Understanding Magazine Subscription Behavior

Ujwal Kumar

11/20/2022

Content

Introduction

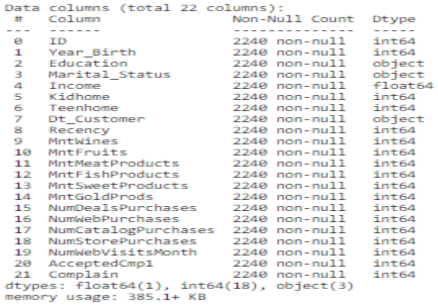
We will do a customer study based on the customer's activity on the website or app, how they spend their money on their lifestyle, and an examination of their education, family, and spending history. We will categorize the types of consumers who are likely to join up for a website's premium membership. We will be able to execute targeting-based marketing or advice to consumers who are likely to sign up for the premium membership plan after evaluating and categorizing the dataset. Furthermore, the organization should concentrate on all aspects of the company in order to get the best results and outcomes.

We will first utilize feature engineering approaches to ensure that we have adequate and valid data for this problem, and then we will divide the data into training and testing phases for prediction. During the modeling phase, we will utilize the scikit-learn library's Logistic Regression and Support Vector Machines (SVM) algorithms.

Data Cleaning

The data collection has around 2240 records and 22 characteristics. Except for Education, Marital Status, and DT Customer, each characteristic has categorical values. The data type of the DT Customer feature is an object, and the values are in the format YYYY-MM-DD. We will process data on these columns as we progress through the Exploratory Data Analysis and Feature Engineering stages.

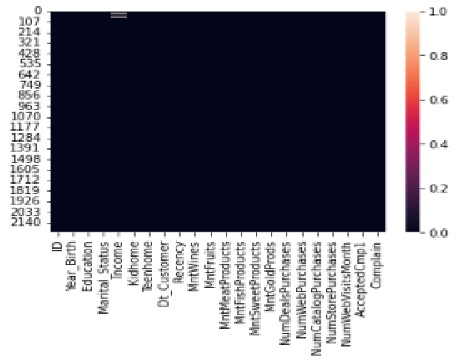
First and foremost, the dataset must be explored and learned about by evaluating the records, such as column information. Most columns have int datatypes, with a few having object datatypes.



```
Data columns (total 22 columns):
 #   Column               Non-Null Count  Dtype  
---  ------               --------------  -----  
 0   ID                   2240 non-null   int64  
 1   Year_Birth           2240 non-null   int64  
 2   Education            2240 non-null   object 
 3   Marital_Status       2240 non-null   object 
 4   Income               2240 non-null   float64
 5   Kidhome              2240 non-null   int64  
 6   Teenhome             2240 non-null   int64  
 7   Dt_Customer          2240 non-null   object 
 8   Recency              2240 non-null   int64  
 9   MntWines             2240 non-null   int64  
 10  MntFruits            2240 non-null   int64  
 11  MntMeatProducts      2240 non-null   int64  
 12  MntFishProducts      2240 non-null   int64  
 13  MntSweetProducts     2240 non-null   int64  
 14  MntGoldProds         2240 non-null   int64  
 15  NumDealsPurchases    2240 non-null   int64  
 16  NumWebPurchases      2240 non-null   int64  
 17  NumCatalogPurchases  2240 non-null   int64  
 18  NumStorePurchases    2240 non-null   int64  
 19  NumWebVisitsMonth    2240 non-null   int64  
 20  AcceptedCmp1         2240 non-null   int64  
 21  Complain             2240 non-null   int64  
dtypes: float64(1), int64(18), object(3)
memory usage: 385.1+ KB
```

(Figure 1: Data Types of Columns)

After reviewing the data types of the columns in the data collection. The next step we took was to look for null values. We were able to detect null values using Python functions, and they were present in the column titled "Income." In the column, there were 24 null values. The figure below depicts the depiction of null values.



(Figure 2: Null Values Visualization)

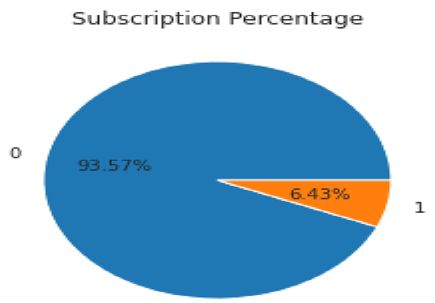
The white line in the accompanying graphic represents null values, which are present in the feature labeled "Income." Following that, we estimated the percentage of null values in the column, which was 1.07%. To be more detailed, we did not directly utilize mean and fill the values while filling the null values with acceptable data. We filled in the "Income" column with null values based on the sort of education they had and then filled in values with the mean of that education group. For example, under the income column, we picked the mean of those who had completed their Ph.D. and filled it with subsequent null values. This was done to obtain the most appropriate income numbers.

Exploratory Data Analysis

To see how Subscription differs from the other variables in the dataset. We will investigate several numbers to discover Subscription count patterns and which factors impact Subscriptions. Kindly refer to **Appendix 1** for EDA graphs and plots.

Some data preparation was required prior to beginning the EDA step. For instance, we have included additional columns from the Dt Customer feature, such as Year and Month, so that we have continuous and relevant data to display and subsequently use in model building. Then we divided the age column into three categories, such as individuals born before 1959 being regarded as Elderly, people born between 1959 and 1977 is considered as Middle Age, and people born after 1977 being considered Young.

We used graphs like count plots, pie charts, and bar plots to examine how Subscription behavior differs with different attributes in the dataset. We also attempted to discover which specific group from a feature is most likely to pick up a subscription, as well as which groups are not taking up subscriptions in order to analyze the pattern. The graph below depicts the proportion of people who have signed up for the subscription. In the diagram, 1 symbolizes the individual who has subscribed, and 0 represents the individual who has not subscribed.



