Comparison of Supervised Models
Boston Housing Data Set

Alejandro Kantor

1 Purpose

Choosing a supervised model for a particular data set depends on several considerations; one of the most important ones is the goodness of fit or performance of the model. In order to facilitate comparing the performance of models for data sets with strictly positive target value and predetermined explanatory variables, we present the program *makeBenchmarking.R* with its corresponding and LaTeX document *documentation.tex*. The R program builds the models detailed in Table 1 and calculates performance statistics, passing them to objects LaTeX can input.

As an example, we run the program on the Boston data set which is described in `https://archive.ics.uci.edu/ml/datasets/Housing`.

Model Short Hand	Description
Linear	Linear model with normal error
Gamma1	Generalized Linear Model with gamma distribution and link inverse
Gamma2	Generalized Linear Model with gamma distribution and link log
Ctree	Conditional Regression Tree from R package *partykit* using default parameters
Cforrest	Conditional Regression Tree from R package *partykit* using default parameters with 100 iterations
SMV1	Support Vector Machine from R package *e1071* using default parameters
SMV2	Support Vector Machine from R package *e1071* using default parameters with $cost = 5$

Table 1: Description of Models

2 Methodology

In order to compare the performance of each model detailed in Table 1, we apply an 8-fold Cross Validation. In particular, we group the original data into 8 groups and then, for each group, we train the model on the remaining 7 groups and test the performance on the selected group. As a result, we have 8 values for each performance measure, which allows for a more robust comparison between the models.

Performance is measured by a few statistics. Our primary statistic is the Mean Square Difference (MSE) between the estimated and observed values. As a complementary measure, we look at the Proportion of Cases with an Absolute Log Difference (PCALD) between the estimated and observed values up to a certain threshold k, for more information see Appendix B.

3 Results

We present the results of running the program on the Boston Data Set in two different formats. We show the average performance indicators for the train and test samples in Tables 2 and 3, respectively. On the other hand, we show the distribution of MSE and absolute log change of 0.05 for the test samples, in Figures 1 and 2

In general, we observe that all models have lower average performance indicators in the test samples compared to the train samples, suggesting over-fitting in the training sample.

SVM2 has the best performance in the test samples with a distribution of MSE in general lower to other models (see Figure 1), averaging at 10.77, as well as a higher distribution of poprchng0.05 (see Figure 2), averaging at 0.42. It is also important to note that the performance of SMV2 is considerably higher than the next best model SMV1 . For example the MSE in the test sample that is 29% higher in SMV1 compared to SMV2. We can see how these two models' estimates compare to the observed value for a test sample in Figure 3.

Model	AIC	MSE	$PCALD_{0.05}$	$PCALD_{0.1}$
Linear	2647.05	20.95	0.27	0.47
Gamma1	2472.86	15.06	0.28	0.52
Gamma2	2506.99	17.71	0.27	0.50
Ctree		11.37	0.34	0.59
Cforest		11.29	0.38	0.66
SMV1		10.70	0.48	0.74
SMV2		6.20	0.57	0.81

Table 2: Mean Summary Statistics of Train Sample

Model	MSE	$PCALD_{0.05}$	$PCALD_{0.1}$
Linear	23.66	0.26	0.45
Gamma1	17.27	0.27	0.50
Gamma2	19.63	0.27	0.49
Ctree	19.32	0.30	0.52
Cforest	16.45	0.30	0.58
SMV1	13.92	0.40	0.64
SMV2	10.77	0.42	0.67

