Logo

Description automatically generated

ALY6020 Module 3 Project

**Authored by: Prateek Singh**

Course: ALY6020 – Predictive Analytics

Submission Date: 05/03/2022

**Instructor’s name: Kristen Drobnis**

# Introduction

There is a drop in subscription for a magazine company in the last year. Magazines had a notion that people staying more at home will help them with increase if number of subscriptions, however that not the case they wanted to do the root cause analysis and suggest the reason behind such a drop. This will help them plan their subscription program by targeting right audience for them. For the purpose of this analysis, we were given the dataset of the subscribers/customers with multiple factors collected like Education level, Marital Status, Expenditure, Income, No. of Kids, Subscription Accepted Status, etc. We did some data cleaning to bring it to the cleansed state where we can perform the data analysis and build the required models.

There are some missing values which were treated accordingly, and the methodology used to replace them is explained in the Analysis section. We merged few redundant behavioral variables as one as well to eliminate the semantic confusion. I have also clubbed few factor variable levels together to reduce the sub-groups and make the analysis clearer. Few columns/variables were dropped too which deemed unnecessary. Some data visualizations were done to understand the correlation between the variables. Finally, the Logistic Regression and Support Vector Machine Model were made to analyze the results and understand the factors affecting the subscription response of the customers.

# Analysis

## Data Profiling & Cleaning

The dataset has 2240 observations and total of 29 columns, out of which 26 are of numeric type and 3 is object type. We can see the Income variable has some missing values.

A screenshot of a computer

Description automatically generated with medium confidenceTable

Description automatically generated

Fig 1: Dataset Summary

There are total of 24 missing values present in Income variable, which needs to be dealt with.

A picture containing text

Description automatically generated

Fig 2: Number of missing values

To replace the missing values, I first established the levels of corresponding Education class where the Income values are missing. This was done because, I wanted to replace the null values with the median of Income values of the similar Education group which make sense because Income and Education level are dependent on each other theoretically and practically. Higher the education you achieve, better salary can be bagged. Hence, I found the Income values are missing from below Education Groups.



Fig 3: Education Group having missing Income values

The median of each group is calculated which will be used to replace the missing values. The median of each Education Groups comes as below. These values were then imputed at the missing values and appended in the data frame. We can see that customer having a ‘PhD’ degree has the highest median Income, then comes our Graduate group consisting of ‘Graduation’ and ‘Master’ degrees and lastly Undergraduate group comprising of ‘2n Cycle’ and ‘Basic’ degrees. This proves our above hypothesis about the relationship between Income and Education level and justifies the methodology of replacing missing values with median of Income from corresponding education group.

Table

Description automatically generated

Fig 4: Median Income Value for Each Education Level

In this section I merged few columns as one and clubbed few levels. Starting with dropping superfluous ‘ID’, ‘Dt\_Customer’, ‘Z\_CostContact’ and ‘Z\_Revenue’ variables, I added all the expenditure of the customer done over ‘MntWines’, ‘MntFruits’, ‘MntMeatProducts’, ‘MntFishProducts’, ‘MntSweetProducts’ and ‘MntGoldProds’ as one and named it as ‘Spent’ column.

‘Married’ and ‘Together’ were grouped under ‘Partner’ marital status and all the others were grouped as ‘Alone’. Two more columns ‘Kidhome’ and ‘Teenhome’ were added together to form ‘Children’ column. As described above the Education levels were also grouped as ‘Undergraduate’, ‘Graduate’ and ‘PhD’ degree holders. Finally, the columns ‘NumDealsPurchases’, ‘NumWebPurchases’, ‘NumCatalogPurchases’ and ‘NumStorePurchases’ were added together to form a ‘Purchase’ column in our dataset. All the columns which were merged were then dropped to clear the data frame of redundancy and hence we got the new data frame having 2240 entries but now with total of 16 columns. We can see here that we still have left with two object variables namely, ‘Education’ and ‘Marital\_Status’ and hence these were converted into numerical variables using Label Encoder.

