Feature Engineering

Chapter 10

Stephen Kimel

Feature Selection

- There is a “desire to have a model that has the best predictive ability *and* is interpretable.”
- “trade-off between predictive performance and model interpretability”
- “A misunderstanding of this trade-off leads to the belief that simply filtering out uninformative predictors will help elucidate which factors are influencing the outcome.”
- Some models are resistant to irrelevant and/or correlated predictors and some are not.
- Goal is to “reduce the number of predictors as far as possible without compromising predictive performance.”

Methods of Feature Selection

Intrinsic/Implicit

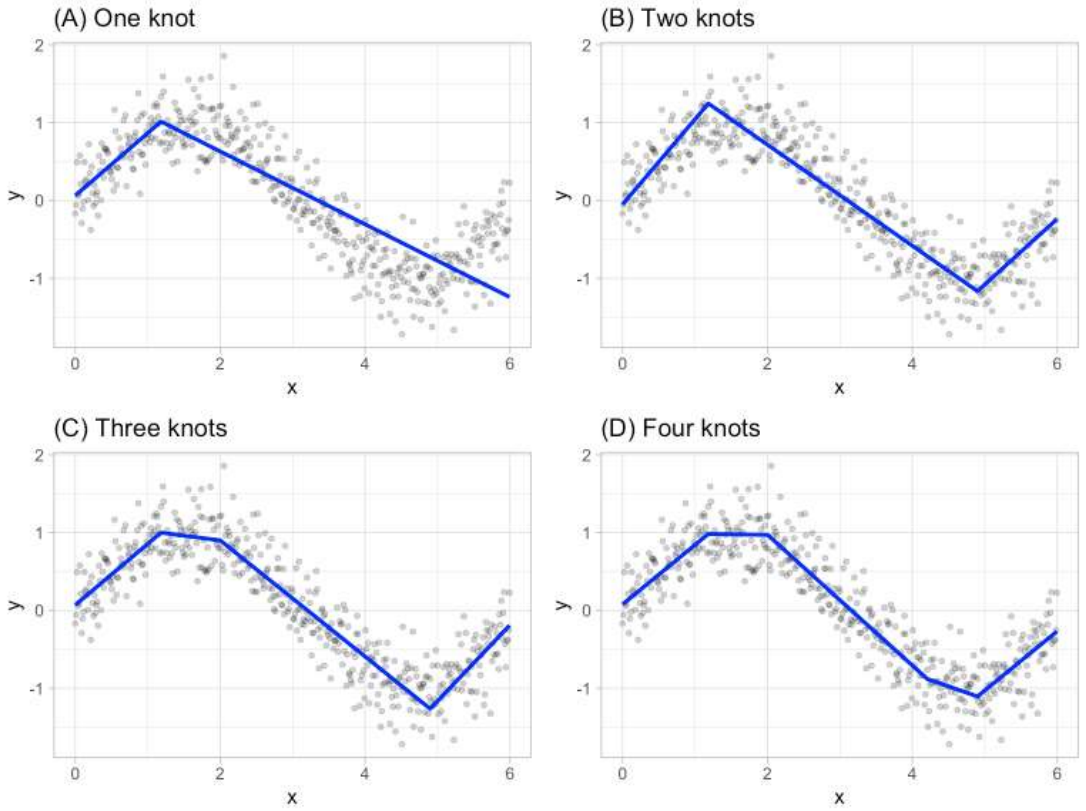
Naturally incorporated into the modeling process

- Tree-based models

Intrinsic/Implicit

Naturally incorporated into the modeling process

- Tree-based models
- MARS Models (earth package in R)



Intrinsic/Implicit

Naturally incorporated into the modeling process

- Tree-based models
- MARS Models (earth package in R)
- Lasso Regression (standardize and penalize)
- Pros:
  - Easy
  - Fast (not as much preprocessing)
  - Connection between objective function (what we are trying to minimize or maximize) and feature selection
- Cons:
  - Can only be used in the specific model

