Feature Construction and Feature Selection

Feature Selection

Feature Selection

- In machine learning and statistics, feature selection, also known as variable selection, attribute selection or variable subset selection, is the process of selecting a subset of **relevant features** (variables, predictors) for use in model construction.

Feature selection techniques are used for several reasons:

- ✓ simplification of models to make them easier to interpret by researchers/users
- ✓ shorter training times,
- ✓ to avoid the curse of dimensionality,
- ✓ enhanced generalization by reducing overfitting

Feature Selection Overview

- Many of the original predictors also may not contain predictive information.
- For many models, predictive performance is degraded as number of uninformative predictors increases.
- Need to appropriately select predictors for modeling
- In supervised feature selection where choice of which predictors to retain is guided by their affect on the outcome.

Goals of Feature Selection

- Trade-off between predictive performance and model interpretability
- Misunderstanding - simply filtering out uninformative predictors will help elucidate which factors are influencing the outcome
- This rationale is problematic for several reasons.
  - Consider case when number of predictors is much greater than number of samples
  - many mutually exclusive subsets of predictors that result in models of nearly equivalent predictive performance (local optima)
- To find the best global solution, i.e. subset of predictors that has best performance, require evaluating all possible predictor subsets and may be computationally infeasible.
- Due to the inherent noise in the available predictors, identified subset may not be true global optimum

Goals of Feature Selection

- Many models are complex - relate the predictors to outcome.
- For such models it is nearly impossible to decipher relationship between any individual predictor and the outcome.
- One approach that attempts to gain insight for an individual predictor in a complex model is to
  - Fix all other selected predictors to a single value,
  - Then observe the effect on the outcome by varying the predictor of interest.
  - This approach is overly simplistic

Goals of Feature Selection

Motivations for removing predictors from a model:

- The primary motivations should be to either mitigate a specific problem in the interplay between predictors and a model, or to reduce model complexity.
  - Some models, notably support vector machines and neural networks, are sensitive to irrelevant predictors.
  - Other models like linear or logistic regression are vulnerable to correlated predictors
  - Removing predictors can reduce the cost of acquiring data or improve the throughput of the software used to make predictions
- Reduce the number of predictors as far as possible without compromising predictive performance.

Classes of Feature Selection Methodologies

- Feature selection mythologies fall into three general classes:
  1. intrinsic (or implicit) methods,
  2. filter methods, and
  3. wrapper methods.
- Intrinsic methods have feature selection naturally incorporated with the modeling process.
- Whereas filter and wrapper methods work to marry feature selection approaches with modeling techniques.

Classes of Feature Selection Methodologies

Intrinsic Methods

- The most seamless and important of the three classes for reducing features are intrinsic methods.
- Some examples include:
  1. Tree- and rule-based models
  2. Multivariate adaptive regression spline (MARS) models
  3. Regularization models
- Relatively fast since selection process is embedded within model fitting process; no external feature selection tool is required.
- Direct connection between selecting features and objective function
- Objective function is statistic that model attempts to optimize.
- **Downside** - It is model dependent.

Classes of Feature Selection Methodologies

- If a model does not have intrinsic feature selection, then some sort of search procedure is required to identify feature subsets that improve predictive performance.
- There are two general classes of techniques for this purpose:
  - Filters and
  - Wrappers

Classes of Feature Selection Methodologies

Filter methods

- Conduct an initial supervised analysis of the predictors to determine which are important features
- Search is performed just once.
- The filter methods often consider each predictor separately.
- Example of filtering - Set of keywords derived from the OkCupid text fields.
- The relationship between occurrence of keyword and outcome was assessed using an odds-ratio.
- The rational to keep or exclude a word (i.e., the filtering rule) was based on statistical significance as well as the magnitude of the odds-ratio.
- The words that made it past filter were then added to logistic regression model.
- Since each keyword was considered separately, they are unlikely to be capturing independent trends in the data.
- For instance words `programming` and `programmer` were both selected using filtering criteria and represent nearly same meaning with respect to outcome.

Classes of Feature Selection Methodologies

Filter methods

- Filters are simple and tend to be fast.
- Effective at capturing the large trends (i.e., individual predictor-outcome relationship) in the data.
- However, prone to over-selecting predictors.
- In many cases, some measure of statistical significance is used to judge "importance" such as a raw or multiplicity adjusted p-value.
- In these cases, there may be a disconnect between objective function for filter method (e.g., significance) and what the model requires (predictive performance).
- In other words, a selection of predictors that meets a filtering criteria like statistical significance may not be a set that improves predictive performance.

