Northeastern University

ALY6020: Predictive Analytics

Assignment 3: Understanding Magazine Subscription Behavior

Abhigna Ramamurthy, 002982276

College of Professional Studies, Northeastern University

Professor: Justin Grosz

Jan 29, 2023

Introduction

Understanding consumer behavior is one of the main factors directly affecting the sales, marketing and even production departments of magazine. This report is aimed to do the same for a company whose prediction of more magazine subscriptions during people working form home has failed. This report aims to understand the behavior of magazine subscribers by analyzing data on subscriber socio-economic status, including age, education, marital status, children, and purchase history, including amounts spent on various categories of items, recency of purchase, and finally complains communicated. The information gathered in this report will be used to inform marketing and sales strategies for magazine publishers. The report will provide an overview of the current state of the magazine industry and identify key trends in subscriber behavior. The findings will be presented in an easy-to-understand format, and recommendations for how publishers can use this information to improve their business will be provided. Overall, the goal of this report is to help publishers better understand and connect with their audience.

Data cleaning

For this analysis, the data was presented in an excel format which was imported into Jupyter Notebook python where the analysis was conducted. Main libraries used include pandas, NumPy, and datetime for data cleaning and manipulation, seaborn, and plotly, for visualization, and statsmodels, and sklearn for data modelling. As soon as the data was loaded the unnecessary columns that were asked to be dropped were eliminated. These included accepted campaigns, revenue, and cost contact details after which there were 2,240 records and 22 attributes to work with. Out of the which 3 were objects, 1 float, and rest were integer values.

The next step was to check for missing values and found that income attribute had 24 missing values which attributed to 1.07% of the total values. As the percentage is so low and income histogram shows a normal distribution with slightly skewed to right, as shown in Figure 1, Appendix, I imputed the missing values with median value of income. The dataset was next tested to check for any duplicates with respect to the attribute ID. The results were in favor and had no duplicates. The ID column was then deleted as it does not hold any analytical value other than uniquely identifying the records.

The feature engineering process was started with checking each attribute of the dataset and make them ready for modeling. Year of birth of the customers was converted to age using datetime library so we can have better analytical leverage than just a date. This can help us understand the age group that the magazine customers fall into. However, there were 3 entries with age greater than 100 which were eliminated as they could be outliers or pending subscriptions, either way not useful for the analysis. Marital status was another object attribute that needed to be converted to numerical format for analysis. The various categories mentioned for the attribute was categorized into three as Married, Divorced, and Other. The three categories were then one-hot encoded for analysis.

The final object attribute was education. Higher education such as Master or PhD were given a value of 1 and everything else were grouped into 0 for the analysis. Dt_Customer is an attribute the shows the customer's enrollment date with the company. Initially this attribute was converted to datetime from string object data type and further converted to number of months of enrollment called months_since_joining to the magazine for easier analysis. The final data cleaning step was to check for outliers. I did find several attributes with values outside of the

upper quartile, but none of them are seemingly invalid values and hence no action against the outliers.

Exploratory Data Analysis

Exploratory data analysis helps visually uncover hidden patterns that can be useful for gaining valuable insights and convert them to magazine's advantage. Figure 1 in Appendix shows a histogram of age which indicates that magazine has a wide range of audience from different ages, 27 to 83, however their highest number of readers fall into 44-54 age group. Distribution of education level of the magazine readers can be seen in Figure 3 of Appendix. High school graduation seems to be the highest education level for most of the users and has the highest response rate in that group. About 39% of the subscribers are also married with no children. Another important observation is that since the joining or enrolment of subscribers to the magazine, we see 2-5 months from enrollment to have highest response rate as shown in Figure 4, Appendix. The correlation plot in Figure 5 of appendix shows some interesting relations.

The correlation of other attributes with 'Response' variable is what I'm most interested to investigate. Higher amount spent on wine and meet products along with a greater number of catalog or website purchases made through the magazine has a high positive correlation with responses. On the other hand, as number of days since the customer's last purchase increases, showcasing inactivity, the response is low. The result also shows that having kids or teenagers at home can contribute to lower response rates. A final interesting insight is that as the subscribers grow older, or has a marital status of married, or have been long since the enrollment for the magazine subscription, the response rate decreases which indicates that there

is an optimal age group and optimal time for reaching out that the magazine could target to increase their subscribers which is further analyzed by implementing machine learning algorithms.

