****

**BUDT758T  
  
DATA MINING AND PREDICTIVE ANALYTICS**

**Homework 3**

**NAME (in capitals): \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_**

* Please submit on Canvas.
* Your submission should consist of this document (with answers filled in in the appropriate places).
* Please ensure that answers are appropriately numbered and clearly legible.
* In the space below please enter the following text and initial below: “I pledge on my honor that I have not given or received unauthorized assistance on this assignment.”

|  |
| --- |
| HONOR PLEDGE:    YOUR INITIALS: |

This assignment is a continuation of Homework 2. The data is the same and you will follow the same steps for data preparation. In particular continue to use the same seed, and R’s **sample**function to partition the data set. Thereafter you will evaluate performance of your classifier for this data set.

**The Assignment**

The data in the accompanying file “VoterPref.csv” (posted on Canvas) contains data from a survey of random sample of registered voters in a state. The subjects were asked whether they were “For” or “Against” a proposal on the ballot to increase the state sales tax by 0.5%, with the stipulation that the additional tax revenues be spent on education. In addition to their position on the proposition, some additional demographic information is collected. The variables in the data set are:

PREFERENCE “For” or “Against”

AGE Years of age at time of survey

INCOME Annual income in thousands of US dollars

GENDER “M” or “F”

The intent of the survey is to develop a strategy to target individuals for a marketing campaign designed to “get out the vote”.

1. Data Preparation
   1. Read the data set in *R*. For the PREFERENCE variable ensure that “Against” is the success class
   2. Set the seed to 71923
   3. Randomly partition the data set into the *training* and *test* data sets. The proportion of observations in the training data set should be 70%. The remaining 30% of observations should be in the test data set.
2. Run a logistic regression model of PREFERENCE on all independent variables. Use only the training data set for estimating the model.
   1. Use a cutoff of 0.5 and do the classification. Compute the confusion matrix for both in-sample and out-of-sample predictions (using the training and test data sets respectively).
   2. Compute the **sensitivity**, **specificity**, **accuracy**, **error rate**, **PPV**, **NPV**.
   3. Plot the ROC curves for both the training and test data sets on the same graph (distinguishing with different colors). What can you infer from a scrutiny of this graph? In particular, is there any evidence of overfitting?
   4. Plot the accuracy against cutoff for both the training and validation data set.
   5. At which value of the cutoff is training accuracy maximized? What is the maximum accuracy value?
   6. What is the accuracy in the validation data set using the cutoff found in (e)?
3. We use the model estimated in (2), but now include misclassification costs. Suppose that there are no costs or benefits associated with correct classification but misclassifying someone who is “For” as being “Against” has a cost of 4, whereas misclassifying someone who is “Against” as being “For” has a cost of 1.
   1. What value of the cutoff minimizes misclassification cost in the training set?
   2. What is the misclassification cost in the training set? In the test set?
   3. Compare your results with the cutoff obtained in (2).
4. Using the model estimated in (2), plot the training data lift chart as well as the validation data lift chart.   
   *Use different graphs for the two data sets.*
5. Using the model estimated in (2), plot the decile-wise lift charts for both the training and validation data sets. *Use different graphs for the two data sets.*
6. Explain in clear language what you deduce about the performance of the classifier from examining the lift charts in (4) and (5).

**Hints:** You may find it useful to look at the scripts for the beer data that I have posted online. Most likely you will be able to reuse code from there. There are some basic control loops (**for**, **ifelse**) and some useful functions (**whichmax**, **max)**. Look at posted slides on basic R programming for reference.