Measuring Performance: Evaluating Variable Reduction Techniques

Spiro Stilianoudakis

Contents

Loading Packages **1**

Setting Working directory **2**

Variable Selection Techniques **3**
Forward Selection 3
Backward Selection 4
Both 6

RFE **8**

Comparing Variable Selection Techniques **10**

Loading Packages

```
library(caret)
```

```
## Warning: package 'caret' was built under R version 3.4.4
```

```
## Loading required package: lattice
```

```
## Loading required package: ggplot2
```

```
## Warning: package 'ggplot2' was built under R version 3.4.4
```

```
#library(data.table)
library(gbm)
```

```
## Loading required package: survival
```

```
##
## Attaching package: 'survival'
```

```
## The following object is masked from 'package:caret':
##
##     cluster
```

```
## Loading required package: splines
```

```
## Loading required package: parallel
```

```
## Loaded gbm 2.1.3
```

```
library(pROC)
```

```
## Warning: package 'pROC' was built under R version 3.4.4
```

```
## Type 'citation("pROC")' for a citation.
```

```
##
## Attaching package: 'pROC'

## The following objects are masked from 'package:stats':
##
##     cov, smooth, var
```

```
library(plyr)
library(dplyr)
```

```
## Warning: package 'dplyr' was built under R version 3.4.4

##
## Attaching package: 'dplyr'

## The following objects are masked from 'package:plyr':
##
##     arrange, count, desc, failwith, id, mutate, rename, summarise,
##     summarize

## The following objects are masked from 'package:stats':
##
##     filter, lag

## The following objects are masked from 'package:base':
##
##     intersect, setdiff, setequal, union
```

```
#library(DMwR)
library(gridExtra)
```

```
##
## Attaching package: 'gridExtra'

## The following object is masked from 'package:dplyr':
##
##     combine
```

```
library(ggplot2)
library(leaps)
library(limma)
#library(DT)
library(knitr)
```

```
## Warning: package 'knitr' was built under R version 3.4.4
```

Setting Working directory

```
#reading in dataset
setwd("C:/Users/Spiro Stilianoudakis/Documents/TAD_data/RData/GM12878")
chr1_gm12878_f <- readRDS("C:/Users/Spiro Stilianoudakis/Documents/TAD_data/RData/GM12878/chr1_gm12878_

#set directory for selection techinques
setwd("C:/Users/Spiro Stilianoudakis/Documents/TAD_data/RData/GM12878/variable_reduction")
```

Variable Selection Techniques

Forward Selection

```
#rds objects for dataset
auc.model.fwd <- readRDS("C:/Users/Spiro Stilianoudakis/Documents/TAD_data/RData/GM12878/variable_reduc
cv.preds.fwd <- readRDS("C:/Users/Spiro Stilianoudakis/Documents/TAD_data/RData/GM12878/variable_reducti

#rds object for roc and aucs
enetlst_fwd <- readRDS("C:/Users/Spiro Stilianoudakis/Documents/TAD_data/RData/GM12878/variable_reductio

auc.fwd <- round(mean(enetlst_fwd[[3]]),3)
auc.fwd
```

```
## [1] 0.81
```

```
#0.810

vars.fwd <- na.omit(cv.preds.fwd[,which.max(auc.model.fwd)])
vars.fwd[grep("_dist",vars.fwd,invert = TRUE)] <- unlist(lapply(vars.fwd[grep("_dist",vars.fwd,invert =

chr1_gm12878_fwd <- chr1_gm12878_f[,which((names(chr1_gm12878_f) %in% vars.fwd) | names(chr1_gm12878_f)=

dim(chr1_gm12878_fwd)
```

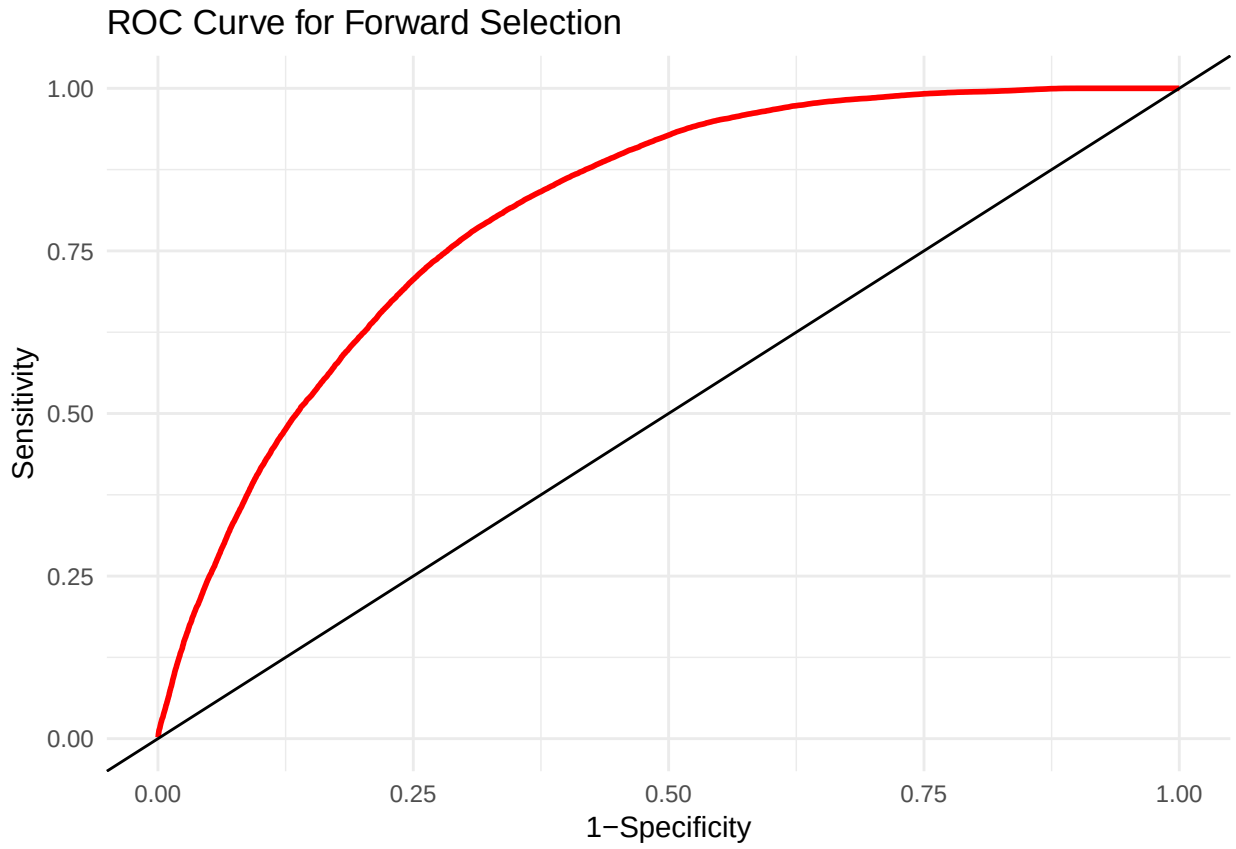
```
## [1] 247632     29
```

```
#247632   29

names(chr1_gm12878_fwd)
```

```
##  [1] "y"                           "complex_dist"
##  [3] "inversion_dist"              "mobile_element_insertion_dist"
##  [5] "low_complexity_dist"         "LTR_dist"
##  [7] "RC_dist"                     "Gm12878_WeakTxn"
##  [9] "Gm12878_Heterochromlo"       "Gm12878_TxnElongation_dist"
## [11] "Gm12878_WeakTxn_dist"        "Gm12878_Heterochromlo_dist"
## [13] "Gm12878_WeakPromoter_dist"   "Gm12878_Insulator_dist"
## [15] "Gm12878_T"                   "Gm12878_CTCF_dist"
## [17] "Gm12878_PF_dist"             "Gm12878_T_dist"
## [19] "Gm12878_TSS_dist"            "Gm12878_DNaseI"
## [21] "Gm12878_DNaseI_dist"         "Gm12878_H3k27ac"
## [23] "Gm12878_H3k9ac"              "Gm12878_H4k20me1"
## [25] "Gm12878_H2az_dist"           "Gm12878_H3k36me3_dist"
## [27] "Gm12878_H3k9ac_dist"         "Gm12878_H3k9me3_dist"
## [29] "Gm12878_H4k20me1_dist"
```

```
setwd("C:/Users/Spiro Stilianoudakis/Documents/TAD_data/RData/GM12878")
saveRDS(chr1_gm12878_fwd, "chr1_gm12878_fwd.rds")

rocdat.fwd <- data.frame(sensitivity=rowMeans(enetlst_fwd[[1]]), specificity=rowMeans(enetlst_fwd[[2]]))
rocdat.fwd$Selection <- "fwd"

ggplot(rocdat.fwd, aes(x=specificity, y=sensitivity)) +
  geom_line(size=1, color="red") +
```

```
  xlab("1-Specificity") +
  ylab("Sensitivity") +
  xlim(0, 1) +
  ylim(0, 1) +
  geom_abline(intercept=0, slope=1) +
  theme_minimal() +
  ggtitle("ROC Curve for Forward Selection")
```



