Module 1 Assignment: Understanding Magazine Subscription Behavior

Dhairyav Jatin Shah

College of Professional Studies, Northeastern University

Professor: Justin Grosz

January 29, 2023

**Introduction**

The consumer subscription patterns of a magazine publisher is used in this project. The company noticed a decrease in subscriptions from the previous year. The corporation is attempting to determine what is and isn't working for the clients using this data set in an effort to increase the number of members. SVM and Logistic Regression are the two models we'll employ to forecast whether a consumer would subscribe or not. The organization can then modify its marketing policies and tactics using the analysis.

**Data cleaning**

There are 2240 records in the dataset, along with 29 variables. The data set has 29 variables total, including 2 categorical date type variable that contains the education and marital status of the customer and 27 numerical variables. As mentioned in the instructions we drop all the columns which are not supposed to be used while making the predictions. These columns **are “Z\_CostContact”, “Z\_Revenue”, “AcceptedCmp1”, “AcceptedCmp2”, “AcceptedCmp3”, “AcceptedCmp4”, “AcceptedCmp5”**. Further we study the data for any null values present. As seen in the **“Income”** we have 24 null values. To understand what method is to be used for imputation we plot a distribution chart. From fig 1 (in appendix 1) we can see the distribution chart of Income and it is skewed towards the right with some outliers towards the left. Looking at this we should not use the mean as the outliers would displace the mean towards the right, hence we will use the median to impute the 24 missing values in the column. Further looking at the data types we observe that the DT\_Customer data type is classified as object we will not use this column to break the date into 3 columns Day, Month and Year which will make it easier to utilize while training the data for modelling.

Next, we study the data set for any outliers present. We use box plot to visualize the outliers in all the numerical columns. From figure 2 (in appendix 1), we can see that there a many outliers in different columns, but among these we can see that outliers in “Year\_birth” and “Income” have outliers which can affect the data while preparing the model. Hence, we eliminate the rows from the “Year\_birth” columns whose birth year is before 1900. Also to study the outliers in “Income” column we use the formula for outliers threshold which is **Q3 + (1.5\*IQR).** Using this formula, we can see that the upper threshold of income is 117416.0. We will eliminate all the rows having income greater than 117416.0. Figure 3 (in appendix 1) represents both the columns after we have treated the outliers.

Once we have treated the continuous variables now we take a look at the categorical variables. We see some abnormal categories in these columns hence we will narrow down these different categories. In the Education column we replace 2n Cycle with masters and in Marital Status we replace Alone, YOLO and Absurd to Single while Together is changes to Married.

**EDA**

Once we have completed the process of data cleaning, we construct a descriptive statistical table to understand the data. From table no 1 (in appendix 2), we can see that our data set contains customers born in the year 1940 till 1996. The average income of the data set is $51,619 with a minimum of $1,730 and a maximum of $113,734. In the columns Kidhome and Teenhome we have 3 values 0,1,2 where they have mentioned in there are any dependents in the household and how many. Average amount spent on Wine is $034.99, Fruits is $26.3, Meat is $165, Fish is $37.64, Sweet is $27 and Gold is $44.11. Further we plot a heat map to understand the collinearity of the dependent variable on the independent variables. From figure (in appendix 1) we can see that our dependent variable Response is correlated the amount spent on different products, income, Year the customer was born.

**Analysis**

**Modeling**

Once the initial analysis is complete, we first split our data into train and test data frames with a 80:20 ratio. First, we build a SVM model where we have dropped Response, Education and Marital Status from our X variable since one a dependent variable and other 2 are categorical variables. We set the kernel = linear and fit the SVM model. From table no 2.1 (in appendix 2), we can see the confusion matrix that we have obtained from the test data set. In this table we can see that we have 378 True Positives and 6 True Negatives, 8 False Negatives and 54 False Positives. We achieved an overall accuracy of 86% . From the classification report in table no 2.2 (in appendix 2), we can see that the model is able to predict if someone is not going to subscribe with a precision of 88% but the precision of predicting if some one is going to subscribe is 43%. We also have a recall of 98% which states that the model is able to predict most of the true vales correctly.

