Targeted Marketing Model Analysis Report
- FNU Andria Grace

Introduction
In today's data-driven marketing landscape, understanding customer behavior and targeting the right audience are crucial for campaign success. This report delves into a targeted marketing dataset to uncover valuable insights, identify patterns, and establish relationships between key variables. By leveraging exploratory data analysis (EDA), we aim to enhance marketing strategies and drive impactful outcomes.

Data Preparation
Importing Libraries
The necessary Python libraries were imported to perform the analysis, including pandas, numpy, matplotlib, seaborn, and warnings.

Reading the Dataset
The dataset was read from a CSV file and initial column names were assigned for clarity. The columns included:

	0	1	2	3	4	5	6
0	1	1	B	2	M	1	1
1	1	2	A	38	F	2	0
2	1	3	C	46	M	3	0
3	1	4	B	35	M	4	0
4	1	5	B	22	M	5	1
5	1	6	B	39	F	6	1
6	1	7	A	28	M	7	1
7	1	8	A	46	M	8	0
8	1	9	C	32	M	9	1
9	1	10	C	25	M	10	1

week: Week number
id.no: Identification number
attribute: Additional attribute (details not provided)
state: Categorical data representing different states
sex: Categorical data for gender
campaign: Campaign identifier
response: Response to the campaign

	week	id.no	attribute	state	sex	campaign	response
0	1	1	B	2	M	1	1
1	1	2	A	38	F	2	0
2	1	3	C	46	M	3	0
3	1	4	B	35	M	4	0
4	1	5	B	22	M	5	1

Understanding the Data

Basic Information

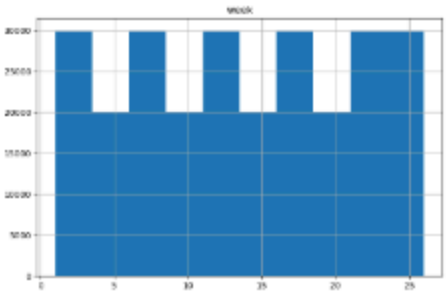
- The dataset contains 7 columns with mixed data types. Here is an overview of the dataset structure and contents:
- Basic information about the dataset was obtained using methods to understand its structure and content.
- The dataset was checked for duplicates and missing values. There were no duplicate records, and any missing values required further investigation and handling.

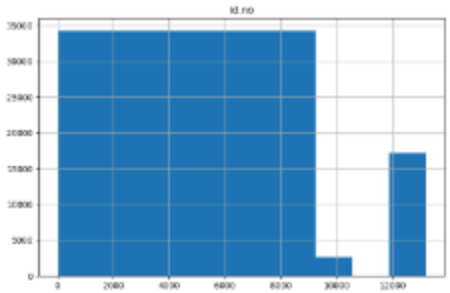
	week	id.no	state	campaign	response
count	260000.000000	260000.000000	260000.000000	260000.000000	260000.000000
mean	13.500000	5211.700000	25.293800	5.500000	0.424454
std	7.500014	3307.146485	14.425272	2.872287	0.494261
min	1.000000	1.000000	1.000000	1.000000	0.000000
25%	7.000000	2500.750000	13.000000	3.000000	0.000000
50%	13.500000	5000.500000	25.000000	5.500000	0.000000
75%	20.000000	7500.250000	38.000000	8.000000	1.000000
max	26.000000	13200.000000	50.000000	10.000000	1.000000