Filter

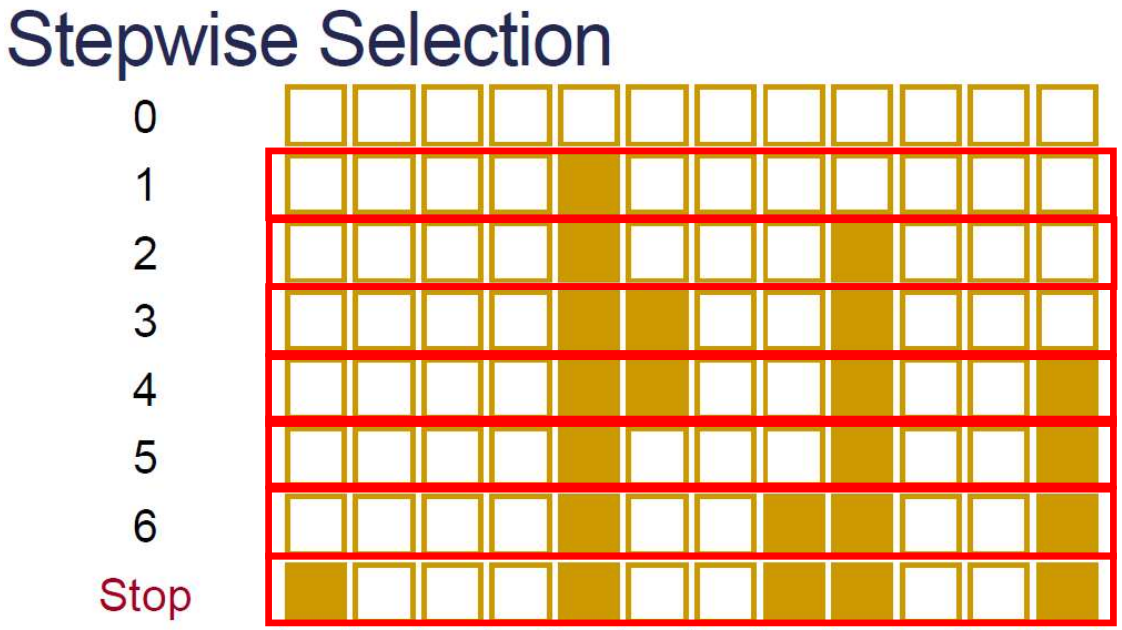
Single search to find important variables

