

Neural Network Model - Cereals Rating

Project

Neural network is an information-processing machine and can be viewed as analogous to human nervous system. A neural network is a model characterized by an activation function, which is used by interconnected information processing units to transform input into output. A neural network has always been compared to human nervous system. This project deals with fitting a neural network model. For this project we use a subset of cereal dataset shared by Carnegie Mellon University (CMU). The details of the dataset are on the following link: <http://lib.stat.cmu.edu/DASL/Datafiles/Cereals.html>. The objective is to predict rating of the cereals variables such as calories, proteins, fat etc. We will be using rating as dependent variable and calories, proteins, fat, sodium and fibre as independent variable.

Library

```
## Loading required package: lattice
## Loading required package: ggplot2
## Warning: package 'neuralnet' was built under R version 3.4.1
##
## Attaching package: 'boot'
## The following object is masked from 'package:lattice':
##
##      melanoma
## Warning: package 'matrixStats' was built under R version 3.4.3
##
## Attaching package: 'matrixStats'
## The following object is masked from 'package:plyr':
##
##      count
```

Data

```
## 'data.frame':   75 obs. of  6 variables:
## $ calories: int  70 120 70 50 110 110 130 90 90 120 ...
## $ protein : int  4 3 4 4 2 2 3 2 3 1 ...
## $ fat      : int  1 5 1 0 2 0 2 1 0 2 ...
## $ sodium   : int 130 15 260 140 180 125 210 200 210 220 ...
## $ fiber    : num 10 2 9 14 1.5 1 2 4 5 0 ...
## $ rating   : num 68.4 34 59.4 93.7 29.5 ...
```

Training and Test Data Sets

Scaling of Data

We will first scale the dataset. The scaling of data is essential otherwise a variable may have large impact on the prediction variable only because of its scale. For this purpose we will use the min-max normalization

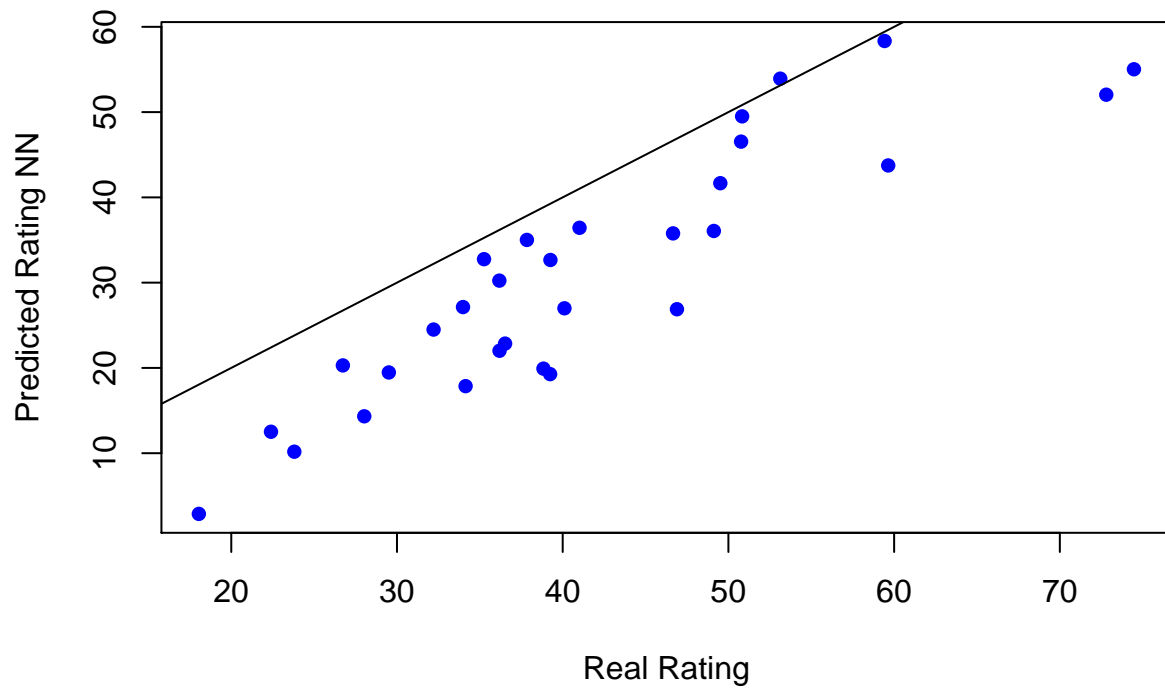
Fit Neural Network

Plot Neural Network

Our model has 3 hidden layers. The black lines show the connections with weights. While the blue line displays the bias term

Predict Rating

We must remember that the predicted rating will be scaled and it must be transformed in order to make a comparison with real rating



Calculate Root Mean Square Error (RMSE)