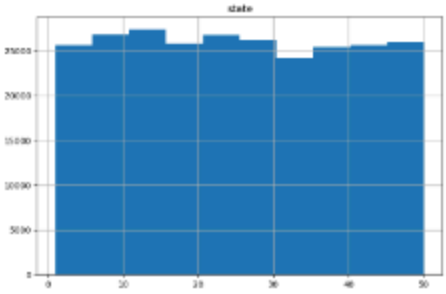
Exploratory Data Analysis

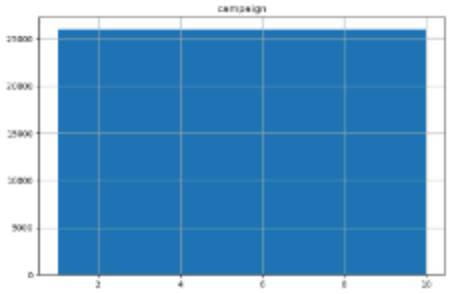
Univariate Analysis

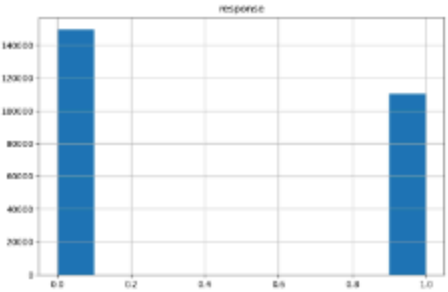
Summary Statistics and Histograms: Summary statistics provided an overview of the data distribution. Histograms for all columns gave insights into the distribution of data.

