Data 624 - Group 2

Homework Part 2

Vinicio Haro
Sang Yoon (Andy) Hwang
Julian McEachern
Jeremy O'Brien
Bethany Poulin

16 December 2019

Contents

Getting Started

Overview

Include details on our process in creating this document.

Dependencies

```
# Predicitve Modeling
libraries("AppliedPredictiveModeling", "mice", "caret", "tidyverse",
    "impute", "pls", "caTools", "mlbench")
# Formatting Libraries
libraries("default", "knitr", "kableExtra", "gridExtra", "sqldf",
    "tibble")
# Plotting Libraries
libraries("ggplot2", "grid", "ggfortify")
```

Assignment 1

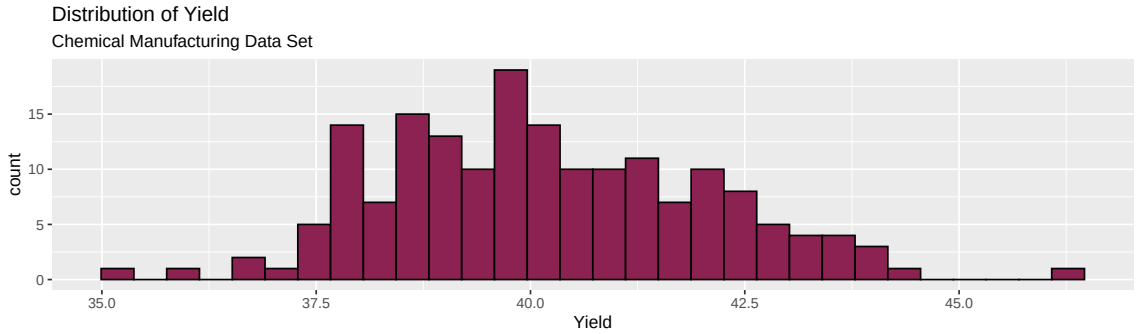
Kuhn and Johnson 6.3

A chemical manufacturing process for a pharmaceutical product was discussed in Sect.1.4. In this problem, the objective is to understand the relationship between biological measurements of the raw materials (predictors), measurements of the manufacturing process (predictors), and the response of product yield. Biological predictors cannot be changed but can be used to assess the quality of the raw material before processing. On the other hand, manufacturing process predictors can be changed in the manufacturing process. Improving product yield by 1% will boost revenue by approximately one hundred thousand dollars per batch:

(a). Start R and use these commands to load the data:

The matrix processPredictors contains the 57 predictors (12 describing the input biological material and 45 describing the process predictors) for the 176 manufacturing runs. yield contains the percent yield for each run.

```
data("ChemicalManufacturingProcess")
```



(b). A small percentage of cells in the predictor set contain missing values. Use an imputation function to fill in these missing values (e.g., see Sect. 3.8).

ManufacturingProcess03 has the largest volume of missing data followed by ManufacturingProcess11. Given that each variable has less than 25% of data missing, we should introduce methods of imputation. For our purposes, we will use the MICE method. MICE is formally known as Multiple Imputation with Chained Equations. On a high level, MICE is built off a tehcnique known as the Gibbs sampler. The Gibbs sampler is a Markov chain based on Monte Carlo. MICE iterates drawing estimates of missing values and parameters related to the distribution of said variables. Chained equations are generally faster than the monte carlo based Gibbs sampler. MICE has 5 imputations listed as its default. predictive mean matching is also a default method for MICE. PMM does a better job at keeping non-linear relationships within individual variables.

In addition to MICE, we drop variables that have near zero variance, however we point out that only one variable was dropped. We still include it as a process step to follow the literature's specifications. After completing MICE, we no longer had missing data in our set. We examined other imputation methods such as KNN but determined that there was no significant change in the summary statistics across different imputation methods.

