**Module 3: Logistic and SVM Model**

ALY6020 – Predictive Analytics, Northeastern University, Boston

Professor Justin Grosz

CRN: 20387

03/12/2023

Submitted by: Abhinav Jain

**Introduction**

This study will examine a magazine publisher who has been seeing a drop in subscribers. Despite the fact that more individuals were spending time at home, the business was not experiencing the rise in readership they had anticipated. We chose to use a combination of logistic regression and SVM modeling to evaluate their subscription data to better understand why this was occurring.

The first step in this procedure was to collect and clean the data before beginning the magazine company's projection. The firm examined a number of factors, such as the kind of material provided, the income of readers, and the readership's demographics. We obtained information on how many individuals spent time at home and how much time they spent doing things like reading.

The magazine firm had access to strong tools, like the logistics and SVM models, to help them comprehend their subscriber data and devise focused measures to reverse the fall in subscriptions. The business was able to enhance its subscription rates by making data-driven decisions and implementing adjustments after evaluating the data and seeing patterns. The consequence was a much higher rate of membership renewals for the business as well as a more enthusiastic and devoted readership. Finally, we'll offer suggestions on how to boost sales to seize future market sales.

**Data Cleaning**

After evaluating the dataset, discovered that there was undesired data present, which might have affected the findings. The initial dataset comprised 2240 records and 29 characteristics. To get better results, we thus deleted several characteristics. In this report, we'll go through how the dataset was cleaned, what we learned from the analysis, and what we deduced from the results.

Several attributes in the dataset, including the life cost contract, revenue, and client purchase date, were deemed to be unimportant after analysis. As a result, we got rid of these attributes to get a more compact dataset. Also discovered that the income variable included 24 null values, which may have affected the outcomes. To make the dataset more balanced, we removed these null values records and were left with 2216 records and 21 characteristics after cleansing the dataset. By eliminating irrelevant information and discarding null values, cleaned up the dataset. After cleaning up the dataset, discovered that marital status and education had significant roles in influencing response variables. These findings can be utilized to create ways to raise income levels for those with less education or single people. Overall, our study offers insightful information about the variables that affect income levels and may be utilized to guide decision-making across a range of businesses. It's probable that in the provided case, the variables contributing to the model's multicollinearity were found using the VIF function. To identify the independent variables that were the source of the multicollinearity, it would be necessary to look at the exact values of the VIF for each independent variable. The Marital Status Divorced 42.8, Marital Status Married 106.8, Marital Status Single 75.7, Marital Status Together 86.7, and Marital Status Widow 15.85 variable were later deleted to alleviate the multicollinearity problem after it was discovered to have a high VIF.

**Exploratory Data Analysis (EDA)**

Analysis shows that the dataset's marital status column contains several inaccurate or irrelevant entries, such as "Alone," "Absurd," and "YOLO." To assure accurate analysis, these values must be eliminated from the dataset which removed these categories from the dataset after converting them to dummy variables.

Chart, bar chart

Description automatically generatedThe distribution of educational levels in the dataset is depicted in the graph. The bulk of the people 1116 had graduated, followed by 481 people who have earned a Doctorate and 365 people who have earned a master's degree. The remaining people in the dataset have the following educational backgrounds: 54 people have finished their basic education, while 200 people have completed a second-cycle program. While assessing the effect of education on the response variable, it was critical to consider the fact that the majority of the persons in the sample have finished higher education programs, as the chart clearly demonstrates.

Chart, bar chart

Description automatically generatedThe distribution of marital status within the dataset is depicted in the graph. The bulk of the population 857 people are married, followed by 573 cohabiting people and 471 single people. The remaining people in the sample are either single or married as follows: 76 people have lost a spouse to widowhood, while 232 people have divorced.

**Logistics Regression Model**

Divided the dataset into a train set and a test set in an 80-20 ratio to assess how well our logistic regression model performed in predicting the response variable. after evaluating the model on the test set after it had been trained on the train set, and the accuracy result was 0.87 (87%).

