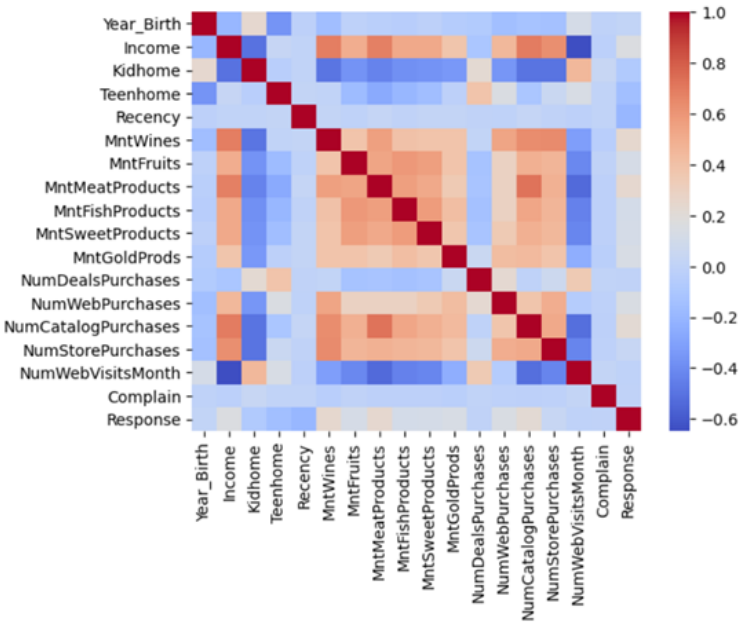
Introduction

The data set contains information about the customers of a magazine from the year 2012 to 2014. There are 2240 rows of information and 29 columns. The goal of this report is to understand why there is a decline in the number of subscriptions. I will be preparing a logistic regression model and a SVM model to accurately predict subscription behavior. Finally, I will compare both these models using overall accuracy, precision and recall statistics of these models. Using this analysis, I will further discuss on the two most significant variables and what the company should focus on next. I will be using various python packages for data cleaning and creating the prediction models such as pandas, numpy, sklearn, bamboolib, scipy, matplotlib, statsmodel.api etc.

Data cleaning

The first step in the analysis is to clean the data and prepare for a logistic regression model. There are 2 categorical variables education and marital status along with a date column Dt_customer. I have dropped all the columns starting with Accepted name as mentioned in the instruction for the project. Next, I checked for duplicates and if there is any dropped them, keeping only the first row. Now there are a few unwanted entries in marital status such as 'absurd', 'YOLO', 'Alone'. I have dropped the rows where marital status is listed as absurd and YOLO as this might be fake or invalid information. For the values where marital status is listed as alone, I updated it to single marital status. Next, I have removed the columns Z_CostContact and Z_Revenue as there are only constant values i.e., 3 and 11 values in all rows respectively for both columns so it makes no sense to keep these two columns in our logistic regression model. Next, I converted the income column from string to a numerical datatype. There is only one column 'Income' containing null values in the dataset. I checked for outlier also in the income column and there are a few outliers for this variable. I dopped the income 666666 value as it is not an expected income and will led to bias in the data. Next in order to replace the null values in this column I checked for distribution of the income data using a histogram. I can see that the data is normally distributed and there is no skewness, so I replaced the null income values with the mean income values.

On further exploring the data I noticed the outliers in the year birth column. There are 3 entries in the data i.e., 1893, 1894 and 1900 which seemed quite odd, as someone born in 1900 would be more than 120 years old by now. So, for the analysis in this report I have removed the values where year born is before 1940.



Next, I plotted a correlation matrix to check how interrelated the variables are to each other. The diagram shows the collinearity between the variables. We can see from the scale on right hand side on corelation matrix plot that if the variables are positively correlated to each other the color is red and if there is a negative correlation it is dispalyed by blue color.

The darker the color the more related the variables are to each other, and the higher will be collinearity.