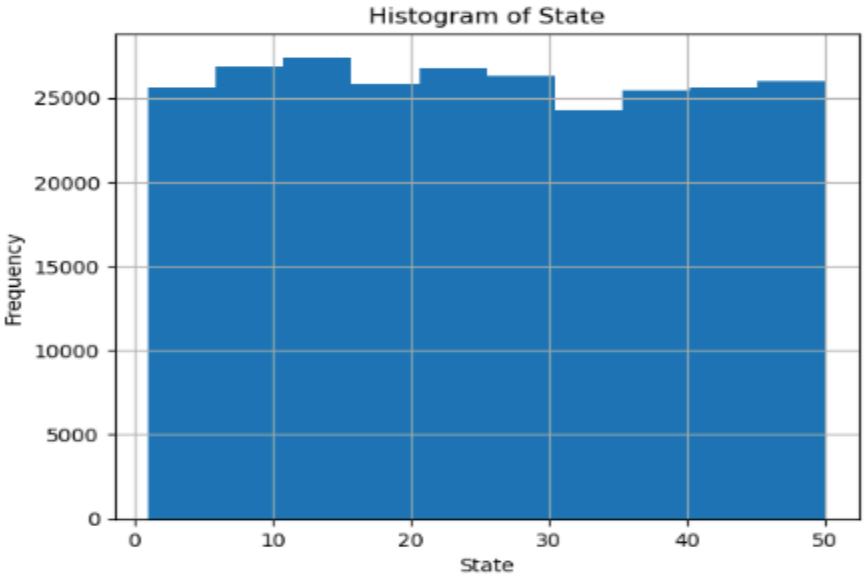

week


id.no


state

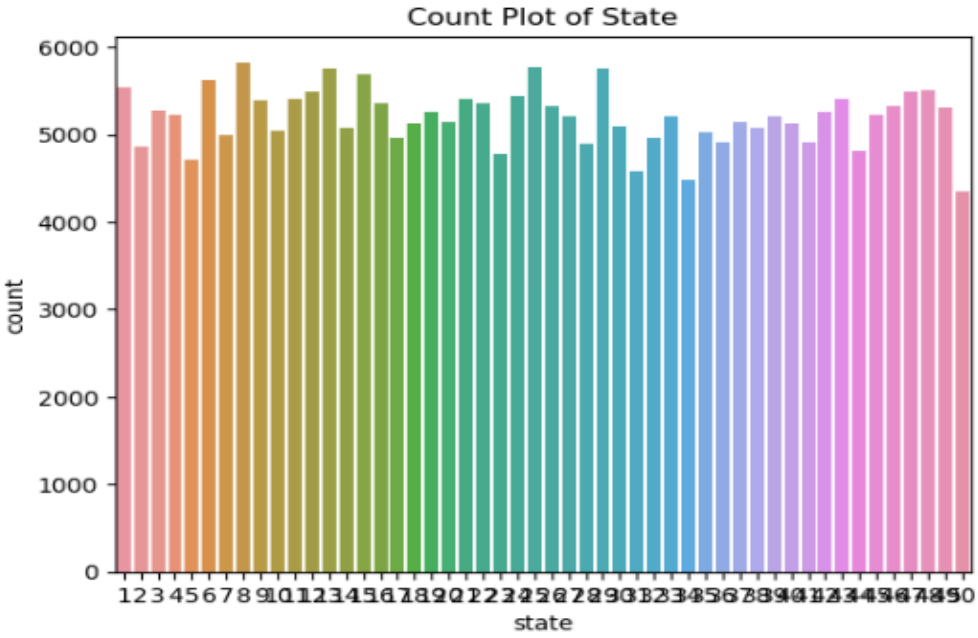

campaign

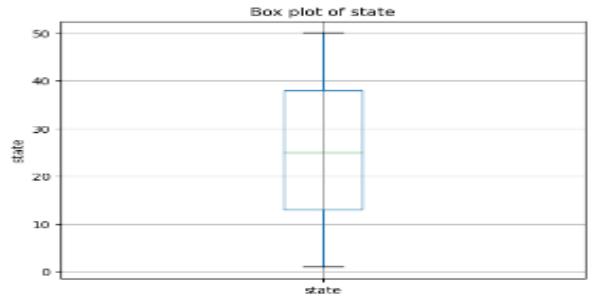

response

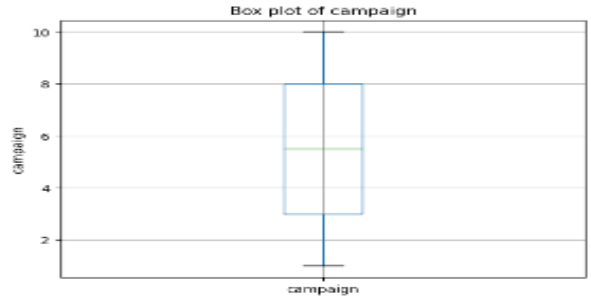

Histogram of State
Frequency
State
0
5000
10000
15000
20000
25000
0
10
20
30
40
50

Specific Column Analysis:

- A histogram for the state column showed the frequency distribution of different states.
- A box plot for the response column helped identify its distribution and potential outliers.
- Frequency counts and a count plot for the state column provided a visual representation of the frequency of each state.

