College of Professional Studies

Northeastern University San Jose

MPS Analytics

Course: ALY6020: Predictive Analytics

Assignment:

Module 3 Project - Understanding Magazine Subscription Behavior

Submitted on:

Feb 1, 2024

Submitted to:

Prof: BEHZAD AHMADI

Submitted by:

NIKSHITA RANGANATHAN

INTRODUCTION

Understanding the dataset:

The dataset provides detailed information on customer profiles which includes demographic data such as age, education, and marital status, as well as spending habits across various product categories. Additionally, the dataset contains information about customers' past purchase history, campaign responses, and other personal characteristics like income level. This rich dataset is designed to enable deep analysis of customer behaviors and preferences, aiding in targeted marketing and personalized customer engagement strategies.

Below are the data descriptions of each variable of the data that briefly describe the contents of the data set:

- **ID**: This column likely represents a unique identifier for each individual in the dataset.
- **Year_Birth**: The year of birth of the individual.
- **Education**: The education level of the individual.
- **Marital_Status**: The marital status of the person.
- **Income**: The annual income of the household for the individual.
- **Kidhome**: The number of kids residing in the individual's household.
- **Teenhome**: The number of teenagers residing in the individual's household.
- **Dt_Customer**: The date when the individual became a customer.
- **Recency**: This might refer to the number of days since the last purchase or interaction.
- **MntWines**: The amount spent on wine over the past two years.
- **MntFruits**: The amount spent on fruits over the past two years.
- **MntMeatProducts**: The amount spent on meat products over the past two years.
- **MntFishProducts**: The amount spent on fish products over the past two years.
- **MntSweetProducts**: The amount spent on sweet products over the past two years.
- **MntGoldProds**: The amount spent on gold products over the past two years.
- **NumDealsPurchases**: The number of purchases made with a discount.
- **NumWebPurchases**: The number of purchases made through the company's website.
- **NumCatalogPurchases**: The number of purchases made using a catalog.
- **NumStorePurchases**: The number of purchases made directly in stores.
- **NumWebVisitsMonth**: The number of visits to the company's website in the last month.
- **AcceptedCmp**[1-5]: Indicates whether the individual accepted the offer in various marketing campaigns.
- **Complain**: Indicates whether the individual has made a complaint over the past two years.
- **Z_CostContact**, **Z_Revenue**: These columns could be related to contact cost and revenue generated, although their specific meanings are unclear without further context.
- **Response**: This could be the response to the latest campaign or a general indicator of responsiveness.

It comprises of 2240 rows and 29 features (26 numerical columns and 3 categorical columns).

DATA CLEANING

- **Handling missing values**
  There are 24 NA values in the "Income" column of the dataset. To address this, mean imputation was applied.

- **Removing duplicate rows**
  Duplicate rows in the dataset can lead to inconsistencies and affect the accuracy of data analysis. It is essential to find and remove any duplicate rows from the dataset before starting the analysis.
  There are no duplicate rows.

- **Calculating customer ages**
  "Year_Birth" column was modified to calculate customer ages by subtracting their birth year from the year 2024.

- **Dropping irrelevant columns**
  Columns 'ID' 'Z_CostContact' 'Z_Revenue' and 'Dt_Customer' were removed because they do not contribute to the analysis or the insights.

- **Simplifying Categorical Data**
  The "Education" column was categorized into broader categories: "Undergraduate," "Graduate," and "Postgraduate" and the "Marital_Status" column into "Single" or "Married" to simplify data interpretation using mapping functions.

- **Combining columns**
  The "Kids" column was created by combining data from the "Kidhome" and "Teenhome" columns.

- **Renaming column names**
  Column names in the dataset were updated for clarity

EDA

Observations from the EDA:

Figure 1:

- The first pie chart shows that half of them have a graduate degree, around a third have postgraduate education, and a smaller portion have an undergraduate degree.
- The other chart indicates that the majority are married, while a significant minority are single. This data is useful for understanding the demographic profile of a population.

