Advanced Business Application Development/ Advanced Programming Application Development

Fall 2022

**Final Exam**

**Points: 140**

**Due: 10 AM December 4**

**There are five major questions in this exam. Make sure you answer each of these questions clearly.**

**Format of your answer:**

For short answer questions (e.g., 1.2) that ask your insights, please provide your insights as clear as possible.

For questions (e.g., 1.3) that ask you to implement the code to analyze, please submit your **R RAW CODE Files (i.e., R script file)** and copy and paste the executed results from your Console. For questions that ask you provide further insights based on your analysis results, you could provide your thoughts right after your analysis results.

**You should submit two types of files**: (1) R raw code files (you should submit five individual R raw files (i.e., Question 1, Question 2, Question 3, Question, 4, and Question 5), and (2) a completed word sheet with the corresponding answers. For those R files, you should also note your codes to indicate what your codes are referring to (e.g., #Q 1.1, #Q1.3). Ensure that your codes are executable so I could assess your answer.

I will take 30 points off directly if you miss to submit any required file.

Note: This is the exam. I will **NOT** answer and review any question regarding your code. I will **ONLY** answer if you need to clarify the meaning of the question.

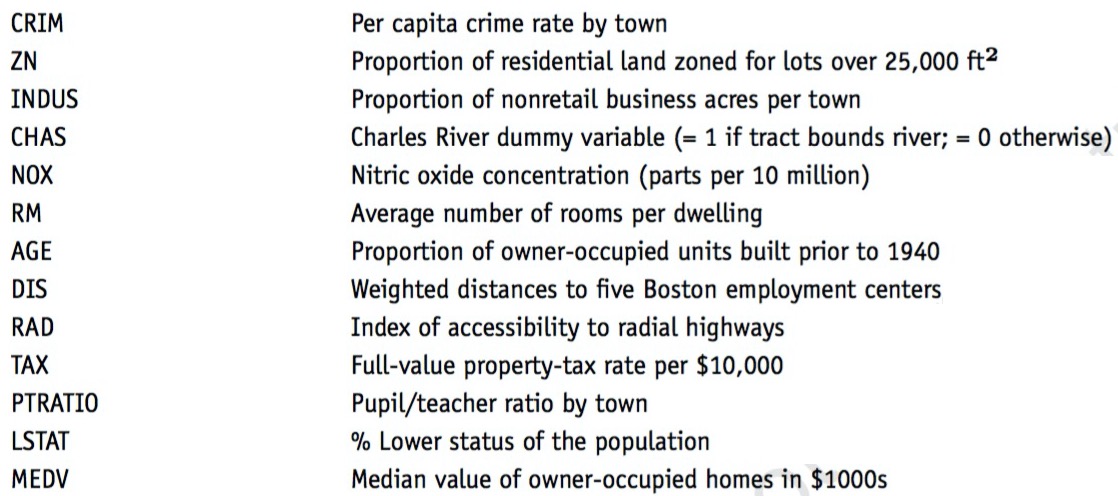
**No Late Submission is allowed. No Exception.**

**Good Luck!**

**Question 1: Predicting Boston Housing Prices. (25 points)**

The file ***BostonHousing.csv*** contains information 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). Table 1 describes each of the predictors and the response.

**Table 1. Description of variables for Boston housing example**



1.1 Compute the correlation table for all variables and show which variables has the strongest positive and which variable has the strongest positive and which variable has the strongest negative correlations with the median house price (MEDV). (1 points)

* 1. Why should the data be partitioned into training and validation sets? What will the training set be used for? What will the validation set be used for? (3 points)
  2. Partition records into 60% for training and 40% for validation sets. Then fit a multiple linear regression to MEDV as a function of CRIM, CHAS, and RM for training sets and show the summarized regression results. Based on your regression results, make a prediction on your validation sets. Be sure to write the equation for predicting the median house price from the predictors in the model and interpret your regression results. (8 points)
  3. Fit another multiple linear regression model to the median house price (MEDV) as a function of LSTAT, INDUS, and NOX for training sets and show the summarized regression results. Then, make a prediction based on this model for your validation sets. (5 points)

* 1. Now get the accuracy metrics for these two predicted models and show their accuracy metrics respectively. Based on their accuracy metrics, select which model is the best model in terms of the accuracy metrics, and why. (8 points)

**Question 2: Financial Condition of Banks (30 points)**

The file ***banks.csv*** includes data on a sample of 20 banks. The “Financial Condition” column records the judgment of an expert on the financial condition of each bank. This outcome variable takes on of two possible values –weak (1) or strong (0)– according to the financial condition of the bank. The predictors are two ratios used in the financial analysis of banks: TotLns&Lses/Assets is the ratio of total loans and leases to total assets and TotExp/Assets is the ratio of total expenses to total assets. The target is to use the two ratios for classifying the financial condition of a new bank.

