**Module 3 Project: Understanding Magazine subscription behaviour**

**Prepared by: Sarvesh Thorve**

**Email: throve.s@northeastern.edu**

**Professor : Justin Grosz**

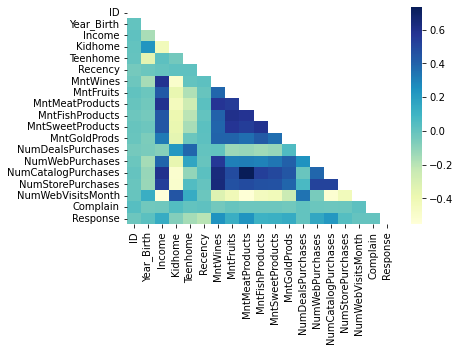
# Introduction

This project seeks to understand magazine subscription behavior of customers and how factors like purchasing habits, products purchased, amount spent on general purchases, income, etc. influence the decision of the customers to take up/continue or drop the subscription in order to provide strong recommendations and practical insights to help address the issue of declining magazine subscription sales. There are 2240 observations and 29 total variables in the dataset. Finally, making predictions while developing and contrasting SVM and Logistic Regression models for improved performance. The two main categorical variables are "Marriage" and "Education," along with other numerical factors like "Income," "Amount Spent on Groceries," and "Complain." Binary variables include "Kids and Teenagers at Home," "Complain," and "Complain" are other significant categorical variables. The target variable is ‘Response’, It tells us about the decision of dropping or continuing the subscription.

# Data cleaning

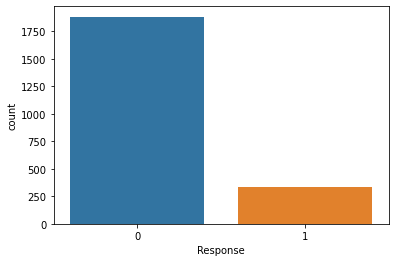
# To eliminate conflicts in our findings, we begin by eliminating factors like date and accepted campaign that are unrelated to our study while also checking the data types and null values. removing the 24 total null values from the "Income" field because they only make up 1% of the overall dataset. The categorical variable "Marriage" was changed to a binary variable for easier analysis. This binary variable had subcategories like "Single" or "Married" as well as subcategories for being alone, YOLO, divorced, and together.

# EDA

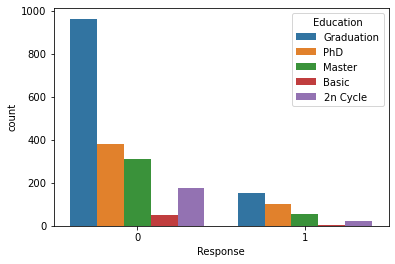
In this section, exploration of what variables may drive the default will be stated.

We begin by using a correlation plot to better comprehend our data. The numerical variables have a strong correlation. Multicollinearity can be brought on by high levels of correlation between variables. The multicollinearity of the model's coefficients may make them difficult to comprehend and liable to instability.

The response variable is connected with the amount spent on wines, the amount spent on meat products, the number of catalog sales, and the number of online transactions, according to the plot.



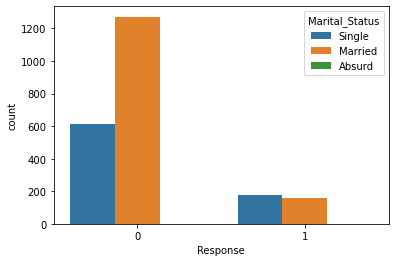
***Plot2. Count plot of Target Variable***



***Plot3. Count plot of Target Variable and ‘Education’***

We decided against utilizing algorithms like PCA and or LDA for dimensionality reduction because we might lose our data. As can be seen in the graph, our Response variable is unbalanced; the number of consumers who aren't subscribing is significantly higher than the number of customers who are.

In addition, we observe that many consumers who have graduated have changed their minds about subscribing.



***Plot4. Count plot of Target Variable and Marital\_status***

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

# Analysis

**Modeling**

# This report will finish further analysis after finishing the initial analysis of the data set by creating models based on the initial deductions made in the EDA. The logistic regression model and the Support Vector Machine (SVM) model will be employed in this section, and once the model has been developed, it will be optimized to help identify the customer behavioral patterns in magazine subscriptions.

