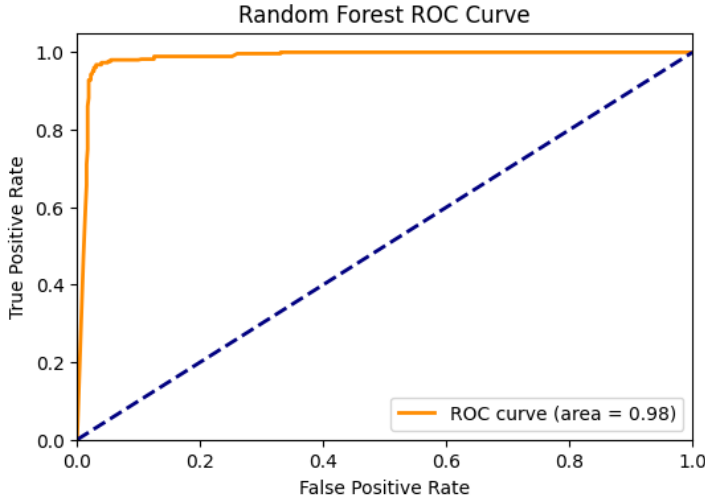
Strong recall for positive responders (98%) but slightly lower recall (89%) and precision for negative responders, performing well but slightly behind Random Forest.

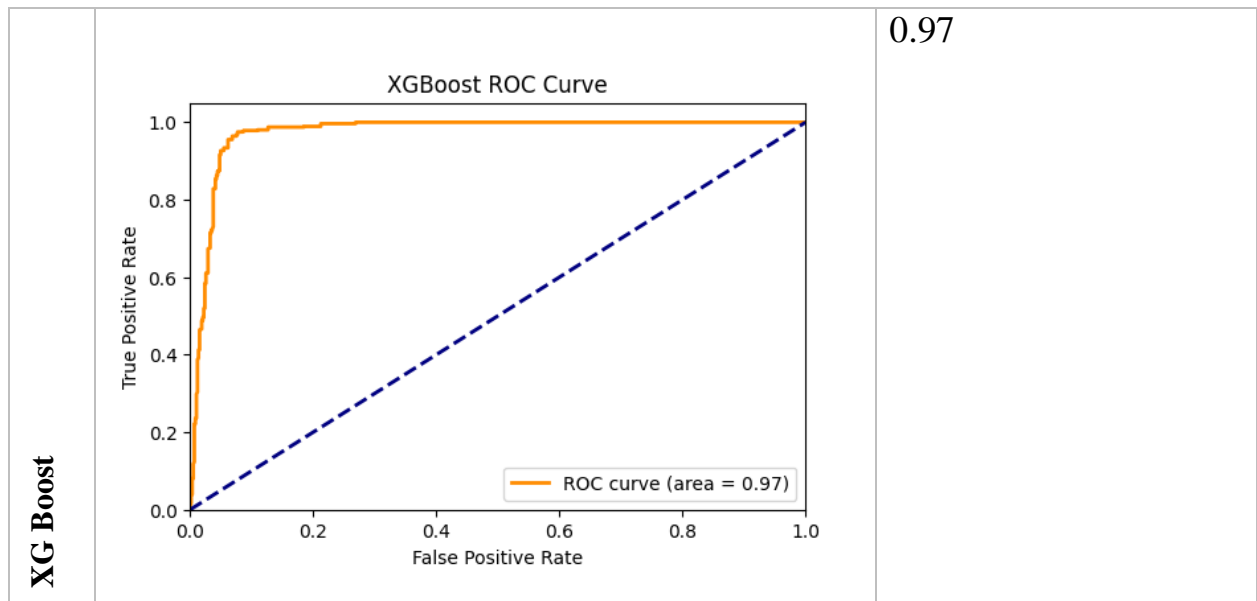
XGBoost Classification Report:				
	precision	recall	f1-score	support
0	0.98	0.89	0.93	584
1	0.89	0.98	0.93	560
accuracy			0.93	1144
macro avg	0.94	0.93	0.93	1144
weighted avg	0.94	0.93	0.93	1144