Further, we construct a Logistic Regression model to predict the response variable. For the first model we include all the numerical independent variables which going to be used the predictor variable Response. Upon fitting the model, we obtain the summary displayed in table no 3.1 (in appendix 2). From the table we can see that column like ID, Year\_birth, Kidhome, MntFruits, MntFishProducts, MntSweetProducts, MntGoldProds, NumWebPurchases, Complain, Day, Month and Year. The overall accuracy of the 1st model is 87%. From table no 3.3 (in appendix 2) we can see the confusion where we have 374 True Positives and 16 True Negatives, 44 False Negetives and 12 False positives. Now we build the 2nd model after eliminating the insignificant variables. From the 2nd model we can obtain a accuracy of 87% which increased 1% since model 1. This model also has some features which are not significant hence they will also be eliminated. MntGoldProds and NumWebVisitsMonth are eliminated from table no 3.1 (in appendix 2). From table no (in appendix 2) we can see the confusion matrix for the 2nd model. Where we can see that it has calculated same number of True Positives but True negatives decreased by 1. The over all accuracy of the model is 87% with a recall of 97% and precision of 89%. Now we go ahead to fit the 3rd model. From table 3.1 (in appendix 2) we can see that all the selected variables are Income, Teenhome, Recency, MntWines, MntMeatProds, NumDealsPurchases, NumCatalogPurchases, NumStorePurchaes. Out of these variables the most significant variable is number deals purchases. Hence we can say that the subscription can increase if better deals and discounts are offered to the customers. Also the feature teenhome plays a negative impact, which states that homes having teen dependents do not purchase magazines so often. Looking at the confusion matrix from table no 3.2 (in appendix 2) we can see that this model was able to predict 376 True Positives and 16 True Negatives, 44 False negatives and 10 False positives. Here we most concerned value is the false negatives which states that the model predicted the customer is going to subscribe but didn’t . The accuracy of the 3rd model is 88% which is the highest out of all the three models with a precision of 90% and a recall of 97% which is a good model to predict if the customer is going to subscribe or not.

**Conclusion**

Based on our analysis and modeling, we created 2 classification models. In SVM our model displayed an overall accuracy of 86%, precision of 88% and recall of 98%. In this model we have 54 false negatives which states that the customer is going to subscribe but doesn’t. In our Logistic Regression model, we obtained an overall accuracy of 88%, precision of 90% and recall of 97%. Our Logistic regression model was able to predict True positives and True Negatives better than the SVM model. Where as it also predicted a total of 44 false negatives and 10 false positives. Which is 10 false negatives less than the SVM model. Which can give us a better understanding of our target audience and how to approach them. The significant features we obtained from Logistic Regression Model also help us understand the trend where we say the coefficient of teenhome is almost -0.9 which states that people with a teen dependent at home is less likely to subscribe and most significant variable was the number of purchases made from deals having the highest coefficient of 1.69.

**Recommendation**

Using our research and modeling, we can give the following recommendations to the company to understand the subscription behavior of their magazine.

1. Homes with teen dependents are least likely to subscribe. The magazine company can work on the content creation of the magazine and make it more attractive towards the teen category.
2. The probability of a customer subscribing to the magazine increases if the company offers better deals to the customers.
3. From our confusion matrix we can recommend the company can pay close attention to the false negatives and invest more time in converting those people towards subscribing to the magazine.

