Module 3 Assignment – Understanding Magazine Subscription Behaviour

Professor: Justin Grosz

Course: ALY6020: Predictive Analytics (CRN 20466)

Submitted by :

Sameeksha Bellavara Santhosh

Date: 28-01-2024

Introduction

The magazine publisher is trying to find the reasons behind the drop in subscriptions. First, we will purify and enhance the data by eliminating specific variables. We will then build a Support Vector Machine (SVM) and logistic regression models to predict subscription behaviour. Finding influential variables will be our main concern. To identify which model is better, we will lastly compare the recall, accuracy, and precision of the two models in addition to identifying important variables that have an impact on business.

Part 1: Data Cleaning

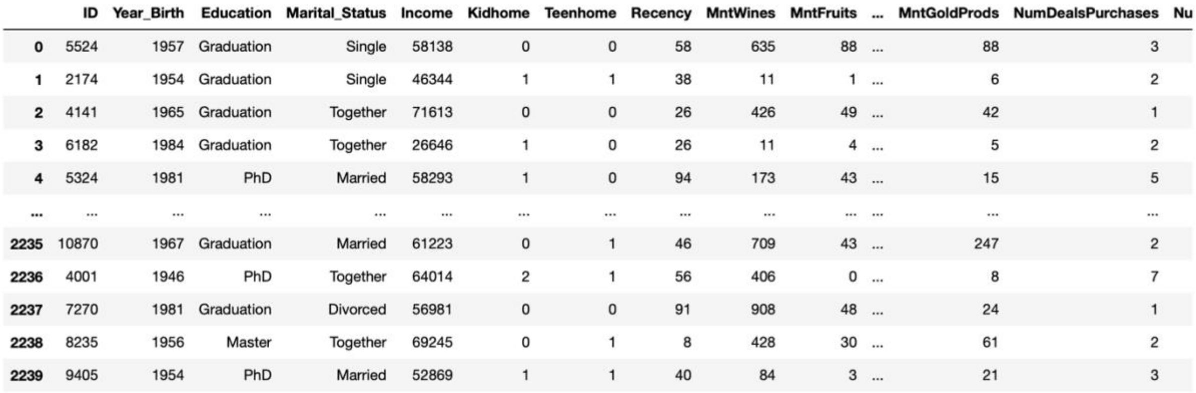
The first and most important step in the data cleaning process was to find and eliminate any unnecessary variables from the dataset. In particular, columns like 'AcceptedCmp1', 'AcceptedCmp2', 'AcceptedCmp3', 'AcceptedCmp4', 'AcceptedCmp5', and 'Dt_Customer' were found to be superfluous for the examination of subscription patterns and were thus removed. By carefully removing unnecessary columns, the focus is improved on important variables, which lowers noise and increases the accuracy of the analyses that follow.

The second step used mean imputation to fill in the missing values in the 'Income' column. The dataset is kept complete without bias introduction by using the mean income value to fill in the missing entries. A popular technique for preserving the variable's overall distribution and enhancing the dataset's dependability for subsequent analysis is mean imputation.

Additionally, the optimization of memory usage within the dataset was considered. The column labelled 'Income' was downcast from its original numeric data type to 'int32' to optimize memory usage. The data type was then purposefully altered to an integer to guarantee consistency and enable more efficient analysis.

Moreover, transformation was applied to non-numeric columns like "Education" and "Marital_Status." Dummy variables were created and these columns were converted to the 'category' data type. This method not only uses less memory but also makes it possible to handle categorical variables in later analyses in an efficient manner. For modelling approaches like logistic regression and SVM, the development of dummy variables is especially helpful as it improves the interpretability and precision of these models.

These data-cleaning methods work together to preserve the quality and integrity of datasets. With this, the dataset is now ready for more in-depth study and modelling projects, such as creating subscription behaviour prediction models and finding important variables with business value.



