Measuring Performance: Evaluating Normalization

Spiro Stilianoudakis

Contents

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(knitr)
```

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

Setting Working directory

```
setwd("C:/Users/Spiro Stilianoudakis/Documents/TAD_data/RData/GM12878/comparing_normalization")
```

Log tranformed and standardized

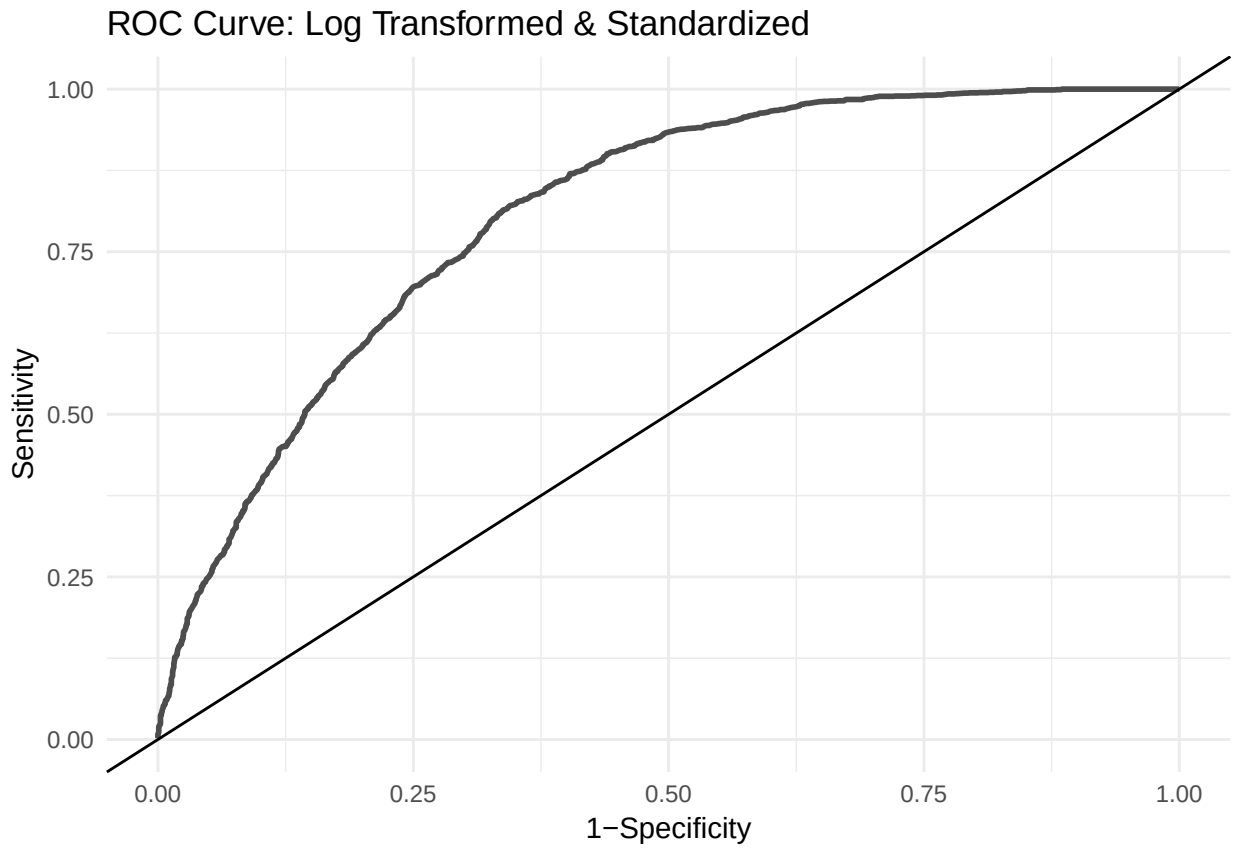
```
enetlst_ls <- readRDS("C:/Users/Spiro Stilianoudakis/Documents/TAD_data/RData/GM12878/comparing_normali

#Mean AUC across 100 bootstrap samples
enetlst_ls[[3]]
```

```
## [1] 0.8153019 0.8013215 0.8070258 0.8138884 0.7944170
```

```
auc.ls <- round(mean(enetlst_ls[[3]]),3)
auc.ls
```

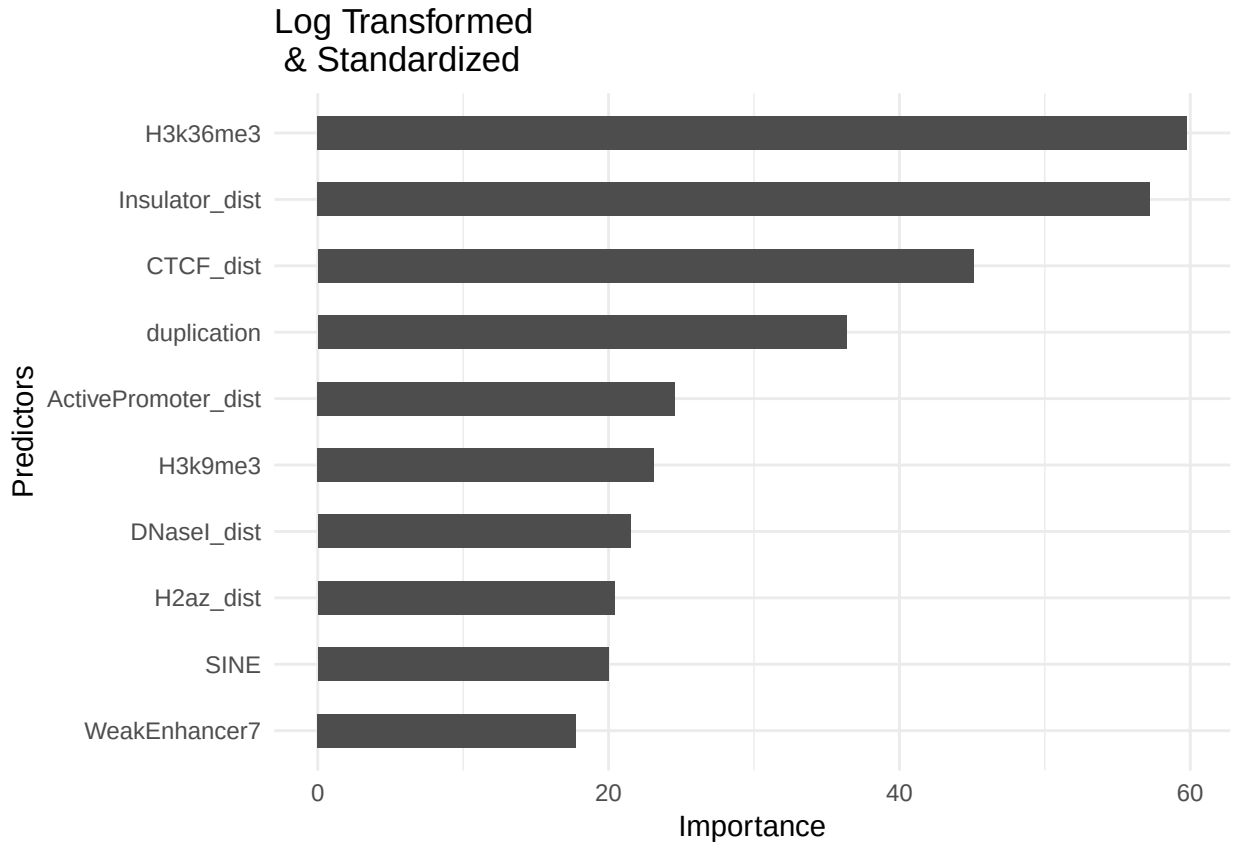
```
## [1] 0.806
```

