**Lab Exercise #2: Can You Determine Leads?**

***Aaryanah Micah Consorte, Kyle Gwyneth Morales | 3CSC***

**I. INTRODUCTION**

**1.1 Scenario and Objectives**

Campaign, such as promotion and discount, is an important aspect of a business. In some instances, the sales of a store can significantly increase when there are grand discounts and promos. However, in the case of ABC Supermarket, the certainty that customers will avail the year-end-sale is not fully determined.

As an approach to the given scenario, and to reduce the cost of the campaign, a machine-learning model that identifies the customers who might purchase the year-end-sale offer is intended to be created. Specifically, for lead identification, the goal is to improve the accuracy, precision, recall, and F1 score metrics.

This documentation contains the methods, trials, and other steps conducted by the students in order to design a model that is not only outstanding in terms of the given metrics, but also reasonable with regards to features used and processes employed.

**1.2 Discussion of Premise and Data Exploration**

Improving the model entails doing feature engineering. Before deciding whether the data may be a good input to the model, it is important to get to know the data first. There are a couple of data profiling and data exploratory analysis steps conducted such as:

1. **Checking for head, tail, and dimensions**

The head and tail offer a quick overview, revealing the first and last

rows to assess data structure, column names, and types. Whereas examining dimensions provides insights into dataset size and complexity. This quick inspection aids in early detection of data quality issues, assists in planning data cleaning.

|  |
| --- |
| df.head(10)  df.tail(10)  df.shape  df.dtypes |

1. **Checking for null and duplicate values**

Checking for null and duplicate values in a dataset is crucial for maintaining data integrity and accuracy. Null values can lead to errors in analysis and model training, impacting the performance of algorithms. Addressing null values ensures completeness and avoids biases in the results. Removing duplicates enhances resource efficiency and supports a cleaner dataset for accurate analyses.

Based on this step, it was discovered that there are null values in ‘Income’. As a response, the null values in this column were replaced by 0.

|  |
| --- |
| df.isnull().sum()  df['Income'].fillna('0', inplace=True) |

1. **Checking for unique values in Marital Status and Education**

This ensures that there are no unexpected or inconsistent entries, aiding in data cleaning and standardization. Knowing the distinct categories is essential for accurate analysis and proper groupings. As such, detecting and addressing unusual entries among unique values helps identify potential errors, contributing to overall data quality.

To deal with the inconsistent entries in marital status such as ‘YOLO’, ‘Alone’, and ‘Absurd’, they were all replaced by ‘Single’.

|  |
| --- |
| unique\_marital\_status\_values = df['Marital\_Status'].unique()  print("Unique values in 'Marital\_Status' column:")  print(unique\_marital\_status\_values)  df['Marital\_Status'] = df['Marital\_Status'].replace({'YOLO': 'Single', 'Alone': 'Single', 'Absurd': 'Single'})  print(df[['ID', 'Marital\_Status']]) |

As per the ‘education’ values check, no inconsistent entries were found, hence it was decided to retain the original values.

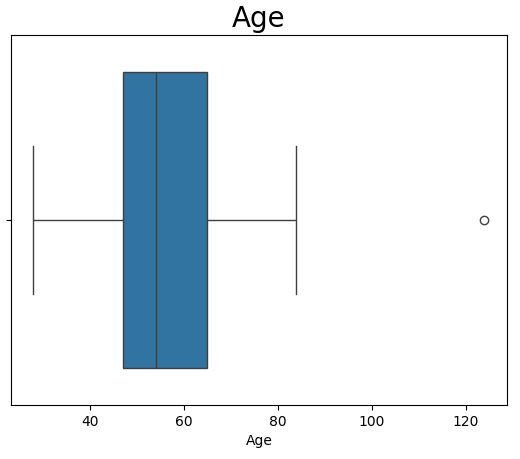
|  |
| --- |
| unique\_education\_values = df['Education'].unique()  print("Unique values in 'Education' column:")  print(unique\_education\_values) |

1. **Checking for Outliers in Data Values**

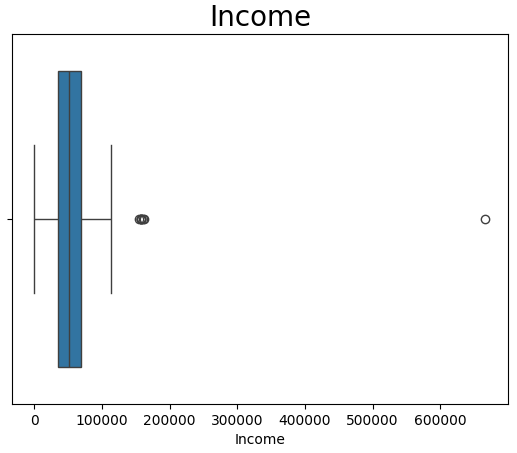
Checking for outliers is crucial to identify data points that deviate significantly from the main group. This helps ensure data quality and prevents these outliers from skewing the analysis, leading to misleading conclusions. This is evident in how much the metrics changed after removing the outliers in 'Year\_Birth'.

Specifically, the following observations were made after creating boxplots for different features in the dataset:

1.) *The variables ‘Age’ and ‘Income’ have outliers. Each of which have been handled differently depending on their nature.*