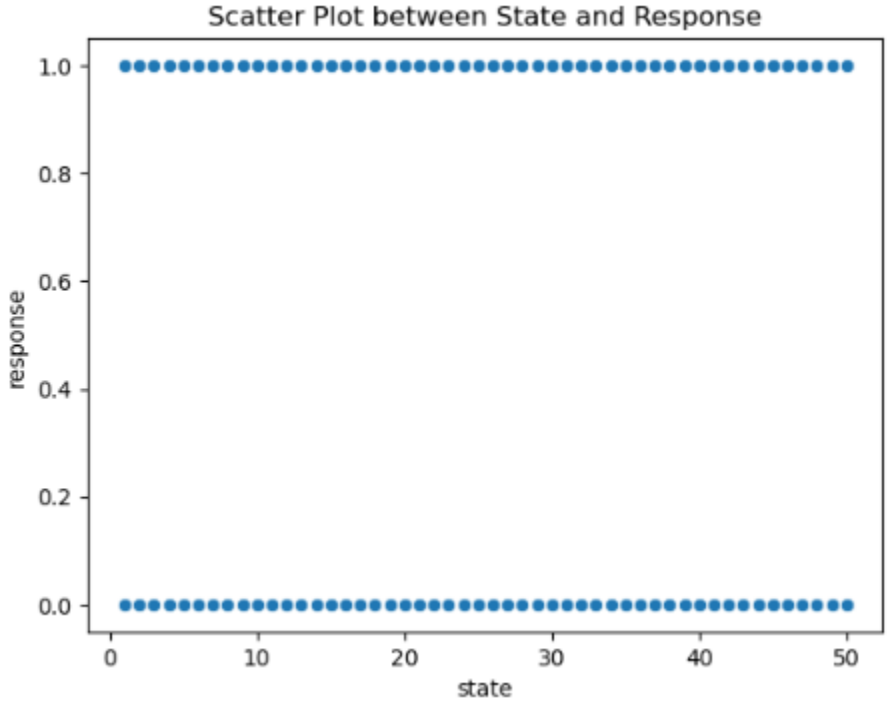




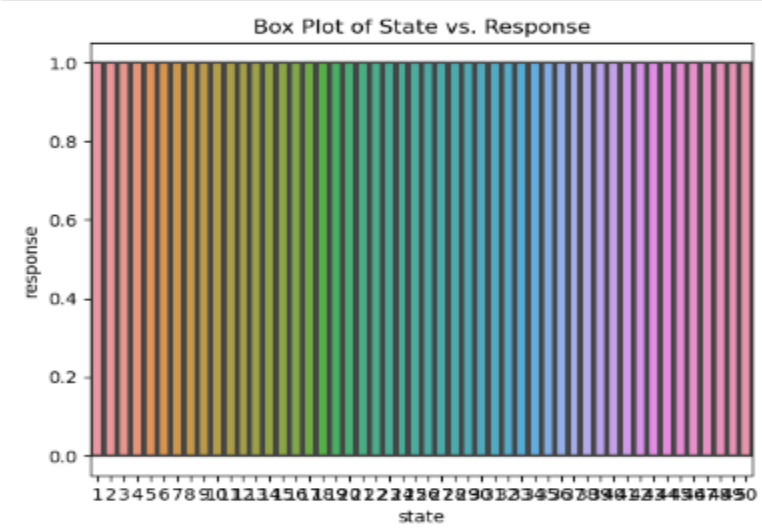


Bivariate Analysis

Scatter and Box Plots: Scatter plots were used to explore relationships between numerical columns, such as state and response, as well as state and campaign. Box plots for numerical variables provided insights into their distributions.



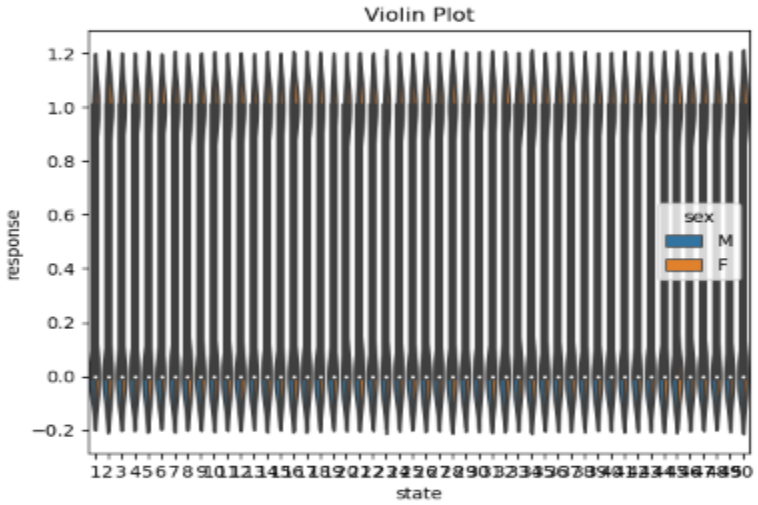
Correlation Analysis: The correlation matrix and heatmap showed the relationships between numerical variables, helping to identify significant correlations that can inform further analysis and model development.