Classes of Feature Selection Methodologies

Wrapper methods

- Make use of iterative search procedures that repeatedly supply predictor subsets to model.
- Then use resulting model performance estimate to guide selection of next subset to evaluate.
- If successful, a wrapper method will iterate to a smaller set of predictors that has better predictive performance than the original predictor set.
- Wrapper methods can take either a **greedy** or **non-greedy approach**.
- A greedy search is one that chooses the search path based on the direction that seems best at the time in order to achieve best immediate benefit.
- A non-greedy search method would reevaluate previous feature combinations and would have ability to move in a direction that is initially unfavorable.
- This allows the non-greedy approach to escape being trapped in a local optima.

Classes of Feature Selection Methodologies

Greedy Wrapper Method

- Example - backwards selection (Recursive Feature Elimination or RFE).
- Predictors are initially ranked by some measure of importance.
- An initial model is created using the complete predictor set.
- The next model is based on a smaller set of predictors where the least important have been removed.
- This process continues down a prescribed path until a very small number of predictors are in model.
- Performance estimates are used to determine when too many features have been removed.
- RFE is greedy in that it considers variable ranking as the search direction.
- It does not re-evaluate the search path at any point
- This approach to feature selection will likely fail if there are important interactions between predictors where only one of predictors is significant in presence of the other(s).

Classes of Feature Selection Methodologies

Non-Greedy Wrapper Method

- Examples of non-greedy wrapper methods are genetic algorithms (GA) and simulated annealing (SA).
- The SA method is non-greedy since it incorporate randomness into the feature selection process.
- The random component of the process helps SA to find new search spaces that often lead to more optimal results.

Classes of Feature Selection Methodologies

Wrapper Methods

- Wrappers have potential advantage of searching a wider variety of predictor subsets.
- Most potential to find the globally best predictor subset (if it exists).
- The primary drawback is the computational time required for these methods to find the optimal or near optimal subset.
- The computational time problem can be further exacerbated by the type of model with which it is coupled.
- For example, the models that are in most need of feature selection (e.g., SVMs and neural networks) can be very computationally taxing themselves.
- Another disadvantage of wrappers is that they have most potential to over fit predictors to training data and require external validation.

Classes of Feature Selection Methodologies

Given advantages and disadvantages of different types of feature selection methods, how should one utilize these techniques?

- A good strategy which we use in practice is to start feature selection process with one or more intrinsic methods.
- Note that it is unrealistic to expect that models using intrinsic feature selection would select the same predictor subset, especially if linear and nonlinear methods are being compared.
- If a non-linear intrinsic method has good predictive performance, then we could proceed to a wrapper method that is combined with a non-linear model.
- Similarly, if a linear intrinsic method has good predictive performance, then we could proceed to a wrapper method combined with a linear model.
- If multiple approaches select large predictor sets then this may imply that reducing the number of features may not be feasible.

Effect of Irrelevant Features

How much do extraneous predictors hurt a model?

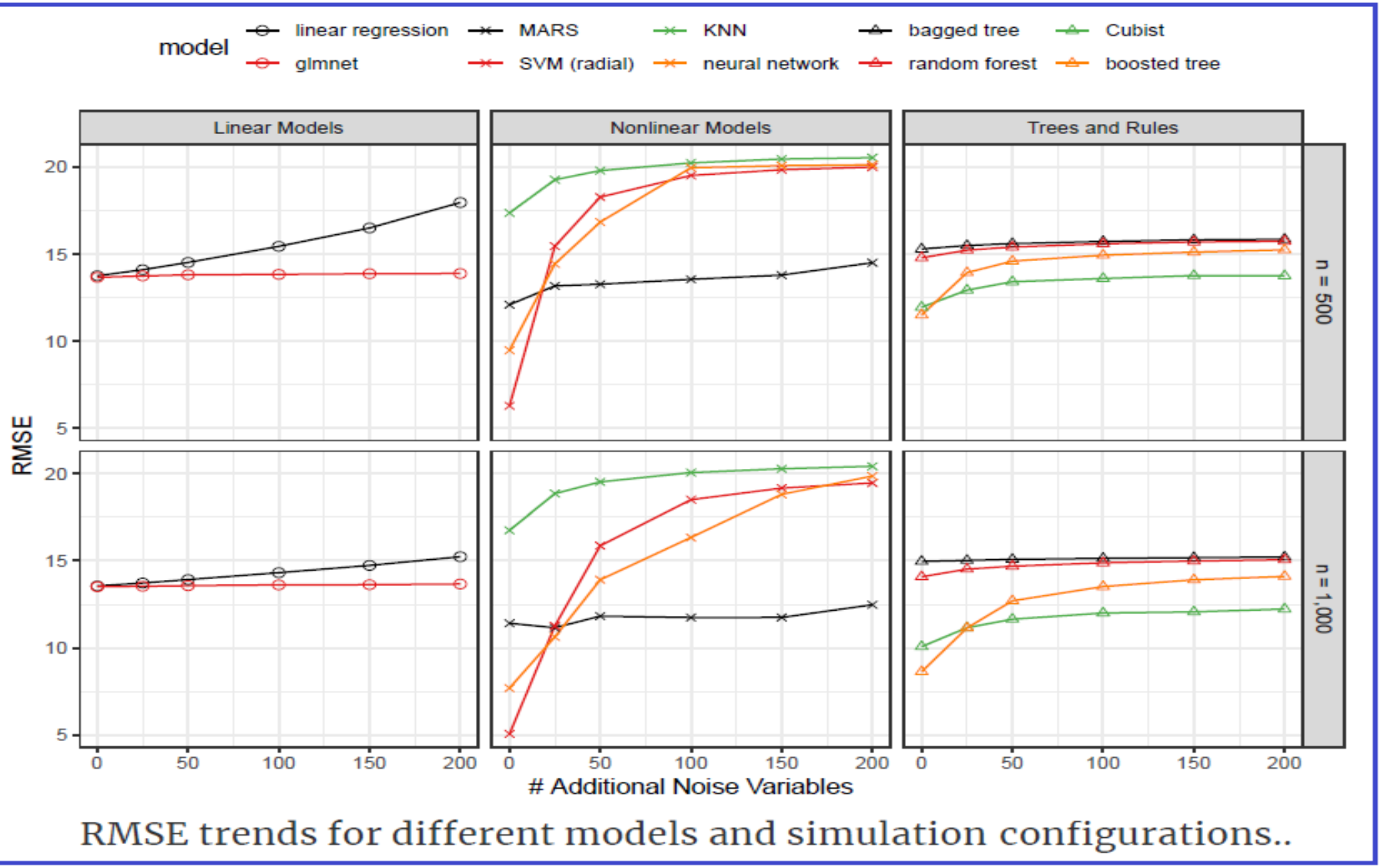
- Predictably, that depends on
  - ✓ type of model,
  - ✓ nature of the predictors,
  - ✓ ratio of the size of the training set to the number of predictors (a.k.a p:n ratio).
- To investigate this, a simulation was used to emulate data with varying numbers of irrelevant predictors and to monitor performance.
- The simulation system consists of a nonlinear function of the 20 relevant predictors:

$$y = x_1 + \sin(x_2) + \log(|x_3|) + x_4^2 + x_5 x_6 + I(x_7 x_8 x_9 < 0) + I(x_{10} > 0) \cdot$$
$$x_{11} I(x_1 1 > 0) + \sqrt{(|x_{12}|)} + \cos(x_{13}) + 2x_{14} + |x_{15}| + I(x_{16} < -1) +$$
$$x_{17} I(x_{17} < -1) - 2x_{18} - x_{19} x_{20} + \epsilon$$

Effect of Irrelevant Features

- To evaluate the effect of extra variables
  - ✓ varying numbers of random standard normal predictors (with no connection to the outcome) were added.
  - ✓ Between 10 and 200 extra columns were appended to the original feature set.
  - ✓ The training set either n=500 or n=1000.
  - ✓ The root mean squared error (RMSE) was used to measure quality of model using a large simulated test set.
- A number of models tuned and trained for each of simulates sets including linear, nonlinear, and tree/rule-based models

Effect of Irrelevant Features



RMSE trends for different models and simulation configurations..

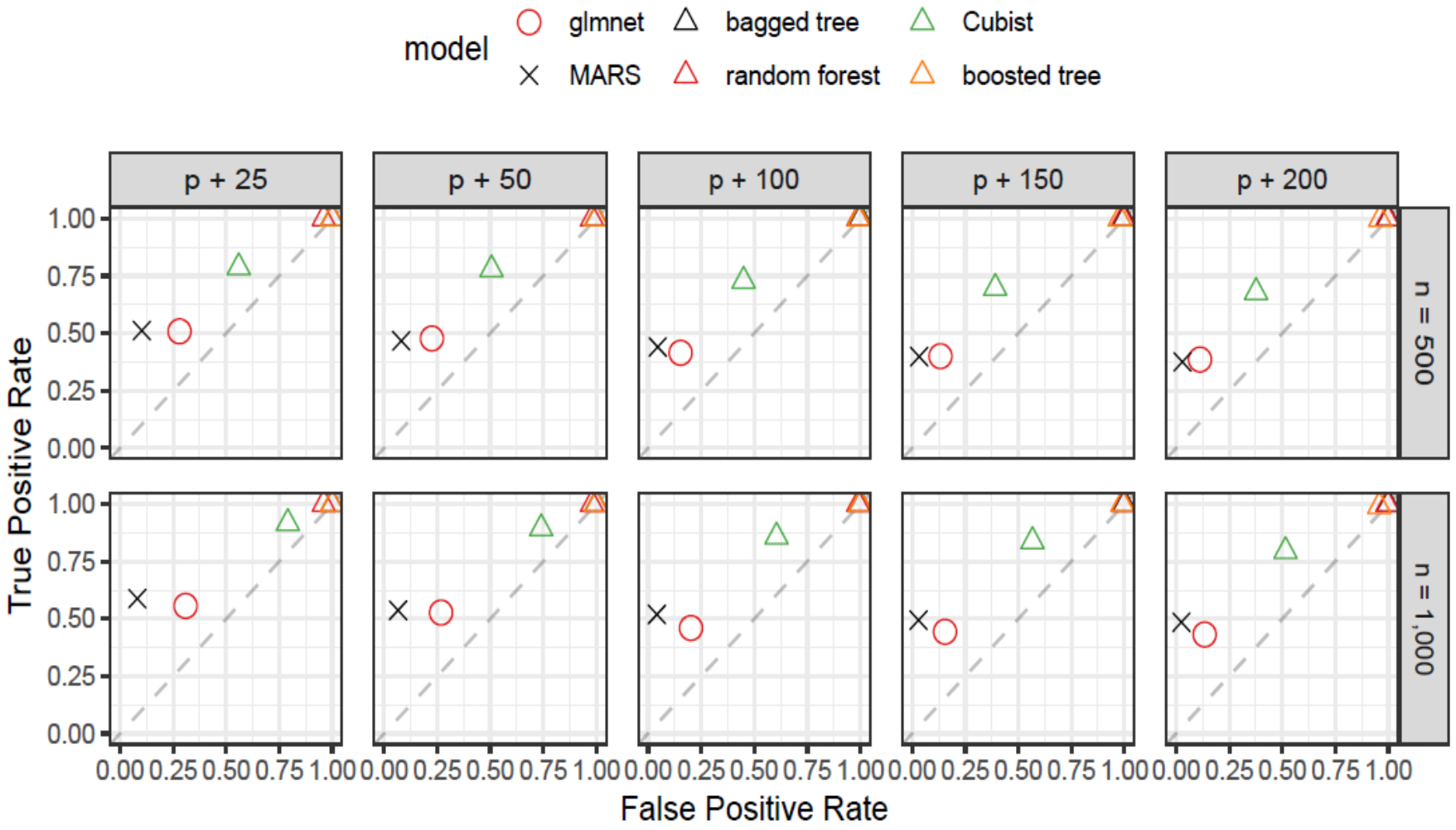
Effect of Irrelevant Features

- Results clearly show that
  - There are a number of models that may require a reduction of predictors to avoid a decrease in performance.
  - For models such as random forest or glmnet, it appears that feature selection may be useful to find a smaller subset of predictors without affecting the model's efficacy.

Effect of Irrelevant Features

- Need to assess how models with built-in feature selection performed at finding the correct predictor subset.
- There are 20 relevant predictors in the data.
- Based on which predictors were selected, a sensitivity-like proportion can be computed that describes the rate at which the true predictors were retained in the model.
- Similarly, number of irrelevant predictors that were selected gives a sense of false positive rate (i.e., one minus specificity) of the selection procedure.

Effect of Irrelevant Features



ROC-like plots for feature selection results in simulated data sets.

Overfitting to Predictors and External Validation

- Model may overfitting the available data during selection of model tuning parameters.
- Risk of finding tuning parameter values that over-learn the relationship between the predictors and the outcome in the training set.