- Simple and fast
- Odds-ratio threshold
- More in Chapter 11

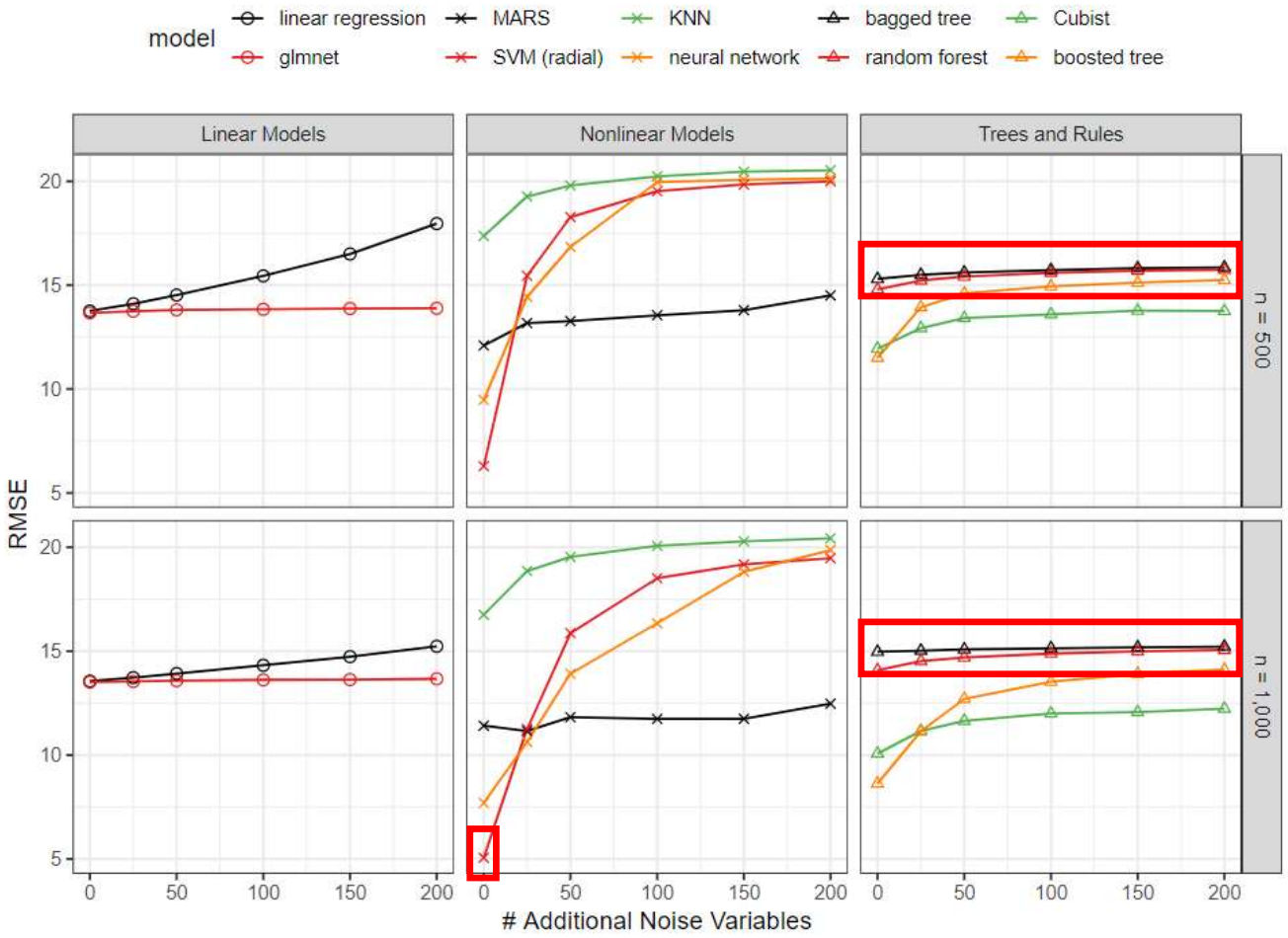
Wrapper

Multiple steps adding and/or removing variables

- Greedy – What is the best decision right now?
- Non-greedy – Look back at previous decision to determine if they are still the right ones
- Tend to overfit the training set
- More in chapters 11 and 12



Effects of Irrelevant Predictors



Did the Models Choose the Right Predictors?

