**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* ()

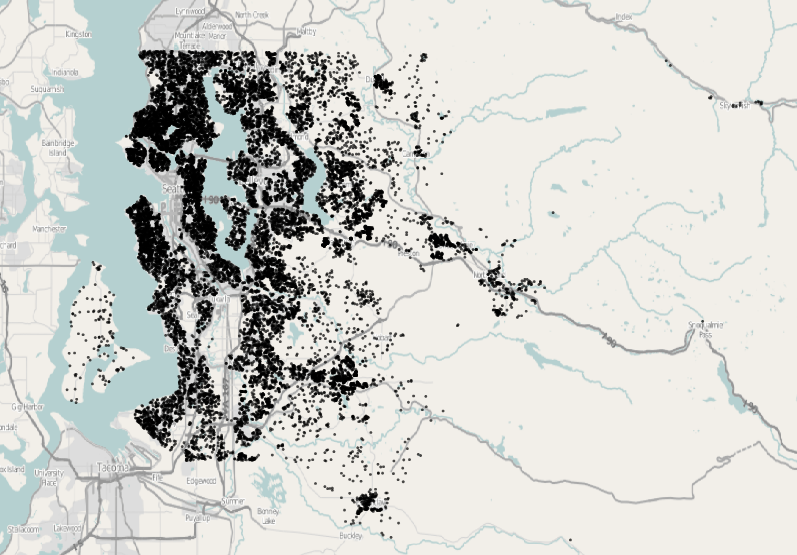
Data source: I-Cheng Yeh, "*Modeling of strength of high performance concrete using artificial neural networks*," Cement and Concrete Research, Vol. 28, No. 12, pp. 1797-1808 (1998)

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.)
2. 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 |  |  |  |  |
| MARS best |  |  |  |  |
| Neural Net |  |  |  |  |

**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.)

**Model creation process on price in the original scale:**

1. Model 1: First, we fit the model with the default settings.
2. Model 2: Next, we decided to add the 2nd hidden layer, adjusting for skewness and robustness (with Transformed Covariates and Robust Fit checked). The following models always included Transformed Covariates and Robust Fir checked.
3. Model 3-6: Next we experimented a lot with different numbers of nodes in each layers and added weight decay penalty method with some number of tours.
   1. We were switching between Gaussian and TanH activation functions and also used a combination of both. At the end, the best result was always given by the TanH function.
   2. We used 3-5 number of tours with weight decay as penalty method. Model 4 (3 tours) and Model 6 (5 tours) gave us the highest R^2. Those specific number of tours did not take long to fit the neural network; we decided to use weight decay to reduce overfitting of our model – to prevent the weights from growing too large.
   3. We saw some overfitting problems in previous models; the best result, with the least variation, gave us Model 6.
   4. Actual vs. predicted plots were not looking so great as at a particular point at Predicted values, the data points seemed to be reaching some kind of a limit (Model 2, 3)

**Model 1:**

A screenshot of a social media post

Description automatically generated

**Model 2:**

A screenshot of a social media post

Description automatically generated

**Model 3:**

A screenshot of a cell phone

Description automatically generated

**Model 4:** A screenshot of a social media post

Description automatically generated

**Model 5:** A screenshot of a social media post

Description automatically generated

**Model 6:**

A screenshot of a social media post

Description automatically generated

We applied a similar process in finding the best model with the transformed response (logged). The actual vs. predicted plots look much better in here.

**Model 1:**

A close up of a piece of paper

Description automatically generated

**Model 2:**

A close up of a piece of paper

Description automatically generated

**Model 3:**

A close up of a piece of paper

Description automatically generated

**Model 4:**

A close up of text on a white background

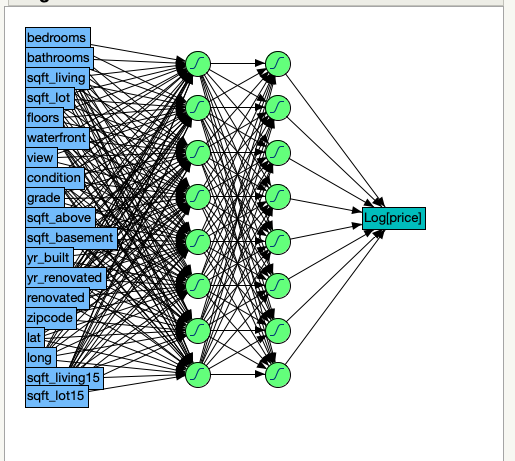
Description automatically generated

**The final model** is Model 4 above with the price logged. The actual vs. predicted plots look like they would not overfit the data as much as the models when the response is in the original scale.

These are the final Model 4 parameters:

A screenshot of a cell phone

Description automatically generated



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.)

A screenshot of a cell phone

Description automatically generated

The above summary report shows the total effect of predictors’ importance in predicting the log selling price. We used the resampled inputs because the distributions are not uniform. We wanted to generate the dependent resampled inputs; however, JMP crushed every time we tried to run it as it was using k-nearest neighbors approach to account for correlation.

We see that yr\_renovated and renovated are the most important, followed by latitude and longitude variables.

1. 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.)