```
#roc curve
fpr.ls <- rowMeans(enetlst_ls[[2]])
tpr.ls <- rowMeans(enetlst_ls[[1]])
rocdat.ls <- data.frame(fpr=fpr.ls, tpr=tpr.ls)
ggplot(rocdat.ls, aes(x=fpr, y=tpr)) +
  geom_line(size=1, color="#4D4D4D") +
  xlab("1-Specificity") +
  ylab("Sensitivity") +
  xlim(0, 1) +
  ylim(0, 1) +
  geom_abline(intercept=0, slope=1) +
  theme_minimal() +
  ggtitle("ROC Curve: Log Transformed & Standardized")
```




```
varimp.ls <- as.vector(rowMeans(enetlst_ls[[4]]))
Labels <- rownames(enetlst_ls[[4]])
Labels[grep("Gm12878_", Labels)] <- gsub("Gm12878_","",Labels[grep("Gm12878_", Labels)])
varimp.ls.df <- data.frame(Feature=Labels,
                           Importance=varimp.ls)
varimp.ls.df <- varimp.ls.df[order(varimp.ls.df$Importance),]
varimp.ls.df <- varimp.ls.df[(dim(varimp.ls.df)[1]-9):dim(varimp.ls.df)[1],]
```

```
varimp.ls.df$Feature <- factor(varimp.ls.df$Feature,
                                    levels=varimp.ls.df$Feature)
p.ls <- ggplot(varimp.ls.df, aes(x=Feature, y=Importance)) +
  xlab("Predictors") +
  ylab("Importance") +
  #ggtitle("Importance Plot for Gradient Boosting Machine") +
  geom_bar(stat="identity",
           width=.5,
           position="dodge",
           fill="#4D4D4D") +
  coord_flip() +
  theme_minimal() +
  ggtitle("Log Transformed \n & Standardized")
p.ls
```



Log tranformed and un-standardized

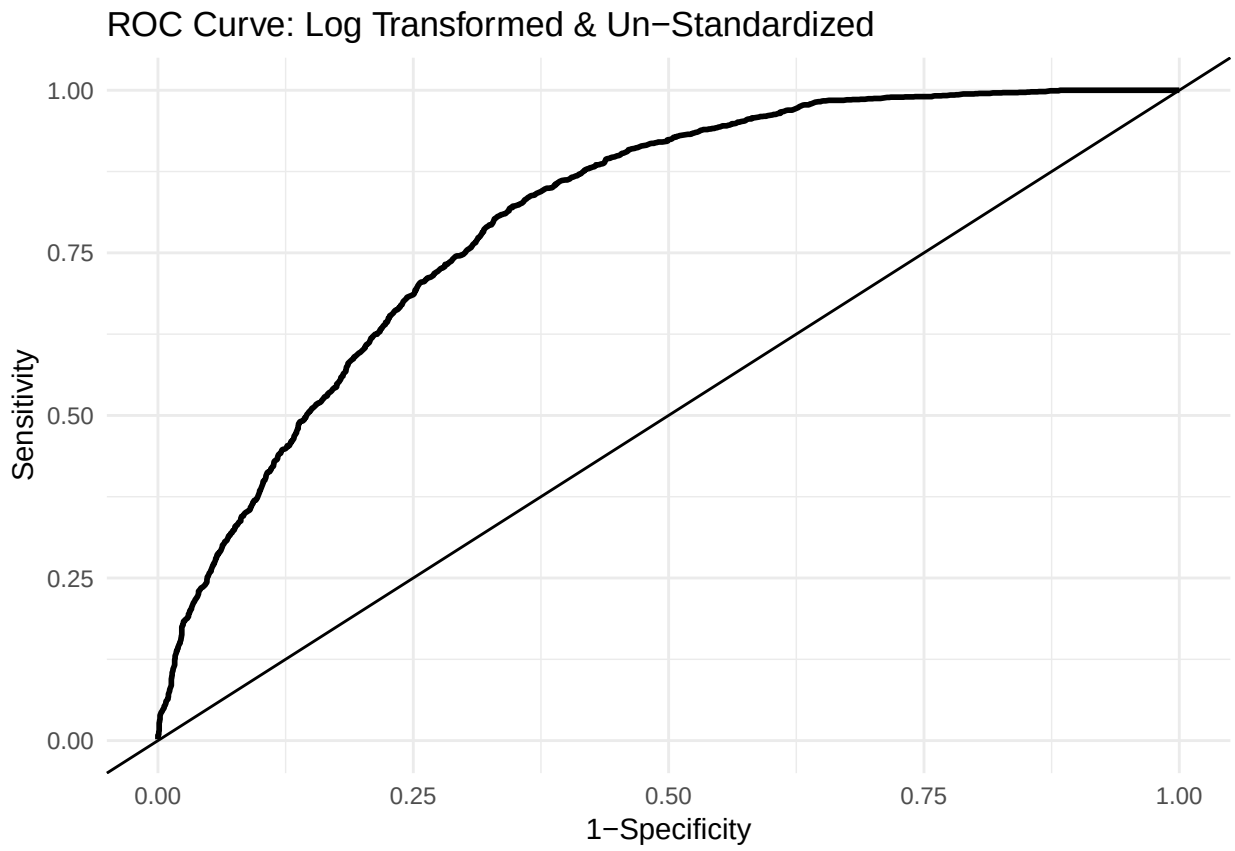
```
enetlst_lns <- readRDS("C:/Users/Spiro Stilianoudakis/Documents/TAD_data/RData/GM12878/comparing_normali