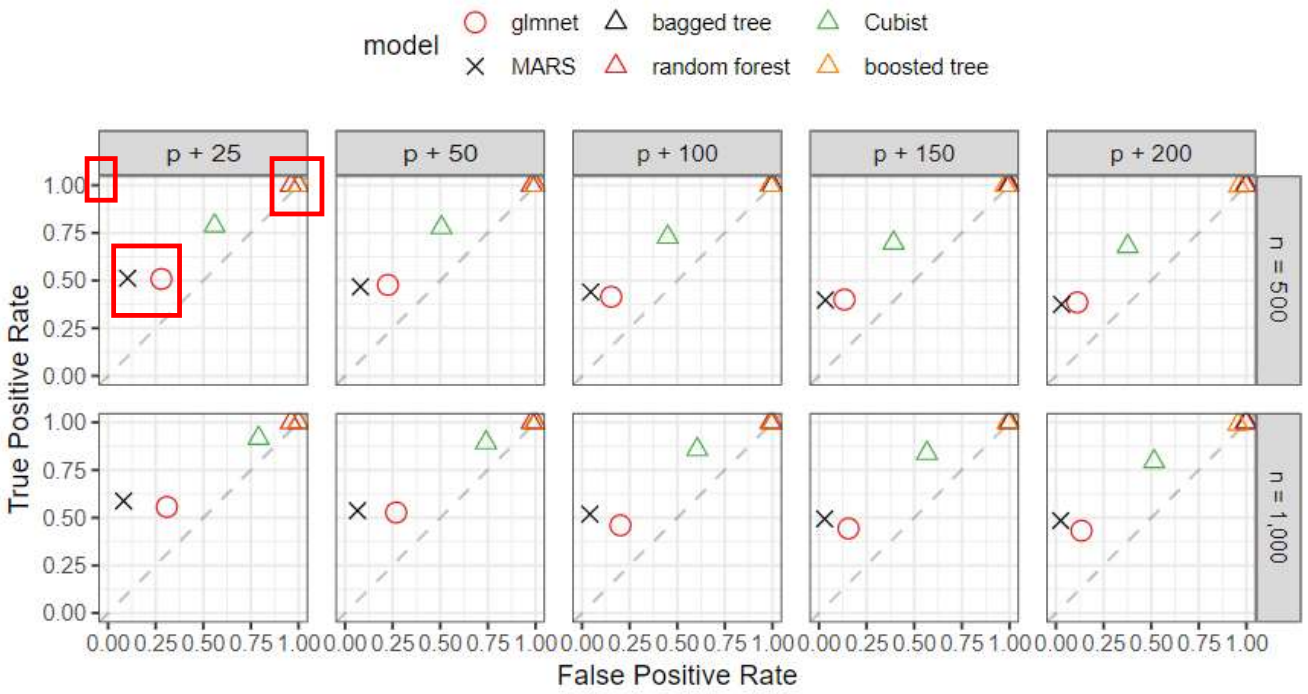


Figure 10.3: ROC-like plots for the feature selection results in the simulated data sets.

Cautions with Feature Selection

```
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
```

