A Deep Dive into Customer Personality Analysis Using Machine Learning Algorithms

Amin Bhavisha, Jinwala Maaz, Narain Dakshita, Nasim Maha, Patel Shivam

***Abstract*— A key component of marketing strategy is customer personality analysis, which aids companies in comprehending the attitudes, preferences, and actions of their target market. The objective of this research is to investigate the relationship between a range of consumer characteristics, including age, income, marital status, education, and marketing campaign response. The distribution of these characteristics and how they affect consumer behavior and responses will also be examined in the investigation. Businesses can use the information to refine their marketing tactics and raise consumer satisfaction levels.**

Keywords- customer personality analysis, marketing strategy optimization, consumer characteristics, relationship analysis, consumer insights.

I. Introduction

1.1 Background

A company's ideal customers are thoroughly examined through customer personality analysis. It assists a company in better understanding its customer base and customizing its offerings to meet the unique requirements, preferences, and worries of various customer segments [1].

The personality, background, and upbringing of a customer have a big impact on how they behave at a firm. It is essential to comprehend consumer behavior to know where your target audience falls in this category [2].

A quantifiable way to comprehend the personality of consumers is through the Big Five concept. It includes the following five essential personality traits of a customer: conscientiousness, extraversion, agreeableness, openness, and neuroticism [3].

1.2 Objectives

This report aims to investigate a company's customer personality by looking at a range of characteristics, including age, income, marital status, education, and more. To comprehend how these characteristics and consumer reactions to marketing campaigns are related, an analysis will be conducted. This will assist the business in better adjusting its marketing tactics to suit the requirements of various consumer groups.

1.3 Importance of Customer Personality Analysis in Marketing

Customer personality analysis is significant in marketing as it allows businesses to create highly personalized content, which is increasingly expected by customers [2]. Understanding consumer personality traits can help identify and draw conclusions about consumer behavior, including preferences, habits, and motivations [3].

By understanding the personalities of their prospects and customers, companies can tailor their marketing communication to build trust and break down resistance to buying [4]. Furthermore, customer personality analysis can help businesses identify key benefits for each group, understand the context of customers' needs, and improve the customer experience [2].

II. Literature Review

1.1 Marketing Correlation

When two or more factors are related, it's called correlation in the marketing context. It is a statistical metric that shows how closely two variables' movements are related. Correlation analysis can be used in marketing to determine how various client attributes relate to one another and how they affect marketing initiatives. Correlation, however, does not always indicate a connection thus, just because two variables are associated, it does not automatically imply that one caused the other.

1.2 Analysis of Customer Personality

Understanding consumer behavior, attitudes, and traits is possible through the application of customer personality analysis. To help firms customize their marketing tactics to match the demands and tastes of their target audience, it entails examining consumer data to find patterns and trends. Businesses may create more successful marketing strategies by using customer personality analysis to better understand the behaviors, preferences, and reactions of their target audience.

1.3 Customer Experience

The term "client experience" describes the thoughts, emotions, and reactions a consumer has when utilizing the goods, services, or personnel of a business. It is a crucial factor in determining the satisfaction and loyalty of customers. Numerous elements, such as the success of marketing efforts, the simplicity of use of a product or service, and the level of customer support, can affect the customer experience. Businesses may increase customer happiness and improve their goods and services by understanding the consumer experience.

1.4 *Characteristics of Personality Impacting Retail Buying Channel*

The choice of retail purchase channels might be influenced by personality factors. For example, because it's more convenient, extroverted customers might favor internet purchasing, but introverted customers would prefer in-store shopping because of the opportunity for face-to-face connection. By knowing how personality qualities affect retail purchase channels, firms may better customize their marketing efforts to appeal to a variety of client personalities.

III. Methodology

1.1 Data Collection

The dataset for this analysis was collected from Kaggle, encompassing information on customer attributes such as ID, Year_Birth, Education, Marital_Status, Income, Kidhome, Teenhome, Dt_Customer, Recency, Wines, Campaign_5, Campaign_1, Campaign_2, Complain, Z_CostContact, Z_Revenue, Response, Age, TotalResponses, and ChildOrTeen. The dataset spans a specific time frame, allowing for a comprehensive analysis of customer behavior.

1.2 Data Preprocessing

The dataset was preprocessed, which included handling missing values, eliminating duplicates, and making sure the data was consistent before it was analyzed. To improve the reliability of the findings, outliers were addressed and categorical variables were appropriately encoded.

1.3 Correlation Analysis Method

A correlation analysis was conducted to explore relationships between customer attributes. The Pearson correlation coefficient was utilized to quantify the strength and direction of linear relationships. Subsequent analyses were informed by the method's insights into which variables show significant correlations.

1.4 Measurement of Campaign Response

Response rates for Campaign_5, Campaign_1 and Campaign_2 were calculated in order to assess consumer reactions to marketing campaigns. This made it easier to investigate in-depth the effectiveness of campaigns and how different customer attributes are related to them.