(Figure 3: Percentage of Subscriptions Taken)

The preceding statistic plainly shows that the proportion of individuals who have taken the subscription offer is far lower than the percentage of people who have not taken the subscription offer. Individuals who are married are more likely to take a subscription offer, followed by couples who live together and people who are single, as we went with our visualization using the Marital Status feature. Individuals who have completed graduation are the most likely to take up a subscription, followed by Ph. D holders in the Education column. Following that, we determined which groups of people are eligible for subscriptions depending on their family size. We had two options for it: Kids Home and Teen Home. One similar tendency in both features was that people are more inclined to accept subscriptions when they do not have children or teenagers at home, but not when they do. The number of subscription offer taken is rapidly decreasing as they have children or teens at home.

After examining the subscription count with features, we looked at where people were spending their money if they weren't on subscriptions. To begin with, a similar pattern was noticed in subgroups such as Married, Together, Single, Divorced, and so on. The most money was spent in the last two years on wine, followed by meat products, gold products, and every group spent the least on fruit.

Now looking at the purchase type according to the age of the people, across age groups a similar pattern was observed and people are more likely to purchase it from stores and then online. Then we progressed to check for complaints, the most complaints are recorded who are older or young.

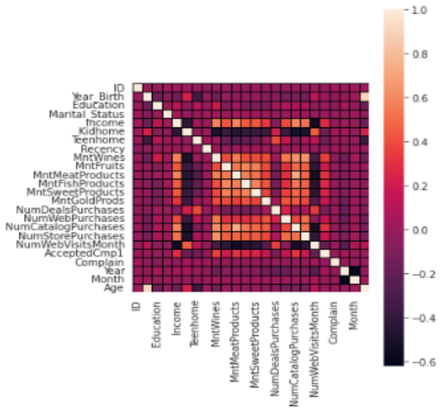
All of this was done to see how subscription acceptance changes with different aspects in the data set, which group we can target more, and which group the organization has to work harder on so that in the near future they also choose subscriptions in the first place. Then we looked at where everyone was spending their money and what kind of purchases they were making.

Feature Engineering

We now have sufficient knowledge about the properties of the data sets when the data cleaning and EDA stages are completed. We have transformed columns in this step that will be more useful in the model-building stage.

To completely depict the dataset, we employed multiple ways to encode the columns that were in the dataset or that we developed during the EDA step. To encode categorical information such as Education, Marital Status, and Age column, Label Encoding was employed. Following the conclusion of categorical characteristics, we began investigating numerical factors. We looked for outliers in numerical columns like income, the amount spent on wine, meat, gold, and so forth. Because the data set we have only comprised 2240 records, we decided to delete the values that are outside of the 5 to 95% range for a given column and classify them as outliers before going to the modeling step.

After we have our data ready for modeling, we first look at how the characteristics in the data are associated with one another, so we know which features have some association with our goal feature which is whether or not the Subscription offer is taken.



(Figure 4: Correlation Matrix)

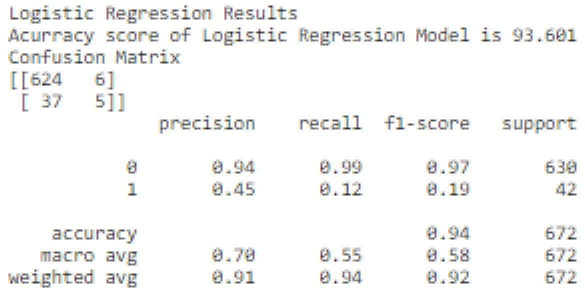
Modeling

After concluding the preliminary evaluation of the data set, this report will perform additional analysis by developing models based on the preliminary findings found in the EDA. In this section, the data will be separated into two parts: training and testing. Following that, we will run the data through two Machine Learning Algorithms: Logistic Regression and Support Vector Machines. **Appendix 2** contains all of the graphs, classification reports, and confusion matrices used to visually examine the findings.

Logistic Regression and Support Vector Machine Model

Following the division of the data into train and test data. We utilized two techniques to show the findings, and we will determine which algorithm to take out of the two.

We fitted the training data after declaring the model objects, then utilized the models for predictions and assessed the outcomes. The models' output is shown below.

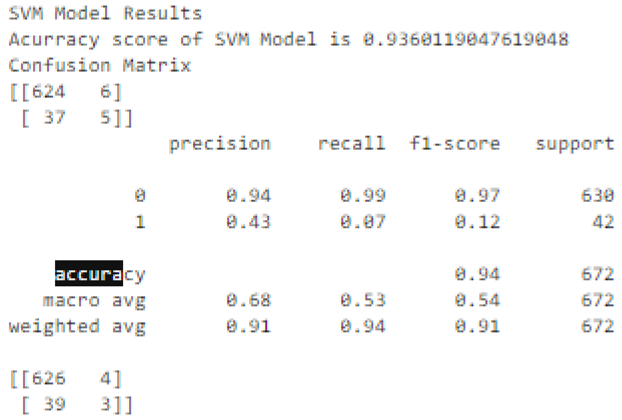


```
Logistic Regression Results
Acurracy score of Logistic Regression Model is 93.601
Confusion Matrix
[[624   6]
 [ 37   5]]
              precision    recall  f1-score   support

           0       0.94      0.99      0.97       630
           1       0.45      0.12      0.19        42

    accuracy                           0.94       672
   macro avg       0.70      0.55      0.58       672
weighted avg       0.91      0.94      0.92       672
```

(Figure 5: Logistic Regression Model Results)



```
SVM Model Results
Acurracy score of SVM Model is 0.9360119047619048
Confusion Matrix
[[624   6]
 [ 37   5]]
              precision    recall  f1-score   support

           0       0.94      0.99      0.97       630
           1       0.43      0.07      0.12        42

    accuracy                           0.94       672
   macro avg       0.68      0.53      0.54       672
weighted avg       0.91      0.94      0.91       672

[[626   4]
 [ 39   3]]
```

(Figure 6: SVM Model Results)

According to the figure above, the findings for both models are similar. Before delving into the details of the model findings, let's go over the elements that are driving them. Using SHAP, p values of features, we figured out that, Income, MntWines, MntFishProducts, and MntMeatProducts are driving the results for both models. That is, for a person to receive a subscription, the most essential factor is his or her income, followed by how he or she spends his or her income on other things such as lifestyle and home items.