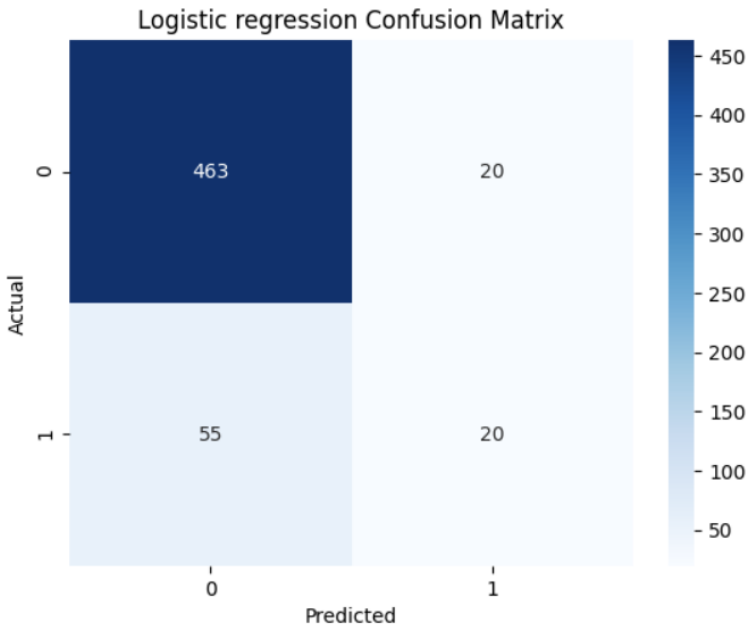
Now for the catagorical variables Education and Marital Status, I have used the dummy variables concept to convert them into binary format. Next I dropped the newly created columns Education_2n Cycle and Marital_Status_Together columns. Since these are leading to multicolinarity we've to drop them before creating the predictive model.

Logistic Regression Model

Logistic regression is a type a classification model used to predict whether an event is likely to happen or not. It is used to predict probability of an event occurring, given a set of input features. The prediction variables or the dependet variable for this model is reposnse and rest other variables will be treated as independent variables. Next step is to divide the datset into test and train data using the train_test_split funcationality of sklearn python package. I have divided the data set in to 75:25 ratio in train and test respectively. Now using the statsmodels.api package I will create the logistic regression model. The summary of the model is shown in appendix. The degree of freedom of the logistic regression model is 25. Considering a confidence interval of 95%, all the variables having p-values less than 0.05 are significant for the model i.e., Teenhome, Recency, MntWines, MntMeatProducts, MntGoldProds, NumCatalogPurchases, NumStorePurchases, NumWebVisitsMonth, Marital_Status_Divorced, Marital_Status_Single, Marital_Status_Widow. Marital status married has a p value of greater than 0.05 while rest all marital status have significant p-values meaning that married people are less likely to invest on magazine subscriptions.

The variables Year_Birth, teenhome, recency, MntFruits, MntFishProducts, NumStorePurchases, Education_Basic, Education_Graduation have a negative coefficient meaning that it has a negative impact on the the likeliness or the probability of subscribing the maganize. The accuracy of the model is 87%, precision is 50% and the recall is 27%.

To create the confusion matrix I have converted the predicted values to binary values using a threshold value of 0.5 where anything above 0.5 will be considered as 1 i.e., the customer response is yes to magazine subscription and anything less than 0.5 is considered 0 i.e, customer did not subscribe to the magazine.



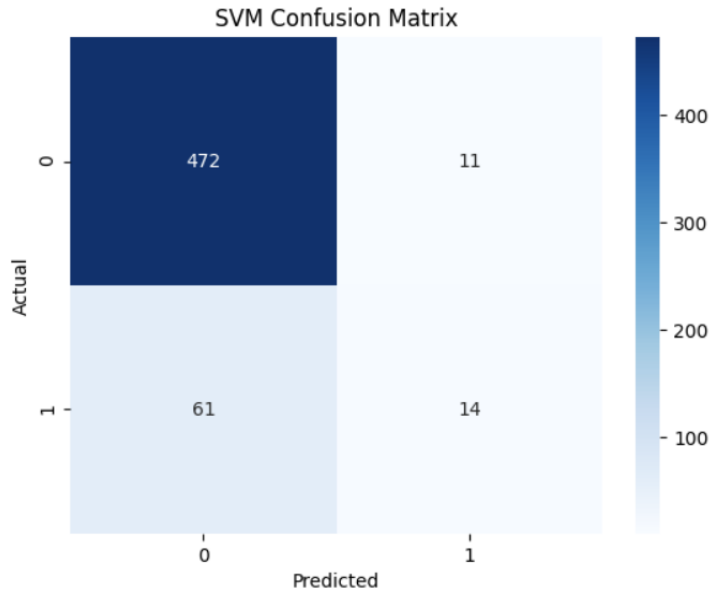
The below confusion matrix shows the overall performance of the logistic regression model. The accuracy, recall and precision of the model are calculated below using the confusion matrix. Accuracy is the measure of how often the model's predictions are correct, precision is the measure of how many of the predicted positive instances are actually positive and the recall is the measure of how many of the actual positive instances are predicted as positive.

