# 11.3 Recursive Feature Elimination

As previously noted, recursive feature elimination (RFE, Guyon et al. (2002)) is basically a backward selection of the predictors. This technique begins by building a model on the entire set of predictors and computing an importance score for each predictor. The least important predictor(s) are then removed, the model is re-built, and importance scores are computed again. In practice, the analyst specifies the number of predictor subsets to evaluate as well as each subset's size. Therefore, the subset size is a *tuning parameter* for RFE. The subset size that optimizes the performance criteria is used to select the predictors based on the importance rankings. The optimal subset is then used to train the final model.

Section 10.4 described in detail the appropriate way to estimate the subset size. The selection process is resampled in the same way as fundamental tuning parameter from a model, such as the number of nearest neighbors or the amount of weight decay in a neural network. The resampling process *includes* the feature selection routine and the *external* resamples are used to estimate the appropriate subset size.

Not all models can be paired with the RFE method, and some models benefit more from RFE than others. Because RFE requires that the initial model uses the full predictor set, then some models cannot be used when the number of predictors exceeds the number of samples. As noted in previous chapters, these models include multiple linear regression, logistic regression, and linear discriminant analysis. If we desire to use one of these techniques with RFE, then the predictors must first be winnowed down. In addition, some models benefit more

from the use of RFE than others. Random forest is one such model (Svetnik et al. 2003) and RFE will be demonstrated using this model for the Parkinson's disease data.

Backwards selection is frequently used with random forest models for two reasons. First, as noted in Chapter 10, random forest tends not to exclude variables from the prediction equation. The reason is related to the nature of model ensembles. Increased performance in ensembles is related to the diversity in the constituent models; averaging models that are effectively the same does not drive down the variation in the model predictions. For this reason, random forest coerces the trees to contain sub-optimal splits of the predictors using a random sample of predictors[83]. The act of restricting the number of predictors that could possibly be used in a split increases the likelihood that an irrelevant predictor will be used in the split. While such a predictor may not have much direct impact on the performance of the model, the prediction equation is then functionally dependent on that predictor. As our simulations showed, tree ensembles may use every possible predictor at least once in the ensemble. For this reason, random forest can use some *post hoc* pruning of variables that are not essential for the model to perform well. When many irrelevant predictors are included in the data and the RFE process is paired with random forest, a wide range of subset sizes can exhibit very similar predictive performances.

The second reason that random forest is used with RFE is because this model has a well-known internal method for measuring feature importance. This was described previously in Section 7.4 and can be used with the first model fit within RFE, where the entire predictor set is used to compute the feature rankings.

One notable issue with measuring importance in trees is related to multicollinearity. If there are highly correlated predictors in a training set that are useful for predicting the outcome, then which predictor is chosen for partitioning the samples is essentially a random selection.

It is common to see that a set of highly redundant and useful predictors are *all* used in the splits across the ensemble of trees. In this scenario, the predictive performance of the ensemble of trees is unaffected by highly correlated, useful features. However, the redundancy of the features dilutes the importance scores. Figure 11.4 shows an example of this phenomenon. The data were simulated using the same system described in Section 10.3 and a random forest model was tuned using 2,000 trees. The largest importance was associated with feature $x_4$. Additional copies of this feature were added to the data set and the model was refit. The figure shows the decrease in importance for $x_4$ as copies are added[84]. There is a clear decrease in importance as redundant features are added to the model. For this reason, we should be careful about labeling predictors as unimportant when their permutation scores are not large since these may be masked by correlated variables. When using RFE with random forest, or other tree-based models, we advise filtering out highly correlated features prior to beginning the routine.