#Mean AUC across 100 bootstrap samples
enetlst_lns[[3]]
```

```
## [1] 0.8150594 0.8014177 0.8032703 0.8138675 0.7917615
```

```
auc.lns <- round(mean(enetlst_lns[[3]]),3)
auc.lns
```

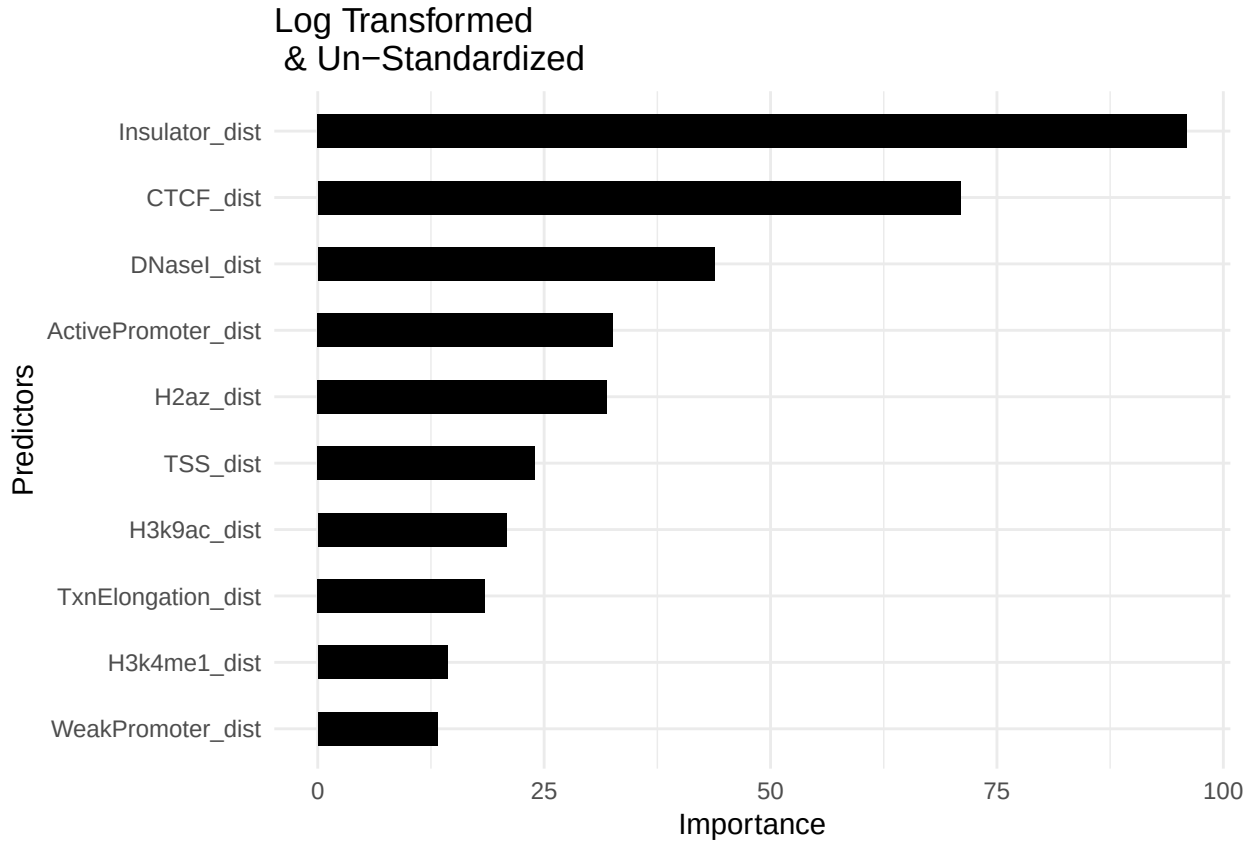
```
## [1] 0.805
```

```
#roc curve
fpr.lns <- rowMeans(enetlst_lns[[2]])
tpr.lns <- rowMeans(enetlst_lns[[1]])
rocdat.lns <- data.frame(fpr=fpr.lns, tpr=tpr.lns)
ggplot(rocdat.lns, aes(x=fpr.lns, y=tpr.lns)) +
  geom_line(size=1, color="#000000") +
  xlab("1-Specificity") +
  ylab("Sensitivity") +
  xlim(0, 1) +
  ylim(0, 1) +
  geom_abline(intercept=0, slope=1) +
  theme_minimal() +
  ggtitle("ROC Curve: Log Transformed & Un-Standardized")
```