Let's go further into the model findings now. To begin with, both models are equally accurate. Next, we'll look into accuracy and recall. Precision is measured by dividing the number of true positives by the total number of true positives and false positives. Precision values for both models when predicting which individual will accept the subscription are 0.45 (Logistic Regression) and 0.43 (SVM Model). This suggests that the Logistic Regression model produces fewer false positives than the SVM Model. Recall is computed by dividing true positives by the total of true positives and false negatives. When forecasting that a certain individual would choose a subscription, the recall values for both models are 0.12 (Logistic Regression) and 0.07. (SVM Model). That is, the SVM model has a greater number for false negatives than the Logistic Regression model.

Following an examination of precision, recall, and accuracy. Because the accuracy of both models was the same. Precision and recall were the key factors in the report**. Logistic Regression Model** outperformed SVM Model in terms of precision and recall since it had fewer false positives and false negatives when predicting whether or not a certain user will sign up for a subscription.

Conclusion

The following approach should be considered by the magazine company to increase the subscription count:

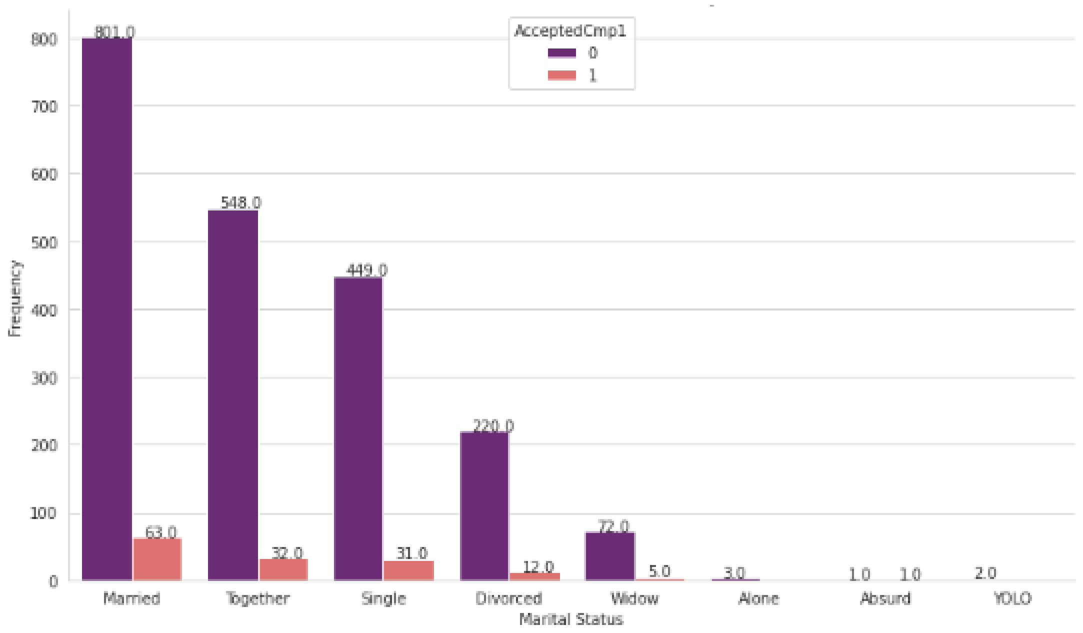
- People are spending more money on goods like wine, meat, and groceries. The organization should spend on promotion and collaborate with stores that clients frequently visit. They should maintain a small portion of the personnel on hand to promote their publication and the benefits of signing up for a membership. This would result in a favorable effect since anytime consumers go to malls or stores to buy food or home items, they will see an advertisement or meet with a corporate representative. So, subconsciously, the magazine subscription will be in their mind, and it is likely that they will check it out and subscribe to the magazine.
- As Income was the most important feature in our case study, the company should categorize people based on their earnings. To begin increasing the number of subscriptions, they should concentrate on those who make a good amount of living since they are more likely to take the subscription. Following this procedure, the wider picture will be to reach out to all segments of society and conduct advertisements in such a way that those earning less money or who have families would also subscribe.

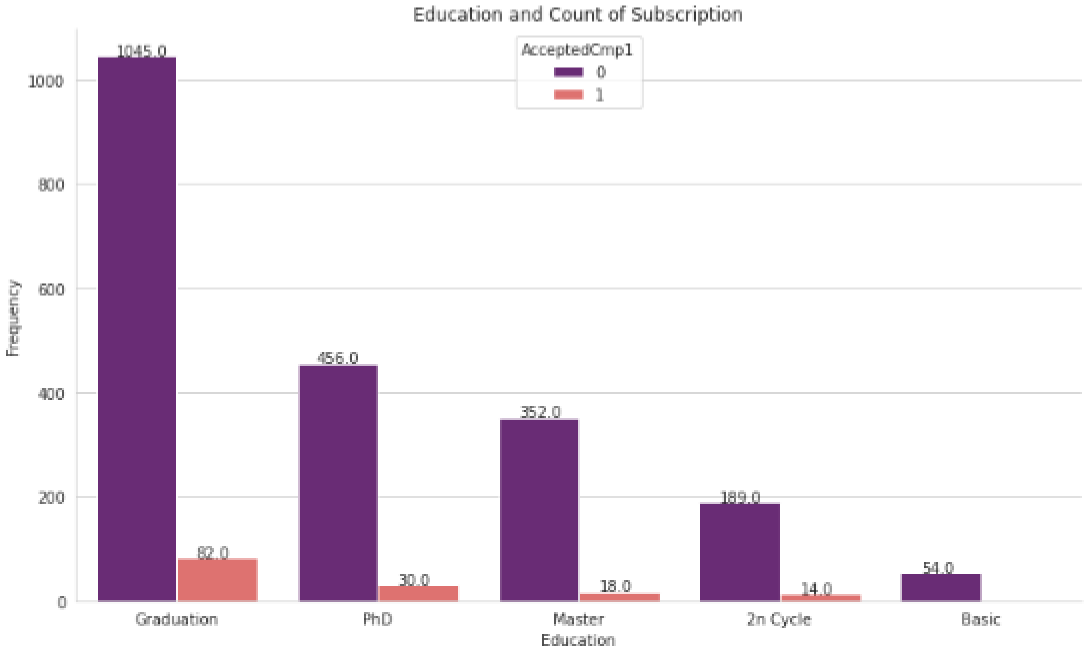
References

- *Sklearn.svm.SVC*. scikit. (n.d.). Retrieved November 20, 2022, from https://scikit-learn.org/stable/modules/generated/sklearn.svm.SVC.html
- *Sklearn.neighbors.kneighborsclassifier*. scikit. (n.d.). Retrieved November 6, 2022, from *What is logistic regression?* Statistics Solutions. (2022, June 14). Retrieved November 20, 2022, from https://www.statisticssolutions.com/free-resources/directory-of-statistical-analyses/what-is-logistic-regression/

