HARISH SAIFI
GL6134
23DSMSA112

LAB 2 - LOGISTIC REGRESSION MODEL FITTING.

Objective

In this lab we are going to perform the Logistic Regression on the data to classify/predict the labels for a consumer to be Loyal or Not Loyal. Since the classification is binary,1 will be assigned as Loyal and 0 for Not Loyal.

Baseline Model

This is the summary section of the Baseline model which includes zero predictors/variables. In other words we are drawing the metrics from the data itself and we are not given any new values.

Classification Table

	Observed		Predicted		
			Loyalty		Percentage Correct
			0	1	
Step 0	Loyalty	0	0	15	.0
		1	0	15	100.0
	Overall Percentage				50.0

We can see that there are 15 Loyal and 15 Not Loyal customers so that makes our baseline model 50% accurate.

Variables in the Equation

		B	S.E.	Wald	df	Sig.	Exp(B)
Step 0	Constant	.000	.365	.000	1	1.000	1.000

The above table shows the summary of the baseline model. Where the Wald statistic is zero.Which means that the regression coefficients are significantly not different from zero or in hypothesis word we are accepting null hypothesis.

Variables not in the Equation					
			Score	df	Sig.
Step 0	Variables	Brand	13.264	1	.000
		Production	.462	1	.497
		Shopping	4.727	1	.030
	Overall Statistics		14.019	3	.003

From the above table we can see that there are two predictors(Brand and Shopping) which will prove to be significant for our Logistic Regression in future.

After Fitting Logistic Regression Model.

Below are the summary tables for different statistics to assess the accuracy of the fitted model. Here we have Cox % Snell R Square and Nagelkerke R Square statistics which tells us about the variation explained by the fitted model.

Model Summary			
Step	-2 Log likelihood	Cox & Snell R Square	Nagelkerke R Square
1	23.471[a]	.453	.604

Our Logistics Model has predicted about 60.4% of the data variability, which is considered to be the appropriate amount to consider that model to be accurate and ready to be deployed for new data or unseen data.

Confusion Matrix For The Model.

Classification Table					
	Observed		Predicted		
			Loyalty		Percentage Correct
			0	1	
Step 1	Loyalty	0	12	3	80.0
		1	3	12	80.0
	Overall Percentage				80.0

The above table is known as a confusion matrix which tells us about the classification done by the model. The sum of the diagonal elements tells us the total correct predictions(sensitivity) while the sum of off-diagonal elements tells us the wrong predictions(specificity).

Overall we have about $(12 + 12)/30 = 0.8 * 100 = 80\%$ accuracy by the model itself which is far better than the Baseline Model.

Variables in the Equation							
		B	S.E.	Wald	df	Sig.	Exp(B)
Step 1[a]	Brand	1.274	.479	7.075	1	.008	3.575
	Production	.186	.322	.335	1	.563	1.205
	Shopping	.590	.491	1.442	1	.230	1.804
	Constant	-8.642	3.346	6.672	1	.010	.000

From the above table we have the Wald statistics for Brand(7.075), Production(0.335), Shopping(0.590) and Constant or Intercept(6.671) are all greater than zero which suggest that all of the values are significantly different from zero.

But we have only one variable which is significant for our model that is “Brand” for predicting the “Loyal” or “Not Loyal” nature of the customer.

The coefficients of the Logistic Regression are as follows:

$$\beta_0 = -\ 8.642$$
$$\beta_1 = \ 1.274$$

Since all other regression coefficients are not significant as per the table so we will not include them in the model.

The Logistic Model will be as follows:

$$P(Y\ =\ 1)\ =\ \frac{1}{1+e^{-(\ \beta_0 + \ \beta_1 X_1 + \ \beta_2 X_2 + \ \cdots\cdots\ +\ \beta_p X_p\ )}}$$

Putting the values of β_0 and β_1 will gives the model which will calculate the probability such that if the probability comes out to be $\geq\ 0.500$ then the corresponding class to that record will be equal to “1” that is “Loyal” otherwise for $<\ 0.500$ the corresponding class to that record will be “0” that is “Not Loyal”.

$$P(Y\ =\ 1)\ =\ \frac{1}{1+e^{-(-8.642\ +\ 1.274\ *\ Brand\ )}}$$