Module 3 Assignment: Understanding Magazine subscription behaviour

Omkar Sadekar

College of Professional Studies, Northeastern University

Professor: Justin Grosz

Jan 29, 2023

Introduction

The aim of this project is to come up with strong recommendations and actionable insights to contribute towards the problem of declining Magazine Subscription sales while understanding Magazine Subscription Behavior of Customers and how factors like buying habits, products purchased, amount spent on purchasing general items and income etc. influences the decision of the customers to take up/continue or drop the subscription. The Dataset contains total 29 variables and 2240 observations. Eventually, predicting the outcome while building and comparing Logistic Regression and SVM model for better performance. There are two major categorical variables 'Marriage' and 'Education' and others are numerical variables like Income and amount spent on groceries and other items and binary variables including kids and teenagers at home and 'Complain'. The target variable is 'Response', It tells us about the decision of dropping or continuing the subscription.

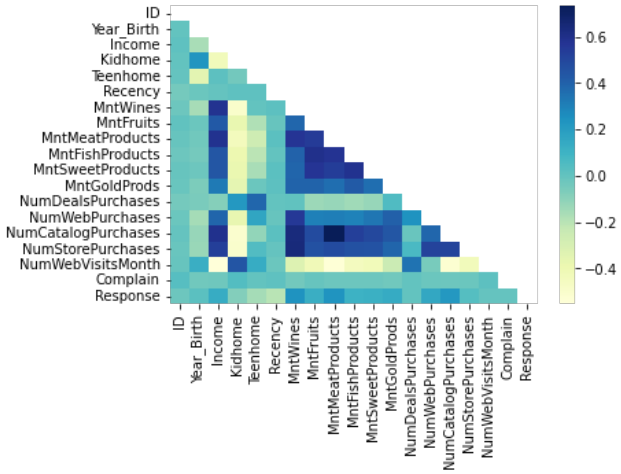
Data cleaning

We start by dropping the variables like date and accepted campaign which are irrelevant to our study while Checking the Data types and null values to avoid discrepancies in our results. Dropping the total 24 null values in the 'Income' as it contributes towards just 1% of the entire Dataset. The Categorical variable 'Marriage' was converted to binary variable, which had subcategories like single, alone, YOLO, divorced and together which essentially means "Single" or "Married" for better analysis.

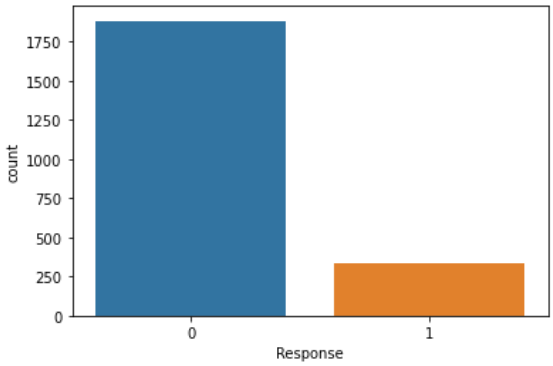
EDA

In this section, exploration of what variables may drive the default will be stated.

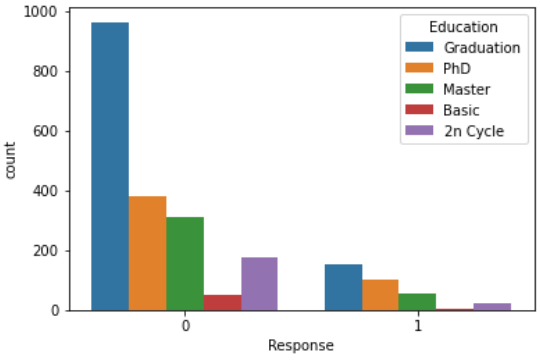
We start by understanding our data better using a correlation plot. The Numerical variables are highly correlated with each other. As, high correlation between variables can cause multicollinearity. The model's coefficients may be challenging to interpret and may become unstable due to multicollinearity. From the plot we get to know that the response variable is correlated with amount spend on wines, amount spend on meat products and number of catalog purchases, number of web purchases. Our Response variable is unbalanced as seen in the graph, the number of customers not subscribing are way more than number of customers subscribing, that is reason behind not using techniques like PCA and or LDA for dimensionality reduction as we might lose our data. Further, we can see that a lot of customers who are graduated have dropped the subscription decision. From the count plot of 'Marriage' and 'Response' we can say that a lot of married customers are dropping the subscriptions compared to customers who are single



