Week 3 Assignment Solutions (50 Points)

|  |  |
| --- | --- |
| **PROPENSITIES AND ACTUAL CLASS MEMBERSHIP**  **FOR VALIDATION DATA** | |
| **Propensity** | **Actual** |
| 0.03 | 0 |
| 0.52 | 0 |
| 0.38 | 0 |
| 0.82 | 1 |
| 0.33 | 0 |
| 0.42 | 0 |
| 0.55 | 1 |
| 0.59 | 0 |
| 0.09 | 0 |
| 0.21 | 0 |
| 0.43 | 0 |
| 0.04 | 0 |
| 0.08 | 0 |
| 0.13 | 0 |
| 0.01 | 0 |
| 0.79 | 1 |
| 0.42 | 0 |
| 0.29 | 0 |
| 0.08 | 0 |
| 0.02 | 0 |

1. Table below shows a small set of predictive model validation results for a classification model, with both actual values and propensities. Calculate error rates, accuracy, sensitivity and specificity using cutoffs of 0.25, 0.5, and 0.75. (10 Points)

**Below are the predicted values at .25, .5 and .75:**

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Propensity** | **Actual** | **Predicted .25 cutoff** | **Predicted .5 cutoff** | **Predicted .75 cutoff** |
| 0.03 | 0 | 0 | 0 | 0 |
| 0.52 | 0 | 1 | 1 | 0 |
| 0.38 | 0 | 1 | 0 | 0 |
| 0.82 | 1 | 1 | 1 | 1 |
| 0.33 | 0 | 1 | 0 | 0 |
| 0.42 | 0 | 1 | 0 | 0 |
| 0.55 | 1 | 1 | 1 | 0 |
| 0.59 | 0 | 1 | 1 | 0 |
| 0.09 | 0 | 0 | 0 | 0 |
| 0.21 | 0 | 0 | 0 | 0 |
| 0.43 | 0 | 1 | 0 | 0 |
| 0.04 | 0 | 0 | 0 | 0 |
| 0.08 | 0 | 0 | 0 | 0 |
| 0.13 | 0 | 0 | 0 | 0 |
| 0.01 | 0 | 0 | 0 | 0 |
| 0.79 | 1 | 1 | 1 | 1 |
| 0.42 | 0 | 1 | 0 | 0 |
| 0.29 | 0 | 1 | 0 | 0 |
| 0.08 | 0 | 0 | 0 | 0 |
| 0.02 | 0 | 0 | 0 | 0 |

**.25 cutoff calculations:**

|  |  |  |
| --- | --- | --- |
| Classification Matrix .25 Cutoff | | |
|  | Predicted | |
| Actual | 1 | 0 |
| 1 | TP = 3 | FN = 0 |
| 0 | FP = 8 | TN = 9 |

Error = (n1,2 + n2,1)/n = 8/20 = .4 = 40%

Accuracy = 1 – n = 1 - .4 = .6 = 60%

Sensitivity = n1,1/(n1,1 + n1,2) = 3/(3 + 0) = 1 = 100%

Specificity = n2,2/(n2,2 + n2,1) = 9/(9 + 8) = .5294 = 52.94%

**.5 cutoff calculations:**

|  |  |  |
| --- | --- | --- |
| Classification Matrix .5 CutOff | | |
|  | Predicted | |
| Actual | 1 | 0 |
| 1 | 3 | 0 |
| 0 | 2 | 15 |

Error = (n1,2 + n2,1)/n = 2/20 = .1 = 10%

Accuracy = 1 – n = 1 - .1 = .9 = 90%

Sensitivity = n1,1/(n1,1 + n1,2) = 3/(3 + 0) = 1 = 100%

Specificity = n2,2/(n2,2 + n2,1) = 15/(15 + 2) = .8824 = 88.24%

**.75 cutoff calculations:**

|  |  |  |
| --- | --- | --- |
| Classification Matrix .75 CutOff | | |
|  | Predicted | |
| Actual | 1 | 0 |
| 1 | 2 | 1 |
| 0 | 0 | 17 |

Error = (n1,2 + n2,1)/n = 1/20 = .05 = 5%

Accuracy = 1 – n = 1 - .05 = .95 = 95%

Sensitivity = n1,1/(n1,1 + n1,2) = 2/(2 + 1) = .6667 = 66.67%

Specificity = n2,2/(n2,2 + n2,1) = 17/(0+17) = 1 = 100%

1. A large number of insurance records are to be examined to develop a model for predicting fraudulent claims.

Of the claims in the historical database, 1% were judged to be fraudulent. A sample is taken to develop a model, and oversampling is used to provide a balanced sample in light of the very low response rate. When applied to this sample (n = 800), the model ends up correctly classifying 310 frauds, and 270 non-frauds. It missed 90 frauds, and classified 130 records incorrectly as frauds when they were not. (10 Points)

A. Produce the confusion matrix for the sample as it stands.

#classification confusion matrix. We know fraudulent = 1

Explanation: when we get Y = 400 from the entire dataset which are all fraudulent, That means the 1% of the dataset is 400 therefore the size of dataset is X = 400/0.01 and X = 40,000.