Backward Selection

```
#rds object for dataset
auc.model.bwd <- readRDS("C:/Users/Spiro Stilianoudakis/Documents/TAD_data/RData/GM12878/variable_reduc
cv.preds.bwd <- readRDS("C:/Users/Spiro Stilianoudakis/Documents/TAD_data/RData/GM12878/variable_reducti

#rds object for roc and aucs
enetlst_bwd <- readRDS("C:/Users/Spiro Stilianoudakis/Documents/TAD_data/RData/GM12878/variable_reductio

auc.bwd <- round(mean(enetlst_bwd[[3]]),3)
auc.bwd
```

```
## [1] 0.81
```

```
#0.81

vars.bwd <- na.omit(cv.preds.bwd[,which.max(auc.model.bwd)])
```

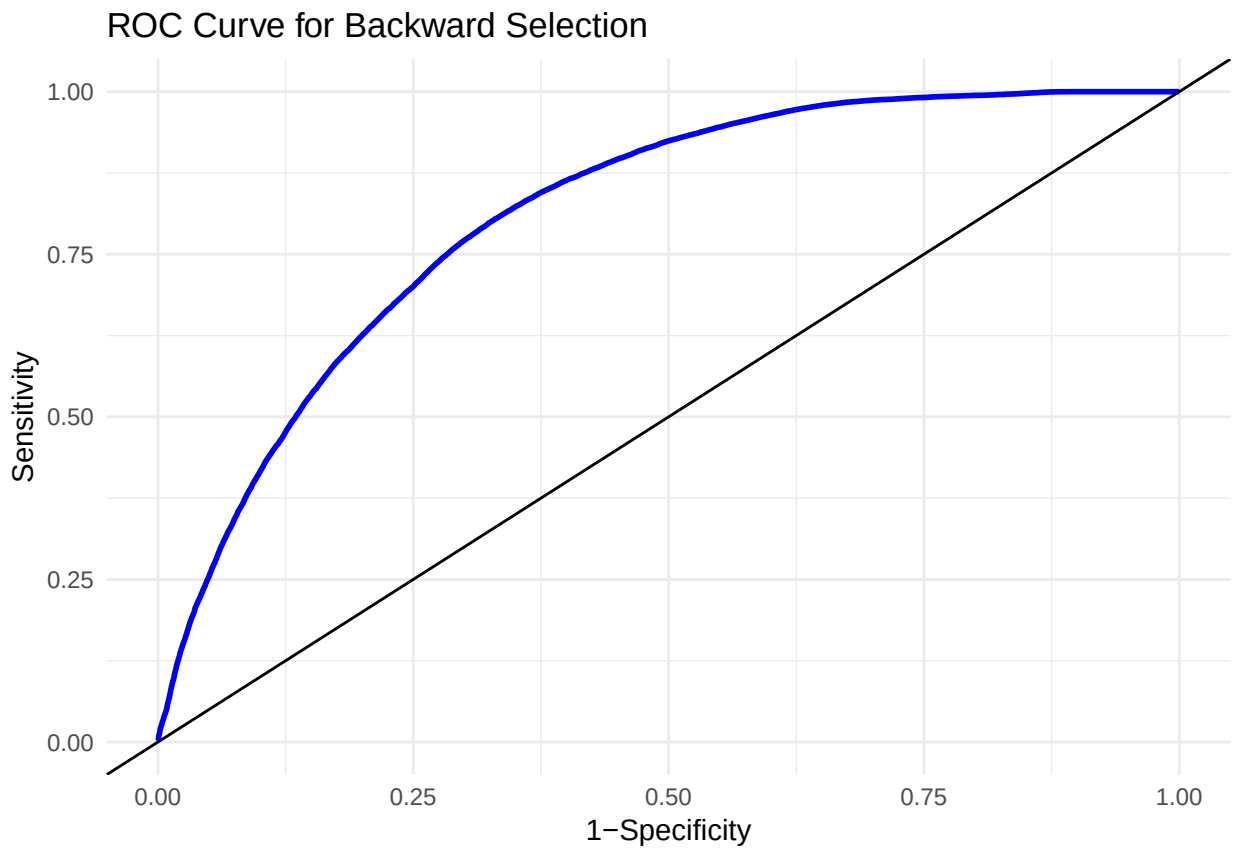
```r
vars.bwd[grep("_dist",vars.bwd,invert = TRUE)] <- unlist(lapply(vars.bwd[grep("_dist",vars.bwd,invert =

chr1_gm12878_bwd <- chr1_gm12878_f[,which((names(chr1_gm12878_f) %in% vars.bwd) | names(chr1_gm12878_f)=

dim(chr1_gm12878_bwd)
```

```
## [1] 247632     35
```

```r
#247632     35

names(chr1_gm12878_bwd)
```

```
##  [1] "y"                                "A"
##  [3] "B"                                "complex_dist"
##  [5] "inversion_dist"                   "mobile_element_insertion_dist"
##  [7] "SINE"                             "low_complexity_dist"
##  [9] "LTR_dist"                         "RC_dist"
## [11] "SINE_dist"                        "Gm12878_WeakTxn"
## [13] "Gm12878_Repressed"                "Gm12878_Heterochromlo"
## [15] "Gm12878_TxnElongation_dist"       "Gm12878_WeakTxn_dist"
## [17] "Gm12878_Repressed_dist"           "Gm12878_Heterochromlo_dist"
## [19] "Gm12878_WeakPromoter_dist"        "Gm12878_Insulator_dist"
## [21] "Gm12878_T"                        "Gm12878_CTCF_dist"
## [23] "Gm12878_T_dist"                   "Gm12878_TSS_dist"
## [25] "Gm12878_DNaseI"                   "Gm12878_DNaseI_dist"
## [27] "Gm12878_H3k27ac"                  "Gm12878_H3k36me3"
## [29] "Gm12878_H3k9ac"                   "Gm12878_H4k20me1"
## [31] "Gm12878_H2az_dist"                "Gm12878_H3k4me2_dist"
## [33] "Gm12878_H3k9ac_dist"              "Gm12878_H3k9me3_dist"
## [35] "Gm12878_H4k20me1_dist"
```

```r
setwd("C:/Users/Spiro Stilianoudakis/Documents/TAD_data/RData/GM12878")
saveRDS(chr1_gm12878_bwd, "chr1_gm12878_bwd.rds")

rocdat.bwd <- data.frame(sensitivity=rowMeans(enetlst_bwd[[1]]), specificity=rowMeans(enetlst_bwd[[2]]))
rocdat.bwd$Selection <- "bwd"

ggplot(rocdat.bwd, aes(x=specificity, y=sensitivity)) +
  geom_line(size=1, color="blue") +
  xlab("1-Specificity") +
  ylab("Sensitivity") +
  xlim(0, 1) +
  ylim(0, 1) +
  geom_abline(intercept=0, slope=1) +
  theme_minimal() +
  ggtitle("ROC Curve for Backward Selection")
```