Analysis

Data Pre-processing

Upon data cleaning and exploratory data analysis, I did several steps to get the data ready for modelling. Multicollinearity check was the first step where correlation between the predictor variables is quantified using variance inflation factor and helps get rid of unnecessary attributes to improve the model performance. At this stage removal of dummy variable Marital_Status_Other was necessary to bring down the vif values of other correlation values. The dataset was further divided into training and testing data as an 80-20 split for model training and testing respectively. The training data was further scaled using standard scaler to ensure all the independent attributes are in the same range.

Data Modeling

Logistic Regression: This is the first model to be implemented on the dataset. The aim of this analysis is to understand the behavior of the subscribers of the magazine. This analysis provides a more concrete understanding than the exploratory data analysis as logistic regression considers all the interactions between the various independent attributes. The logistic regression model output displays significant variables which can be determined by the P(>|t|), p-value, and depicts the positive and negative impact over dependent variable price through the coefficients. The Significance of the variable is determined if the P-Value is less than the significance level (0.05 A significance level of 0.05 indicates a 5% risk) then it terms that the

model fits the data well. The results of the regression model are shown in Figure 6 of Appendix. The two most significant attributes obtained in this analysis are “amount spent on wine purchases in the last 2 years” which has a coefficient of +0.36 which means 36% increase in response rates for every amount spent on wines and the second most significant attribute is “number of purchases made directly in store” with coefficient of -0.35 which indicates 35% decrease in response rate for every purchase made directly in store instead of through the magazine or its website.

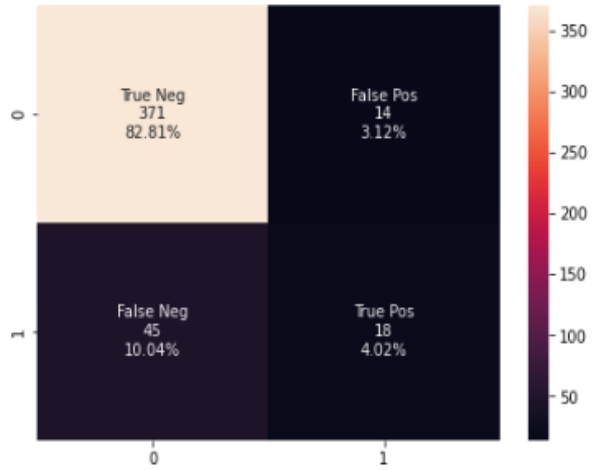

Figure 7: Confusion Matrix for Logistic Reg.

The confusion matrix shown here represents the summary of performance for the model. As seen in the Confusion Matrix result, our Model has an accuracy of 86.83%. The model also yielded 56.2% precision which means the model makes a correct prediction 56.2% of the times and recall of the model is 28.6% which means if the response is 1 then the model can predict it 28.6% of the times. The F1-score is 0.849 which indicates good model performance. AUC, area under the curve for logistic regression is 0.62 which indicates the model has an average performance. It's correctly identifying more true positive cases than false positive cases. It's correctly identifying more true negatives than false negatives. But it's not having a perfect performance. It's still leaving room for improvement. The mean squared error value is 0.132 which tells how close a regression line is to a set of points. This can be used in comparison of the models.

SVM: Support Vector Machine is a supervised machine learning algorithm which goal is to find the best boundary (or "hyperplane") that separates the different classes in a dataset. This boundary is chosen so that it maximally separates the classes while also maximizing the margin, which is the distance between the boundary and the closest points from each class. Once the boundary is determined, new data can be classified by seeing which side of the boundary it falls on. SVMs can also handle non-linearly separable data by using a technique called kernel trick which helps to project the data into a higher dimension space. One disadvantage of SVM is that the significant features of the model are not easy to extract. Hence the comparison to logistic regression can only be made using the benchmarking metrics.