	ID	Year_Birth	Education	Marital_Status	Income	Kidhome	Teenhome	Recency	MntWines	MntFruits	...	MntGoldProds	NumDealsPurchases	Nu
0	5524	1957	Graduation	Single	58138	0	0	58	635	88	...	88	3	
1	2174	1954	Graduation	Single	46344	1	1	38	11	1	...	6	2	
2	4141	1965	Graduation	Together	71613	0	0	26	426	49	...	42	1	
3	6182	1984	Graduation	Together	26646	1	0	26	11	4	...	5	2	
4	5324	1981	PhD	Married	58293	1	0	94	173	43	...	15	5	
...	...	...	...	...	...	...	...	...	...	...	...	...	...	
2235	10870	1967	Graduation	Married	61223	0	1	46	709	43	...	247	2	
2236	4001	1946	PhD	Together	64014	2	1	56	406	0	...	8	7	
2237	7270	1981	Graduation	Divorced	56981	0	0	91	908	48	...	24	1	
2238	8235	1956	Master	Together	69245	0	1	8	428	30	...	61	2	
2239	9405	1954	PhD	Married	52869	1	1	40	84	3	...	21	3	

Figure 1: Cleaned Dataset

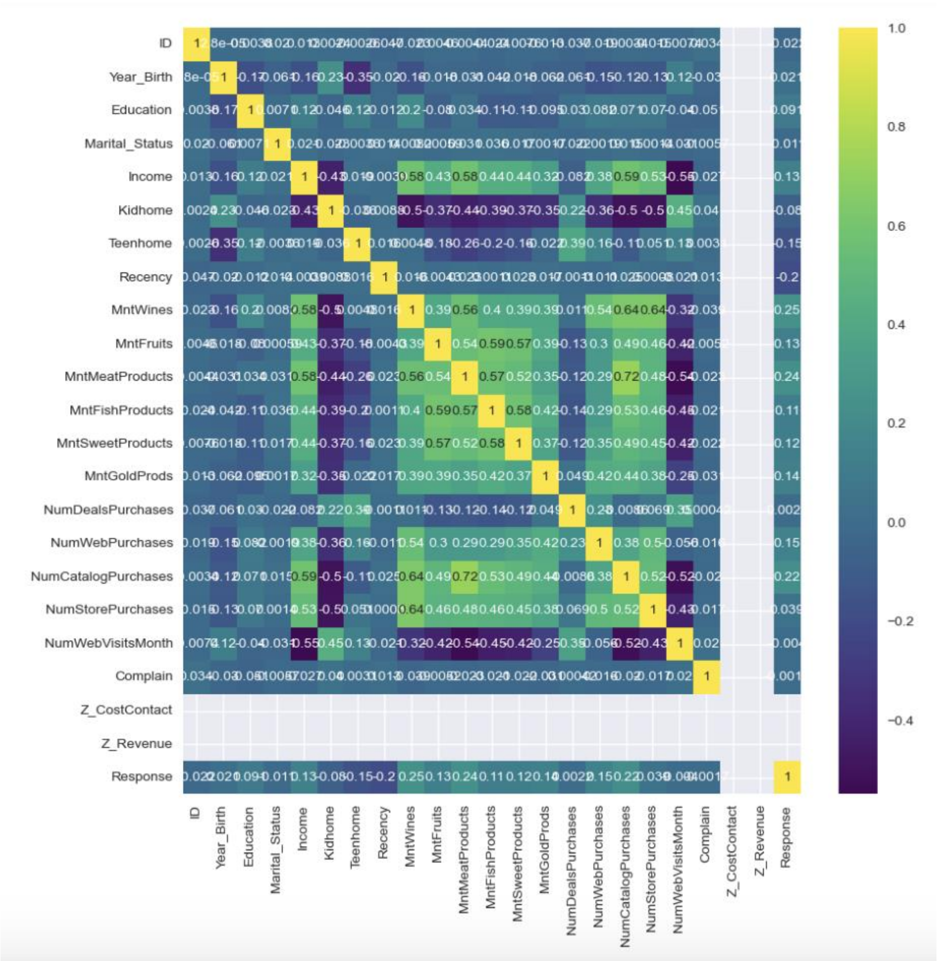


Figure 2: Correlation Heatmap

The correlation between the various variables in the dataset is visually represented by the heatmap that was generated. Lighter colours (closer to 0) suggest no correlation at all, while darker colours (closer to 1 or -1) indicate stronger positive or negative correlations, respectively. For reference, the precise correlation coefficients are given in the numerical annotations. Understanding the linear relationships between variables is aided by this analysis, which can also direct decisions about the construction of models or additional exploratory data analysis.

Part 2: Model Development

Logistic Regression Model: By analysing the model's results, we can determine the significance of various variables and their possible influence on business. The coefficients for each variable are shown in the table's "coef" column, and the corresponding p-values are shown in the "P>|t|" column.

Regression Coefficients and Significance:

Teenhome: The coefficient for Teenhome is -1.1438, with a significant p-value of < 0.001. This negative coefficient suggests that an increase in the number of teenagers at home is associated with a decreased likelihood of subscription. Customers with more teenagers in their households are less inclined to subscribe to the magazine.