- When models over-interpret patterns in the training set, the predictive performance suffers with new data.
- The solution to this problem is to evaluate tuning parameters on a data set that is not used to estimate model parameters.
- An analogous problem can occur when performing feature selection.
- For many data sets it is possible to find a subset of predictors that has good predictive performance on the training set but has poor performance when used on a test set or other new data set.
- The solution to this problem is similar to the solution to the problem of overfitting: feature selection needs to be part of the resampling process.

Overfitting to Predictors and External Validation

- Unfortunately, practitioners often combine feature selection and resampling inappropriately.
- Most common mistake is to only conduct resampling inside of feature selection procedure.

```
1 Rank the predictors using the training set;
2 for subset sizes 5 to 1 do
3     for each resample do
4         Fit model with subset on the analysis set.;
5         Predict the assessment set.;
6     end
7     Determine the best subset using resampled performance;
8     Fit the best subset using the entire training set;
9 end
```

Overfitting to Predictors and External Validation

There are two key problems with this procedure:

1. Since the feature selection is external to the resampling, resampling cannot effectively measure the impact (good or bad) of the selection process
2. The same data are being used to measure performance and to guide the direction of the selection routine. This is analogous to fitting a model to the training set and then re-predicting the same set to measure performance

Overfitting to Predictors and External Validation

- A better way of combining feature selection and resampling is to make feature selection a **component of the modeling process**.
- Feature selection should be incorporated the same way as preprocessing and other engineering tasks.
- What we mean is that an appropriate way to perform feature selection is to do this inside of the resampling process .