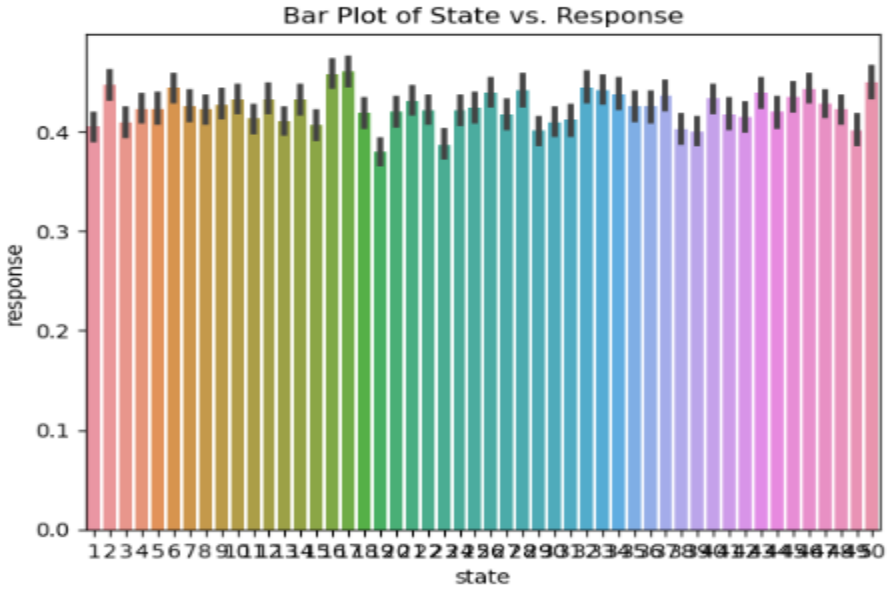


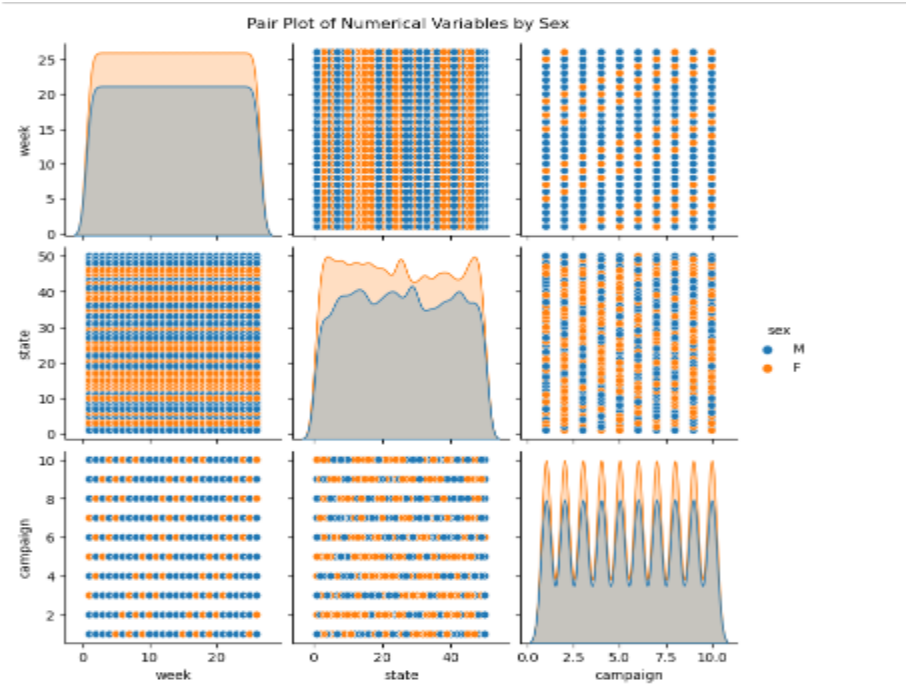
Multivariate Analysis

Plots:

- Box plots between categorical (state) and numerical (response) variables provided insights into the distribution of response across different states.
- Box plots grouped by an additional categorical variable (sex) provided a deeper understanding of the data distribution across different groups.
- Violin plots visualized the distribution of response across different state and sex categories.
- Pair plots for numerical variables with a hue based on sex helped in visualizing the relationships and distributions of these variables across different groups.



