Feature Engineering Chapter 11
Greedy Search methods

Ranking and selecting features

Simple filters

- Screen predictors by testing if there is a relationship before including them
- Ranked and keep either top *p* or based on some threshold
- Mix of predictor types can mean converting everything to a p-value can be appropriate

Can be done individually or using all features at once (e.g. importance scores from random forests)

A summary of these simple filters can be found in Figure 11.1.

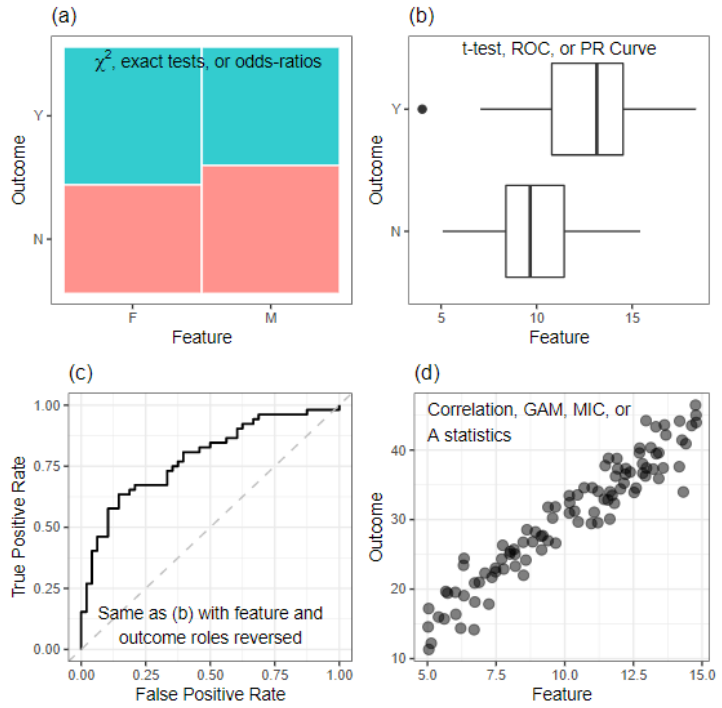


Figure 11.1: A comparison of the simple filters that can be applied to feature and outcome type combinations: (a) categorical feature and categorical outcome, (b) continuous feature and categorical outcome, (c) categorical feature and continuous outcome, and (d) continuous feature and continuous outcome.

Prevent false positives

Simple filters

- Consider adjusting p-values to account for number
- As usual… use appropriate resampling methods

"When using simple screening filters, selecting both the subset of features and model tuning parameters cannot be done in the same layer of cross-validation, since the filtering must be done independently of the model tuning. Instead, we must incorporate another layer of cross-validation. The first layer, or external layer, is used to filter features. Then the second layer (the "internal layer") is used to select tuning parameters."