Table 3: Mean Summary Statistics of Test Sample

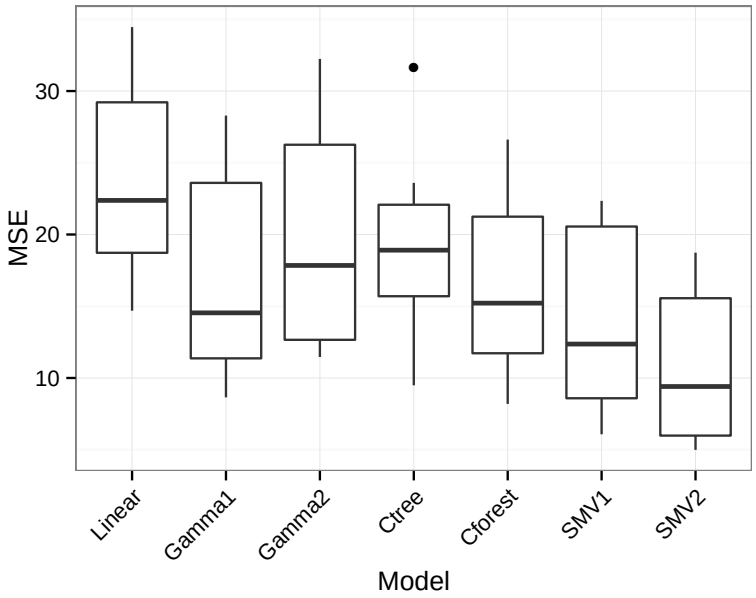

Figure 1: Distribution of MSE in Test Sample by Model

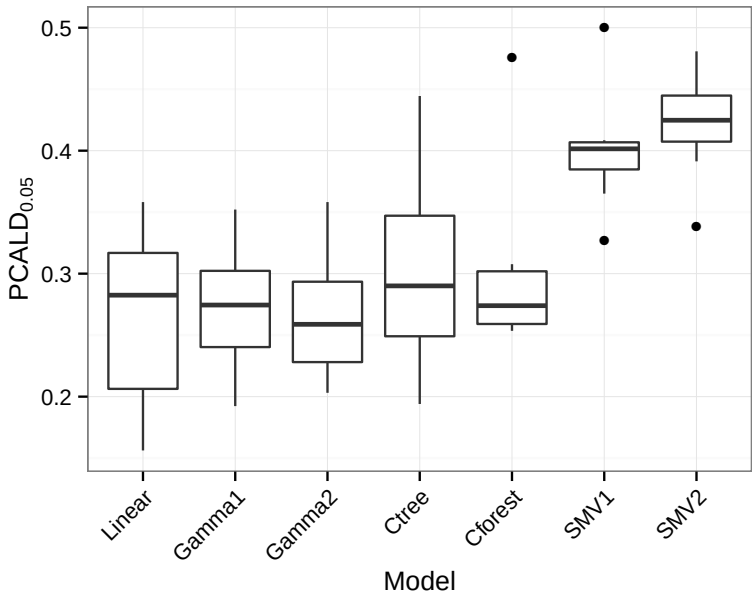

Figure 2: Distribution of $PCALD_{0.05}$ in Test Sample by Model

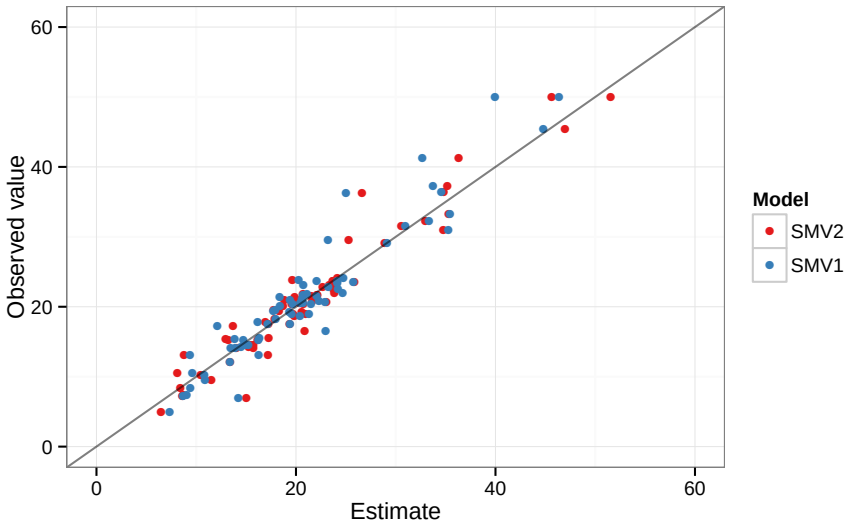

Figure 3: Scatter Plot of Estimates by Models with the Best Performance the Observed Values

4 Conclusions

The main conclusion of this analysis is that the program successfully allows for the comparison of the performance of the models detailed in Table 1 with the Boston Housing Data Set.