Both

```
#rds objects for datasets
auc.model.both <- readRDS("C:/Users/Spiro Stilianoudakis/Documents/TAD_data/RData/GM12878/variable_redu
cv.preds.both <- readRDS("C:/Users/Spiro Stilianoudakis/Documents/TAD_data/RData/GM12878/variable_reduc

#rds object for roc and aucs
enetlst_both <- readRDS("C:/Users/Spiro Stilianoudakis/Documents/TAD_data/RData/GM12878/variable_reducti

auc.both <- round(mean(enetlst_both[[3]]),3)
auc.both
```

```
## [1] 0.809
```

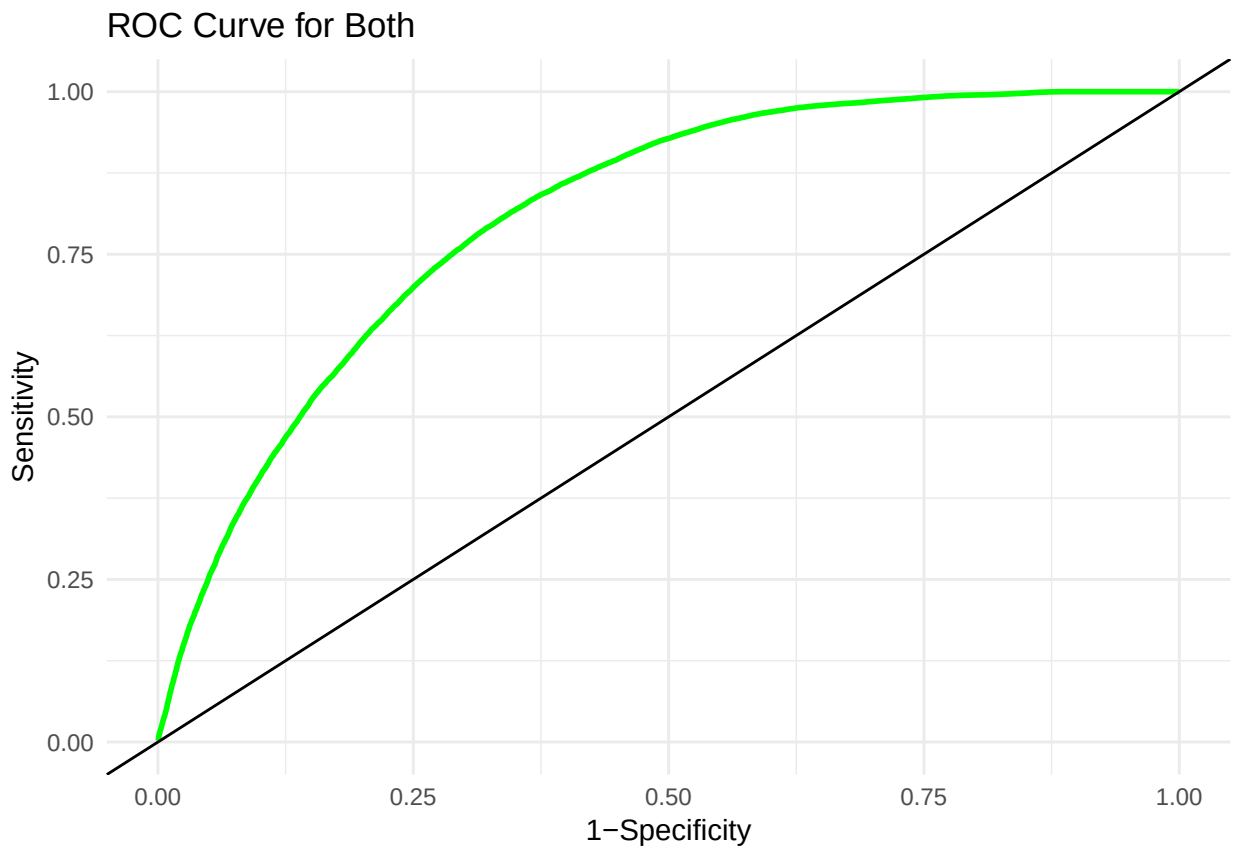
```
#0.809

vars.both <- na.omit(cv.preds.both[,which.max(auc.model.both)])
vars.both[grep("_dist",vars.both,invert = TRUE)] <- unlist(lapply(vars.both[grep("_dist",vars.both,inve

chr1_gm12878_both <- chr1_gm12878_f[,which((names(chr1_gm12878_f) %in% vars.both) | names(chr1_gm12878_

dim(chr1_gm12878_both)
```

```
## [1] 247632     27
```

```
#247632     27

names(chr1_gm12878_both)
```