Figure 2:

Key observations from the histograms:

- Amounts spent on various products indicate that more money is typically spent on wine, while expenditure on other categories like fruits, meat, fish, sweets, and gold is lower.
- The histograms for different types of purchases (deals, web, catalog, and store purchases) show varying distributions, indicating different preferences in purchasing channels among customers.
- The histogram for complaints shows that the vast majority of customers have not filed complaints, suggesting general customer satisfaction or low reporting.

Figure 3:

- There is a strong positive correlation observed between customer income and their spending behavior.
- Additionally, purchasing wine, meat, as well as using catalog and in-store channels for shopping, show significant positive correlations with both income and overall spending.

LOGISTIC REGRESSION & SVM MODEL

We started with encoding the categorical values present in the data. Encoding is a crucial data preprocessing step in machine learning and statistical analysis, particularly when working with categorical variables. Without encoding, models may not consider categorical data, leading to errors. This process also ensures consistency in representing categories and plays a vital role in improving model performance by allowing models to learn from categorical data.

Then, the dataset is split into two parts: one containing the independent variables and the other with the target variable.
Subsequently, the dataset is further divided into training and testing subsets, with a distribution of 70% for the training set and 30% for the testing set.

The feature data was standardized to ensure that all numerical features are on a similar scale and help to prevent features with large scales from dominating the model.

A **logistic regression model** is constructed. The model's accuracy is calculated, and a confusion matrix and classification report are generated to assess its performance in detail.

The **Logit** function is then applied to the data, and a summary of the logistic regression analysis is produced, providing coefficients, standard errors, z-values, p-values, and confidence intervals for each predictor. This summary helps to understand the significance of each predictor and the model's overall fit to the data.

Accuracy : The model attained an accuracy rate of roughly 88.5%, signifying the ratio of accurate predictions to the total number of predictions made..

Confusion Matrix: The confusion matrix provides insights into the model's performance in terms of true positives (564), true negatives (31), false positives (13), and false negatives (64).

The logistic regression analysis revealed significant factors impacting subscription behavior:

- Marital Status: Married individuals are more likely to subscribe, suggesting targeted marketing to increase subscriptions.
- Recency: Recent activity negatively affects subscriptions, emphasizing customer retention for better subscription rates.
- Amount - Meat: Higher meat spending boosts subscriptions, an opportunity for meat-related promotions.
- NumWebVisitsMonth: More visits lead to higher subscriptions, stressing online customer experience.
- Marketing Campaign Responses: Positive responses drive subscriptions, indicating campaign continuation and expansion.

- NumStorePurchases: Fewer in-store purchases relate to higher subscriptions, requiring re-engagement of physical store shoppers.
- Household Composition: Fewer children in households increase subscriptions, potential for tailored communication.

Regular model monitoring and adjustments are essential to ensure the ongoing effectiveness of these strategies.

The **Support Vector Machine (SVM)** model achieved an accuracy of **87.9%** in predicting subscription behavior.

CONCLUSION

The logistic regression model achieved a slightly higher overall accuracy of 0.885 compared to the SVM model's accuracy of 0.879.

It also a notably higher recall for Class 1, indicating it is better at identifying all potential positive cases. The SVM model, while having a marginally higher precision, is less capable of identifying all positive instances.

Logistic Regression revealed key predictors such as **Marital Status, Recency, and NumWebVisitsMonth**.