##means if .01 of the entire dataset is fraudulent (being 1) is 400 then, what is the size of dataset

The 800 observation dataset to be evaluated by the model and we get the following CM

|  |  |  |
| --- | --- | --- |
| Classification matrix for 800 obs. | | |
|  | Predicted | |
| Actual | 1 | 0 |
| 1 | 310 | 90 |
| 0 | 130 | 270 |

#Misclassification rate = (90 + 130) / 800 = 0.275 = 27.5%

#The model ends up classifying (310 + 130) / 800 = 0.55 = 55% of the records

#as fraudulent (which 130 of them not really fraudulent!!.

B. Find the adjusted misclassification rate (adjusting for the oversampling).

#Now we need to add enough zeros to Y to create a balance dataset. This makes our training dataset 800.

#In the original dataset, fraudulent (1’s) 1% and non-fraudulent (0's) constitute 99% of the

#Number of zeros (number of non-fraudulent) = 0.99 \* 40, 000 = 39600 (total number of non-fraudulent observations)

#classification confusion matrix corrected accordingly:

|  |  |  |  |
| --- | --- | --- | --- |
|  | Predicted Class | |  |
| Actual Class | 1 | 0 | Total |
| 1 | 310 | 90 | 400 |
| 0 | 12870 = (130 x 39600)/400 | 26730 = (270 x 39600)/400 | 39600 |
| Total | 13180 | 26820 |  |

For the FP (False Fraudulent) cell: if 130 of 400 is predicted FP then how much of (40000 – 400) is FP

For the TN (non-fraudulent) cell : if 270 of 400 is predicted TN then how many of 39600 is TN?

C. What percentage of new records would you expect to be classified as fraudulent?

#Misclassification rate = (90 + 12870) / 40000 = 0.324 = 32.4%

miscl.rate2 <- (90 + 12870) / 40000

miscl.rate2

#> miscl.rate2

#[1] 0.324

#The model ends up classifying (310 + 12870) / 40, 000 = 0.3295 = 32.95% of

#the records as fraudulent.

#From the above calculations, we expect 32.95% of the records to be classified as frauds.

1. (10 Points) Do Exercise on page 132 (Table 5.1) and page 135 (Figure 5.2). These two exercises are related. Interpret your work. Avoid unnecessary explanations.
2. (10 Points) The Institute for Statistics Education at *Statistics.com* offers online courses in statistics and analytics, and is seeking information that will help in packaging and sequencing courses. Consider the data in the file Course-Topics.csv, These data are for purchases of online statistics courses at Statistics.com. Each row represents the courses attended by a single customer. The firm wishes to assess alternative sequencings and bundling of courses. Use association rules to analyze these data, and interpret several of the resulting rules.
   1. Get Items frequency
   2. Generate rules with highest lift and supp= 0.01, conf = 0.1
   3. Generate rules with highest lift and supp= 0.01, conf = 0.1 and 0.5

The dataset is “Coursetopics.csv”

See posted R code script file for answers

1. Predicting Boston Housing Prices. (10 Points)

The dataset mlba::BostonHousing contains information collected by the US Bureau of the Census concerning housing in the area of Boston, Massachusetts. The dataset includes information on 506 census housing tracts in the Boston area. The goal is to predict the median house price in new tracts based on information such as crime rate, pollution, and number of rooms. The dataset contains 13 predictors, and the response is the median house price (MEDV). The following Table describes each of the predictors and the response. Do not use the CAT.MEDV!

1. Why should the data be partitioned into training and holdout sets? What will the training set be used for? What will the holdout set be used for?
2. Fit a multiple linear regression model to the median house price (MEDV) as a function of CRIM, CHAS, and RM. Write the equation for predicting the median house price from the predictors in the model.
3. Using the estimated regression model, what median house price is predicted for a tract in the Boston area that does not bound the Charles River, has a crime rate of 0.1, and where the average number of rooms per house is 6?
4. Create a correlation matrix of INDUS, NOX, and TAX. Interpret the result
5. Create a Correlation Matrix and identify correlated predictors (use all predictors except CHAS)
6. Build the model using all attributes except CHAS, INDUS, and AGE
7. Run the model on holdout and display the accuracy.