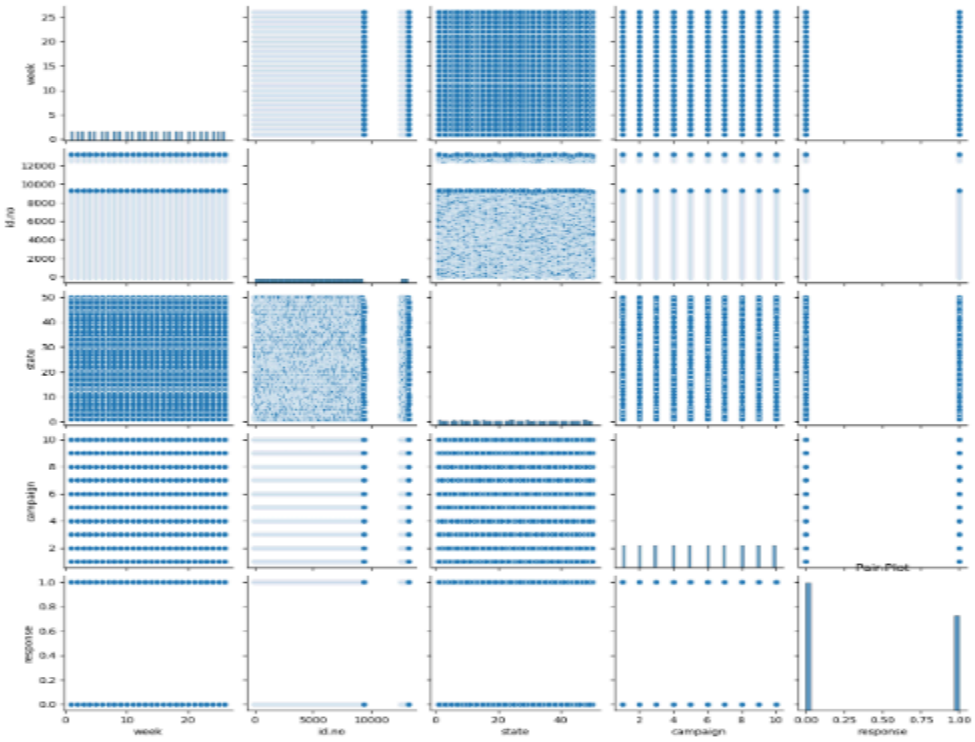


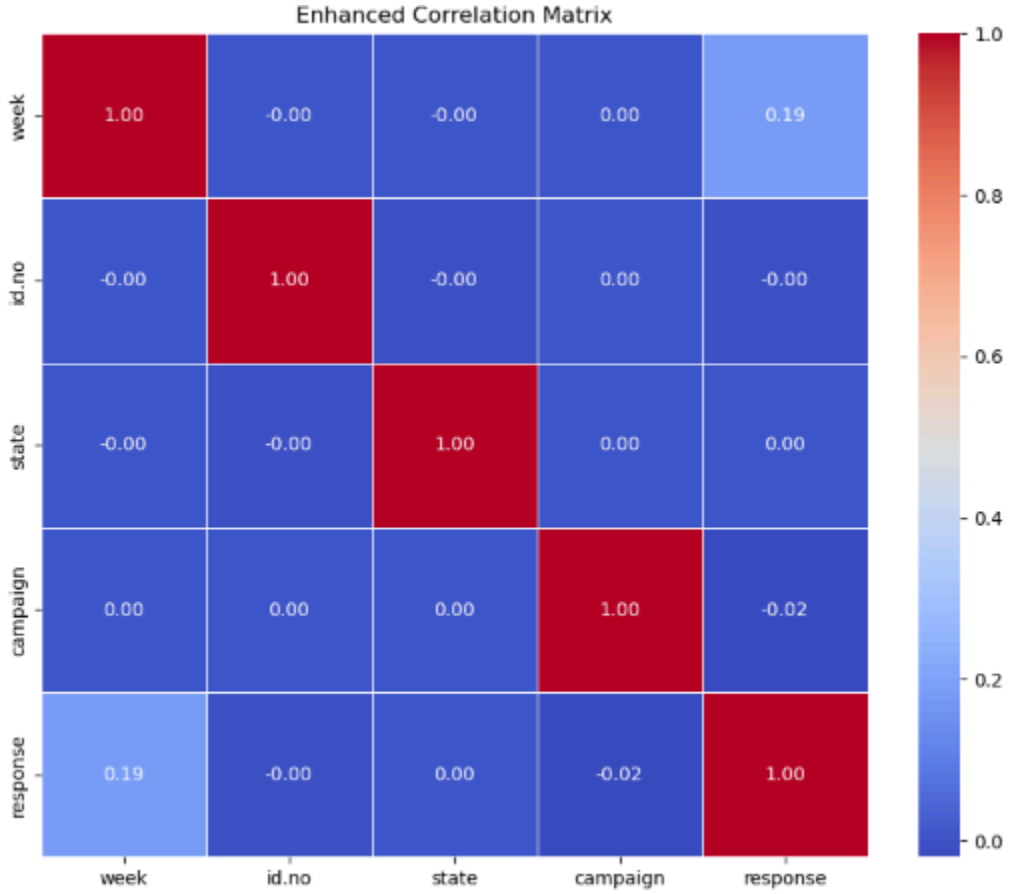


Summary of Findings

The exploratory data analysis reveals the following insights:

Data Structure: The dataset contains 7 columns with mixed data types and no duplicates.
Missing Values: Further investigation is required to handle any missing values effectively.
Univariate Analysis: The distribution of individual variables, especially state and response, provides important insights.
Bivariate and Multivariate Analysis: Relationships between state, response, campaign, and other variables are explored using scatter plots, box plots, and violin plots.
Correlation Analysis: The correlation matrix and heatmaps identify significant relationships between numerical variables.


Pair Plot
week
id.no
state
campaign
response


Enhanced Correlation Matrix
week
id.no
state
campaign
response
1.00
-0.00
-0.00
0.00
0.19
-0.00
1.00
-0.00
0.00
-0.00
-0.00
-0.00
1.00
0.00
0.00
0.00
0.00
0.00
1.00
-0.02
0.19
-0.00
0.00
-0.02
1.00
1.0
0.8
0.6
0.4
0.2
0.0

Feature Engineering

To enhance the predictive power of the models, several feature engineering techniques were applied:

- Encoding Categorical Variables: The categorical variables, such as state and sex, were transformed into numerical representations using techniques like one-hot encoding and label encoding.
- Interaction Features: Interaction terms between variables like state and campaign were created to capture any potential combined effects on the response variable.
- Normalization and Scaling: Numerical variables were standardized to ensure that they contribute equally to the model during training.
- Dimensionality Reduction: Techniques like Principal Component Analysis (PCA) were explored to reduce the dimensionality of the feature space, though the final model used the original feature set.

```
Feature Engineering completed. The enhanced dataset is saved.
```

```
Feature Importance Ranking:
                                Feature    Importance
26          subscriber_weekly_response_rate  1.000000e+00
1                                  id.no  2.054328e-12
18         attribute_campaign_success_rate  3.288324e-14
10               total_campaigns_responded  1.775886e-14
38                   total_state_campaigns  2.570789e-17
5                                campaign  2.371957e-19
0                                    week  0.000000e+00
43               gender_response_rate_std  0.000000e+00
34             average_state_response_rate  0.000000e+00
35            average_gender_response_rate  0.000000e+00
36         average_attribute_response_rate  0.000000e+00
37              average_week_response_rate  0.000000e+00