Table 1: Distribution of Missing data in Chemical Manufacturing Process

Variable	Count	Variable	Count
ManufacturingProcess03	15	BiologicalMaterial01	0
ManufacturingProcess11	10	BiologicalMaterial02	0
ManufacturingProcess10	9	BiologicalMaterial03	0
ManufacturingProcess25	5	BiologicalMaterial04	0
ManufacturingProcess26	5	BiologicalMaterial05	0
ManufacturingProcess27	5	BiologicalMaterial06	0
ManufacturingProcess28	5	BiologicalMaterial07	0
ManufacturingProcess29	5	BiologicalMaterial08	0
ManufacturingProcess30	5	BiologicalMaterial09	0
ManufacturingProcess31	5	BiologicalMaterial10	0
ManufacturingProcess33	5	BiologicalMaterial11	0
ManufacturingProcess34	5	BiologicalMaterial12	0
ManufacturingProcess35	5	ManufacturingProcess09	0
ManufacturingProcess36	5	ManufacturingProcess13	0
ManufacturingProcess02	3	ManufacturingProcess15	0
ManufacturingProcess06	2	ManufacturingProcess16	0
ManufacturingProcess01	1	ManufacturingProcess17	0
ManufacturingProcess04	1	ManufacturingProcess18	0
ManufacturingProcess05	1	ManufacturingProcess19	0
ManufacturingProcess07	1	ManufacturingProcess20	0
ManufacturingProcess08	1	ManufacturingProcess21	0
ManufacturingProcess12	1	ManufacturingProcess32	0
ManufacturingProcess14	1	ManufacturingProcess37	0
ManufacturingProcess22	1	ManufacturingProcess38	0
ManufacturingProcess23	1	ManufacturingProcess39	0
ManufacturingProcess24	1	ManufacturingProcess42	0
ManufacturingProcess40	1	ManufacturingProcess43	0
ManufacturingProcess41	1	ManufacturingProcess44	0

(c). Split the data into a training and a test set, pre-process the data, and tune a model of your choice from this chapter. What is the optimal value of the performance metric?

We will build a PLS model also known as partial least squares. PLS is a statistical method that fits a linear regression model by projecting the feature variables and response variable to some new space via a mapping function. Because of this projection mechanisim, for both predictors and the response, the method becomes bilinear or simply known as

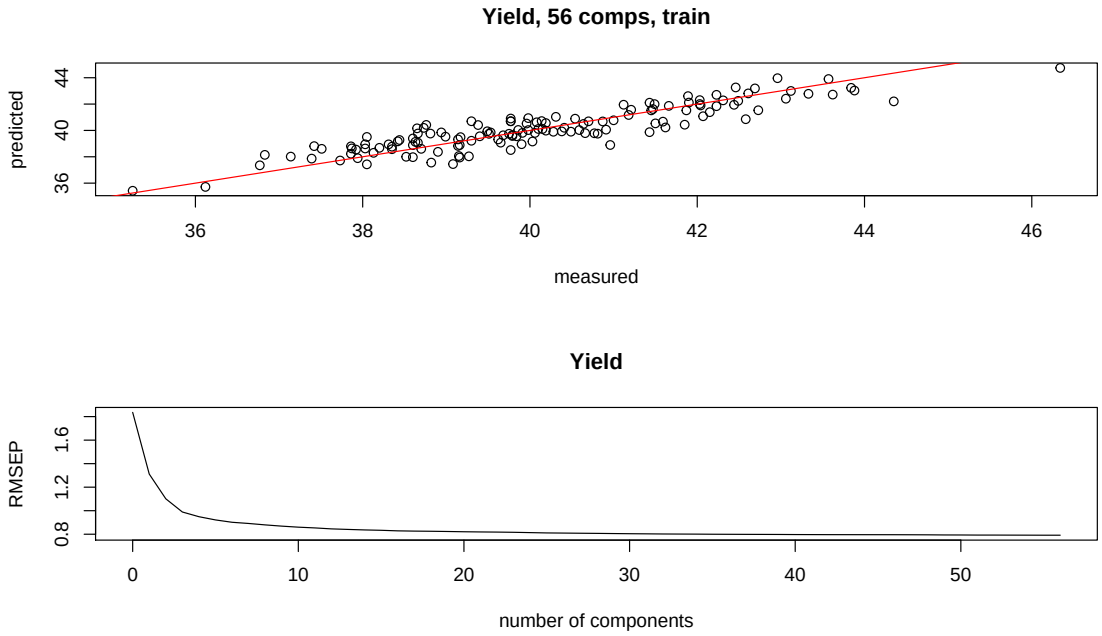
linear with respect to to each of the variable types. PLS also has certain advantages over other methods such as being more robust to dealing with issues arising from multicolinearity.