```
varimp.lns <- as.vector(rowMeans(enetlst_lns[[4]]))
Labels <- rownames(enetlst_lns[[4]])
Labels[grep("Gm12878_", Labels)] <- gsub("Gm12878_","",Labels[grep("Gm12878_", Labels)])
varimp.lns.df <- data.frame(Feature=Labels,
                            Importance=varimp.lns)
varimp.lns.df <- varimp.lns.df[order(varimp.lns.df$Importance),]
varimp.lns.df <- varimp.lns.df[(dim(varimp.lns.df)[1]-9):dim(varimp.lns.df)[1],]
varimp.lns.df$Feature <- factor(varimp.lns.df$Feature,
```

```
                                      levels=varimp.lns.df$Feature)
p.lns <- ggplot(varimp.lns.df, aes(x=Feature, y=Importance)) +
  xlab("Predictors") +
  ylab("Importance") +
  #ggtitle("Importance Plot for Gradient Boosting Machine") +
  geom_bar(stat="identity",
           width=.5,
           position="dodge",
           fill="#000000") +
  coord_flip() +
  theme_minimal() +
  ggtitle("Log Transformed \n & Un-Standardized")
p.lns
```



Not Log tranformed and Standardized

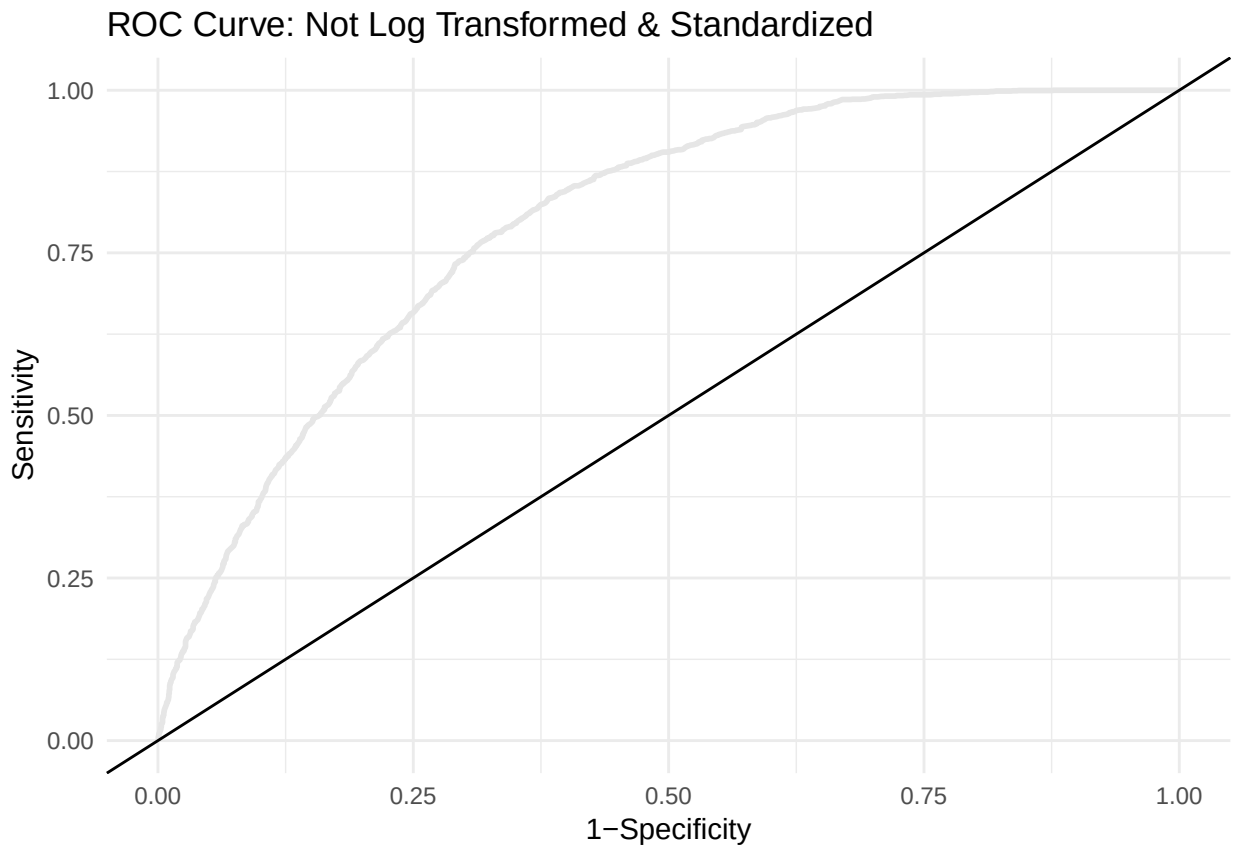
```
enetlst_nls <- readRDS("C:/Users/Spiro Stilianoudakis/Documents/TAD_data/RData/GM12878/comparing_normali

