****

**BUDT758T  
  
DATA MINING AND PREDICTIVE ANALYTICS**

**Individual Assignment 2**

**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 introduce you to classification concepts. You will develop (1) a linear probability model and (2) a logistic regression model. You will need to create random partitions of a data set, build your model on the training data set and then compute prediction errors using the test data set. There are a couple of helpful hints at the end of the assignment.

**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 (i.e. the class with higher level – e.g. “1” for binary variable)
   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. Exploratory analysis of the *training* data set
   1. Construct boxplots of INCOME and AGE (broken up by values of PREFERENCE). Present the plot as **Exhibit A**. What do you observe?
   2. Construct a table for PREFERENCE showing proportions for and against.
   3. Construct a two-way table for count of PREFERENCE broken up by GENDER (i.e. what are the numbers of men and women who are for and against the proposition).
3. Run a linear regression model of PREFERENCE on the demographic variables. Use only the training data set for fitting the model.
   1. Compute the average error, RMSE and the mean absolute error (MAE) for both in-sample predictions (i.e. for the training data set) and the out-of-sample predictions (i.e. for the test data set). Use predicted values from the regression equation (do **not** do the classification for this yet).
   2. For which data set are these errors smaller?
   3. Use a cutoff of 0.5 and do the classification (i.e. make the class predictions). What proportions of predicted classes are for and against in each data set?
   4. What proportion of class predictions are in error in each of the training and test data set?
4. Run a logistic regression model of PREFERENCE on the demographic variables. Use only the training data set for this.
   1. Present the output as **Exhibit B**.
   2. Provide a precise interpretation of the coefficient of AGE.
   3. Provide a precise interpretation of the coefficient of the gender variable.
   4. Use a cutoff of 0.5 and do the classification. What proportion of predicted classes are in error (in the training and test data set)?
   5. Compare these error rates with those in question 3d (linear regression).
   6. Compute the predicted probability for voting *against* the proposition for an individual who is a female, is 36 years old, and has an income $70,000.

**Hints:** You may find the *R***ifelse** function convenient for classification. Finally, the **predict** function that was used for regression will also work for the logistic case. Note however that, by default, it will give you the predicted logit. If you pass it an additional argument (type = “response”) you will get predicted probabilities. E.g.

**p <- predict(fit, newdata=df, type = "response")**