****

**BUDT758T  
  
DATA MINING AND PREDICTIVE ANALYTICS**

**Individual Assignment 4**

**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: |

The goal of this homework is to guide you through developing Neural Network models for (1) classification, and for   
(2) numerical prediction. You will change the key parameters (the number of hidden layers and nodes) and develop an understanding of how this affects results.

**Data**

The data set for this assignment is the Airlines data set used in Assignment 1.

The data in the accompanying file “Airline data 2.csv” (posted on Canvas) was assembled by Professor Robert Windle of the Smith School with assistance from Oliver Yao. The file contains information on 627 air routes in the United States. A route refers to a pair of airports. Note that some cities are served by more than one airport. In such cases, the airports are distinguished by their 3-letter code. The data was collected for the third quarter of 1996 (3Q96). The variables in the data set are:

1. S\_CODE: starting airport’s code
2. S\_CITY: starting city
3. E\_CODE: ending airport’s code
4. E\_CITY: ending city
5. COUPON: average number of coupons (a one-coupon flight is a non-stop flight, a two-coupon flight is a one stop flight, etc.) for that route
6. NEW: number of new carriers entering that route between Q3-96 and Q2-97
7. VACATION: whether a vacation route (Yes) or not (No); Florida and Las Vegas routes are generally considered vacation routes
8. SW: whether Southwest Airlines serves that route (Yes) or not (No)
9. HI: Herfindahl Index – airlines use this as a measure of market concentration (a larger value indicates greater concentration)
10. S\_INCOME: starting city’s average personal income
11. E\_INCOME: ending city’s average personal income
12. S\_POP: starting city’s population
13. E\_POP: ending city’s population
14. SLOT: whether either endpoint airport is slot controlled or not; this is a measure of airport congestion
15. GATE: whether either endpoint airport has gate constraints or not; this is another measure of airport congestion
16. DISTANCE: distance between two endpoint airports in miles
17. PAX: number of passengers on that route during period of data collection
18. FARE: average fare on that route

We will **not** use the first four attributes (S\_CODE, S\_CITY, E\_CODE, and E\_CITY) in our analysis.

1. Data Preparation
   1. Read the data set in *R*. We will only use the following variables in the analysis: SW, VACATION, SLOT, FARE, DISTANCE, HI, GATE, and PAX.
   2. For Neural Networks all explanatory variables have to be numerical. So convert all factors to binary dummy variables.
   3. Rescale all numerical variables to lie in the 0-1.
   4. Delete all missing observations.
   5. Set the seed to 71923
   6. Randomly partition the data set into the *training* and *test* data sets. The proportion of observations in the training data set should be 60%. The remaining 40% of observations should be in the test data set.

* *Use the “sample” function to partition the data.*

1. You will first build a model to predict what routes Southwest will choose to enter. Run a logistic regression model of SW on all the other variables enumerated in 1(a) above. Use only the training data set for this.
   1. Present the output as **Exhibit A**.
   2. Use a cutoff of 0.5 and compute the classification matrix. What are the error rate, the sensitivity and the specificity of the classifier? (Compute these for both the training and test data.)
2. Build a Neural Network with one hidden layer, and 4 nodes in that layer. Use the default settings of the *neuralnet* function except that, since this is a classification problem, set *err.fct = “ce”* and *linear.output = FALSE*.

* *Set the seed to 13 (the only randomness here is in the choice of initial weights)*
* *Use the following order of variables: SW~VACATION+SLOT+DISTANCE+FARE+HI+GATE+PAX*
  1. Plot the net and attach as **Exhibit B**.
  2. Compute the confusion matrix for both the training and test data sets.
  3. Compute the error rate, sensitivity, and specificity in each case.
  4. Compare your results with the logistic regression classification model in (2).

**NOTE:** Convergence of the algorithm can sometimes be an issue. If that happens, you can consider changing the seed   
above. A different start point often results in convergence (you could also increase the *stepmax* and *threshold* parameters of the function, but I have rarely found this to be helpful). If numerous attempts at this fail, report failure to converge in your answer and attach your code.

1. In this section of the assignment you will explore changing the specification of the hidden layer and determine how this affects the performance of your model.
   1. Compare models with one hidden layer with between 0 to 7 nodes in that layer. Use the error rate as the basis for your comparison. Compare both the training and test error rates.
   2. Now compare models with two hidden layers. Set the number of nodes in the first layer at 4, and let the number of nodes in the second layer vary from 1 to 4. Use the error rate as the basis for your comparison. Compare both the training and test error rates.
   3. Which, among the models examined in 4(a) and 4(b), is the best model? Why?
   4. What do you observe from changing the layers and number of nodes? Summarize your findings.
   5. Using the model that you judge to be the best, plot the ROC curve. On the same graph, plot the ROC curve for the logistic regression model in (2). Use the test data for this. Comment.
2. The Neural Network algorithm can also be used with continuous dependent variables. In this section, you will build a model to predict FARE based on the other variables listed in (1) above. Include SW, but this time as a predictor.

* *ReSet the seed to 13 (the only randomness here is in the choice of initial weights)*
* *Use the following order of variables: FARE~ SW+VACATION+SLOT+DISTANCE+HI+GATE+PAX*
  1. As benchmark, run a multiple regression model for FARE, and compute the RMSE for both the training and test data sets.
  2. Now build a Neural Network with one hidden layer and 4 nodes in that layer. Set *err.fct = “sse”* (this is the default, so really only need to ensure that you do not set this to anything else) and *linear.output = TRUE*. Plot the net and attach as **Exhibit C**.
  3. Compute the RMSE for both the training and test data set, and compare with your regression model in 5(a).
  4. Change the number of nodes as described in 4(a) and 4(b). Only, this time, use the RMSE as the basis for your comparison (for both the training and test error rates). Describe what you find.