This methodology ensured a systematic approach to analyzing the dataset, allowing for robust insights into customer behavior and responses to marketing initiatives.

IV. Exploratory Data Analysis

The first step to analyzing this dataset is to import the required libraries. After importing the libraries, the dataset is loaded into the Panda data frame.

V. Data Cleaning and Preprocessing

After the dataset is loaded, the columns 'ID', 'Recency', 'Z_CostContact', and 'Z_Revenue' are dropped as these columns do not contribute to the purpose of our analysis.

To clean the dataset further, all the columns in the dataset are renamed for easier interpretation and analysis of the data. The campaign names are renamed for clear comprehension of the campaign results. “AcceptedCmp1' as 'Campaing_1', 'AcceptedCmp2' as 'Campaing_2', 'AcceptedCmp3' as 'Campaing_3', 'AcceptedCmp4' as 'Campaing_4' and 'AcceptedCmp5' as 'Campaing_5'. Other column names are renamed as well for ease of readability.

After renaming the columns, we check for missing and null data. Any missing values are filled in by using the median value of each column in the data frame. After filling in the missing values, we drop all the null values from the dataset. Now, the dataset is ready for performing data visualization and data algorithm models on it.

VI. Data Visualization

First, a correlation matrix is created to recognize the correlation between the attributes in the dataset. Columns like 'Year_Birth', 'Income', 'Kidhome', 'Teenhome', 'Wines', 'Fruits', 'Meat', 'Fish', 'Sweets', 'Gold', 'Deals', 'Web', 'Catalog', 'Store', 'Web_Visits' are all included in the correlation matrix. The values in the cells represent the correlation coefficient between pairs of these numerical variables.

The catalog seems to be highly positively correlated with meat (+0.72), indicating that the company has specific catalog promotions that heavily feature or target meat products. Customers interested in meat may be more responsive to these catalog promotions, leading to a positive correlation. The correlation matrix is shown in Figure 1.

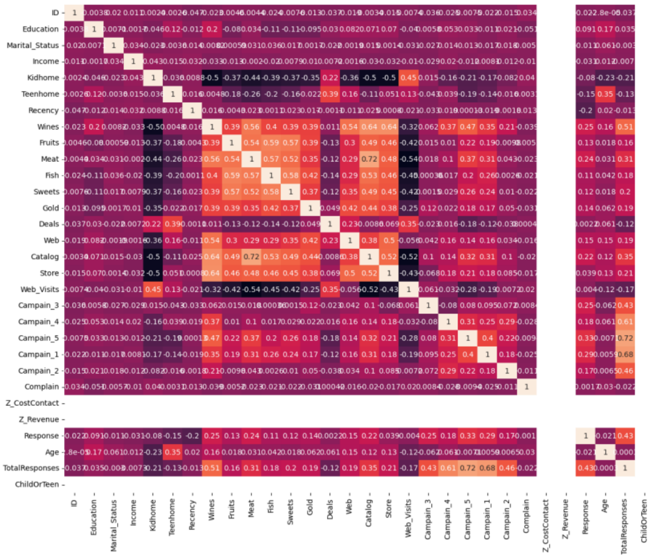

Figure 1. Correlation matrix between each Attribute

Now we tried to assess the age distribution of the people in our dataset. We started by first converting the variable 'Year_Birth' column in the DataFrame df to DateTime format, interpreting the values as years in the '%Y' format.

Then we calculated the age of the people by taking the current year and subtracting the year of birth. It gives us an age distribution plot as shown in Figure 2. It can be noted that ages 40 - 60 form the majority of subjects in our study, with 50 and above people from that age range included in our study.

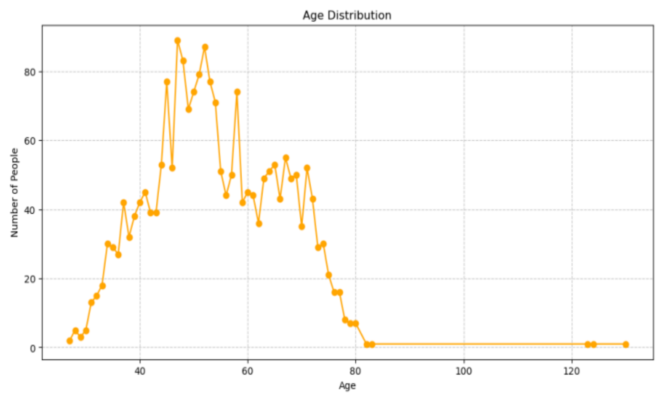

Figure 2. Age Distribution in the dataset

After determining the age distribution present in the dataset, we determine the 'marital status' of the people. To get this data, we first counted the occurrences of each marital status present in the dataset. Then, we calculated the percentage of these counts.

