Nathan Lin, Andrew Ton, Mansoor Syed
DS 4559
12/4/2016

Project #2 Responses

Required responses and supplementary information for questions 1 and 2 can be found below.

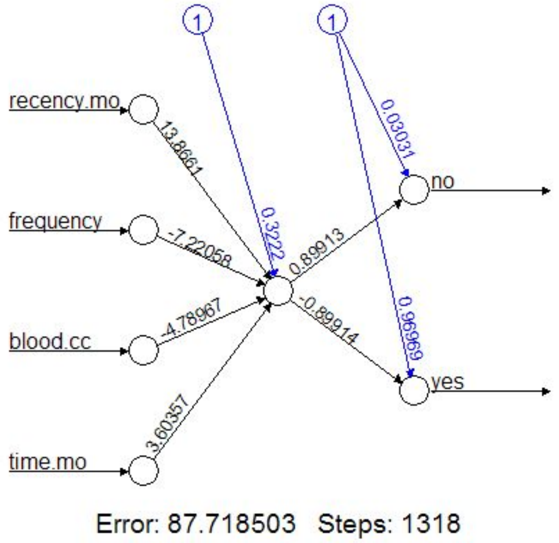
Our submission contains 2 files:

- Question 1
  - Lin_Ton_Syed_Project2_Q1.R
- Question 2
  - Lin_Ton_Syed_Project2_Q2.R

Question 1

This dataset contains a portion of the donor database for the Blood Transfusion Service Center in Taiwan (random sample of 748 observations). The center drives a bus to a university every 3 months to collect blood, and the dataset is used to build a variation of the RFM (Recency, Frequency, Monetary Value) marketing model in analyzing customer value. In this case, the "customer" is a blood donor, with their value measured in the frequency/volume of donations. Blood donations are vital for emergency and elective surgeries (there was recently a shortage in the US), so identifying target donors and those who may deliver the best future value is extremely important for a blood bank.

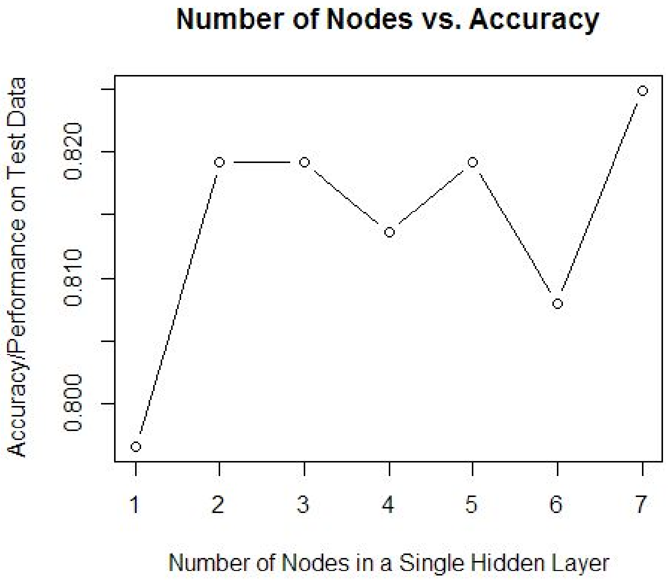
Network with 1 hidden layer/node



The accuracy of 1 hidden node was 79.66%.

Number of nodes vs. accuracy, output

This output was generated by looping through the same neural net 7 times, each time added a new node. Note that with 8 nodes, the network would not converge.



Our best performance was with 7 nodes, with an accuracy of 82.49%.

<u>Question 2</u>

The aim was to determine the performance of C50 and random forest trees in predicting the malignancy of breast cancer from a Wisconsin data-set. Predicted classes were compared to actual classes given a number of cancer attributes such as clump size and shape.

We ran a 10-fold cross-validation for random forest and C50 on the Wisconsin BC data and the results of the models are reported in the datatable below. The Random Forest model had extremely accurate results. It scored 100% accuracy for more than six hundred patients and had an AUC of .9935, almost optimal behavior.

Although the C50 model was less accurate than the random forest model, it was still fairly accuracy at ~94%.