- True Positive (When we are predicting Yes and, actual is also Yes): 20
- True Negative (When we are predicting No and, actual is also No): 463
- False Positive (When we are predicting Yes but in actual it is No): 20
- False Negative (When we are predicting No but in actual it is Yes): 55
- Accuracy of the Model =87%
- Precision= TP/(TP+FP) = 20/ (20+20) =0.5 ~ 50%
- Recall = TP/(TP+FN) = 20/ (20+55) =0.2666 ~ 27%
- Specificity= TN/ (TN + FP) = 463/ (463+20) = 0.958 ~ 96%
- Total Predictions: 558
- Correct Predictions: 483
- Incorrect Predictions: 75

Support Vector Machine Model

Support vector machine is a machine learning algorithm used for classification and regression analysis. For analysis in this report I will be using the same dependent and independent

variables and the same test and train data as per the logistic regression model. The line that optimizes the separation between the nearest data points of various classes is known as the hyperplane. The basic idea behind SVM is to find the hyperplane that best separates the data into different classes. The decision boundary of the model is defined by the support vectors, which are the nearest data points. For analysis in this report I have taken svm as linear. The model accuracy for the svm model is 87%, precision is 56% and the recall is 19%.



Below confusion matrix shows the performance of the svm model. The accuracy, recall and precision of the model are calculated below using the confusion matrix.

The accuracy of the model is almost same as that of logistic regression model.

- True Positive (When we are predicting Yes and, actual is also Yes): 14
- True Negative (When we are predicting No and, actual is also No): 472
- False Positive (When we are predicting Yes but in actual it is No): 11
- False Negative (When we are predicting No but in actual it is Yes): 61
- Accuracy of the Model =87%
- Precision= TP/(TP+FP) = 14/ (14+11) =0.56 ~ 56%
- Recall = TP/(TP+FN) = 14/ (14+61) =0.18666 ~ 19%
- Specificity= TN/ (TN + FP) = 472/ (472+11) = 0.977 ~ 98%
- Total Predictions: 558
- Correct Predictions: 486
- Incorrect Predictions: 72

As per the above results we can interpret that the model predicted correctly 472 times that the customer is not subscribing the magazine and 14 times correctly predicitng that the customer will subscribe to it. Moreover 61 times we predicted that customer will not subscribe to the magazine but anyways the customer ended up subscribing it.

Model Comparison

Both the logistic regression model and the SVM model have almost same accuracy at 87% but the number of correct predictions count is slightly higher in came of SVM i.e., 486 in comparison to 483 in logistic regression model. The number of false predictions should be as low as low as possible for the best possible model. The precision of SVM model is slightly better than logistic model, the SVM has 56% precision while the logistic has 50% precision.

