Contents

[1. Executive Summary 4](#_Toc194036900)

[1.1. Objective 4](#_Toc194036901)

[1.2. Data Overview 4](#_Toc194036902)

[1.3. Methodology 4](#_Toc194036903)

[1.4. Key Findings 5](#_Toc194036904)

[1.5. Marketing Simulation & Business Impact 5](#_Toc194036905)

[2. Introduction 5](#_Toc194036906)

[2.1. Context & Objectives 5](#_Toc194036907)

[2.2. Importance of Customer Personality Analysis 5](#_Toc194036908)

[2.3. Predictive Modeling for Wine & Gold Buyers 5](#_Toc194036909)

[3. Data Overview 6](#_Toc194036910)

[3.1. Description of Project Dataset (Project\_Dataset\_X.sav) 6](#_Toc194036911)

[3.2. Description of Deployment Dataset (Project\_Dataset\_Base.sav) 6](#_Toc194036912)

[4. Exploratory Data Analysis (EDA) 7](#_Toc194036913)

[4.1. Summary Statistics & Data Distribution 7](#_Toc194036914)

[4.2. Relationship Between Variables & Target Variables (Wine Buyer, Gold Buyer) 8](#_Toc194036915)

[4.3. Cross-tabulation Analysis 10](#_Toc194036916)

[5. Model Selection & Training 11](#_Toc194036917)

[5.1. Introduction to Model Selection 11](#_Toc194036918)

[5.2. Brief Explanation of Each Model 11](#_Toc194036919)

[5.3. Training Process 12](#_Toc194036920)

[6. Evaluation of Model Performance 12](#_Toc194036921)

[6.1. EvaluationResults 12](#_Toc194036922)

[6.2. Key Findings 13](#_Toc194036923)

[6.3. Conclusion 13](#_Toc194036924)

[7. Calculating Combined Scores for Deployment Base 13](#_Toc194036925)

[7.1. Overview 13](#_Toc194036926)

[7.2. Selection of Best Models 14](#_Toc194036927)

[7.3. Preparing the Deployment Dataset 14](#_Toc194036928)

[7.4. Feature Scaling: 14](#_Toc194036929)

[7.5. Probability Predictions**:** 14](#_Toc194036930)

[7.6. Combining the Scores**:** 14](#_Toc194036931)

[7.7. Final Output**:** 15](#_Toc194036932)

[8. Simulating Operational Marketing Action on the Top 20% Customers 15](#_Toc194036933)

[8.1. Overview 15](#_Toc194036934)

[8.2. Sorting the Customers by Combined Score**:** 15](#_Toc194036935)

[8.3. Selecting the Top 20% of Customers**:** 15](#_Toc194036936)

[8.4. Defining the Threshold for Expected Buyers**:** 15](#_Toc194036937)

[8.5. Calculating Expected Buyers in the Top 20%**:** 16](#_Toc194036938)

[8.6. Calculating the Proportion of Expected Buyers: 16](#_Toc194036939)

[8.7. Output**:** 16](#_Toc194036940)

[9. Customer Profile Summary for Targeted Clients (Top 20%): 16](#_Toc194036941)

[9.1. Age Distribution: 17](#_Toc194036942)

[9.2. Education Level: 17](#_Toc194036943)

[9.3. Income: 18](#_Toc194036944)

[9.4. Family Dynamics: 18](#_Toc194036945)

[9.5. Marital Status: 18](#_Toc194036946)

[9.6. Recency (Last Purchase Activity): 18](#_Toc194036947)

[9.7. Web Visits (Customer Engagement): 18](#_Toc194036948)

[9.8. SW1 Probability (Wine Buyer Probability): 19](#_Toc194036949)

[9.9. SG1 Probability (Gold Buyer Probability): 19](#_Toc194036950)

[10. Conclusion: 19](#_Toc194036951)

# Executive Summary

## **Objective**

This project focuses on **Customer Personality Analysis** to identify which customers are most likely to buy **wine** and **gold**. By analyzing customer characteristics and purchasing behavior, businesses can optimize marketing strategies by targeting the right customer segments, improving conversion rates, and reducing unnecessary marketing costs.

## **Data Overview**

The analysis is based on **Project\_Dataset\_X.sav**, which contains **1,250 observations** with customer demographics, purchasing habits, and engagement with previous marketing campaigns. Key variables include:

* **Demographics**: Year of birth, education, marital status, income.
* **Household Information**: Number of kids and teenagers in the household.
* **Customer Engagement**: Recency of last purchase, number of web visits per month, responses to past campaigns.
* **Target Variables**: Whether a customer is a **Wine Buyer** or a **Gold Buyer**.

For deployment, the **Project\_Dataset\_Base.sav** is used, which contains the same features but lacks the target labels (**Wine Buyer, Gold Buyer**).

## **Methodology**

To predict customer likelihood of purchasing wine and gold, multiple machine learning models were used:

1. **Logistic Regression** – Establishes a probability score based on linear relationships.
2. **k-Nearest Neighbors (k-NN)** – Classifies customers based on similarity to others.
3. **Naïve Bayes** – Uses probability distributions for classification.

Model performance was evaluated using **accuracy, precision, recall, F1-score, and AUC** to determine the best-performing model for deployment.