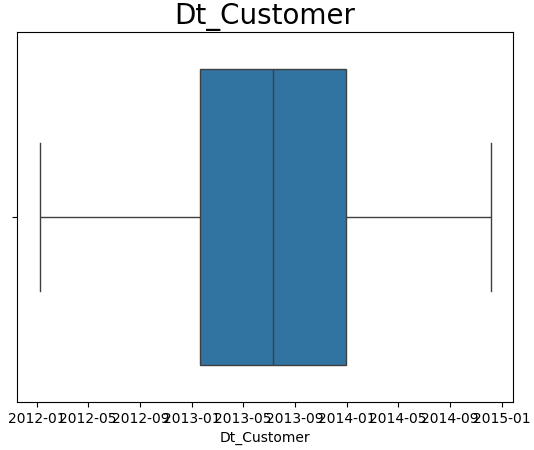


**Figure 1.** Age Boxplot



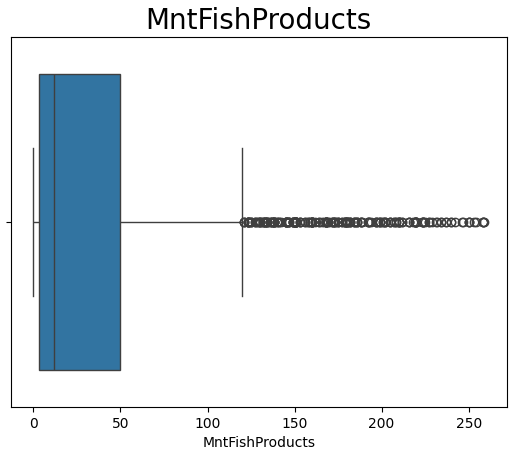
**Figure 2.** Income Boxplot

2.) *There are no outliers in Dates.*

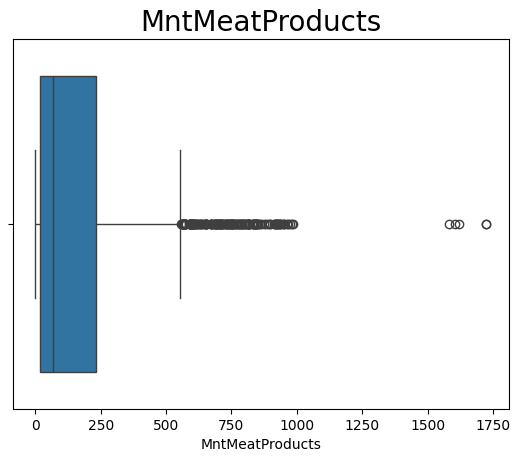


**Figure 3.** Date Boxplot

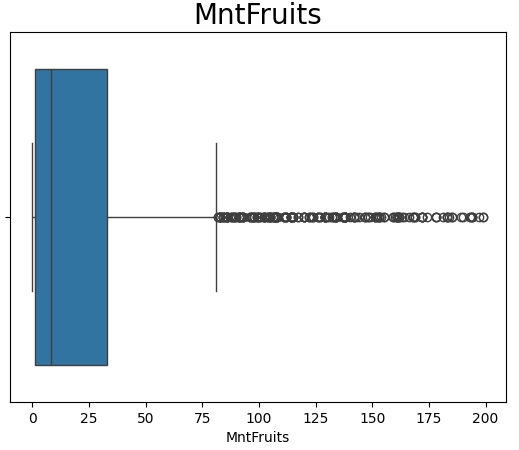
*2.) There are no outliers in MntFishProducts, MntMeatProducts, MntFruits, MntWines, MntSweetProducts, and MntGoldProds*