# Logistic regression model

The first is the logistic regression model. In this situation, the logistic regression model offers two benefits: First off, logistic regression is excellent at handling binary classification problems because the dependent variable in the topic under study, "Response," is a problem of binary classification. Second, the model has great interpretability. By examining the weight of the features, we can understand how different features have an impact on the outcome. We split 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 prevent overfitting issues because there are very few observations of customers with subscriptions. Categorical variables are used to produce dummy variables. Ultimately, a prediction model using logistic regression is created. By analyzing the P value and coefficient for each variable in Appendix 2, we can come to certain conclusions. Aside from "Receny," "Teenhome," "MntWines," "MntMeatProducts," "MntGoldProds," "NumDealsPurchase," "NumWebPurchase," "NumCatalogPurchase," and other numbers

The variable "Response" is significantly impacted by "NumStorePurchase," while other coefficients have a favorable association. A low p-value (often less than 0.05) suggests that there is enough data to support the alternative hypothesis that the coefficients are not equal to zero and to reject the null hypothesis. Our standard of significance is exceeded by the P-values for income and number of deals purchased. We cannot, however, ignore these variables as being insignificant because we are aware that multicollinearity lowers the statistical power of the regression model and our data is highly correlated as a result of multiple factors influencing the factors involved in predicting the magazine subscription. Class 0 precision is 89%, and class 1 precision is 65%.

|  |  |  |
| --- | --- | --- |
|  | P | N |
| T | 732 | 22 |
| F | 92 | 41 |

***Fig 1. Confusion matrix for logistic regression***

It gauges how well the model foresees successful results. A high precision means that the model rarely predicts something that will actually happen. The recall rates for classes 0 and 1 are 97% and 31%, respectively. It gauges how well the model can find each instance of positive in the data. A high recall rate indicates that the model has plenty of True positive examples. The F-1 score, which is the weighted average of precision, recall, accuracy, is 93% for class 0 and 42% for class 1, accuracy is 87%.

# SVM Model

SVMs are a well-liked supervised learning technique for tasks involving classification and regression. SVMs are especially helpful when the dataset contains a large number of features or when the connection between the features and the target variable is non-linear. In certain situations, SVMs can perform better than conventional techniques like logistic regression. Because we have roughly 19 features, we utilize the SVM model to examine the data and determine whether it is superior to the logistic regression model.

|  |  |  |
| --- | --- | --- |
|  | P | N |
| T | 729 | 25 |
| F | 109 | 24 |

After applying the model, we get the following outcomes: Precision for classes 0 and 1 is 87% and 49%, respectively. It gauges how well the model foresees successful results. A high precision means that the model rarely predicts something that will actually happen. The recall rates for classes 0 and 1 are 97% and 18%, respectively. It calculates the capacity of the model to identify each instance of positive in a data. A high recall rate indicates that the model has plenty of actual positive examples. The F-1 score, which is the weighted average of precision, recall, and accuracy, is 92% for class 0 and 26% for class 1, and the model's total accuracy is 85%. Lastly, the table below compares the F1 score, which is the weighted average of precision, recall, and accuracy.

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

The accuracy of both models is slightly lower, which leads us to the conclusion that it is challenging to forecast whether or not a client will subscribe based on these factors. A high F1 score means that the model's precision and recall are satisfactory. In other words, the model predicts with a high proportion of true positives and a low proportion of false positives. Both models have strong F-1 scores in class 0. As a result, the model is able to anticipate it correctly. Due to imbalanced data, F-1 scores for class "1" are exceedingly low. Yet, depending on the investigations that came before, we may still infer if a person is likely to subscribe. First, based on the model's findings, we may ascertain the main causes of the customer's resistance to subscribing using the important factors indicated in the logistic regression model. For magazine subscriptions, it's crucial to consider consumer spending patterns on goods like wine, steak, and gold as well as the volume of transactions made and the frequency of website visits. While each model has benefits and drawbacks, it might be challenging to decide between them, however logistic regression surpassed the SVM model in terms of the metrics.

# Recommendation

We can provide the business with a set of guidelines from this research for comprehending magazine subscription behavior and identifying the most important areas to concentrate on to boost sales.

