Chapter 9
A Summary of Solubility Models

Across the last few chapters, a variety of models have been fit to the solubility data set. How do the models compare for these data and which one should be selected for the final model? Figs. 9.1 and 9.2 show scatter plots of the performance metrics calculated using cross-validation and the test set data.

With the exception of poorly performing models, there is a fairly high correlation between the results derived from resampling and the test set (0.9 for the RMSE and 0.88 for R^2). For the most part, the models tend to rank order similarly. K-nearest neighbors were the weakest performer, followed by the two single tree-based methods. While bagging these trees did help, it did not make the models very competitive. Additionally, conditional random forest models had mediocre results.

There was a "pack" of models that showed better results, including model trees, linear regression, penalized linear models, MARS, and neural networks. These models are more simplistic but would not be considered interpretable given the number of predictors involved in the linear models and the complexity of the model trees and MARS. For the most part, they would be easy to implement. Recall that this type of model might be used by a pharmaceutical company to screen *millions* of potential compounds, so ease of implementation should not be taken lightly.

The group of high-performance models include support vector machines (SVMs), boosted trees, random forests, and Cubist. Each is essentially a black box with a highly complex prediction equation. The performance of these models is head and shoulders above the rest so there is probably some value in finding computationally efficient implementations that can be used to predict large numbers of new samples.

Are there any real differences between these models? Using the resampling results, a set of confidence intervals were constructed to characterize the differences in RMSE in the models using the techniques shown in Sect. 4.8. Figure 9.3 shows the intervals. There are very few statistically significant

M. Kuhn and K. Johnson, *Applied Predictive Modeling*,
DOI 10.1007/978-1-4614-6849-3_9,

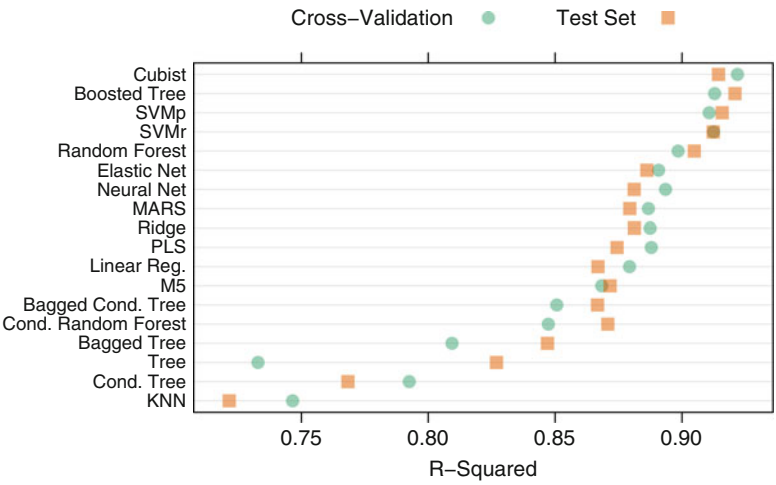

Fig. 9.1: A plot of the R^2 solubility models estimated by 10-fold cross-validation and the test set

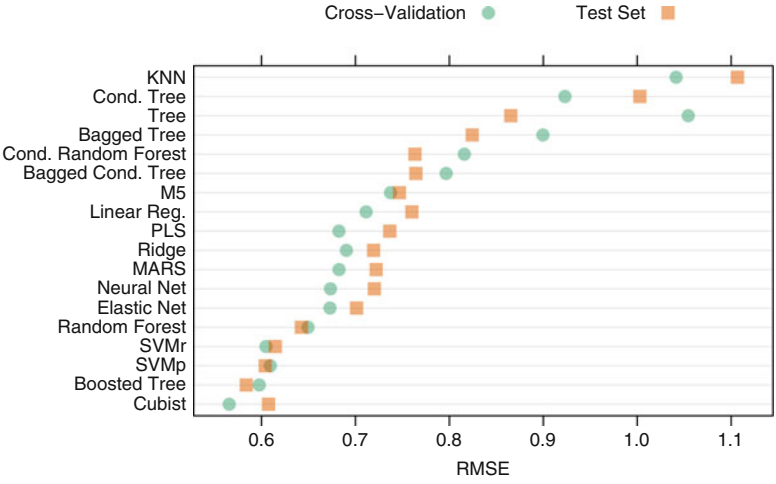

Fig. 9.2: A plot of the RMSE solubility models estimated by 10-fold cross-validation and the test set

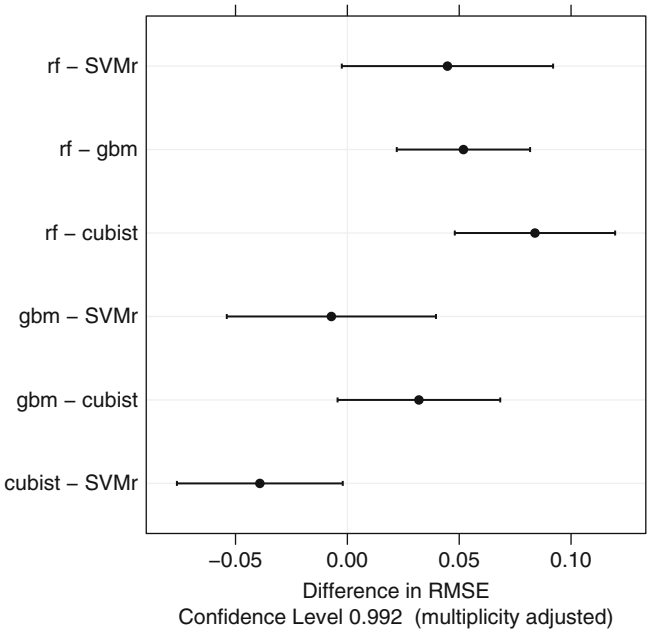

Fig. 9.3: Confidence intervals for the differences in RMSE for the high-performance models

differences. Additionally, most of the estimated mean differences are less than 0.05 log units, which are not scientifically meaningful. Given this, any of these models would be a reasonable choice.