**Alex Riles**

**DSCI 425 – Supervised Learning (62 pts.)**

**Assignment 3 – Neural Networks for Regression**

**predicting strength of concrete**

Concrete is the most important material in civil engineering. The concrete compressive strength is thought to be a highly nonlinear function of age and ingredients.

**Variable Information:**  
Given below are the variables contained the file **Concrete.csv** on course website. These data come from a collection of 17 experiments where the compressive strength (MPa) of concrete was determined under different formulations and length of curing (days). These data consist of n = 1030 observations on nine variables (8 predictors and 1 response). There are no cases with missing values!  
  
Name / Data Type / Description/Measurement Units (red denotes variable has zeroes)

* Cement () - continuous – kg of cement per cubic meter of concrete
* Blast Furnace Slag () - continuous – kg of slag per cubic meter of concrete
* Fly Ash () - continuous -- kg of fly ash per cubic meter of concrete
* Water () - continuous -- kg of water per cubic meter of concrete
* Superplasticizer () - continuous -- kg of superplasticizer per cubic meter of concrete
* Coarse Aggregate () - continuous -- kg of course aggregate per cubic meter of concrete
* Fine Aggregate () - continuous -- kg of fine aggregate per cubic meter of concrete
* Age - discrete – age of concrete measured in days (1-365)
* Concrete compressive strength - continuous – compressive strength in *megapascals* ()

Data source: I-Cheng Yeh, "*Modeling of strength of high performance concrete using artificial neural networks*," Cement and Concrete Research, Vol. 28, No. 12, pp. 1797-1808 (1998)

1. Develop a neural network for these data using the nnet package in R. Use some form of cross-validation to choose an “optimal” neural network model fit to these data. **Explain you model development process including supporting R code/results.** Include a plot the predicted and actual values from your neural network model, both in the transformed and untransformed scales, assuming you used a transformation of the compressive strength of the concrete. (15 pts.)

concrete.nnet <- nnet(Strength ~., data = concrete, size = 10, lineout = T, skip = T, maxit = 1000, decay = .001)

con.x <- concrete[,-1]

con.y <- concrete[,1]

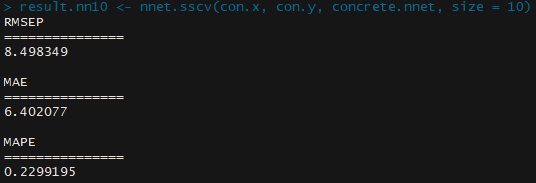
result.nn10 <- nnet.sscv(con.x, con.y, concrete.nnet, size = 10)

result.nn5 <- nnet.sscv(con.x, con.y, concrete.nnet, size = 5)

result.nn3 <- nnet.sscv(con.x, con.y, concrete.nnet, size = 3)

trendscat(concrete$Strength, fitted(concrete.nnet))

cor(concrete$Strength,fitted(concrete.nnet))^2



With the code above for running the model without transforming the data. I found a MAPEP of 22.992% with the size being set at 10. I didn’t test size 5 and 3 but they both had a MAPEP higher than size ten. When looking deeper into size 10 there wasn’t any values at seemed out of the norm of the other validations done. The range was 13 to 34 with could be high but the values seemed normal looking at the numbers.

con.train <- concrete

con.bc <- preProcess(con.train, method = "BoxCox")

con.bc$bc

con.train$Strength <- con.train$Strength^(6/10)

con.train$Cement <- con.train$Cement^(1/5)

con.train$Water <- con.train$Water^(8/10)

con.train$CourseAgg <- con.train$CourseAgg^(11/10)

con.train$FineAge <- con.train$FineAge^(18/10)

con.train$Age <- log(con.train$Age)

con.train.nn <- nnet(Strength ~., data = con.train, size = 10, lineout = T, skip = T, maxit = 10000, decay = .001)

con.train.x <- con.train[,-1]

con.train.y <- con.train[,1]

result.trainNn10 <- nnet.sscv(con.train.x, con.train.y, con.train.nn, size = 10)