Overfitting to Predictors and External Validation

```
Split data into analysis and assessment sets;
for each resample do
    Rank the predictors using the analysis set;
    for subset sizes 5 to 1 do
        Fit model with subset on the analysis set;
        Predict the assessment set.;
    end
    Average the resampled performance for each model and subset size;
    Choose the model subset with the best performance;
    Fit the best subset using the entire training set;
end
```

Filter Methods

- Simple univariate filters and
- Recursive Feature Elimination (RFE)

Illustrative Data: Predicting Parkinson's Disease

- A group of 252 patients, 188 of whom had a previous diagnosis of Parkinson's disease, were recorded speaking a particular sound three separate times.
- Several signal processing techniques were then applied to each replicate to create 750 numerical features.
- The objective was to use the features to classify patients' Parkinson's disease status.
- Many of the features consisted of related sets of fields produced by each type of signal processor (e.g., across different sound wavelengths or sub-bands).
- Resulting data have extreme amount of multi-collinearity;
  - about 10,750 pairs of predictors have absolute rank correlations greater than 0.75.

Simple Filters

- Most basic approach - screen the predictors to see relationship with the outcome prior to including them in a model.
- To do this, a numeric scoring technique is required to quantify strength of relationship.
- Using the scores, the predictors are ranked and filtered with either a threshold or by taking the top predictors.
- Scoring the predictors can be done separately for each predictor, or can be done simultaneously across all predictors (depending on the technique that is used).
- If the predictors are screened separately, there are a large variety of scoring methods.
- Techniques used depends on the type of predictor and outcome.

Simple Filters

When screening individual categorical predictors, there are several options depending on the type of outcome data.

- When the outcome is categorical,
  - Relationship between the predictor and outcome forms a contingency table.
  - When there are three or more levels for the predictor, the degree of association between predictor and outcome can be measured $\chi 2$ with statistics such as (chi-squared) tests or exact methods
  - When there are exactly two classes for the predictor, the odds-ratio can be an effective choice.

Simple Filters

When screening individual categorical predictors, there are several options depending on the type of outcome data.

- When the outcome is numeric, and the categorical predictor has two levels, then a basic t-test can be used to generate a statistic.
- ROC curves and precision-recall curves can also be created for each predictor and the area under the curves can be calculated.
- When the predictor has more than two levels, the traditional ANOVA -statistic can be calculated.

Simple Filters

When the predictor is numeric, the following options exist:

- When outcome is categorical, the same tests can be used in the case above where the predictor is categorical and the outcome is numeric.
- Roles are simply reversed in the t-test, curve calculations and t-test.
- When there are a large number of tests or if the predictors have substantial multicollinearity, the correlation adjusted t-scores and are a good alternative to simple ANOVA statistics.
- When outcome is numeric, a simple pairwise correlation (or rank correlation) statistic can be calculated.
- Alternatively, a generalized additive model (GAM) can be used.

