Week 5 Individual Assignment

Welcome to your week 5 individual assignment. Remember this is an **individual assignment** to be completed alone. It is important that you develop your own data mining and programming skills, part of which is to learn how to use online resources to solve your problems. Facing questions that cannot be answered using the code in the tutorials is a great opportunity to develop important programming and data analysis skills. Search for solutions on your own for instance on Google or ChatGTP on what you are trying to do (e.g., “how to delete a variable in R”).

Upload your R script directly into the assignment task on Blackboard by **Monday, February 12th at midnight**.

Preparation

To answer the assignment questions, complete the following steps (code below).

1. **Create a new R script document for your assignment named `wk5_assignment.R`**. The assignment will not run on the grading system if you use a different name.
2. Load the `HousingBoston_train.csv` into a **variable called `dat_train`**. Make sure that **categorical variables (and strings of text) are treated as factors**. Do not change the file name.
3. Load the dataset called **`HousingBoston_test.csv`** into a variable `dat_test`. Make sure that **categorical variables (and strings of text) are treated as factors.** Do not change the file name.
4. **Change factors into binary** to help interpreting your plots and logistic regression output. R is assigning values 1 and 2 for our `mvalue` variable when set to factors. For the interpretation of the logistic regression, and especially its plots, it is easier to be the target variable of a binary class with value 0 and 1. In this way, the plots relate to the probabilities between class 0 and 1.

```r
# Load train and test dataset
dat_train <- read.csv("HousingBoston_train.csv", stringsAsFactors = TRUE)
dat_test <- read.csv("HousingBoston_test.csv", stringsAsFactors = TRUE)

# change factors to binary with values of 0 and 1
dat_train$mvalue <- 1*(dat_train$mvalue == "below")
dat_test$mvalue <- 1*(dat_test$mvalue == "below")
```

Very important: Your script is going to be directly graded based on the R code

To ensure that your R script can be interpreted, you need to correctly format your script and label your solution values:

- Every solution needs to be stored in a new variable called: **“q+question number.”** For example:
  - If question 7 is “What is 5 - 2?”, then you would write a line in your script `q7 <- 5 - 2`
  - If question 11 is “What is the average of the”cars” variable ?“, then you would write a line in your script `q11 <- mean(dat$cars)`
- For **multiple choice questions** the answer value is the number written beside the option.
  - For example if the right answer for question 17 is the second option, write `q17 <- 2`
  - For questions with multiple answers, use the combine function, For example, if the answer for question 18 is both the first and second options, write `q18 <- c(1, 2)`

- Make sure to **NOT include `setwd()`**. Your local path will be different to the blackboard R installation. I highly recommend setting up a R project that will manage the working directory automatically.
- **Do not round** any numbers.
- **Do not use the `view()`** command.
- **Do not `set.seed()`** unless instructed.
- **Keep percentages unconverted**, as a number between 0 and 1. In other words, 100% equals to 1 and 50% would be written as 0.5.
- It is **critical that you do not install or use any packages other than those specified in the instructions**. We cannot guarantee that other packages will run on the server. If you decide to take the risk, make sure that you include a line where you install all packages that are not part of instructions!

In addition, **you must include the R code to support your responses**. For example, if you have to analyze a plot and answer multiple-choice questions, you must include the plot code as well as your selected multiple-choice response. Do not store plots into a variable just showing how you execute it is fine.

Please post any questions to the Teams channel. Each question below is worth 1 point.

Questions

As in the last assignment, we are going to study Boston house prices for a local real estate agency. They are interested in whether houses sell above or below the median house price value based on economic indicators.

The client provided us with a description for each of the variables. You can get the names of the variables in the dataset using`names()`. **Make sure you familiarize yourself with the the adjusted `mvalue` interpretation as this is important to interpret the output of your analysis (and answers for this assignment).**

- **crim**: Per capita crime rate by town
- **zn**: Proportion of residential land zoned for lots over 25,000 ft^2
- **indus**: Proportion of nonretail business acres per town
- **chas**: Charles River dummy variable (1 if tract bounds river, 0 # otherwise)
- **nox**: Nitric oxide concentration (parts per 10 million)
- **rm**: Average number of rooms per dwelling
- **age**: Proportion of owner-occupied units built prior to 1940
- **dis**: Weighted distances to five Boston employment centers
- **rad**: Index of accessibility to radial highways. Larger index denotes better accessibility (1-8,24)
- **tax**: Full-value property-tax rate per $10,000
- **ptratio**: Pupil/teacher ratio by town
- **lstat**: Percentage lower status of the population
- **mvalue**: 1 (=Below) if the median home value is below the overall median, otherwise 0 (=Above)

The real estate manager has indicated to you that he suspects the number of rooms in the building is one of the key determinants of the house price. Therefore, we decide to investigate the relationship between value and the number of rooms.

