Homework 4

Important Note: Please make sure to include screenshots of SAS codes as well as the output in Word file. TAs are grading many assignments, and going back and forth between your report and SAS file made it difficult for TAs to fully understand your code and outputs.

The SAS dataset HeinzHunts has data on grocery store purchases of Hunts and Heinz ketchup. Each observation corresponds to one purchase occasion (of one of these brands) and consists of the following variables:

1. **Heinz** : =1 if Heinz was purchased, =0 if Hunts was purchased
2. **PriceHeinz** : Price of Heinz
3. **PriceHunts** : Price of Hunts
4. **DisplHeinz** : = 1 if Heinz had a store display, =0 if Heinz did not have a store display
5. **DisplHunts** : = 1 if Hunts had a store display, =0 if Hunts did not have a store display
6. **FeatureHeinz** : = 1 if Heinz had a store feature, =0 if Heinz did not have a store feature
7. **FeatureHunts** : = 1 if Hunts had a store feature, =0 if Hunts did not have a store feature

1. Create a variable LogPriceRatio = log (PriceHeinz/PriceHunts). Also create interaction terms between display and feature for both brands (e.g., DisplHeinz * FeatureHeinz for Heinz, and the same for Hunts).

2. Randomly select 80% of the data set as the training sample, remaining 20% as test sample. Use the seed= option to set random seed to a value of 10.

3. Estimate a logit probability model for the probability that Heinz is purchased – using LogPriceRatio, DisplHeinz, FeatureHeinz, DisplHunts, FeatureHunts, and interaction terms (from part 1) as the explanatory variables.

4. Interpret all parameter estimates. What promotional methods (feature / display) are effective for Hunts? For Heinz? How would you interpret the results for the interaction effects? Interpret all your results in in terms of odds ratio. Refer to lecture notes for examples.

5. Based on the estimated model, and using the logit probability formula, calculate the change in predicted probability that Heinz is purchased if LogPriceRatio changes from 0.3 to 0.4 and Heinz does not use a feature or display, while Hunts uses a feature and a display.

   Recall that in the logit model: $\Pr(Y = 1) = \frac{e^{\beta X}}{1+e^{\beta X}}$, where Y is the outcome variable, X are the predictor variables, and β are the estimated model coefficients.

6. The estimated model is to be used for targeting customers for Hunts coupons to build loyalty for the brand. Coupons are to be sent to customers who are likely to buy Hunts, and not to customers who

are likely to buy Heinz. In other words, the coupons should be sent to customers whose predicted probability of buying Heinz is below a certain threshold level that needs to be determined based on the costs of misclassifications (incorrectly sending / not sending a coupon)

The following information about the costs of incorrect classification is available: The cost of incorrectly sending a coupon to a customer who would have bought Heinz is $1 per customer, and the cost of incorrectly failing to send a coupon to a customer who would have bought Hunts is $0.25 per customer.

Based on these costs, what is the optimal threshold probability level that should be used with the estimated model to decide which consumers should receive coupons.

(HINT: Step 1: Using appropriate SAS commands, create an ROC table for the test data from the estimated model. The ROC table provides the number of false positive and false negative classifications for each possible probability threshold.

Step 2: Using the cost information, calculate the total cost of misclassification for each probability threshold.

Total Cost = # of False Positives * False Positive Cost + # of False Negatives * False Negative Cost

Think carefully as to what is false positive and negative in this context. The definition changes depending on how you set up your question.

Step 3: Choose the probability threshold that leads to the lowest total cost.)