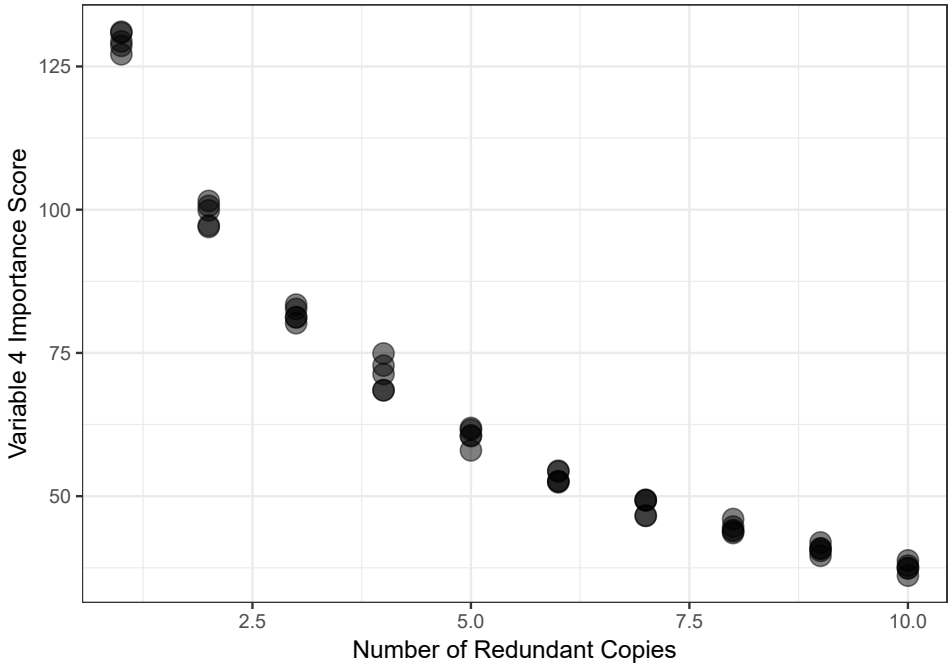


Figure 11.4: The dilution effect of random forest permutation importance scores when redundant variables are added to the model.

For the Parkinson's data, backwards selection was conducted with random forests,[85] and each ensemble contained 10,000 trees[86]. The model's importance scores were used to rank the predictors. Given the number of predictors and the large amount of correlation between them, the subset sizes were specified on the $\log_{10}$ scale so that the models would more thoroughly investigate smaller sizes. The same resampling scheme was used as the simple filtering analysis.

Give our previous comments regarding correlated predictors, the RFE analysis was conducted with and without a correlation filter that excluded enough predictors to coerce all of the absolute pairwise correlations to be smaller than 0.50. The performance profiles, shown in the left-hand panel of Figure 11.6, illustrates that the models creating using all predictors show a slight edge in model performance; the ROC curve AUC was 0.064 larger with a confidence interval for the difference being (0.037, 0.091). Both curves show a fairly typical pattern; performance stays fairly constant until relevant features are removed. Based on the unfiltered models, the numerically best subset size is 377 predictors, although the filtered model could probably achieve similar performance using subsets with around 30 predictors.

What would happen if the rankings were based on the ROC curve statistics instead of the random forest importance scores? On one hand, important predictors that share a large correlation might be ranked higher. On the other hand, the random forest rankings are based on the simultaneous presence of all features. That is, all of the predictors are being considered at once. For this reason, the ranks based on the simultaneous calculation might be better. For example, if there are important predictor interactions, random forest would account for these contributions, thus increasing the importance scores. To test this idea, the same RFE search was used with the individual areas under the ROC curves (again, with and without a correlation filter). The results are shown in the right-hand panel of

Figure 11.6. Given the experimental noise, the performance values are about the same although the filtered model that ranks using ROC curves had slightly better resampled performance values. Interestingly, the elimination process here precipitated a more rapid loss of performance. Based on the filtered model shown here, a subset of 128 predictors could produce an area under the ROC curve of about 0.786. Using another 100 random samples of subsets of size 128, in this case will lower correlations, the optimized model has higher ROC scores than 32% of the random subsets.