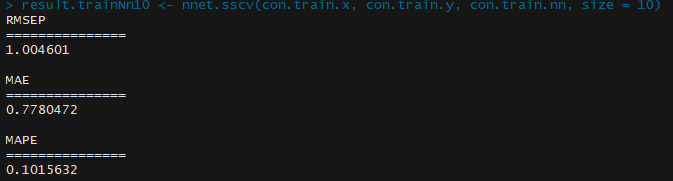
result.trainNn5 <- nnet.sscv(con.train.x, con.train.y, con.train.nn, size = 5)

result.trainNn3 <- nnet.sscv(con.train.x, con.train.y, con.train.nn, size = 3)

trendscat(con.train$Strength,fitted(con.train.nn))

cor(con.train$Strength,fitted(con.train.nn))

I used Box – Cox to find what variables needed to be transformed. There were 4 predictors that were transformed with different square roots and the predictor AGE was Log transformed. The response Strength was transformed with the square root of 3/5th. The MAPEP that was found with size 10 was 10.194%. I did check the size 3 and 5 of found them with a higher MAPEP. The range for MAPEP was between .075 and .112 which is a small range than the untransformed data.

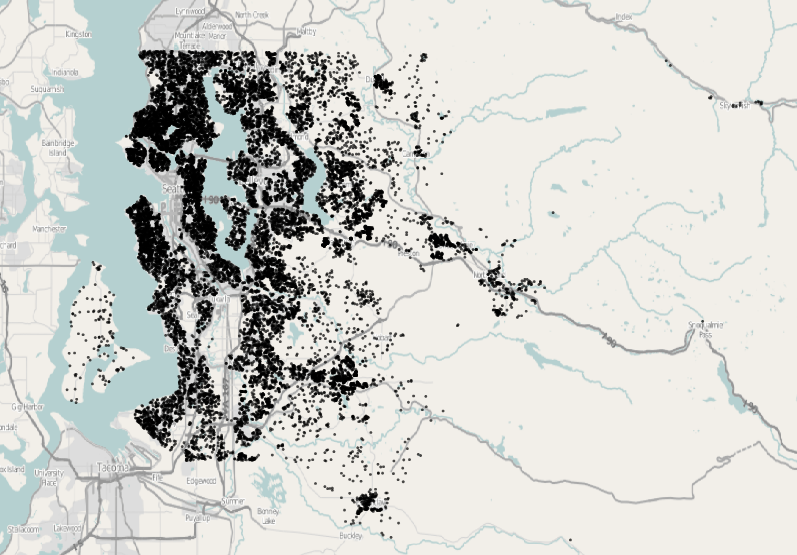


1. Using training (66%) and validation (33%) sets compare the predictive performance (RMSEP, MAE, MAPE) of your neural network model from part (a) and your best MLR and MARS models from Assignment 2. Use the same training/validation set for all three, thus you will need to run your best MLR and MARS models again using the same training and test sets used for the neural network model. (12 pts.)

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Model Method | (training data) | (MPa) | (MPa) | (%) |
| MLR best |  | 3.224 | 2.594 | 35.793 |
| MARS best | .8795 | 5.623255 | 4.304911 | 14.52817 |
| Neural Net |  | 1.005 | .778 | 10.156 |

**Problem 2 – PREDICTING SELLING PRICE OF HOMES IN KING COUNTY, WA**

The data for these sales comes from the official public records of home sales in the King County area, Washington State. The data set contains 21,606 homes that sold between May 2014 and May 2015. The table below gives variable names and descriptions. The map below shows the location of all 21,606 homes you will be working with.

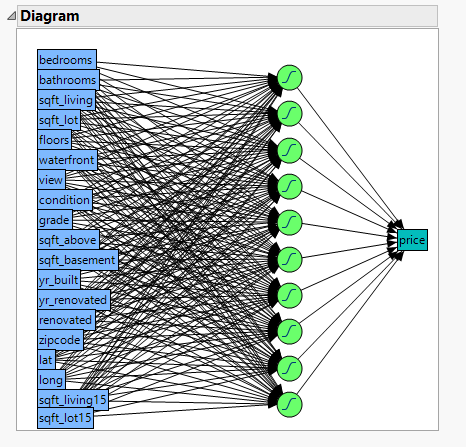
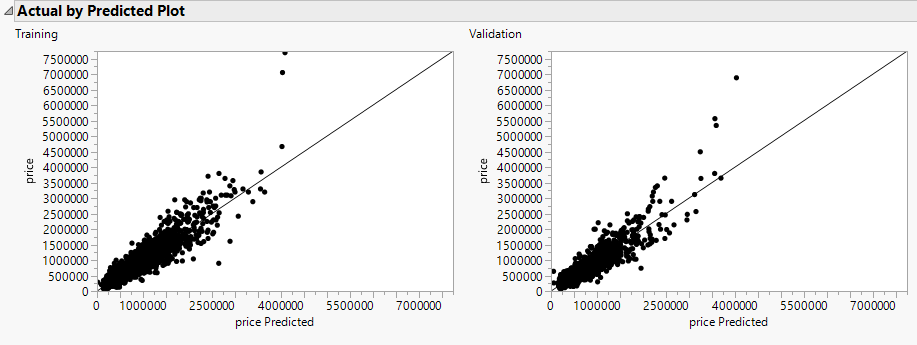
****

