Sean McLean

ALY 6020

Module 3 Assignment

Understanding Market Subscription Behavior

Introduction

The mission of the module 3 assignment is to understand magazine subscription trends and patterns from the past year. With the decline in overall subscriptions, the goal is to look at the dataset provided and research why people who have spent more time at home are not subscribing to magazines for reading consumption. The assignment consists of four parts which will break down the process by data cleaning, modeling, and creating data visualizations. The dependent variable that was selected is the 'response' attribute which indicates whether a customer responded positively or not to the marketing campaign.

Part 1

After importing the dataset into Python and installing the necessary libraries and packages, the head function is executed to provide a glimpse at the first five rows of each attribute in the dataset. The binary classifications are evident in many of the variables where either zero or one is displayed, indicating that the marketing campaign is asking some yes or no questions to the people being surveyed. The questions that are being asked in the marketing campaign pertain to demographic information like income and education, how much is being spent on particular products per month, and how recently they have been a magazine subscriber. Other attributes in the dataset pertain to the survey and whether the individual subscribed afterwards and if they had a positive response to the marketing campaign.

The summary statistics were employed by using the info and describe functions to look at the overall structure of the dataset. There are 2240 entries and most of the attributes are integers with all of them being non-null values. The income variable was the only attribute that had missing values, with 24 cells containing no information on the individual's income. Methods used to fill in then the blank cells were to employ the median value in the income range which equaled 51,381.50. This was verified by looking at the first 30 rows of the dataset because within that range there were two missing values, and each one was filled in with the median value. The 'isnull' function was also printed to check for any remaining missing values, with zero missing values remaining after the data cleaning process was completed (Anaconda, 2024).

Part 2

With the dependent variable set, the dataset was split into training and testing sets with the size set to 0.2. The features are then standardized so they all make an equal contribution to the analyzing process. Implementing the logistic regression function, the logistic regression model was built with the training and testing sets. The overall accuracy of the model is 0.84375 which is a strong percentage, but the overall precision (0.491) and recall (0.377) are both subpar. This would indicate that the predicted outcomes and actual outcomes were not very high and that adjustments might need to be made to the model. The response variable with a value of zero had strong marks in precision, recall, and f1 scores, while the response variable with a value of one had below average marks in all the categories. A confusion matrix was constructed to show the results of the testing set and its high accuracy rating (See Appendix A).

The attributes with the highest coefficient values were 'mntmeatproducts,' 'AcceptedCmp5', 'Education_PhD', 'AcceptedCmp4', and 'NumCatalogPurchases'. The most significant variables paint a picture of the type of individual that had a positive experience from the marketing campaign that was conducted. They are well educated, purchase more products from catalogs, spend a certain amount per month on meat products, and are more likely to subscribe to a magazine at certain points of the year or respond to the campaign. These qualities are important to study in that it identifies a key demographic that had a positive experience with the survey and can be used to attract more subscribers in the future. It also shows that there are certain points of the year where people will more likely respond and subscribe to a publication. Because it is the last two campaigns of the cycle then it is possible that people could subscribe more during the holidays when people are buying gifts, so placing more of an emphasis on the holiday season could be beneficial to the company. The attributes that pertain to meat products and catalog purchases are areas to research as well since it is where people are spending their money, so connecting magazines that could cater to these attributes can lead to higher subscription sales due to these current trends.

The lowest coefficient values in the dataset were ‘MntWines,’ ‘Dt_Customer_2013-04-29', ‘NumStorePurchases’, ‘Teenhome’, and ‘Recency.’ The types of people that this could be are people with or without teenagers in the house, people that are not drinking wine or spending their money in stores and have not bought any subscriptions or responded to the campaign in a while. This could suggest that if households have teenagers in the home, then maybe they are too busy to go spend their money in stores and are instead purchasing products online or in other ways. The negative impact on wine purchasing could also indicate that these are families that are not drinking wine as much as people that do not have teenagers in the house. These findings are useful in that they provide the company with some input into which variables have the most negative effect on its campaign and where to invest in its business model in the future.