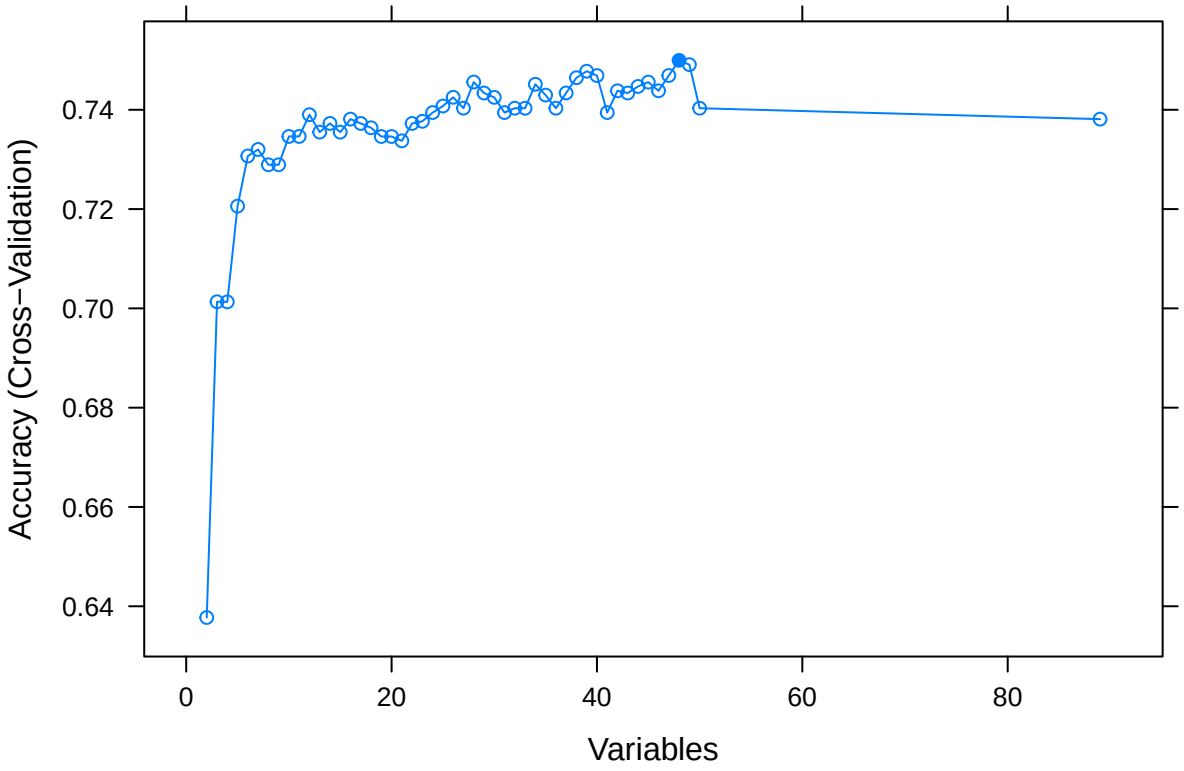
```
##  [1] "y"                              "complex_dist"
##  [3] "mobile_element_insertion_dist" "low_complexity_dist"
##  [5] "LTR_dist"                       "Gm12878_Heterochromlo"
##  [7] "Gm12878_TxnElongation_dist"     "Gm12878_WeakTxn_dist"
##  [9] "Gm12878_Heterochromlo_dist"     "Gm12878_Insulator_dist"
## [11] "Gm12878_R"                      "Gm12878_T"
## [13] "Gm12878_CTCF_dist"              "Gm12878_PF_dist"
## [15] "Gm12878_TSS_dist"               "Gm12878_DNaseI"
## [17] "Gm12878_DNaseI_dist"            "Gm12878_H3k27ac"
## [19] "Gm12878_H3k79me2"               "Gm12878_H3k9ac"
## [21] "Gm12878_H3k9me3"                "Gm12878_H4k20me1"
## [23] "Gm12878_H2az_dist"              "Gm12878_H3k36me3_dist"
## [25] "Gm12878_H3k4me2_dist"           "Gm12878_H3k9ac_dist"
## [27] "Gm12878_H4k20me1_dist"
```

```
setwd("C:/Users/Spiro Stilianoudakis/Documents/TAD_data/RData/GM12878")
saveRDS(chr1_gm12878_both, "chr1_gm12878_both.rds")


rocdat.both <- data.frame(sensitivity=rowMeans(enetlst_both[[1]]), specificity=rowMeans(enetlst_both[[2]
rocdat.both$Selection <- "both"

ggplot(rocdat.both, aes(x=specificity, y=sensitivity)) +
  geom_line(size=1, color="green") +
  xlab("1-Specificity") +
  ylab("Sensitivity") +
  xlim(0, 1) +
  ylim(0, 1) +
  geom_abline(intercept=0, slope=1) +
  theme_minimal() +
  ggtitle("ROC Curve for Both")
```



RFE

```
rfeModel <- readRDS("C:/Users/Spiro Stilianoudakis/Documents/TAD_data/RData/GM12878/variable_reduction/
roc.rfeModel <- readRDS("C:/Users/Spiro Stilianoudakis/Documents/TAD_data/RData/GM12878/variable_reducti

auc.rfe <- round(pROC::auc(roc.rfeModel),3)
#0.7986

plot(rfeModel, type="b")
```