REFERENCES

GeeksforGeeks. (2023a, January 10). *Logistic Regression using Statsmodels*.

https://www.geeksforgeeks.org/logistic-regression-using-statsmodels/

APPENDIX

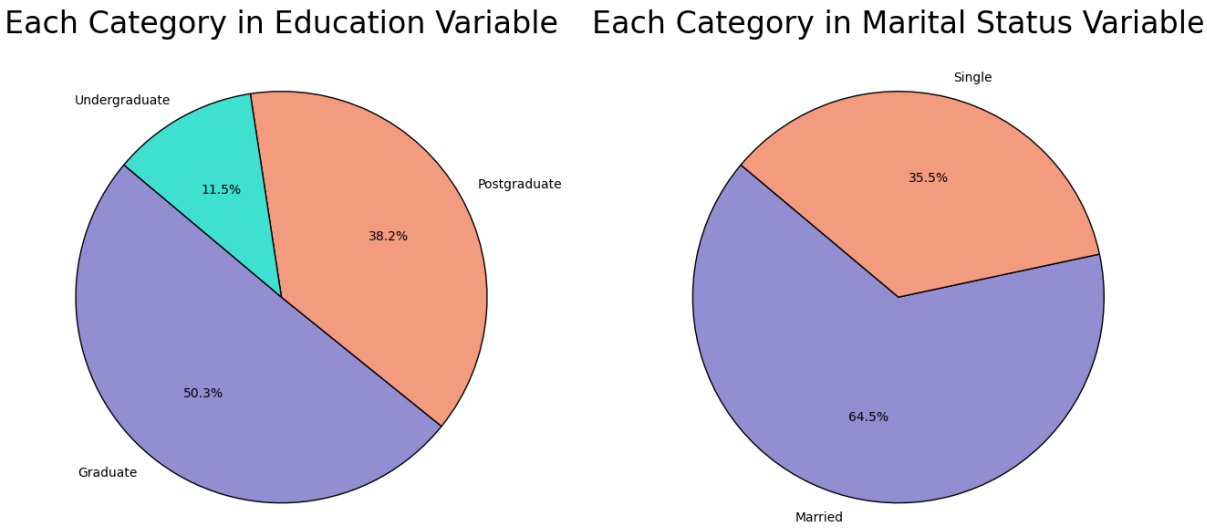

Figure 1 – Education and Marital status

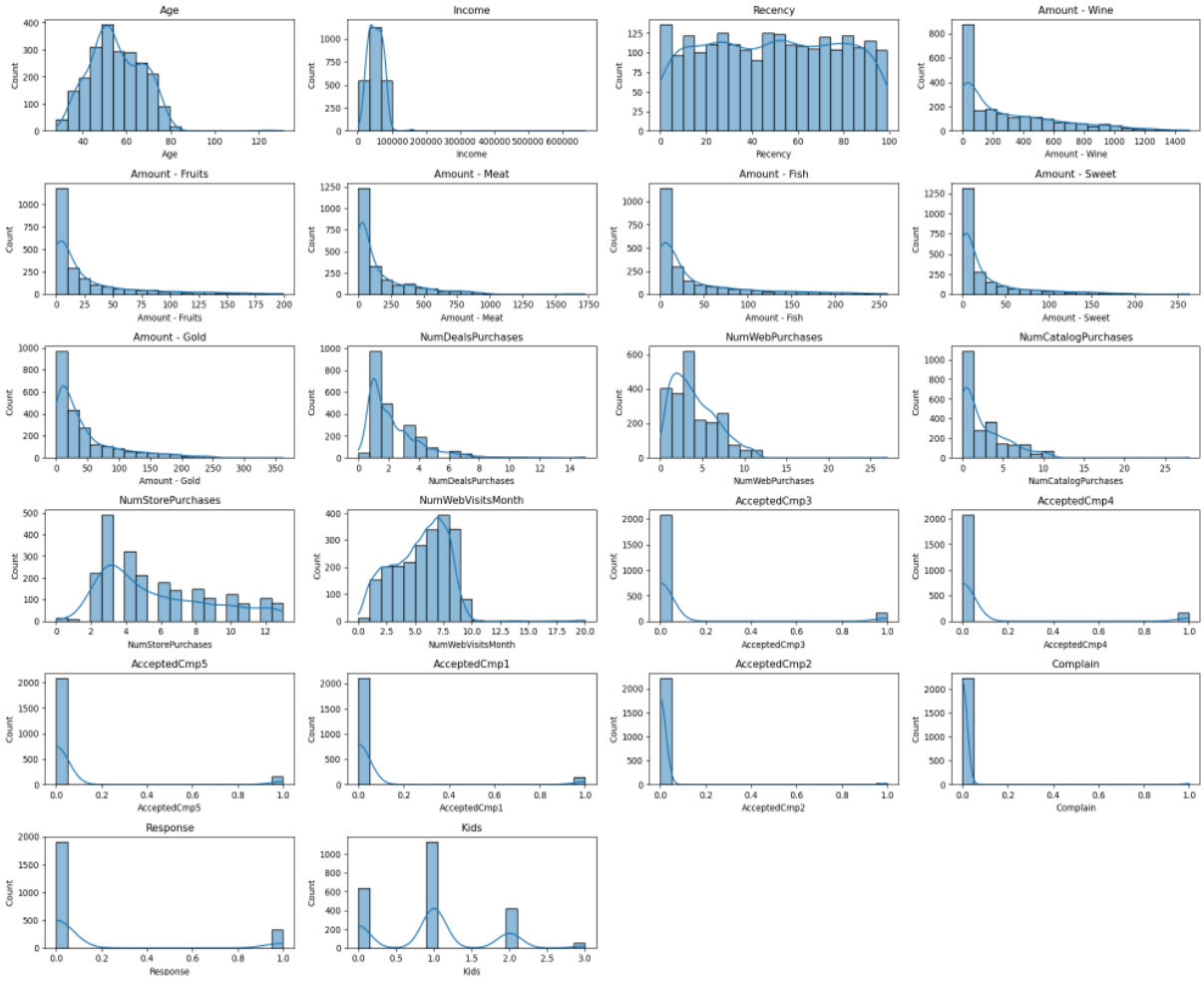

Figure 2 – Histograms for understanding distribution

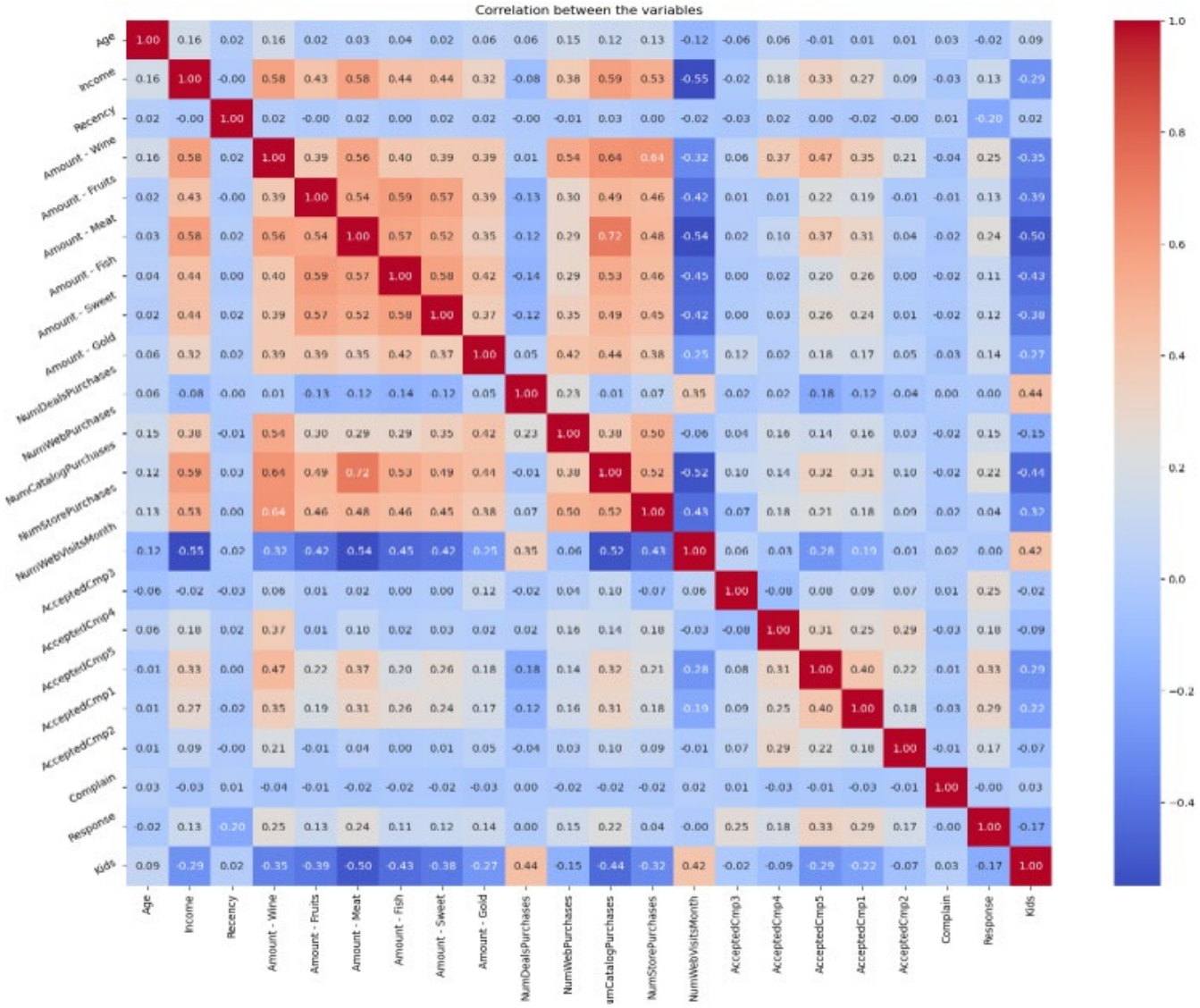

Figure 3 – Correlation matrix

```
Accuracy: 0.885
Confusion Matrix:
 [[564  13]
 [ 64  31]]
Classification Report:
              precision    recall  f1-score   support

         0.0       0.90      0.98      0.94       577
         1.0       0.70      0.33      0.45        95

    accuracy                           0.89       672
   macro avg       0.80      0.65      0.69       672
weighted avg       0.87      0.89      0.87       672
```

Figure 4 – Logistic Regression Model

```
Optimization terminated successfully.
         Current function value: 0.281541
         Iterations 8
                           Logit Regression Results
==============================================================================
Dep. Variable:               Response   No. Observations:                 2240
Model:                          Logit   Df Residuals:                     2216
Method:                           MLE   Df Model:                           23