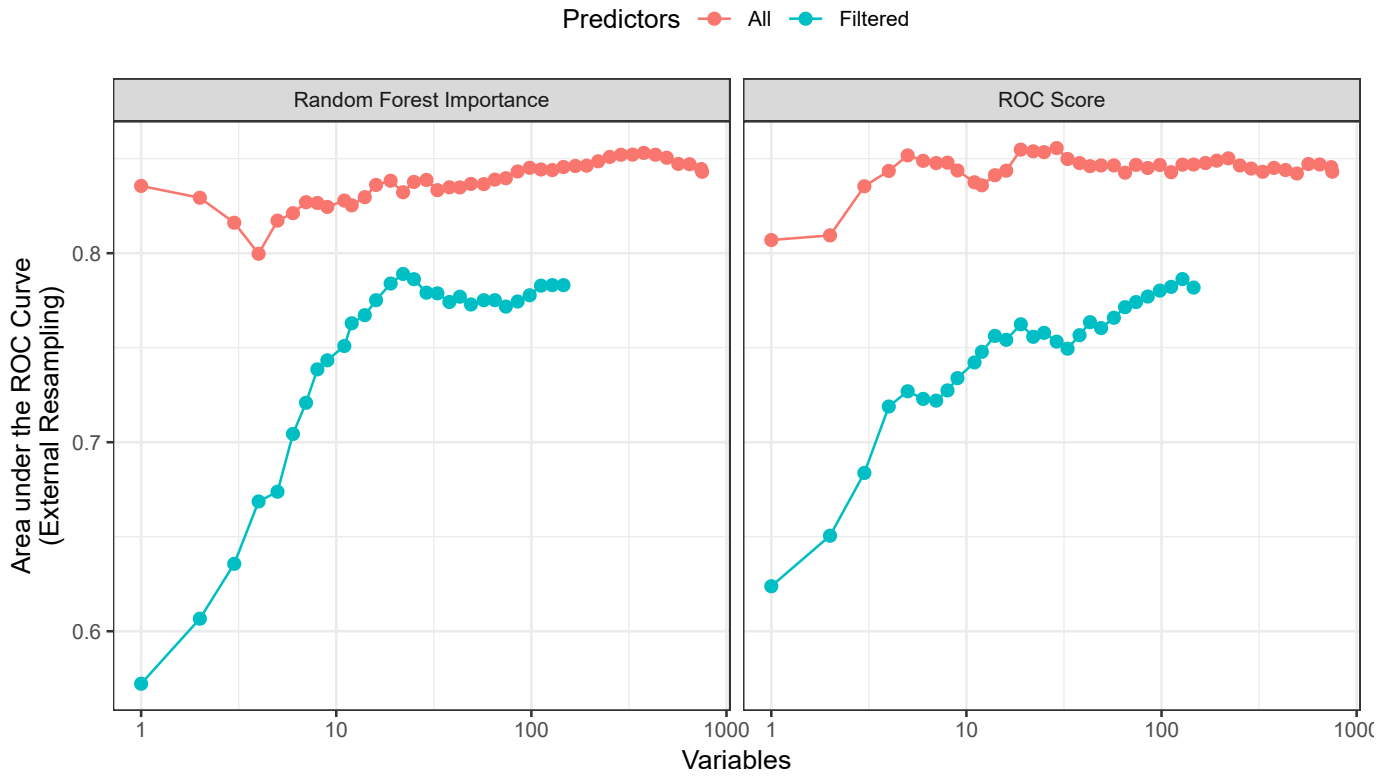


Figure 11.5: The external resampling results for RFE using random forests. The panels reflect how the predictor importances were calculated.

How consistent were the predictor rankings in the data? Figure 11.6 shows the ROC importance scores for each predictor. The points are the average scores over resamples and the bands reflect two standard deviations of those values. The rankings appear reasonably consistent since the trend is not a completely flat horizontal line. From the

confidence bands, it is apparent that some predictors were only selected once (i.e., no bands), only a few times, or had variable values across resamples.