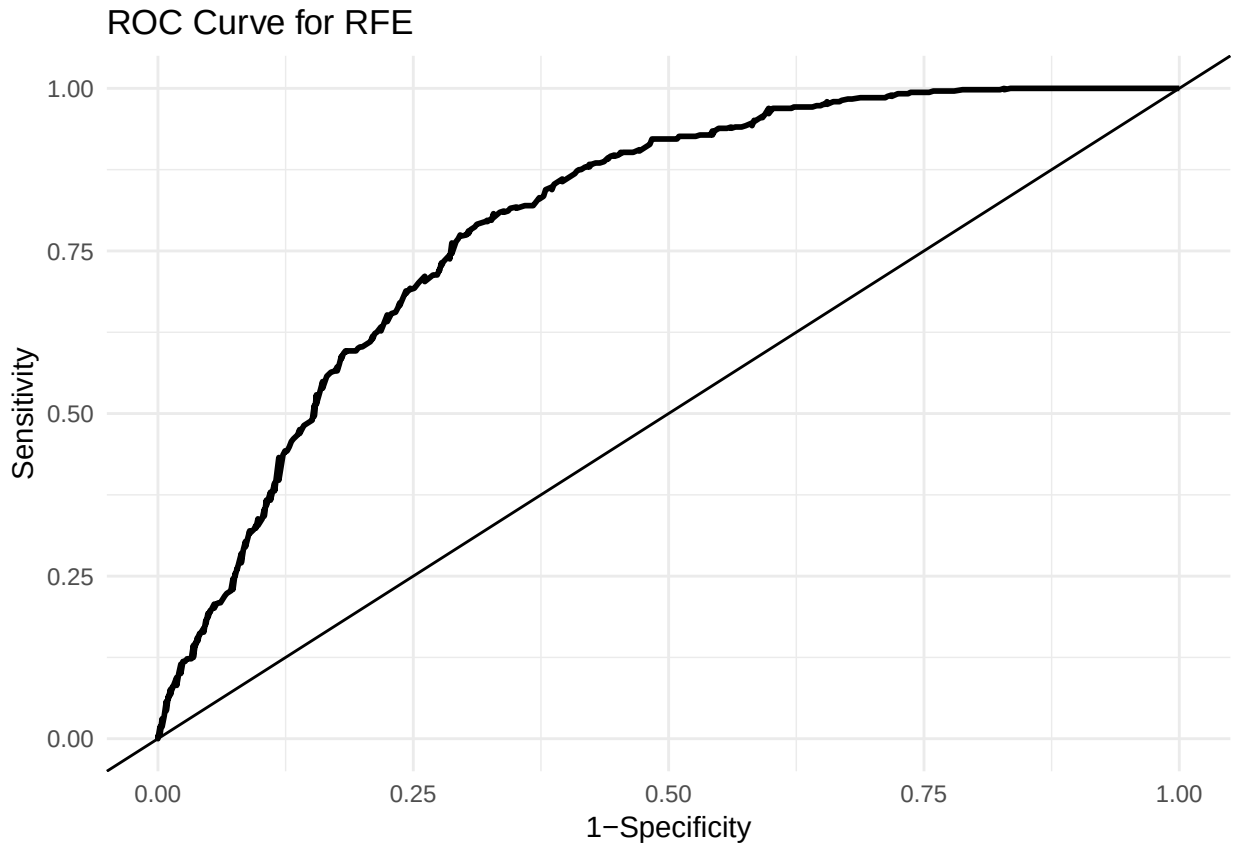



```
accdat <- rfeModel$results
accdat <- accdat[order(accdat$Accuracy, decreasing = TRUE),]

predictors(rfeModel)
```

```
##  [1] "Gm12878_Insulator_dist"         "Gm12878_CTCF_dist"
##  [3] "Gm12878_DNaseI_dist"            "Gm12878_H2az_dist"
##  [5] "Gm12878_TSS_dist"               "Gm12878_H3k9ac_dist"
##  [7] "Gm12878_PF_dist"                "Gm12878_WeakPromoter_dist"
##  [9] "Gm12878_ActivePromoter_dist"    "Gm12878_H3k79me2_dist"
## [11] "Gm12878_H3k4me2_dist"           "Gm12878_H3k27ac_dist"
## [13] "Gm12878_WeakTxn_dist"           "Gm12878_H3k36me3_dist"
## [15] "Gm12878_TxnElongation_dist"     "Gm12878_StrongEnhancer5_dist"
## [17] "Gm12878_StrongEnhancer4_dist"   "Gm12878_H3k4me1_dist"
## [19] "se_GM12878_dist"                "Gm12878_WE_dist"
## [21] "Gm12878_H3k4me3_dist"           "Gm12878_H4k20me1_dist"
## [23] "VMR_dist"                       "Gm12878_WeakEnhancer6_dist"
## [25] "Gm12878_H3k27me3_dist"          "Gm12878_TxnTransition_dist"
## [27] "Gm12878_E_dist"                 "Gm12878_Repressed_dist"
## [29] "DNA_dist"                       "Gm12878_R_dist"
## [31] "Gm12878_H3k9me3_dist"           "Gm12878_Heterochromlo_dist"
## [33] "A"                              "Gm12878_PoisedPromoter_dist"
## [35] "UCNE_dist"                      "Gm12878_DNaseI"
## [37] "Gm12878_T_dist"                 "SINE_dist"
## [39] "simple_repeat_dist"             "sequence_alteration_dist"
## [41] "novel_sequence_insertion_dist" "Gm12878_WeakEnhancer7_dist"
```

```
## [43] "Gm12878_H3k36me3"              "duplication_dist"
## [45] "Gm12878_H2az"                  "Gm12878_RepetitiveCNV15_dist"
## [47] "LTR_dist"                      "tandem_duplication_dist"
```

```
rocdat.rfe <- data.frame(sensitivity=roc.rfeModel$sensitivities, specificity=1-roc.rfeModel$specificiti
rocdat.rfe$Selection <- "rfe"

ggplot(rocdat.rfe, aes(x=specificity, y=sensitivity)) +
  geom_line(size=1, color="black") +
  xlab("1-Specificity") +
  ylab("Sensitivity") +
  xlim(0, 1) +
  ylim(0, 1) +
  geom_abline(intercept=0, slope=1) +
  theme_minimal() +
  ggtitle("ROC Curve for RFE")
```



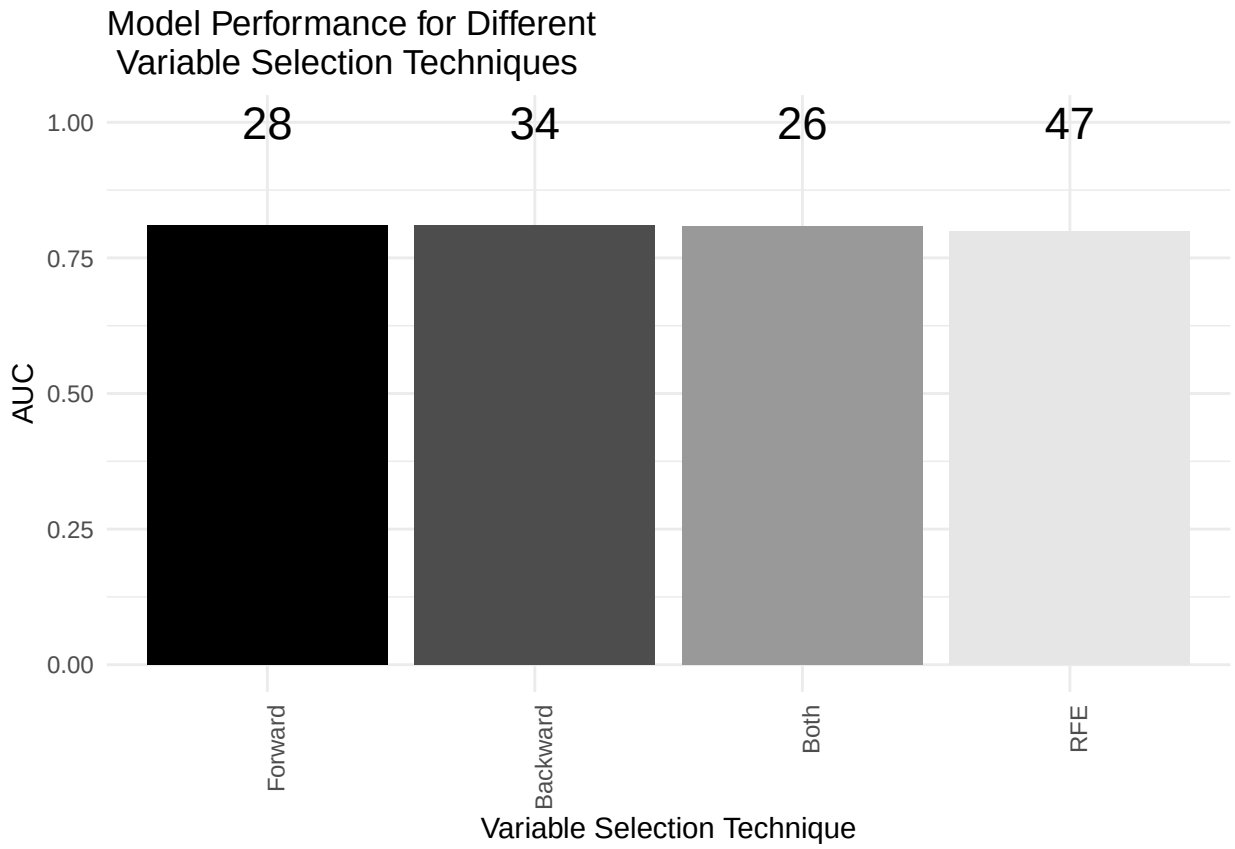
Comparing Variable Selection Techniques

```
auc.plot <- data.frame(Selection=c("Forward",
                                   "Backward",
                                   "Both",
                                   "RFE"),
                       auc=c(auc.fwd,
```

