DATA624 Group 3

**TEAM:**

## Kuiete Tchoupou Alain

## Littlejohn Jeffrey

## Malhotra Samriti

## Mishra Rajwant

## Ng Jimmy

Homework Assignment 2

Table of Contents

[Kuhn & Johnson 6 2](#__RefHeading___Toc10800_1911604043)

[HW\_KJ\_6.3 2](#__RefHeading___Toc6530_19116040431)

[Kuhn & Johnson 7 3](#__RefHeading___Toc6528_1911604043)

[HW\_KJ\_7.2 3](#__RefHeading___Toc6530_1911604043)

[Tune several models on these data. For example: 4](#__RefHeading___Toc6275_1911604043)

[MARS 6](#__RefHeading___Toc6277_1911604043)

[SVM 7](#__RefHeading___Toc6279_1911604043)

[NNET 8](#__RefHeading___Toc6281_1911604043)

[Which models appear to give the best performance? Does MARS select the informative predictors (those named X1-X5)? 9](#__RefHeading___Toc6283_1911604043)

[HW\_KJ\_7.5 13](#__RefHeading___Toc6532_1911604043)

[(a). Which nonlinear regression model gives the optimal resampling and test set performance? 14](#__RefHeading___Toc6287_1911604043)

[(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? 15](#__RefHeading___Toc6289_1911604043)

[(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? 16](#__RefHeading___Toc6291_1911604043)

[Kuhn & Johnson 8 20](#__RefHeading___Toc6534_1911604043)

[HW\_KJ\_8.1 20](#__RefHeading___Toc6536_1911604043)

[8.1. Recreate the simulated data from Exercise 7.2: 21](#__RefHeading___Toc6295_1911604043)

[(a) Fit a random forest model to all of the predictors, then estimate the variable importance scores: 21](#__RefHeading___Toc6297_1911604043)

[Did the random forest model significantly use the uninformative predictors (V6 – V10)? 21](#__RefHeading___Toc6299_1911604043)

[(b) Now add an additional predictor that is highly correlated with one of the informative predictors. For example: 22](#__RefHeading___Toc6301_1911604043)

[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? 22](#__RefHeading___Toc6303_1911604043)

[(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? 23](#__RefHeading___Toc6305_1911604043)

[(d) Repeat this process with different tree models, such as boosted trees and Cubist. Does the same pattern occur? 25](#__RefHeading___Toc6307_1911604043)

[HW\_KJ\_8.2 29](#__RefHeading___Toc6538_1911604043)

[8.2. Use a simulation to show tree bias with different granularities. 29](#__RefHeading___Toc6309_1911604043)

[HW\_KJ\_8.3 31](#__RefHeading___Toc6540_1911604043)

[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: 31](#__RefHeading___Toc10802_1911604043)

[(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? 31](#__RefHeading___Toc10804_1911604043)

[(b) Which model do you think would be more predictive of other samples? 32](#__RefHeading___Toc6542_1911604043)

[(c) How would increasing interaction depth affect the slope of predictor importance for either model in Fig. 8.24 32](#__RefHeading___Toc6317_1911604043)

[HW\_KJ\_8.7 32](#__RefHeading___Toc10806_1911604043)

[Kuhn & Johnson 10 33](#__RefHeading___Toc10808_1911604043)

[HW\_KJ\_10 33](#__RefHeading___Toc6530_19116040432)

# Kuhn & Johnson 6

## HW\_KJ\_6.3

|  |
| --- |
|  |
|  |
|  |

# Kuhn & Johnson 7

## HW\_KJ\_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*=10sin(*π**x*1*x*2)+20(*x*3−0.5)2+10*x*4+5*x*5+*N*(0,*σ*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).  The package mlbench contains a function called mlbench.friedman1 that simulates these data: |
| **library**(mlbench)  set.seed(200)  trainingData <- mlbench.friedman1(200, sd = 1)  *## We convert the 'x' data from a matrix to a data frame*  *## One reason is that this will give the columns names.*  trainingData$x <- data.frame(trainingData$x)  *## Look at the data using*  featurePlot(trainingData$x, trainingData$y) |
|  |