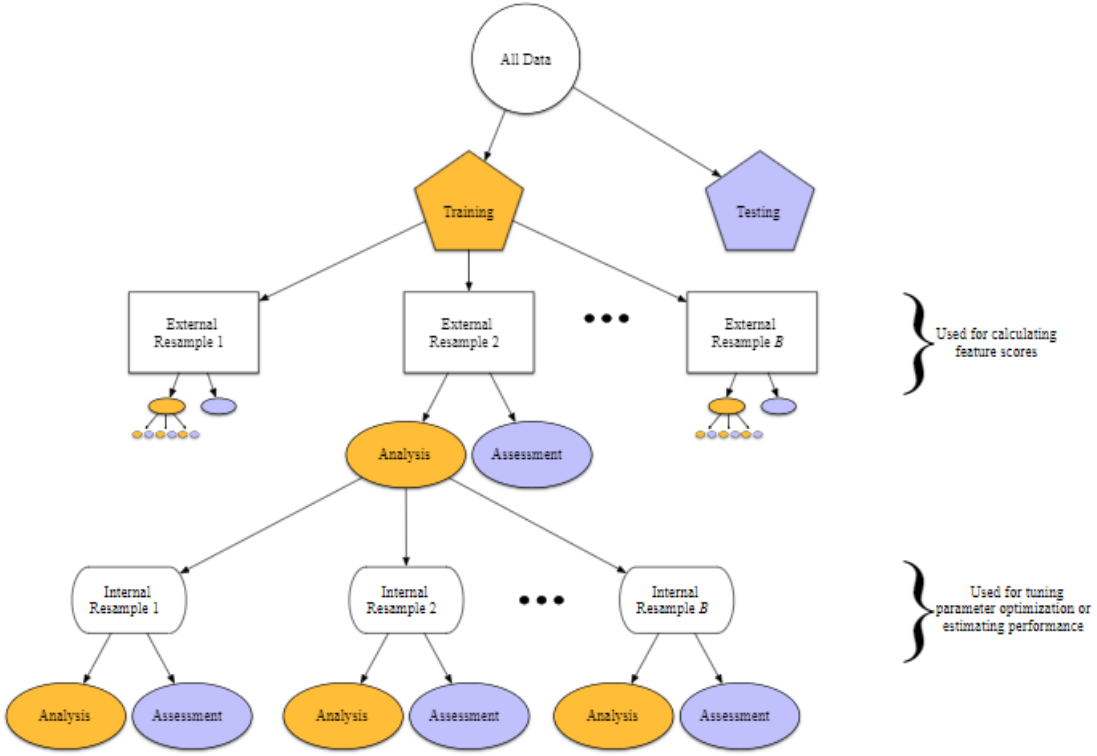


Figure 11.2: A diagram of external and internal cross-validation for simple filters.

Dealing with multicollinearity

Simple filters

- Can use methods like Partial least squares that will handle multi-collinearity
- Can set-up a collinearity filter

How would collinearity filter look?

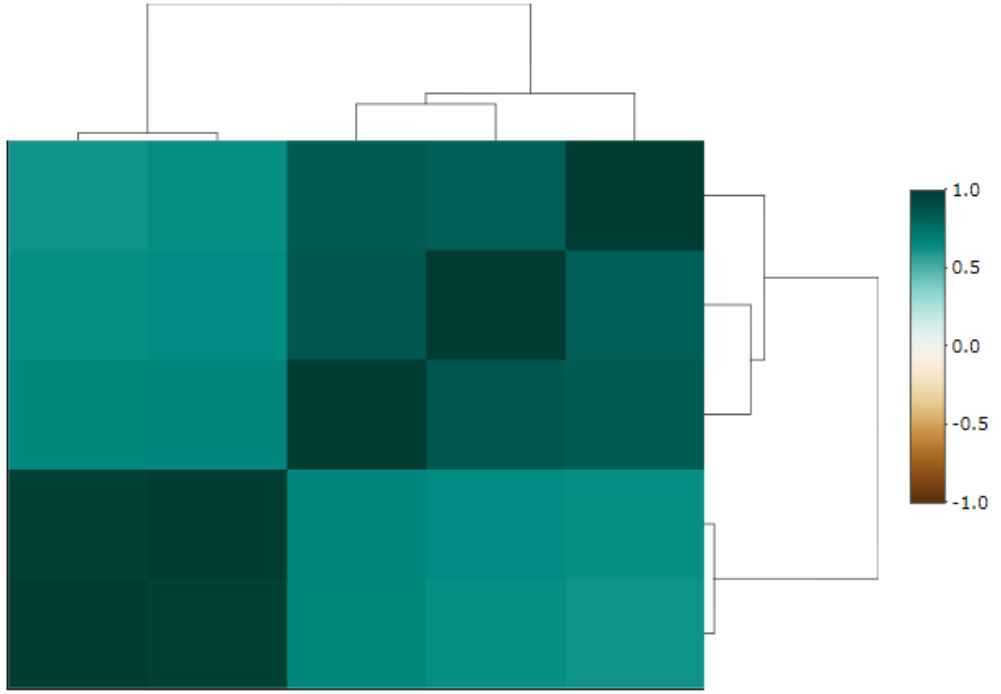


Figure 11.3: The correlation matrix of the predictors chosen by the filtering process based on a threshold on the area under the ROC curve. Hovering over

Testing if there was improvement

Simple filters

- E.g. use t-test of performance of cross-validated samples against model built w/o filter
- Build lots of models with p components (selecting p randomly) and test difference in performance

Recursive feature elimination

AKA backwards selection

- “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.”
- Often used with linear models as well as Random Forest models…

Note: Subset size is a *tuning parameter* …

Reasons for RFE with random forest

"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."