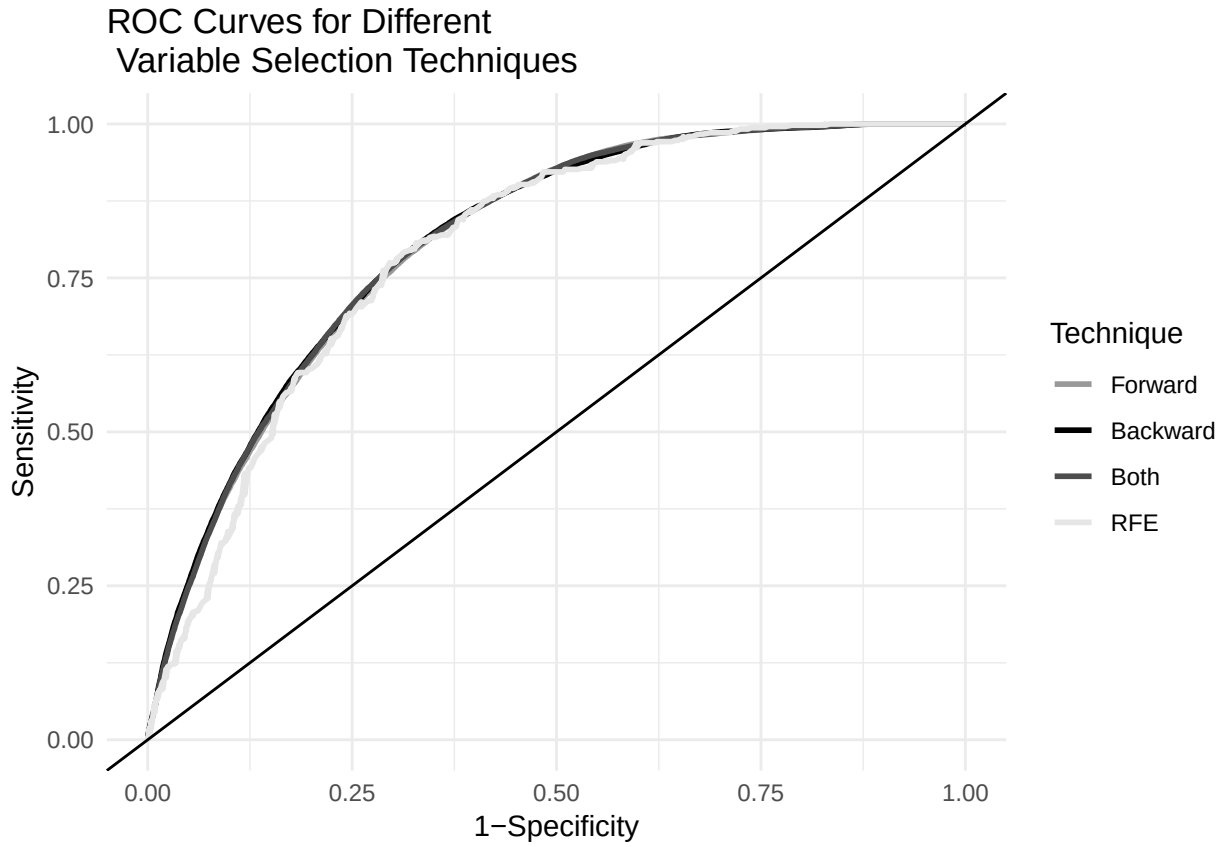
```
                              auc.bwd,
                              auc.both,
                              auc.rfe))

auc.plot <- auc.plot[order(auc.plot$auc, decreasing=TRUE),]

auc.plot$Selection <-factor(auc.plot$Selection,
                                      levels=auc.plot$Selection)

#datatable(auc.plot)
kable(auc.plot)
```

Selection	auc
Forward	0.810
Backward	0.810
Both	0.809
RFE	0.799

```
p<-ggplot(data=auc.plot, aes(x=Selection, y=auc, fill=Selection)) +
  xlab("Variable Selection Technique") + ylab("AUC") +
  geom_bar(stat="identity") + ylim(0,1) +
  scale_fill_manual(values=gray(c(0,.3,.6,.9)), guide=FALSE) +
  annotate("text", x=1, y=1, label= "28", size=6) +
  annotate("text", x=2, y=1, label= "34", size=6) +
  annotate("text", x=3, y=1, label= "26", size=6) +
  annotate("text", x=4, y=1, label= "47", size=6) +
  theme_minimal() +
  theme(axis.text.x = element_text(angle = 90, hjust = 1)) +
  ggtitle("Model Performance for Different \n Variable Selection Techniques")
p
```




```
allrocdat <- rbind.data.frame(rocdat.fwd, rocdat.bwd, rocdat.both, rocdat.rfe)

ggplot(data=allrocdat, aes(x=specificity, y=sensitivity, color=Selection)) +
  geom_line(size=1) +
  scale_colour_manual(name="Technique",
    labels=c("Forward",
             "Backward",
             "Both",
              "RFE"),
    values=c("#999999", "#000000", "#4D4D4D", "#E6E6E6")) +
  xlab("1-Specificity") +
  ylab("Sensitivity") +
  xlim(0, 1) +
  ylim(0, 1) +
  geom_abline(intercept=0, slope=1) +
  theme_minimal() +
  ggtitle("ROC Curves for Different \n Variable Selection Techniques")
```




```
intersect(vars.fwd,intersect(vars.bwd,vars.both))
```

```
##  [1] "Gm12878_DNaseI_dist"           "Gm12878_Insulator_dist"
##  [3] "Gm12878_TSS_dist"              "Gm12878_DNaseI"
##  [5] "Gm12878_H2az_dist"             "Gm12878_H3k9ac"
##  [7] "Gm12878_H3k9ac_dist"           "Gm12878_Heterochromlo_dist"
##  [9] "Gm12878_Heterochromlo"         "Gm12878_CTCF_dist"
## [11] "Gm12878_H4k20me1_dist"         "Gm12878_H4k20me1"
## [13] "low_complexity_dist"           "Gm12878_H3k27ac"
## [15] "Gm12878_TxnElongation_dist"    "Gm12878_WeakTxn_dist"
## [17] "Gm12878_T"                     "LTR_dist"
## [19] "complex_dist"                  "mobile_element_insertion_dist"
```

```
intersect(vars.fwd,intersect(vars.bwd,predictors(rfeModel)))
```