Converting these counts into percentages helped to create a pie chart, as shown in figure 3, to demonstrate the 'marital status' of the people. From the figure above, we see that the highest percentage belong to the marital status "married" of 38.6 % and the lowest percentage belongs to "YOLO" and "Absurd" of 0.1% in the total distribution.

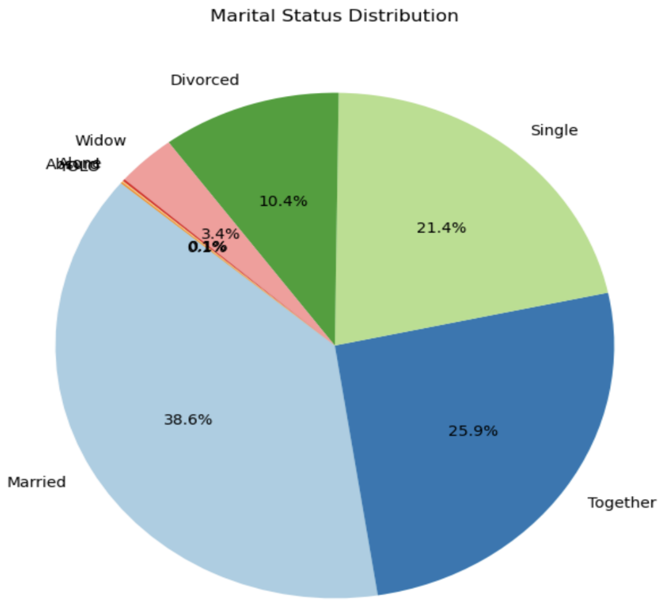

Figure 3. Pie Chart for marital status distribution

Following our pie chart analysis, the team then formed a grouped bar chart to determine the response rate of people with different marital statuses to various marketing campaigns. It can be seen that "Alone" responded positively to Campaign 3, with a response rate of 30%. Absurd responded positively to Campaign 1 and 5, with a response rate of 50%. There is no response for Campaign 2, 4, and from "alone", and for Campaign 2, 3, and 4 from "Absurd". The other categories, "Single", "Together"' "Married", "Divorced" and "Widow" responded to all the campaigns.

The graph shown in Figure 4 assists in unraveling the connection between marital status and the inclination to engage with different promotional campaigns, providing valuable insights into customer behavior.

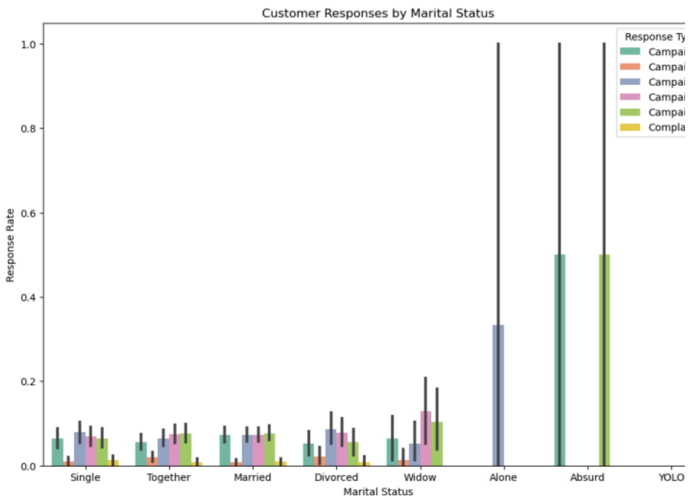

Figure 4. Customer Response to Campaigns by Marital Status

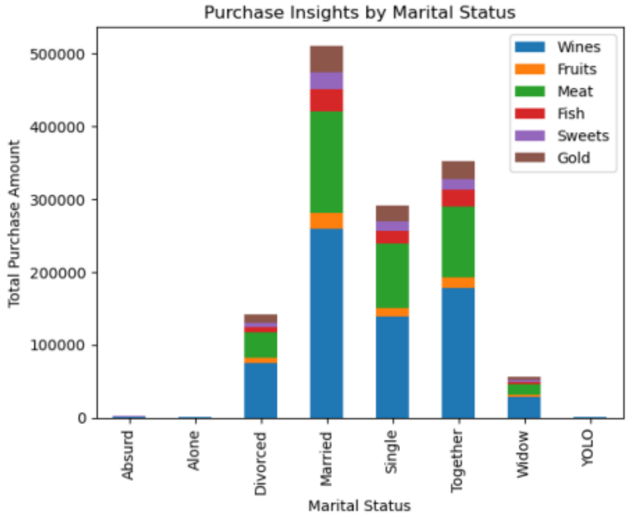

Figure 5. Purchase Insights by Marital Status

Following our analysis of promotional campaign responses according to marital status, the team then engaged in forming a stacked bar chart. Here, we aim to closely examine the shopping preferences of customers with varying marital statuses to gain insights into their purchasing behavior for different items. Gold, wine and meat form the major consumption among all marital statuses.