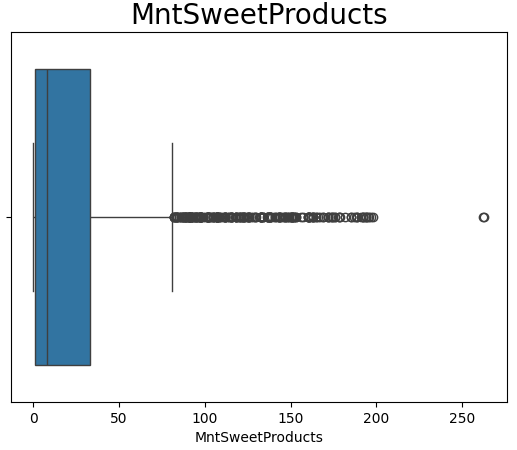
**Figure 4.** *MntFishProducts* Boxplot



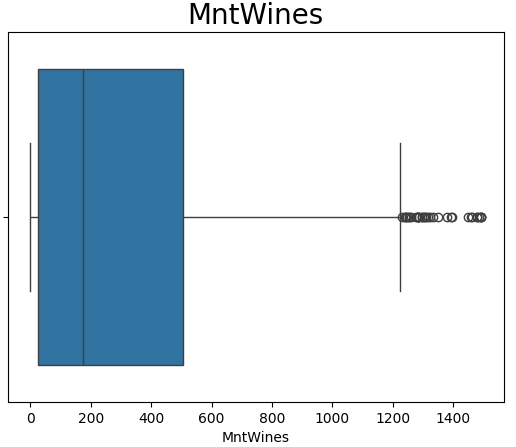
**Figure 5.** *MntMeatProducts* Boxplot



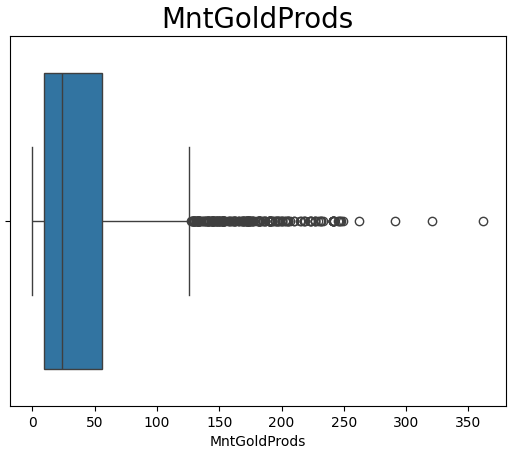
**Figure 6.** *MntFruits* Boxplot



**Figure 7.** *MntSweetProducts* Boxplot

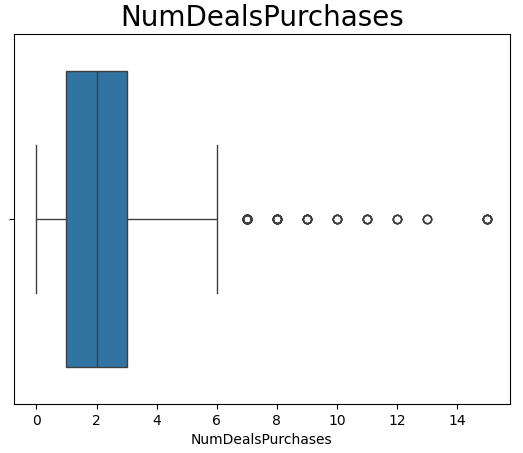


**Figure 8.** *MntWines* Boxplot

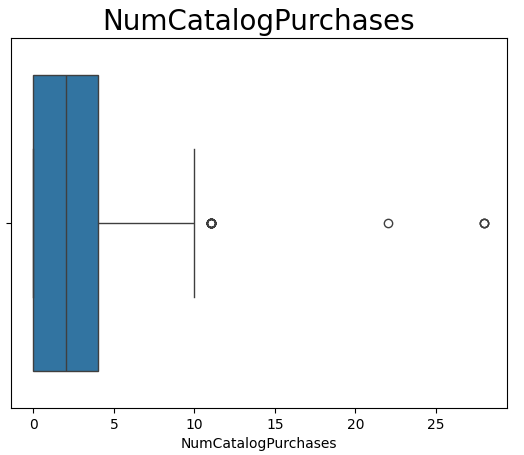


**Figure 9.** *MntGoldProds* Boxplot

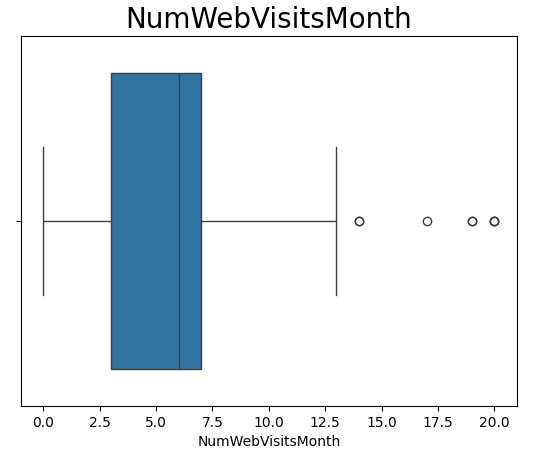
3.) *There are no outliers in NumDealsPurchases, NumCatalogPurchases, and NumWebVisitsMonth*



**Figure 10.** *NumDealsPurchases* Boxplot

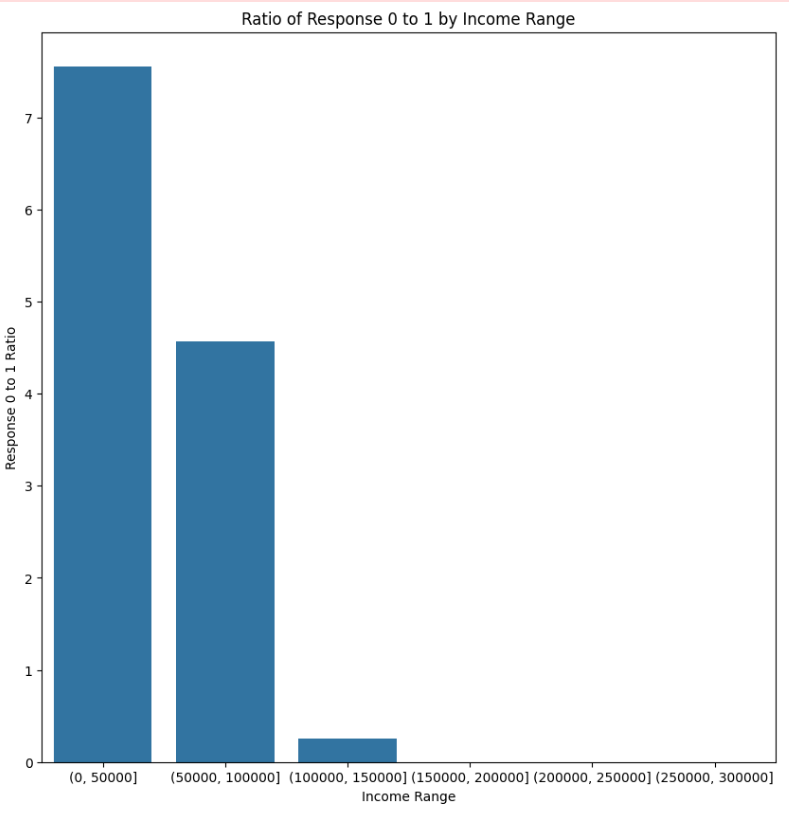


**Figure 11.** *NumCatalogPurchases* Boxplot



**Figure 12.** *NumWebVisitsMonth*Boxplot

1. **Income vs Response**

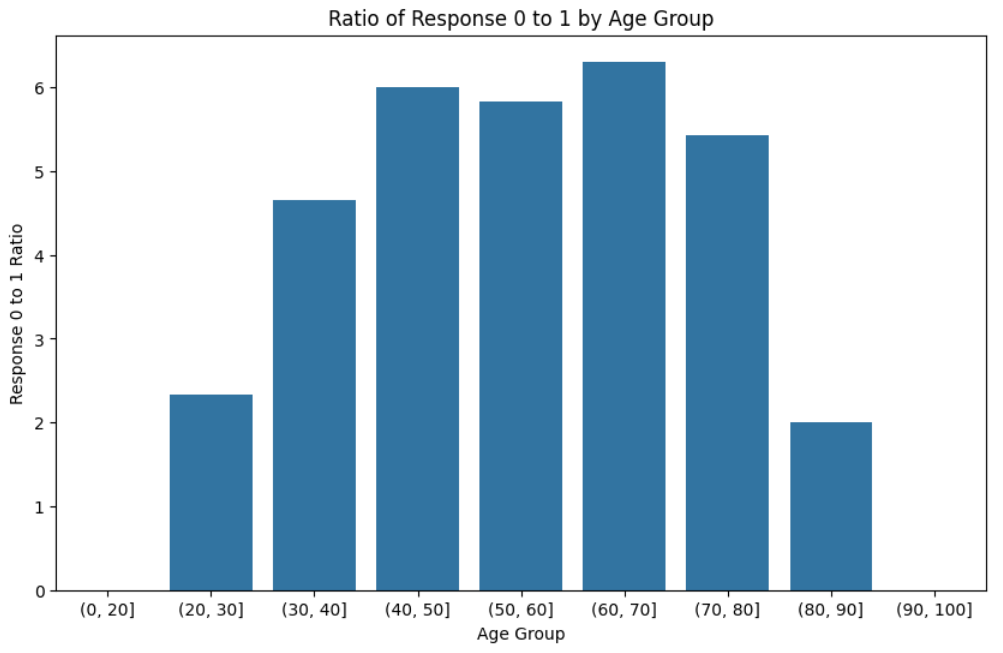


**Figure 13.** Income vs Response Ratio

The relationship between the income and response was also explored. To better compare the differences, the ratio of the responses were computed. The bar plot shows that there is a high ratio, hence high difference between responses on customers with income ranging to 0 to 50,000. On the other hand, the ratio is low on customers with 100,000 - 150,000 income.