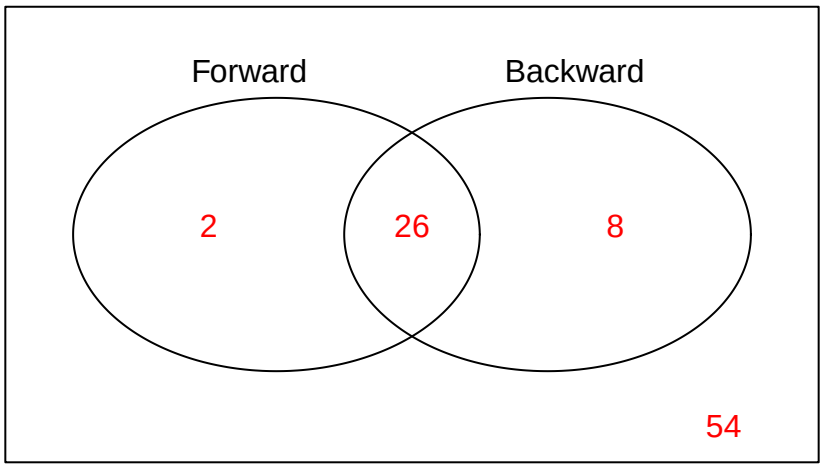
```
##  [1] "Gm12878_DNaseI_dist"        "Gm12878_Insulator_dist"
##  [3] "Gm12878_TSS_dist"           "Gm12878_DNaseI"
##  [5] "Gm12878_H2az_dist"          "Gm12878_H3k9ac_dist"
##  [7] "Gm12878_Heterochromlo_dist" "Gm12878_CTCF_dist"
##  [9] "Gm12878_H4k20me1_dist"      "Gm12878_H3k9me3_dist"
## [11] "Gm12878_TxnElongation_dist" "Gm12878_WeakTxn_dist"
## [13] "Gm12878_WeakPromoter_dist"  "LTR_dist"
## [15] "Gm12878_T_dist"
```

```
intersect(intersect(vars.fwd,intersect(vars.bwd,vars.both)),
          intersect(vars.fwd,intersect(vars.bwd,predictors(rfeModel))))
```

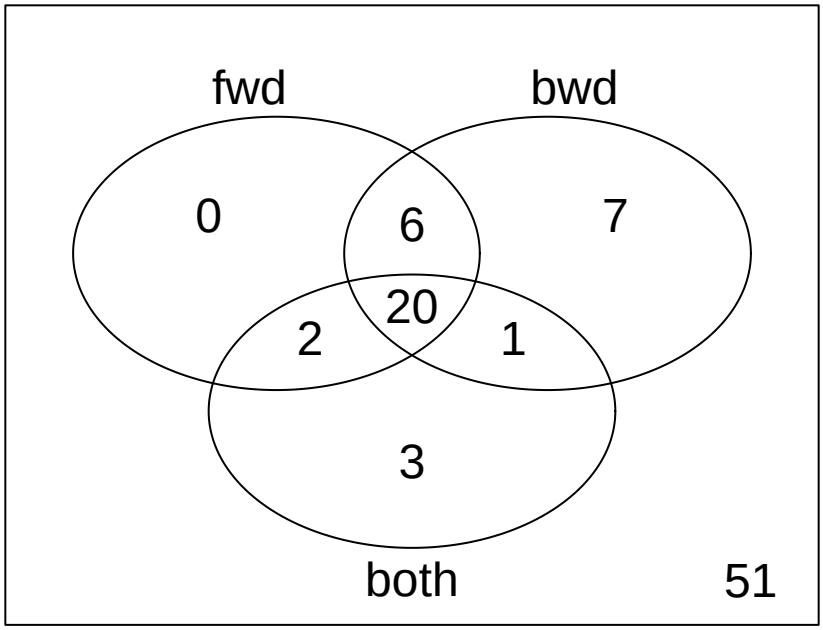
```
##  [1] "Gm12878_DNaseI_dist"        "Gm12878_Insulator_dist"
```

```
##  [3] "Gm12878_TSS_dist"            "Gm12878_DNaseI"
##  [5] "Gm12878_H2az_dist"           "Gm12878_H3k9ac_dist"
##  [7] "Gm12878_Heterochromlo_dist" "Gm12878_CTCF_dist"
##  [9] "Gm12878_H4k20me1_dist"       "Gm12878_TxnElongation_dist"
## [11] "Gm12878_WeakTxn_dist"        "LTR_dist"
```

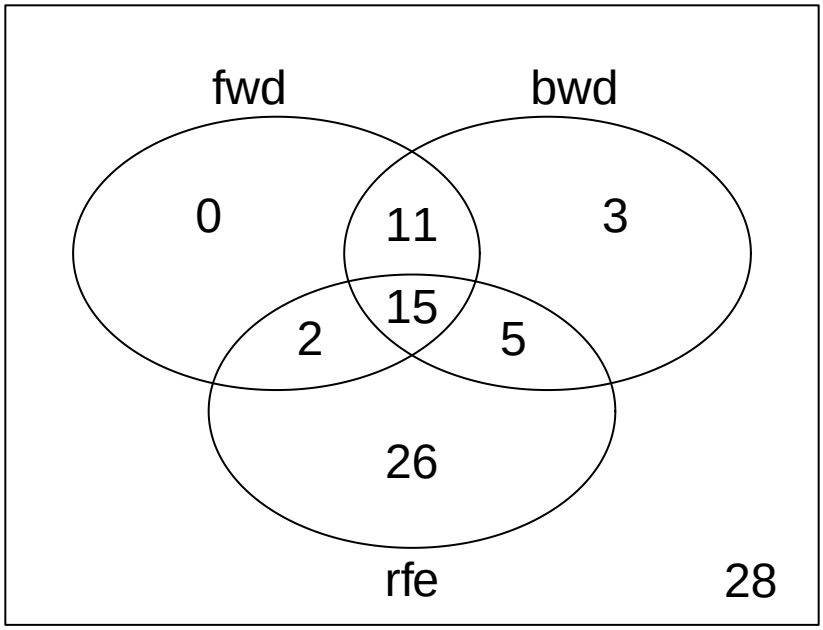
```
fwd <- (names(chr1_gm12878_f) %in% vars.fwd)
bwd <- (names(chr1_gm12878_f) %in% vars.bwd)
both <- (names(chr1_gm12878_f) %in% vars.both)
rfe <- (names(chr1_gm12878_f) %in% predictors(rfeModel))

#fwd compared to bwd
venndatfb <- cbind(fwd,bwd)
fb <- vennCounts(venndatfb)
vennDiagram(fb, include = "both",
  names = c("Forward", "Backward"),
  cex = 1, counts.col = "red")
```