Date:                Thu, 01 Feb 2024   Pseudo R-squ.:                  0.3315
Time:                        02:48:12   Log-Likelihood:                -630.65
converged:                       True   LL-Null:                       -943.39
Covariance Type:            nonrobust   LLR p-value:                2.077e-117
==============================================================================
                         coef    std err          z      P>|z|      [0.025      0.975]
------------------------------------------------------------------------------
const                 -3.0473      0.501     -6.086      0.000      -4.029      -2.066
Age                    0.1829      0.634      0.289      0.773      -1.059       1.425
Education              0.0967      0.110      0.880      0.379      -0.119       0.312
Marital_Status         1.1731      0.151      7.781      0.000       0.878       1.469
Income                -3.2644      4.804     -0.679      0.497     -12.681       6.152
Recency               -2.7245      0.279     -9.779      0.000      -3.271      -2.178
Amount - Wine          0.5021      0.510      0.985      0.325      -0.497       1.502
Amount - Fruits        0.4800      0.458      1.049      0.294      -0.417       1.377
Amount - Meat          3.9922      0.810      4.928      0.000       2.405       5.580
Amount - Fish         -0.1046      0.449     -0.233      0.816      -0.985       0.776
Amount - Sweet        -0.0892      0.567     -0.157      0.875      -1.200       1.021
Amount - Gold          0.5062      0.548      0.924      0.356      -0.568       1.580
NumDealsPurchases      1.5416      0.689      2.238      0.025       0.191       2.892
NumWebPurchases        2.0205      0.822      2.458      0.014       0.409       3.632
NumCatalogPurchases    1.5611      1.084      1.440      0.150      -0.563       3.686
NumStorePurchases     -1.8526      0.431     -4.296      0.000      -2.698      -1.007
NumWebVisitsMonth      4.2334      0.912      4.644      0.000       2.447       6.020
AcceptedCmp3           1.7600      0.217      8.106      0.000       1.334       2.186
AcceptedCmp4           0.8304      0.268      3.099      0.002       0.305       1.355
AcceptedCmp5           1.7513      0.271      6.460      0.000       1.220       2.283
AcceptedCmp1           1.1868      0.263      4.505      0.000       0.670       1.703
AcceptedCmp2           1.3119      0.541      2.427      0.015       0.252       2.371
Complain               0.1798      0.865      0.208      0.835      -1.515       1.875
Kids                  -1.1729      0.456     -2.573      0.010      -2.067      -0.279
==============================================================================
```