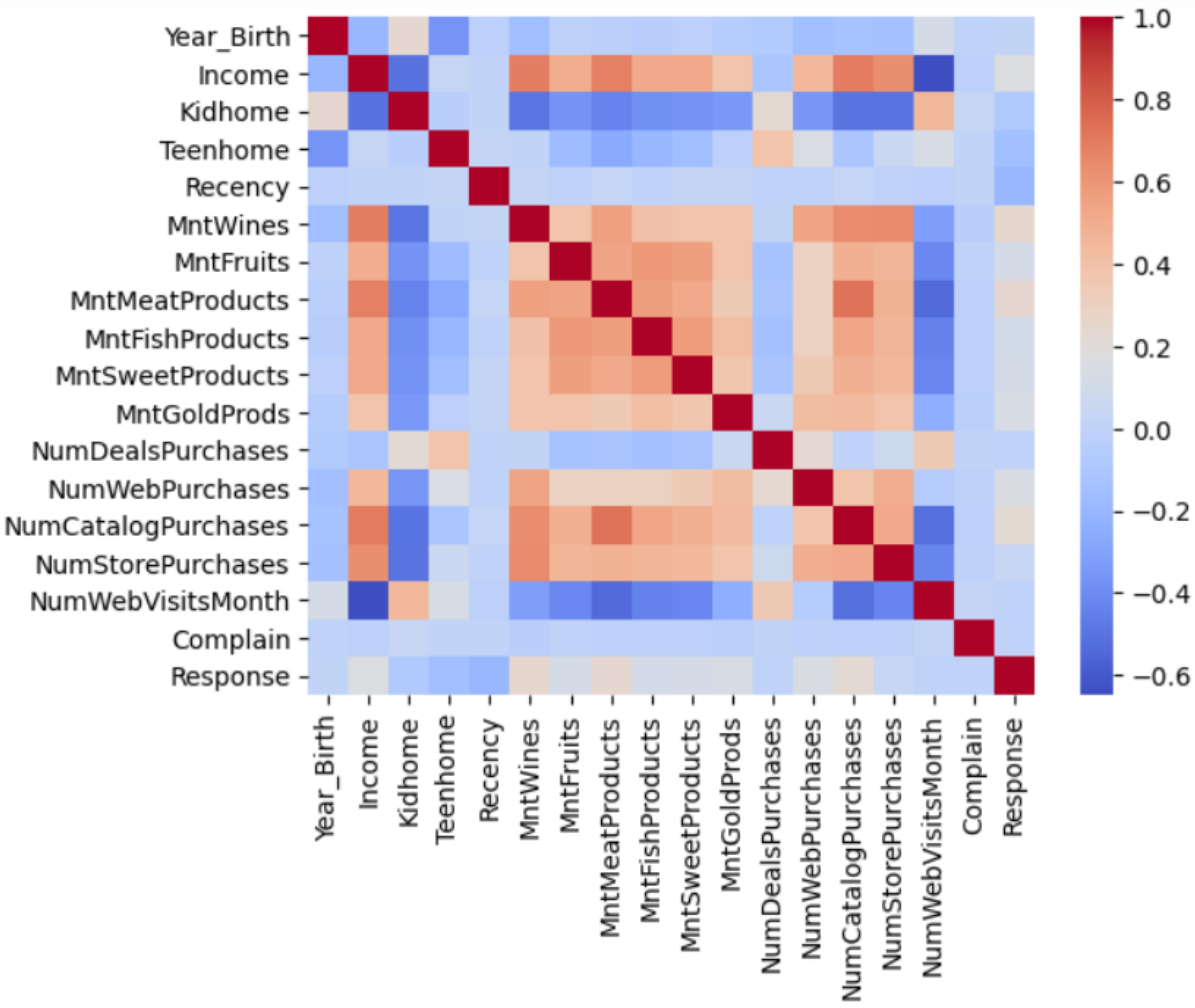
Since SVM accounts for more resources and run time as compared to logistic regression model if the company can efford all these without any expense than I would recommend to use the SVM model as this has better accuracy and precision as compared to logistic model. This model will surely help the business as the the false positve count is very less as comapred to logistic model. In this scenario we would like to have a minimun of false postive values as they are the customer that we predicted will be subscribing to the magazine but they ended up not doing that. In terms of the significance of the variabels I would say the two most significant variables are the education and the marital status. As the education level increases from basic to PHD the coefficient of variabels changes from negative to postive indicating that it is likely to increase the probability of a customer subscribing to the magazine. Similarly the coefficient of variable married couples is least and for singles, divorced and window the coeficient value is higher again indicating that they are more likely to subscribe to the magazine. Considering all these factors I think the company should focus on campaigns towards educated customers and singles, widow and divorced individuals