We find that the SMV2 model has the best performance followed by the SMV1 model. Thus, if our only consideration is performance, we would suggest using this model for making prediction for the given data set.

Appendix

A TODOs & Improvements

We suggest the following improvements to the program

- generalize the process so we can choose which models and performance statistics the program runs and thus allow for classification models and other regression models to be included;
- modify the code so that Artificial Neural Networks can also be fitted;
- further modulate the code reducing its repetition;
- make a more general latex template for different types of outputs from *makeBenchmarking.R*.

B Proportion of Cases with Log Difference $\leq$ k

The Absolute Log Difference for a give observation i is defined by the following equation.

$$ALD_i = \|\log(y_i) - \log(\hat{y}_i)\| \tag{1}$$

where y_i is the observed value and $\hat{y}_i$ is the estimated value.

The Proportion of Cases with Log Difference $\leq$ k is defined by counting proportion of cases which satisfy $ALD_i \leq k$ as shown in the following equation.

$$PCALD_k = \frac{\sum_{i=1}^{n} (\mathbb{1}_{ALD_i \leq k})}{n} \tag{2}$$

where n is the number of cases in the data set.

C Scatter Plot Examples

In this Section, we present the scatter plots of the estimated values by each model with respect to the observed value. This analysis is performed on one of the test samples.