Simple Filters

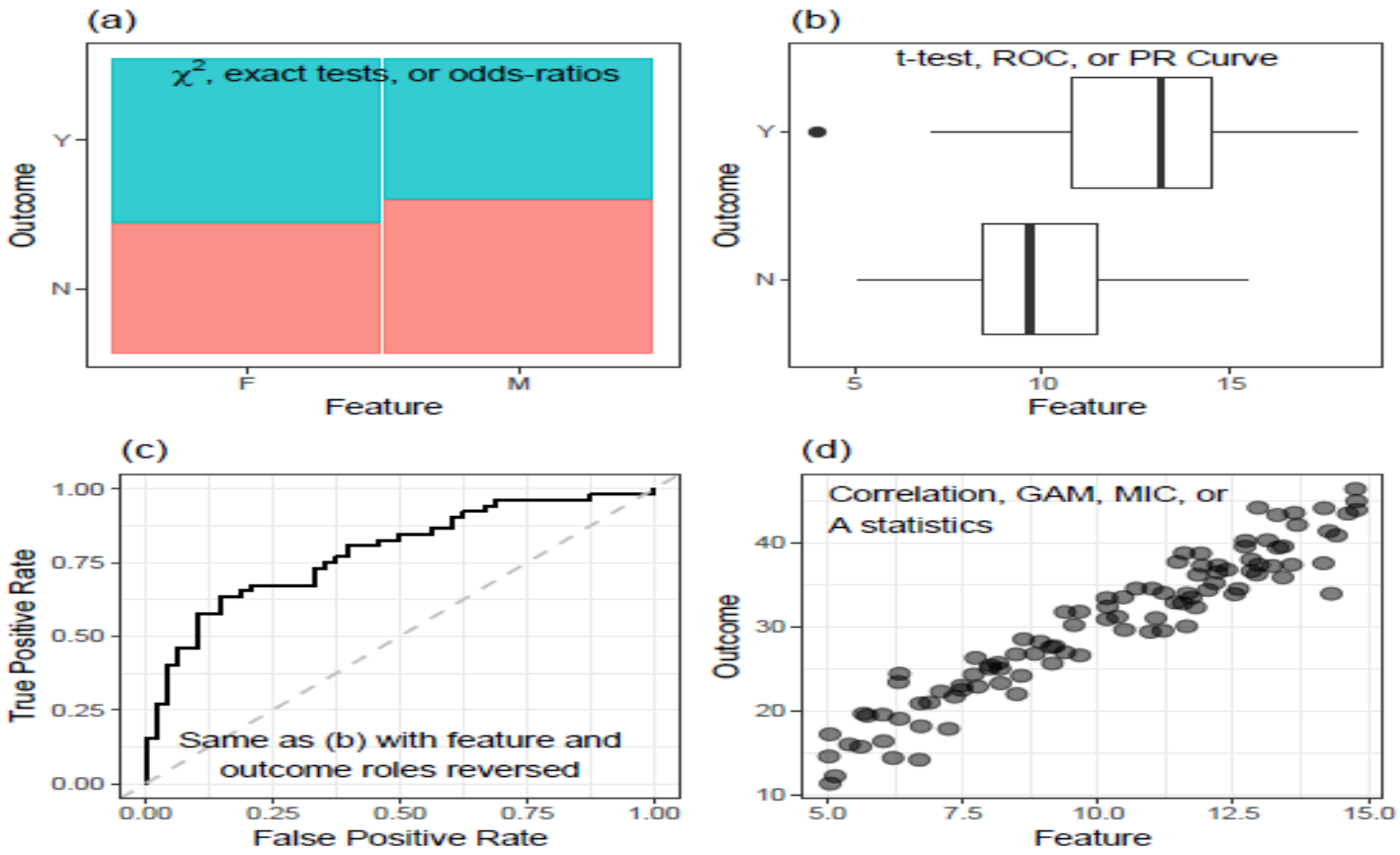


Figure 11.1: A comparison of the simple filters that can be applied to feature and outcome type combinations: (a) categorical feature and categorical outcome, (b) continuous feature and categorical outcome, (c) categorical feature and continuous outcome, and (d) continuous feature and continuous outcome.

Simple Filters

- Most data sets contain a mix of predictor types.
- Challenging to arrive at a ranking of the predictors since their screening statistics are on different scales.
- For example, an odds-ratio and a t-statistic are not compatible since they are on different scales.
- In many cases, each statistic can be converted to a p-value so that there is a commonality across the screening statistics.
- A p-value stems from statistical framework of hypothesis testing.
- Each statistic can be converted to a p-value, but this conversion is easier for some statistics than others.
- For instance, converting a t-statistic to a p-value is a well-known process.
- It is not easy to convert an AUC (Area under the Receiver Operator Characteristic) Curve to a p-value
- A solution to this problem is by using a permutation method.

Simple Filters

Permutation Method for converting a -statistic to a p-value

- Can be applied to any statistic to generate a p-value
- For a selected predictor and corresponding outcome, the predictor is randomly permuted, but the outcome is not
- Statistic of interest is then calculated on the permuted data.
- Disconnects observed predictor and outcome relationship, thus creating no association between the two
- Same predictor is randomly permuted many times to generate a distribution of statistics (distribution of no association).
- Statistic from the original data can then be compared to distribution of no association to get a probability, or p-value

(Refer Robnik-Sikonja, M, and I Kononenko. 2003. "Theoretical and Empirical Analysis of Relieff and Rrelieff." *Machine Learning* 53 (1): 23–69 for more details).

Simple Filters

- Simple filters are effective at identifying individual predictors that are associated with the outcome.
- However, these filters are very susceptible to finding predictors that have strong associations in the available data but do not show any association with new data.
- In the statistical literature, these selected predictors are labeled as false positives.
- An entire sub-filed of statistics has been devoted to developing approaches for minimizing the chance of false positive findings, especially in the context of hypothesis testing and p-values.
- One approach to reducing false positives is to adjust the p-values to effectively make them larger and thus less significant.