#Mean AUC across 100 bootstrap samples
enetlst_nls[[3]]
```

```
## [1] 0.8084435 0.8001046 0.7821470 0.7936183 0.7838993
```

```
auc.nls <- round(mean(enetlst_nls[[3]]),3)
auc.nls
```

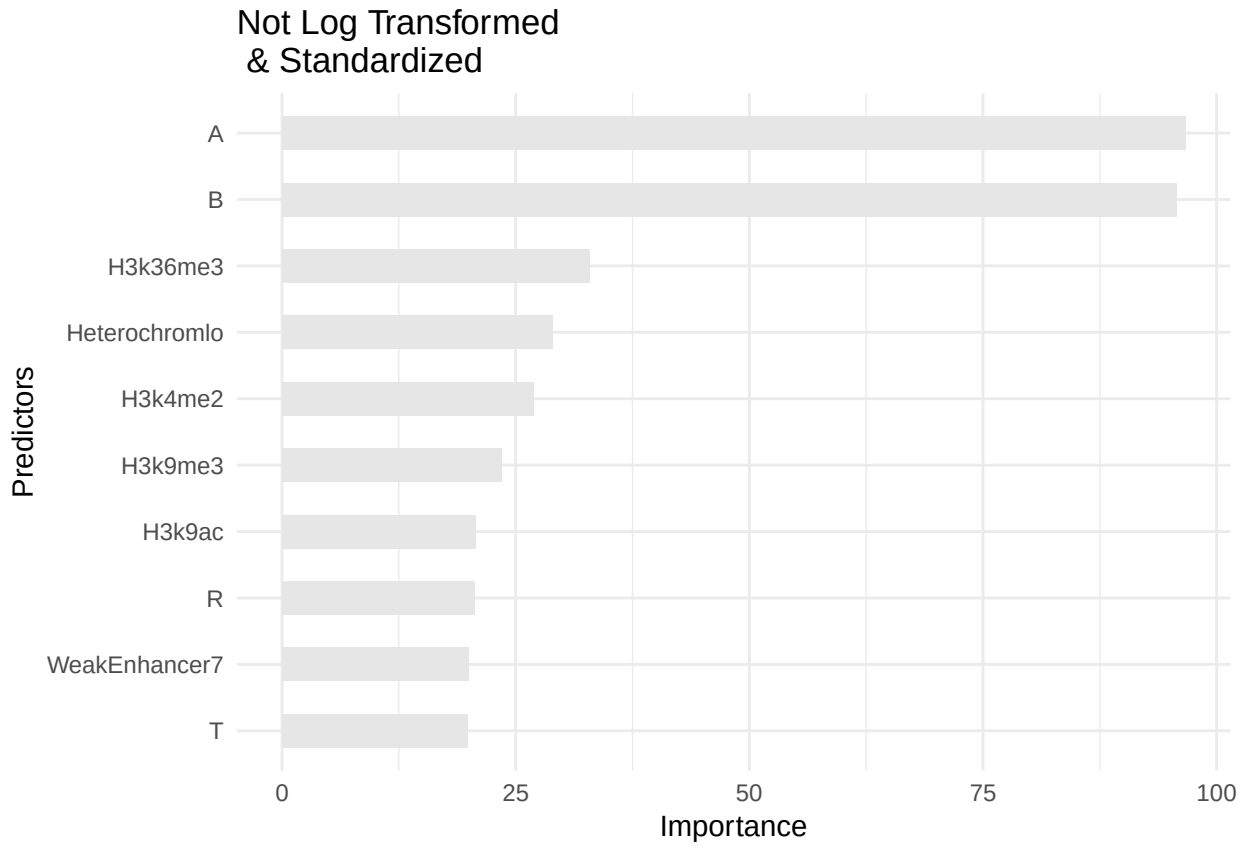
```
## [1] 0.794
```

```
#roc curve
fpr.nls <- rowMeans(enetlst_nls[[2]])
tpr.nls <- rowMeans(enetlst_nls[[1]])
rocdat.nls <- data.frame(fpr=fpr.nls, tpr=tpr.nls)
ggplot(rocdat.nls, aes(x=fpr.nls, y=tpr.nls)) +
  geom_line(size=1, color="#E6E6E6") +
  xlab("1-Specificity") +
  ylab("Sensitivity") +
  xlim(0, 1) +
  ylim(0, 1) +
  geom_abline(intercept=0, slope=1) +
  theme_minimal() +
  ggtitle("ROC Curve: Not Log Transformed & Standardized")
```




```
varimp.nls <- as.vector(rowMeans(enetlst_nls[[4]]))
Labels <- rownames(enetlst_nls[[4]])
Labels[grep("Gm12878_", Labels)] <- gsub("Gm12878_","",Labels[grep("Gm12878_", Labels)])
varimp.nls.df <- data.frame(Feature=Labels,
                            Importance=varimp.nls)
varimp.nls.df <- varimp.nls.df[order(varimp.nls.df$Importance),]
varimp.nls.df <- varimp.nls.df[(dim(varimp.nls.df)[1]-9):dim(varimp.nls.df)[1],]
varimp.nls.df$Feature <- factor(varimp.nls.df$Feature,
```