**Reference**

1. Chouinard.J (May 2022) *How to use Confusion Matrix in Scikit-Learn (with Example)* <https://www.jcchouinard.com/confusion-matrix-in-scikit-learn/>
2. Chouinard.J (May 2022) *How to use Classification Report in Scikit-learn (Python)* <https://datatofish.com/statsmodels-linear-regression/>

**Appendix 1**

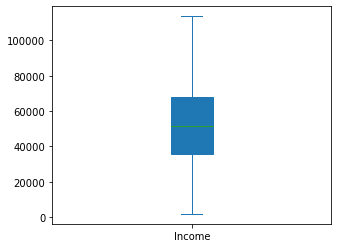
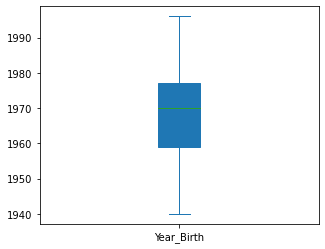
**Figure 1 : Hist plot**

**Graphical user interface, application

Description automatically generatedA picture containing shoji, window, train, crossword puzzle

Description automatically generatedFigure 2 : Box plot for Outliers**

**Figure 3 : Box plot after treating outliers.**



**Chart, treemap chart

Description automatically generatedFigure 4: Heatmap**

**Appendix 2**

**Table no 1: Descriptive Statistics**

|  |  |  |  |
| --- | --- | --- | --- |
|  | **Mean** | **Min** | **Max** |
| **ID** | 5589.29 | 0.0 | 11191.0 |
| **Year\_Birth** | 1968.88 | 1940.0 | 1996.0 |
| **Income** | 51619.50 | 1730.0 | 113734.0 |
| **Kidhome** | 0.44 | 0.0 | 2.0 |
| **Teenhome** | 0.50 | 0.0 | 2.0 |
| **Recency** | 49.10 | 0.0 | 99.0 |
| **MntWines** | 304.99 | 0.0 | 1493.0 |
| **MntFruits** | 26.34 | 0.0 | 199.0 |
| **MntMeatProducts** | 165.28 | 0.0 | 1725.0 |
| **MntFishProducts** | 37.64 | 0.0 | 259.0 |
| **MntSweetProducts** | 27.16 | 0.0 | 263.0 |
| **MntGoldProds** | 44.11 | 0.0 | 362.0 |
| **NumDealsPurchases** | 2.31 | 0.0 | 15.0 |
| **NumWebPurchases** | 4.10 | 0.0 | 27.0 |
| **NumCatalogPurchases** | 2.63 | 0.0 | 28.0 |
| **NumStorePurchases** | 5.8 | 0.0 | 13.0 |
| **NumWebVisitsMonth** | 5.33 | 0.0 | 20.0 |
| **Complain** | 0.008 | 0.0 | 1.0 |
| **Response** | 0.14 | 0.0 | 1.0 |
| **Day** | 15.64 | 1.0 | 31.0 |
| **Month** | 6.46 | 1.0 | 12.0 |
| **Year** | 2013.0 | 2012.0 | 2014.0 |

**Table 2: SVM Model**

**Table 2.1 : Confusion Matrix of SVM**

|  |  |  |
| --- | --- | --- |
|  | **Yes** | **No** |
| **Yes** | 378 | 8 |
| **No** | 54 | 6 |

**Table 2.2 : Classification Report of SVM**

**Table

Description automatically generated**

**Table 3 : Logistic Regression Model**

**Table 3.1 : Regression Results.**

**1st Model**

**Table

Description automatically generated**

**2nd Model**

**Table

Description automatically generated**

**3rd Model**

**Table

Description automatically generated**

**Table 3.2 : Confusion Matrix for all 3 models.**

|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
|  | **1st Model** | |  | **2nd Model** | |  | **3rd Model** | |
|  | **Yes** | **No** |  | **Yes** | **No** |  | **Yes** | **No** |
| **Yes** | 374 | 12 | **Yes** | 374 | 12 | **Yes** | **376** | **10** |
| **No** | 44 | 16 | **No** | 45 | 15 | **No** | **44** | **16** |

**Table 3.3 : Classification report of all 3 models.**

**Table

Description automatically generated with medium confidence1st Model**

**Table

Description automatically generated with medium confidence2nd Model**

**Table

Description automatically generated with medium confidence3rd Model**