A Better Way

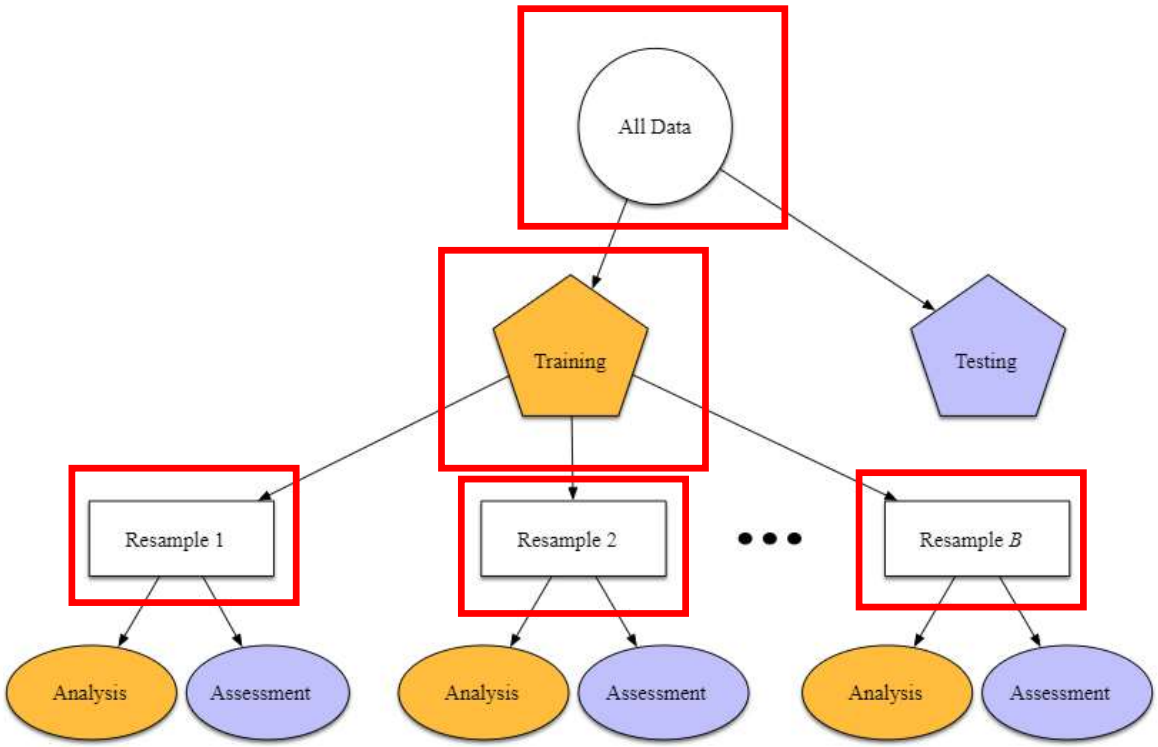
0 Split data into Train and Test

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

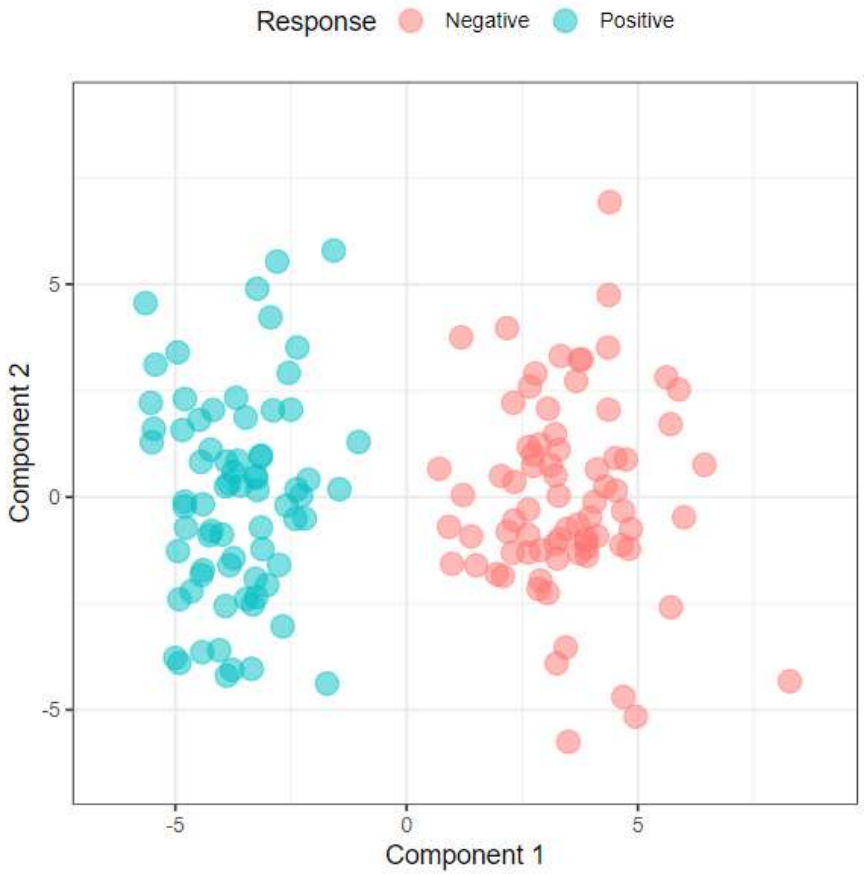
Case Study

- “In this problem, a researcher had collected 75 samples from each of the two classes. There were approximately 10,000 predictors for each data point. The ultimate **goal** was to attempt to identify a subset of predictors that had the ability to classify samples into the correct response with an **accuracy of at least 80%**.”
- “The researcher chose to use 70% of the data for a training set, 10-fold cross-validation for model training, and the implicit feature selection methods of the glmnet and random forest.”
- Accuracy was under 60% ☹
- “The logic was then to first identify and select predictors that had a univariate signal with the response. A t-test was performed for each predictors, and predictors were ranked by significance of separating the classes. The top 300 predictors were selected for modeling.”

When Were the t-tests Done?



What What!?


Response
Negative
Positive
Component 2
5
0
-5
-5
0
5
Component 1

What What!?

“Regrettably, because feature selection was performed outside of resampling, the apparent signal that was found was simply due to random chance.”