| Tune several models on these data. For example: **library**(caret)  knnModel <- train(x = trainingData$x,  y = trainingData$y,  method = "knn",  preProc = c("center", "scale"),  tuneLength = 10)  knnModel  *## k-Nearest Neighbors*  *##*  *## 200 samples*  *## 10 predictor*  *##*  *## Pre-processing: centered (10), scaled (10)*  *## Resampling: Bootstrapped (25 reps)*  *## Summary of sample sizes: 200, 200, 200, 200, 200, 200, ...*  *## Resampling results across tuning parameters:*  *##*  *## k RMSE Rsquared MAE*  *## 5 3.466085 0.5121775 2.816838*  *## 7 3.349428 0.5452823 2.727410*  *## 9 3.264276 0.5785990 2.660026*  *## 11 3.214216 0.6024244 2.603767*  *## 13 3.196510 0.6176570 2.591935*  *## 15 3.184173 0.6305506 2.577482*  *## 17 3.183130 0.6425367 2.567787*  *## 19 3.198752 0.6483184 2.592683*  *## 21 3.188993 0.6611428 2.588787*  *## 23 3.200458 0.6638353 2.604529*  *##*  *## RMSE was used to select the optimal model using the smallest value.*  *## The final value used for the model was k = 17.*  knnPred <- predict(knnModel, newdata = testData$x)  *## The function 'postResample' can be used to get the test set*  *## performance values*  postResample(pred = knnPred, obs = testData$y)  *## RMSE Rsquared MAE*  *## 3.2040595 0.6819919 2.5683461* |
| MARS arsGrid <- expand.grid(degree = 1:2, nprune = seq(2, 14,by = 2))  marsModel <- train(x = trainingData$x,  y = trainingData$y,  method = 'earth',  tuneGrid = marsGrid,  trControl = trainControl(method = "cv"))  ggplot(marsModel) + labs(title = "MARS Cross-Validated RMSE Profile") + theme\_classic() |
|  |
| SVM svmModel <- train(x = trainingData$x,  y = trainingData$y,  method = 'svmRadial',  tuneLength = 14,  trControl = trainControl(method = "cv"))  ggplot(svmModel) + labs(title = "SVM Cross-Validated RMSE Profile") + theme\_gray() |
|  |
| NNET nnetGrid <- expand.grid(.size = c(1:10), .decay = c(0, 0.01, .1))  nnetModel <- train(trainingData$x,  trainingData$y,  method = "nnet",  tuneGrid = nnetGrid,  trControl = trainControl(method="cv"),  preProc = c("center", "scale"),  linout = TRUE,  trace = FALSE,  MaxNWts = 10 \* (ncol(trainingData$x) + 1) + 10 + 1,  maxit = 500)  ggplot(nnetModel) + labs(title = "NNET Cross-Validated RMSE Profile") + theme\_linedraw() |
|  |
| Which models appear to give the best performance? Does MARS select the informative predictors (those named X1-X5)? |
| As displayed in the summary table, MARS appears to be supreme comparing to KNN, SVM and NNET in terms of RMSE, R squared and MAE on the test set. The variable importance plot for MARS indicates that the variables X1 through 5 are the most important. X1 appears to be the most important variable. MARS appears to have successfully selected the most informative predictors and outperformed the other models in this exercise.  *# Prediction*  *# knnPred is already given to us from the textbook*  marsPred <- predict(marsModel, newdata = testData$x)  svmPred <- predict(svmModel, newdata = testData$x)  nnetPred <- predict(nnetModel, newdata = testData$x)  *# Performance*  knnPerformance <- postResample(pred = knnPred, obs = testData$y)  marsPerformance <- postResample(pred = marsPred, obs = testData$y)  svmPerformance <- postResample(pred = svmPred, obs = testData$y)  nnetPerformance <- postResample(pred = nnetPred, obs = testData$y)  *# Summary*  rbind(  *# KNN*  "KNN Train" = c("RMSE" = max(knnModel$results$RMSE),  "RSquared" = max(knnModel$results$RMSE),  "MAE" = max(knnModel$results$RMSE)),  "KNN Test" = knnPerformance,  *# MARS*  "MARS Train" = c("RMSE" = max(marsModel$results$RMSE),  "RSquared" = max(marsModel$results$Rsquared),  "MAE" = max(marsModel$results$MAE)),  "MARS Test" = marsPerformance,  *# SVM*  "SVM Train" = c(max(svmModel$results$RMSE),  max(svmModel$results$Rsquared),  max(svmModel$results$MAE)),  "SVM Test" = svmPerformance,  *# NNET*  "NNET Train" = c(max(nnetModel$results$RMSE),  max(nnetModel$results$Rsquared),  max(nnetModel$results$MAE)),  "NNET Test" = nnetPerformance) %>%  kable(caption = "Model Performance", digits = 3) %>%  kable\_styling() |
|  |