For our PLS model, we partitioned the data by taking 80% of the data as training and the remaining 20% as testing subsets. We also apply center and scaling arguments set to true. We built a standard PLS model and evaluated the root mean summary areas to determine the optimal number of components to select. In our case, 41 components seemed to cause the root meean summary area to flatten. Any changes after 41 were marginal. We generate performance metrics agasint the training data for our baseline PLS model with all components.

Table 2: PLS Performance Metrics on Training Subset

	x
RMSE	1.3116314
Rsquared	0.4896581
MAE	1.0584889

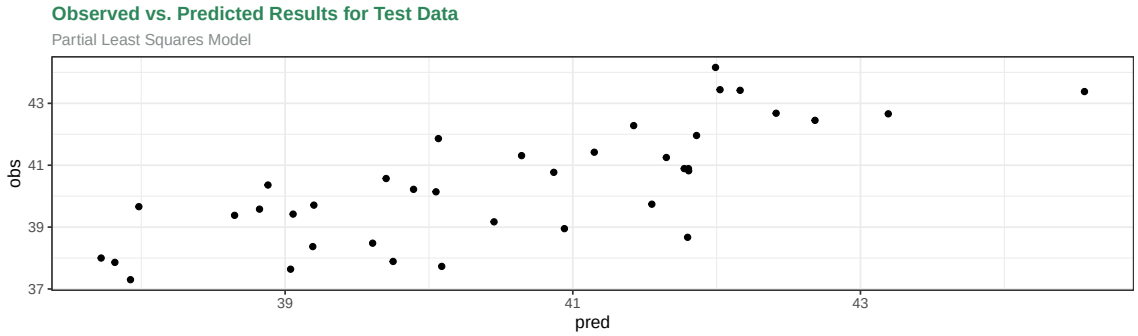
Our Baseline PLS model generates a RMSE of 1.3. In addition, the model captures 49% of data variability. We include the visualizations pertaining to the RMSEP values against Yield. At roughly 41 components, the RMSEP flattens.



(d). Predict the response for the test set. What is the value of the performance metric and how does this compare with the resampled performance metric on the training set?

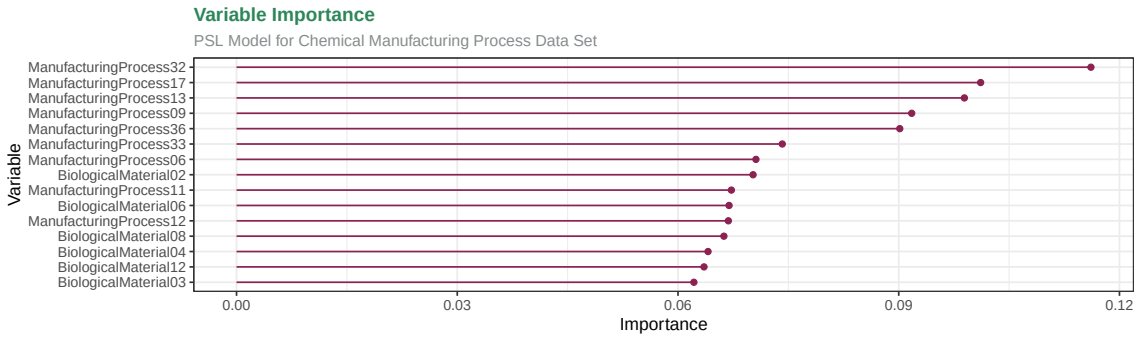
Table 3: PLS Performance Metrics on Test Subset