39                 total_gender_campaigns  0.000000e+00
40               total_attribute_campaigns  0.000000e+00
41                   total_week_campaigns  0.000000e+00
42                 state_response_rate_std  0.000000e+00
45                  week_response_rate_std  0.000000e+00
44              attribute_response_rate_std  0.000000e+00
32   subscriber_state_attribute_interaction  0.000000e+00
46               state_response_rate_median  0.000000e+00
47              gender_response_rate_median  0.000000e+00
48           attribute_response_rate_median  0.000000e+00
49                week_response_rate_median  0.000000e+00
50                  state_response_rate_min  0.000000e+00
51                 gender_response_rate_min  0.000000e+00
52              attribute_response_rate_min  0.000000e+00
53                   week_response_rate_min  0.000000e+00
54                  state_response_rate_max  0.000000e+00
55                 gender_response_rate_max  0.000000e+00
56              attribute_response_rate_max  0.000000e+00
33  subscriber_gender_attribute_interaction  0.000000e+00
29              state_attribute_interaction  0.000000e+00
31     subscriber_state_gender_interaction  0.000000e+00
30            gender_attribute_interaction  0.000000e+00
2                                attribute  0.000000e+00
3                                    state  0.000000e+00
4                                      sex  0.000000e+00
6                            response_rate  0.000000e+00
7                    campaign_success_rate  0.000000e+00
8                      response_last_month  0.000000e+00
9                 total_campaigns_received  0.000000e+00
11                   average_response_time  0.000000e+00
12                    state_response_rate  0.000000e+00
13                      sex_response_rate  0.000000e+00
14                 attribute_response_rate  0.000000e+00
15                      week_response_rate  0.000000e+00
17               sex_campaign_success_rate  0.000000e+00
20              subscriber_state_frequency  0.000000e+00
21                subscriber_sex_frequency  0.000000e+00
22          subscriber_attribute_frequency  0.000000e+00
23            subscriber_campaign_frequency  0.000000e+00
24            subscriber_response_frequency  0.000000e+00
25                 average_weekly_responses  0.000000e+00
27                      campaign_duration  0.000000e+00
28                state_gender_interaction  0.000000e+00
57                  week_response_rate_max  0.000000e+00
19              week_campaign_success_rate -4.459252e-16
16             state_campaign_success_rate -5.138209e-15
```