|  |
| --- |
| *# Variable Importance*  marsImportance <- caret::varImp(marsModel)  marsImportance$importance %>%  as.data.frame() %>%  tibble::rownames\_to\_column() %>%  dplyr::mutate(name = forcats::fct\_inorder(rowname)) %>%  arrange(desc(Overall)) %>%  head(10) %>%  ggplot(., aes(x = reorder(name, Overall), y = Overall)) +  geom\_point() +  geom\_segment(aes(x = name, xend = name, y = 0, yend = Overall)) +  ggtitle("Variable Importance: MARS for Simulated Data Set") +  labs(x = "Variable", y = "Importance") +  coord\_flip() +  theme\_minimal() |
|  |
|  |

## HW\_KJ\_7.5

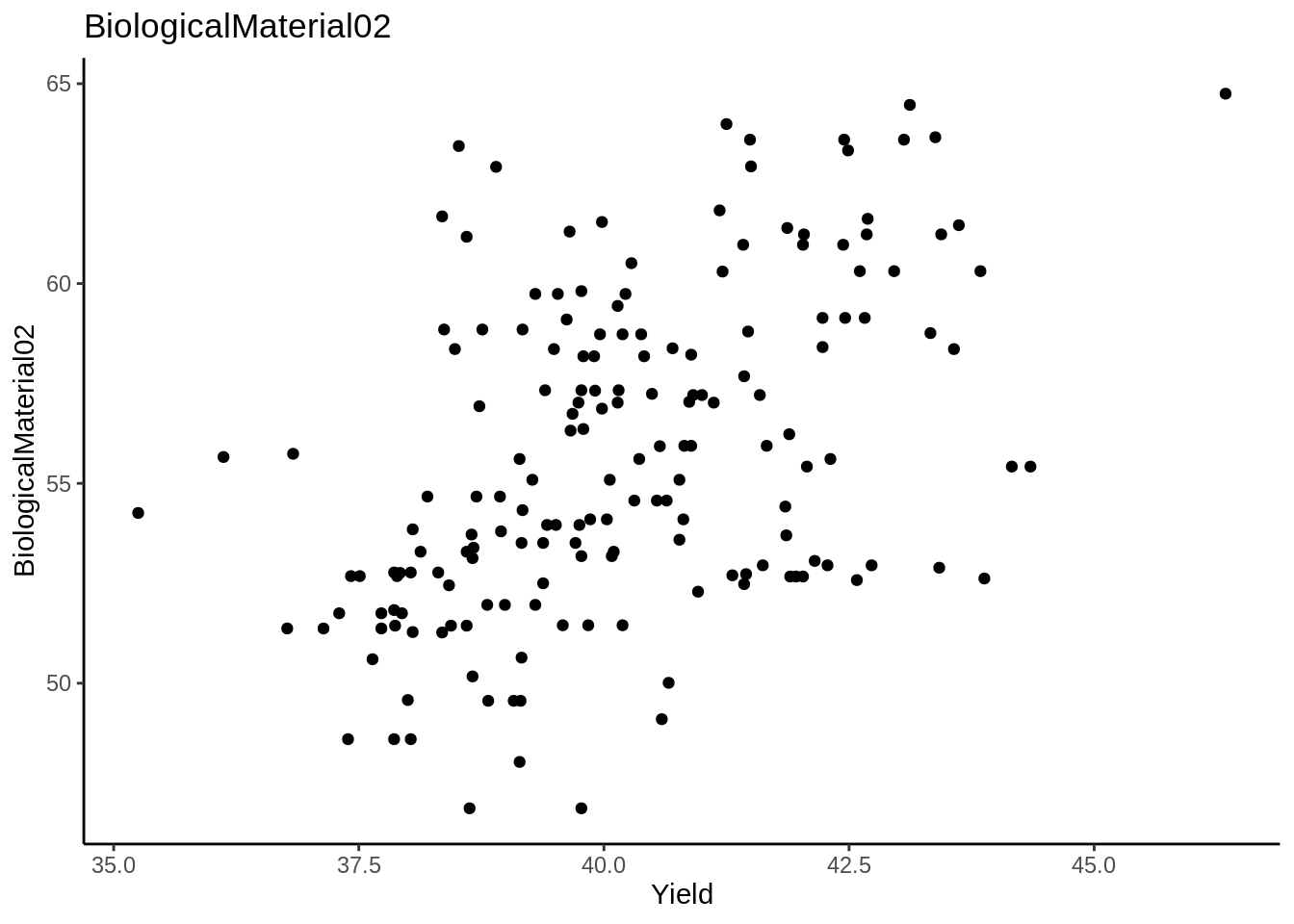
|  |
| --- |
| Exercise 6.3 describes data for a chemical manufacturing process. Use the same data imputation, data splitting, and preprocessing steps as before and train several nonlinear regression models. (a). Which nonlinear regression model gives the optimal resampling and test set performance? *# data prep - using the same methodology from exercise 6.3*  data(ChemicalManufacturingProcess)  impu\_data <- kNN(ChemicalManufacturingProcess, imp\_var = FALSE)  n <- nrow(impu\_data)  i.training <- sort(sample(n,round(n\*0.8)))  L.training <- impu\_data[i.training,]  L.test <- impu\_data[-i.training,]  *# check if any missing left*  **if**(!any(colSums(is.na(impu\_data)) %>% unlist >0)){print("there is no more missing after successful imputation")}  *## [1] "there is no more missing after successful imputation"*  *# dim*  sapply(list(impu\_data, L.training, L.test), dim) %>%  as.data.frame %>%  cbind(dim = c("# of row", "# of column")) %>%  dplyr::select(dim, imput\_data = V1, L.training = V2, L.test = V3)  *## dim imput\_data L.training L.test*  *## 1 # of row 176 141 35*  *## 2 # of column 58 58 58*  *# KNN*  knnModel2 <- train(Yield~.,  data = L.training,  method = "knn",  preProc = c("center", "scale"),  tuneLength = 10)  *# MARS*  marsGrid2 <- expand.grid(degree = 1:2, nprune = 1:20)  marsModel2 <- train(Yield~.,  data = L.training,  method = "earth",  tuneGrid = marsGrid2,  trControl = trainControl(method = "cv"))  *# SVM*  svmModel2 <- train(Yield~.,  data = L.training,  method = "svmRadial",  preProc = c("center", "scale"),  tuneLength = 14,  trControl = trainControl(method = "cv"))  *# NNET*  nnetGrid2 <- expand.grid(decay = c(0, 0.01, .1), size = c(1:10))  nnetModel2 <- train(Yield~.,  data = L.training,  method = "nnet",  tuneGrid = nnetGrid2,  trControl = trainControl(method = "cv"),  preProc = c("center", "scale"),  linout = TRUE,  trace = FALSE,  maxit = 500)  *# Prediction*  knnPred2 <- predict(knnModel2, newdata = L.test)  marsPred2 <- predict(marsModel2, newdata = L.test)  svmPred2 <- predict(svmModel2, newdata = L.test)  nnetPred2 <- predict(nnetModel2, newdata = L.test)  *# Performance*  knnPerformance2 <- postResample(pred = knnPred2, obs = L.test$Yield)  marsPerformance2 <- postResample(pred = marsPred2, obs = L.test$Yield)  svmPerformance2 <- postResample(pred = svmPred2, obs = L.test$Yield)  nnetPerformance2 <- postResample(pred = nnetPred2, obs = L.test$Yield)  *# Summary*  rbind(  *# KNN*  "KNN Train" = c("RMSE" = max(knnModel2$results$RMSE),  "RSquared" = max(knnModel2$results$Rsquared),  "MAE" = max(knnModel2$results$MAE)),  "KNN Test" = knnPerformance2,  *# MARS*  "MARS Train" = c("RMSE" = max(marsModel2$results$RMSE),  "RSquared" = max(marsModel2$results$Rsquared),  "MAE" = max(marsModel2$results$MAE)),  "MARS Test" = marsPerformance2,  *# SVM*  "SVM Train" = c(max(svmModel2$results$RMSE),  max(svmModel2$results$Rsquared),  max(svmModel2$results$MAE)),  "SVM Test" = svmPerformance2,  *# NNET*  "NNET Train" = c(max(nnetModel2$results$RMSE),  max(nnetModel2$results$Rsquared),  max(nnetModel2$results$MAE)),  "NNET Test" = nnetPerformance2) %>%  kable(caption = "Model Performance on ChemicalManufacturing Data", digits = 3) %>%  kable\_styling() |



