**INDIVIDUAL ASSIGNMENT-4**

1) Revisiting the Conjoint Study for All-in-one Printers (3 Points)Recall the Conjoint Analysis survey on All-in-one printers from Individual Assignment 1. Also recall, that in that survey, ratings data (on a scale of 0 - 100) for the 16 profiles were collected from 35 respondents. The relative attribute importance for the respondents, resulting from their utility functions, are provided to you in the file, Printers\_RAI.csv. Use this data to run a Cluster Analysis (Don’t standardize the variables). Find k-means segments based on individuals’ relative attribute importance. Consider the three-segment solution. Describe the three segments in plain English.

A1)











Based on the segments created, we can see that:

S1: 10 i.e 40%

S2: 9 i.e. 14%

S3: 16 i.e. 46%

The segments, based on their characteristics can be described as follows:

Segment 1 prefers having a brand name and requires the product to have a high speed. Segment one users however are not ready to pay very high prices for the product but are not price sensitive either and hence prefer moderate price. Their keen-ness towards wireless devices is almost negligible. I would call this segment as the Techies who are well informed about new devices and are tech savvy.

Segment 2 is not brand conscious, but they do need a wireless device. Although they are not willing to spend a lot of money to buy this device. They also don’t consider speed to be a deciding factor when buying the product.

Based on their needs and attitudes I would call this segment as price sensitive.

Segment 3 is willing to pay a higher price for a segment. Although they do not have a brand preference or a speed or for the product having a wireless functionality.

Although they don’t have a brand preference, they are still not comfortable with completely unheard-of brands. I would call this segment the indiscriminate segment as they are not choosy and fussy.

2) Recall the ConneCtor PDA problem from Individual Assignment 3. Suppose, based on the Needs of the respondents in the survey (X1 – X10), they were divided into two broad segments. The segment memberships of the respondents, and their demographic information (Z1 – Z5), are provided in the file PDA\_2seg.csv. Discuss how well you can use the demographic variables to discriminate between the two segments (estimate a suitable binary logistic regression model to predict consumers’ segment membership using the Stepwise variable selection approach and report the classification accuracy). Interpret the estimated model.

Note: Split the data set into a ‘Training Sample’ (70%) and a ‘Testing Sample’ (30%).Note: Use any five of the six occupation category dummies as potential inputs in your regression model.

A2) **ESTIMATION OF THE MODEL WITH THE TRAINING DATASET:**

Deviance Residuals:

Min 1Q Median 3Q Max

-1.7827 -0.5252 -0.2124 0.3876 3.0274

Coefficients: (1 not defined because of singularities)

Estimate Std. Error z value Pr(>|z|)

(Intercept) -1.275e+00 2.602e+00 -0.490 0.62421

ID 5.680e-02 1.363e-02 4.168 3.07e-05 \*\*\*

Z1 2.885e-05 2.594e-02 0.001 0.99911

Z2 -1.365e+00 5.046e-01 -2.706 0.00682 \*\*

Z3 -4.884e-02 1.698e-02 -2.876 0.00402 \*\*

Z4 6.520e-01 2.656e-01 2.455 0.01410 \*

Z5\_1 -6.344e-01 1.513e+00 -0.419 0.67498

Z5\_2 1.288e+00 2.043e+00 0.630 0.52852

Z5\_3 3.055e-01 9.077e-01 0.337 0.73643

Z5\_4 -2.126e-02 9.280e-01 -0.023 0.98172

Z5\_5 5.519e-01 8.901e-01 0.620 0.53519

Z5\_6 NA NA NA NA

---

Signif. codes: 0 ‘\*\*\*’ 0.001 ‘\*\*’ 0.01 ‘\*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

Null deviance: 142.104 on 104 degrees of freedom

Residual deviance: 75.142 on 94 degrees of freedom

AIC: 97.142

Number of Fisher Scoring iterations: 6

**STEP WISE SELECTED MODEL SUMMARY:**

Deviance Residuals:

Min 1Q Median 3Q Max

-1.8477 -0.5911 -0.2051 0.3643 2.9506

Coefficients:

Estimate Std. Error z value Pr(>|z|)

(Intercept) -0.98624 2.10372 -0.469 0.63921

ID 0.05627 0.01292 4.355 1.33e-05 \*\*\*

Z4 0.62793 0.25257 2.486 0.01291 \*

Z3 -0.04827 0.01675 -2.881 0.00397 \*\*

Z2 -1.36834 0.48384 -2.828 0.00468 \*\*

---

Signif. codes: 0 ‘\*\*\*’ 0.001 ‘\*\*’ 0.01 ‘\*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

Null deviance: 142.104 on 104 degrees of freedom

Residual deviance: 76.398 on 100 degrees of freedom

AIC: 86.398

Number of Fisher Scoring iterations: 6

**HIT RATIO FOR THE STEP WISE MODEL OF THE TRAINING DATASET:**

0.866667

Accuracy for training is 86.67%

**HIT RATIO FOR THE STEP WISE MODEL OF THE TESTING DATASET:**

0.7333

Accuracy for testing is 73.33%

Thus the model needs to be trained over more iterations and the input parameters need to be adjusted to make sure that the testing accuracy > training accuracy.