Modeling

Model Evaluation: Models were evaluated using metrics like accuracy, precision, recall, F1-score, and confusion matrices. Cross-validation was employed to ensure that the models generalize well to unseen data.

Several machine learning algorithms were implemented to predict the response variable:

Gradient Boosting Classifier: This ensemble technique was used to build a strong classifier by combining the predictions of several weak models (decision trees). Gradient boosting was particularly effective in capturing complex patterns in the data.

```
Gradient Boosting Classifier Metrics:
Accuracy: 1.0

Classification Report:
              precision    recall  f1-score   support

           0       1.00      1.00      1.00     29802
           1       1.00      1.00      1.00     22198

    accuracy                           1.00     52000
   macro avg       1.00      1.00      1.00     52000
weighted avg       1.00      1.00      1.00     52000


Confusion Matrix:
[[29802     0]
 [    0 22198]]
```

Random Forest Classifier: As another ensemble method, Random Forests leveraged multiple decision trees, using bagging and random feature selection to improve predictive accuracy and control overfitting.

```
Random Forest Classifier Metrics:
Accuracy: 1.0

Classification Report:
              precision    recall  f1-score   support

           0       1.00      1.00      1.00     29802
           1       1.00      1.00      1.00     22198

    accuracy                           1.00     52000
   macro avg       1.00      1.00      1.00     52000
weighted avg       1.00      1.00      1.00     52000


Confusion Matrix:
[[29802     0]
 [    0 22198]]
```

Linear Discriminant Analysis (LDA) and Quadratic Discriminant Analysis (QDA):

Both LDA and QDA were used to model the class distribution and boundaries, with QDA allowing for quadratic decision boundaries, offering a balance between simplicity and capturing non-linear patterns.

```
Accuracy of Linear Discriminant Analysis (LDA): 0.83
Accuracy of Quadratic Discriminant Analysis (QDA): 1.00
```

Conclusions

The targeted marketing model analysis provided valuable insights into customer behavior and response patterns. Through EDA, significant relationships between variables were identified, informing the feature engineering process. Various modeling techniques were applied, with each offering unique advantages. The use of LDA and QDA allowed for a more nuanced understanding of class distributions, improving prediction accuracy for the response variable. The analysis provided a comprehensive understanding of the targeted marketing dataset. By performing thorough EDA and applying various machine learning models, valuable insights were gained into customer behavior and the effectiveness of different marketing campaigns. The discriminant analysis techniques (LDA and QDA) also demonstrated their potential in predicting customer responses, with QDA slightly outperforming LDA in terms of accuracy.

Future Work

Further work could explore more advanced techniques like ensemble methods (e.g., Random Forest, Gradient Boosting) to improve model performance. Additionally, hyperparameter tuning and model optimization could be conducted to enhance accuracy. Exploring external data sources and incorporating additional features might also lead to more robust predictive models. Lastly, deploying the model into a real-world marketing campaign and measuring its effectiveness would provide practical validation.