Appendix

2,236 rows and 20 columns

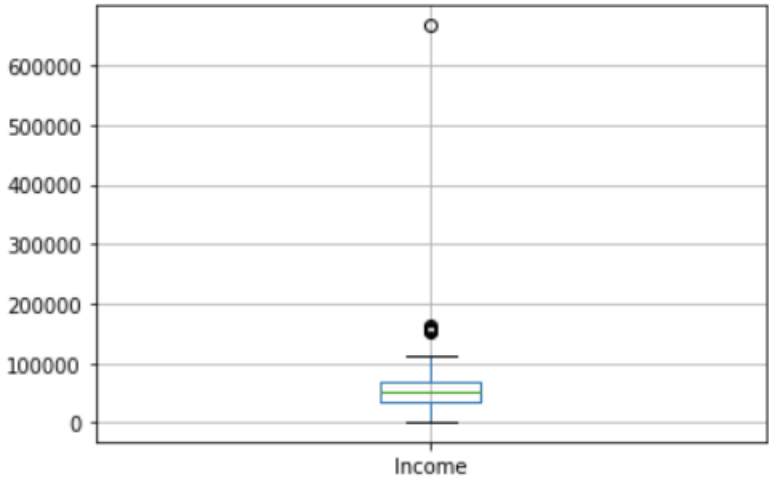
Show [All columns] Update

Column	Data type	Unique values	Missing values
Year_Birth	int64	59	0
Education	object	5	0
Marital_Status	object	5	0
Income	int64	1,973	24 - 1.1%
Kidhome	int64	3	0
Teenhome	int64	3	0
Recency	int64	100	0
MntWines	int64	776	0
MntFruits	int64	158	0
MntMeatProducts	int64	558	0
MntFishProducts	int64	182	0
MntSweetProducts	int64	177	0
MntGoldProds	int64	212	0
NumDealsPurchases	int64	15	0
NumWebPurchases	int64	15	0
NumCatalogPurchases	int64	14	0
NumStorePurchases	int64	14	0
NumWebVisitsMonth	int64	16	0
Complain	int64	2	0
Response	int64	2	0



	precision	recall	f1-score	support
0	0.89	0.96	0.93	483
1	0.50	0.27	0.35	75
accuracy			0.87	558
macro avg	0.70	0.61	0.64	558
weighted avg	0.84	0.87	0.85	558

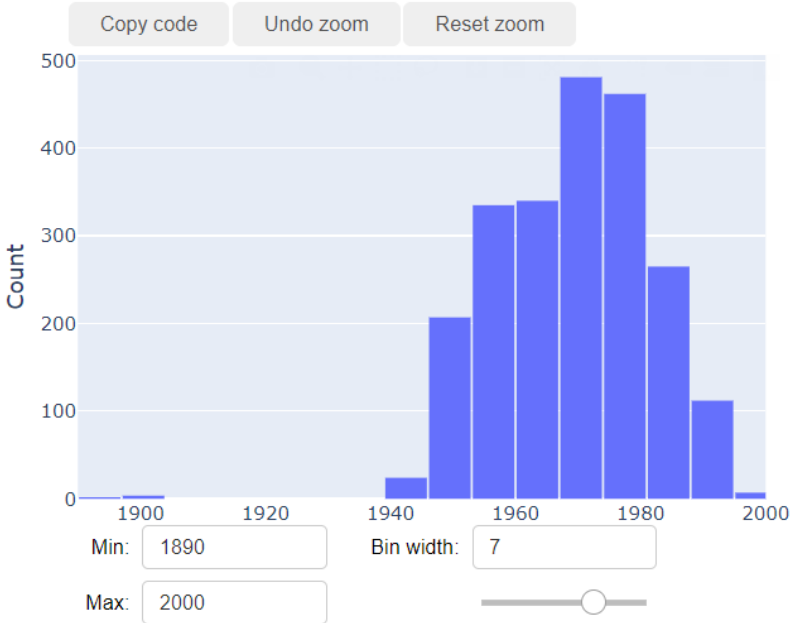
	precision	recall	f1-score	support
0	0.87	0.96	0.91	468
1	0.57	0.26	0.35	90
accuracy			0.85	558
macro avg	0.72	0.61	0.63	558
weighted avg	0.82	0.85	0.82	558