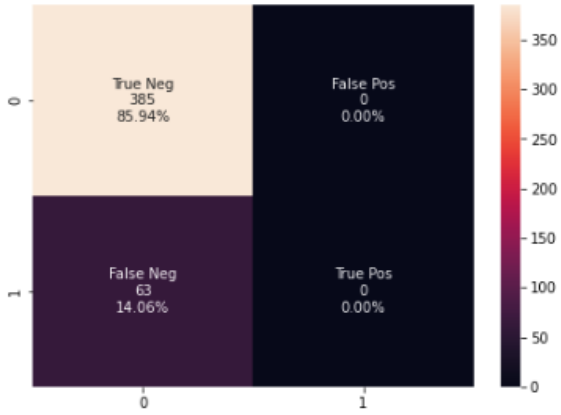

Figure 8: Confusion Matrix for SVM

The confusion matrix shown here represents the summary of performance for the model for the SVM model. As seen in the Confusion Matrix result, our Model has an accuracy of 85.9% which is almost same performance level as logistic regression, however accuracy is not the only metric to investigate. The model also yielded 0% precision and 0% recall which is concerning as number of true positives and false positives are 0 in the test set which nullifies the accuracy of the model as a performance metric. The F1-score is 0.794 which indicates good model performance in the true negative or false negative front. AUC, area under the curve for logistic regression is 0.62 which indicates the model has an average performance. The mean squared error value is 0.141 which higher than that of logistic regression. The reason we see a higher performance when the response if 0 is because the

dataset is highly imbalanced w.r.t response dependent variable. More than 80% of the data was no response. Hence by the mechanism of SVM, the division of the hyperplane favored more on no response value than a response value of 1.

Model comparison using benchmarking metrics

Model comparison is a crucial aspect of choosing which model works best for the given data parameters. The benchmarking metrics chosen for evaluation in this report are accuracy, precision, recall, mean squared error, F1 score, and false positive rate. False positive rate (FPR) is a measure of the proportion of negative instances that are incorrectly classified as positive by the models. In the context of the magazine subscription, FPR indicates the model to have predicted a response of 1 where the actual was 0.

Model	***Accuracy***	***Precision***	***Recall***	***MSE***	***F1 score***	***FPR***	***AUC***
Logistic Regression	86.8%	56.2%	28.6%	0.13	0.85	3.12%	0.62
SVM	85.9%	0%	0%	0.14	0.79	0%	0.62

Table 1: Benchmarking metrics used in the two models

Based on the above model results shown in Table 1, Logistic Regression comes out as the best model on all aspects. However, FPR is higher than that of SVM model but knowing that SVM favors response value of 0 as seen in Figure 8 confusion matrix, logistic regression model is a more well-rounded model for the given data. Logistic also has lower mean squared error value with higher precision, recall, accuracy, as well as F1 score. Hence logistic regression is the model chosen for this analysis indicating wine purchases, and store purchases are significant variables to consider for betterment of magazine subscriptions.

Conclusion

From the above analysis we can conclude that higher performance is obtained from the logistic regression model. The attributes “MntWines” (amount spent on wine purchase in the last 2 years) and “NumStorePurchases” (number of purchases made directly in stores) are the two most significant features that drive magazine subscription. It also means, more money spent on wine purchases, and lesser direct store purchases can increase the responses to the magazine subscription.

Amount spent on wine showcases a bigger area of improvement for the magazine that can be taken advantage of for gaining more subscribers for the magazine. Given the age group of magazine users, we see that magazine is seen as a more relaxing lifestyle choice at the same time keeping in mind that majority of the customers have median income ranges. Showcasing more cost-effective ways to enjoy their free time, including products like wine, and other food cultures can increase interests in reading the magazine. This can help improve magazine content and at the same time do more targeted approach to outreach efforts of magazine subscriptions.

On the other hand, number of purchases in the store directly showcases limited interest for the customers and will largely drive by the look of the magazine, hence driving down the subscribers. Customers who make frequent purchases in a store may have a specific interest in the products or services offered by the store and may not be interested in receiving information about other products or services. This might also indicate that such persona of customers might prefer digital content and suggests the magazine to make more digital accessibility to customers more seamless. On a final note, this information can help magazine

business owners to make informed decisions on where to focus their efforts to drive up their outreach efforts for acquiring more subscribers.