	x
RMSE	1.222201
Rsquared	0.579859
MAE	0.988292



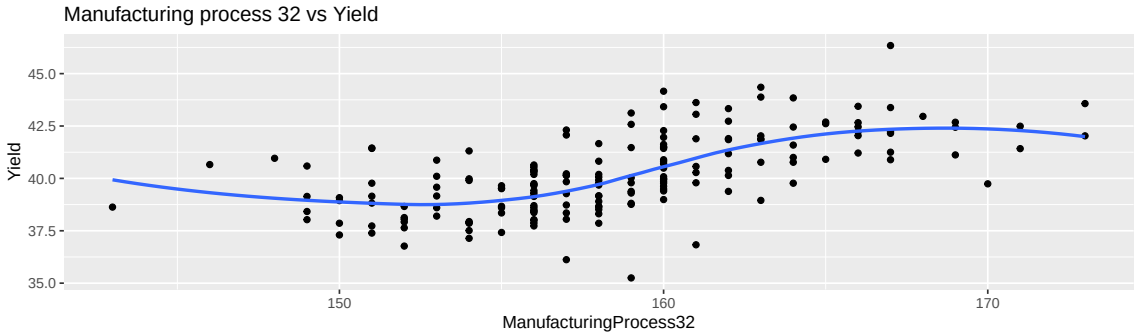
We see an increased R squared against the test data with 56 percent of the data variability accounted for. We also see the RMSE decrease to 1.2 from 1.3. There is also a slight decrease in the MAE. Overall, it seems that using PLS with 41 components imporved the performance metrics vs our baseline PLS model with all components.

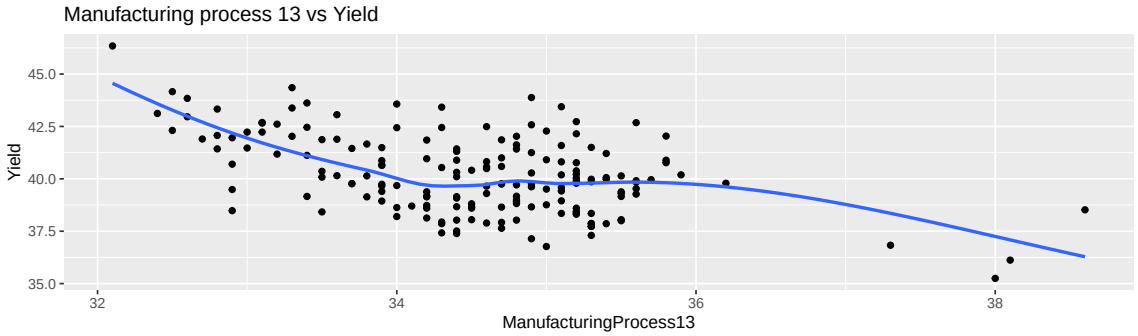
(e). Which predictors are most important in the model you have trained? Do either the biological or process predictors dominate the list?



VarImp allows us to identify the variables by name and compute their importance. ManufacturingProcess32 was flagged as the most important predictor overall and within the group of other Manufacturing Process variables. BiologicalMaterial02 was the most important variable within the BiologicalMaterial group but ranks 6th overall. The variable importance rankings are dominated by Manufacturing Process 9 to 6 within the top 15 predictors.

(f). Explore the relationships between each of the top predictors and the response. How could this information be helpful in improving yield in future runs of the manufacturing process?





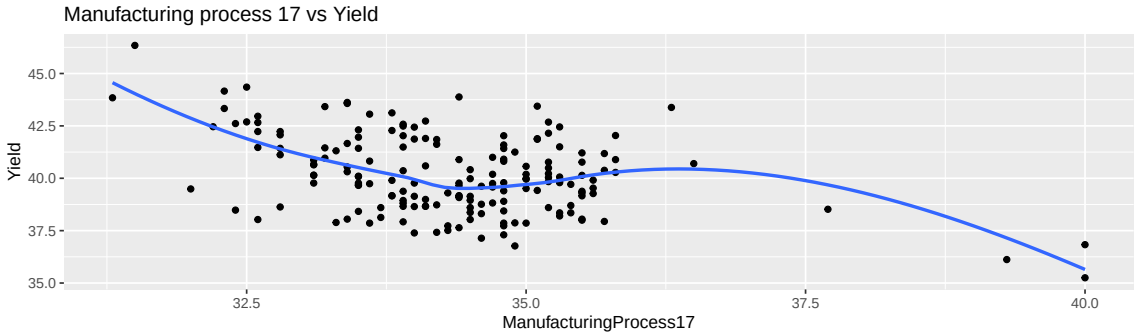


Table 4: Correlation of top 5 Important Variables vs Yield

	Yield
Yield	1.0000000
ManufacturingProcess32	0.6083321
ManufacturingProcess17	-0.4258069
ManufacturingProcess13	-0.5036797
ManufacturingProcess36	-0.4907962
ManufacturingProcess09	0.5034705