Figure 1: Comparison of AUC and accuracy for Random Forest testing and C50

	AUC	*Accuracy*
Random Forest	.9935	1.0
C50	.9411	.94

Figure 2: ROC Curve for Random Forest Model (10-fold cross validation)˙

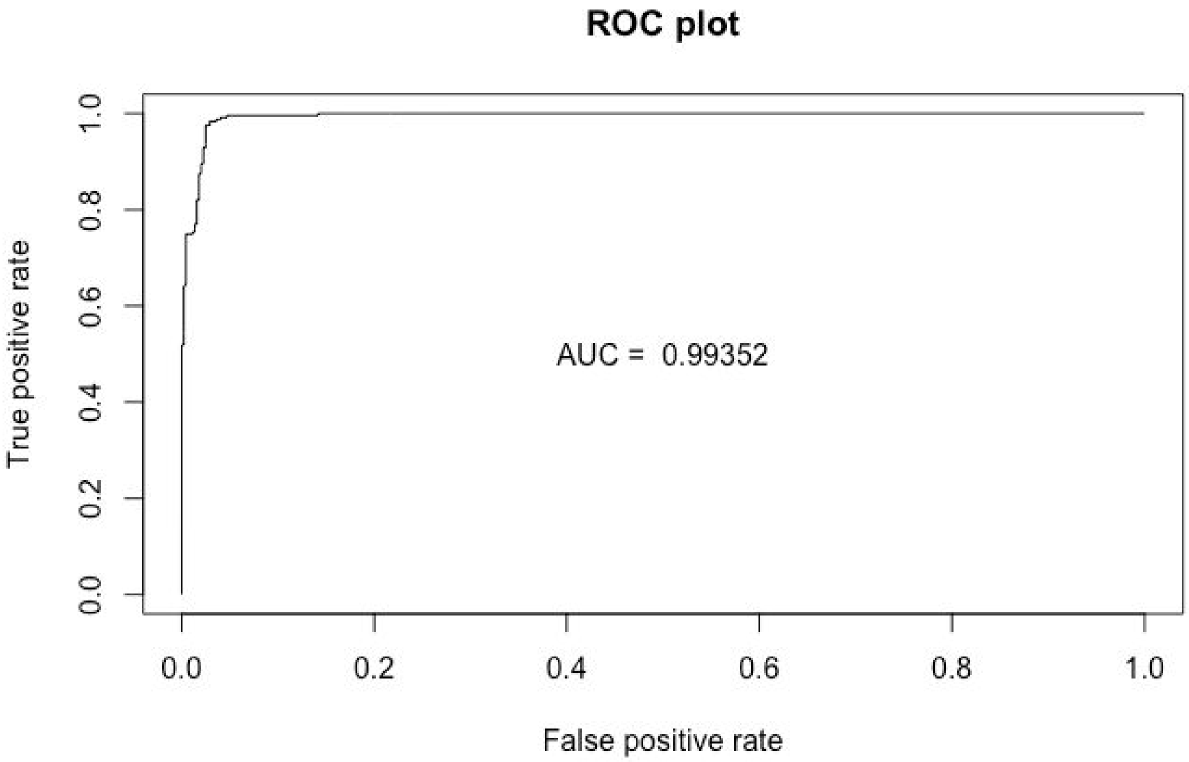
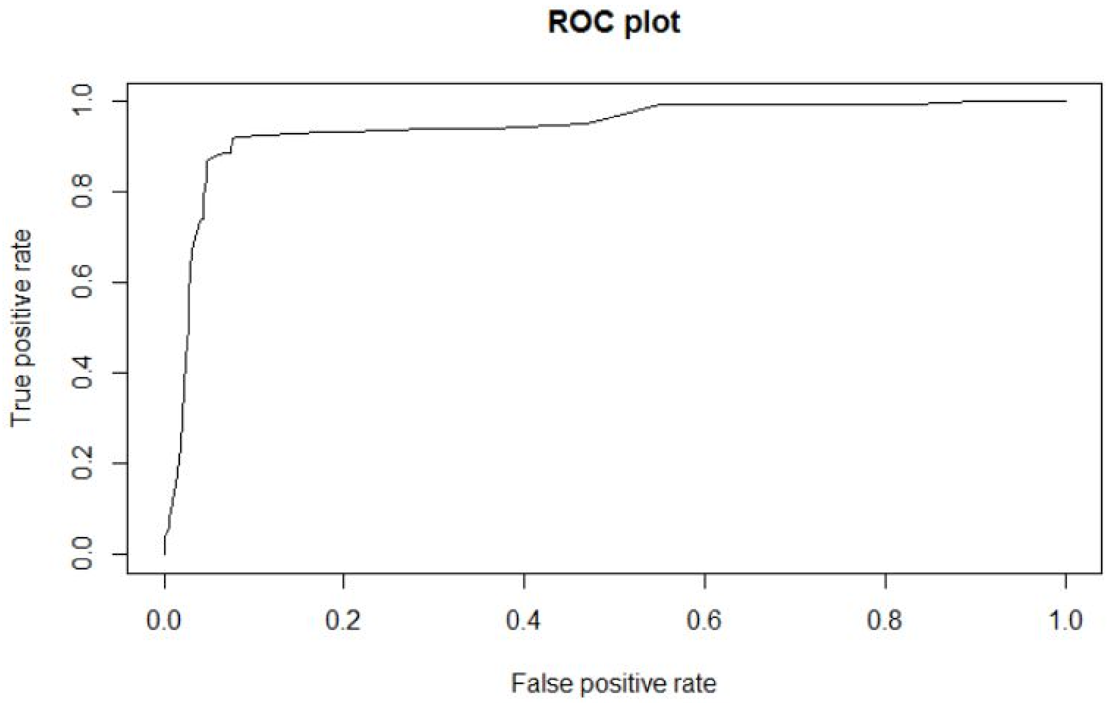


Figure 3: ROC Curve for C50 Model (10-fold cross validation)


ROC plot
True positive rate
False positive rate