Table

Description automatically generated with medium confidence

Fig 5: Cleaned Dataset Summary

Finally, we have our dataset to be used for our analysis purpose. We can see that average population is from the Year 1969 (Current Age ~ 53 years) which is normally distributed. Maximum of the population holds Graduate degrees and have a marital status of Married. Average Income gathered is $52243 and looks normally distributed just for one anomaly of $666666 as an outbound outlier. Recency looks normally distributed. Accepted Campaign Results and Complain values doesn’t show any sort of relations with the dataset. Spent amount is right-skewed, however, purchase is normally distributed. We can also observe that most of the customers have chose to withdraw the subscription in the last campaign.

Table

Description automatically generated with medium confidence

Fig 6: Statistical Data Set Analysis

Pair plot is plotted to see the relationship between each variable, however it is not very visible in the screenshot below. Subscriptions are more when the ‘Spent’ values are on the higher side. Visually it is significant that the customer is more likely to have subscribed in the latest campaign if they have subscribed in any of the campaigns before.

Chart, schematic

Description automatically generated with medium confidence

Fig 7: Pair-Plot to see the relationship between variables

## Data Visualization

It looks like there is no significant effect of Year of Birth of Customers on Subscription Count. However, we can see that there is maximum subscription from the customers from the age group between ‘1969’ and ‘1977’ and then from ‘1982’ and ‘1984’. We saw a continuous decline after that meaning younger generations are not much into magazine subscriptions.

Chart, bar chart

Description automatically generated

Fig 8: Effect of Age on Subscription

We saw that in each Education group there is a very big difference amongst number of subscribers and non-subscribers. Because we have more population of ‘Graduate’ level we can see that the Subscribers count are more there, however the total percentage of subscribers comes out to be mere 14%. Second biggest population is of ‘PhD’ holders with almost 21% population as subscribers. Lastly, ‘Undergraduate’ population have only 9.33% subscribers. With the increase in education level we saw the increase in subscription rate.

Chart, bar chart

Description automatically generated

Fig 9: Effect of Education on Subscription

We have a higher population with relationship status as committed and fewer who are single. However, we can see that the probability of single customers subscribing for the magazines are higher than the committed fellows. Which can be predicted as they might have more self-time and might be lonelier than couples.

Chart

Description automatically generated

Fig 10: Effect of Marital Status on Subscription

From the below image we can conclude that Subscribers from any Education group have higher median Income, which shows money is one of the factors affecting the subscription. With pandemic hitting and more people sitting at home, many lost their jobs and were out of income hence this could be one of the factors which saw a decline in subscription in recent times. There are few outliers however all belonged to the non-subscriber group.

A picture containing timeline

Description automatically generated

Fig 11: Regression Model with all the Independent Variables

There is not much difference but slightly subscription is more on higher Income level.

Table

Description automatically generated

Fig 12: Income vs Subscription Distribution Plot

We cannot see definite affect happening because of the recency of days since customer's last purchase. The Subscription seems to go down as customers stops purchasing items much. Non-Subscribers pretty much remain constant throughout.

Chart, histogram

Description automatically generated

Fig 13: Recency vs Subscription Distribution Plot

We cannot find any correlation between the amount spent by customers and subscription status. However, we saw a bit of increase in the subscribers between the amount spent of $1100 to $1800 in last two years.

Chart

Description automatically generated

Fig 14: Spent Amount vs Subscription Distribution Plot

We cannot see very high correlation between response and other variables. Highest being 0.33 positive relation with Accepted Campaign 5 and -0.20 negative relation with Recency on each end. We have the highest positive correlation between ‘Purchase’ and ‘Spent’ of 0.75 and negative correlation between ‘Number of Web visits per month’ and ‘Income’ of -0.55. ‘Income’, ‘Spent’, ‘Web Visits’, ‘Purchase’ and ‘Children’ seems like most important factors. We will investigate this further in our model analysis.

Timeline

Description automatically generated

Fig 15: Correlation Heat Map amongst variables

We calculated the VIF Score for checking the multicollinearity and couldn’t find one between the variables.