References

Classifying data using Support Vector Machines (SVMs) in Python - GeeksforGeeks. (2023). GeeksforGeeks. https://www.geeksforgeeks.org/classifying-data-using-support-vector-machinessvms-in-python/

How to Import an Excel File into Python using Pandas – Data to Fish. (2022). Data to Fish – Data Science Tutorials. https://datatofish.com/read_excel/

Li, S. (2017, September 29). Building A Logistic Regression in Python, Step by Step. Medium. https://towardsdatascience.com/building-a-logistic-regression-in-python-step-by-step-becd4d56c9c8

Log in - Loading Session Information. (2023). Log in - Loading Session Information. https://northeastern.instructure.com/courses/131156/pages/lesson-3-5-coding-differences-between-svm-and-logistics-regression?module_item_id=8228171

Narkhede, S. (2018, May 9). Understanding Confusion Matrix. Medium. https://towardsdatascience.com/understanding-confusion-matrix-a9ad42dcfd62

Appendix

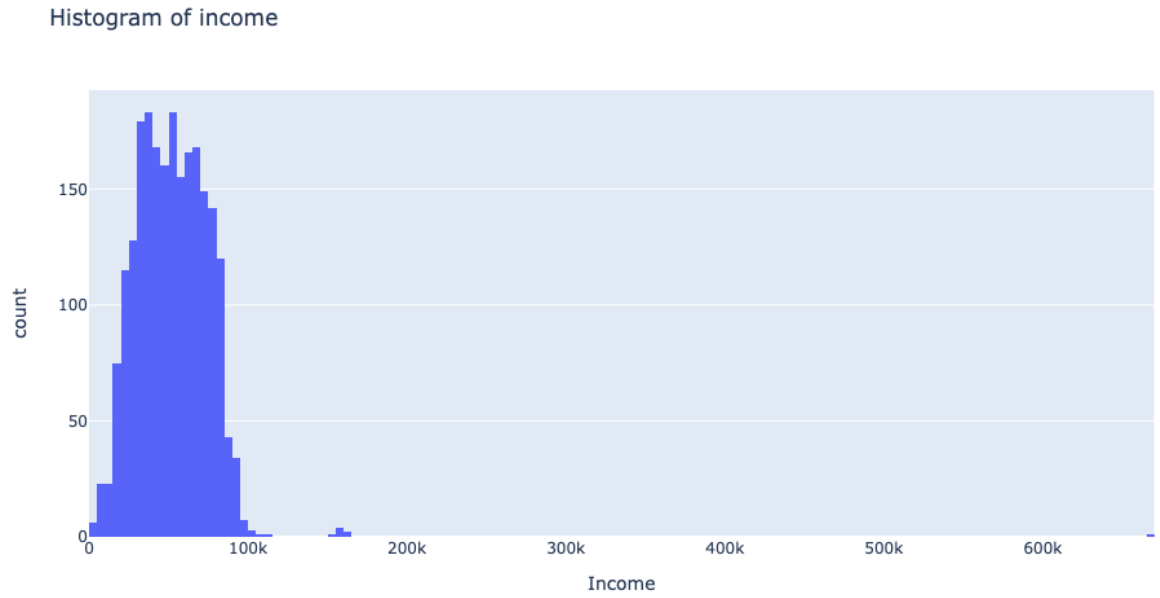

Figure 1: Histogram of income.

