Assignment 1
(Due Tue 9/21)

ISOM 674, Fall 2021

The data we will use for this assignment is data on the sale price of houses in a Midwestern city in the U.S. The dataset has 2930 observations.

The data came from home sales during the period 2006 to 2010. Because I do not want to add the complication of time series into this assignment, I have "de-trended" the data and adjusted all of the prices so that you can treat them as all having occurred in May of 2010. To help make sure that there is no confusion about using time, I have also deleted the date of sale variables from the dataset.

The data files are as follows:

- The data dictionary for the dataset:
  GradedHW1-DataDocumentation.txt:
- The entire data set:
  GradedHW1-All-Data.csv:
- Training, validation and test data sets based on a 50%/25%/25% split:
  GradedHW1-Train-Data.csv
  GradedHW1-Validation-Data.csv
  GradedHW1-Test-Data.csv

All of these files are comma delimited. Missing values are shown by blanks.

Questions:

1. The goal is to randomly split the entire data set into training, validation, and test samples using only single family homes and then using a 50%/25%/25% split.

   - First remove all observations except those that are single family homes.
   - Next, randomly draw the training sample. If the training sample fraction does not result in an integer number of observations, round up.
   - From the remaining observations, randomly draw the validation sample. Again rounding up if the split fraction does not result in an integer value.
   - Finally, use the remaining observations for the test sample.

   Sample code to do this has been given to you in the file HW1-Q1-Sample.r. This code may or may not be correct.

   Answer the questions on the Google form at the link provided.

For the next problems, use only the following x-variable to predict the y-variable SalePrice:

Lot.Area
Total.Bsmt.SF
Gr.Liv.Area
Full.Bath
Bedroom.AbvGr
Building Age

Note that you will have to compute the Building Age (in years) as of 2010.

Also, please use the training, validation, and test samples I have provided (not the ones you created in question 1. We will also be restricting our attention to only single family homes.

2. In any analysis that you do, you should always examine the individual variables (both the x's and the y). Examine the variables. Answer the questions on the Google form.

3. Without transforming or standardizing the variables, fit a k-NN regressions to the data for $k = 1, 2, \ldots, 40$. Make of plot of the $\sqrt{\text{MSE}}$ calculated from the validation data against k. Answer the questions on the Google form. Turn in the plot of the $\sqrt{\text{MSE}}$ vs. k as instructed below.

4. Determine the best k (in question 3) and then determine the $\sqrt{\text{MSE}}$ using the test data.

5. Repeat question 3 but standardize the variables. Do not transform the variables, just standardize them. Again, answer the questions on the Google form and turn in the plot of the $\sqrt{\text{MSE}}$ vs. k as instructed below.

6. Repeat question 4 using the results from question 5 (untransformed but standardized variables).

7. Next transform the variables as appropriate and refit the k-NN regression without standardizing the variables. That is, repeat question 3 but using transformed variables. Turn in the plot of the $\sqrt{\text{MSE}}$ vs. k as instructed below.

8. Now determine the best k in question 7 and then determine the $\sqrt{\text{MSE}}$ calculated using the test data. (You are repeating question 4 but using the transformed data.)

9. Repeat question 7 using standardized transformed variables. Turn in the plot of the $\sqrt{\text{MSE}}$ vs. k as instructed below.

10. Repeat question 8 using standardized transformed variables.

11. Which of the 4 "best" models that you determined in questions 4,6, 8, and 10 is the best overall model? Why?

Instructions for Submitting the Plots

In addition to answering the questions on the Google form, you need to turn in 4 plots of the $\sqrt{\text{MSE}}$ vs. k. Please turn these plots in using a pdf file with the following file name:

HW1-Plots-EmoryNetID.pdf

Please substitute your Emory Net ID for "EmoryNetID" in the file name above.

This pdf file should begin with your name and have two plots per page (so the pdf file will be a total of 2 pages). Please make sure that the plots are large enough to be easily legible (i.e., fill or nearly fill the width of the page) and are labeled with both the question number and a descriptive title.

One way to do this is to create a MS Word document (with the file name format as above), add your name and then copy the graphs from R to the Word document. You can create the labels either in R (e.g., using the title( ) function) or add them in the Word document. The export the Word document as a pdf.