"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. "

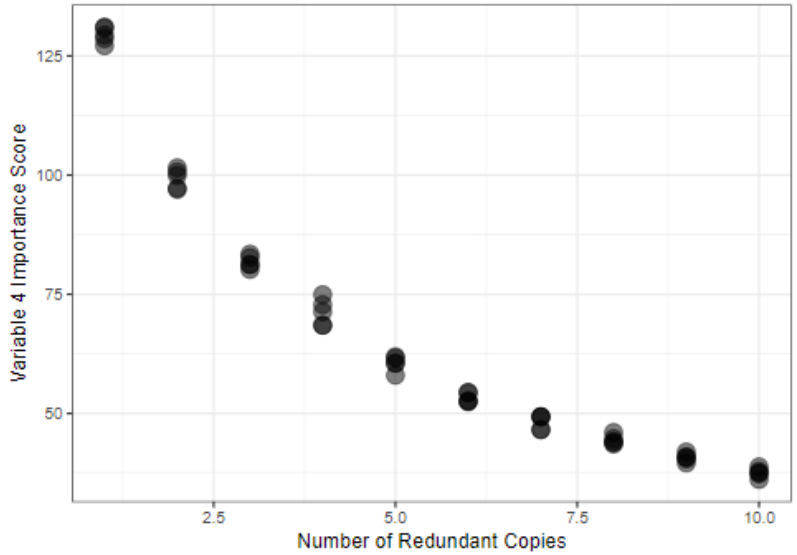


Figure 11.4: The dilution effect of random forest permutation importance scores when redundant variables are added to the model.

Random Forests and filtering techniques

Two things:

- Performance was better without a “correlation filter” applied (in both cases)
- Using the simultaneous feature importance scores performed better than the individual importance scores