# Analyze customer web visits to improve online presence.

# Target high-spending customers and optimize sales.

# Personalize marketing efforts based on purchase history.

# Focus on customer retention through loyalty programs and excellent service.

# Use data-driven insights to inform marketing and sales strategies.

# Reference

1. scikit-learn library's documentation:

https://scikit-

learn.org/stable/modules/generated/sklearn.linear\_model.LogisticRegression.htm

l

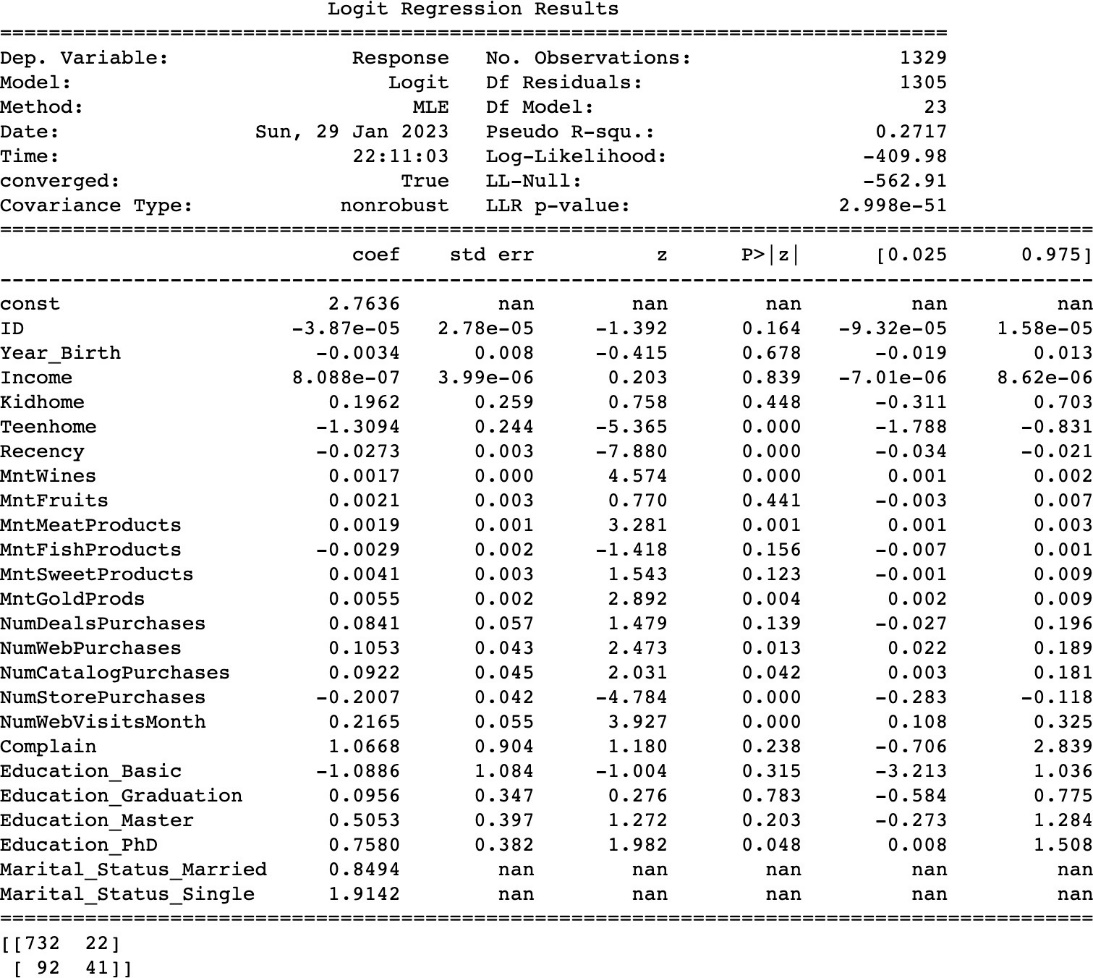
1. A comprehensive tutorial:

https://[www.datacamp.com/community/tutorials/svm-](http://www.datacamp.com/community/tutorials/svm-)

classification-scikit-learn-python

# Appendix 1

“Logistic Regression Results”



“Classification Report for Logistic Regression”