The major consumers for these items are "married," followed by the marital status "together.". The total purchase amount for gold for married consumers was $500,000 and $390,000 for single consumers.

The graph helps in understanding which product categories contribute more significantly to the overall purchase amounts for different marital statuses and the stacked structure allows for a quick visual comparison of the total purchase amounts across item categories within each marital status group.

This will assist in our marketing approaches and recommendations of different products to people with varying marital statuses. Figure 5 depicts the stack bar chart for Purchase Insights according to marital status.

.

Now, we tried to assess the average consumption of different items based on age groups. As shown in figure 6, the x-axis of the graph represents different age groups, categorized as '0-20', '21-40', '41-60', '61-80', and '81-100'. Each bar in the chart corresponds to a specific age group. It appears that wine formed the highest consumption among all age groups, with the major consumers of wine being the age groups 81-100 and purchase quantities being 650. This was followed by meat with a purchase quantity of approximately 350 for age groups 81-100. Fruit had the lowest consumption among all.

The graph provides insights into how average consumption patterns vary across different age groups. It helps identify which product categories are more popular or preferred by individuals in specific age brackets. This information can be useful for targeted marketing, inventory management, and tailoring products to specific age demographics.

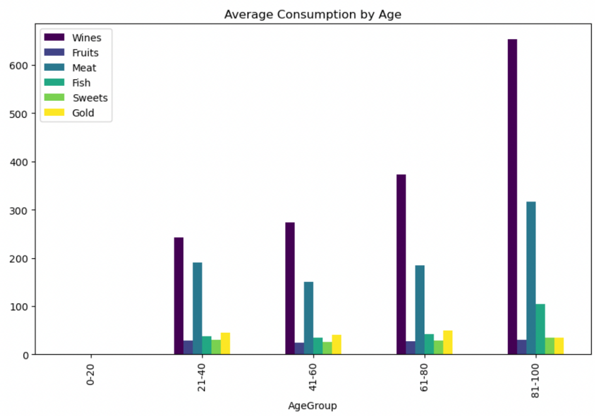

Figure 6. Average Consumption by age

Once we determined the average consumption by age group, we then proceeded to identify which platforms used for promotion have the most engagement amongst specific age groups and are preferred for purchasing items. We have three platforms, namely "Web", "Catalog" and "Store".

We plotted a grouped bar chart, as shown in Figure 7, inferring that the store platform has the maximum engagement rate among all age groups, with age groups 81-100 preferring it the most for their buying decisions and the catalog platform has the least engagement rate among all age groups. This will be helpful to identify which marketing platforms generate the highest engagement among different age groups to enhance our marketing efforts for future campaigns.

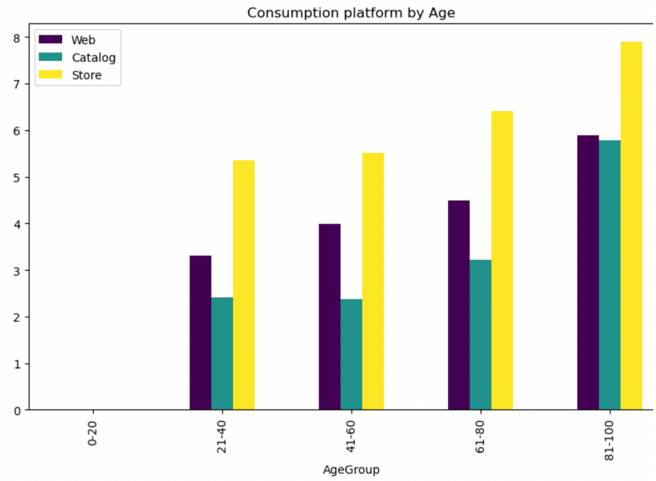

Figure 7. Engagement of platforms by age groups

The graph in figure 8 visualizes the total responses to marketing campaigns categorized by marital status. Each bar corresponds to a specific marital status category and shows the total responses from individuals within that category. It appears that Absurd and Yolo had the highest response rate among all marital statuses, with approximately 50% total response rate. "Together" had the lowest total response rate, which was approximately 10%.

This graph allows for a quick visual comparison of the effectiveness of marketing campaigns across different marital statuses. It would help to identify which marital status groups are more responsive to the marketing efforts. It can guide strategic decisions related to campaign targeting, messaging, and optimization based on the observed responses within each marital status category.