"Year_Birth"

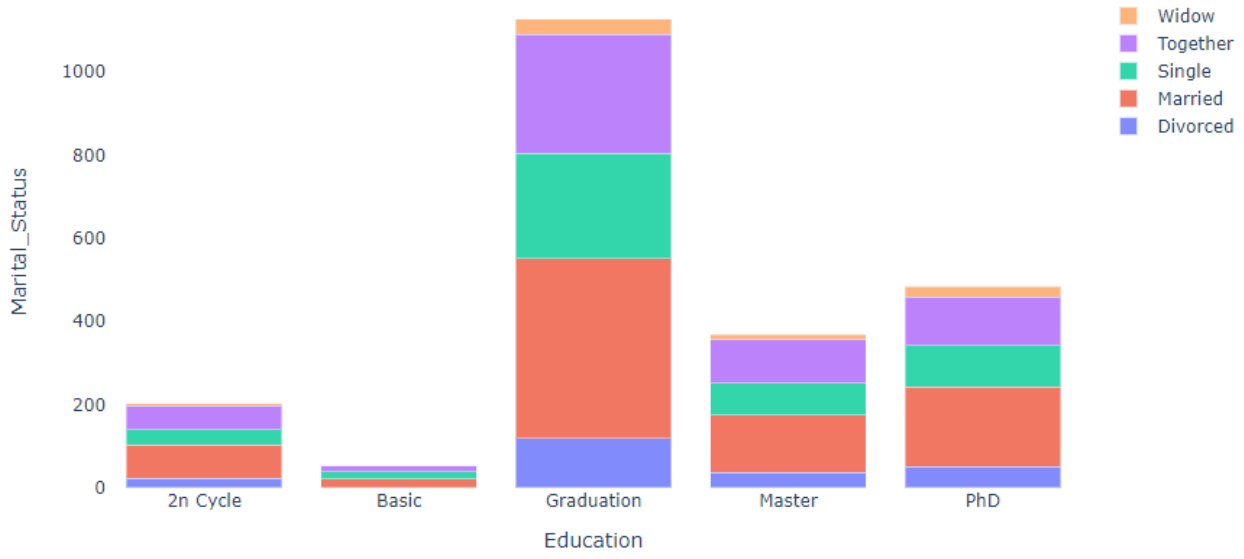
Pandas dtype	Valid	Uniques	Missing values
int64	2236	59	0

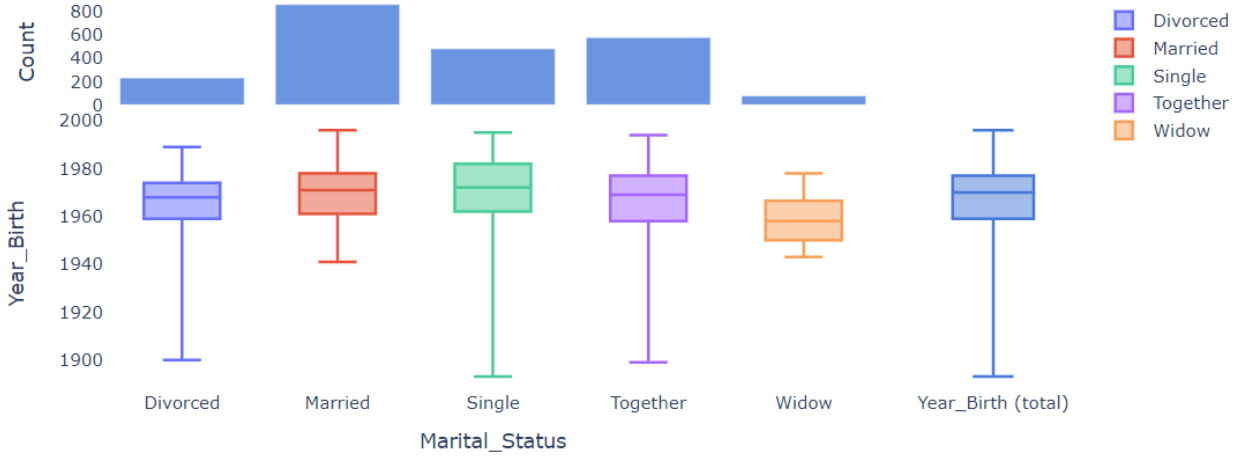
Histogram



Statistics	
Min	1893.000000
Q1	1959.000000
Median	1970.000000
Q3	1977.000000
Max	1996.000000
Mean	1968.796512
StdDev	11.980604