Comparing Confusion Matrix

Model	Confusion Matrix	Interpretation									
Decision Tree	<p>Decision Tree Confusion Matrix</p>  <table border="1"> <thead> <tr> <th>Actual \ Predicted</th> <th>0</th> <th>1</th> </tr> </thead> <tbody> <tr> <th>0</th> <td>496</td> <td>88</td> </tr> <tr> <th>1</th> <td>12</td> <td>548</td> </tr> </tbody> </table>	Actual \ Predicted	0	1	0	496	88	1	12	548	<ul style="list-style-type: none"> The Decision Tree model has a high number of false positives (88), suggesting it often misclassifies negative responders as positive once. However, it performs well in predicting accepting customers with very few false negatives (12).
Actual \ Predicted	0	1									
0	496	88									
1	12	548									
Random Forest	<p>Random Forest Confusion Matrix</p>  <table border="1"> <thead> <tr> <th>Actual \ Predicted</th> <th>0</th> <th>1</th> </tr> </thead> <tbody> <tr> <th>0</th> <td>530</td> <td>54</td> </tr> <tr> <th>1</th> <td>11</td> <td>549</td> </tr> </tbody> </table>	Actual \ Predicted	0	1	0	530	54	1	11	549	<ul style="list-style-type: none"> Random Forest shows a strong performance with low false positives (54) and very few false negatives (11), indicating it accurately identifies both classes effectively. The model's overall accuracy is enhanced by its ability to correctly classify the majority of cases.
Actual \ Predicted	0	1									
0	530	54									
1	11	549									
XG Boost	<p>XGBoost Confusion Matrix</p>  <table border="1"> <thead> <tr> <th>Actual \ Predicted</th> <th>0</th> <th>1</th> </tr> </thead> <tbody> <tr> <th>0</th> <td>517</td> <td>67</td> </tr> <tr> <th>1</th> <td>11</td> <td>549</td> </tr> </tbody> </table>	Actual \ Predicted	0	1	0	517	67	1	11	549	<ul style="list-style-type: none"> XGBoost also demonstrates strong classification of accepting customers with only 11 false negatives. However, it has a higher number of false positives (67) compared to Random Forest, indicating more misclassifications of negative responders as positive.
Actual \ Predicted	0	1									
0	517	67									
1	11	549									

Receiver Operating Characteristic Curve (ROC) and Area Under the Curve (AUC)

Model	ROC Curve	ROC Score
Decision Tree		0.91
Random Forest		0.98 The high ROC scores suggest that all models are effective at classifying customers based on their acceptance of offers, but Random Forest has the highest discrimination ability, making it the most reliable among the three.



Overall Recommendation

Best Choice: Random Forest is the preferred model for this classification task due to its high performance, demonstrating the best balance of accuracy, precision, recall, and overall reliability. For further steps, it has been utilized as the primary model for implementation.

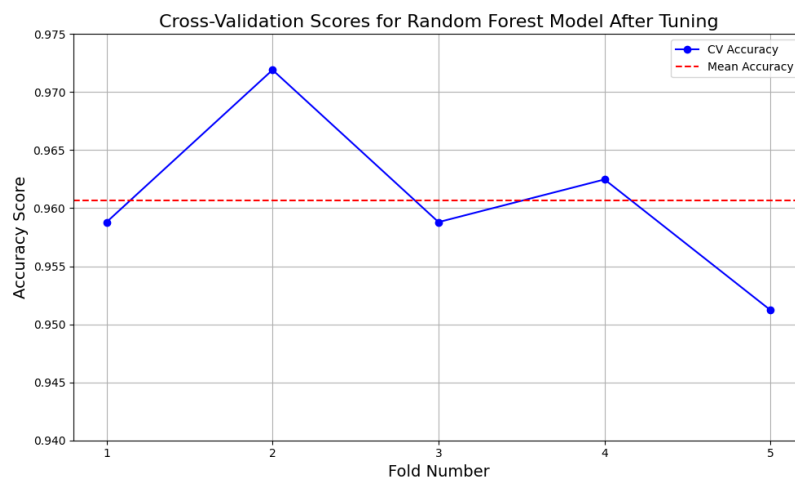
Model Optimization

Tuning Process

To enhance the performance of the Random Forest Classifier has employed **RandomizedSearchCV**.