## **Key Findings**

* Certain factors, such as **income, number of web visits, and previous campaign responses**, were strong indicators of wine and gold purchases.
* Logistic Regression showed the best balance of accuracy and interpretability, making it the preferred model for deployment.
* Customers were scored based on their likelihood to purchase wine (SW) and gold (SG).

## **Marketing Simulation & Business Impact**

* A **top 20% scoring segment** was selected for targeted marketing efforts.
* Customers in this segment had **higher income, lower recency values, and active online engagement**.
* Based on this analysis, businesses can implement **personalized campaigns**, such as exclusive promotions or loyalty rewards, to increase sales efficiency.

# Introduction

## **Context & Objectives**

This project aims to identify customers most likely to purchase **wine** and **gold** by analyzing their demographics, behaviors, and engagement with previous marketing campaigns. The goal is to build predictive models that help businesses target the right customer segments and optimize marketing efforts.

## **Importance of Customer Personality Analysis**

Customer Personality Analysis allows businesses to focus marketing on high-potential customers, improving efficiency and reducing costs. By segmenting customers based on their likelihood to purchase specific products, companies can enhance product customization and customer engagement.

## **Predictive Modeling for Wine & Gold Buyers**

Using machine learning models like **Logistic Regression**, **k-NN**, and **Naïve Bayes**, we predict the likelihood of customers purchasing wine (SW) and gold (SG). These models will guide targeted marketing strategies, maximizing sales potential by focusing on the most promising segments.

# Data Overview

## **Description of Project Dataset (Project\_Dataset\_X.sav)**

The **Project\_Dataset\_X.sav** contains **1,250 customer observations** with a variety of demographic, household, and behavioral information. This dataset includes two key target variables: **Wine\_buyer** and **Gold\_buyer**, which indicate whether a customer is likely to purchase wine or gold, respectively.

#### **Key Variables in the Dataset:**

* **ID**: Unique identifier for each customer.
* **Year\_Birth**: Year of birth of the customer.
* **Education**: Education level of the customer (e.g., Primary, Secondary, Tertiary).
* **Marital\_Status**: Marital status of the customer (e.g., Single, Married).
* **Income**: Annual income of the customer.
* **Kidhome**: Number of children in the household.
* **Teenhome**: Number of teenagers in the household.
* **Dt\_Customer**: Date when the customer became a client.
* **Recency**: Number of days since the customer's last purchase.
* **NumWebVisitsMonth**: Number of web visits per month.
* **AcceptedCmp1, AcceptedCmp2, AcceptedCmp3**: Binary variables indicating whether the customer accepted each of three previous campaigns.
* **Complain**: Whether the customer has lodged a complaint (binary).
* **Response**: Whether the customer responded to the last marketing campaign (binary).
* **Wine\_buyer**: Target variable indicating if the customer is likely to buy wine.
* **Gold\_buyer**: Target variable indicating if the customer is likely to buy gold.

## **Description of Deployment Dataset (Project\_Dataset\_Base.sav)**

The **Project\_Dataset\_Base.sav** contains the same set of features as the **Project\_Dataset\_X.sav**, excluding the target variables (**Wine\_buyer** and **Gold\_buyer**). This dataset is used for deploying the model and calculating the predicted scores for wine and gold buyers.

#### **Key Variables**

* The dataset includes **demographics, household information, and customer engagement features**, which can be used to predict customer behavior.
* **No target variables** (Wine\_buyer, Gold\_buyer) are present in this dataset, as these are to be predicted based on the features.

# Exploratory Data Analysis (EDA)

## **Summary Statistics & Data Distribution**

In this section, we will provide an overview of the **summary statistics** and **data distribution** of the key variables in the dataset. The summary statistics will give insights into the central tendency, spread, and skewness of the data.

#### **Key Variables Summary**

**Summary of Key Continuous Variables:**

1. **Income**:

Average income is 51,855, with a wide range (1,730 to 162,397), reflecting significant income variability across customers.

1. **Recency** (Days since last purchase):

Average of 48.92 days, with a high spread (standard deviation = 29.26), indicating both recent and more distant customer engagements.

1. **NumWebVisitsMonth**:

On average, customers visit the website 5.30 times per month, suggesting moderate online engagement.

1. **Wine\_buyer**:

About 40% of customers buy wine, reflecting a notable segment of wine purchasers.

1. **Gold\_buyer**:

Around 37% of customers purchase gold, showing a strong interest in this product category as well.

**Categorical Variables**:

1. **Education:**

The education level of most customers is high, with **Graduation** being the most common category (627 out of 1250 customers).

1. **Marital\_Status:**

The majority of customers are **Married** (484 out of 1250), followed by other categories such as **Single** or **Divorced**.

## **Relationship Between Variables & Target Variables (Wine Buyer, Gold Buyer)**

Categorical Variables Encoding

**Education Encoding**: We applied ordinal encoding to the **'Education'** variable, which has an inherent order. The mapping was as follows:

* 'Basic' → 0
* '2n Cycle' → 1
* 'Graduation' → 2
* 'Master' → 3
* 'PhD' → 4

**Marital Status Encoding**: We applied one-hot encoding to the **'Marital\_Status'** variable to convert it into numeric format, with separate columns for each marital status category (e.g., 'Marital\_Single', 'Marital\_Married').