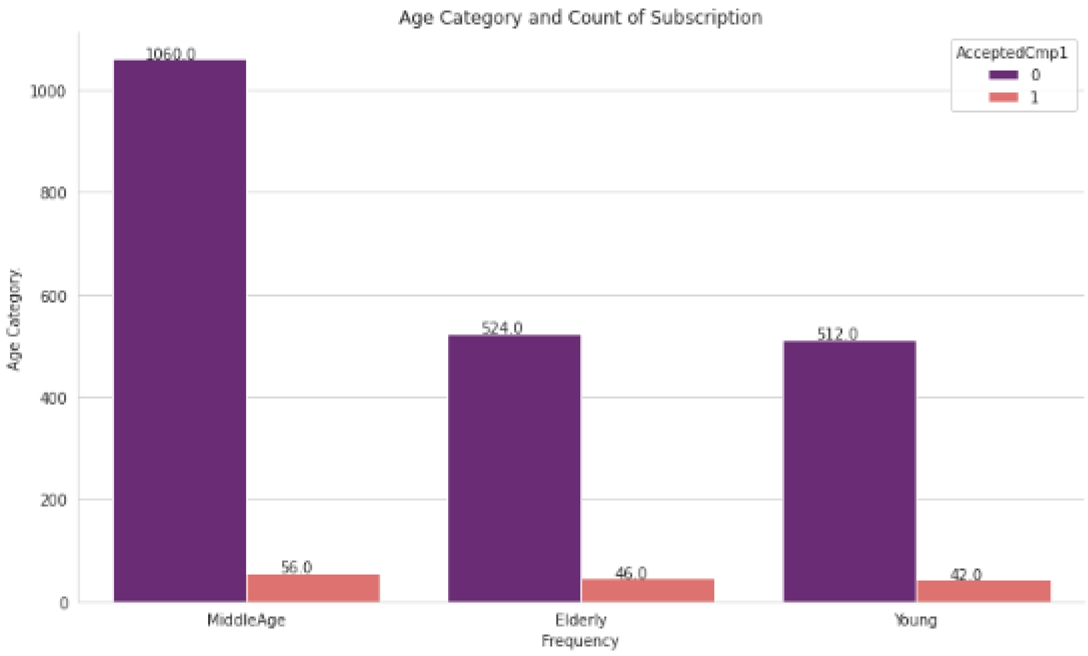
Appendix 1

Count Plot of Marital Status

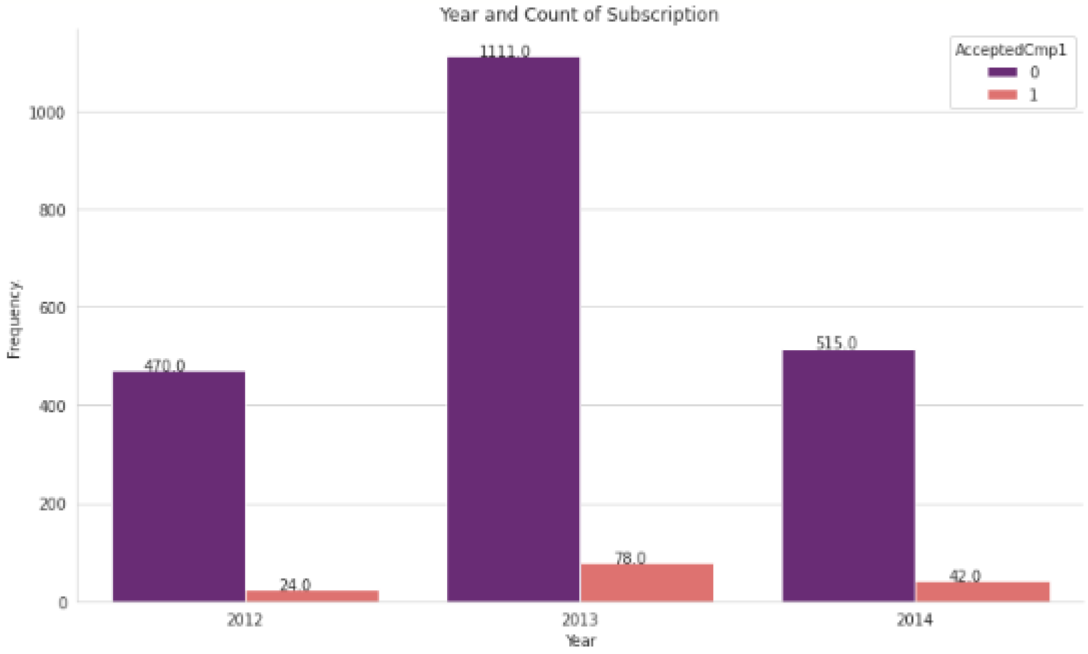



Education and Count of Subscription

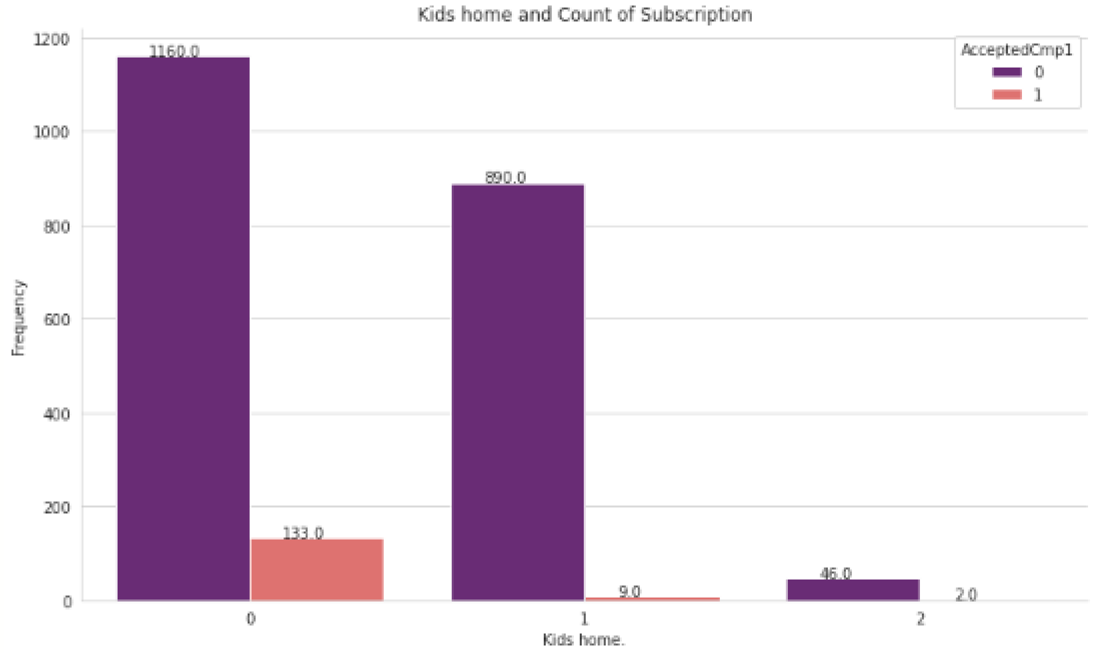
Count Plot of Age Category



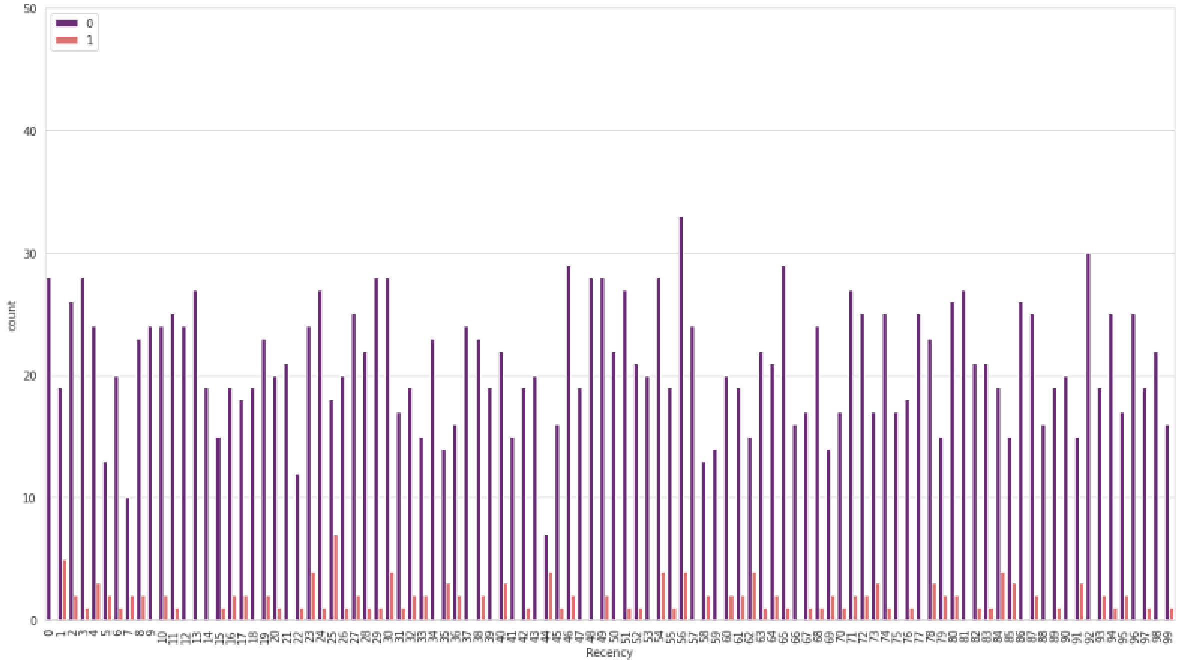
Count Plot of Year Feature



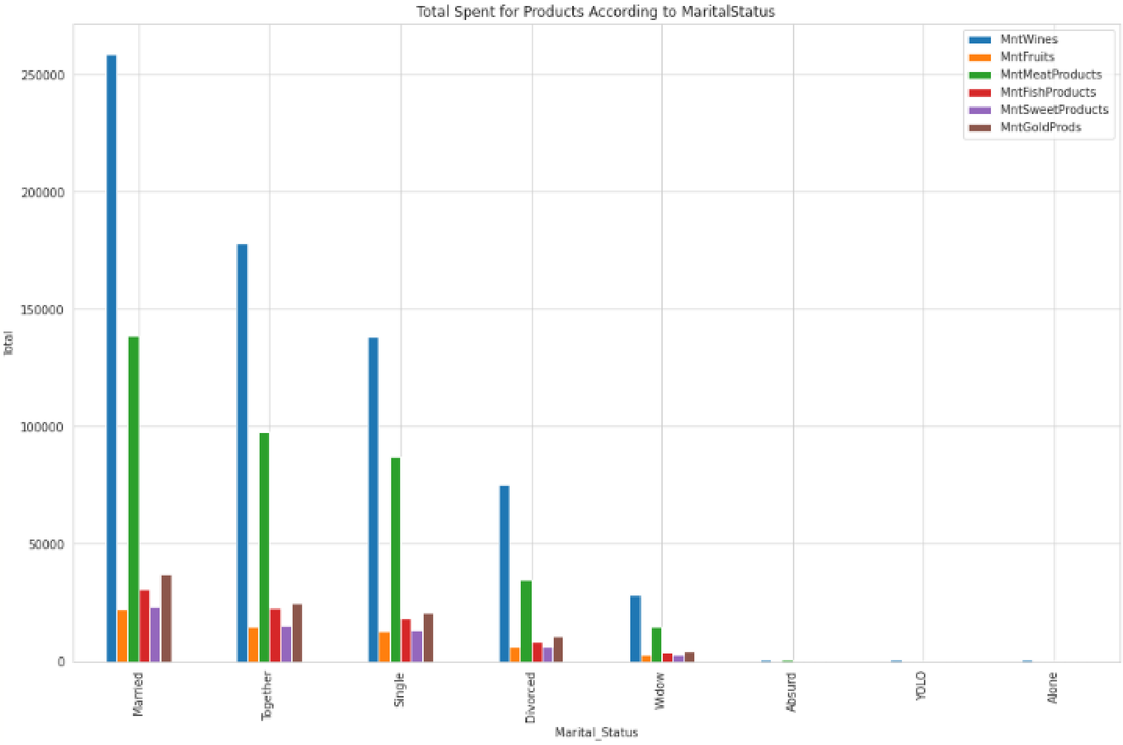
Count Plot of Kids Home feature



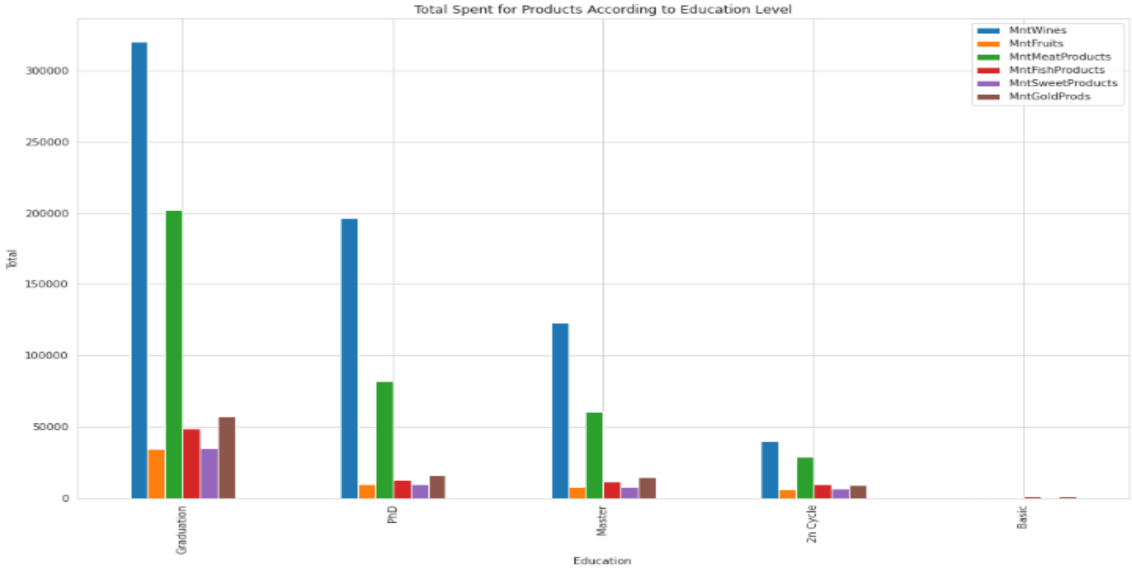
Count Plot of Work class feature



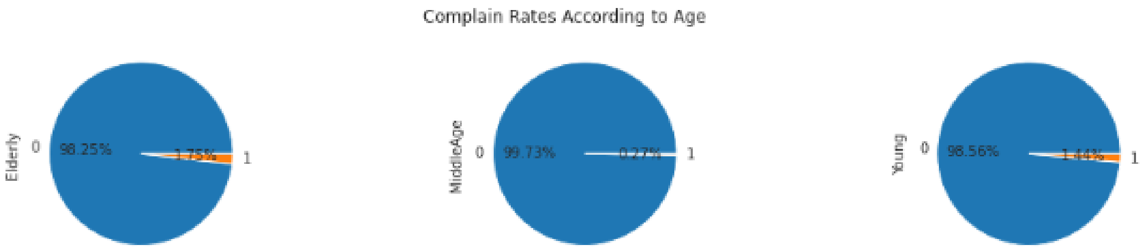
Amount Spent by Marital Status People



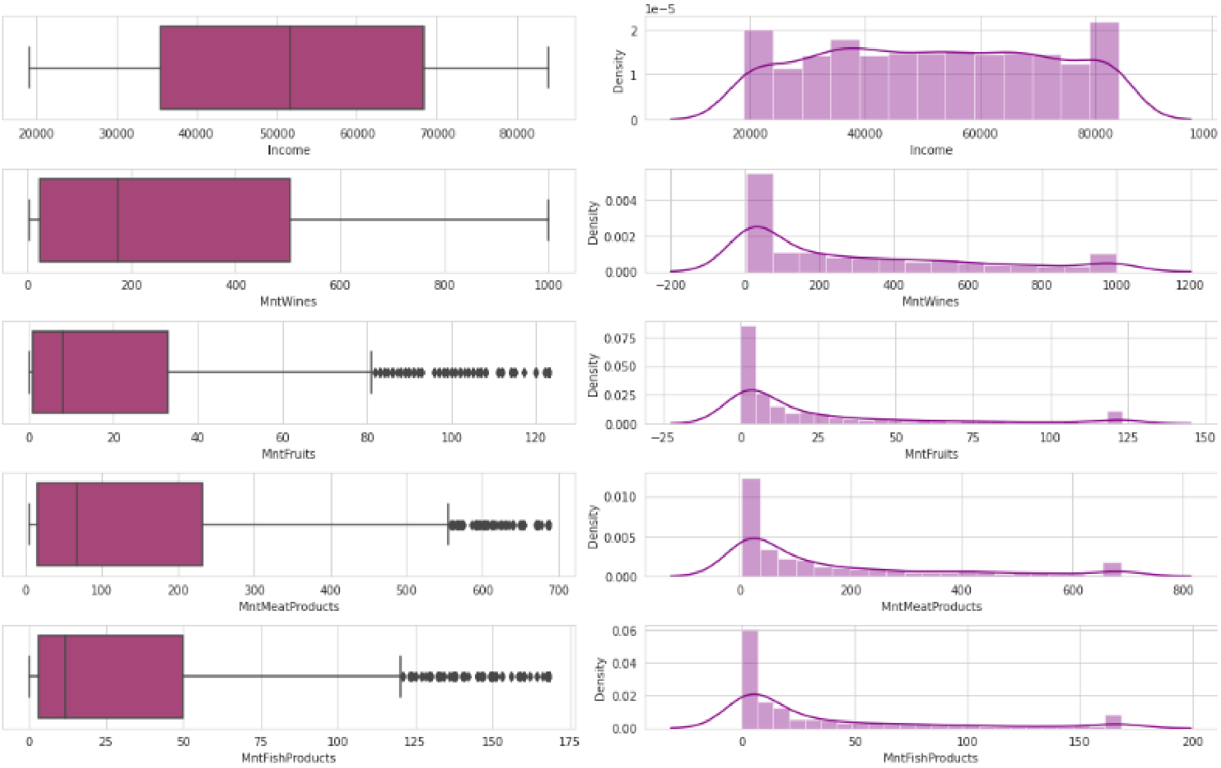
Amount Spent by Education Feature



Complain Rates According to Age

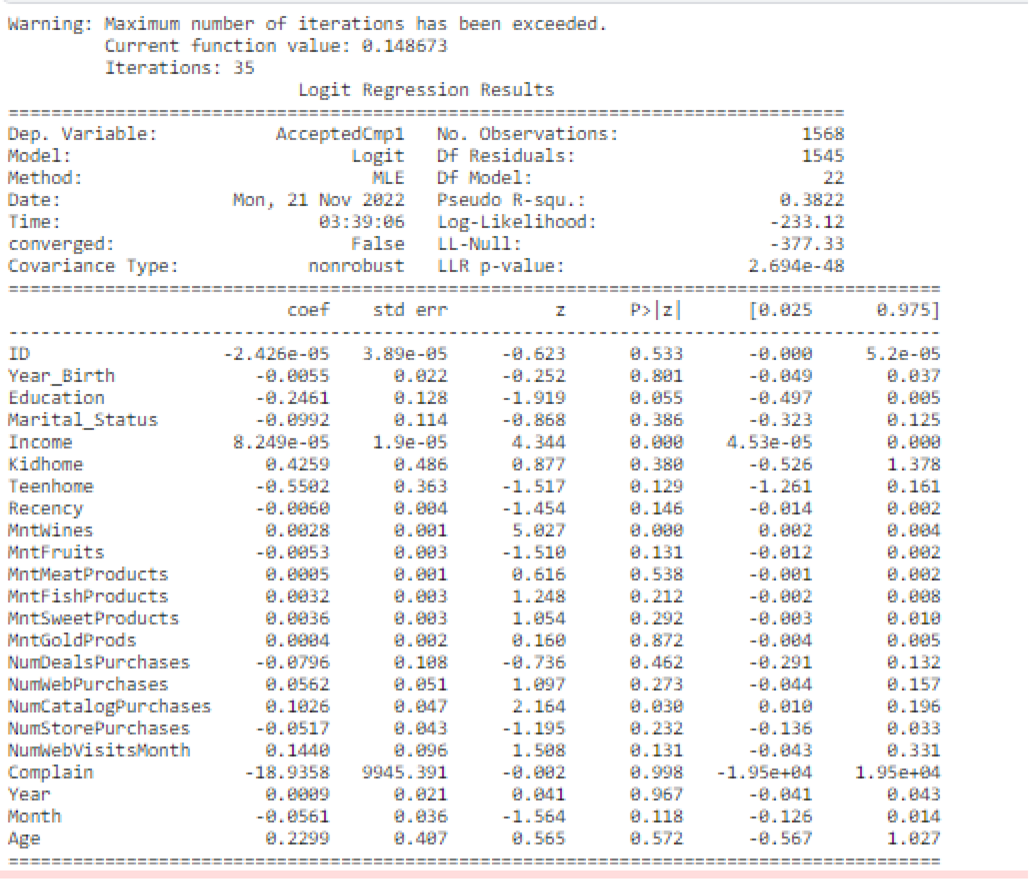


BoxPlot and DisPlot of numerical Features



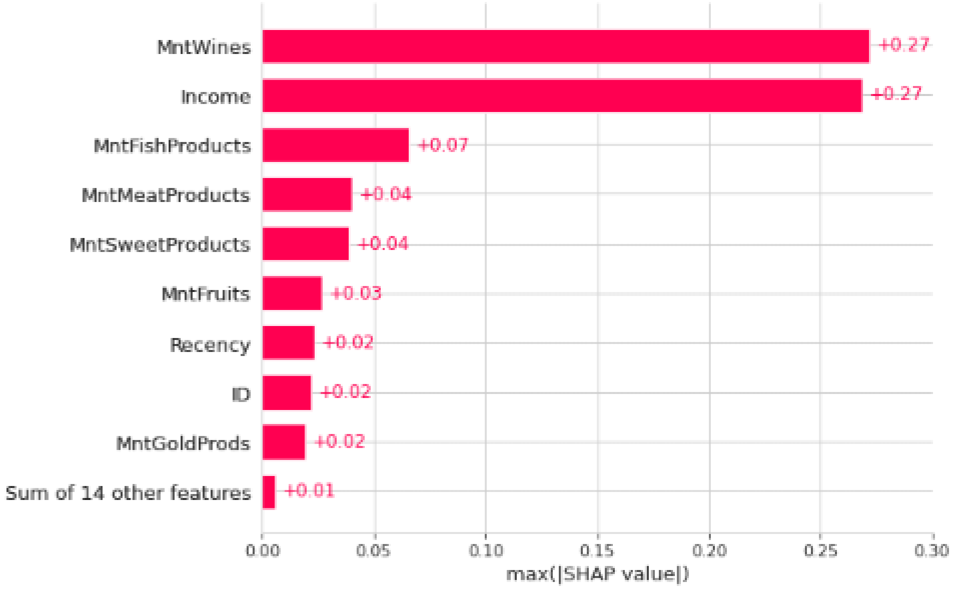
Appendix 2

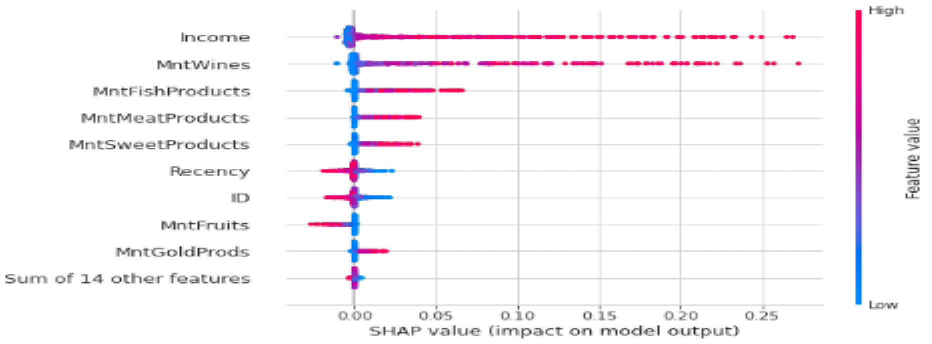
Logistic Regession Results

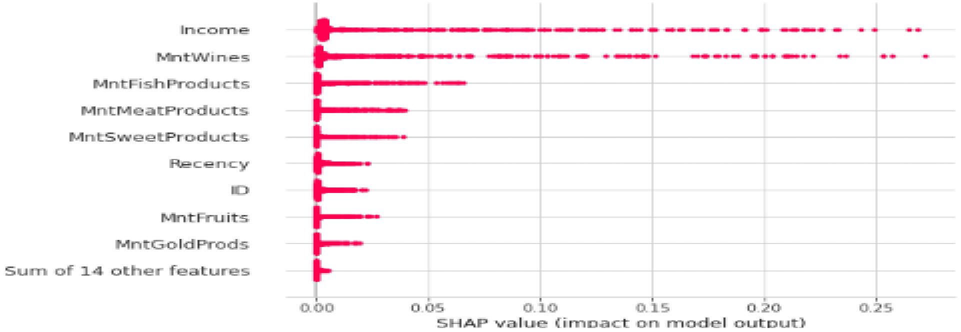


```
Warning: Maximum number of iterations has been exceeded.
