Logistical Regression to Classify Customer Segments

Overview

The aim of our regression analysis is to find which customers are individual buyers or wholesalers based on the quantity of items they have purchased. Customers who buy at least 5 units per product will be eligible for a 3.5% discount under the 'Flora for Business' marketing campaign. The logistic regression classifying algorithm accurately predicted that 81% of the customers are wholesalers. With a recall value of 1.00 we can decipher that the correct positive predictions relative to the total actual positives were identified correctly as wholesalers. Thus, the model will allow the newly created customer center to accurately identify the customers to email and call for the promotion.

Analysis

Aim

A household decoration seller, Flora, wants to classify its customers into two groups of personal users and wholesalers. Their new marketing campaign aims to identify these target segments to launch a new sub-product called "Flora for Business" which will offer a 3.5% discount on selected products when ordering more than five units per product.

Method

For our analysis we used a logistical regression to create distinct categories from the input values given for training. The model uses one or more independent variables to determine a best-fitting relationship between the dependent variable and a set of independent variables (Waseem, 2022).

Confusion Matrix Results

Training data

```
[[608   0]
 [122   0]]
              precision    recall  f1-score   support

           0       0.83      1.00      0.91       608
           1       0.00      0.00      0.00       122

    accuracy                           0.83       730
   macro avg       0.42      0.50      0.45       730
weighted avg       0.69      0.83      0.76       730
```

The F-1 scores of 0.91 shows that the training model classified most of the observations for wholesalers into the correct class. The model had a perfect recall value of 1.00, correctly predicting actual positives. Overall, the model predicted 83% of the classifications correctly.

Test data

```
[[197   0]
 [ 47   0]]
              precision    recall  f1-score   support

  Wholesaler       0.81      1.00      0.89       197
    Personal       0.00      0.00      0.00        47

    accuracy                           0.81       244
   macro avg       0.40      0.50      0.45       244
weighted avg       0.65      0.81      0.72       244
```

The F-1 score of 0.89 for wholesalers shows the preciseness of our model. Since 0.89 is close to 1 (perfect prediction), our model classifies most of the observations for wholesalers into the correct class. With a recall value of 1.00 we can decipher that the correct positive predictions relative to the total actual positives was identified correctly for wholesalers. In all, the model predicted 81% of the classifications correctly.

Advantages of the Classification Algorithm

Our model had a recall score of 1.00 for wholesalers, identifying all actual positives correctly for both the models. The F-1 and accuracy scores were high for both the models adding onto its preciseness.

Limitations of the Classification Algorithm

In our confusion matrix the true negative value is 0, indicating that when the model predicted a non-wholesaler the classification was not accurate; questioning the model's predicting ability. Lastly, the model performed better during training than in testing suggesting that the model is overfitted.

Business Scenarios

TikTok

The video creating and sharing platform is growing with thousands of new users every day. Classifying the current and incoming users into distinct groups based on their contribution towards the growth of the platform, will allow TikTok to cater towards their specific needs. Adapting the application to the users' preference can drive more growth and profit as the platform promotes content that the target audience resonates with.

Identifying risk factors for diseases

Most of the epidemiological and clinical research is based on risk assessment. 'When the outcome variable of interest is dichotomous, a tool popular in assessing the risk of exposure or the benefit of a treatment is a logistic regression (Ismail & Anil, 2014). This classifying algorithm is deemed as the standard method to analyze a patient's potential risk factors by identifying the presence or absence of a disease (Ismail & Anil, 2014). Accurately predicting illnesses will lead to patients returning for other treatments or recommending others to visit the clinic for check ups or emergencies.

Teamwork

Working as a team

Unlike previous courses, we were informed about our teams just a couple of days before the assignment was due. It created a bit of a challenge to get to know each team member, their preferred style of working and accommodating everyone's schedule. However, each member was understanding of the situation and made sure to put their best foot forward and worked efficiently towards the deadline.

Karam and Manuel worked on the SQL queries to extract data and worked diligently to update the code to prevent forming a Cartesian product. Esther and Madhuri worked on the Python regression analysis and the paper. Everyone created a stress-free environment and encouraged each other to slowdown and have fun with the process.

Understanding the assignment

There was a delay in learning the course material required to complete the assignment and creating a great deliverable in the given timeframe. This gap in knowledge motivated each of us to quickly learn skills needed for the business case. Karam and Manuel devoted time to learn about the complex queries, and Esther and Madhuri studied how to use scikit-learn's modeling framework for model development.

References

Ismail, B. & Anil, M. (2014). Regression methods for analyzing the risk factors for a lifestyle disease among the young population of India. *Indian Heart Journal, 66*(6): 587-592. 10.1016/j.ihj.2014.05.027

Waseem, M. (2022, Jan 05). How to implement classification in machine learning? Edureka. https://www.edureka.co/blog/classification-in-machine-learning/#log.