From the observations, and given the gap between the value of ratios, it can be inferred that the responses vary depending on the income of the customer.

1. **Age vs Response**

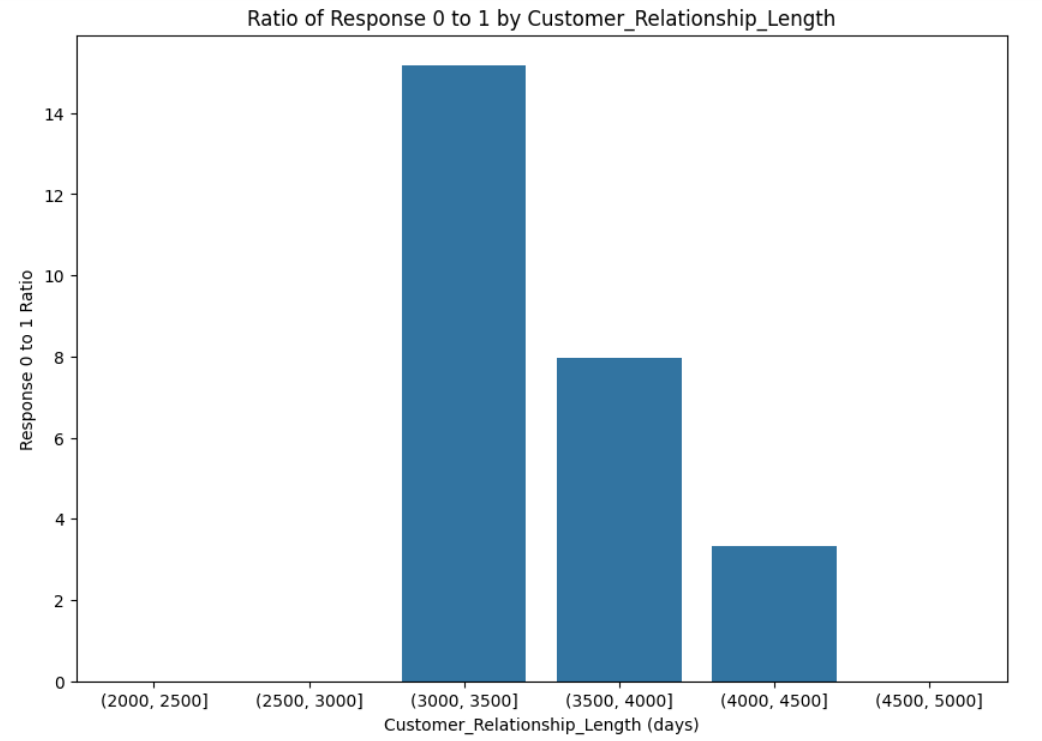


**Figure 14.** Age vs Response Ratio

The figure above shows the ratio of customers who availed and did not avail the sale. It can be implied that the ratios are highest from age brackets 60-70, and are lowest on age brackets 80-90.

Although the values of ratios appear to be close together in the bar plot, there’s still a huge difference between the ratio of responses from ages 20-30 and 80-90 with the rest of the ratio — hence, implying that age can be a factor in the possibility of the customer availing the sale.

1. **Customer Relationship Length vs Response**



**Figure 15.** Customer Relationship Length vs Response Ratio

Another connection that was explored was Customer Relationship Length, or how long the customer has been with the supermarket and their response in participating in the year-end sale. Based on the figure above, the ratio is high for customers for 3000-3500 days and lowest on 4000 - 4500 days. This suggests that the customer relationship length may have an impact with the response given the difference between the ratios.

**II. METHODOLOGY**

**2.1 Feature Engineering**

1. **Adding the ‘Age’**

Customer age, extracted from birthdays, plays a vital role in understanding customer demographics and informing strategic business decisions. It is helpful in segmenting the customer base for targeted marketing campaigns and tailoring products and services according to the specific age groups' preferences (Brown, 2021).

1. **Adding the ‘Customer Relationship Length’**

The duration of the customer's relationship with the company allows for assessing customer loyalty, predicting behavior patterns, and informing strategic decisions (Lemon & Verhoef, 2016). By analyzing this information, it is easier to identify which customers are more likely to stay with the company long-term and potentially benefit from targeted marketing efforts, such as discount campaigns.

1. **Adding the ‘Recency’**

Knowing how recently a customer purchased something helps in understanding current customer behavior, predicting who might stop buying, and grouping customers for targeted marketing (Lemon & Verhoef, 2016). By adding this to the features, the model gains an insight on patterns related to customer engagement over time.

1. **Adding ‘Income’ and Keeping the Outlier**

Another feature added is the income of the customer. Adding it as a feature is due to the prospect that the decision to purchase the year-end sale is connected to the financial capacity of the customer. It also results in a significant increase in the metrics, as shown in III. Experiments: Trial 5.

During data exploration and profiling, it has been discovered that there is an outlier in the income data.

However, further analysis results in the decision to not remove this outlier. The occurrence of a very high income in the data can be classified as a True Outlier. True outliers are those that do not occur as a result of human errors. These data items are real and can have a variety of explanations for their existence (Khan, 2022). Similarly, high income is possible in real life, and it is reasonable for a customer to have such income, hence, it has been decided to retain that data.

1. **Adding ‘NumCatalogPurchases’**

This was added to the features because it can be leveraged by machine learning models to predict which customers are more likely to be receptive to future discount campaigns.