         Current function value: 0.148673
         Iterations: 35
                           Logit Regression Results
==============================================================================
Dep. Variable:           AcceptedCmp1   No. Observations:                 1568
Model:                          Logit   Df Residuals:                     1545
Method:                           MLE   Df Model:                           22
Date:                Mon, 21 Nov 2022   Pseudo R-squ.:                  0.3822
Time:                        03:39:06   Log-Likelihood:                -233.12
converged:                      False   LL-Null:                       -377.33
Covariance Type:            nonrobust   LLR p-value:                 2.694e-48
======================================================================================
                          coef    std err          z      P>|z|      [0.025      0.975]
--------------------------------------------------------------------------------------
ID                  -2.426e-05   3.89e-05     -0.623      0.533      -0.000     5.2e-05
Year_Birth             -0.0055      0.022     -0.252      0.801      -0.049       0.037
Education              -0.2461      0.128     -1.919      0.055      -0.497       0.005
Marital_Status         -0.0992      0.114     -0.868      0.386      -0.323       0.125
Income               8.249e-05    1.9e-05      4.344      0.000    4.53e-05       0.000
Kidhome                 0.4259      0.486      0.877      0.380      -0.526       1.378
Teenhome               -0.5502      0.363     -1.517      0.129      -1.261       0.161
Recency                -0.0060      0.004     -1.454      0.146      -0.014       0.002
MntWines                0.0028      0.001      5.027      0.000       0.002       0.004
MntFruits              -0.0053      0.003     -1.510      0.131      -0.012       0.002
