**DSBA/MBAD 6211 Assignment 1 (Logistic Regression – 50 pts)**

**Instructions:** This is an individual assignment. The submitted solution and answers should be your own. The data file for this homework is Organics.csv, which is to be downloaded from Canvas. Use Python to build a logistic regression model. Create a new Word document and save it as Logistic\_xxxx (where xxxx is your ninernet login name). Write your full name on the first page of the Word document. Where required, write your answers or paste screenshots in this Word document. You need to submit both the Word document and Python Code file. **Your Python code should run correctly for your assignment to be graded. Code that generates error will result in loss of points (up to a maximum of 20%)**

**Problem description and questions:** A supermarket offers a new line of organic products. The supermarket’s management wants to determine which customers are likely to purchase these products. The supermarket has a customer loyalty program. As an initial buyer incentive plan, the supermarket provided coupons for the organic products to all of the loyalty program participants and collected data that includes whether these customers purchased any of the organic products. Based on the data collected, the supermarket wants to understand the behavior of their customers and their likelihood of purchase of organic products. The **ORGANICS** data set contains 8 variables as shown in the table below and more than 22,000 observations.

|  |  |
| --- | --- |
| **Variable Name** | **Description** |
| **ID** | Customer loyalty identification number |
| **DemAffl** | Demographics - Affluence grade on a scale from 1 to 30 (higher values indicate more affluent customer) |
| **DemAge** | Demographics - Age, in years |
| **DemGender** | Demographics - M = male, F = female, U = unknown |
| **PromClass** | Loyalty status: tin, silver, gold, or platinum |
| **PromSpend** | Total amount spent in $ |
| **PromTime** | Time as loyalty card member in years |
| **TargetBuy** | Organics purchased? 1 = Yes, 0 = No |

# Variable and model naming requirements:

* Include your ***name initials*** to the data frame names as well as model names in your Python coding. This is required for your work to be graded.
* For instance, my initials are **CS**, and in my coding, I would name the data frames as ***dfCS, dfCS.train***, and ***dfCS.test.*** I would also name the models as ***logisticCS***, etc.

# Tasks and Questions to answer in the Word document:

1. What variable would you consider as the target variable? Explain your reason. (2 pts)

I believe the target variable would be 'TargetBuy' as it is the only variable to return 0 or 1, with logistic regression, we want to predict the outcome of the binary variable based on the other independents in the model.

1. Select any 5 variables to consider as independent variables in the model. Explain the reasons for your selection. (5 pts)

- PromSpend:

- The overall amount of spenifing according to a lifetime, it varies and intially could be thought as good predictor, however it was determined later not to be statisically significant.

- DemGender\_M:

- After dummy coding DemGender into three, DemGender\_M was one of the best to be reported in the model, having extremely low P-values and Z-value that was statistacally significant

- DemGender\_F:

- After dummy coding DemGender into three, DemGender\_F was one of the best to be reported in the model, having extremely low P-values and Z-value that was statistacally significant

- DemAffl:

- How much a wealth a person has on a 30 point scales, shows what economic bracket people would be in to more likely buy the organic line.

- DemAge:

- How old the person will be if they by the organics, could be a good indicator on what age range would try to eat healthier.

1. Are there any variables which cannot be used in your model? Why? (3 pts)

- ID would not make sense as it just serves as a unique identifier for each of the people's data in each row. It uniquely identifying everything would create a perfect model which wouldn’t help in us predicting.

1. What variables need to be dummy coded before you run your logistic regression model? Explain what new dummy coded columns you created. (6 pts)

- DemGender and PromClass would need to be dummy coded as logistic regression cannot take in text values. I did both to run multiple models to determine which value would overall be better as predictors for our target y. This split gender into three: Male, Female, and Unknown which i dropped overall. PromClass was split into four: PromClass\_Gold, PromClass\_Tin, PromClass\_Silver, and PromClass\_Platinum with me dropping tin. You also need to drop one of the variables to make sure multicollinearity doesn't occur (n-1).