Part 3

The support vector machine model is executed with the training and testing sets by using the SVC function. It is then predicted, evaluated, and printed to show its accuracy and the impact the attributes have on the dependent variable. The results using the SVM model are like the logistic regression model in that the overall accuracy of the model and response of zero to the campaign is high but the response of one to the campaign is low. The precision value (0.44) which measures the proportion of true positive predictions out of all the positive predictions made by the model was slightly lower than the precision value in the logistic regression model. The recall value (0.62), however, which measures the proportion of actual positive instances that the model correctly identified, was noticeably higher than the value in the logistic regression model. A ROC model which shows the measure of separability was constructed to show the results of the testing set and its high accuracy rating (See Appendix B). The area under the curve of 0.82 indicates that the SVM model has a strong positive rate overall (Narkhede, 2018).

The attributes with the highest coefficient values were 'mntmeatproducts,' 'Education_PhD,' 'NumCatalogPurchases,' 'MntGoldProds,' and 'Education_Master.' The type of individual or household that this describes is well-educated, that purchases more items from catalogs, and invests in certain items like meat and gold products. This is vital for the campaign in that it identifies a key demographic like the logistic regression model detailed and where and what the person or household is purchasing. The results are valuable for the company because it can put resources in the future into what type of person or household is more likely to subscribe or respond to the survey. It also reveals what kind of products people or households will be more likely to invest in, which can lead to an increase in revenue.

The lowest coefficient values in the dataset were 'Dt_Customer_2013-01-06,' 'Dt_Customer_2013-04-29', 'Teenhome', 'NumStorePurchases', and 'Recency.' The lowest coefficient values are almost identical to the values in the logistic regression model, indicating that there is consistency between the two models. The only difference is that the 'MntWines' has less of a negative impact than the logistic regression model and that the 'Dt_Customer_2013-01-06' variable has a more negative effect on the survey. The type of individual or household with the most negative attributes pertains to whether a teenager is in the home or not, there are not as many store purchases, and the period from the last time they responded or subscribed to a magazine. Like the previous model, this could suggest that because they have teenagers in the house that they spend more time online than in-person due to less time and more convenience and that there is less interest in magazine subscriptions. This is useful information in that the company could divest from this type of person or household or establish a period when if someone has not responded or subscribed for a long time then they will not be contacted again in the future.

Part 4

Analyzing the two models shows that they are both similar in their results with one response having higher marks than the other response. While each model has qualities that would be beneficial to the company, I would recommend incorporating the support vector model for future business and marketing planning. This is despite executing the 'if else' command and 'accuracy score' function which recommends using the logistic regression model instead based off the results. The model is similar in overall accuracy and precision value to the logistic regression model but has a much higher recall value. It also has higher marks when looking at the individual responses in the dependent variable. Comparing the most positive and negative coefficients of the variables, the SVM model fares slightly better in that it is close to the other model in the top attribute's values, and the most negative attributes are closer to zero. The top five variables in terms of positive and negative significance are crucial to the future course of action for increasing magazine subscriptions by showing who will subscribe and who will not.

Conclusion

The binary classification models used in the assignment were critical to identifying demographics, purchasing habits and behaviors, and response times for the company. It has helped solve problems that stemmed from the previous year in the decline of magazine subscriptions and how to rectify the issues. The business model can be revised from the survey results and future adjustments can be made if these results do not achieve better sales. These recommendations I believe will be a major asset to future business revenue and more subscribers for the company.