1. **Adding the ‘Education’**

This feature enhances the model's predictive capabilities and provides valuable socio-demographic insights. Education serves as a key factor influencing customer behavior, preferences, and decision-making processes (Kotler & Keller, 2021). The inclusion of this variable allows the model to create meaningful insights based on educational levels and uncover nuanced relationships between education and other features.

1. **Adding the ‘NumWebVisitsMonth’**

The number of web purchases is an important factor in forecasting whether a consumer will take advantage of a promotion since it provides useful information about the client's online buying habit, engagement level, and loyalty.

Customers who make frequent online purchases show a comfort with online transactions and are more inclined to respond to incentives.

Data on web purchases aids in understanding customer preferences, allowing firms to adapt discounts to match individual purchasing habits. Furthermore, the past frequency of web purchases serves as a predictor, allowing firms to forecast the response of customers.

1. **Adding the ‘NumDealsPurchases’**

Incorporating the number of purchases made with discount into a prediction model for customer discount availability is critical. It identifies clients who are sensitive to promotional pricing, allowing firms to tune discounts for a good reaction.

Analyzing this data provides insights into client purchasing behavior, separating opportunistic purchasers from those who enjoy cheap things. This information assists in successful client segmentation and targeted marketing based on unique preferences.

1. **Excluding the ‘Complain’**

When applying the Logistic Regression Model, the coefficients were also estimated. The coefficients when exponentiated represent the odd ratios. Moreover, the direction of the coefficients also implies if the log-odds of the event (target variable being 1) decrease or increase.

To get the coefficients of each features, the following code was implemented:

|  |
| --- |
| print("Estimated Coefficients:")  print(clf.coef\_)  print("Intercept:", clf.intercept\_) |

The results show that the Complain feature has a coefficient of -0.05, Since it has a negative sign, it also has a negative association and it may imply that there is a decrease in certain parameters.

Although some features like the customer relationship length and recency have a negative sign (negative association), it only makes sense to remove the ‘Complain’ data.

Aside from the increase in the accuracy, precision, recall and F1 when ‘Complain’ is removed (as seen in trial 3), the reason to include it as a feature is also not sound since the act of complaining does not automatically mean that the customers will be uninterested in participating in the sale. There are other factors that influence both the complaint and the non-participation.

**2.2 Other Machine Learning Model**

In order to explore the other options and see if the metrics will be further improved, machine learning models, aside from logistic regression, were applied.

Given that the value of responses are 1 or 0 (participating in the sale or not), **Decision Trees** can be a good machine learning model for making a prediction.

Another model is the **Naive Bayes Classifier.** It is a machine learning algorithm that performs well in categorical features. Using it in this context may also be relevant in customer response prediction.

**2.3 Changing the Hyperparameters**

During the trials using Decision Trees and Naive Bayes Classifier, the initial values of some metrics are 0. This presents a possible need for tuning of hyperparameters. Specifically, GridSearchCV object is applied in the Decision Trees Model and Naive Bayes Classifier. Tuning the hyperparameters results in improved and reasonable values of the metrics.

**2.4 Analyzing the Metrics**

One of the main objectives is to improve the accuracy, precision, recall, and F1 score metrics for lead identification. To do that, the student first assesses what are the realistic and ideal values to achieve.

Accuracy is said to be the ratio of the total number of correct predictions to the total number of predictions generated. An accuracy of 70% or more is seen as a model with good performance (Barkved, 2022). Applying it in the context of predicting the response of customers, high accuracy implies that the model can actually predict if the customer will purchase the model or not.

While it is a good metric, it may not be the best way to evaluate if the model is successful in identifying the customers who might purchase the year-end sale offer.

Other metrics that are frequently associated are precision and recall. In this instance, recall indicates out of all the customers who actually responded positively to the offer (true positive + false negative), how many were correctly identified as such by the model (true positive). Precision, on the other hand, focuses on the ratio of the number of customers correctly recognized as availing the promo (true positive) among all the predictions generated by the model that indicates customers would avail the promo (true positive + false positive).

Lastly, F1 score is the harmonic mean of precision and recall. A good F1 score indicates that there is a balance between the two metrics, also reflecting that the model can deliver a well-rounded performance.

Based on the analysis of what each metrics determine in relation to the given context, the students focused on **aiming to achieve a higher precision**, rather than a higher recall. In simple words, the goal is to count the true positives (customers predicted as purchasing the year-end sale and have actually purchased it) out of the sum of true positives and false positives (all predictions that indicates customer purchasing the sale). Having a higher precision means the model makes fewer false positives, making it reliable in identifying customers who are likely to participate in the sale.

According to research, a 70% precision is ideal and realistic (StudyCountry, 2023). Hence, the goal was to achieve a precision value that is approximately close to 70%. Moreover, the students also worked on improving the accuracy of the model. Despite focusing on achieving a high precision, a model with high accuracy still equates to a good model performance

**III. EXPERIMENTS**

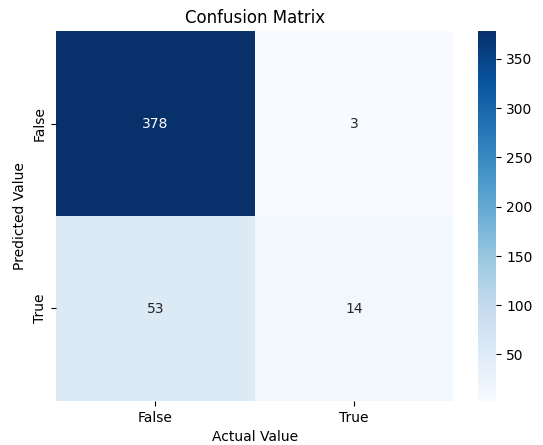
The method of improving the model’s results undergoes numerous trials, that are as follows:

Table 1. *Trials Conducted*

| **TRIAL 0**: Initial Metrics from given code | | |
| --- | --- | --- |
| 0 | **Features:**  **-** MntFishProducts  - MntMeatProducts  - MntFruits  - MntSweetProducts  - MntWines  - MntGoldProds  **Test size:** 0.3 | Accuracy: 0.8512  Precision: 0.5000  Recall: 0.0200  F1: 0.0385  AUC: 0.5083 |
| **TRIAL 1:** Adding ‘NumDealsPurchases’, ‘NumWebVisitsMonths’, ‘Age’, ‘Customer Relationship Length’, ‘Complain’ and ‘Recency’  **Observation**: Accuracy, Precision, Recall, F1, and AUC increased | | |
| 1 | **Model:** Logistic Regression  **Features:**  **-** NumDealsPurchases  - NumWebVisitsMonth  **-** MntFishProducts  - MntMeatProducts  - MntFruits  - MntSweetProducts  - MntWines  - MntGoldProds  - Age  - CustomerRelLength  - Complain  - Recency  **Test size:** 0.3 | Accuracy: 0.8586 **⬆**  Precision: 0.5714 **⬆**  Recall: 0.2 **⬆**  F1: 0.2963 **⬆**  AUC: 0.5869 **⬆** |
| **TRIAL 2**: Same features, decreasing test size  **Observation**: Accuracy, Precision, Recall, F1, and AUC increased | | |
| 2 | **Model:** Logistic Regression  **Features:**  **-** NumDealsPurchases  - NumWebVisitsMonth  **-** MntFishProducts  - MntMeatProducts  - MntFruits  - MntSweetProducts  - MntWines  - MntGoldProds  - Age  - CustomerRelLength  - Complain  - Recency  **Test size** = 0.2 | Accuracy: 0.8616 **⬆**  Precision: 0.6000 **⬆**  Recall: 0.2239 **⬆**  F1: 0.3261**⬆**  AUC: 0.5988 **⬆** |
| **TRIAL 3**: Removing Complain  **Observation**: Accuracy, Precision, Recall, F1, and AUC increased | | |
| 3 | **Model:** Logistic Regression  **Features:**  **-** NumDealsPurchases  - NumWebVisitsMonth  **-** MntFishProducts  - MntMeatProducts  - MntFruits  - MntSweetProducts  - MntWines  - MntGoldProds  - Age  - CustomerRelLength  - ~~Complain~~  - Recency  **Test size** = 0.2 | Accuracy: 0.8638 **⬆**  Precision: 0.6154 **⬆**  Recall: 0.2388 **⬆**  F1: 0.3441 **⬆**  AUC: 0.6063 **⬆** |
| **TRIAL 4**: Removing Recency and Age  **Observation**: Accuracy, Precision, Recall, F1, and AUC decreased | | |
| 4 | **Model:** Logistic Regression  **Features:**  **-** NumDealsPurchases  - NumWebVisitsMonth  **-** MntFishProducts  - MntMeatProducts  - MntFruits  - MntSweetProducts  - MntWines  - MntGoldProds  - ~~Age~~  - CustomerRelLength  - ~~Recency~~  **Test size** = 0.2 | Accuracy: 0.8527 **⬇**  Precision: 0.5556  **⬇**  Recall: 0.0746  **⬇**  F1: 0.1316  **⬇**  AUC: 0.5321  **⬇** |
| **TRIAL 5**: Re-adding the age and recency, Adding the Income  **Observation**: Accuracy, Precision, Recall, F1, and AUC increased | | |
| 5 | **Model:** Logistic Regression  **Features:**  **-** NumDealsPurchases  - NumWebVisitsMonth  **-** MntFishProducts  - MntMeatProducts  - MntFruits  - MntSweetProducts  - MntWines  - MntGoldProds  - Age  - CustomerRelLength  - Recency  - Income  **Test size** = 0.2 | Accuracy: 0.8661**⬆**  Precision: 0.6400 **⬆**  Recall: 0.2388 **⬆**  F1: 0.3478 **⬆**  AUC: 0.6076 **⬆** |
| **TRIAL 6**: Adding Kidhome and Teenhome  **Observation**: Precision decreased, Recall, F1, and AUC increased, Accuracy stays the same | | |
| 6 | **Model:** Logistic Regression  **Features:**  **-** NumDealsPurchases  - NumWebVisitsMonth  **-** MntFishProducts  - MntMeatProducts  - MntFruits  - MntSweetProducts  - MntWines  - MntGoldProds  - Age  - CustomerRelLength  - Recency  - Income  - Teenhome  - Kidhome  **Test size** = 0.2 | Accuracy: 0.8661  Precision: 0.6296 **⬇**  Recall: 0.2537 **⬆**  F1: 0.3617 **⬆**  AUC: 0.6137 **⬆** |
| **TRIAL 7**: Adding Education  **Observation**: Precision increased; Accuracy, Recall, F1, and AUC decreased | | |
| 7 | **Model:** Logistic Regression  **Features:**  **-** NumDealsPurchases  - NumWebVisitsMonth  **-** MntFishProducts  - MntMeatProducts  - MntFruits  - MntSweetProducts  - MntWines  - MntGoldProds  - Age  - CustomerRelLength  - Recency  - Income  - Teenhome  - Kidhome  - Education  **Test size** = 0.2 | Accuracy: 0.8616 **⬇**  Precision: 0.6316 **⬆**  Recall: 0.1791 **⬇**  F1: 0.2791**⬇**  AUC: 0.5804 **⬇** |
| **TRIAL 8**: Removing Kidhome, and Teenhome; Adding Num Catalog Purchases  **Observation**: Precision increased; Accuracy, Recall, F1, and AUC decreased | | |
| 8 | **Model:** Logistic Regression  **Features:**  **-** NumDealsPurchases  - NumWebVisitsMonth  **-** MntFishProducts  - MntMeatProducts  - MntFruits  - MntSweetProducts  - MntWines  - MntGoldProds  - Age  - CustomerRelLength  - Recency  - Income  - Education  ~~- Teenhome~~  ~~- Kidhome~~  - NumCatalogPurchases  **Test size** = 0.2 | Accuracy: 0.8638 **⬆**  Precision: 0.6364 **⬆**  Recall: 0.2090 **⬆**  F1: 0.3146 **⬆**  AUC: 0.5940 **⬆** |
| **TRIAL 9**: Adding Web Purchase Visit Rate  **Observation**: Precision increased; Accuracy, Recall, F1, and AUC decreased | | |
| 9 | **Model:** Logistic Regression  **Features:**  **-** NumDealsPurchases  - NumWebVisitsMonth  **-** MntFishProducts  - MntMeatProducts  - MntFruits  - MntSweetProducts  - MntWines  - MntGoldProds  - Age  - CustomerRelLength  - Recency  - Income  - Education  - NumCatalogPurchases  - WebPurchaseVisitRate  **Test size** = 0.2 | Accuracy: 0.8594 **⬇**  Precision: 0.5909 **⬇**  Recall: 0.1940 **⬇**  F1: 0.2921 **⬇**  AUC: 0.5852 **⬇** |
| **TRIAL 10:** Removing data with birthdays before 1900 and Web Purchase Visit Rate  **Observation**: Accuracy, Precision, Recall, F1, and AUC increased | | |
| 10 | **Model:** Logistic Regression  **Features:**  **-** NumDealsPurchases  - NumWebVisitsMonth  **-** MntFishProducts  - MntMeatProducts  - MntFruits  - MntSweetProducts  - MntWines  - MntGoldProds  - Age  - CustomerRelLength  - Recency  - Income  - Education  - NumCatalogPurchases  ~~- WebPurchaseVisitRate~~  **Test size** = 0.2 | Accuracy: 0.8750 **⬆**  Precision: 0.8235 **⬆**  Recall: 0.2090 **⬆**  F1: 0.3333 **⬆**  AUC: 0.6005 **⬆** |
| **TRIAL 11**: Trying Naive Bayes Model with the same features.  **Observation:** Accuracy and Precision decreased; Recall, F1, and AUC increased. | | |
| 11 | **Model:** Naive Bayes  **Features:**  **-** NumDealsPurchases  - NumWebVisitsMonth  **-** MntFishProducts  - MntMeatProducts  - MntFruits  - MntSweetProducts  - MntWines  - MntGoldProds  - Age  - CustomerRelLength  - Recency  - Income  - Education  - NumCatalogPurchases  **Test size** = 0.2 | Accuracy: 0.6920 **⬇**  Precision: 0.2463 **⬇**  Recall: 0.4714 **⬆**  F1: 0.3235 **⬆**  AUC: 0.6021 **⬆** |
| **TRIAL 12:** Still Naive Bayes, Removing MntFishProducts, MntFruits, MntMeatProducts, MntWines, MntSweetProducts, and MntGoldProds  **Observation**: Accuracy, Precision, Recall, F1, and AUC increased | | |
| 12 | **Model:** Naive Bayes  **Features:**  **-** NumDealsPurchases  - NumWebVisitsMonth  - Age  - CustomerRelLength  - Recency  - Income  - Education  - NumCatalogPurchases  **Test size** = 0.2 | Accuracy: 0.7656 **⬆**  Precision: 0.3267 **⬆**  Recall: 0.4714 **⬆**  F1: 0.3860 **⬆**  AUC: 0.6458 **⬆** |
| **TRIAL 13:** Trying Decision Trees Model  **Observation:** Recall increased; Precision, Recall, F1, and AUC decreased. | | |
| 13 | **Model:** Decision Trees  **Features:**  **-** NumDealsPurchases  - NumWebVisitsMonth  - Age  - CustomerRelLength  - Recency  - Income  - Education  - NumCatalogPurchases  **Test size** = 0.2 | Accuracy: 0.8393 **⬇**  Precision: 0.4545 **⬇**  Recall: 0.1429 **⬇**  F1: 0.2174 **⬇**  AUC: 0.5556 **⬇** |