1. Do you have to consider missing values in your dataset? How did you handle the presence of missing values, if any? (4 pts)

There were originally, but then I dropped them at the beginning after loading in the data, and then again later to be sure there were none randomly created from the dummies.

#*Calculate the nulls and then drop them*

model\_data\_JB.isnull().sum()

clean\_JB = model\_data\_JB.dropna()

clean\_JB.info()

print(clean\_JB.isnull().sum())

clean\_JB

Data columns (total 8 columns):

# Column Non-Null Count Dtype

--- ------ -------------- -----

0 ID 17272 non-null int64

1 DemAffl 17272 non-null float64

2 DemAge 17272 non-null float64

3 DemGender 17272 non-null object

4 PromClass 17272 non-null object

5 PromSpend 17272 non-null float64

6 PromTime 17272 non-null float64

7 TargetBuy 17272 non-null int64

dtypes: float64(4), int64(2), object(2)

memory usage: 1.2+ MB

ID 0

DemAffl 0

DemAge 0

DemGender 0

PromClass 0

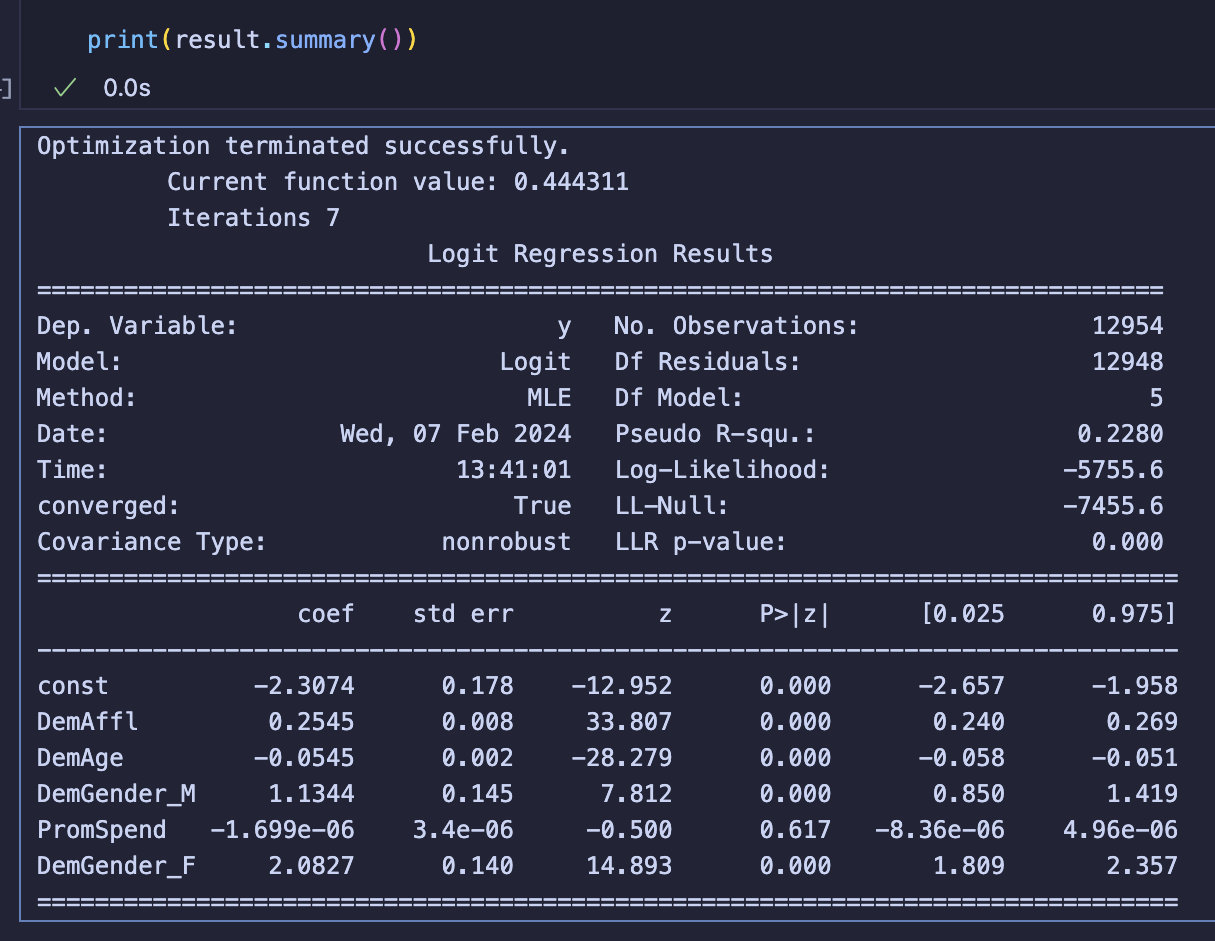
PromSpend 0

PromTime 0

TargetBuy 0

dtype: int64

1. Provide following screenshots from your logistic regression model. (5 pts)
   1. The model result summary, along with the coefficient table



* 1. The classification (confusion) matrix output

A screen shot of a computer program

Description automatically generated

1. Is the overall model statistically significant? Explain how you arrived at your conclusion. (4 pts)

- Overall most of the model is statisically significant with the following variables passing both the P-value and Z-test:

- DemAffl

- DemAge

- DemGender\_M

- DemGender\_F

- The ones that did not past this test are:

- PromSpend

- If I could, I would drop PromSpend completely from the model because to me it doesn't make sense to keep a statistically insignificant predictor in the model. After checking all the other variables (PromTime, PromClass\_Gold, PromClass\_Silver, PromClass\_Tin, PromClass\_Platinum), all were determined to be statistically insignificant. This makes me want to omit the 5th predictor completely.

1. Interpret the impacts of each independent variable on the target variable. (10 pts)

- For every level/point that the Target variable goes up by:

- DemAffl increases by 0.2545 holding other variables constant. This indicates a positive relationship between affluence and the likelihood of purchasing organic product.

- DemAge decrease by -0.0545, holding other variables are held constant. This suggests a negative relationship between age and the likelihood of purchasing organic products.

- DemGender\_M increases by 1.1344 compared to the baseline, holding all else constant, indicatinng a small postive likelihood of purchasing organic.

- Demgender\_F increases by 2.0827 compared to the baseline, holding all else constant. This suggests a stronger positive impact of being female on the likelihood of purchasing organic products compared to being male.

- Promspend would decrease by -3.11e-06 however it would not be considered because its not statistically significant.

1. From the confusion matrix, compute the accuracy, precision, recall and F1 score for this model. (6 pts)

- Accuracy: (2964 + 493)/ (2964 + 180 + 681 + 493) = 0.80060213061

- Precision: 493 / (493 + 180) = 0.73254086181

- Recall: 493 / (493 + 681) = 0.41993185689

- F1 Score: 2∗(0.7324∗0.4199)/(0.7324+0.4199) = 0.5331

1. Summarize your findings to the Director of Marketing for this company. Based on your summary, provide two recommendations that could address the company’s problem described earlier. (Limit this answer to a paragraph of not more than 200 words) (5 pts)

As the Director of Marketing, I would target a younger crowd specifically more towards women but not to the point where it would dis-include men from the marketing campaign. Some ways I could bring them in would be with social media campaigns, store displays, or more personalized marketing to resonate with them.

Another recommendation would be to target the more affluent customers as well, trying to draw in people with more classy campaigns or bring a perception about the food that causes the more affluent customers to notice them.

A third reason (although not really necessary and just my thoughts) would be to see why Spending is not significant to the model and find a way for Organic food to become more viable based on peoples spending strategies and influence perhaps better decisions.