MGS 616 PREDICTIVE ANALYTICS FINAL REPORT

Customer Behavior Analysis

Introduction

The goal of this project is to conduct a comprehensive customer behavior study, which is a thorough examination of a company's most valuable customers. It will help businesses learn more about their customers and will make it easier for them to tailor their products to various customers' needs, habits, and concerns based on prediction models. The purpose of consumer behavior analysis is to identify methods to better accommodate customer behavior patterns, establish product-market fit and boost conversion rate. (Custer, 2022)

The aim is to estimate a customer's spending potential using their income, and to understand the key influencers among several variables, including education level, marital status, and the number of children. The project also seeks to assess the probability of a customer's involvement in campaigns, considering their marital status and educational background using customer behavior data. To achieve these objectives, we have employed methods like Linear Regression for predicting spending trends, Random Forest for identifying key variables, and Naive Bayes for projecting the likelihood of campaign participation.

Methods

The first stage of our approach was to perform exploratory data analysis which entails pre-processing the dataset to guarantee its suitability for building models. This stage includes removing duplicates and null values, and excluding columns containing singular values that lack analytical significance as a part of data cleaning process. Following this, the next phase involves analysing and comprehending the dataset.

```
── Data Summary ────────────────────────
                           Values
Name                       Customer_df
Number of rows             5000
Number of columns          29
_________________________
Column type frequency:
  character                3
  numeric                  26
_________________________
Group variables            None

── Variable type: character ──────────────────────────────────────────────────────────────────────────────────────
  skim_variable            n_missing complete_rate min max empty n_unique whitespace
1 Customer_Education_Level         0             1   3  10     0        5          0
2 Customer_Marital_Status          0             1   4   8     0        8          0
3 Customer_enrolment_Date          0             1  10  10     0      701          0

── Variable type: numeric ────────────────────────────────────────────────────────────────────────────────────────
   skim_variable          n_missing complete_rate       mean         sd   p0    p25    p50     p75    p100 hist
 1 Customer_Id                    0             1  5594.       3224.      10  2821.  5622.   8402.  11181 ▇▇▇▇▇
 2 Birth_Year                     0             1  1945.         30.2   1893  1919   1945    1971    1996 ▇▇▇▇▇
 3 Yearly_Income                  0             1 80041.      97398.    1804 10325  32437  124932. 649248 ▇▂▁▁▁
 4 Children_count                 0             1     1.49       1.13     0     0      1       3       3 ▇▁▁▁▇
 5 Teenagers_count                0             1     1.51       1.12     0     1      1       3       3 ▇▁▁▁▇
 6 Purchase_recency               0             1    49.4       29.1      0    24     49      75      99 ▇▇▇▇▇
 7 Expense_wine                   0             1  1993.       2961.      0   229    775    2376   28306 ▇▁▁▁▁
 8 Expense_fruits                 0             1  2800.       4260.      0   333.  1103    3216.  41206 ▇▁▁▁▁
 9 Expense_meat                   0             1  3239.       5002.      0   384.  1248.   3699   51212 ▇▁▁▁▁
10 Expense_fish                   0             1   815.       1222.      0    96    312     967   12596 ▇▁▁▁▁
11 Expense_sweets                 0             1  1192.       1808.      0   140   445    1376.  16653 ▇▁▁▁▁
12 Expense_gold                   0             1  2819.       4326.      0   335  1034    3226   39717 ▇▁▁▁▁
13 Purchases_discount             0             1     7.64       4.63     0     4      8      12      15 ▇▇▇▇▆
14 Purchases_website              0             1    19.3       11.4      0     9     19      29      39 ▇▇▇▇▆
15 Purchases_catalogue            0             1    21.7       12.7      0    11     22      32      43 ▇▇▇▇▆
16 Purchases_stores               0             1     7.46       4.63     0     3      8      12      15 ▇▇▇▇▆
17 Monthly_Webvisits              0             1    13.6        8.08     0     7     14      21      27 ▇▇▇▇▆
18 Accepted_campaign3             0             1     0.0258     0.159    0     0      0       0       1 ▇▁▁▁▁
19 Accepted_campaign4             0             1     0.047      0.212    0     0      0       0       1 ▇▁▁▁▁
20 Accepted_campaign5             0             1     0.0404     0.197    0     0      0       0       1 ▇▁▁▁▁
21 Accepted_campaign1             0             1     0.0566     0.231    0     0      0       0       1 ▇▁▁▁▁
22 Accepted_campaign2             0             1     0.0102     0.100    0     0      0       0       1 ▇▁▁▁▁
23 Customer_complaint             0             1     0.0184     0.134    0     0      0       0       1 ▇▁▁▁▁
24 Z_CostContact                  0             1     2          0        2     2      2       2       2 ▁▁▇▁▁
25 Z_Revenue                      0             1    13          0       13    13     13      13      13 ▁▁▇▁▁
26 Accepted_last_campaign         0             1     0.137      0.343    0     0      0       0       1 ▇▁▁▁▁
> |
```

- To ensure the models function smoothly, we made several adjustments to the dataset:
- We removed the columns 'z_costcontact' and 'z_revenue' from the dataset as they contained uniform values.
- We standardized the education levels, replacing '2n cycle' with 'Master'.
- We categorized 'Marital Status' into two groups: 'Single' and 'In a Relationship'.
- We grouped individuals according to their ages using the 'Age' column.
- We used the 'Total_Spent' variable to determine the total expenditure of customers on all products.
- We calculated the frequency of purchases made by each customer in the past two years.
- We leveraged the 'Dt_Customer' variable to ascertain when customers first engaged with the company.
- We computed the total number of accepted deals for each customer.