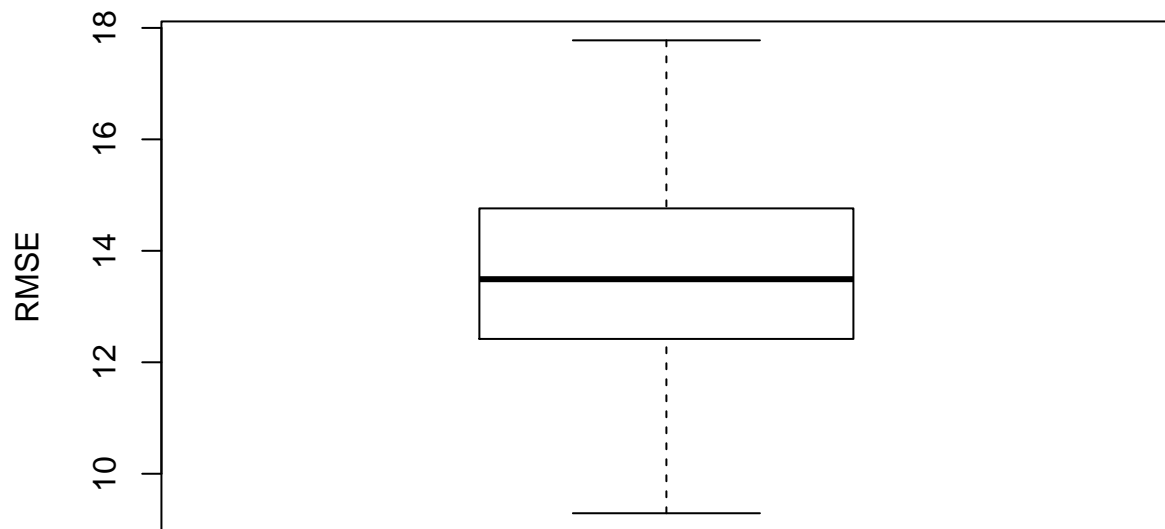
Cross-Validation of Neural Network

K-Fold cross-validation

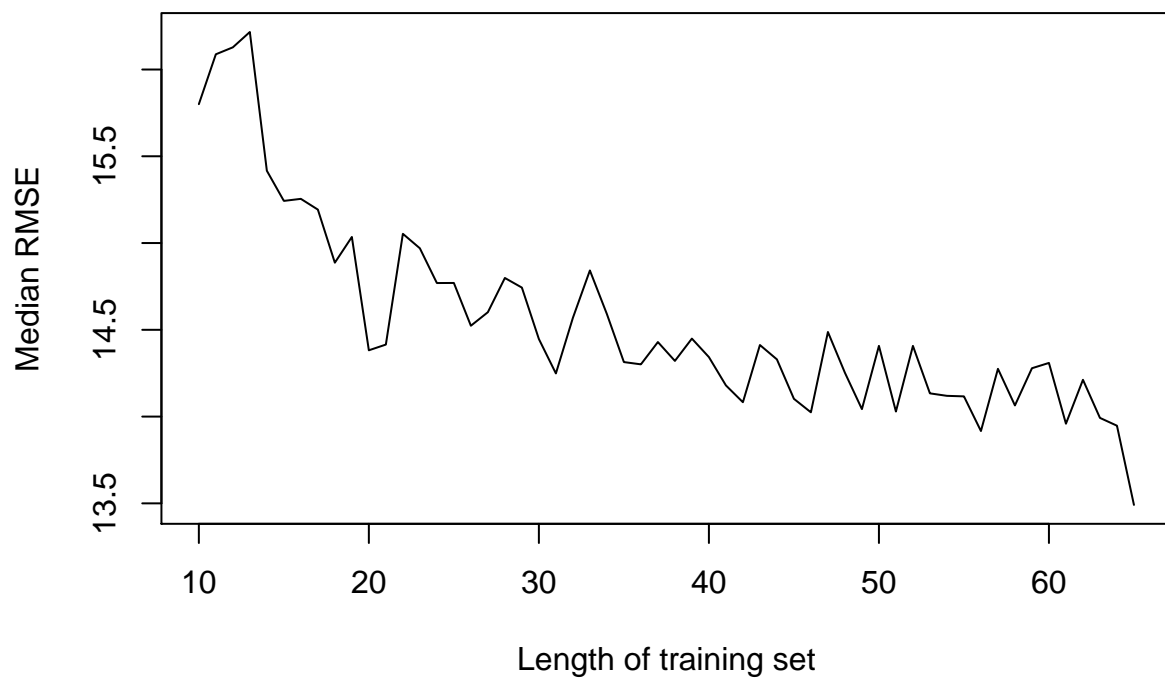
Box Plot

The boxplot shows that the median RMSE across 100 samples when length of the training set is fixed to 65 is 13.7

RMSE Boxplot(length of training set = 65)



Variation of RMSE with length of training set



Conclusion

The plot shows that the median RMSE of our model decreases as the length of the training data set increases. So the model accuracy is dependent on the length of the training data set. This shows that the performance of neural network model is sensitive to training - test split