## Logistic Regression Model

We build the model of accuracy 85.5% with almost all the predictor variables as significant variables but ‘Education’, ‘Income’, ‘Accepted Campaign 2’, ‘Complain’ and ‘Purchase’ as insignificant. We can see that Year of Birth has negative affect on subscription meaning newer generation is not into magazines anymore. Marital Status, Recency and Children also have the negative affect on subscription status which speaks a little about availability in terms of time for reading.

Table

Description automatically generated

Fig 16: Regression Model Summary

Our Model Accuracy and Precision value is displayed below.

Text

Description automatically generatedChart, treemap chart

Description automatically generated

Fig 17: Logistic Regression Confusion Matrix

## Support Vector Machine Model

Chart, treemap chart

Description automatically generatedOur SVM Model accuracy is seen a little lower than Logistic Regression Model and comes out to be 84.5%. Our Model Accuracy and Precision value is displayed below.

Text

Description automatically generated

Fig 18: Support Vector Machine Confusion Matrix

# Conclusion

1. Logistic Regression is slightly a better model than Support Vector Machine model in this case.
2. Our Logistic Regression model has 85.5% accuracy and 60.60% precision.
3. Family is the biggest negative influence on the subscription rate. We saw married couples having children are less likely to subscribe for magazines.
4. As the days of customer not purchasing anything increases chances of unsubscribing increase as well.

# Reference

1. CHOUDHARY, P. R. I. T. A. M. (2020, March 13). Bank personal loan modelling. Kaggle. Retrieved April 29, 2022, from <https://www.kaggle.com/code/pritech/bank-personal-loan-modelling/notebook>
2. cosine. (2021, September 14). Logistic regression using Statsmodels. GeeksforGeeks. Retrieved April 29, 2022, from <https://www.geeksforgeeks.org/logistic-regression-using-statsmodels/>
3. Get the percentage of a column in pandas python. DataScience Made Simple. (2019, November 13). Retrieved April 29, 2022, from <https://www.datasciencemadesimple.com/get-percentage-column-pandas-python-2/>
4. Grosz, J. (n.d.). Implementing Logistics Regression. Canvas. Retrieved April 29, 2022, from <https://northeastern.instructure.com/courses/105748/pages/lesson-3-4-implementing-logistics-regression?module_item_id=7143789>
5. Kang, C. (2020, July 5). Applying logistic regression and SVM. Chan`s Jupyter. Retrieved May 3, 2022, from <https://goodboychan.github.io/python/datacamp/machine_learning/2020/07/05/01-Applying-logistic-regression-and-SVM.html#Applying-logistic-regression-and-SVM>
6. Worrell, N. (2020, June 11). Predictive analytics on customer behavior with support vector machines (SVM). Medium. Retrieved May 3, 2022, from <https://towardsdatascience.com/predictive-analytics-on-customer-behavior-with-support-vector-machines-svm-7e68fd2be610>
7. Haripriyan. (2022, March 7). Marketing data preprocessing and eda\_ongoing... Kaggle. Retrieved May 3, 2022, from <https://www.kaggle.com/code/haripriyan/marketing-data-preprocessing-and-eda-ongoing>
8. Greshmapk. (2022, March 2). Marketing campaign. Kaggle. Retrieved May 3, 2022, from <https://www.kaggle.com/code/greshmapk/marketing-campaign/data>
9. Richandata. (2022, February 7). Classification - marketing campaign. Kaggle. Retrieved May 3, 2022, from <https://www.kaggle.com/code/richandata/classification-marketing-campaign>
10. Yesilyurttemre. (2022, January 18). Customer segmentation in the marketing campaign. Kaggle. Retrieved May 3, 2022, from <https://www.kaggle.com/code/yesilyurttemre/customer-segmentation-in-the-marketing-campaign>
11. Mehakfatima. (2018, April 30). Logistic regression and SVM. Kaggle. Retrieved May 3, 2022, from <https://www.kaggle.com/code/mehakfatima/logistic-regression-and-svm/data>

# Appendix

Note: Code is attached separately.