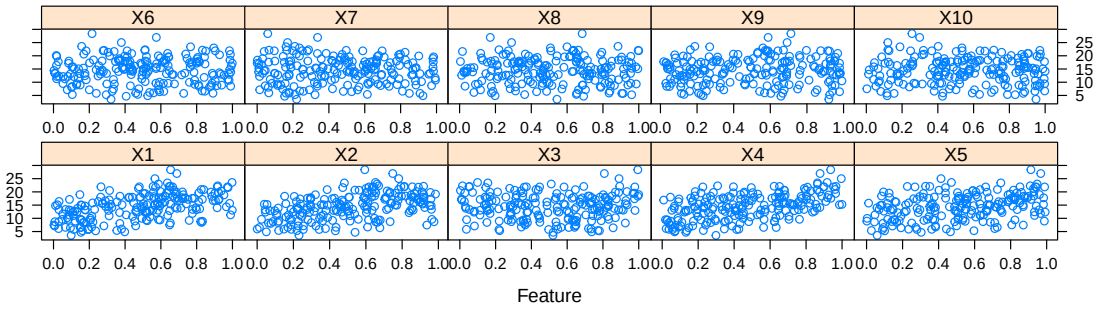
Out of the top three important variables, only ManufacturingProcess32 has a positive correlation with yield. We can drill down to the extent of the correlation with ManufacturingProcess32,ManufacturingProcess17, and Manufacturing-

Process13 vs Yield. Please note we also applied a loess smoother in our visualizations. ManufacturingProcess13 and ManufacturingProcess17 have a moderate negative correlation with yield, meaning that there exists an inverse relationship between these predictors and yield. From a business point of view, our aim is to increase yield since we know that yield ties into revenue. We do not have insight into what mechanics go into each manufacturing process but we can use this knowledge to adjust the processes to emulate manufacturing process 32.

Assignment 2

Kuhn and Johnson 7.2

Friedman (1991) introduced several benchmark data sets create by simulation. One of these simulations used the following nonlinear equation to create data: $y = 10\sin(\pi x_1 x_2) + 20(x_3 - 0.5)^2 + 10x_4 + 5x_5 + N(0,\sigma^2)$; where the x values are random variables uniformly distributed between $[0, 1]$ (there are also 5 other non-informative variables also created in the simulation).

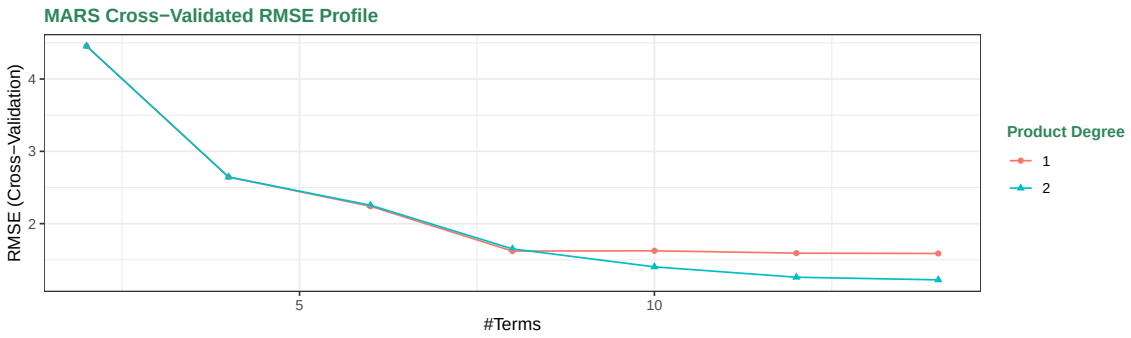


(a). Tune several models on these data. For example:

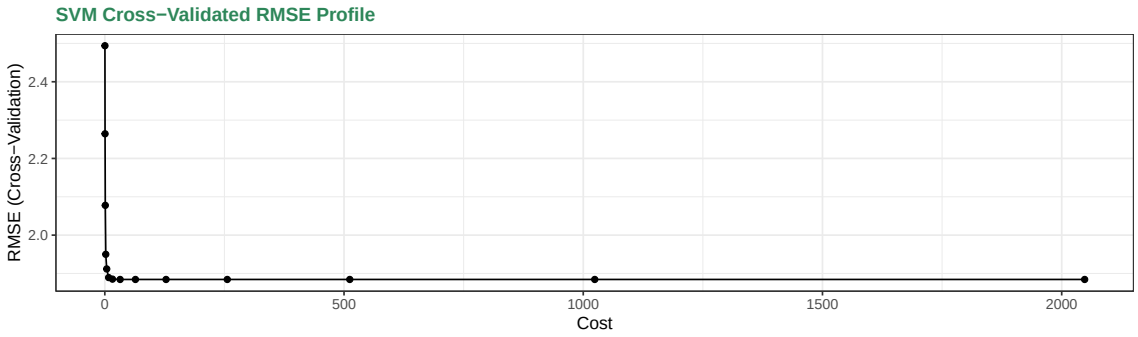
Model 1 MARS Regression: MARS, otherwise known as multivariate adapitive regression splines is a non parametric regression technique that automatically captures non-linearity and interaction between predictors. The basic MARS model has the following form:

$$\hat{f} = \sum_{i=1}^{k} c_i B_i(x)$$

The model computes the sum of basis functions B multiplied by constant coefficients c.The basis function can either be a constant, a hinge function, or a product of hinge functions. By definition, a hinge function is a piecewise function that converges at a point known as a knot.



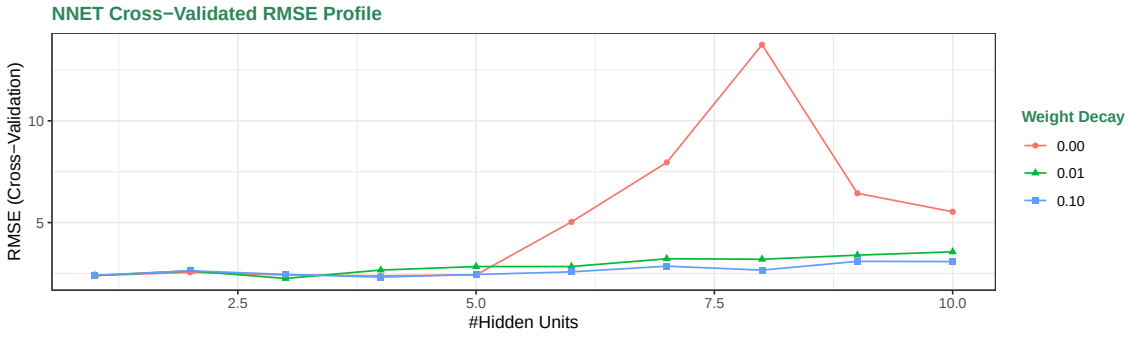
Model 2 SVM: SVM, also known as support vector machine is a method that can be applied to classification and regression tasks. On a high level, SVM creates a hyperplane in n dimensional space. This hyperplane acts like a classification boundary which can be linear or non linear. This boundary classifies information from a feature space.



Model 3 NNET:

NNEt otherwise known as a Neural Network, is a method inspired by a biological neuron system. It uses a system of nodes that are parallel to the way neurons work. It is ideal for capturing non-linear relationships that would otherwise be complicated in a multiple linear regression model. NNET evolves internally based on the calculated weights of each input. The basic structure is shown below:

$$Y = \sum (weight \quad * \quad input) \quad + \quad bias$$

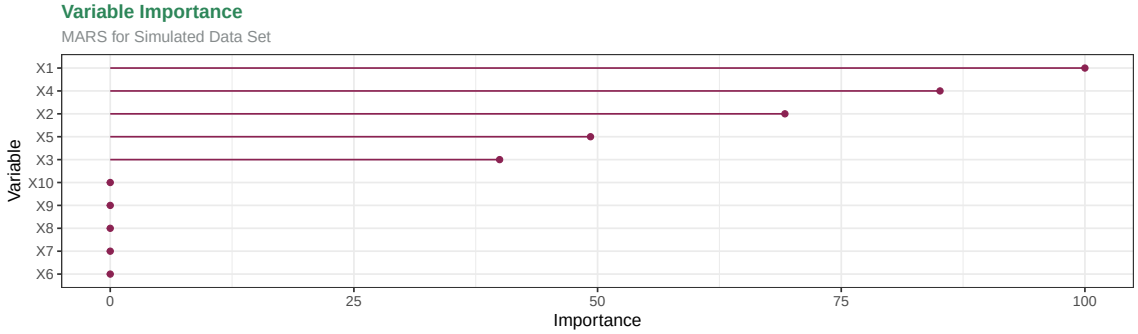


(b). Which models appear to give the best performance? Does MARS select the informative predictors (those named X1-X5)?

Table 5: Model Performance

	RMSE	RSquared	MAE
knnTrain	3.6521	3.6521	3.6521
knnTest	3.2041	0.6820	2.5683
MARSTrain	4.4548	0.9389	3.7461
MARSTest	1.1723	0.9449	0.9325
SVMTrain	2.4940	0.8561	1.9931
SVMTest	2.0699	0.8263	1.5723
NNETTrain	**13.7411**	**0.8021**	**5.5596**
NNETTest	**2.5669**	**0.7427**	**1.9506**

MARS appears to give the best performance based on RMSE, R squared and MAE on test set. The above table shows how our other selected models stack up against the best performing MARS model. We now evaluate the variable importance for our best performing model.



The variable importance table for MARS indicates that variables X1 through X5 were picked as the most important. Out of our collection of important variables used in MARS, X1 is the most important.

It is very likely that the lack of contribution allotted to the X6-X10 variables which bolster the R Squared and RMSE performance and noise from these variables did not reduce the predictive strength of this model as it does in small quantities in the other three models.

Kuhn and Johnson 7.5

Exercise 6.3 describes data for a chemical manufacturing process. Use the same data imputation, data splitting, and pre-processing steps as before and train several nonlinear regression models.

We pulled in the data processing method from 6.3. This includes imputation and removal of zero processing. Please refer to 6.3 for a more detailed look at the EDA involved with this data set. We tunned a KNN model, NNET model, MARS, and SVM model using specifications from the literature.

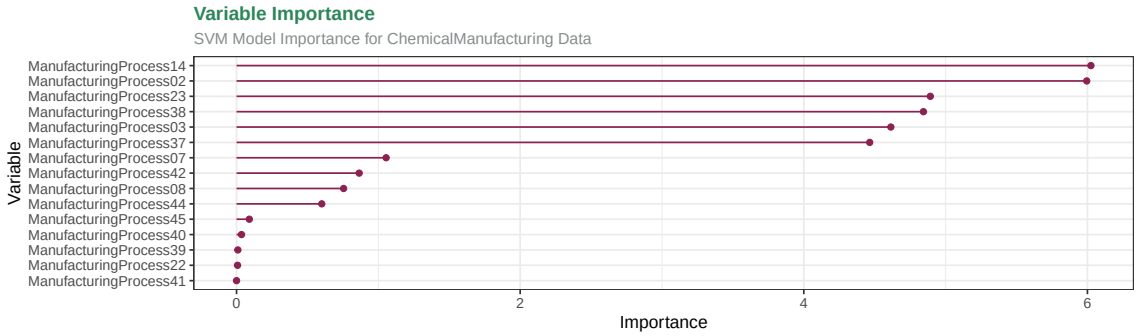
(a). Which nonlinear regression model gives the optimal resampling and test set performance?