**Reason for Encoding**: Categorical variables need to be encoded before correlation analysis and model training to ensure they are represented numerically, allowing them to be used in statistical analysis and machine learning models.

#### Key Correlation Findings:

1. **Income is a strong predictor of purchases**

It has a **high positive correlation with Wine Buyer (0.67)** and **Gold Buyer (0.40)**, meaning higher-income customers are more likely to buy these products.

1. **Children in the household reduce buying likelihood**

**Kidhome has a strong negative correlation** with both **Wine Buyer (-0.51)** and **Gold Buyer (-0.39)**, suggesting customers with young children are less likely to purchase.

1. **Website visits and online engagement impact buying**

**NumWebVisitsMonth is negatively correlated** with **Wine Buyer (-0.37)** and **Gold Buyer (-0.23)**, indicating that frequent website visits do not necessarily translate into purchases.

1. **Marketing campaign acceptance relates to wine and gold purchases**

Customers who accepted **previous campaigns (AcceptedCmp1: 0.29, AcceptedCmp3: 0.02 for Wine Buyer)** were more likely to be buyers, with **AcceptedCmp1 showing a stronger link** to both wine and gold purchases.

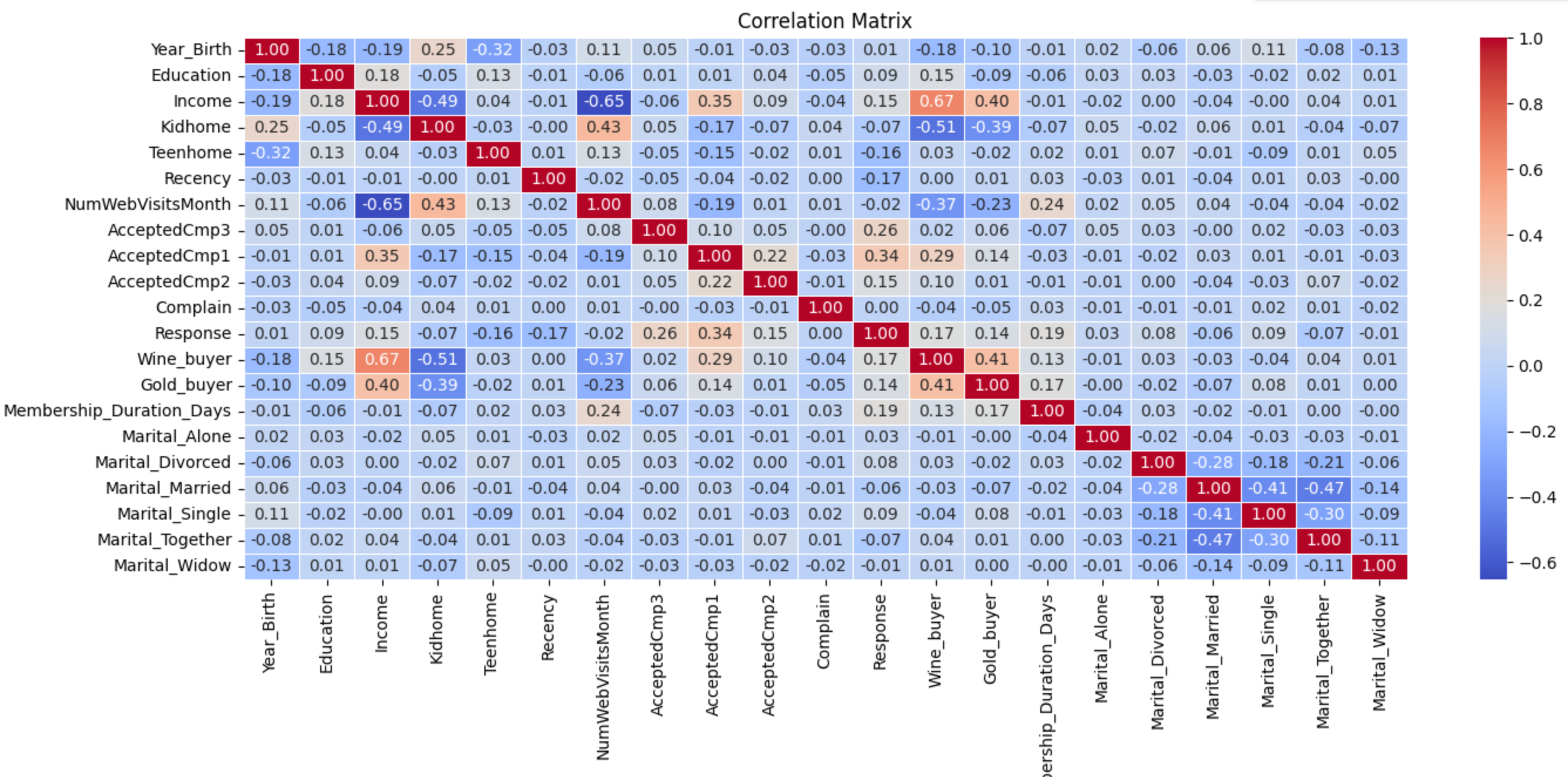
1. **Age plays a role in purchasing behavior**

**Year of Birth is negatively correlated with both Wine Buyer (-0.18) and Gold Buyer (-0.10)**, suggesting older customers are more likely to make purchases.