Parameter	Possible Estimator values
n_estimators	chosen from 100 to 500
max_features	options included 'sqrt', 'log2' and 'None'
max_depth	From 10 to 50, including None.
min_samples_split	2, 5, and 10.
min_samples_leaf	Selected from 1, 2, and 4.
bootstrap	Tested with both True and False.

Using a **stratified K-Fold cross-validation** approach with **5 splits** ensured consistent class distributions across folds. Additionally, shuffling data with **shuffle=True** promotes a random distribution in the folds, enhancing model evaluation and reducing bias in class representation.



After fitting the model, the best hyperparameters were identified as:

Parameter	Best hyperparameters identified
n_estimators	400
max_features	log2
max_depth	20
min_samples_split	2
min_samples_leaf	1
bootstrap	False

Performance After Tuning

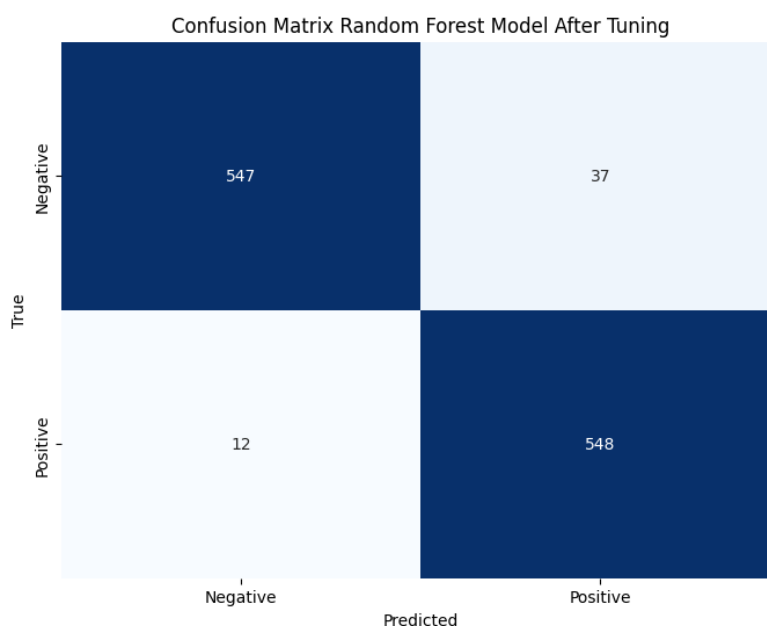
Criteria	Random Forest	Insight
Testing Accuracy	0.9571	The tuned Random Forest model demonstrates high reliability and accuracy, with a testing accuracy of 95.71% and a cross-validation accuracy of 96.06%, making it well-suited for customer prediction tasks.
Mean Cross Validation Accuracy	0.9606	

```
Classification Report:
              precision    recall  f1-score   support

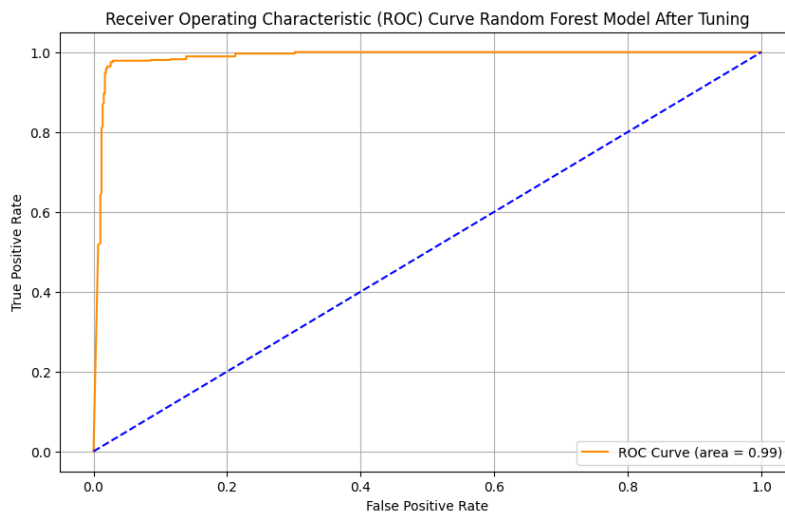
     0           0.98       0.94       0.96         584
     1           0.94       0.98       0.96         560

 accuracy              0.96
 macro avg              0.96
weighted avg              0.96
```