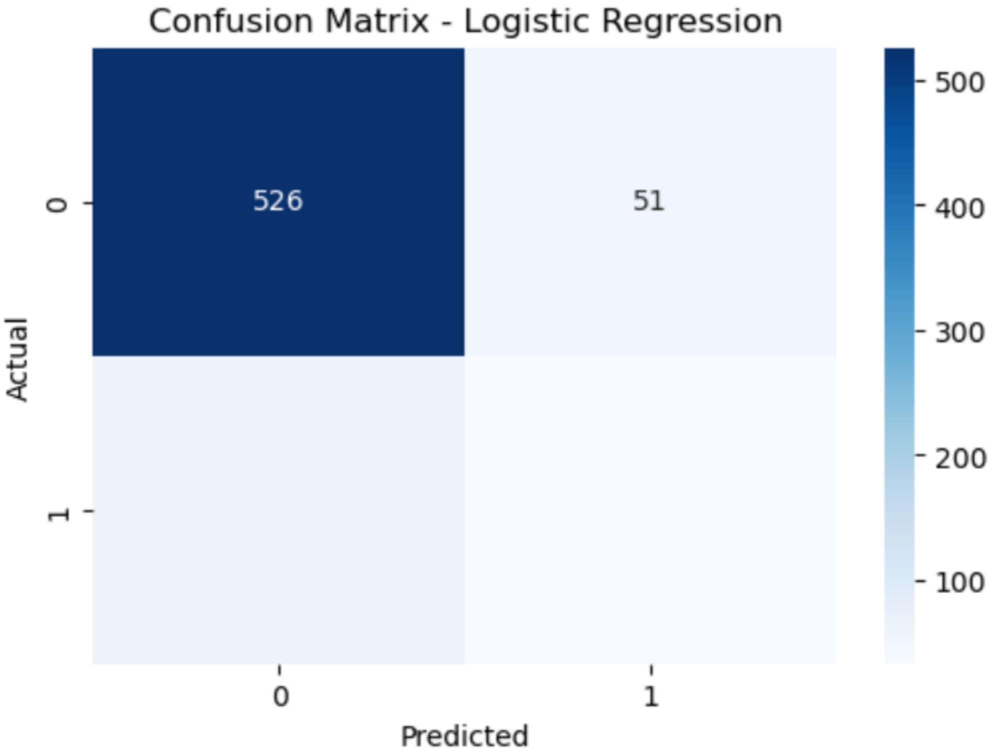
References

Anaconda Distribution. (2024). Anaconda (Version 4.12.3). Retrieved July 27, 2024, from https://www.anaconda.com

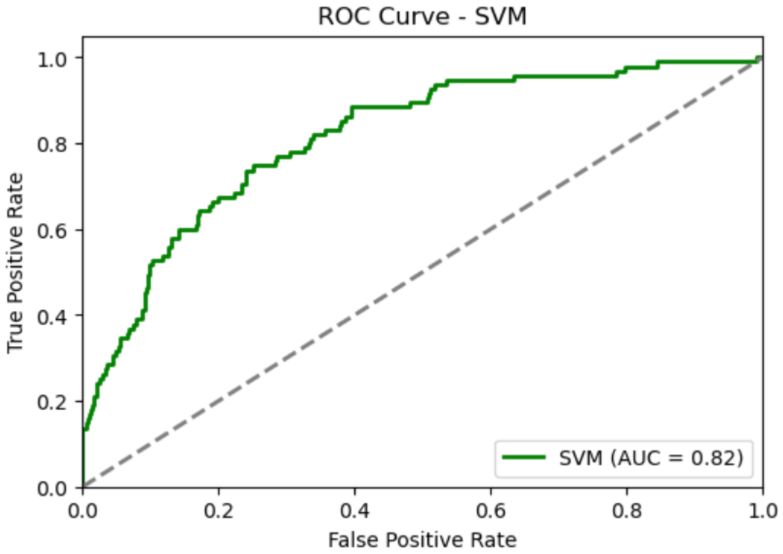
Narkhede, S. (2018, Jun. 26th). *Understanding AUC-ROC Curve.* Medium. https://towardsdatascience.com/understanding-auc-roc-curve-68b2303cc9c5

Appendix

Appendix A



Appendix B


ROC Curve - SVM
True Positive Rate
False Positive Rate
SVM (AUC = 0.82)