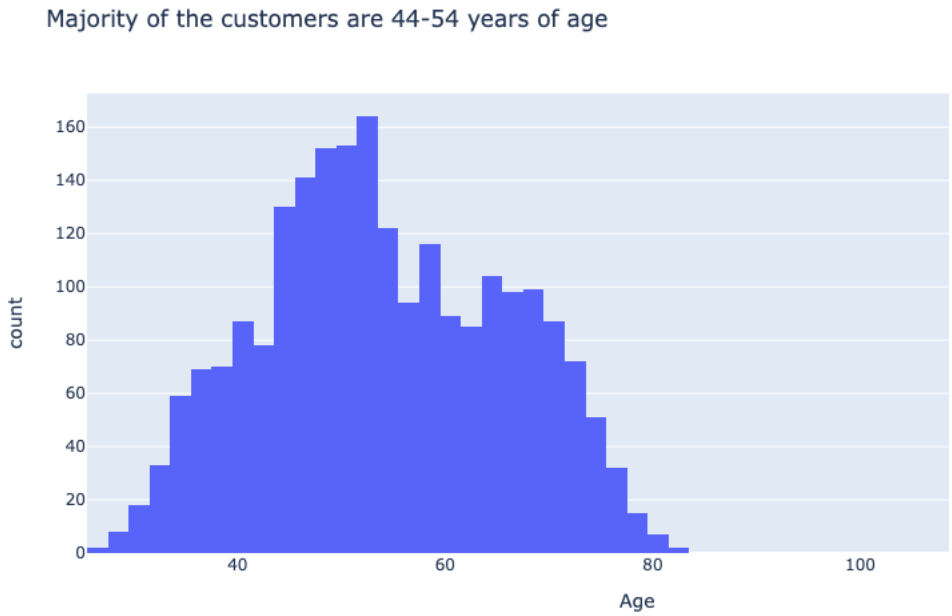

Figure 2: Histogram of age.