Value counts	
Positive values	2236
Zero values	0
Negative values	0


Widow
Together
Single
Married
Divorced
Marital_Status
1000
800
600
400
200
0
2n Cycle
Basic
Graduation
Master
PhD
Education

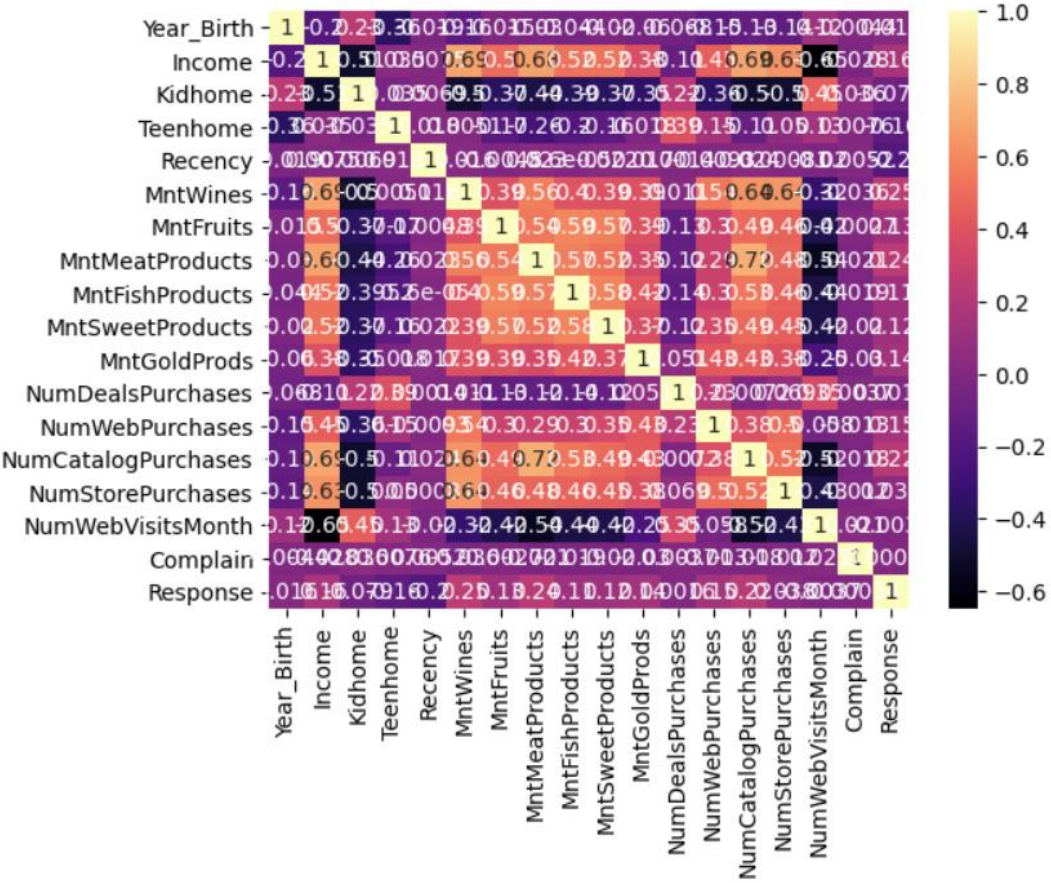

Divorced
Married
Single
Together
Widow
Count
800
600
400
200
0
Year_Birth
2000
1980
1960
1940
1920
1900
Divorced
Married
Single
Together
Widow
Year_Birth (total)
Marital_Status