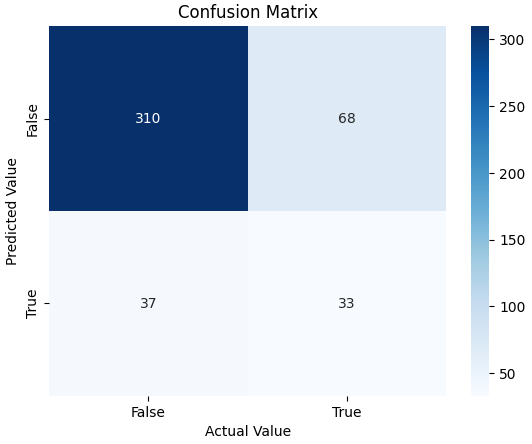
**IV. RESULTS AND ANALYSIS**

Three different machine learning models were tested such as: Logistic Regression, Naive Bayes, and Decision Trees.



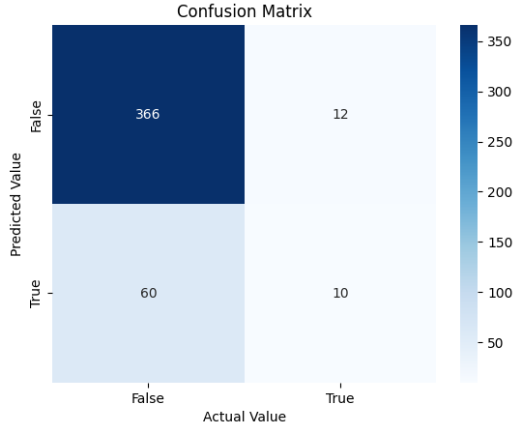
**Figure 16.** Confusion Matrix using Logistic Regression

The confusion matrix reveals 378 true negatives, indicating accurate predictions of the absence of a condition. There are 3 false negatives, representing instances of incorrect predictions of absence. Additionally, 14 true positives and 53 false positives denote correct and incorrect predictions of presence, respectively. This assessment underscores the model's proficiency in identifying negative instances but indicates room for improvement in positive predictions.



**Figure 17.** Confusion Matrix using Naive Bayes

The values in the confusion matrix show the model's strength in identifying negative instances but suggest a need for improvement in minimizing false negatives and false positives.



**Figure 18.** Confusion Matrix using

Decision Trees

The confusion matrix above shows that the model exhibits proficiency in identifying negatives, but there is room for improvement in reducing false positives and false negatives.

Among the various models evaluated, Logistic Regression stands out as the best choice overall, primarily due to its high precision. This metric aligns well with the students’ goal of accurately classifying whether the customers will reply with a positive or negative response.