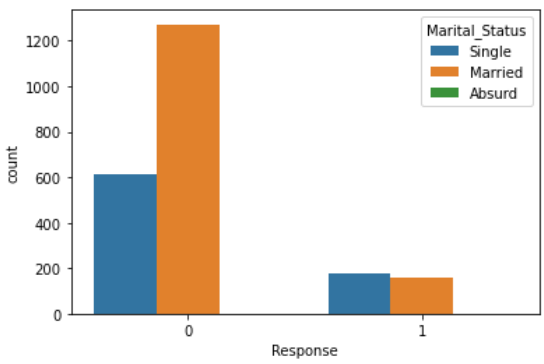
Plot1. Correlation matrix



Plot2. Count plot of Target Variable



Plot3. Count plot of Target Variable and 'Education'



Plot4. Count plot of Target Variable and Marital_status

Analysis

Modeling

After completing the preliminary analysis of the data set, this report will complete further analysis by establishing models based on the preliminary conclusions drawn in the EDA. In this section, the logistic regression model and Support Vector Machine(SVM) model will be used and after the model is established, the model will be optimized so that it can better determine the customer behavioral patterns in Magazine Subscriptions.

Logistic regression model

The logistic regression model is the first. The logistic regression model has two advantages in this case: First, the dependent variable 'Response' in the problem being studied is a binary classification problem, and logistic regression is very good at dealing with binary classification problems. Second, the model's interpretability is excellent. We can see the influence of different features on the result by looking at the weight of features. Because there are very few observations of customers with subscriptions, we divide the data set into a training data set and a testing data set in a 3:2 split rather than a 4:1 split to avoid overfitting problems. Dummy variables are created from categorical variables. Finally, a logistic regression model for prediction is established. We can draw some conclusions by interpreting the P value and coefficient of each variable in Appendix 2. Among the numerical variables, variables "Receny", "Teenhome", "MntWines", "MntMeatProducts", "MntGoldProds", "NumDealsPurchase", "NumWebPurchase", "NumCatalogPurchase", and

"NumStorePurchase" have significant effects on the variable "Response", while other coefficients have a positive correlation. A low p-value (usually less than 0.05) indicates that there is sufficient evidence to reject the null hypothesis and support the alternative hypothesis that the coefficients are not equal to zero. Income, NumDealsPurchased P-values are greater than our level of significance. However, because we know that multicollinearity reduces the statistical power of the regression model and our data is highly corelated due to multiple factors influencing the factors involved in predicting the magazine subscription, we cannot dismiss these variables as insignificant. Class 0 and class 1 precision are 89% and 65%, respectively. It measures how well the model predicts successful outcomes. A high precision indicates that the model makes few false positive predictions. Class 0 and class 1 recall rates are 97% and 31%, respectively. It measures the model's ability to detect every instance of positivity in the data. According to a high recall rate, the model does not lack many True positive examples. The model's overall accuracy is 87%, and the F-1 score, which is the weighted average of precision, recall, and accuracy, is 93% for class 0 and 42% for class 1.

	P	N
T	732	22
F	92	41

Fig 1. Confusion matrix for logistic regression

SVM Model

SVMs are a popular supervised learning algorithm for classification and regression problems. SVMs are particularly useful when there are many features in the dataset or when the features and the target variable have a non-linear relationship. SVMs can outperform traditional algorithms like logistic regression in such cases. We use the SVM model to analyze the results and see if they are better than the logistic regression model because we have about 19 features. We obtain the following results after implementing the model: Class 0 and class 1 precision are 87% and 49%, respectively. It measures how well the model predicts successful outcomes. A high precision indicates that the model makes few false positive predictions. Class 0 and class 1 recall rates are 97% and 18%, respectively. It measures the model's ability to detect every instance of positivity in the data. According to a high recall rate, the model does not lack many genuine positive examples. The model's overall accuracy is 85%, and the F-1 score, which is the weighted average of precision, recall, and accuracy, is 92% for class 0 and 26% for class 1.

	P	N
T	729	25
F	109	24

Fig 2. Confusion matrix for svm model

Finally, we can compare the F1 score which is the weighted average of precision and recall and accuracy in the table below

Table 1. Table for metrics

Model	Accuracy	F-1 Score (0)	F-1 Score(1)

Logistic Regression	0.87	0.93	0.42
SVM	0.85	0.92	0.26

Conclusion

Based on the above analysis, we can conclude that it is difficult to predict whether a customer will subscribe or not based on these features because the accuracy of both models is slightly lower. A high F1 score indicates that the model achieves an acceptable level of precision and recall. That is, the model makes a high proportion of true positive predictions while making a low proportion of false positive predictions. In class 0, both models have a high F-1 score. As a result, the model can correctly predict it. F-1 scores for class "1" are extremely low due to unbalanced data. However, based on the preceding series of analyses, we can still make a reasonable guess as to whether a person is likely to subscribe. First, based on the model's conclusion, we can determine why the customer is unwilling to subscribe based on significant variables listed in the logistic regression model. Customers' spending habits on items such as wine, meat, and gold, as well as the number of purchases made and the frequency of website visits, are extremely important for magazine subscriptions. It is difficult to choose between the two models because each has advantages and disadvantages, but based on the metrics, logistic regression outperformed the SVM model.