Column	Data type	Unique values	Missing values
Year_Birth	int32	56	0
Income	int32	1,970	0
Kidhome	int32	3	0
Teenhome	int32	3	0
Recency	int32	100	0
MntWines	int32	775	0
MntFruits	int32	158	0
MntMeatProducts	int32	557	0
MntFishProducts	int32	182	0
MntSweetProducts	int32	177	0
MntGoldProds	int32	212	0
NumDealsPurchases	int32	15	0
NumWebPurchases	int32	15	0
NumCatalogPurchases	int32	14	0
NumStorePurchases	int32	14	0
NumWebVisitsMonth	int32	16	0
Complain	int32	2	0
Response	int32	2	0
Education_Basic	int32	2	0
Education_Graduation	int32	2	0
Education_Master	int32	2	0
Education_PhD	int32	2	0
Marital_Status_Divorced	int32	2	0
Marital_Status_Married	int32	2	0
Marital_Status_Single	int32	2	0
Marital_Status_Widow	int32	2	0

Logit Regression Results

Dep. Variable:	Response	No. Observations:	1674
Model:	Logit	Df Residuals:	1648
Method:	MLE	Df Model:	25
Date:	Sun, 12 Mar 2023	Pseudo R-squ.:	0.2710
Time:	17:23:35	Log-Likelihood:	-523.29
converged:	True	LL-Null:	-717.77
Covariance Type:	nonrobust	LLR p-value:	5.620e-67

	coef	std err	z	P>|z|	[0.025	0.975]
const	-1.1490	14.417	-0.080	0.936	-29.406	27.108
Year_Birth	-0.0012	0.007	-0.166	0.868	-0.015	0.013
Income	4.889e-06	8.31e-06	0.589	0.556	-1.14e-05	2.12e-05
Kidhome	0.2345	0.222	1.056	0.291	-0.201	0.670
Teenhome	-1.1212	0.211	-5.324	0.000	-1.534	-0.708
Recency	-0.0267	0.003	-8.826	0.000	-0.033	-0.021
MntWines	0.0019	0.000	5.424	0.000	0.001	0.003
MntFruits	-0.0011	0.002	-0.460	0.646	-0.006	0.004
MntMeatProducts	0.0015	0.001	2.952	0.003	0.001	0.003
MntFishProducts	-0.0002	0.002	-0.109	0.913	-0.004	0.003
MntSweetProducts	0.0041	0.002	1.817	0.069	-0.000	0.009
MntGoldProds	0.0043	0.002	2.723	0.006	0.001	0.007
NumDealsPurchases	0.0717	0.050	1.440	0.150	-0.026	0.169
NumWebPurchases	0.0325	0.034	0.964	0.335	-0.034	0.099
NumCatalogPurchases	0.1177	0.042	2.776	0.005	0.035	0.201
NumStorePurchases	-0.1639	0.035	-4.731	0.000	-0.232	-0.096
NumWebVisitsMonth	0.2453	0.051	4.821	0.000	0.146	0.345
Complain	0.4273	0.900	0.475	0.635	-1.336	2.191
Education_Basic	-222.9361	1.17e+48	-1.91e-46	1.000	-2.29e+48	2.29e+48
Education_Graduation	-0.0362	0.298	-0.121	0.903	-0.620	0.548
Education_Master	0.1967	0.339	0.580	0.562	-0.468	0.862
Education_PhD	0.5817	0.326	1.783	0.075	-0.058	1.221
Marital_Status_Divorced	1.1352	0.288	3.945	0.000	0.571	1.699
Marital_Status_Married	0.2292	0.228	1.003	0.316	-0.219	0.677
Marital_Status_Single	1.1319	0.237	4.782	0.000	0.668	1.596
Marital_Status_Widow	1.4281	0.401	3.565	0.000	0.643	2.213


Year_Birth
Income
Kidhome
Teenhome
Recency
MntWines
MntFruits
MntMeatProducts
MntFishProducts
MntSweetProducts
MntGoldProds
NumDealsPurchases
NumWebPurchases
NumCatalogPurchases
NumStorePurchases
NumWebVisitsMonth
Complain
Response
1.0
0.8
0.6
0.4
0.2
0.0
−0.2
−0.4
−0.6