```
                                    levels=varimp.nls.df$Feature)
p.nls <- ggplot(varimp.nls.df, aes(x=Feature, y=Importance)) +
  xlab("Predictors") +
  ylab("Importance") +
  #ggtitle("Importance Plot for Gradient Boosting Machine") +
  geom_bar(stat="identity",
           width=.5,
           position="dodge",
           fill="#E6E6E6") +
  coord_flip() +
  theme_minimal() +
  ggtitle("Not Log Transformed \n & Standardized")
p.nls
```



Not Log tranformed and Un-Standardized

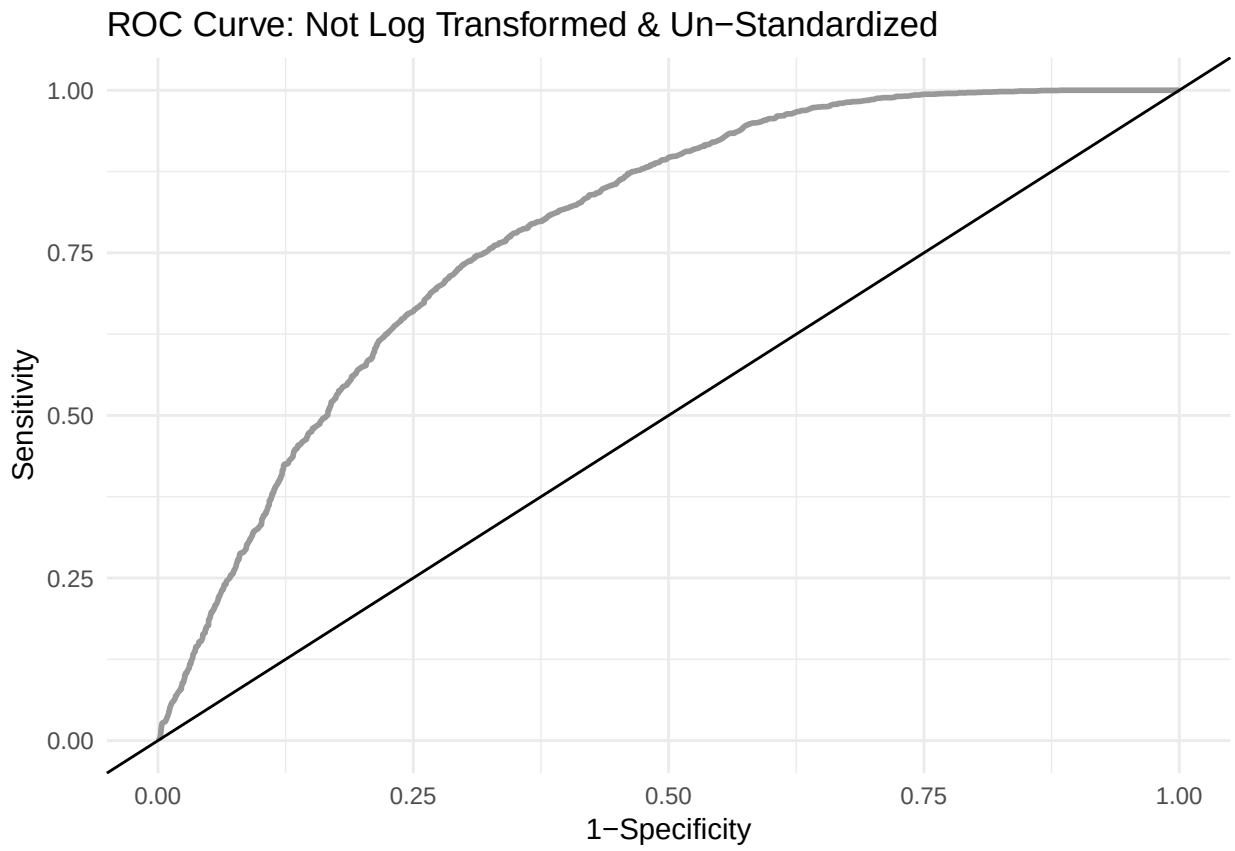
```
enetlst_nlns <- readRDS("C:/Users/Spiro Stilianoudakis/Documents/TAD_data/RData/GM12878/comparing_normal

#Mean AUC across 100 bootstrap samples
enetlst_nlns[[3]]
```

```
## [1] 0.7940741 0.7707093 0.7912596 0.7875418 0.7750335
```

```
auc.nlns <- round(mean(enetlst_nlns[[3]]),3)
auc.nlns
```

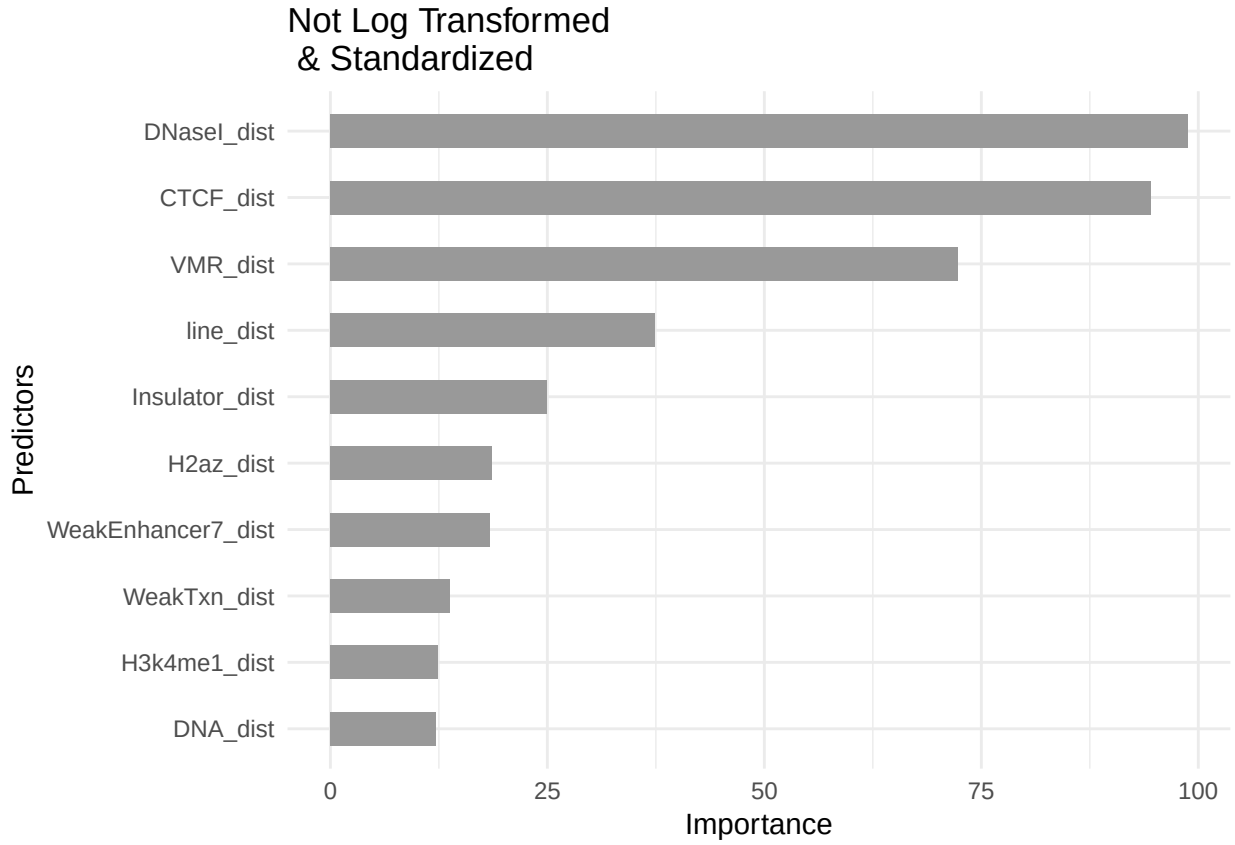
```
## [1] 0.784
```