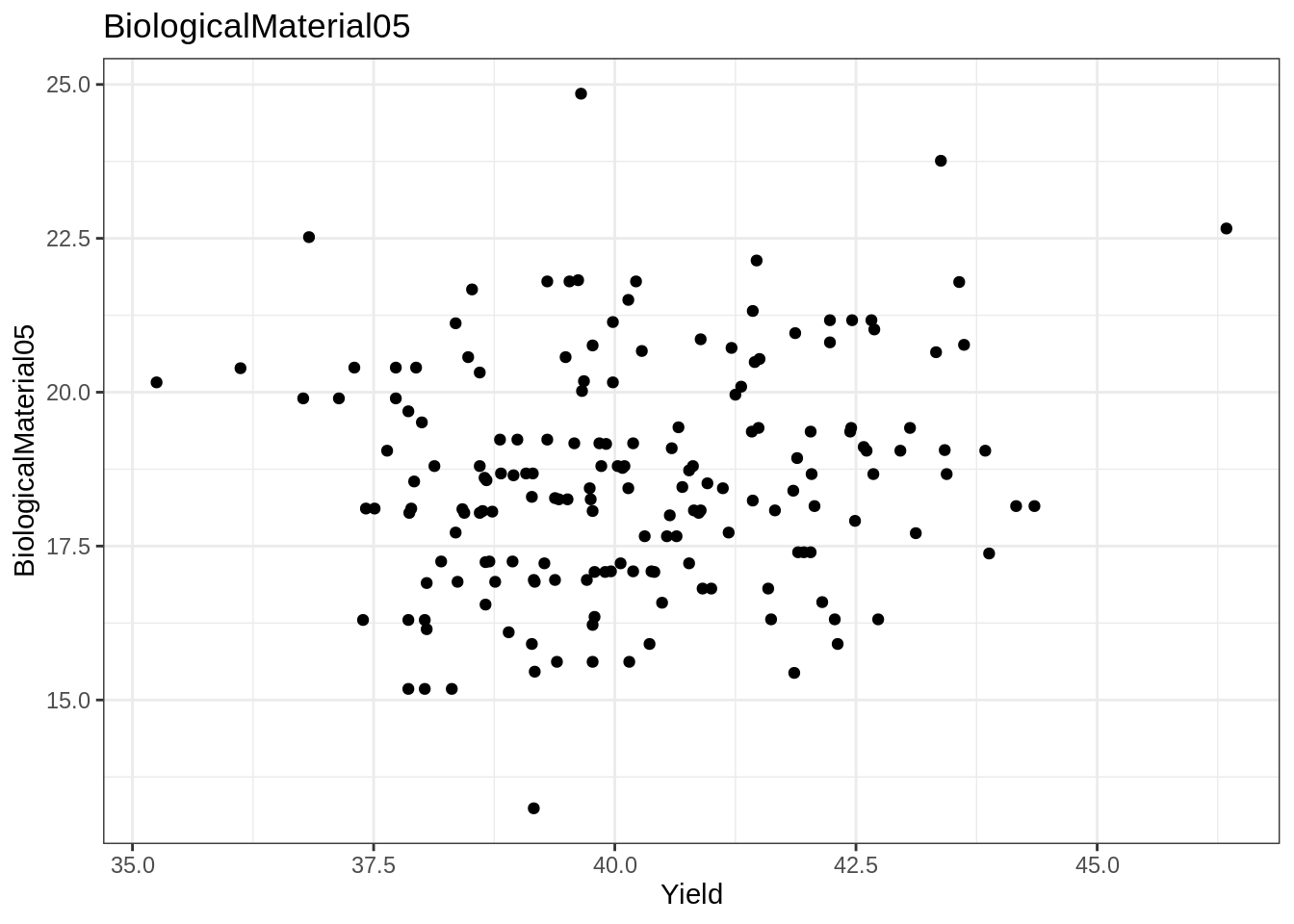
|  |
| --- |
| (b). Which predictors are most important in the optimal nonlinear regression model? Do either the biological or process variables dominate the list? How do the top ten important predictors compare to the top ten predictors from the optimal linear model? |
| svmImportance2 <- caret::varImp(svmModel2)  svmImportance2$importance %>%  as.data.frame() %>%  tibble::rownames\_to\_column() %>%  dplyr::mutate(name = forcats::fct\_inorder(rowname)) %>%  arrange(desc(Overall)) %>%  head(10) %>%  ggplot(., aes(x = reorder(name, Overall), y = Overall)) +  geom\_point() +  geom\_segment(aes(x = name, xend = name, y = 0, yend = Overall)) +  ggtitle("Variable Importance: SVM for ChemicalManufacturing Data") +  labs(x = "Variable", y = "Importance") +  coord\_flip() +  theme\_minimal() |
|  |
| (c). Explore the relationships between the top predictors and the response for the predictors that are unique to the optimal nonlinear regression model. Do these plots reveal intuition about the biological or process predictors and their relationship with yield? In general, the SVM model pointed to the same direction, conclusion as the linear model (Elastic net) as both indicated that the ManufacturingProcess variables seemed to be more important, e.g. ManufacturingProcess32, 09 tops the list. However, the SVM found that several biological variables important that were not found in the linear model and vice versa, e.g. BiologicalMaterial02 was found in SVM only, whereas BiologicalMaterial05 was found in the linear model only. It seemed that both variables positively correlated with Yield, but the relationship between BiologicalMaterial02 and Yield seemed to be much more obvious and stronger than 05.  features <- svmImportance2$importance %>%  as.data.frame() %>%  tibble::rownames\_to\_column() %>%  dplyr::mutate(name = forcats::fct\_inorder(rowname)) %>%  arrange(desc(Overall)) %>%  head(10) %>%  .$rowname %>%  as.vector  subset <- impu\_data %>%  dplyr::select(all\_of(c("Yield", features)))  corSubset <- cor(subset) %>%  as.matrix %>%  as.data.frame %>%  tibble::rownames\_to\_column(., "variable") %>%  arrange(desc(Yield)) %>%  dplyr::select(variable, Yield)  corSubset %>%  kable(caption = "Correlation") %>%  kable\_styling() |



|  |
| --- |
| *# BiologicalMaterial02*  impu\_data %>%  dplyr::select(Yield, BiologicalMaterial02) %>%  ggplot(., aes(Yield, BiologicalMaterial02)) +  geom\_point() +  ggtitle("BiologicalMaterial02") +  theme\_classic() |



|  |
| --- |
| *# BiologicalMaterial05*  impu\_data %>%  dplyr::select(Yield, BiologicalMaterial05) %>%  ggplot(., aes(Yield, BiologicalMaterial05)) +  geom\_point() +  ggtitle("BiologicalMaterial05") +  theme\_bw() |



