St Lawrence College, School of Business

Graduate Certificate Program, Business Analytics

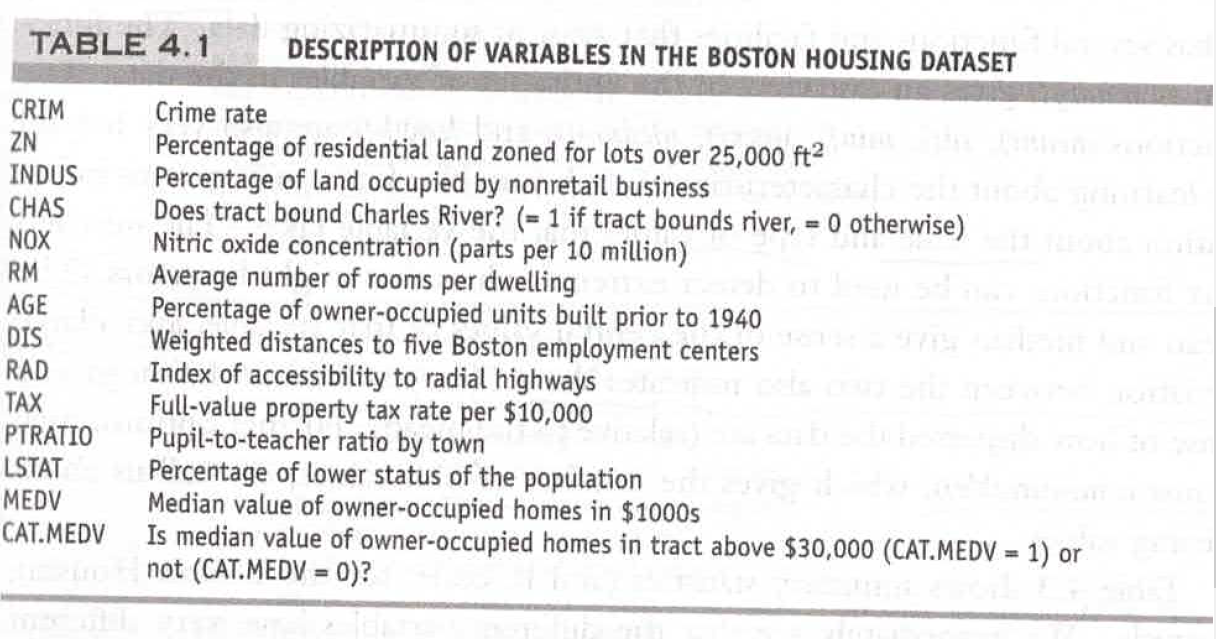
ADMN5003: **Data Acquisition, Analysis and Modelling Techniques** (Fall 2022)

12 Marks

* Open book assignment to be completed independently or in a group of 2
* Submit a R file on BlackBoard in the folder of “Assignment 4”. Name it as “assign4R\_firstname1\_and\_firstname2”. Write all your answers and coding in the same file. Clearly label question numbers.
* Late submission penalty is 10% per day

\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

1. Use “BostonHousing.csv” to predict the median house price in new tracts based on information such as crime rate, pollution, number of rooms and distance to employment centers. The dataset contains 13 predictors, and the response is the median house price (MEDV). (Total 5.5 marks)



1. Partition the data. Use training data to build the following 2 models. (0.5 mark)
2. Fit a **multiple linear regression** model to the median house price (**MEDV**) as a function of CRIM, CHAS, RM and DIS. (0.5 mark) Which predictors are significant? (0.5 mark)
3. Find the best model of the above model using the “backward” method. (0.5 mark) Write the equation of the best model. (0.5 mark) output variable = intercept + coeff 1 X1, coeff 2 X2…)
4. Use correlation matrix to discuss the relationships among INDUS, NOX and TAX. (0.5 mark) 1 bonus mark if you can use one of the heatmaps we learned in class.
5. Fit a **logistic regression model** to the categorized median house price (**CAT.MEDV**) as a function of CRIM, CHAS, RM and DIS. (0.5 mark) Which predictors are significant? (0.5 mark)
6. List 3 differences between multiple linear regression and logistic regression. (1.5 marks)

|  |  |  |
| --- | --- | --- |
|  | Multiple linear regression | Logistic regression |
|  |  |  |
|  |  |  |
|  |  |  |

1. Use “BostonHousing.csv” 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). (Total 3 marks)
2. Perform a k-NN prediction with all 12 predictors (the output variable is MEDV, ignore the CAT.MEDV column), trying values of k from 1 to 5. Partition your data. (1 mark) Normalize the data. (0.5 marks) Choose function knn() from package **class** rather than package **FNN**. (0.5 mark) To make sure R is using the class package when both packages are loaded, use class::knn().
3. What is the best k? (0.5 mark)
4. What does the best k mean (i.e. how many nearest neighbors shall be used to determine the predicted value of MEDV, and how is the value determined)? (0.5 mark)
5. Neural Nets (Total 3.5 marks)
6. A neural net typically starts out with random coefficients; hence, it produces essentially random predictions when presented with its first case. What is the key ingredient by which the net evolves to produce a more accurate prediction? (0.5 mark)
7. Use data “ToyotaCorolla.csv” to predict the price of a used Toyota Corolla based on its specifications. Use the three predictors Age\_08\_04, KM, and Fuel\_Type to build a neural network model. (1 mark) Convert the categorical variable to dummies. (0.5 mark) Scale the numerical predictor and outcome variables to a 0-1 scale. (0.5 mark) Fit a neural network model to the data. Use a single hidden layer with 2 nodes. (0.5 mark) Plot the neural network. (0.5 mark)