```
#roc curve
fpr.nlns <- rowMeans(enetlst_nlns[[2]])
tpr.nlns <- rowMeans(enetlst_nlns[[1]])
rocdat.nlns <- data.frame(fpr=fpr.nlns, tpr=tpr.nlns)
ggplot(rocdat.nlns, aes(x=fpr.nlns, y=tpr.nlns)) +
  geom_line(size=1, color="#999999") +
  xlab("1-Specificity") +
  ylab("Sensitivity") +
  xlim(0, 1) +
  ylim(0, 1) +
  geom_abline(intercept=0, slope=1) +
  theme_minimal() +
  ggtitle("ROC Curve: Not Log Transformed & Un-Standardized")
```




```
varimp.nlns <- as.vector(rowMeans(enetlst_nlns[[4]]))
Labels <- rownames(enetlst_nlns[[4]])
Labels[grep("Gm12878_", Labels)] <- gsub("Gm12878_","",Labels[grep("Gm12878_", Labels)])
varimp.nlns.df <- data.frame(Feature=Labels,
                             Importance=varimp.nlns)
varimp.nlns.df <- varimp.nlns.df[order(varimp.nlns.df$Importance),]
varimp.nlns.df <- varimp.nlns.df[(dim(varimp.nlns.df)[1]-9):dim(varimp.nlns.df)[1],]
varimp.nlns.df$Feature <- factor(varimp.nlns.df$Feature,
```

```
                                    levels=varimp.nlns.df$Feature)
p.nlns <- ggplot(varimp.nlns.df, aes(x=Feature, y=Importance)) +
  xlab("Predictors") +
  ylab("Importance") +
  #ggtitle("Importance Plot for Gradient Boosting Machine") +
  geom_bar(stat="identity",
           width=.5,
           position="dodge",
           fill="#999999") +
  coord_flip() +
  theme_minimal() +
  ggtitle("Not Log Transformed \n & Standardized")
p.nlns
```



Comparing additional performance metrics across all normalization techniques

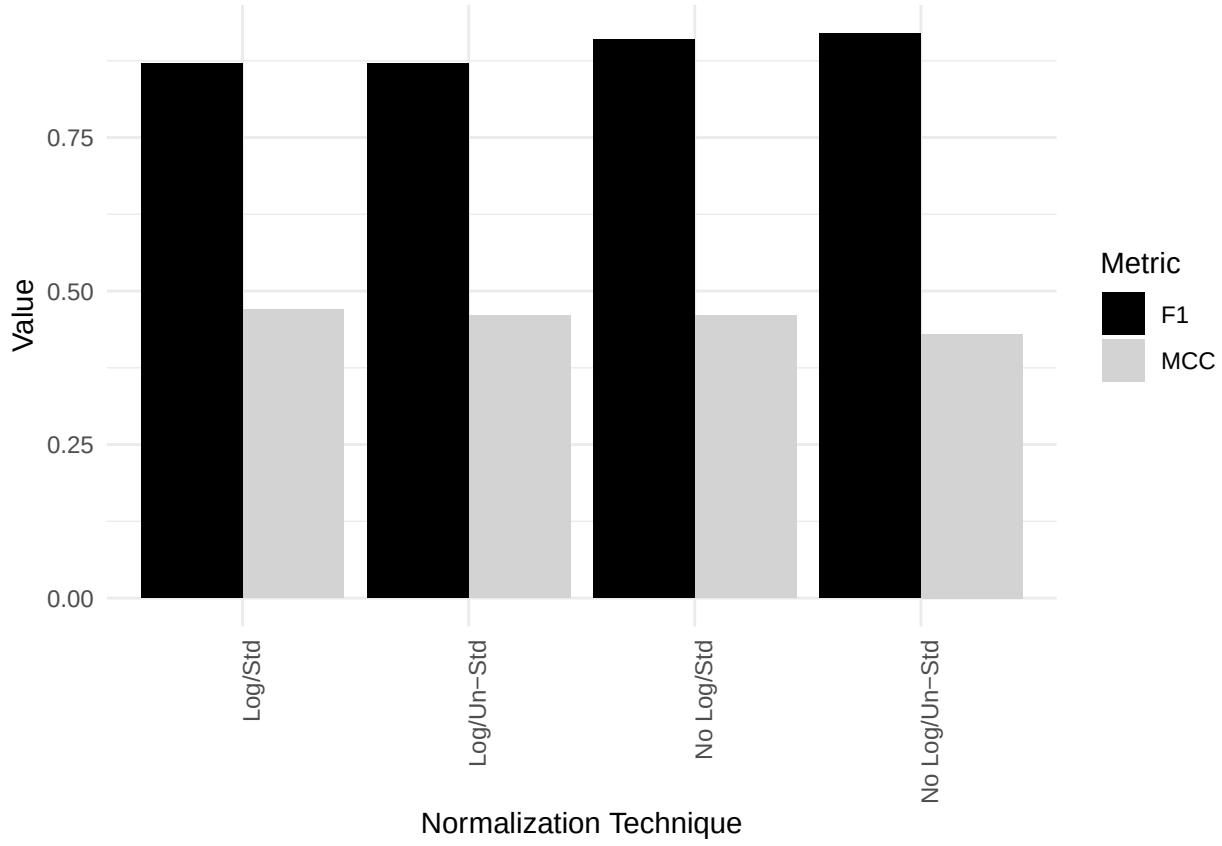
```
options(scipen = 999)

enetperf_ls <- readRDS("C:/Users/Spiro Stilianoudakis/Documents/TAD_data/RData/GM12878/comparing_normali

enetperf_lns <- readRDS("C:/Users/Spiro Stilianoudakis/Documents/TAD_data/RData/GM12878/comparing_normal
```