MntMeatProducts         0.0005      0.001      0.616      0.538      -0.001       0.002
MntFishProducts         0.0032      0.003      1.248      0.212      -0.002       0.008
MntSweetProducts        0.0036      0.003      1.054      0.292      -0.003       0.010
MntGoldProds            0.0004      0.002      0.160      0.872      -0.004       0.005
NumDealsPurchases      -0.0796      0.108     -0.736      0.462      -0.291       0.132
NumWebPurchases         0.0562      0.051      1.097      0.273      -0.044       0.157
NumCatalogPurchases     0.1026      0.047      2.164      0.030       0.010       0.196
NumStorePurchases      -0.0517      0.043     -1.195      0.232      -0.136       0.033
NumWebVisitsMonth       0.1440      0.096      1.508      0.131      -0.043       0.331
Complain              -18.9358   9945.391     -0.002      0.998   -1.95e+04    1.95e+04
Year                    0.0009      0.021      0.041      0.967      -0.041       0.043
Month                  -0.0561      0.036     -1.564      0.118      -0.126       0.014
Age                     0.2299      0.407      0.565      0.572      -0.567       1.027
======================================================================================
```

Feature Importance Using Shap For the models


MntWines
Income
MntFishProducts
MntMeatProducts
MntSweetProducts
MntFruits
Recency
ID
MntGoldProds
Sum of 14 other features
+0.27
+0.27
+0.07
+0.04
+0.04
+0.03
+0.02
+0.02
+0.02
+0.01
0.00
0.05
0.10
0.15
0.20
0.25
0.30
max(|SHAP value|)


Income
MntWines
MntFishProducts
MntMeatProducts
MntSweetProducts
Recency
ID
MntFruits
MntGoldProds
Sum of 14 other features
SHAP value (impact on model output)
High
Feature value
Low


Income
MntWines
MntFishProducts
MntMeatProducts
MntSweetProducts
Recency
ID
MntFruits
MntGoldProds
Sum of 14 other features
SHAP value (impact on model output)