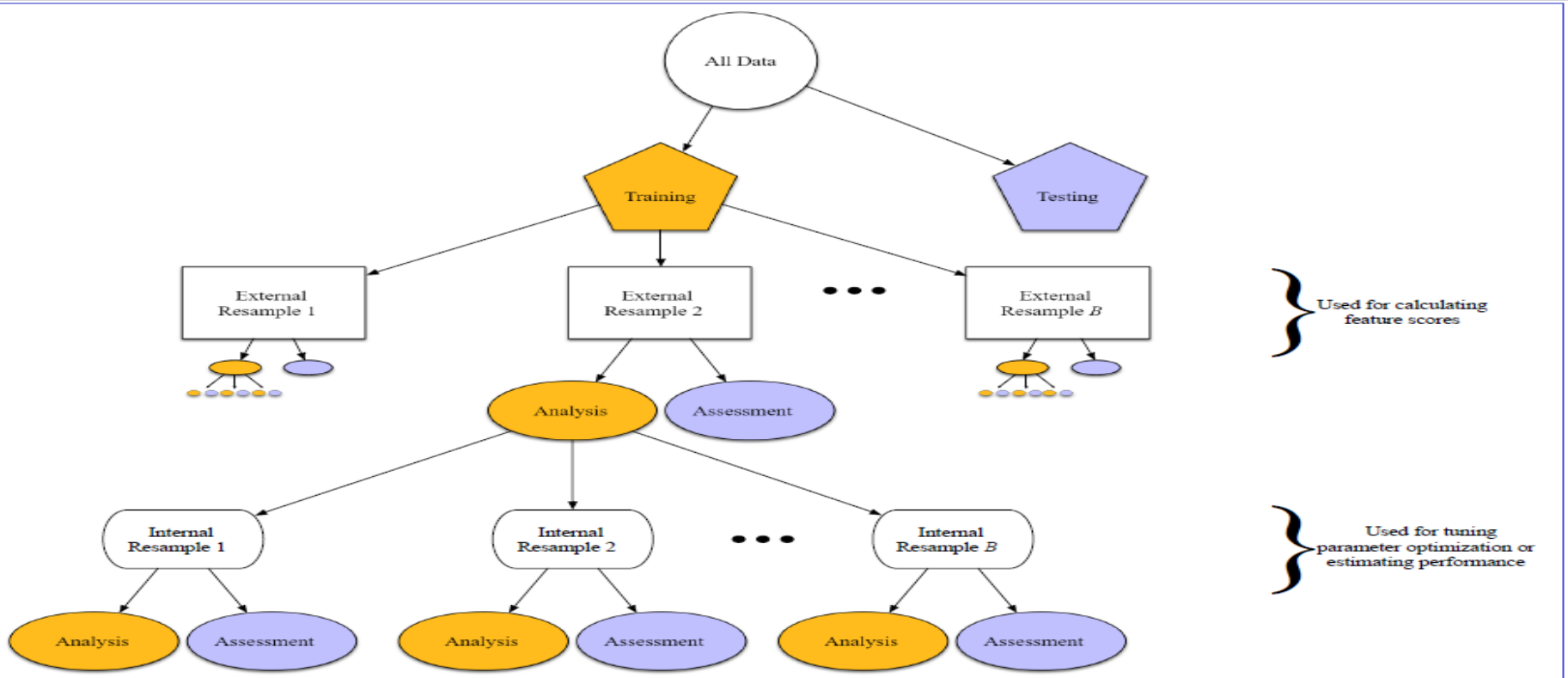
Simple Filters

- False positive findings can be minimized by using an independent set of data to evaluate the selected features.
- Parallel to the context of identifying optimal model tuning parameters
- Cross validation is used to identify an optimal set of tuning parameters and avoid overfit.
- Model building process needs to accomplish two objectives:
  1. To identify an effective subset of features, and
  2. To identify the appropriate tuning parameters such that the selected features and tuning parameters do not overfit.

Simple Filters

- 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.
- Filtering must be done independently of the model tuning.
- Instead, must incorporate another layer of cross-validation.
- The first layer, or external layer, is used to filter features.
- Then second layer (the “internal layer”) is used to select tuning parameters.

Simple Filters



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

- Conducting feature selection can be computationally costly.
- No. of models constructed & evaluated is **I x E x T**,
  where I is the number if internal resamples, E is the number of external resamples, and T is the total number of tuning parameter combinations

Recursive Feature Elimination (RFE)

- RFE is 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, 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 predictors based on the importance rankings.
- The optimal subset is then used to train the final model.