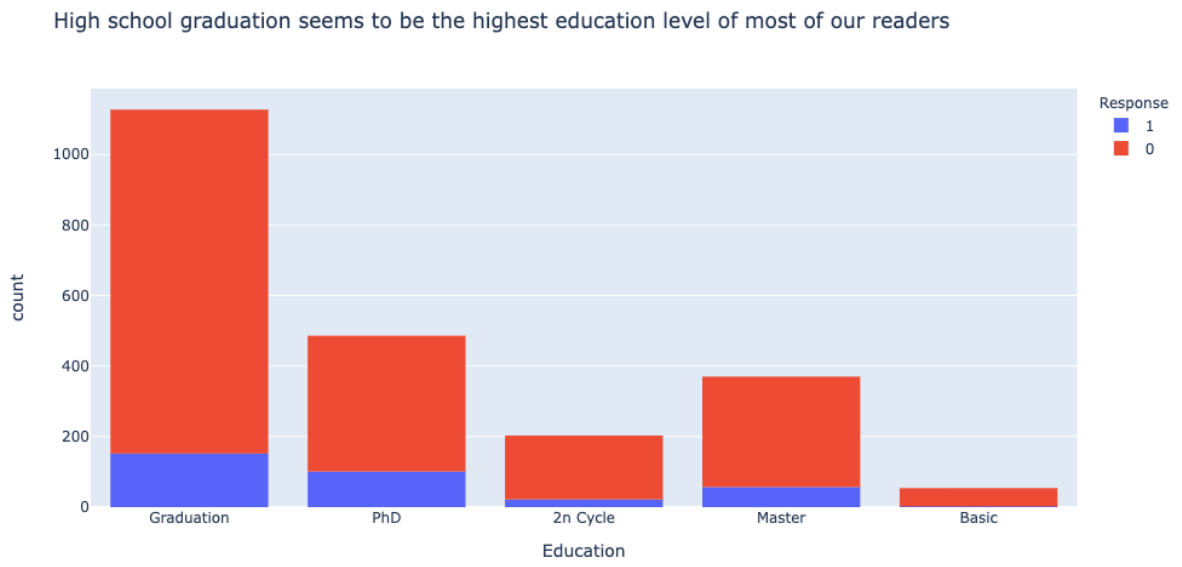

Figure 3: Histogram of education

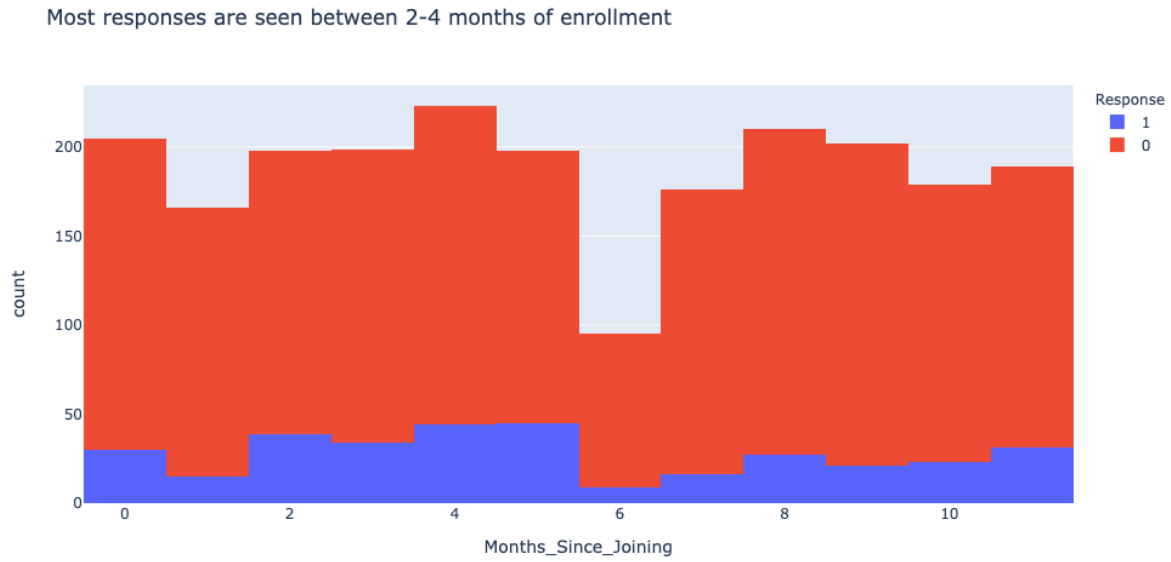

Figure 4: Histogram of months since joining.