# Kuhn & Johnson 8

## HW\_KJ\_8.1

|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **library**(AppliedPredictiveModeling) 8.1. Recreate the simulated data from Exercise 7.2: **library**(mlbench)  set.seed(200)  simulated <- mlbench.friedman1(200, sd = 1)  simulated <- cbind(simulated$x, simulated$y)  simulated <- as.data.frame(simulated)  colnames(simulated)[ncol(simulated)] <- "y" (a) Fit a random forest model to all the predictors, then estimate the variable importance scores: **library**(randomForest)  ## randomForest 4.6-14  ## Type rfNews() to see new features/changes/bug fixes.  **library**(caret)  rfmodel1 <- randomForest(y ~ ., data = simulated,  importance = TRUE,  ntree = 1000)  rfImp1 <- varImp(rfmodel1, scale = FALSE)  rfImp1  ## Overall  ## V1 8.732235404  ## V2 6.415369387  ## V3 0.763591825  ## V4 7.615118809  ## V5 2.023524577  ## V6 0.165111172  ## V7 -0.005961659  ## V8 -0.166362581  ## V9 -0.095292651  ## V10 -0.074944788 Did the random forest model significantly use the uninformative predictors (V6 – V10)? The importance values of predictors (V6 - V10) are either very low and positive or very low and negative. This means that the random forest model really did not use those variables.  varImpPlot(rfmodel1) | | | | | | | |  |
|  | | | | | | | |  |
| (b) Now add an additional predictor that is highly correlated with one of the informative predictors. For example: simulated2 <- simulated  simulated2$duplicate1 <- simulated2$V1 + rnorm(200) \* .1  cor(simulated2$duplicate1, simulated2$V1)  ## [1] 0.9460206 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? rfmodel2 <- randomForest(y ~ ., data = simulated2,  importance = TRUE,  ntree = 1000)  rfImp2 <- varImp(rfmodel2, scale = FALSE)  rfImp2  ## Overall  ## V1 5.69119973  ## V2 6.06896061  ## V3 0.62970218  ## V4 7.04752238  ## V5 1.87238438  ## V6 0.13569065  ## V7 -0.01345645  ## V8 -0.04370565  ## V9 0.00840438  ## V10 0.02894814  ## duplicate1 4.28331581  The variable V1 importance value has been affected, its importance value has decreased. The model chose randomly which variable to use among the two highly correlated variable. There are two variables with the same type of information. This introduces instability in the model coefficients. | | | | | | | |  |
|  | | | | | | | |  |
| (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? **library**(party)  bagCtrl <- cforest\_control(mtry = ncol(simulated) - 1)  baggedTree <- cforest(y ~ ., data = simulated)  **library**(caret)  crfImp <- varimp(baggedTree, conditional = TRUE)  crfImp  ## V1 V2 V3 V4 V5 V6  ## 5.471457361 5.166826657 0.020994281 6.689072245 1.256076719 0.004925215  ## V7 V8 V9 V10  ## -0.008184439 -0.009141850 -0.012594617 -0.013200299  crf.importance <- data.frame(var = names(crfImp), y =crfImp)  **library**(dplyr)  crf.imp <- crf.importance %>% arrange(desc(y))  crf.imp  ## var y  ## V4 6.689072245  ## V1 V1 5.471457361  ## V2 V2 5.166826657  ## V5 V5 1.256076719  ## V3 V3 0.020994281  ## V6 V6 0.004925215  ## V7 V7 -0.008184439  ## V8 V8 -0.009141850  ## V9 V9 -0.012594617  ## V10 V10 -0.013200299  crf.imp %>% dplyr::mutate(var = reorder(var, y)) %>%  ggplot(aes(x = var, y= y)) +  geom\_point(show.legend = FALSE) +  labs(x = "Important Variables", y = NULL) +  coord\_flip() | | | | | | | |  |
|  | | | | | | | |  |
| (d) Repeat this process with different tree models, such as boosted trees and Cubist. Does the same pattern occur? **Boosted trees models**  **library**(gbm)  ## Loaded gbm 2.1.5  gbmModel <- gbm(y ~ ., data = simulated, distribution = "gaussian")  gbmImp <- varImp(gbmModel, numTrees = 100, scale = FALSE)  gbmImp  ## Overall  ## V1 4071.65288  ## V2 3550.71836  ## V3 1156.48989  ## V4 4416.45306  ## V5 1791.01067  ## V6 81.00046  ## V7 0.00000  ## V8 0.00000  ## V9 0.00000  ## V10 0.00000  The most important predictors are V1, V2, V3, V4, V5 The V6 predictor is less important than the first four V7, V8, V9, and V10 have not been used by the model  summary(gbmModel)  ## var rel.inf  ## V4 V4 29.3114601  ## V1 V1 27.0230635  ## V2 V2 23.5656846  ## V5 V5 11.8867193  ## V3 V3 7.6754823  ## V6 V6 0.5375902  ## V7 V7 0.0000000  ## V8 V8 0.0000000  ## V9 V9 0.0000000  ## V10 V10 0.0000000  The same pattern is repeated. The important variables are V1-V5  **Random Forest Model**  rfmodel2 <- randomForest(y ~ ., data = simulated2,  importance = TRUE,  ntree = 1000)  rfImp2 <- varImp(rfmodel2, scale = FALSE)  rfImp2  ## Overall  ## V1 5.49787495  ## V2 6.23713508  ## V3 0.63801571  ## V4 7.00094583  ## V5 2.03385902  ## V6 0.15776399  ## V7 0.01694912  ## V8 -0.05427651  ## V9 -0.02525938  ## V10 -0.07556684  ## duplicate1 4.66387457  Random Forest presents the same pattern as previous where variables V6-V7 are not important.  **Cubist Model**  simulatedX <- dplyr::select(simulated, V1:V10)  **library**(Cubist)  cubistModel <- cubist(simulatedX, simulated$y, committees = 100, neighbors = 0.01)  cubistImp <- varImp(cubistModel, scale = FALSE)  cubistImp  ## Overall  ## V1 71.5  ## V3 47.0  ## V2 58.5  ## V4 48.0  ## V5 33.0  ## V6 13.0  ## V7 0.0  ## V8 0.0  ## V9 0.0  ## V10 0.0  cubistTuned <- train(simulatedX, simulated$y, method = "cubist", )  varImp(cubistTuned)  ## cubist variable importance  ##  ## Overall  ## V1 100.00  ## V2 75.69  ## V4 68.06  ## V3 58.33  ## V5 55.56  ## V6 15.28  ## V9 0.00  ## V8 0.00  ## V10 0.00  ## V7 0.00  Cubist present the same pattern as the GBM model We observe practically the same pattern. the variables (V6 - V10)are not important, though not use by all those models. | | | | | | | |  |