1. **Longer membership duration shows a slight positive impact**

**Membership Duration correlates positively with both Wine Buyer (0.13) and Gold Buyer (0.17)**, indicating loyalty over time might contribute to purchasing behavior.

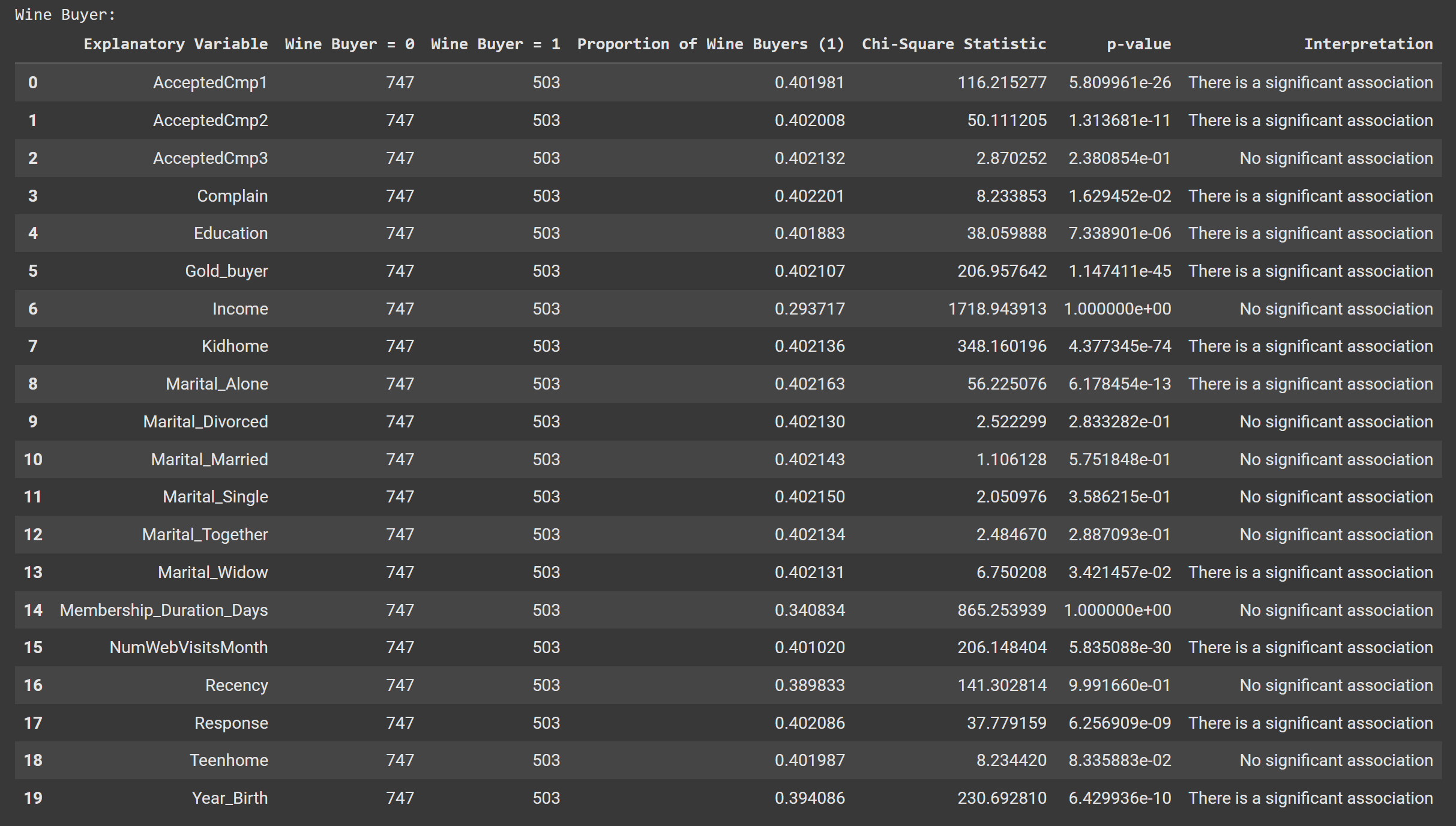
#### Correlation Heatmap



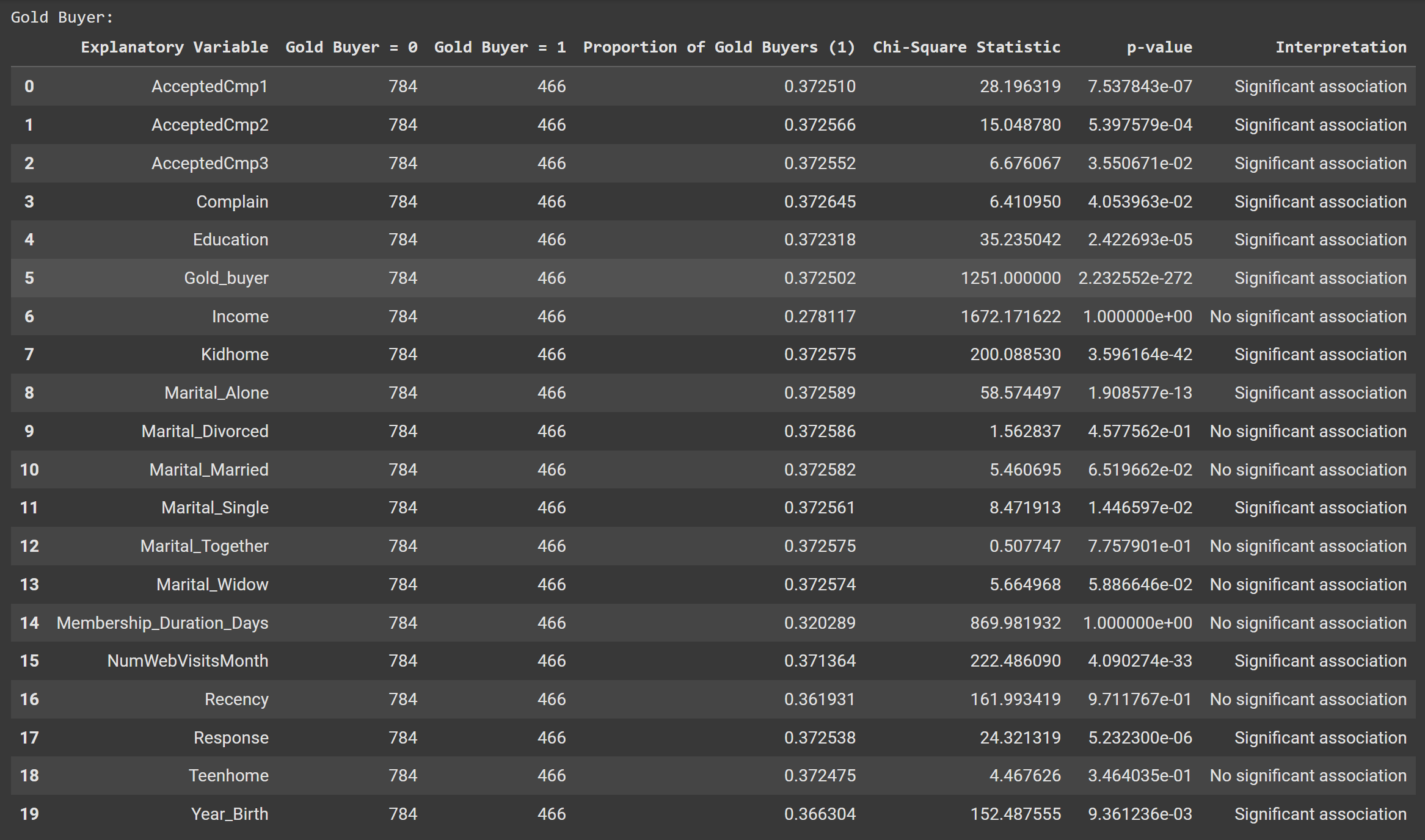
## **Cross-tabulation Analysis**

To further explore the relationships between categorical variables and the target variables, **cross-tabulation analysis** will be used. This analysis shows the frequency distribution of the categories of two variables and provides insights into their joint distributions.

#### Wine Buyer Cross Tabulation



#### Gold Buyer Cross Tabulation



# Model Selection & Training

## **Introduction to Model Selection**

To classify customers as **Wine Buyers (SW1)** and **Gold Buyers (SG1)**, multiple models were tested to compare their performance and determine the most effective algorithm. Since different models have unique advantages, using multiple approaches allows for a more comprehensive evaluation.

* **Logistic Regression**
* **k-Nearest Neighbors (k-NN)**
* **Naïve Bayes**

By applying these models, we aim to determine which one performs best in predicting customer buying behavior.

## **Brief Explanation of Each Model**

* **Logistic Regression** estimates the probability of a binary outcome using the logistic function, assuming a linear relationship between input variables and the log-odds of the target, making it suitable for datasets with clear linear separation.
* **k-Nearest Neighbors (k-NN)** is a non-parametric, instance-based method that classifies data based on the majority class of the nearest **k** neighbors, where a lower **k** captures finer details but may overfit, while a higher **k** generalizes better; it works well with non-linear decision boundaries but requires proper feature scaling.
* **Naïve Bayes** is a probabilistic classifier based on **Bayes' Theorem**, assuming feature independence; despite this assumption being rarely true, it performs well in practice, is highly efficient even with large datasets, and is particularly useful for categorical data classification.

## **Training Process**

#### **Data Splitting**

* The dataset was split into **training (80%)** and **testing (20%)** sets using train\_test\_split.
* The same training and test sets were used for both target variables (**Wine Buyer** and **Gold Buyer**) to ensure consistency in evaluation.

#### **Feature Scaling**

* Since **Logistic Regression** and **k-NN** are sensitive to feature scales, StandardScaler was applied to normalize the input variables.
* **Naïve Bayes** does not require scaling, but for uniformity, it was trained on the scaled data as well.