The students first applied feature engineering using Logistic Regression. The experiments aimed at refining the predictive model for customer responses to a promotional offer revealed key insights. Initial trials (TRIAL 0) set the baseline, with subsequent enhancements showing improvements. The addition of crucial features like 'Age,' 'Customer Relationship Length,' 'Complain,' 'Income,' and 'Recency' significantly boosted the model's performance.

Whereas the removal of 'Recency' and 'Age' (TRIAL 4) led to a decline, highlighting their importance. Notably, cleaning data by removing birthdays before 1900 (TRIAL 10) resulted in the highest accuracy and precision.

Table 2. *Comparison of Metrics Results using different Models*

|  | **Logistic Regression** | **Naive Bayes** | **Decision Trees** |
| --- | --- | --- | --- |
| **A** | 0.8750 | 0.7656 | 0.8393 |
| **P** | 0.8235 | 0.3267 | 0.4545 |
| **R** | 0.2090 | 0.4714 | 0.1429 |
| **F1** | 0.3333 | 0.3860 | 0.2174 |

Another machine learning model used is Naive Bayes model. Applying this model, and using the features ‘Age’, ‘Customer Relationship Length’, ‘Recency’, ‘Income’, ‘Education’, and ‘Num Catalog Purchases’ have generated higher recall and F1 score (TRIAL 12).

Lastly, Decision Trees Model is applied using the same features. Compared to the previous models, the metrics using this model are all lower.

Table 3. *Best Model with respect to Machine Learning Model*

|  | **Model** |
| --- | --- |
| **Accuracy** | Logistic Regression |
| **Precision** | Logistic Regression |
| **Recall** | Naive Bayes |
| **F1** | Naive Bayes |

To summarize, in terms of accuracy and precision, the Logistic Regression model gained the highest value. While for recall and F1, Naive Bayes seemed to be more effective.

The results and the experimentation with different models just showed that metrics can vary depending on the model and the features. Although the same features are used, it may not be as accurate, or as precise as the other model. Similarly, some combinations may also result in a different recall and f1 score.

Further trials demonstrated that balance is required in feature selection and model tuning. Experiments with different models indicated the importance of a tailored model selection.

**V. CONCLUSIONS AND RECOMMENDATIONS**

Overall, based on the trials conducted, it can be implied that Trial 10 or the use of Logistic Regression with features NumDealsPurchases, NumWebVisitsMonth, MintFish Products, Mtn Meat Products, Mint Sweet Products, Mnt Fruits, Mnt Wines, MntGoldProds, Age, CustomerRelLength, Recency, Income, Education, and NumCatalogPurchases gathered the best model performance with 0.8235 precision and an accuracy of 0.8750.

Although the Naive Bayes model (Trial 12) gathered a higher value of recall and f1, the analysis of metrics led to the conclusion that in the context of predicting the response of customers, it’s better to aim for a **higher precision** rather than a higher recall. With increased precision, the model generates fewer false positives, making it a more reliable tool in identifying customers who are likely to engage in the sale.

To sum up the overall process of designing a model that predicts the response of customers, it was discovered that the selection of features is important as it greatly affects the number of true and false positives, and true and false negatives. It is also crucial to note that features that function well for one model may not necessarily translate to higher metrics for another model.

Moreover, aside from considering the metrics, the students aimed to carefully consider the reasons as to why the specific feature is needed in the model. For instance, it is notable to consider adding the customer relationship length as a feature since it can influence their response towards participating in the year-end sale. Customer age aids in demographic understanding and targeted marketing. Recency of purchases helps predict customer behavior and enables grouping for marketing strategies. Income is considered due to its connection to purchase decisions, with an outlier retained for its genuine nature. NumCatalogPurchases and NumWebVisitsMonth are added to predict receptiveness to future discount campaigns based on past behaviors. The education feature enhances socio-demographic insights, while NumDealsPurchases is crucial for identifying customers sensitive to promotional pricing, aiding in effective segmentation and targeted marketing efforts.

Despite the achieved high precision and accuracy, there were a couple of recommendations. During the selection of features, the students were solely using Logistic Regression. This led to better metrics for the said model. Moving forward, having a simultaneous testing of each model and the effect of each feature is suggested, especially in the Decision Trees model since only one trial of feature selection is conducted using this model. This also suggests that more trials of feature engineering is recommended to achieve better results.

However, even if feature selection is not conducted using other models, the features added were carefully considered and were reasonable to add regardless of the value of metrics. In addition, a more meticulous cleaning of the data set is advised. This is crucial to obtaining good results in Accuracy, Precision, Recall, F1, and AUC. Lastly, usage of other Machine Learning models such as SVM and K-neighbors classification could be considered.

With respect to the evaluation of the business aspect in designing the entire model, given the precision and accuracy of the Logistic Regression model, it predicts that 334 customers will participate in the year-end sale, while 1904 will not. With this prediction, ABC supermarket can prioritize reaching out to the identified 334 customers. It's important to note that, being a predictive model, there's no guarantee that those classified by the model will necessarily avail the gold membership and the promo. Nevertheless, this information can contribute to reducing campaign cost.

**VI. REFERENCES**

Brown, J. (2021, February 5). How to tailor your product or service to different age groups. Forbes. https://www.forbes.com/sites/forbesbusinesscouncil/2024/02/05/unlocking-growth-seven-ways-to-tailor-marketing-plans/

Khan, M. (2022, April 7). A Basic Guide to Outliers. Godatadrive.com; DataDrive. https://godatadrive.com/blog/outliers-101#:~:text=True%20Outlier%3A%20Outliers%20that%20are,many%20reasons%20behind%20their%20existence.

How To Know if Your Machine Learning Model Has Good Performance | Obviously AI. (2022).Obviously.ai.https://www.obviously.ai/post/machine-learning-model-performance

Kotler, P., & Keller, K. L. (2021). Marketing Management, Global Edition. Pearson Education Limited.

Lemon, K. N., & Verhoef, P. C. (2016). Understanding Customer Experience Throughout the Customer Journey. Journal of Marketing, 80(6), 69–96. Sagepub. https://doi.org/10.1509/jm.15.0420

What is a good precision value? (2023). Studycountry.com.https://www.studycountry.com/wiki/is-0-6-precision-good