2.1 Partition records into 60% for training and 40% for validation sets. Then fit a logistic regression to Financial\_Condition as function of TotLns&Lses/Assets and TotExp/Assets on training sets and show summarized logistic results. (10 points)

2.2 Now based on the logistic regression results, make predicted probability for validation set and show their predicted probability. (10 points)

2.3 Now, create a data frame for the first 5 actual records of validation sets along with their predicted probability. (10 points)

**Question 3: Predicting Delayed Flights (30 points)**

The file ***FlightDelays.csv*** contains information on all commercial flights departing the Washington, DC area and arriving at New York during January 2004. For each flight, there is information on the distance of the route, the scheduled time and date of the flight, and so on. The variable that we are trying to predict is whether or not a flight is delayed (Fight\_Status). Table 2 describes variables in this file.

**Table 2. Description of variables for Flight Delays example**

|  |
| --- |
| **Variable**  **Definition** |

CRS\_DEP\_TIME Scheduled departure time

CARRIER The airline

DEP\_TIME Actual departure time

DISTANCE Flight distance in miles

FL\_DATE Flight date

Weather Whether the weather is inclement (1) or not (0)

DAY\_WEEK Day of week (1= Mon, 2=Tus, 3=Wed….)

DAY\_OF\_MONTH Day of month (1= the first day of month; 2= the second day of month….)

Flight\_Status Whether the flight was delayed or on time (defined as arriving within 15 min of scheduled

time)

|  |
| --- |
|  |

3.1 Partition records into 60% for training and 40% for validation sets. Fit a classification tree to the flight delay variable using all the relevant predictors in *FlightDelays.csv* on training sets with maximum of 8 levels and set up cp =0.001 and then plot the tree. (5 points)

**Note**: cp refers to complexity parameter

3.2 In the setting of decision tree, there is a technique called pruning the tree. Discuss the purpose of pruning the tree and why we may need to prune the tree.

Finally, then prune 3.1. tree and plot the pruned tree. (10 points)

3.3 Fit a new classification tree to the flight delay variable using all the relevant predictors on training sets, excluding the Weather predictor. Set cp=0.001 and maximum =6. Plot this new classification tree. (10 points)

3.4 Based on the tree from 3.3, do predictions for both training and validations sets and report their confusion matrix respectively. (5 points)

**Question 4: Cosmetics Purchase (20 points)**

The file ***Cosmetics.csv*** contains 1000 transaction information about cosmetics (1: purchased; 0: no purchased)

* 1. Draw an item frequency plot and answer which two are the most popular items. (2 points)
  2. Build an association rule model and set the support value as 0.01 and the confidence value as 0.1. Based on your association rule results, show the first eight rules and sort by their lift values and ensure to interpret your results. (9 point)
  3. Now, in your 4.2 results, you can see “support”, “confidence”, and “lift”. Please explain the meaning of support, confidence, and life and discuss their implications in the setting of association rules. (9 point)

**Question 5: Department Store Sales Time-Series (35 points)**

The file ***DepartmentStorSales.csv*** contains data on the quarterly sales for a department store over a 6-year period from 2000 to 2005

* 1. We discuss four major components of time series (level, trend, seasonality, and noise).

Discuss the meaning of these four major components. (3 points)

* 1. Show the data in a time-series format and create a well-performed time plot of the data. (2 points)
  2. Decompose your time-series data into four major components (level, trend, seasonality, and noise) and plot these four components individually. Ensure to and explain your observations about different trends from your plots. (6 points)

* 1. We discuss different trend models in time serious, including the linear trend regression model, exponential trend model, and polynomial trend model. Discuss and compare these three models. (4 points)
  2. Partition the first 18 records as training sets and the rest of 6 records as validation sets. Fit a **linear trend model** on training sets. Based on this linear trend model, make a forecast for validation sets and show the results. (7 points)
  3. Fit a polynomial trend model with seasonal effects on training sets. Based on this model, make a forecast for validation sets and show the results. Check accuracy measures of the forecasted model and show these accuracy measures. (8 points)
  4. Finally, based on your decomposition of four major components from 5.3, compare your 5.5 and 5.6 models and discuss which model is more appropriate in this particular setting and why. (5 points)