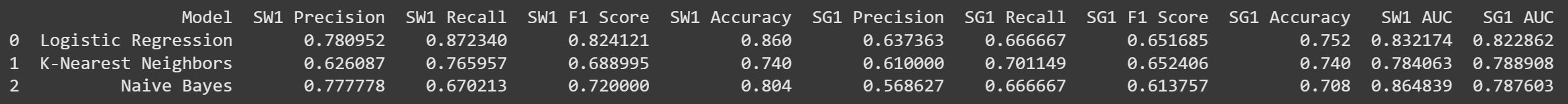
#### **Model Training**

* Each model was trained separately on **Wine Buyer (SW1)** and **Gold Buyer (SG1)** labels.
* Hyperparameters were kept at default values initially, with potential tuning for improved performance.

# Evaluation of Model Performance

## **Evaluation Results**

To assess the performance of different classification models for predicting **Wine Buyers (SW1)** and **Gold Buyers (SG1)**, we analyzed their precision, recall, F1-score, and accuracy. The results for each model are summarized in the table below:



## **Key Findings**

#### **Key Observations**:

* **Logistic Regression** performed well on **SW1 (Wine Buyer)** with high **Precision** (0.781) and **Recall** (0.872), achieving an **F1 Score** of 0.824 and **Accuracy** of 0.860. For **SG1 (Gold Buyer)**, the performance was slightly lower but still respectable, with an **Accuracy** of 0.752 and **AUC** of 0.832.
* **K-Nearest Neighbors** (KNN) showed a balanced performance across both targets. It achieved **SW1** accuracy of 0.740 and **SG1** accuracy of 0.740. However, its **Precision** for **SW1** was lower compared to Logistic Regression. The **AUC** scores were 0.784 for **SW1** and 0.789 for **SG1**, indicating reasonable model discrimination.
* **Naive Bayes** demonstrated solid performance for **SW1**, with a high **AUC** (0.864839), but it was less effective for **SG1** with **Precision** and **F1 Score** lower than the other models. Still, its **Accuracy** and **AUC** for both targets were commendable.

## **Conclusion**

* **Logistic Regression** performed the best for **SW1** and **SG1**, with the highest **F1 Score** for both and a strong **AUC** for both target variables.
* **K-Nearest Neighbors** showed competitive **AUC** scores but didn't excel in **Precision** for **SW1** compared to **Logistic Regression**.
* **Naive Bayes** provided the highest **SW1 AUC** but didn't perform as well for **SG1**.

# Calculating Combined Scores for Deployment Base

## **Overview**

In this step, we focused on applying the **best models** for **SW1 (Wine Buyer)** and **SG1 (Gold Buyer)** to the deployment dataset and calculating **combined scores** for each record in the dataset. Here’s a breakdown of the actions taken:

## **Selection of Best Models**

We identified the best-performing models for both **SW1** and **SG1** based on evaluation metrics such as **Precision**, **Recall**, **F1 Score**, and **AUC**. After thorough evaluation, **Logistic Regression** was selected as the best model for both targets.

## **Preparing the Deployment Dataset**

To apply the trained models on the deployment dataset (df\_base), we ensured that the dataset had the same features as the training dataset. This is crucial to ensure the models would work properly with the deployment data.

## **Feature Scaling:**

The features in the deployment dataset were scaled using the same **StandardScaler** used during training. This step ensures that the feature values are in the same scale as the training data, which is critical for the model to perform optimally.

## **Probability Predictions:**

For each record in the deployment dataset, we obtained **probability predictions** from the models. Specifically:

* **SW1\_Prob**: This represents the probability that the individual is a **Wine Buyer** (target = 1) based on the trained **Logistic Regression** model for **SW1**.
* **SG1\_Prob**: This represents the probability that the individual is a **Gold Buyer** (target = 1) based on the trained **Logistic Regression** model for **SG1**.

The predict\_proba() function of the **Logistic Regression** model was used to obtain the probabilities. The probabilities for the positive class (class = 1) are extracted from the results, as these represent the likelihood of the individual being a buyer.

## **Combining the Scores:**

To derive a final score that takes into account both the **Wine Buyer** and **Gold Buyer** probabilities, a **combined score** was calculated. This was done by averaging the **SW1\_Prob** and **SG1\_Prob** for each individual in the dataset:

* **Combined\_Score** = (SW1\_Prob + SG1\_Prob) / 2

This combined score provides a unified measure of the likelihood that an individual is a buyer, considering both wine and gold purchasing behavior.

## **Final Output:**

The final result was added as two new columns to the df\_base dataframe:

* **SW1\_Prob**: Probability for Wine Buyer
* **SG1\_Prob**: Probability for Gold Buyer
* **Combined\_Score**: Average of both probabilities

This approach allows us to make more comprehensive predictions by combining insights from both target variables.

# Simulating Operational Marketing Action on the Top 20% Customers

## **Overview**

In this step, we simulated an operational marketing action focused on the top 20% of customers based on their **combined score**. Here's a breakdown of the actions taken and the calculation of the proportion of **expected buyers** within the selected customer base:

## **Sorting the Customers by Combined Score:**

We first sorted the deployment dataset (df\_base) based on the **combined score** in **descending order**. This allows us to identify the customers with the highest likelihood of being a buyer, considering both wine and gold purchasing behaviors.