So far we often created plots where we plotted 2 variables against one another. The `plot()` function also allows you to input a single variable, which plots the index against the variable. Let's first create a simple plot of the variable `rm`, where `rm` is on the y-axis and the x-axis are observations (row number). We color the points based on the `mvalue` to highlight the cases where the price is below the median (class 1). Note that in R, the color corresponding to the value 1 is black. Therefore, we turn the target value inside the plot function back into a factor.

```r
# Plot Average number of rooms per dwelling
# To ensure correct color coding we revert to factors
plot(dat_train$rm, col = as.factor(dat_train$mvalue),
     ylab = "Average number of rooms per dwelling", xlab = "Observations")
legend('bottomright', legend = unique(as.factor(dat_train$mvalue)),
       pch = 1, col = unique(as.factor(dat_train$mvalue)))
```

1. What does the plot tell you about the potential use of the number of rooms (`rm`) as a predictor for the home price (`mvalue`)? (multiple-choice, single value)
   - □ The number of rooms is a good predictor for classification. A large number of observations of class 0 (above median home value) is associated with smaller rooms. (value 1)
   - □ The number of rooms is a good predictor for classification. A large number of observations of class 1 (below median home value) are associated with smaller rooms. (value 2)
   - □ The Number of rooms is a bad predictor for classification. There is no clear tendency of class separation. (value 3)
   - □ The number of rooms is a bad predictor for classification. Only a few observations of class 1 (below median home value) are associated with smaller rooms. (value 4)

Next, let us run a logistic regression model on the training data with the target variable `mvalue` against `rm` (numbers of rooms) as you have learned in the tutorial.

2. What coefficient do you get for $\hat{\beta}_1$?

From that model estimates and its coefficients, the manager is interested in finding out the probability of a house with 5.5 rooms selling for a price below the median house price.

3. What is the probability for a house with 5.5 rooms to be **below** the median house price?

Although this number sounds plausible to the manager, he doubts how well this model predicts. We decide to investigate the performance through a truth table (aka confusion matrix) on the test set. The manager tells us they are particularly interested in understanding when there is a 20% chance of the house selling below the median price.

4. How many predicted house prices are correctly classified as BELOW median using a threshold of 0.2 on **the test set** using `rooms`?

5. What is the corresponding misclassification error at a threshold of 0.2 on **the test set**?

When the manager hears about the misclassification error, he starts to doubt whether that threshold (of around 20% probability of a house being sold below the Median price) is the right choice. He tells us they would really like to have a small error in the below-average objects.

6. What threshold should we use to predict the most BELOW median house prices (class = 1) correctly from your `rm` model? (multiple-choice, single answer)
   - □ The initial choice is correct; he should use a low threshold, such as 0.2, for the best prediction. (value 1)
   - □ The initial choice is wrong; he should use a high threshold, such as 0.8, for the best prediction. (value 2)
   - □ The initial choice is wrong; he should use a threshold of 0.5 for the best prediction. (value 3)
   - □ None of the above. (value 4)

While the number of rooms seems to give us a pretty good separation of prices, the real estate managers suggest that we should also not forget about the age of a building.

Build another logistic regression model. This time regressing `age` and `rooms` on the target `mvalue` on the **training data set**. (multiple-choice, multiple answers)

7. What can you interpret the model summary (coefficients) and their significance? (multiple-choice, multiple answers)
   - □ Holding rooms constant, one unit change in age will decrease the log-odds of the price being below median by 0.04. (value 1)
   - □ Holding age constant, one unit change in rooms will decrease the log-odds of the price being below median by 2.42 (value 2)
   - □ At the average age and number of rooms, the log-odds of the price being below median is 2.46. (value 3)

- □ Both predictors are highly correlated, and we cannot trust the model coefficients. (value 4)
- □ Both predictors are highly significant in predicting median house price. (value 5)

8. How many predicted observations of house prices are correctly classified as BELOW median using a threshold of 0.2 on the **test set** using `age` and `rooms`?

We also want to visualize the decision boundary of the new model. For this, create a plot of `rm` against `age` and draw three decision boundaries, for threshold 0.2, 0.5 and 0.8, respectively.

9. What does changing the decision boundary impact? (multiple-choice, multiple answers)
   - □ Lowering the threshold shifts the decision boundary line to the left (x-axis being the number of rooms). (value 1)
   - □ Increasing the threshold shifts the decision boundary line to the left (x-axis being the number of rooms). (value 2)
   - □ A threshold of 0.2 will almost perfectly classify all true below the median price in the training set. (value 3)
   - □ A lower threshold leads to classifying objects with fewer number of rooms and lower age as being below the median price. (value 4)
   - □ A lower threshold leads to classifying objects with a higher number of rooms and higher age as the higher number as being below the median price. (value 5)
   - □ Decision boundaries cannot be drawn with two predictors (features). (value 6)

Even though the manager seems pretty happy with the second model, we believe that we can push the model performance even further by including all variables.

Create a third logistic regression model. This time **regressing all available predictors** against `mvalue` on your **training data set**.

10. What is the misclassification error improvement on the **test set** from your **full model** (all variables) over the one **containing only the room variable** at a threshold of 0.2 (i.e., $MisclError_{full} - MisclError_{rm}$)?