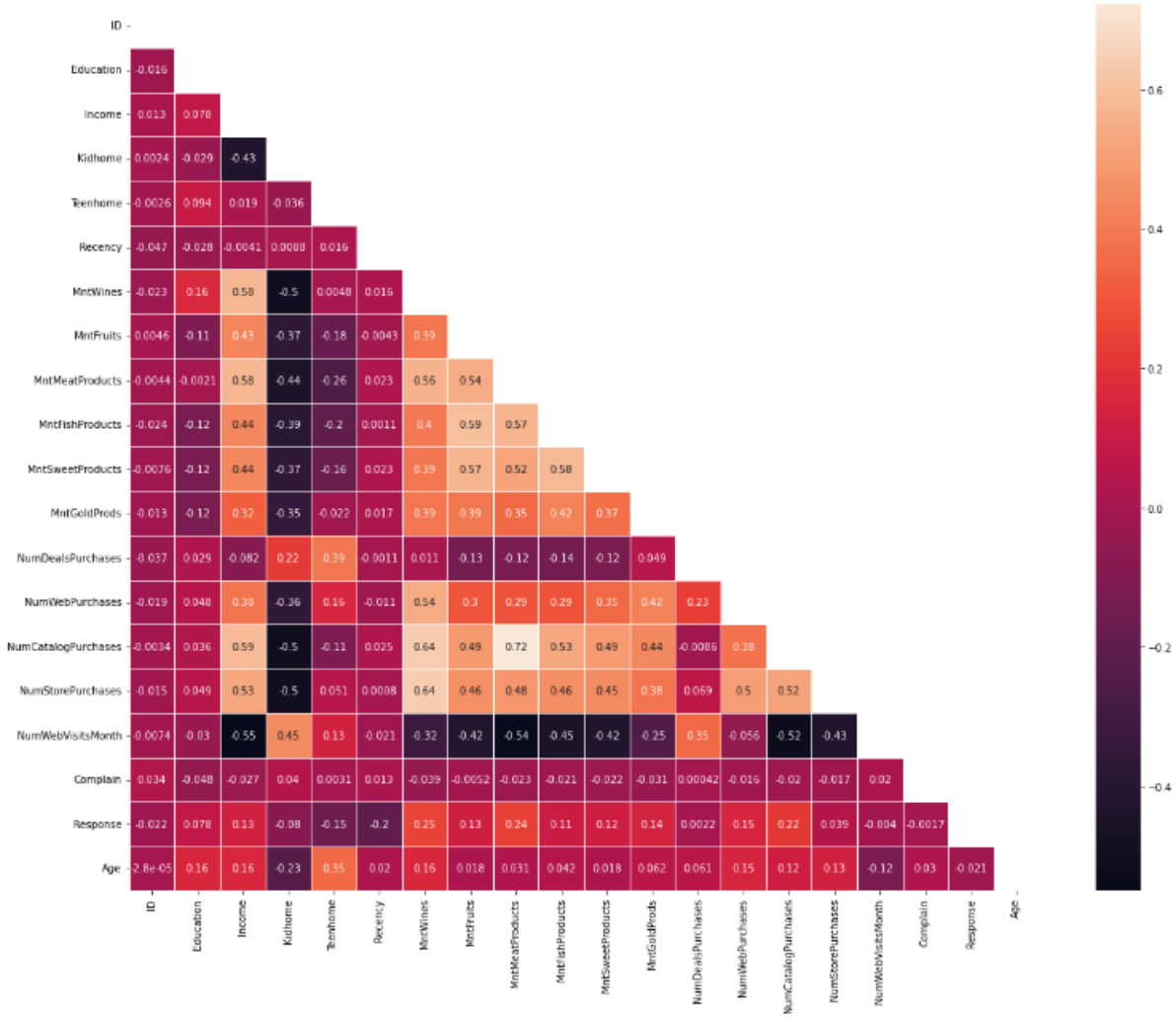

Figure 5: Correlation plot.

```
Optimization terminated successfully.
         Current function value: 0.641498
         Iterations 5
                           Logit Regression Results
==============================================================================
Dep. Variable:                Response   No. Observations:                 1792
Model:                           Logit   Df Residuals:                     1771
Method:                            MLE   Df Model:                           20
Date:                 Sun, 29 Jan 2023   Pseudo R-squ.:                -0.5100
Time:                         11:39:10   Log-Likelihood:               -1149.6
converged:                       True   LL-Null:                      -761.30
Covariance Type:            nonrobust   LLR p-value:                    1.000
===============================================================================================
                                  coef    std err          z      P>|z|      [0.025      0.975]
-----------------------------------------------------------------------------------------------
Education                       0.0953      0.053      1.785      0.074      -0.009       0.200
Income                         -0.0083      0.069     -0.120      0.904      -0.144       0.127
Kidhome                         0.0196      0.069      0.284      0.777      -0.116       0.155
Teenhome                       -0.1815      0.063     -2.862      0.004      -0.306      -0.057
Recency                        -0.3332      0.051     -6.578      0.000      -0.433      -0.234
MntWines                        0.3564      0.089      3.990      0.000       0.181       0.531
MntFruits                       0.0230      0.073      0.317      0.751      -0.119       0.165
MntMeatProducts                 0.2644      0.093      2.852      0.004       0.083       0.446
MntFishProducts                 0.0011      0.075      0.014      0.989      -0.145       0.147
MntSweetProducts                0.0412      0.070      0.590      0.555      -0.096       0.178
MntGoldProds                    0.0844      0.063      1.342      0.180      -0.039       0.208
NumDealsPurchases               0.0239      0.066      0.360      0.719      -0.106       0.154
NumWebPurchases                 0.0592      0.069      0.855      0.392      -0.077       0.195
NumCatalogPurchases             0.1637      0.096      1.702      0.089      -0.025       0.352
NumStorePurchases              -0.3527      0.079     -4.474      0.000      -0.507      -0.198
NumWebVisitsMonth               0.2507      0.078      3.197      0.001       0.097       0.404
Complain                        0.0278      0.050      0.557      0.577      -0.070       0.126
Age                            -0.0155      0.056     -0.275      0.784      -0.126       0.095
Marital_Status_Divorced         0.0803      0.052      1.546      0.122      -0.021       0.182
Marital_Status_Married         -0.0935      0.052     -1.795      0.073      -0.196       0.009
Months_Since_Joining           -0.0303      0.050     -0.602      0.547      -0.129       0.068
===============================================================================================
```

Figure 6: Logistic Regression results.