Recency: The coefficient for Recency is -0.0289, and the p-value is < 0.001. This negative coefficient indicates that customers who made recent purchases are less likely to subscribe. The more recent the customer's interaction with the company, the lower the likelihood of subscription.

MntWines: The coefficient for MntWines is 0.0018, and the p-value is < 0.001. A positive coefficient suggests a positive correlation between the amount spent on wine purchases and the likelihood of subscription. Customers who spend more on wines are more likely to be subscribers.

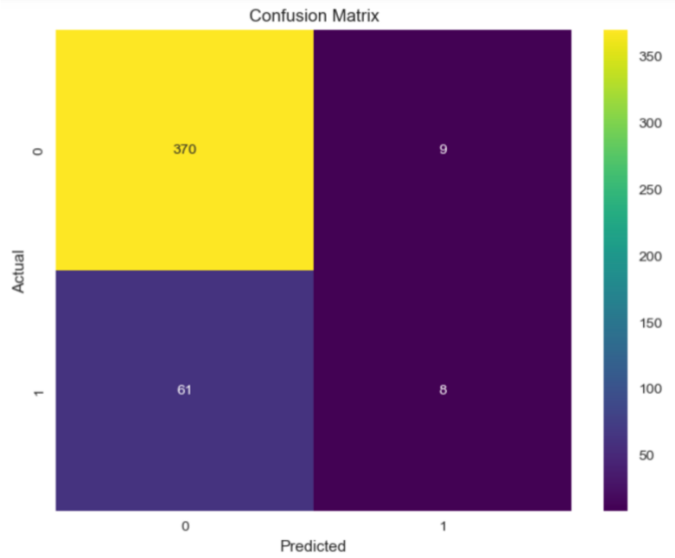
NumCatalogPurchases: The coefficient for NumCatalogPurchases is 0.0941, and the p-value is 0.015. This positive coefficient implies that an increase in the number of catalogue purchases is associated with a higher likelihood of subscription. Customers who engage in more catalogue purchases are more likely to subscribe.

Insignificant Variables:

Several variables, such as 'ID,' 'Year_Birth,' 'Income,' and categorical variables like 'Education' and 'Marital_Status,' do not demonstrate significant coefficients (p > 0.05). These variables may have minimal impact on subscribers' behaviour based on the current analysis.

The factors influencing subscription behaviour can be understood through the use of the logistic regression model. Businesses can more effectively target specific customer segments with their marketing strategies by knowing the importance of each variable, which may lead to an increase in subscription rates. A thorough assessment of the model's predictive abilities is given by its performance metrics, which include accuracy, confusion matrix, and classification report. Over time, the model's accuracy and applicability can be further improved by routinely checking and updating it with fresh data.

Confusion Matrix for Logistic Regression:



The logistic regression model achieved an accuracy of 0.84375 based on the confusion matrix. The business can improve its marketing strategies by focusing on particular consumer segments and modifying its tactics in response to the important variables associated with subscription behaviour that have been identified. The company can make educated decisions if it has a comprehensive grasp of variables like the proportion of teenagers living in households, current shopping patterns, spending on wine (MntWines), and participation in catalogue purchases. Using this information, the business can put in place strategies that will boost subscriptions and, in turn, improve overall performance.

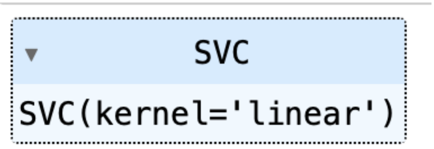
Part 3: Building an SVM model

Confusion matrix for SVM Model:

	Positive	Negative
Positive	367	12
Negative	60	9

The confusion matrix shows that the model correctly predicted 360 instances of non-subscription (true negatives) and 9 cases of subscription (true positives). However, it misclassified 60 non-subscription cases as subscriptions (false positives) and 12 subscription cases as non-subscriptions (false negatives).

The SVM model's total score of 0.8392 indicates a moderate level of accuracy in subscription behaviour prediction. This score also reflects the model's overall accuracy. The total number of true positives and true negatives divided by the total number of instances yields this accuracy score.