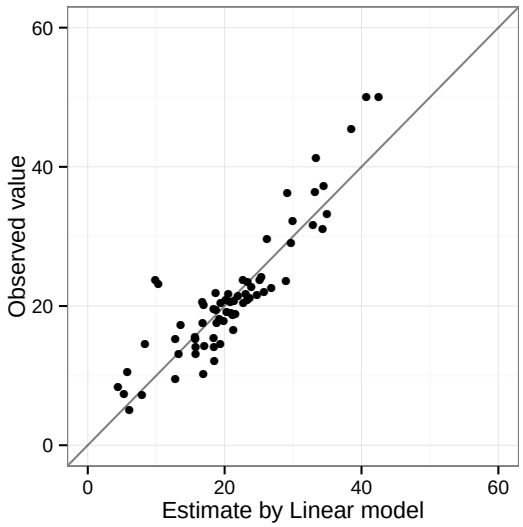

Figure 4: Linear

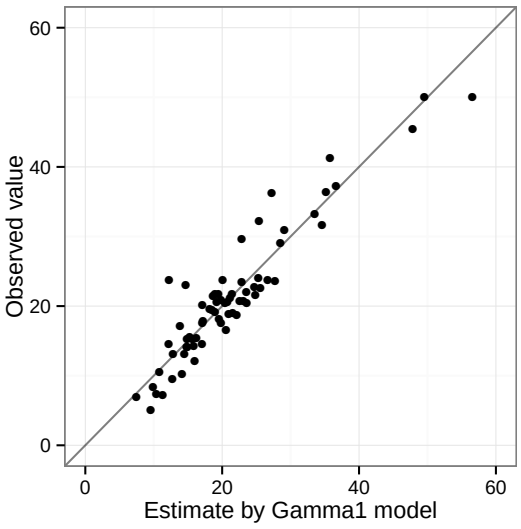

Figure 5: Gamma1

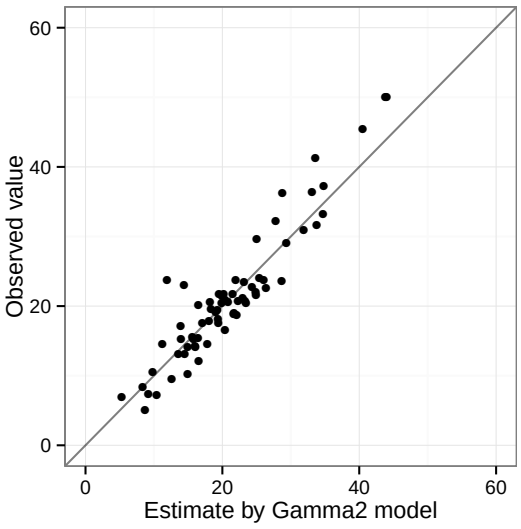

Figure 6: Gamma2

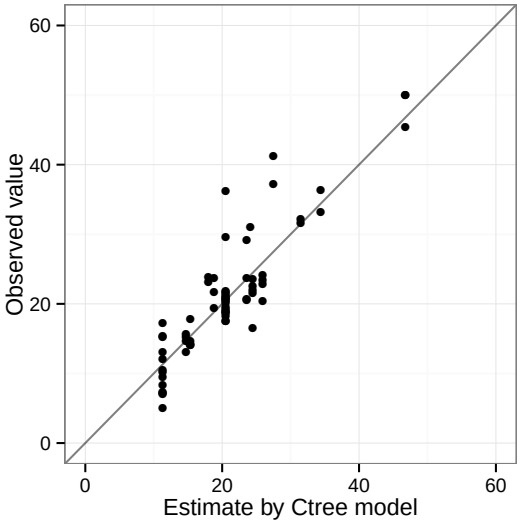

Figure 7: Ctree

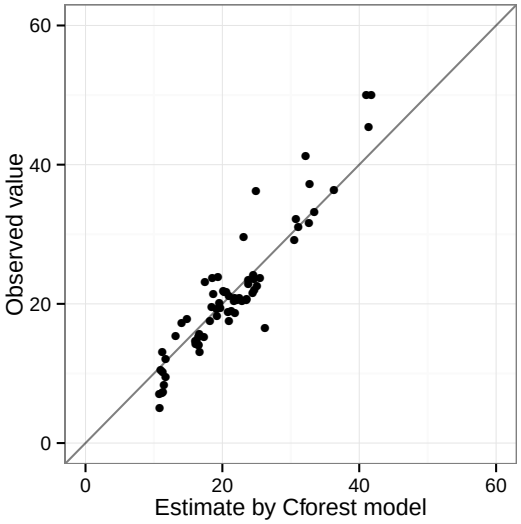

Figure 8: Cforest

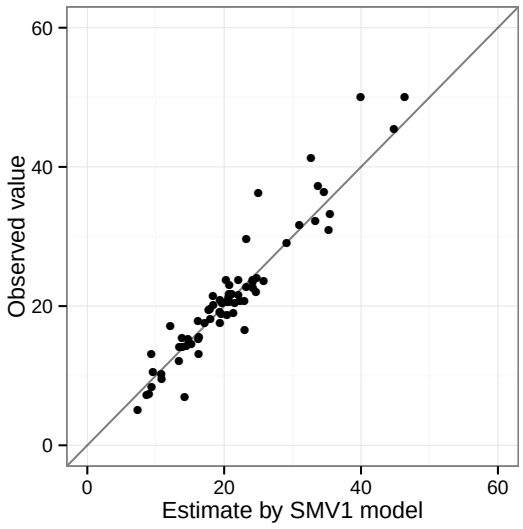

Figure 9: Svm1

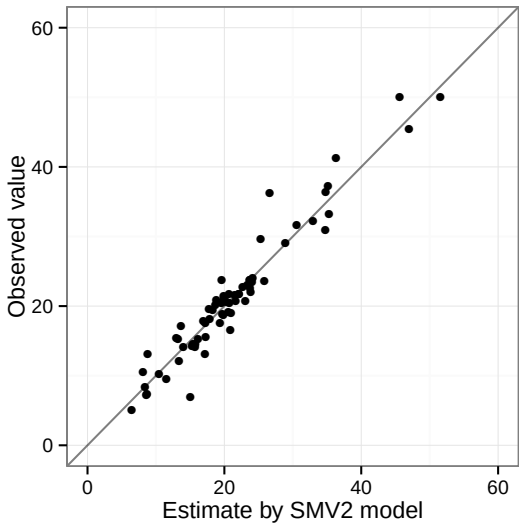

Figure 10: Svm1