## HW\_KJ\_8.2

|  |
| --- |
| 8.2. Use a simulation to show tree bias with different granularities. **Generate random variables Predictors with different granularities Generate also the response variable colinear to some of them**  X1 <- sample(c(3, 5), 100, replace = TRUE)  X2 <- sample(c(3, 5, 7, 11, 13, 17, 19, 23, 29, 31, 37, 41), 100, replace = TRUE)  X3 <- sample(10:1000/10, 100)  y <- 7\*X1 + 3\*X2 + rnorm(100)  df <- data.frame(X1, X2, X3, y)  head(df)  ## X1 X2 X3 y  ## 1 5 3 80.1 44.95411  ## 2 3 19 54.6 76.88111  ## 3 3 11 45.0 51.92110  ## 4 5 7 54.5 55.32196  ## 5 3 5 51.2 34.74864  ## 6 3 19 66.0 76.96352  **Fit a model and Compare the predictors importance**  rfBias <- randomForest(y ~ ., data = df,  importance = TRUE,  ntree = 100)  varImportance <- varImp(rfBias, scale = FALSE)  varImportance  ## Overall  ## X1 125.24888  ## X2 1536.94859  ## X3 6.56273  The predictor X2 with more unique values is more used by the random forest model than the predictor X2 with less unique values. But the variable X3 with more distinct values that is non correlated and has no relation with the response variable was also used by the model. |

## HW\_KJ\_8.3

|  |
| --- |
| 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? The gradient boosting model employs the greedy strategy of choosing the optimal weak learner at each stage. Ridgeway (2007) suggests that small values of the learning parameter (< 0.01) work best. High values of learning rate will result in increasing the contribution of few variables at each stage of the process and then shrinking the number of important variables.  Learning rate controls the fraction of the predictions of each tree being added. A higher learning rate means that larger fraction of each tree’s predictions is added to the final prediction. This effectively means that a higher learning rate increases the dependent / correlation structure. More of the same predictors will be selected among the trees. Therefore, the right-hand plot has its importance focus on just the first few of the predictors and look very steep.  Bagging fraction is the fraction of data being used in each iteration of the trees. When you have a small bagging fraction, say 0.1, on each iteration just 10% of the full data is randomly sampled. So, each tree may be built using very different dataset. Since the dataset are very different, the trees will be splitting very differently from each other. On the contrast, when you have large bagging fraction, say 0.9, essentially on each iteration the trees are seeing the same dataset - they will likely split similarly. This means that larger bagging fraction increases the dependent / correlated structure in the boosting trees. Therefore, the right-hand plot with a larger bagging fraction has its importance focus on just the first few of the predictors. |
| (b) Which model do you think would be more predictive of other samples? Greedy models are less likely to select the optimal global model and are prone to over-fitting. Stochastic models reduce prediction variance. Therefore, the less greedy model on the left (with a 0.1 learning rate) that is also more random (due to only selecting 0.1 of the training set observations to propose the next tree in the expansion) would be more predictive of other samples.  Learning rate and bagging fraction are important parameters to control the overfitting of the gradient boosting model that requires tuning. A smaller learning rate and bagging fraction leads to better generalization ability over unseen samples. If I must guess, the model with 0.1 learning rate and bagging fraction will be more predictive of out of bag samples. However, since this involves a tradeoff between bias-variance, I can only confirm using cross validation or a test set. (c) How would increasing interaction depth affect the slope of predictor importance for either model in Fig. 8.24 The author comment stated that the larger value of shrinkage has an impact on reducing RMSE for all choices of tree depth and number of trees.  Figure 8.20 presents the cross-validated RMSE results for boosted trees across tuning parameters of tree depth (1–7), number of trees (100–1,000), and shrinkage (0.01 or 0.1); the bagging fraction in this illustration was fixed at 0.5. When examining this figure, the larger value of shrinkage (right-hand plot) has an impact on reducing RMSE for all choices of tree depth and number of trees. Also, RMSE decreases as tree depth increases when shrinkage is 0.01. The same pattern holds true for RMSE when shrinkage is 0.1 and the number of trees is less than 300. Using the one-standard-error rule, the optimal boosted tree has depth 3 with 400 trees and shrinkage of 0.1. |

## HW\_KJ\_8.7

|  |
| --- |
|  |
|  |
|  |
|  |

# Kuhn & Johnson 10

## HW\_KJ\_10

|  |
| --- |
|  |
|  |
|  |

## Market basket analysis / recommender systems

## Introduction

I am assigning one simple problem on market basket analysis / recommender systems.

Imagine 10000 receipts sitting on your table. Each receipt represents a transaction with items that were purchased. The receipt is a representation of stuff that went into a customer’s basket – and therefore ‘Market Basket Analysis’.

That is exactly what the Groceries Data Set contains: a collection of receipts with each line representing 1 receipt and the items purchased. Each line is called a transaction and each column in a row represents an item.

Here is the dataset = GroceryDataSet.csv (comma separated file)

#### Your assignment is to use R to mine the data for association rules. You should report support, confidence and lift and your top 10 rules by lift. Turn in as you would the other problems from HA and KJ. Due 07/18/20 with the packaged set, HW #2.

### MARKET BASKET ANALYSIS

## Data preprocessing and exploring

## citrus.fruit semi.finished.bread margarine  
## 1 tropical fruit yogurt coffee  
## 2 whole milk   
## 3 pip fruit yogurt cream cheese   
## 4 other vegetables whole milk condensed milk  
## 5 whole milk butter yogurt  
## 6 rolls/buns   
## ready.soups X X.1 X.2 X.3 X.4 X.5 X.6 X.7  
## 1   
## 2   
## 3 meat spreads   
## 4 long life bakery product   
## 5 rice abrasive cleaner   
## 6