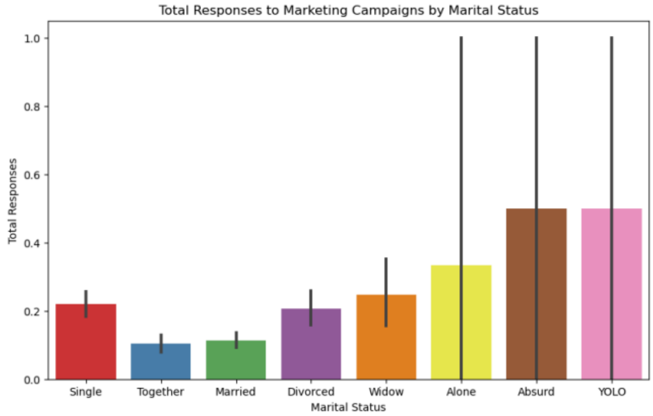

Figure 8. Response to Marketing Campaign by Marital Status

Then we see the educational qualifications of the people of each marital status. The graph is a count plot created using Seaborn, depicting the distribution of individuals with graduation education based on their marital status. The x-axis represents different marital statuses (e.g., single, together, married). Each group on the x-axis corresponds to a specific marital status category.

The y-axis represents the count of individuals. The height of each bar corresponds to the number of individuals within each marital status category based on their educational qualifications. It can be seen that "graduation" was the highest education qualification achieved among all marital statuses, with the "married" category having the highest number of individuals with a graduation degree (approximately 450), the next highest qualification achieved is PHD.

The lowest qualification appears to be the 2nd cycle. This would allow us to determine the different income levels of our target consumers based on the level of education they have achieved. Knowledge of income levels helps in developing an effective pricing strategy. Marketers can set prices that align with the perceived value of the product for customers within specific education and income brackets

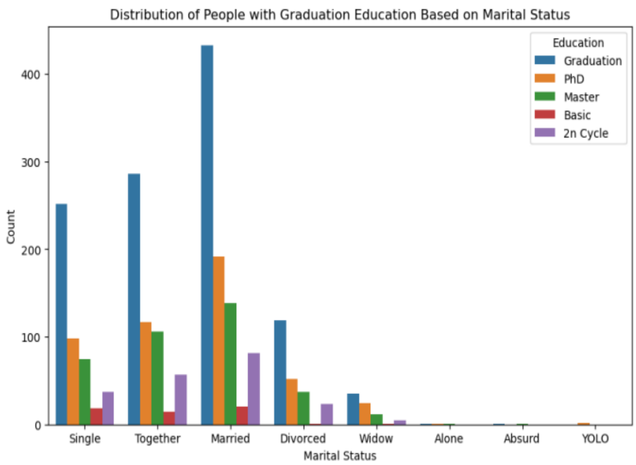

Figure 9. Educational qualifications based on marital status

VII. DATA ALGORITHMS MODEL IMPLEMENTATION AND ANALYSIS

To implement data algorithm models, we used 80% of the data for training and 20% for testing.

A) K- Nearest Neighbor

Classification Report K-Nearest Neighbor

	Precision	*Recall*	*F1-Score*	*Support*
0	0.86	0.95	0.91	379
1	0.40	0.17	0.24	69
Accuracy			0.83	448
Macro Average	0.63	0.56	0.57	448
Weighted Average	0.79	0.83	0.8	448

Table 1. K- Nearest Neighbour

Here we are trying to assess how different customers responded to different campaigns based on their demographic features such as income, age, teen or kid home, marital status, and income levels.

Class 0: Corresponds to one category of the 'Response' variable, and it's represented numerically as 0. In this case, it could be a negative response or an absence of a response.

Class 1: Corresponds to the other category of the 'Response' variable, and it's represented numerically as 1. In this case, it is a positive response or the presence of a response.

Our problem is predicting whether a customer responds positively to a campaign (1) or not (0). Class 1 represents customers who responded positively, and Class 0 represents customers who did not respond positively.

Looking closely at the classification report above, the overall accuracy of our KNN model is 83.25%, meaning the model correctly predicts the class for 83% of the instances.

For class 0, the precision is 0.86. This means that when the model predicts class 0, it is correct 86% of the time. For class 1: precision is 0.40. This means that when the model predicts class 1, it is correct 40% of the time.

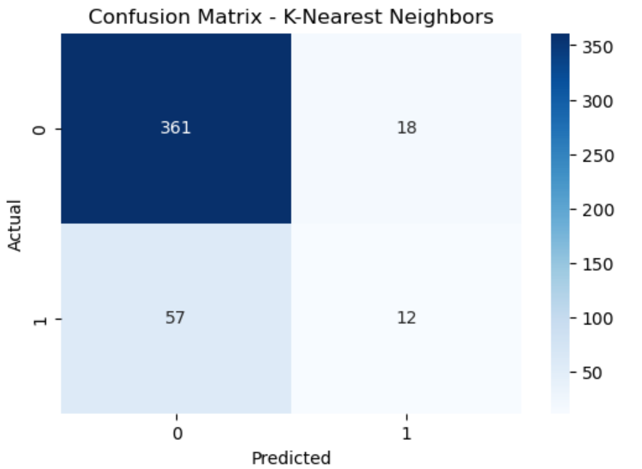

Figure 10. Confusion Matrix for K-Nearest Neighbors

Now, following the accuracy analysis using our KNN algorithm, we further built a confusion matrix to reinforce our understanding as given above. The heatmap visually represents the confusion matrix, making it easier to interpret true positives, true negatives, false positives, and false negatives. The confusion matrix is a table that summarizes the performance of a classification model. It provides a detailed breakdown of the model's predictions compared to the actual class labels. The interpretation is as follows:

True Negatives: The top-left cell (361) indicates the number of instances correctly predicted as no response - Class 0.