Results

Linear Regression and Correlation between Income and Expenses

The correlation coefficient between Income and Total_Spent variables is 0.949. This demonstrates a strong positive correlation between income and expenditure, implying that as income increases, expenditure tends to increase as well. To further investigate this relationship, we executed a linear regression model, which yielded an intercept of 111.846 and a slope (X) of 0.158. Using this model, we attempted to predict a customer's spending capacity if they have an income of $160,000. The model predicted a spending capacity of $25,386.92. Supporting images for the correlation and linear regression are provided below.

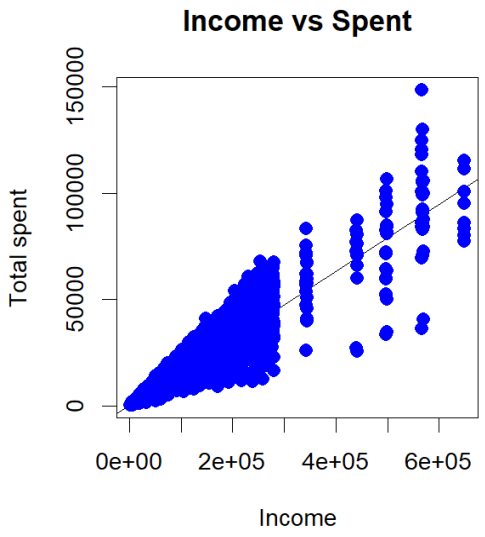

Fig. 1: Plot showing the variation of total spends vs income

Subsequently, we calculated the average spending by customers in relation to their education level, marital status, age ranges, and purchased products. We then utilized a Random Forest model to identify the factor with the most significant influence on a customer's spending potential.

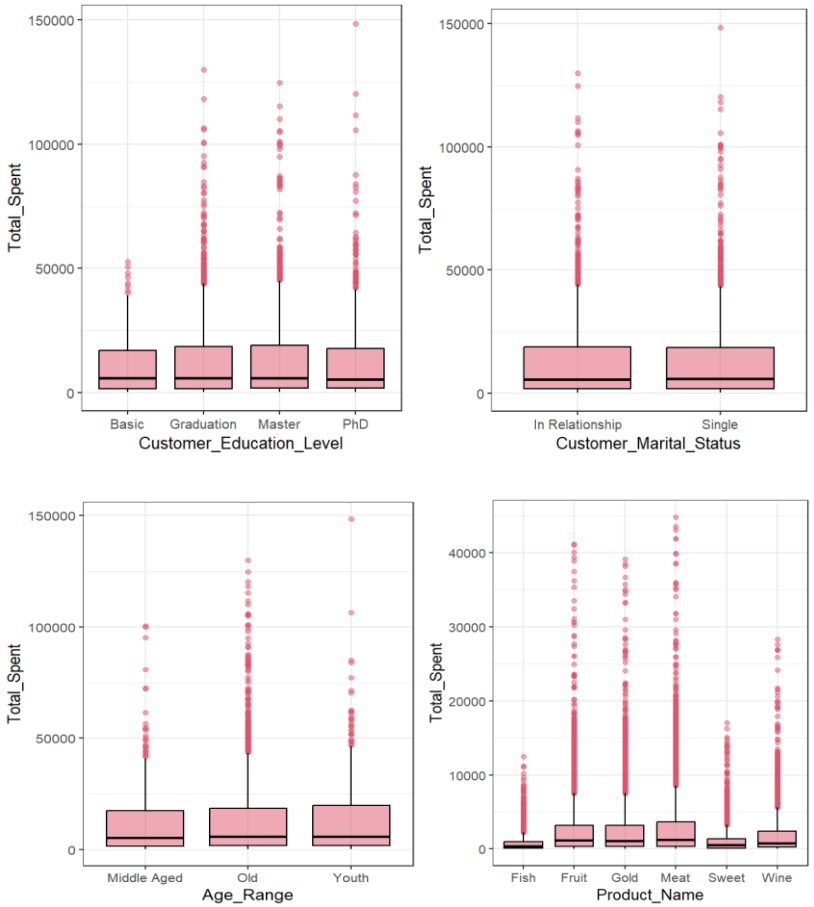

Fig. 2: Plots showing the distribution of customers

As a part of exploratory data analysis, the above 4 boxplots and dot- plots represent total spent distribution of customers based on demographic factors like Education level, Marital Status, Age and nature of product.