The tuned Random Forest model achieved a balanced and high performance across classes, with a precision and recall of 0.94-0.98 for both accepting and non-accepting customers, resulting in an overall accuracy of 96%.



The tuned Random Forest model achieves high accuracy and robustness in identifying both accepting and non-accepting customers, with minimal misclassifications.

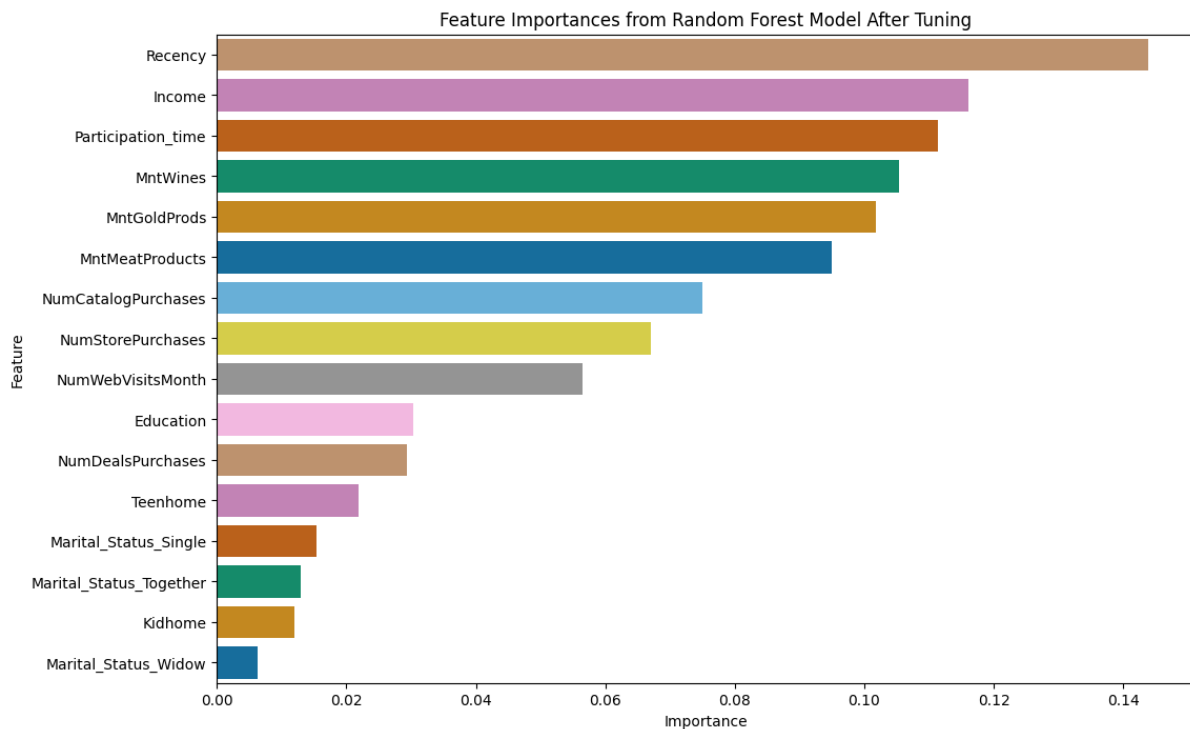


The ROC score of 0.99 demonstrates exceptional discrimination ability, reinforcing that the tuned model effectively distinguishes between accepting and non-accepting customers.