Table 6: Model Performance on ChemicalManufacturing Data

	RMSE	RSquared	MAE
knnTrain	1.4835	0.4245	1.1746
knnTest	1.4078	0.4477	1.1041
MARSTrain	1.6367	0.6488	1.1614
MARSTest	1.3726	0.4682	1.1300
SVMTrain	**1.4142**	**0.7144**	**1.1448**
SVMTest	**1.2061**	**0.5715**	**0.9833**
NNETTrain	9.9082	0.3704	6.2548
NNETTest	1.4683	0.3987	1.1936

Radial SVM outperformed the other models across all key KPI's. The next best model was MARS. In part b, we will address what variables are dominant in our SVM model.

(b). Which predictors are most important in the optimal nonlinear regression model? Do either the biological or process variables dominate the list? How do the top ten important predictors compare to the top ten predictors from the optimal linear model?



`ManufacturingProcess` Variables dominate the ranking of important variables with ManufacturingProcess14 at the top. ManufacturingProcess32 was at the top of the important variables list when it came to our linear model with some Biological Process within the top 10.

(c). Explore the relationships between the top predictors and the response for the predictors that are unique to the optimal nonlinear regression model. Do these plots reveal intuition about the biological or process predictors and their relationship with yield?

We examined the top 5 predictors that were flagged as being the most important before the importance measure dropped.There are some pretty clear differences in the data which might explain both the overall poor performance of

Table 7: Correlation

VALUE	Yield
Yield	1.0000000
ManufacturingProcess14	-0.0099574
ManufacturingProcess38	-0.0864593
ManufacturingProcess03	-0.1189520
ManufacturingProcess37	-0.1593141
ManufacturingProcess02	-0.1953977

the linear models as well as the improved significance of Process-Based variables in the non-linear models.

Of the `ManfuacturingProcess` variables, they appear to be either tight clusters or discrete values which predict an array of possible Yields, which is directly opposed the the definition of linearly separable data base on earlier examination of correlation plots.

Assignment 3

Kuhn and Johnson 8.1

Recreate the simulated data from Exercise 7.2:

(a). **Fit a random forest model to all of the predictors, then estimate the variable importance scores. Did the random forest model significantly use the uninformative predictors (V6-V10)?**

(b). **Now add an additional predictor that is highly correlated with one of the informative predictors. Fit another random forest model to these data. Did the importance score for V1 change? What happens when you add another predictor that is also highly correlated with V1? For example:**

(c). **Use the 'cforest' function in the party package to fit a random forest model using conditional inference trees. The party package function 'varimp' can calculate predictor importance. The 'conditional' argument of that function toggles between the traditional importance measure and the modified version described in Strobl et al. (2007). Do these importances show the same pattern as the traditional random forest model?**

(d). **Repeat this process with different tree models, such as boosted trees and Cubist. Does the same pattern occur?**

Kuhn and Johnson 8.2

Use a simulation to show tree bias with different granularities.

Kuhn and Johnson 8.3

In stochastic gradient boosting the bagging fraction and learning rate will govern the construction of the trees as they are guided by the gradient. Although the optimal values of these parameters should be obtained through the tuning process, it is helpful to understand how the magnitudes of these parameters affect magnitudes of variable importance. Figure 8.24 provides the variable importance plots for boosting using two extreme values for the bagging fraction (0.1 and 0.9) and the learning rate (0.1 and 0.9) for the solubility data. The left-hand plot has both parameters set to 0.1, and the right-hand plot has both set to 0.9:

(a). **Why does the model on the right focus its importance on just the first few of predictors, whereas the model on the left spreads importance across more predictors?**

(b). **Which model do you think would be more predictive of other samples?**

(c). **How would increasing interaction depth affect the slope of predictor importance for either model in Fig.8.24?**

Kuhn and Johnson 8.7

Refer to Exercises 6.3 and 7.5 which describe a chemical manufacturing process. Use the same data imputation, data splitting, and pre-processing steps as before and train several tree-based models:

(a). **Which tree-based regression model gives the optimal resampling and test set performance?**

(b). **Which predictors are most important in the optimal tree-based regression model? Do either the biological or process variables dominate the list? How do the top 10 important predictors compare to the top 10 predictors from the optimal linear and nonlinear models?**

(c). **Plot the optimal single tree with the distribution of yield in the terminal nodes. Does this view of the data provide additional knowledge about the biological or process predictors and their relationship with yield?**

Assignment 4

TBD

R Script