**CHECKING FOR THE POTIENTIAL OVER-FITTING IN THE DATASETS:**

**TRAINING**: 0.7428571

**TESTING:** 0.6444444

Thus we can see that after 6 iterations, the training set has a higher accuracy than the testing set. To overcome this it is essential to choose the right factors and create a model that has a better testing accuracy than training.

3) Dermaglow, a skincare spa chain wants to be able to identify its loyal customers (Those who visit only Dermaglow for their skincare needs). They have data for 348 customers who are known to be either loyal or not loyal. The data is available in the file,Dermaglow.csv. It includes the following variables: Loyal: Coded as 1 if customer is known to be loyal and 0 if customer is not loyal

Avg.spent: Avg. amount spent per visit by the customer at a Dermaglow spa (in $)

Intervisit.time: Average time interval (in weeks) between visits to a Dermaglow spa

Mincome: Monthly income of customer (in $1000s)

Rating: Customer’s satisfaction rating for Dermaglow (on a scale of 1-100)

Note: Split the data set into a ‘Training Sample’ ( 70%) and a ‘Testing Sample’ ( 30%).

a) Estimate a binary logistic regression model to predict Dermaglow’s loyal customers using all input variables in the model. Report the classification accuracy in the Training and Testing samples.

b) Now, use the Stepwise Model Selection approach. Report the classification accuracy in the Training and Testing samples for the selected model. Report and interpret the selected

model. Using this model, calculate the predicted probability that a customer that spends an average of $50 per visit, visits once in six weeks, has a monthly income of $6,000, and has a 70% satisfaction rating, will be a Loyal customer.

c)Plot the ROC curves, and report the AUC numbers for the models in 3(a) and 3(b)?

A3)

BINARY LOGISTICAL MODEL :

Call:

glm(formula = Loyal ~ ., family = binomial, data = Derma.train)

Deviance Residuals:

Min 1Q Median 3Q Max

-2.0607 -0.9472 0.5137 0.8962 1.8368

Coefficients:

Estimate Std. Error z value Pr(>|z|)

(Intercept) -10.885397 2.472299 -4.403 1.07e-05 \*\*\*

Avg.spent 0.006561 0.005999 1.094 0.274

Intervisit.time -0.174012 0.043326 -4.016 5.91e-05 \*\*\*

MIncome -0.033550 0.047883 -0.701 0.484

Rating 0.175279 0.035285 4.968 6.78e-07 \*\*\*

---

Signif. codes: 0 ‘\*\*\*’ 0.001 ‘\*\*’ 0.01 ‘\*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

Null deviance: 334.05 on 243 degrees of freedom

Residual deviance: 279.84 on 239 degrees of freedom

AIC: 289.84

Number of Fisher Scoring iterations: 4

Accuracy for:

Training: 0.4549

Testing: 0.7211

Hence we can see that the training accuracy is almost 45.49% while the testing accuracy is 72.11%

STEPMOD SUMMARY:

Call:

glm(formula = Loyal ~ Rating + Intervisit.time, family = binomial,

data = Derma.train)

Deviance Residuals:

Min 1Q Median 3Q Max

-2.1468 -0.9710 0.5332 0.9245 2.0332

Coefficients:

Estimate Std. Error z value Pr(>|z|)

(Intercept) -10.30924 2.36772 -4.354 1.34e-05 \*\*\*

Rating 0.16941 0.03471 4.881 1.05e-06 \*\*\*

Intervisit.time -0.17871 0.04017 -4.449 8.62e-06 \*\*\*

---

Signif. codes: 0 ‘\*\*\*’ 0.001 ‘\*\*’ 0.01 ‘\*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

Null deviance: 334.05 on 243 degrees of freedom

Residual deviance: 281.64 on 241 degrees of freedom

AIC: 287.64

Number of Fisher Scoring iterations: 4

Thus from the accuracy values we can see that the model has been trained well as we are getting a higher testing value over training which is the desired outcome. We can also see that it takes almost 4 iterations to train.

Using this model, we can calculate the predicted probability that a customer that spends an average of $50 per visit, visits once in six weeks, has a monthly income of $6,000, and has a70% satisfaction rating, will be a Loyal customer.

V = -10.885 + 0.006561(Avg.Spent) -0.174 (Intervisit time)-0.03355(MIncome) +0.175279 (Rating)

V = -10.885 + 0.006561(50) - 0.174 (0.1667) - 0.03355(6000) + 0.175279 (70)

V = -10.885 + 0.328 -0.029 – 201.3 + 12.27

V = - 199.616

Thus

P= exp (V)/ 1+ exp(V)

P = 4.0635228e-87

From the model we can also see that Intervisit time and Ratings are not significant factors in determining if a customer is loyal or not as they have extremely low p-values.



The values for Area Under the Curve (AUC) is the same for both 3a and 3b aka 0.5070755.