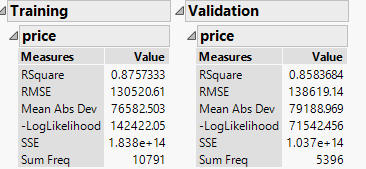
**Variables in King County, WA Datasets**

* ID – id number (DO NOT USE IN YOUR MODELS!)
* **price** - Price of each home sold
* **bedrooms** - Number of bedrooms
* **bathrooms** - Number of bathrooms, where .5 accounts for a room with a toilet but no shower.
* **sqft\_living** - Square footage of the apartments interior living space.
* **sqft\_lot** - Square footage of the land space.
* **floors** - Number of floors.
* **waterfront** - A categorical variable for whether the apartment/home was overlooking the waterfront or not (1 = yes, 0 = no).
* **view** - An ordinal index from 0 to 4 of how good the view of the property has.
* **condition** - An index from 1 to 5 on the condition of the apartment**.**
* **grade** - An ordinal index from 1 to 13, where 1-3 falls short of building construction and design, 7 has an average level of construction and design, and 11-13 have a high quality level of construction and design.  Other intermediary values indicate conditions in between these descriptors.
* **sqft\_above** - The square footage of the interior housing space that is above ground level.
* **sqft\_basement** - The square footage of the interior housing space that is below ground level.
* **yr\_built** - The year the house was initially built.
* **yr\_renovated** - The year of the house’s last renovation, 0 indicates it has not been renovated.
* **renovated** – indicator of whether or not the home has been renovated (1 = yes, 0 = no)
* **zipcode –** ZIP code area the house is in (Note: ZIP codes are NOT numeric!)
* **lat -** Lattitude of the home
* **long**- Longitude of the home
* **sqft\_living15**- The mean square footage of the interior living space of the nearest fifteen neighboring homes.
* **sqft\_lot15** -The mean square footage of the land lots of the nearest fifteen neighboring homes.
* **Test Set** – denotes whether the home is in the Test Set or the Training Set. These sets are the same as those for Assignment 1.

1. Using the **King County Homes (full).JMP** file on the course website develop a neural network model for predicting home price in JMP. Be sure to use some form of cross-validation to fine-tune your model. DO NOT USE BOOSTING! Also I would NOT recommend using multiple tours to fit the model, as this will take a long time for even a modest number of tours. Include plots of the actual vs. predicted in both the log-scale (assuming you used as the response) and in the original scale (both the predicted and actual prices in $).

Discuss the process you used to arrive at your final model. Include a diagram of your final model from JMP. Include results of the cross-validation for your final model. (20 pts.)



So, I decided to use TanH for the activation type, with only the first layer, and transform covariates. I started with 3 nodes and kept moving up tell I got to 10 nodes. I stopped there when I saw the R squared being at .87 for the training data, then I went up to 11 and saw the r square go down so that when I decided to stop there.

1. Can you identify which variables appear to be the most important in predicting the selling price (or log selling price) from your final model? If so, which variables seem the most important? Can you create plots/visualizations or create summary statistics that show which predictors/terms are most important? I DO NOT HAVE A SPECIFIC THING I AM LOOKING FOR HERE, I JUST WANT TO SEE IF YOU CAN COME UP WITH SOMETHING. (10 pts.)
2. Use your model to predict the selling price of the homes in the test set (denoted by the ***Test Set*** column) and submit those predictions in the same format as your predictions for Assignment 1. I will again create a leaderboard based upon these predictions. I will demonstrate how to do this in class – remind me! (5 pts.)