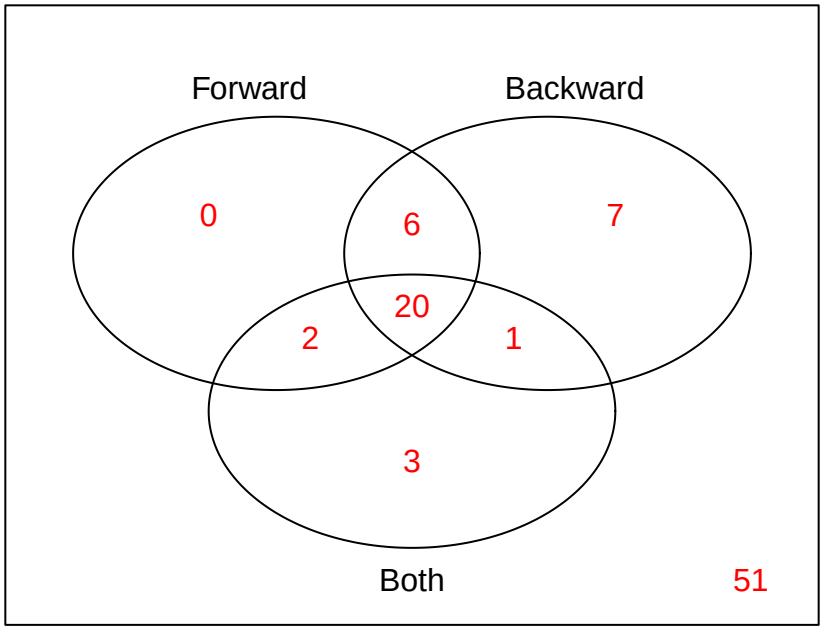

```
venndat1 <- cbind(fwd,bwd,both)
venndat2 <- cbind(fwd,bwd,rfe)
venndat3 <- cbind(fwd,bwd,both,rfe)

a <- vennCounts(venndat1)
b <- vennCounts(venndat2)
c <- vennCounts(venndat3)

vennDiagram(a)
```



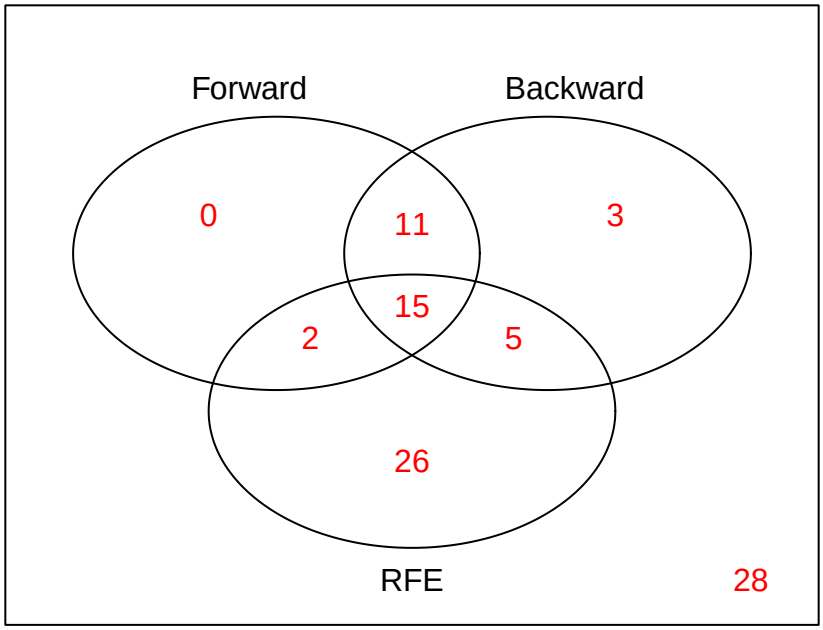

```
vennDiagram(b)
```



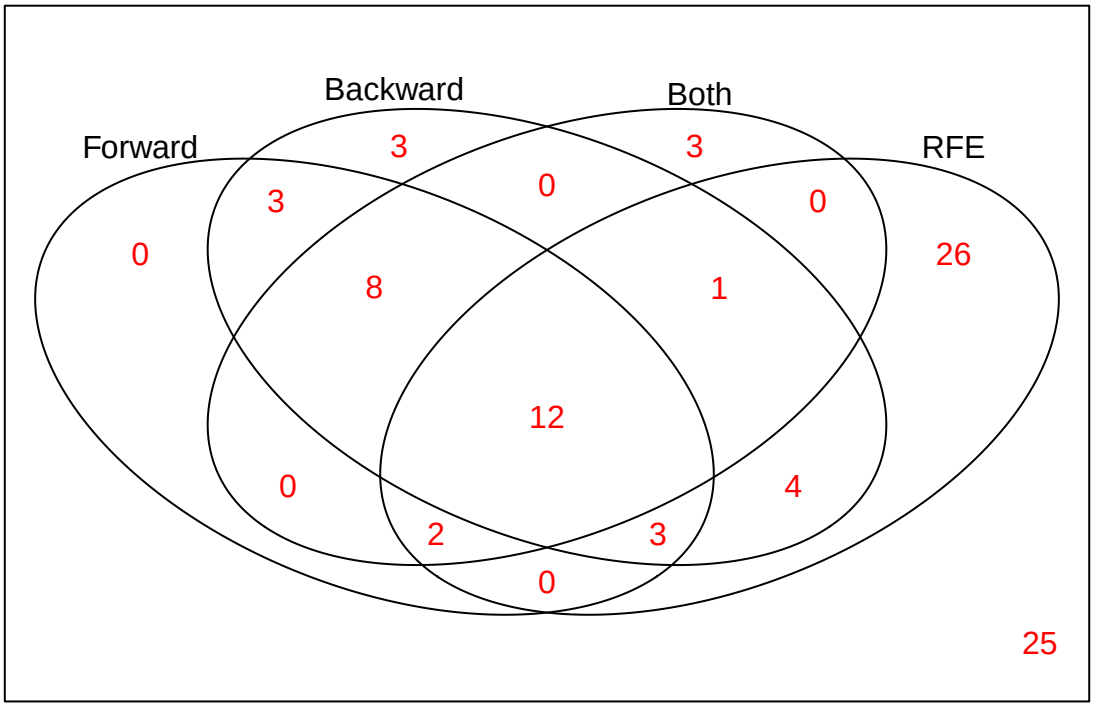

```
vennDiagram(a, include = "both",
  names = c("Forward", "Backward", "Both"),
  cex = 1, counts.col = "red")
```




```
vennDiagram(b, include = "both",
  names = c("Forward", "Backward", "RFE"),
  cex = 1, counts.col = "red")
```




```
vennDiagram(c, include = "both",
  names = c("Forward", "Backward","Both", "RFE"),
  cex = 1, counts.col = "red")
```


Backward
Both
Forward
RFE
3
3
3
0
0
0
26
8
1
12
0
4
2
3
0
25