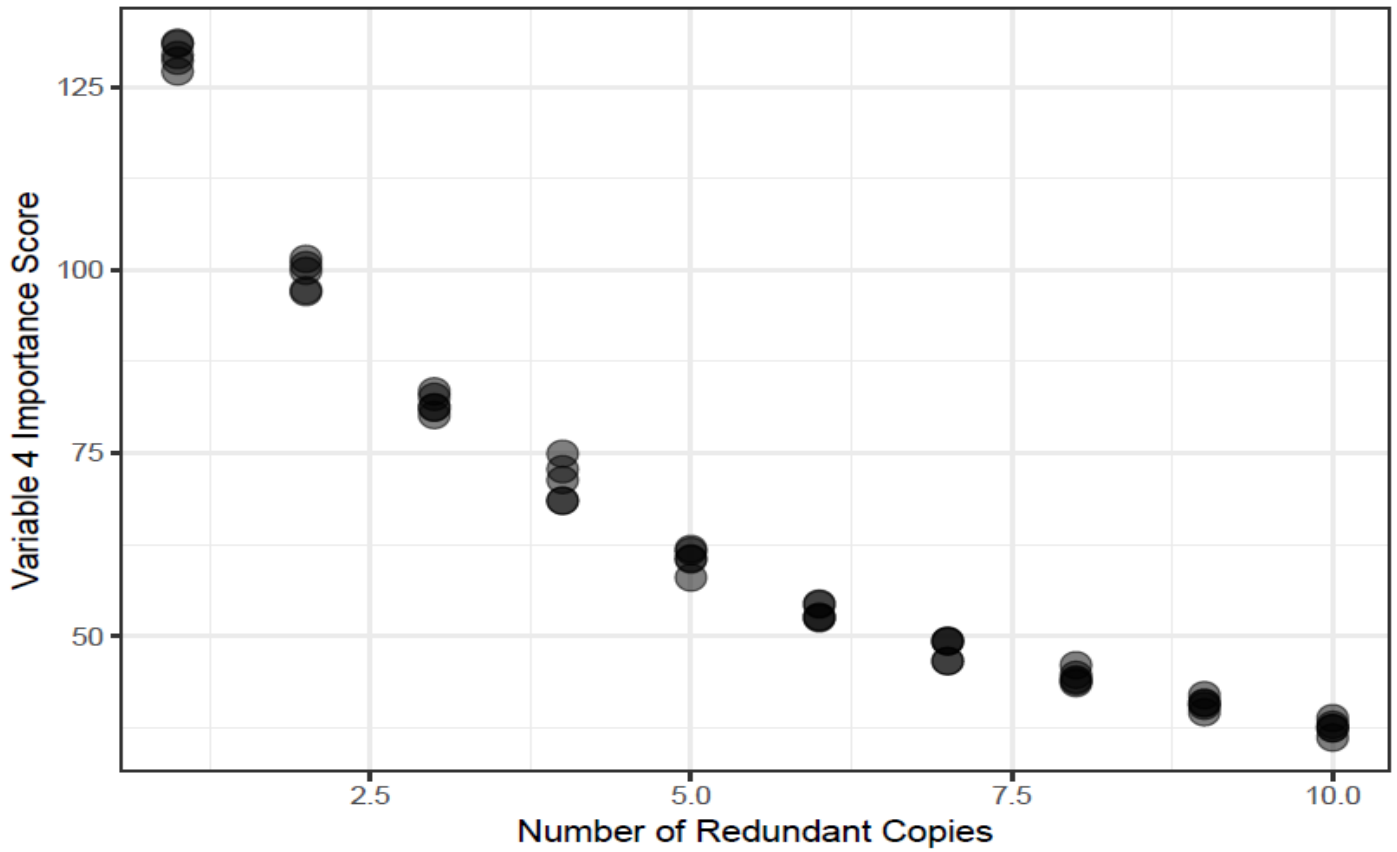
Recursive Feature Elimination (RFE)

- Not all models can be paired with the RFE method
- 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.
- 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.
- Sime models like Random forest benefit more from use of RFE

Recursive Feature Elimination (RFE)

- Backwards selection is used with random forest models for two reasons.
  1. First, random forest tends not to exclude variables from the prediction equation
  2. Model has a well-known internal method for measuring feature importance.
- Notable issue with measuring importance in trees is related to multicollinearity
- If highly correlated predictors, then which predictor is chosen for partitioning samples is essentially a random selection.
- Predictive performance of the ensemble of trees is unaffected by highly correlated, useful features.
- However, redundancy of the features dilutes the importance scores.

Recursive Feature Elimination (RFE)



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

Stepwise Selection

- Developed as a feature selection technique for linear regression models.
- The forward stepwise regression approach uses a sequence of steps to allow features to enter or leave the regression model one-at-a-time.
- Often this procedure converges to a subset of features.
- The entry and exit criteria is commonly based on a p-value threshold.
- A typical entry criterion is that a p-value must be less than 0.15 for a feature to enter the model and must be greater than 0.15 for a feature to leave the model.
- The process begins by creating *p* linear regression models, each of which uses exactly one of the features.
- The importance of the features are then ranked by their individual ability to explain variation in the outcome.

Stepwise Selection

- 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.
- Process begins by creating p linear regression models
- Features are removed based on p-values.
- In the next step, p-1 linear regression models are built.
- These models consist of the feature selected in the first step as well as each of the other features individually.
- Each of the additional features are evaluated, and the best feature that meets the inclusion criteria is added to the selected feature set.

Stepwise Selection

- This approach is problematic for several reasons, and a large literature exists that critiques this method.
- Stepwise selection has two primary faults
  1. Over selection of features
  2. Model overfitting

Thank You !!!