## **Selecting the Top 20% of Customers:**

The next step was to select the **top 20%** of customers from the sorted list. This was done by calculating the top 20% of the total dataset, and then using the .head() function to extract the first 20% of the customers who have the highest combined scores:

* top\_20\_percent = int(0.2 \* len(df\_base\_sorted))
* df\_top\_customers = df\_base\_sorted.head(top\_20\_percent)

## **Defining the Threshold for Expected Buyers:**

We defined a **threshold** of **0.6** for both the **SW1 (Wine Buyer)** and **SG1 (Gold Buyer)** probabilities. This threshold determines whether a customer is considered an **expected buyer** for either product:

* If a customer's **SW1\_Prob** or **SG1\_Prob** exceeds 0.6, they are considered an expected buyer.

## **Calculating Expected Buyers in the Top 20%:**

Using the threshold of 0.6, we filtered the **top 20% customers** who meet the criteria of being an expected buyer (having a **probability greater than 0.6** for either SW1 or SG1):

* expected\_buyers = df\_top\_customers[(df\_top\_customers["SW1\_Prob"] > threshold) | (df\_top\_customers["SG1\_Prob"] > threshold)]
* The shape of the resulting subset was used to determine how many customers in the top 20% are expected to buy either wine or gold.

## **Calculating the Proportion of Expected Buyers:**

To quantify the success of our marketing simulation, we calculated the **proportion of expected buyers** within the top 20%:

* proportion\_expected = expected\_buyers / top\_20\_percent
* This proportion gives us an indication of the effectiveness of targeting the top 20% of customers, as it shows how many of them are expected to buy based on the model's predictions.

## **Output:**

The final output is the **proportion of expected buyers** within the top 20% of customers. This metric provides a useful insight into the likelihood of success for our operational marketing action based on targeting the top 20% with the highest combined scores.

The printed result gives us the proportion:

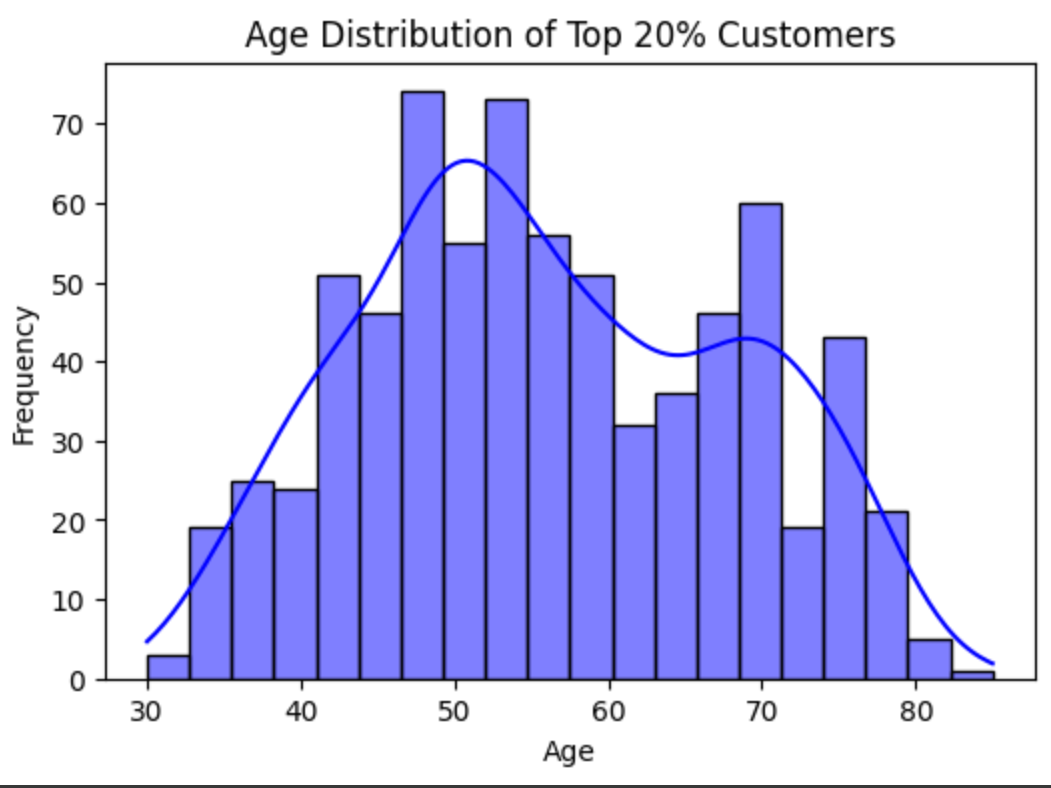


# Customer Profile Summary for Targeted Clients (Top 20%):

Based on the analysis of the top 20% of customers (sorted by the combined SW1 and SG1 scores), we have the following insights:

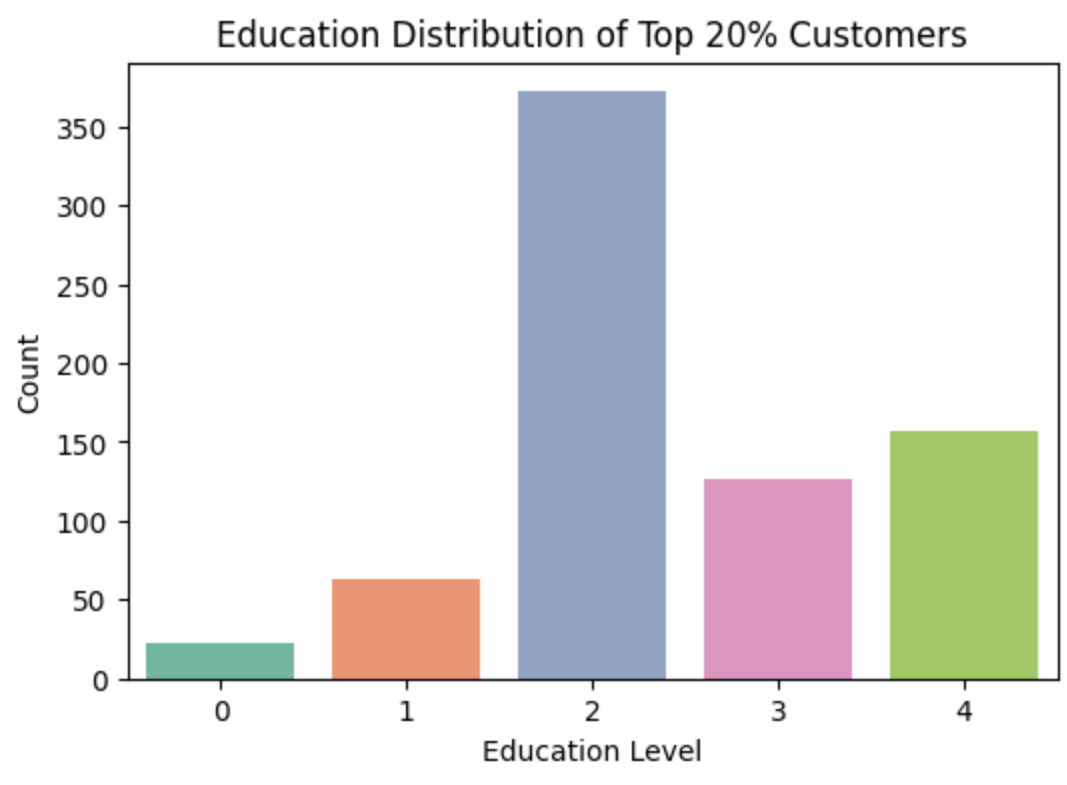
## **Age Distribution:**

* The average age of the targeted customers is **56.4 years**, with a **range from 30 to 82 years**.
* The middle 50% (from the 25th to 75th percentile) are between **47 and 68 years old**.
* This indicates that the targeted customers are primarily older adults, with a large concentration in their late 40s to early 70s.



## **Education Level:**

* **Average education level**: 2.19. (2 is Graduation)
* The majority of customers (25th percentile) have **high school** education, while the top 75% also includes customers with a **higher education degree** (up to 3 or 4 on the scale).
* The encoding of education is the following  
  'Basic': 0, '2n Cycle': 1, 'Graduation': 2, 'Master': 3, 'PhD': 4



## **Income:**

* The average income of the targeted customers is approximately **76,324** with a large variation in the range.
* **25th percentile**: Customers with an income of **62,698**.
* **50th percentile (median)**: The median income is **72,167**.
* **75th percentile**: Customers with an income of **81,950**.
* A significant portion of customers earns over **62,000**, indicating a relatively high-income group.

## **Family Dynamics:**

* **Kidhome** (Children at home): The average number of children at home is very low, at **0.05** (with the vast majority of customers having no children at home). The range is from **0 to 1** child.
* **Teenhome** (Teenagers at home): The average number of teenagers is **0.47**. However, most customers have either **no teenagers** or just **one teenager at home**.
* This suggests that the customers tend to be in the later stages of life, where children may no longer be living at home.

## **Marital Status:**

* **Marital\_Married**: About **34%** of the targeted customers are married, as indicated by the average value of **0.34**. The range goes from **0** (not married) to **1** (married).
* This suggests that the customers may have relatively stable family or relationship statuses.

## **Recency (Last Purchase Activity):**

* The average **Recency** is **54.81** (likely in days), with a large variation of up to **99 days**.
* The 75th percentile of recency is **79 days**, showing that most of these customers have recently engaged in purchasing behavior (likely within the past few months).

## **Web Visits (Customer Engagement):**

* **NumWebVisitsMonth** (number of web visits per month): The average is **4.29** visits per month, with a range from **0 to 9**.
* The customers are fairly active online, with most engaging at least 2 to 6 times a month.

## **SW1 Probability (Wine Buyer Probability):**

* The **SW1\_Prob** (probability of being a wine buyer) ranges from **0.52 to 1.00**, with an average of **0.67**.
* The distribution suggests that most customers have a **moderate to high probability** of purchasing wine, with some near-certainty (values approaching 1.0).

## **SG1 Probability (Gold Buyer Probability):**

* Similarly, the **SG1\_Prob** (probability of being a gold buyer) has an identical average of **0.67**, with values ranging from **0.52 to 1.00**.
* Like wine buyers, these customers show a moderate to high probability of purchasing gold.

# Conclusion:

The targeted customers for both wine and gold buyers tend to be **middle-aged to older** individuals, with relatively **high income levels** and **moderate to high engagement** online. They have stable family dynamics, with fewer children or teenagers at home, and are generally **active buyers** with strong probabilities for purchasing both wine and gold. Their profiles align with **affluent and established individuals**, with most being **married** and involved in more long-term financial decisions.