**ASSIGNMENT 3**

**UNDERSTANDING MAGAZINE SUBSCRIPTION BEHAVIOUR**

Trilok Palla

Northeastern University [College of Professional Studies](https://cps.northeastern.edu/)

**ALY 6020- Predictive Analytics**

Justin Grosz

**INTRODUCTION:**

Magazine subsciptions have become one of the basic requirement over the past 2 decades. Most of them are issued along with newpaper. But due to high technological advancement on our smartphones, we can hardly see any newspapers let alone magazines. People prefer more online issues rather than on paper, which has had a significant effect on its market.

**About Logistic Regression:**

Logistic regression is a supervised classification technique at its core. For a given collection of features (or inputs), X, the target variable (or output), y, can only take discrete values in a classification issue. The model typically constructs a regression model to predict the likelihood that a given data item belongs to the category labeled "1." Logistic regression models the data using the sigmoid function, much as linear regression assumes the data follows a linear distribution.

**About SVM:**

It is a Supervised Machine learning algorithm which is used both for regression and/or classification. Although it is occasionally quite helpful for regression, classification is where it is most often used. In essence, SVM identifies a hyper-plane that establishes a distinction between the various types of data. This hyper-plane is only a line in two-dimensional space. Each dataset item is plotted in an N-dimensional space using SVM, where N is the total number of features and attributes in the dataset. The best hyperplane should then be found to divide the data. SVM can only perform binary classification (i.e., choose between two classes).

**About the data:**

This dataset consists of 2240 data points with 29 features. These 29 features consist of 3 object type data and rest 26 are numeric. With the help of these features, we are trying to train our data using the Logistic Regression and SVM techniques to get which parameter affects AcceptedCmp1 (Dependent Variable) the most.

**Libraries Used:**

* Numpy
* Pandas
* Matplotlib
* Seaborn
* Sklearn methods
* Requests

**Part 1:**

**BASIC DATA CLEANING:**

* The first step here is to check if the data consists of any NULL values in it. From checked there were 24 NULL values found on ‘Income’ column. These missing data points have been filled with mean values. In this manner, we can avoid removing these individuals from the dataset and also not skewing the column values. Although not ideal, this approach enables you to include values that do not affect the entire dataset because the average remains constant regardless of how many averages are added.
* Next duplicates were checked to see any repeated data points. None were found.
* The object Parameter Dt\_customer was found to have date values as such, it has been converted to datetime dtype and has been split onto 3 columns as Year, Month, Date of subscription.
* Unnecessary variables such as 'AcceptedCmp2', 'AcceptedCmp3', 'AcceptedCmp4', 'AcceptedCmp5' , 'Response' have been dropped.
* Also features with only one class or a single value have been dropped, i.e, 'ID', 'Z\_CostContact', 'Z\_Revenue'.

**Part 2:**

**DATA VISUALIZATIONS:**

In this step, few pie charts, bar plots and box plots of object variables have been depicted to get a better understanding of the data in hand. A heatmap has also been put in to get correlation values among the variables.

Given below are the observations that can be seen from the above plots:

* 1 Major Oulier was found in Income column with an income of 666666.
* Most of the customers are graduated.
* A subscription acceptance of only 6.4 was found.

The plot Images have been attached in the appendix.

**DATA PREPROCESSING:**

In this, we have split the training and test data set in a 80 to 20 ratio with ‘AcceptedCmp1’ as the dependent variable under data frame ‘y’ and the rest of the variables under a different data frame ‘x’.

**MODEL BUILDING AND COMPARING:**

The data now which we currently have is having 2240 data points with 23 features ready to be built.

The logistic regression and SVM here can be built using both “sklearn” and “statsmodels”. But with “statsmodels” we will be able to get a better summary to modify our model using various parameters from confusion matrix obtained.