## citrus.fruit semi.finished.bread margarine   
## sausage : 825 :2159 :3802   
## whole milk : 717 whole milk : 654 whole milk : 506   
## frankfurter : 580 other vegetables: 550 other vegetables: 415   
## tropical fruit : 482 root vegetables : 383 rolls/buns : 293   
## other vegetables: 460 rolls/buns : 378 yogurt : 289   
## citrus fruit : 452 tropical fruit : 355 soda : 229   
## (Other) :6318 (Other) :5355 (Other) :4300   
## ready.soups X X.1   
## :5101 :6105 :6960   
## whole milk : 315 rolls/buns : 176 soda : 150   
## other vegetables: 254 soda : 168 rolls/buns : 146   
## rolls/buns : 238 yogurt : 160 shopping bags: 107   
## soda : 211 whole milk : 149 bottled water: 95   
## yogurt : 202 shopping bags: 145 yogurt : 93   
## (Other) :3513 (Other) :2931 (Other) :2283   
## X.2 X.3 X.4   
## :7605 :8150 :8588   
## soda : 120 shopping bags: 76 soda : 61   
## shopping bags: 107 bottled water: 68 shopping bags : 56   
## rolls/buns : 92 newspapers : 66 fruit/vegetable juice: 55   
## newspapers : 68 rolls/buns : 59 bottled water : 54   
## domestic eggs: 57 soda : 59 newspapers : 51   
## (Other) :1785 (Other) :1356 (Other) : 969   
## X.5 X.6   
## :8938 :9184   
## shopping bags : 49 shopping bags: 40   
## soda : 39 newspapers : 36   
## fruit/vegetable juice: 34 pastry : 27   
## newspapers : 33 bottled water: 25   
## bottled water : 26 napkins : 23   
## (Other) : 715 (Other) : 499   
## X.7 X.8   
## :9366 :9483   
## soda : 30 soda : 24   
## shopping bags : 19 shopping bags : 18   
## chocolate : 17 fruit/vegetable juice: 16   
## fruit/vegetable juice: 17 napkins : 14   
## napkins : 17 newspapers : 14   
## (Other) : 368 (Other) : 265   
## X.9 X.10   
## :9561 :9638   
## shopping bags : 18 shopping bags : 16   
## fruit/vegetable juice: 17 napkins : 13   
## newspapers : 14 fruit/vegetable juice: 11   
## soda : 14 hygiene articles : 11   
## napkins : 11 candy : 9   
## (Other) : 199 (Other) : 136   
## X.11 X.12   
## :9693 :9739   
## shopping bags : 11 napkins : 8   
## napkins : 9 chocolate : 5   
## chocolate : 8 newspapers : 5   
## hygiene articles : 6 candy : 4   
## long life bakery product: 6 fruit/vegetable juice: 4   
## (Other) : 101 (Other) : 69   
## X.13 X.14 X.15   
## :9768 :9782 :9796   
## candy : 5 detergent : 4 bottled beer : 3   
## chocolate : 5 fruit/vegetable juice: 4 napkins : 3   
## napkins : 5 shopping bags : 4 pot plants : 3   
## newspapers : 3 chocolate : 3 candy : 2   
## bottled water: 2 bottled water : 2 hygiene articles: 2   
## (Other) : 46 (Other) : 35 (Other) : 25   
## X.16 X.17   
## :9805 :9816   
## napkins : 4 napkins : 2   
## fruit/vegetable juice : 2 baking powder : 1   
## house keeping products: 2 bottled beer : 1   
## hygiene articles : 2 cleaner : 1   
## candles : 1 cling film/bags: 1   
## (Other) : 18 (Other) : 12   
## X.18 X.19   
## :9820 :9826   
## waffles : 2 bottled beer : 2   
## chocolate marshmallow: 1 bottled water: 1   
## cling film/bags : 1 cake bar : 1   
## dental care : 1 coffee : 1   
## dog food : 1 flour : 1   
## (Other) : 8 (Other) : 2   
## X.20 X.21   
## :9827 :9827   
## chocolate : 2 chocolate : 1   
## fruit/vegetable juice : 1 female sanitary products: 1   
## liquor (appetizer) : 1 long life bakery product: 1   
## long life bakery product: 1 margarine : 1   
## pasta : 1 rum : 1   
## white wine : 1 (Other) : 2   
## X.22 X.23   
## :9828 :9829   
## abrasive cleaner : 1 chocolate : 1   
## chocolate : 1 hygiene articles: 1   
## hygiene articles : 2 napkins : 2   
## long life bakery product: 1 sugar : 1   
## specialty fat : 1   
##   
## X.24 X.25 X.26   
## :9830 :9833 :9833   
## cooking chocolate : 1 skin care: 1 hygiene articles: 1   
## house keeping products: 2   
## soups : 1   
##   
##   
##   
## X.27   
## :9833   
## candles: 1   
##   
##   
##   
##

## Association rules for the grocery store

The grocery dataset shows the items that have been bought together. Now we going to see how often they are bought together , and rules.

We can then see how many transactions we have and what they are.

## [1] "Description of the transactions"

## transactions in sparse format with  
## 9835 transactions (rows) and  
## 169 items (columns)

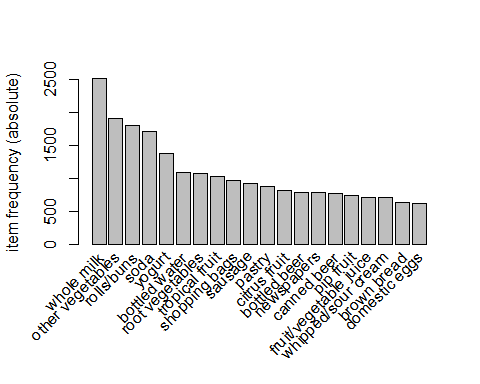
## transactions as itemMatrix in sparse format with  
## 9835 rows (elements/itemsets/transactions) and  
## 169 columns (items) and a density of 0.02609146   
##   
## most frequent items:  
## whole milk other vegetables rolls/buns soda   
## 2513 1903 1809 1715   
## yogurt (Other)   
## 1372 34055   
##   
## element (itemset/transaction) length distribution:  
## sizes  
## 1 2 3 4 5 6 7 8 9 10 11 12 13 14 15   
## 2159 1643 1299 1005 855 645 545 438 350 246 182 117 78 77 55   
## 16 17 18 19 20 21 22 23 24 26 27 28 29 32   
## 46 29 14 14 9 11 4 6 1 1 1 1 3 1   
##   
## Min. 1st Qu. Median Mean 3rd Qu. Max.   
## 1.000 2.000 3.000 4.409 6.000 32.000   
##   
## includes extended item information - examples:  
## labels  
## 1 abrasive cleaner  
## 2 artif. sweetener  
## 3 baby cosmetics

We see 9835 transactions, this is the number of rows as well, and 169 items.

The summary gives us some useful information:

* density: The percentage of non-empty cells in the sparse matrix. In another word, the total number of items that was purchased divided by the total number of possible items in that matrix.
* Transactions sizes There are 2159 transactions with one item, 1643 with 2 items, 1299 with 3 items, 1299 with items. This indicates that most customers buy small number of items on each purchase.
* The data distribution is right skewed.