False Positives: The top-right cell (18) indicates the number of instances incorrectly predicted as a response - Class 1 but are actually in Class 0**.**

False Negatives: The bottom-left cell (57) indicates the number of instances incorrectly predicted as no response - Class 0 but are actually in Class 1.

True Positives: The bottom-right cell (12) indicates the number of instances correctly predicted as a response - Class 1.
Class 0: The model is effective at identifying instances without a response, with a high true negative rate. The false positive rate is relatively low.
Class 1: The model struggles to correctly identify instances with a response, as indicated by a higher false negative rate. The true positive rate is lower, suggesting that the model misses some instances of Class 1.

B) Naive Bayes

Classification Report Naïve Bayes

	Precision	*Recall*	*F1-Score*	*Support*
0	0.89	0.83	0.86	379
1	0.32	0.43	0.37	69
Accuracy			0.77	448
Macro Average	0.60	0.63	0.61	448
Weighted Average	0.80	0.77	0.78	448

Table 2. Naive Bayes

The overall accuracy of the Naive Bayes model on the test set is approximately 77%. Precision is higher for class 0 (customers who did not respond positively to the marketing campaigns - 0.89) compared to class 1 (customers who did respond positively to the marketing campaigns - 0.32). This indicates that when the model predicts class 0, it is more likely to be correct. Recall (sensitivity) is higher for class 0 (0.83) compared to class 1 (0.43). This indicates that the model is better at capturing instances of class 0.

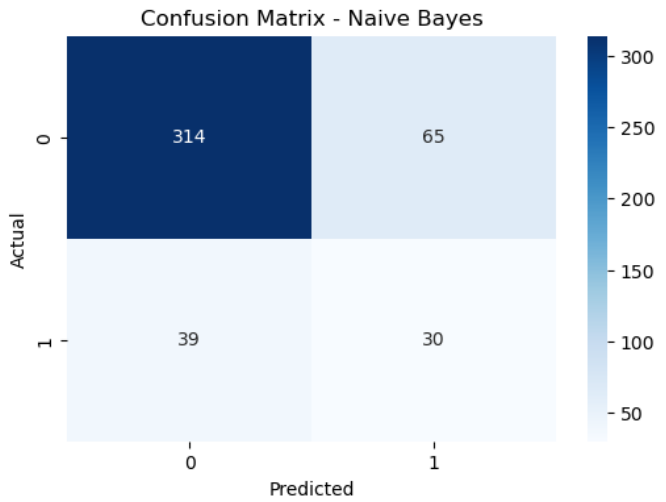

Figure 11. Confusion Matrix for Naive Bayes

The interpretation of the confusion matrix for Naive Bayes is as follows:

True Negatives (TN): The top-left cell (314) instances are actually in no response - Class 0 and correctly predicted as class 0.

False Positives (FP): The top-right cell (65) instances are actually in class 0 but incorrectly predicted as class 1 - response.

False Negatives (FN): The bottom-left cell (39) instances are actually in class 1 but incorrectly predicted as class 0.

True Positives (TP): The bottom-right cell (30) instances are actually in class 1 and correctly predicted as class 1.

The model faces challenges in correctly identifying instances with the condition (Positive Response/Class 1), as indicated by a relatively high number of false negatives and a lower number of true positives.

C) Logistic Regression

Logistic regression is a statistical model used for binary classification problems, where the outcome variable is categorical and has two classes. Despite its name, logistic regression is primarily used for classification rather than regression. The logistic regression model estimates the probability that a given input belongs to a particular class [10].

We first start by standardizing our features using StandardScaler to ensure that all features have a mean of 0 and a standard deviation of 1. Standardization is a common preprocessing step in logistic regression to improve convergence during the optimization process. The accuracy, classification report, and confusion matrix are printed to the console to provide a comprehensive summary of the logistic regression model's performance as below:

Classification Report of Logistic Regression

	Precision	*Recall*	*F1-Score*	*Support*
0	0.87	0.97	0.92	379
1	0.55	0.17	0.26	69
Accuracy			0.85	448
Macro Average	0.71	0.57	0.59	447
Weighted Average	0.82	0.85	0.82	448

Table 3. Logistic Regression Classification

From the classification report above, the accuracy of the logistic regression model on the test set is approximately 85.49%. This indicates that about 85.49% of the instances were correctly classified. Precision for Class 0 is 0.87, meaning that 87% of the time the model predicts class 0 correctly. For class 1, the precision is 0.55. Recall for Class 0 is 0.97, meaning that 97% of Class 0 occurrences are captured by the model. For class 1, the recall is 0.17. The harmonic mean of recall and precision is known as the F1-score. For class 0 it is 0.92, and for class 1 it is 0.26.