**Part 3:**

**Model Building using Logistic Regression:**

Table

Description automatically generated

From the **first round of model building**, given below are the results obtained,

Graphical user interface, text, application

Description automatically generated

Observations:

* Many variables are found to have p-value greater than 0.05.
* Model accuracy of 93% is observed.
* Columns **NumWebPurchases(P- 0.028) , MntSweetProducts(P- 0.057)** are found to have p-values close to 0.05.
* Model is having a pression of 50 %.
* Model is found to have 4 false positives.
* Variables such 'Year\_Birth' ,  'Kidhome' ,  'Recency',  'MntFruits’ , 'MntMeatProducts' , 'MntGoldProds' , 'NumDealsPurchases' ,'NumStorePurchases' ,'NumWebVisitsMonth’ ,'Complain’ ,'Day\_Customer' have been removed for the second round due to **p-value > 0.1.**

For the **second round of regression**, after removing variables P-value greater than 0.1 from the dataframe. Given below are the observations noted,

Table

Description automatically generated

Results of model are as below,

Table

Description automatically generated

Observations:

* A good development on precession has been observed with **75%.**
* **Lower false positives** than previous model(Only 1) from the **Confusion Matrix.**
* Recall also has been decreased.
* Model accuracy has been a been decreased slightly to **92%.**
* An increase in False negatives is observed.

Lower Presision values have been observed from further model building.

**Model Building Using SVM:**

The cleaned data has been picked and split into 80:20 ratio for building model using SVM. Given below are the results obtained.

Table

Description automatically generated

Observations:

* A model accuracy of 92% has been obtained.
* Precision of only 36% is observed.
* More false positives have been observed (7), than logic regression.

**CONCLUSION:**

So, after 2 rounds of using the logic regression model and SVM model on the given dataset, it is clear that logistic regression is having an upper hand, with high precsssion and good accuracy scores over SVM. As such ideal model recommendation would be **logistic regression.**

2 variables have been observed with a good P-Value i.e, **NumWebPurchases(P- 0.028) and MntSweetProducts(P- 0.057)**. These value have reached to having P-Value of 0.01 on the second round of regression. These both are having an ideal p-value and relates the most and directly relate to the target variable ‘**AcceptedCmp1’**.

How acceleration effect MPG:

Technically several factors come into play that an individual is opting to purchase a magazine subscription . But from the analysis done on the given data set, **The higher the NumWebPurchases and MntSweetProducts** happen to have highest impact on High interest responses.(Best First impression)

**Business Recommendations:**

As we have analyzed from the data we have, having more sweet products displayed on the magazines can significantly increase the Magazine subscriptions. More amount should be spend on Sweet products. These products enables to have better impression on people. Majority of people are attracted to sweets. They can also get a better impression by attaching a local made sample toffee to promote it along with magazine to give a great impression. The magazine company should start advertising/marketing more information about sweet products such as pastries, famous deserts around the town, local specialities as such to have a higher subscription rate.

Higher web purchases and visits on web also leads to higher subscription rates. This should be a bit obvious as people who mosly visit the website are either daily users or new customers looking to subscribe online. The site can be enhanced with better webpage designs. Also if the reading from web, people should able to access the magazine in every type of device (Phone, PC, Ipads). Options should be created for better/clear reading of maganizes, this can enable more suggestions to new customers from the current ones.

**Appendix:**

Chart, pie chart

Description automatically generated

*Fig1: A Piechart plot of Education Attribute.*

Chart, pie chart

Description automatically generated

*Fig2: A Piechart plot Marital Status*

A screenshot of a computer

Description automatically generated with low confidence

*Fig3: Correlation Heatmap*

A picture containing histogram

Description automatically generated

*Fig4: Income Distribution Barplot*

*Chart, box and whisker chart

Description automatically generated*

*Fig5: Income Distribution Boxplot*

**References:**

<https://www.geeksforgeeks.org/introduction-to-support-vector-machines-svm/>

<https://www.geeksforgeeks.org/understanding-logistic-regression/>

Thank you,

Trilok Palla