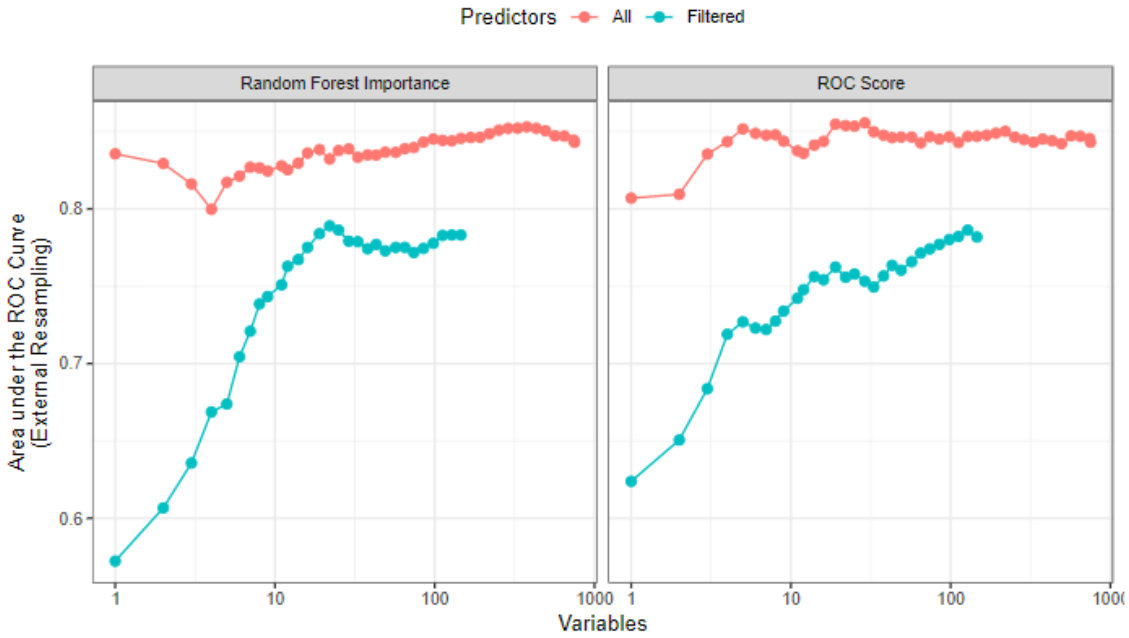


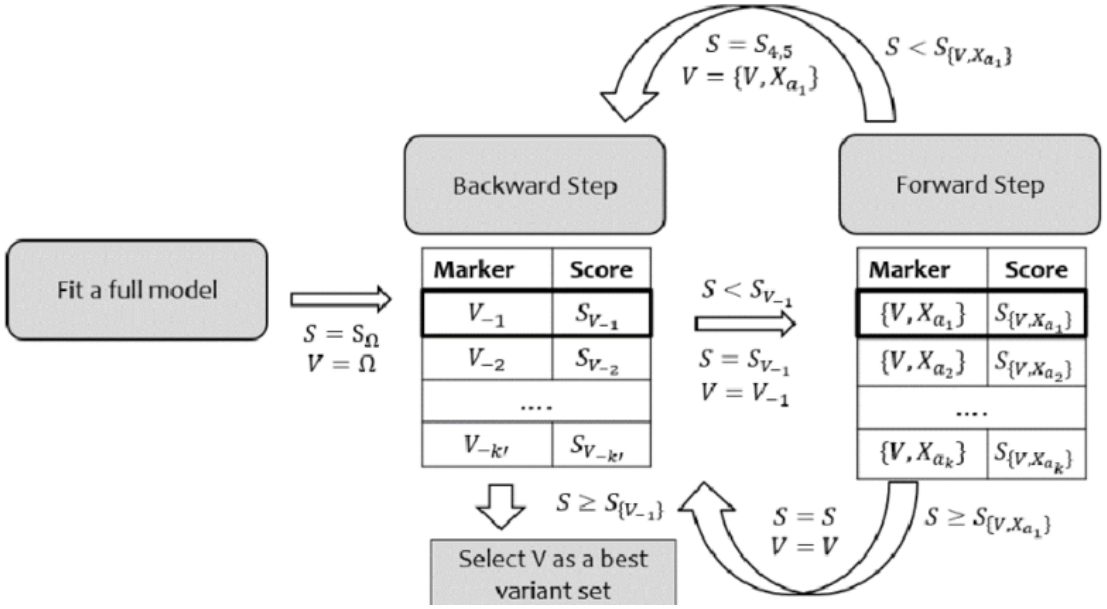
Figure 11.5: The external resampling results for RFE using random forests. The panels reflect how the predictor importances were calculated.

Stepwise Selection

(generally used with linear models)

- “The process begins by creating p linear regression models, each of which uses exactly one of the features.[87] The importance of the features are then ranked by their individual ability to explain variation in the outcome. The amount of variation explained can be condensed into a p-value for convenience. If no features have a p-value less than 0.15, then the process stops. However, if one or more features have p-value(s) less than 0.15, then the one with the lowest value is retained.”
- Generally use AIC or resampled performance (rather than p-values) when using
- “Our recommendation is to avoid this procedure altogether. Regularization methods, such as the previously discussed glmnet model, are far better at selecting appropriate subsets in linear models. If model inference is needed, there are a number of Bayesian methods that can be used (Mallick and Yi 2013; Piironen and Vehtari 2017b, 2017a).”
- Is “less greedy”

Stepwise Selection



term	AIC
+ nerd	36,863
+ firefly	36,994
+ im	37,041
current model	37,059
- white	37,064
- age	37,080
- essay length	37,108

https://www.researchgate.net/profile/Hyunsoo_Kim39/publication/309182017/figure/fig6/AS:667069779542021@1536053237412/Stepwise-selection-B-F-Stepwise-selection-with-backward-to-forward-selects-prediction.png