Through the application of the Random Forest model, we discovered that yearly income is the most significant factor impacting total purchasing behavior as illustrated below.

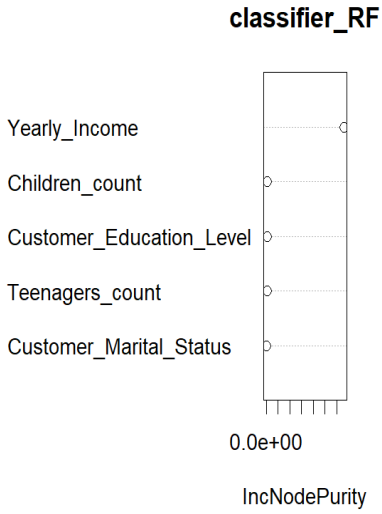

Fig. 3: Random Forest distribution showing the impact of different variables

In the final stage, we sought to comprehend the effectiveness of our campaigns. Utilizing the Naive Bayes model, we predicted the likelihood of a customer accepting offers from a specific campaign, considering their educational background and marital status.

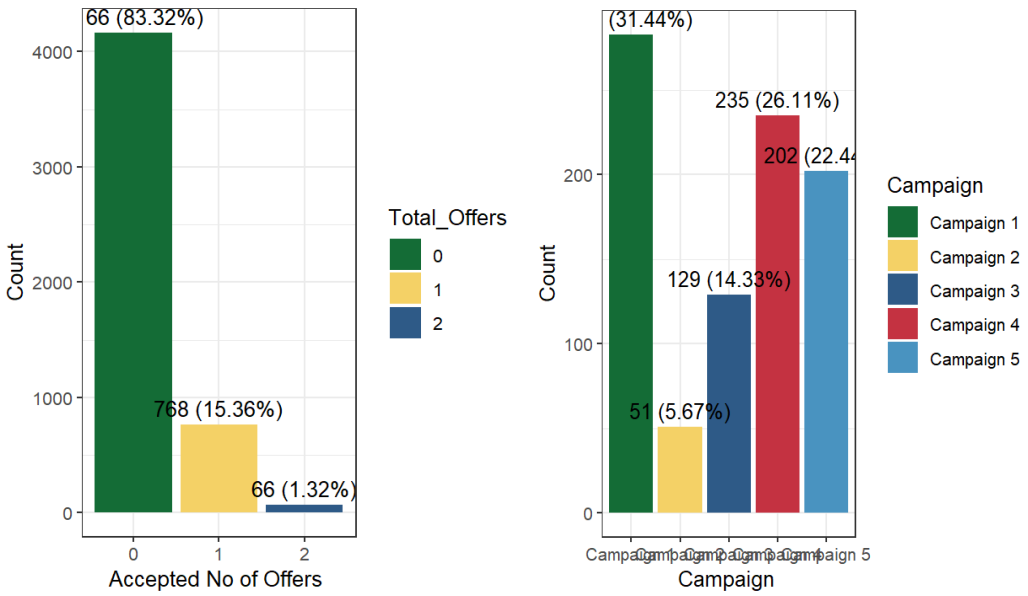

Fig. 4: Frequency distribution of offers accepted by customers based on campaigns

```
A-priori probabilities:
Y
     0      1      2
0.8332 0.1536 0.0132

Conditional probabilities:
   Customer_df$Customer_Education_Level
Y         Basic Graduation     Master        PhD
  0 0.01848296 0.38310130 0.41166587 0.18674988
  1 0.02213542 0.37760417 0.42838542 0.17187500
  2 0.03030303 0.39393939 0.42424242 0.15151515

   Customer_df$Customer_Marital_Status
Y   In Relationship    Single
  0       0.4639942 0.5360058
  1       0.4361979 0.5638021
  2       0.5303030 0.4696970
```

We have used the method of Naive Bayes to predict the probability of a customer of accepting offers from specific campaigns. The priori probabilities are observed to be 83%, 15% and 1% for 0, 1 and 2 accepted offers. We have further tried to understand the conditional probability of customers accepting 0,1 and 2 offers based on their education level and marital status.

The accuracy is seen to be 84% for the above implemented Naive Bayes model.

We can see that all the methods give us results based on the variables selected by us to understand the behavior and patterns of customers.

Conclusion and future scope

This research is significant because it provides businesses with valuable insights and predictions based on preferences and behaviours of their consumers, allowing them to optimize their product offerings, pricing strategies, and marketing campaigns, helping with customer acquisition strategies. (Predoiu, 2022)

To further enhance this study, in future, we would consider incorporating additional variables, such as geographical location, ethnicity, and lifestyle preferences, and expand the horizon in

terms of the independent variables. By perpetually analysing customer behavior and refining our approach, we can ensure that businesses remain successful in a market that is constantly evolving.

References

Custer, C., 2022. *A Complete Guide to Customer Behavior Analysis in 2023.* [Online].

Predoiu, O., 2022. *Customer Behavior Analysis – Improve Acquisition & Retention with Behavioral Data.* [Online].