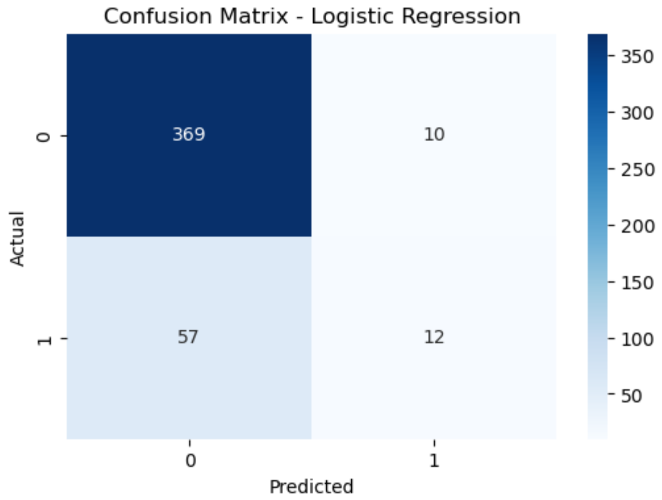

Figure 12. Confusion Matrix for Logistic Regression

The interpretation of the confusion matrix for logistic regression is as follows:

True Negatives (TN): The top-left cell (369) instances are actually in no response - Class 0 and correctly predicted as class 0.
False Positives (FP): The top-right cell (10) instances are actually in class 0 but incorrectly predicted as class 1 - response.
False Negatives (FN): The bottom-left cell (57) instances are actually in class 1 but incorrectly predicted as class 0.
True Positives (TP): The bottom-right cell (12) instances are actually in class 1 and correctly predicted as class 1.

Class 0 (negative response) seems to be well-predicted with a high number of true negatives (TN), indicating that the model is effective at identifying instances without the condition.

Class 1 (positive response) shows challenges, as evidenced by a relatively low number of true positives (TP) and a higher number of false negatives (FN). This suggests that the model has difficulty correctly identifying instances with the condition.

D) Random Forest Classifier

Random Forest classifier is an algorithm for supervised machine learning. It is a method that involves combining numerous base-level models to produce better outcomes [12]. The Random Forest Classifier is well known for overfitting resistance and providing good predictive accuracy[11].

We first set the implementation parameter to set the number of trees in the forest. It is set to 100. A higher number of trees often leads to better performance. The accuracy, classification report and confusion matrix are provided to depict a comprehensive summary of the Random Forest Classifier model's performance, as below:

Classification Report of Random Forest Classifier classification

	Precision	*Recall*	*F1-Score*	*Support*
0	0.88	0.93	0.91	379
1	0.44	0.29	0.35	69
Accuracy			0.84	448
Macro Average	0.66	0.61	0.63	448
Weighted Average	0.81	0.83	0.82	448

Table 4. Random Forest Classifier Classification

From the classification report above, the accuracy of the random forest classifier model on the test set is approximately 84%., which suggests that approximately 84% of the instances were correctly classified. The precision for Class 0 is 0.88, meaning that 88% of the time, the model predicts Class 0 correctly. For class 1, the precision is 0.44. Recall for Class 0 is 0.93, implying that 93% of Class 0 occurrences are captured by the model. For class 1, the recall is 0.29. The harmonic mean of recall and precision is known as the F1-score. For class 0, it is 0.91, and for class 1, it is 0.35.

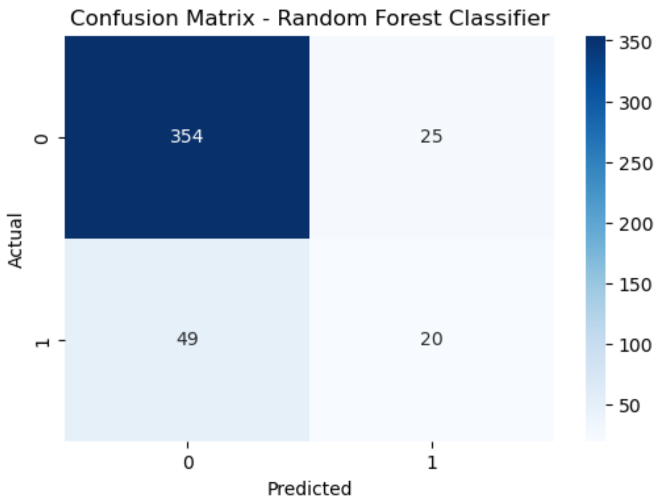

Figure 13. Confusion matrix Random Forest Classifier

The interpretation of the confusion matrix for the random forest classifier is as follows:
True Negatives (TN): The top-left cell (354) instances that are correctly predicted as no response- class 0.