It also computed the accuracy and recall scores to further assess the model's performance. The proportion of true positives among all positive forecasts, or the accuracy score, was 0.69. The proportion of true positives among all real positives, or the recall score, was 0.169.

The confusion matrix reveals that although the model accurately predicted 379 negative values (true negative, TN), it properly predicted just 11 positive values (true positive, TP). The model inaccurately predicted 5 negative values as positive and 54 positive values as negative (false negative, FN) (false positive, FP).

These findings suggest that the logistic regression model predicts negative values (TN) with a high degree of accuracy but struggles to predict positive values (TP). The poor accuracy and recall scores that were previously computed reflect this. We may draw the conclusion that, with an accuracy score of 0.87, our logistic regression model is reasonably accurate in predicting the response variable. The accuracy and recall ratings, which are comparatively poor, still need to be raised. To enhance the model's performance and enhance the prediction of the response variable, future investigations may investigate other modeling methodologies and tactics. Generally speaking, our logistic regression model offers insightful information about the variables that affect the response variable and may be applied to decision-making processes in various industries.

**SVM (Support Vector Machine)**

After evaluating an SVM model to forecast the response variable, the model had an 85.5% test-set accuracy. The SVM model has a precision score of 0.53 and accurately predicted 53% of the actual positive values. The SVM model's recall score, however, was just 0.15, suggesting that it overlooked many genuine positive values.

The confusion matrix demonstrates that 9 negative values were mistakenly predicted as positive whereas 370 negative values were accurately forecasted as negative by the SVM model. Nevertheless, the SVM model only predicted 10 positive values accurately, missing 55 truly positive values.

|  |  |  |  |
| --- | --- | --- | --- |
| Prediction Matrix | Accuracy | Precision | Recall |
| Logistics Regression | 87% | 0.69 | 0.17 |
| SVM | 85% | 0.53 | 0.15 |

The logistic regression model achieved an accuracy of 87% on the test set, with a precision score of 0.69 and a recall score of 0.17. The confusion matrix for the logistic regression model shows that it correctly predicted 374 negative values and 11 positive values but missed 54 actual positive values and incorrectly predicted 5 negative values as positive.

On the other hand, the SVM model achieved an accuracy of 85.5% on the test set, with a precision score of 0.53 and a recall score of 0.15. The confusion matrix for the SVM model shows that it correctly predicted 370 negative values and 10 positive values but missed 55 actual positive values and incorrectly predicted 9 negative values as positive.

**Conclusion:**

Generally, the logistic regression model outperformed the SVM model in terms of accuracy, precision, and recall. However, both models struggled with identifying actual positive values, which suggests that further analyses could explore different modeling techniques and strategies to improve performance. It is also worth noting that the logistic regression model had a higher precision score than the SVM model, indicating that it may be better suited for applications where minimizing false positives is important.

This research suggests that the response variable may be significantly influenced by a person's marital status. To learn more, future investigations may look more closely at how marital status affects income levels and other factors.

One potential recommendation to increase magazine reading and engagement could be to convert the magazine content into an audio format, such as a podcast. This would allow individuals to listen to the content while engaging in other activities, such as exercising or commuting. Additionally, incorporating slide notes or artwork could make the listening experience more visually engaging and increase overall retention of the material. Providing multiple ways for individuals to consume the content, may increase their likelihood of spending more time with the magazine and ultimately lead to increased engagement and loyalty.

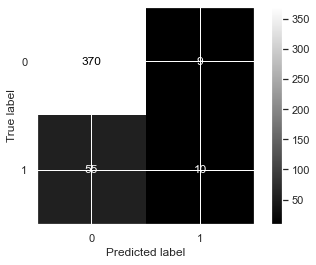
**Appendix:**

**Logistic Regression Model**

Chart

Description automatically generated

**Support vector machine model**



**Correlation matrix:**

A screenshot of a computer

Description automatically generated with low confidence