Let’s have a look item frequency plot.



## Create some rules

* We use the Apriori algorithm in a rules library to mine frequent itemsets and association rules. The algorithm employs level-wise search for frequent itemsets.
* We pass supp=0.001 and conf=0.8 to return all the rules have a support of at least 0.1% and confidence of at least 80%.
* We sort the rules by decreasing confidence.
* Have a look the summary of the rules.

## Apriori  
##   
## Parameter specification:  
## confidence minval smax arem aval originalSupport maxtime support minlen  
## 0.8 0.1 1 none FALSE TRUE 5 0.001 1  
## maxlen target ext  
## 10 rules FALSE  
##   
## Algorithmic control:  
## filter tree heap memopt load sort verbose  
## 0.1 TRUE FALSE TRUE 2 TRUE  
##   
## Absolute minimum support count: 9   
##   
## set item appearances ...[0 item(s)] done [0.00s].  
## set transactions ...[169 item(s), 9835 transaction(s)] done [0.00s].  
## sorting and recoding items ... [157 item(s)] done [0.00s].  
## creating transaction tree ... done [0.00s].  
## checking subsets of size 1 2 3 4 5 6 done [0.02s].  
## writing ... [410 rule(s)] done [0.00s].  
## creating S4 object ... done [0.00s].

## set of 410 rules  
##   
## rule length distribution (lhs + rhs):sizes  
## 3 4 5 6   
## 29 229 140 12   
##   
## Min. 1st Qu. Median Mean 3rd Qu. Max.   
## 3.000 4.000 4.000 4.329 5.000 6.000   
##   
## summary of quality measures:  
## support confidence lift count   
## Min. :0.001017 Min. :0.8000 Min. : 3.131 Min. :10.00   
## 1st Qu.:0.001017 1st Qu.:0.8333 1st Qu.: 3.312 1st Qu.:10.00   
## Median :0.001220 Median :0.8462 Median : 3.588 Median :12.00   
## Mean :0.001247 Mean :0.8663 Mean : 3.951 Mean :12.27   
## 3rd Qu.:0.001322 3rd Qu.:0.9091 3rd Qu.: 4.341 3rd Qu.:13.00   
## Max. :0.003152 Max. :1.0000 Max. :11.235 Max. :31.00   
##   
## mining info:  
## data ntransactions support confidence  
## transactions 9835 0.001 0.8

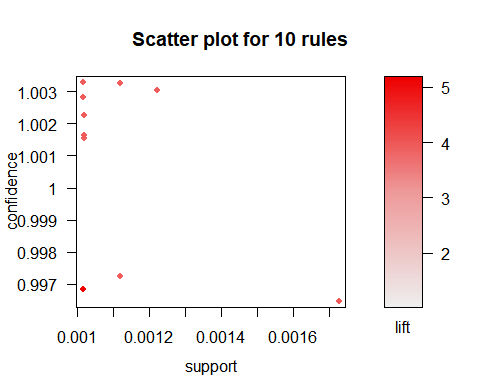
* The number of rules: 9835.
* The distribution of rules by length: Most rules are 4 items long.
* The summary of quality measures: ranges of support, confidence, and lift.
* The information on the data mining: total data mined, and minimum parameters we set earlier.

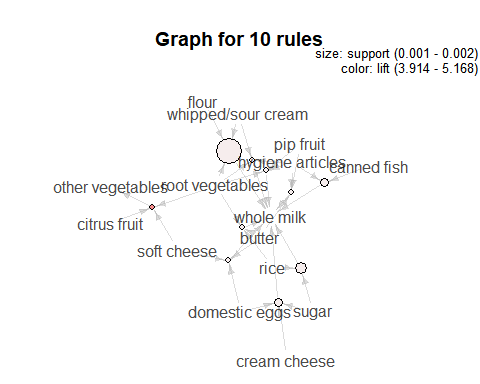
let’s inspect top 10.

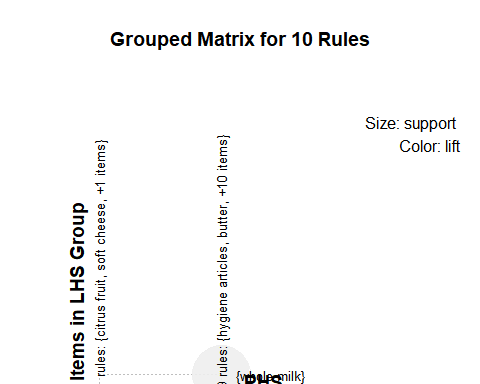
## lhs rhs support confidence lift count  
## [1] {rice,   
## sugar} => {whole milk} 0.001220132 1 3.913649 12  
## [2] {canned fish,   
## hygiene articles} => {whole milk} 0.001118454 1 3.913649 11  
## [3] {butter,   
## rice,   
## root vegetables} => {whole milk} 0.001016777 1 3.913649 10  
## [4] {flour,   
## root vegetables,   
## whipped/sour cream} => {whole milk} 0.001728521 1 3.913649 17  
## [5] {butter,   
## domestic eggs,   
## soft cheese} => {whole milk} 0.001016777 1 3.913649 10  
## [6] {citrus fruit,   
## root vegetables,   
## soft cheese} => {other vegetables} 0.001016777 1 5.168156 10  
## [7] {butter,   
## hygiene articles,   
## pip fruit} => {whole milk} 0.001016777 1 3.913649 10  
## [8] {hygiene articles,   
## root vegetables,   
## whipped/sour cream} => {whole milk} 0.001016777 1 3.913649 10  
## [9] {hygiene articles,   
## pip fruit,   
## root vegetables} => {whole milk} 0.001016777 1 3.913649 10  
## [10] {cream cheese,   
## domestic eggs,   
## sugar} => {whole milk} 0.001118454 1 3.913649 11

* 100% customers who bought {rice, sugar} end up bought {whole milk} as well.
* 100% customers who bought {canned fish,hygiene articles} end up bought {whole milk} as well.

And plot these top 10 rules.







In this post, we have learned how to Perform Market Basket Analysis in R and how to interpret the results.

reference: [R and Data Mining](http://www.rdatamining.com/examples/association-rules)