False Positives (FP): The top-right cell (25) instances were incorrectly predicted as class 1 - response when they belong in class 0.

False Negatives (FN): The bottom-left cell (49) instances are truly in class 1 but incorrectly predicted as class 0.

True Positives (TP): The bottom-right cell (20) instances belong to class 1 and are correctly predicted as class 1.

With a high number of true negatives (TN), class 0 (negative response) appears to be well-predicted, suggesting that the model is successful in identifying instances without the condition.

A higher percentage of false negatives (FN) as compared to true positives (TP) indicates difficulties with Class 1 (positive response). This implies that the model has trouble accurately detecting instances that meet the condition.

VII. To select the most suitable model for implementation

After comparing the four models - KNN, Naive Bayes, Logistic Regression, and Random Forest Classifier—it is apparent that Logistic Regression is the most appropriate choice for the particular scenario. Compared to the other models, logistic regression showcases the highest accuracy of 85.04% and demonstrates superior performance, precision, and recall for both classes 0 and 1. When Logistic Regression predicts a positive response, it is usually correct, as indicated by its higher precision for Class 1.

Likewise, the high recall of the model for Class 0 suggests that it is effective in identifying cases where there is no positive response. Looking at these points, it is evident that logistic regression stands out as the most effective choice for the given scenario.

VIII. Conclusion

In conclusion, we need to capitalize on targeted catalog promotions that highlight meat products to improve customer response and marketing strategies. It is reasonable to develop customized marketing promotions, especially for individuals 50 years of age and older, as the majority of people belong to this age group. For the "married" and "together" segments, we must heavily promote gold, wine, and meat due to their higher consumption patterns in the purchase insights. We need to promote wine in ways that appeal to a range of age groups, organize a variety of tasting events, and highlight diverse wine offerings to increase its popularity.

To improve customer engagement outside the age group of 81-100 and to improve the "Store" platform for marketing campaigns, we must implement exclusive in-store promotions, loyalty programs, and a seamless shopping experience. Given their higher response rates, "Absurd" and "YOLO" marital statuses need more research and promotional strategies to identify their unique selling points. Additionally, we need to tailor our marketing campaigns to the income brackets of the various marital status groups and provide exclusive offers to stimulate their curiosity.

After comparing four models, logistic regression emerges as the optimal choice for classification, exhibiting the highest accuracy (85.04%). The goal is to maximize available resources, increase campaign effectiveness, and ultimately improve overall customer satisfaction and engagement.

REFERENCES

1. Widyadwatmaja, A. (2021, November 23). *Customer personality analysis segmentation (clustering)*. Medium. https://medium.com/@andhikaw.789/customer-personality-analysis-segmentation-clustering-1b68a62a61a2
2. Fontanella, C. (2022, August 18). *A beginner's Guide to Customer Behavior Analysis*. HubSpot Blog. https://blog.hubspot.com/service/customer-behavior-analysis
3. Patton, B. (n.d.). The Trait Theory of Personality in Consumer Behavior. Greenbook. https://www.greenbook.org/insights/market-research-trends/the-trait-theory-of-personality-in-consumer-behavior

4. Wagner, E. (2017, April 7). *How to identify your customer's personality*. i7 Marketing - Full Service Marketing Agency. https://www.i7marketing.com/blog/internetmarketing/identify-customers-personality
5. Bateman, C. R., & Valentine, S. R. (2019). Consumers' personality characteristics, judgment of salesperson ethical treatment, and nature of purchase involvement. *Journal of Business Ethics*, *169*(2), 309–331. https://doi.org/10.1007/s10551-019-04312-7
6. Waqas, M., Hamzah, Z. L., & Salleh, N. A. (2020). Customer experience: A systematic literature review and consumer culture theory-based conceptualization. *Management Review Quarterly*, *71*(1), 135–176. https://doi.org/10.1007/s11301-020-00182-w
7. Li, D., & Yu, D. (2022, September 14). *The impact of consumer positive personality on the purchase behavior of Smart Products*. Frontiers in psychology. https://www.ncbi.nlm.nih.gov/pmc/articles/PMC9519063/
8. Hermes, A., & Riedl, R. (2021, December 3). *Influence of personality traits on choice of Retail Purchasing Channel: Literature Review and Research Agenda*. MDPI. https://www.mdpi.com/0718-1876/16/7/179
9. Thorn, James. (2020). Logistic Regression Explained. https://towardsdatascience.com/logistic-regression-explained-9ee73cede081
10. Plaisant, E. (2023). Discovering Random Forest Classifier and Tuning it. Medium. https://medium.com/@eymericplaisant/discovering-random-forest-classifier-and-tuning-it-6142fbf154f5
11. Serokell. (2022, June 27). Random forest classifier: Basic principles and applications. Medium. https://ai.plainenglish.io/guide-to-random-forest-classification-and-regression-algorithms-8e2d578d406b