Week 3 Assignment (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:**

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | **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 |

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.

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

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

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. No points will be given!
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?

Not partitioning the dataset into training and holdout sets means using the entire dataset for model training, without any independent data left for evaluating the model's performance. To build a reliable model, it's crucial to train it using a distinct **training dataset** and then assess its performance using a separate **holdout dataset**. This separation allows us to gauge the model's ability to generalize to new, unseen data, a fundamental step in the modeling process.

1. 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.
2. 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?
3. Create a correlation matrix of INDUS, NOX, and TAX. Interpret the result
4. Create a Correlation Matrix and identify correlated predictors (use all predictors except CHAS)
5. Build the model using all attributes except CHAS, INDUS, and AGE
6. Run the model on holdout and display the accuracy.

See posted R code script file for answers

Load required libraries

if (!require(mlba)) {

library(devtools)

install\_github("gedeck/mlba/mlba", force=TRUE)

}

options(scipen=999)

# datasets

library(mlba)

library(tidyverse)

#Association Rule

Install/packages(“arules”

install.packages("recommenderlab")

library(arules)

library(arulesViz)

library(recommenderlab

# for gains and lift chart

library(gains)

# model building

library(caret)

# for producing a graphical display of a correlation matrix

library(corrgram)

# arranging multiple graphs

library(gridExtra)