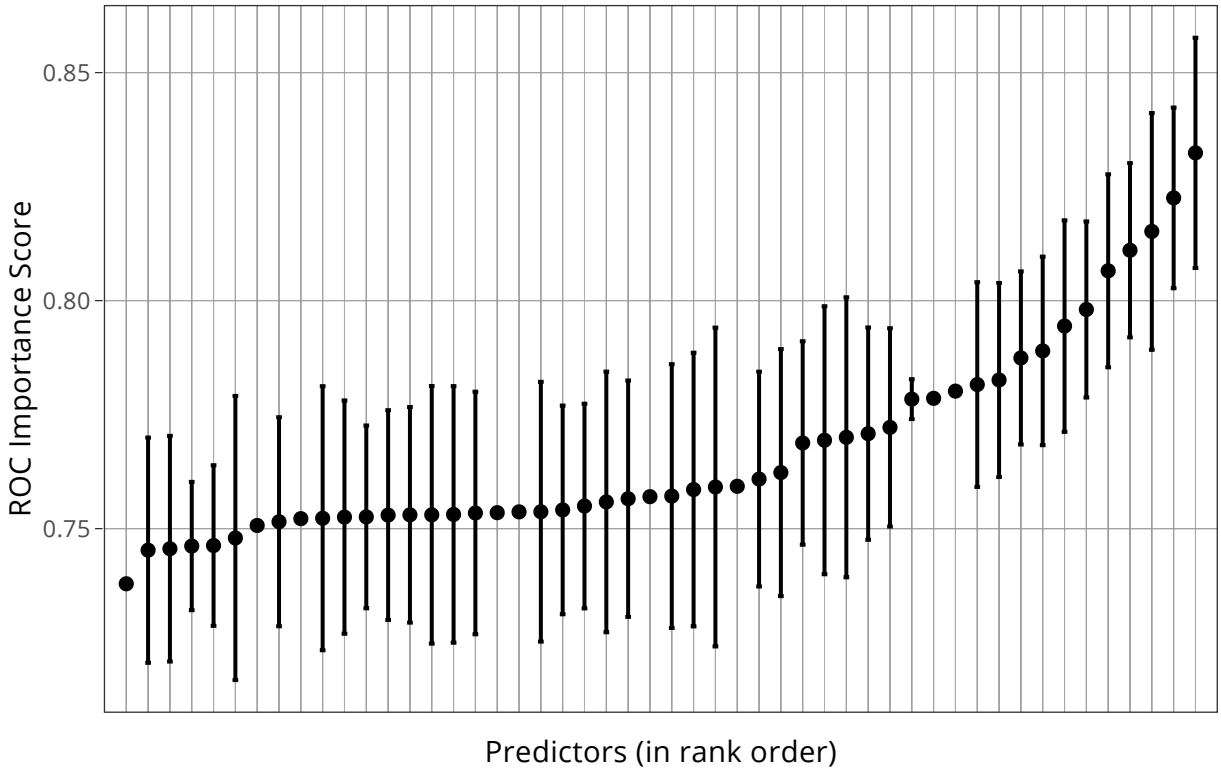


Figure 11.6: Mean and variability of predictor rankings.

RFE can be an effective and relatively efficient technique for reducing the model complexity be removing irrelevant predictors. Although it is a greedy approach, it is probably the most widely used method for feature selection.

83. Recall the size of the random sample, typically denoted as $m_{try}$, is the main tuning parameter↩

84. Each model was replicated five times using different random number seeds.↩

85. While $m_{try}$ is a tuning parameter for random forest models, the default value of $m_{try} \approx sqrt(p)$ tends to provide good overall performance. While tuning this parameter may have

led to better performance, our experience is that the improvement tends to be marginal.↩

86. In general, we do not advise using this large ensemble size for every random forest model. However, in data sets with relatively small (but wide) training sets, we have noticed that increasing the number of trees to 10K–15K is required to get accurate and reliable predictor rankings.↩