Before after comparison

Criteria	Before Tuning	After Tuning	Direction Of Impact
Testing Accuracy	0.9432	0.9571	↑
ROC Score	0.98	0.99	↑
Classification Report Summary	Balanced performance with precision and recall of 0.91-0.98	Improved performance with precision and recall of 0.94-0.98	↑
Confusion Matrix Summary	Higher false positives	Reduced false positives	↑

Insights and Business Recommendations



Based on the hyperparameter tuned random forest-based classification model, the following can be concluded:

The Random Forest model reveals that Recency (days since last purchase) is the top predictor of customer acceptance, underscoring the importance of recent engagement. Income and Participation_time (duration of customer engagement) are also highly influential, suggesting that wealthier, long-term customers may be more receptive. Spending on specific products like wines, gold products, and meat significantly affects acceptance, indicating that high spenders in these categories are valuable targets. Engagement through various purchase channels, particularly catalog and in-store purchases, also contributes to positive responses, with monthly website visits providing moderate influence. Education and deal usage show some impact but are less critical. Lastly, household and marital status factors like Teenhome, Kidhome, and Marital_Status have minimal influence. This suggests that to increase acceptance, the plan should focus on recently active, high-income, long-term customers with diverse purchasing habits.

Based on the results of the model, the following recommendations can be provided to reduce the chance of negative responses:

1. Prioritize High-Potential Customers

Use the model to identify high-potential customers who, based on purchasing behavior, engagement frequency, and with longer participation times, are more likely to accept the offer. This focus can help reduce resource expenditure on less responsive customers.

2. Personalize Promotions Based on Spending Habits

Design promotions around high-spending categories, such as wines, gold products, and meat, to appeal to customers with a demonstrated interest in these items.

Use feature importance analysis to identify high spenders in specific categories, such as wine and meat, and customers with frequent store visits, to create personalized offers aligned with their purchasing patterns.

3. Tailor Messaging to Household Needs

Customize marketing messages to address specific household needs (e.g., families with children) to increase relevance and response rate.

4. Enhance Engagement Across Multiple Channels

Strengthen catalog and in-store marketing efforts, as these channels have shown a positive influence on acceptance and can complement online engagement strategies.

5. Offer Exclusive, Targeted Deals

Provide targeted discounts and loyalty incentives to educated, deal-seeking customers who appreciate value-driven offers.

6. Leverage Recency and Frequency for Targeting

Customers with recent engagement and higher visit frequency may be more receptive to offers, indicating readiness to spend. Focusing on this group can improve return on marketing investment.

7. Proactively Address Potential Concerns

Although the complaint rate is low, proactively addressing service quality and customer satisfaction, especially for loyalty program members, could improve response rates among hesitant customers.

References

- GeeksforGeeks. (2024, May 17). *Decision tree*. GeeksforGeeks. <https://www.geeksforgeeks.org/decision-tree/>
- Narkhede, S. (2022, March 5). Understanding AUC - ROC Curve - Towards Data Science. *Medium*. <https://towardsdatascience.com/understanding-auc-roc-curve-68b2303cc9c5>
- *Machine Learning Random Forest Algorithm* - Javatpoint. (n.d.-b). [www.javatpoint.com. https://www.javatpoint.com/machine-learning-random-forest-algorithm](https://www.javatpoint.com/machine-learning-random-forest-algorithm)
- GeeksforGeeks. (2023, February 6). *XGBoost*. GeeksforGeeks. <https://www.geeksforgeeks.org/xgboost/>

Appendix

Colab note book:

<https://colab.research.google.com/drive/1jFY0ehFPd5C8DkcU3r8Vj8kmMfAjoisl?usp=sharing>

Kaggle link to Dataset:

<https://www.kaggle.com/datasets/ahsan81/superstore-marketing-campaign-dataset/data>