Recommendation

From this report we can give the company a set of procedures for understanding magazine subscription behavior and finding key areas to focus on to increase the sales.

- Analyzing customer web visit data: Track and analyze each customer's web visits to determine the effectiveness of the company's web marketing initiatives. This data can help the company improve its online presence and reach a larger audience.
- Targeting high-spending customers by analyzing their spending habits on items such as wine, meat, and gold. Customers who buy more in these categories should be targeted, and their sales should be optimized by offering them relevant products and discounts.
- Personalizing marketing efforts: By tracking customer purchase history, the company can personalize marketing efforts and target customers with offers and promotions that they are most likely to find appealing.
- Concentrating on customer retention: Retain current customers by providing loyalty programs, special promotions, and excellent customer service.
- Using data-driven insights: Use the report's data to inform decisions and drive the company's marketing and sales strategies. Monitor and analyze customer behavior on an ongoing basis to identify trends and make informed decisions.

Reference

1. scikit-learn library's documentation: https://scikit-learn.org/stable/modules/generated/sklearn.linear_model.LogisticRegression.html
2. A comprehensive tutorial: https://www.datacamp.com/community/tutorials/svm-classification-scikit-learn-python

Appendix 1

"Logistic Regression Results"

Logit Regression Results

Dep. Variable:	Response	No. Observations:	1329
Model:	Logit	Df Residuals:	1305
Method:	MLE	Df Model:	23
Date:	Sun, 29 Jan 2023	Pseudo R-squ.:	0.2717
Time:	22:11:03	Log-Likelihood:	-409.98
converged:	True	LL-Null:	-562.91
Covariance Type:	nonrobust	LLR p-value:	2.998e-51

	coef	std err	z	P>|z|	[0.025	0.975]
const	2.7636	nan	nan	nan	nan	nan
ID	-3.87e-05	2.78e-05	-1.392	0.164	-9.32e-05	1.58e-05
Year_Birth	-0.0034	0.008	-0.415	0.678	-0.019	0.013
Income	8.088e-07	3.99e-06	0.203	0.839	-7.01e-06	8.62e-06
Kidhome	0.1962	0.259	0.758	0.448	-0.311	0.703
Teenhome	-1.3094	0.244	-5.365	0.000	-1.788	-0.831
Recency	-0.0273	0.003	-7.880	0.000	-0.034	-0.021
MntWines	0.0017	0.000	4.574	0.000	0.001	0.002
MntFruits	0.0021	0.003	0.770	0.441	-0.003	0.007
MntMeatProducts	0.0019	0.001	3.281	0.001	0.001	0.003
MntFishProducts	-0.0029	0.002	-1.418	0.156	-0.007	0.001
MntSweetProducts	0.0041	0.003	1.543	0.123	-0.001	0.009
MntGoldProds	0.0055	0.002	2.892	0.004	0.002	0.009
NumDealsPurchases	0.0841	0.057	1.479	0.139	-0.027	0.196
NumWebPurchases	0.1053	0.043	2.473	0.013	0.022	0.189
NumCatalogPurchases	0.0922	0.045	2.031	0.042	0.003	0.181
NumStorePurchases	-0.2007	0.042	-4.784	0.000	-0.283	-0.118
NumWebVisitsMonth	0.2165	0.055	3.927	0.000	0.108	0.325
Complain	1.0668	0.904	1.180	0.238	-0.706	2.839
Education_Basic	-1.0886	1.084	-1.004	0.315	-3.213	1.036
Education_Graduation	0.0956	0.347	0.276	0.783	-0.584	0.775
Education_Master	0.5053	0.397	1.272	0.203	-0.273	1.284
Education_PhD	0.7580	0.382	1.982	0.048	0.008	1.508
Marital_Status_Married	0.8494	nan	nan	nan	nan	nan
Marital_Status_Single	1.9142	nan	nan	nan	nan	nan

```
[[732  22]
 [ 92  41]]
```

"Classification Report for Logistic Regression"

	precision	recall	f1-score	support
0	0.89	0.97	0.93	754
1	0.65	0.31	0.42	133
accuracy			0.87	887
macro avg	0.77	0.64	0.67	887
weighted avg	0.85	0.87	0.85	887

"SVM Results- Classification report"

```
[[729  25]
 [109  24]]
              precision    recall  f1-score   support

           0       0.87      0.97      0.92       754
           1       0.49      0.18      0.26       133

    accuracy                           0.85       887
   macro avg       0.68      0.57      0.59       887
weighted avg       0.81      0.85      0.82       887
```