Figure 5 – Logistic model result summary (statsmodels)

```
Accuracy: 0.879
Confusion Matrix:
 [[568   9]
 [ 72  23]]
Classification Report:
              precision    recall  f1-score   support

         0.0       0.89      0.98      0.93       577
         1.0       0.72      0.24      0.36        95

    accuracy                           0.88       672
   macro avg       0.80      0.61      0.65       672
weighted avg       0.86      0.88      0.85       672
```

Figure 6 – SVM Model

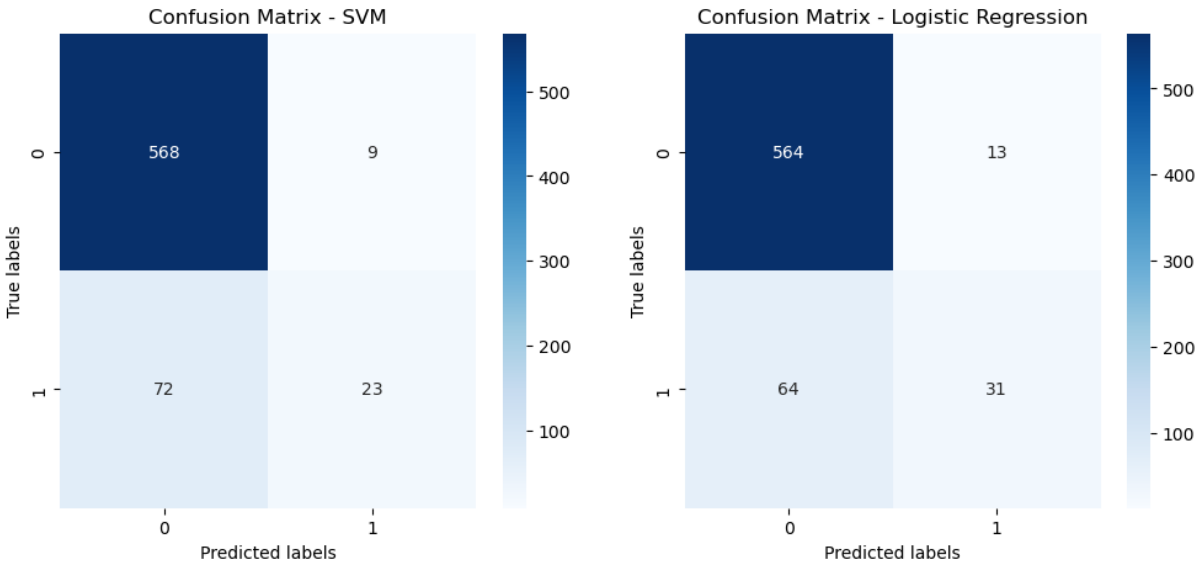

Figure 7 – Confusion matrix for both models