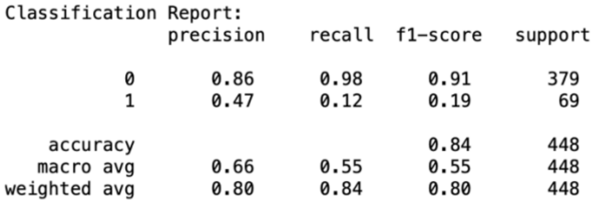


```
▾ SVC
SVC(kernel='linear')
```

The SVM model was modified using kernel tricks, but it is unclear how this affected the model's ability to improve.

Part 4: Comparison of Accuracy

Classification Report for Logistic Regression:

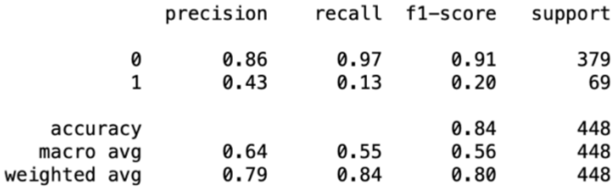


```
Classification Report:
              precision    recall  f1-score   support

           0       0.86      0.98      0.91       379
           1       0.47      0.12      0.19        69

    accuracy                           0.84       448
   macro avg       0.66      0.55      0.55       448
weighted avg       0.80      0.84      0.80       448
```

Classification Report for SVM Model:



```
              precision    recall  f1-score   support

           0       0.86      0.97      0.91       379
           1       0.43      0.13      0.20        69

    accuracy                           0.84       448
   macro avg       0.64      0.55      0.56       448
weighted avg       0.79      0.84      0.80       448
```

In the comparison of the models, it is clear that the logistic regression model outperforms the SVM model in terms of accuracy, precision, and recall for both classes. The logistic regression model achieves a superior accuracy of 0.84375 compared to the SVM model.

In precision, the logistic regression model excels with a precision of 0.86 for class 0 (non-subscription), while the SVM model lags with a precision of only 0.43. Similarly, for class 1 (subscription), the logistic regression model attains a precision of 0.47, surpassing the SVM model's precision of 0.13. This suggests that the logistic regression approach more accurately identifies both non-subscription and subscription instances.

In terms of recall, the logistic regression model surpasses the SVM model, achieving recalls of 0.98 for class 0 and 0.12 for class 1, while the SVM model achieves recalls of 0.97 for class 0 and 0.13 for class 1. This underscores the logistic regression model's superior ability to accurately capture both subscription and non-subscription scenarios.

Considering accuracy, precision, and recall, the logistic regression model exhibits overall superior performance compared to the SVM model. It demonstrates greater precision and a superior ability to distinguish between the two classes, highlighting its advantage in anticipating subscription behaviour.

Part 5: Recommendations

To determine why both models have trouble with recall, particularly when it comes to subscription prediction, more research is required. This could entail addressing potential data imbalances, experimenting with various modelling approaches, or feature engineering.
To possibly enhance model performance, take into consideration applying more sophisticated strategies, such as ensemble methods or investigating non-linear kernels for SVM.
Improving the dataset and working with domain experts to obtain insights into customer behaviour could lead to more accurate predictive models.
Keep in mind that the particular objectives and limitations of the current business problem determine the model and variables to be used. Further improvements can be guided by domain knowledge and experimentation.

Conclusion:

In summary, the logistic regression model exhibits better accuracy, precision, and recall for both subscription and non-subscription instances, outperforming the SVM model in the prediction of subscription behaviour. Significant variables that have been identified, like the presence of teenagers, recentness, wine spending, and catalogue purchases, offer insightful information for marketing strategies that are specifically targeted. In this case, the logistic regression model is the best option for forecasting and comprehending subscription behaviour due to its better performance and significant variable insights. Taking these insights into account can help boost overall subscription rates and maximize marketing efforts.

References

1. Nettleton, D. (2014). Data modeling. In *Elsevier eBooks* (pp. 137–157). https://doi.org/10.1016/b978-0-12-416602-8.00009-1
2. Narkhede, S. (2021, June 15). Understanding Confusion Matrix - towards Data science. *Medium*. https://towardsdatascience.com/understanding-confusion-matrix-a9ad42dcfd62
3. Saini, A. (2024, January 23). *Guide on Support Vector Machine (SVM) Algorithm*. Analytics Vidhya. https://www.analyticsvidhya.com/blog/2021/10/support-vector-machinessvm-a-complete-guide-for-beginners/
4. Wilimitis, D. (2021, December 7). The kernel trick in support vector classification - towards data science. *Medium*. https://towardsdatascience.com/the-kernel-trick-c98cdbcaeb3f