```
enetperf_nls <- readRDS("C:/Users/Spiro Stilianoudakis/Documents/TAD_data/RData/GM12878/comparing_normal

enetperf_nlns <- readRDS("C:/Users/Spiro Stilianoudakis/Documents/TAD_data/RData/GM12878/comparing_norma

lstab <- round(as.matrix(rowMeans(enetperf_ls)),2)
lnstab <- round(as.matrix(rowMeans(enetperf_lns)),2)
nlstab <- round(as.matrix(rowMeans(enetperf_nls)),2)
nlnstab <- round(as.matrix(rowMeans(enetperf_nlns)),2)

lstab[1:5,1] <- round(lstab[1:5,1],0)
lnstab[1:5,1] <- round(lnstab[1:5,1],0)
nlstab[1:5,1] <- round(nlstab[1:5,1],0)
nlnstab[1:5,1] <- round(nlnstab[1:5,1],0)

perfdat <- cbind.data.frame(rownames(enetperf_ls),
                            lstab,
                            lnstab,
                            nlstab,
                            nlnstab)
rownames(perfdat) <- NULL
colnames(perfdat) <- c("Metric", "Log/Std", "Log/Un-Std", "No Log/Std", "No Log/Un-Std")

kable(perfdat)
```

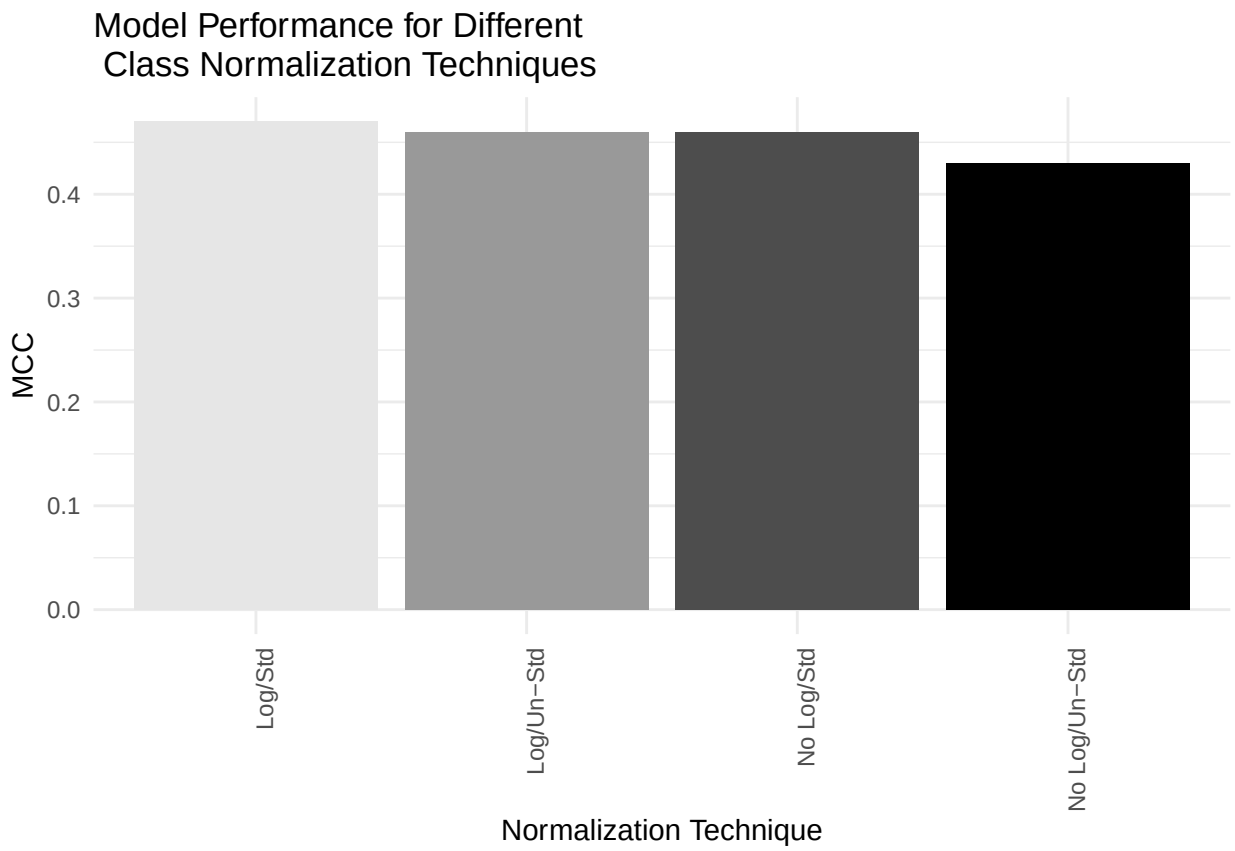
Metric	Log/Std	Log/Un-Std	No Log/Std	No Log/Un-Std
TN	335.00	333.00	301.00	278.00
FN	109.00	109.00	82.00	76.00
FP	153.00	155.00	187.00	210.00
TP	381.00	381.00	408.00	414.00
Total	978.00	978.00	978.00	978.00
Sensitivity	0.78	0.78	0.83	0.84
Specificity	0.69	0.68	0.62	0.57
Kappa	0.46	0.46	0.45	0.41
Accuracy	0.73	0.73	0.72	0.71
Precision	0.71	0.71	0.69	0.66
FPR	0.31	0.32	0.38	0.43
FNR	0.22	0.22	0.17	0.16
FOR	0.25	0.25	0.21	0.22
NPV	0.75	0.75	0.79	0.78
MCC	0.47	0.46	0.46	0.43
F1	0.87	0.87	0.91	0.92

```
mccf1 <- data.frame(Metric = c(rep("MCC",4), rep("F1",4)),
                    Technique = rep(c("Log/Std",
                                      "Log/Un-Std",
                                      "No Log/Std",
                                      "No Log/Un-Std"), 2),
                    Value = c(as.numeric(perfdat[15,2:5]), as.numeric(perfdat[16,2:5])))

ggplot(data=mccf1, aes(x=Technique, y=Value, fill=Metric)) +
  geom_bar(stat="identity", position=position_dodge()) +
  scale_fill_manual(values=c('black','lightgray')) +
```

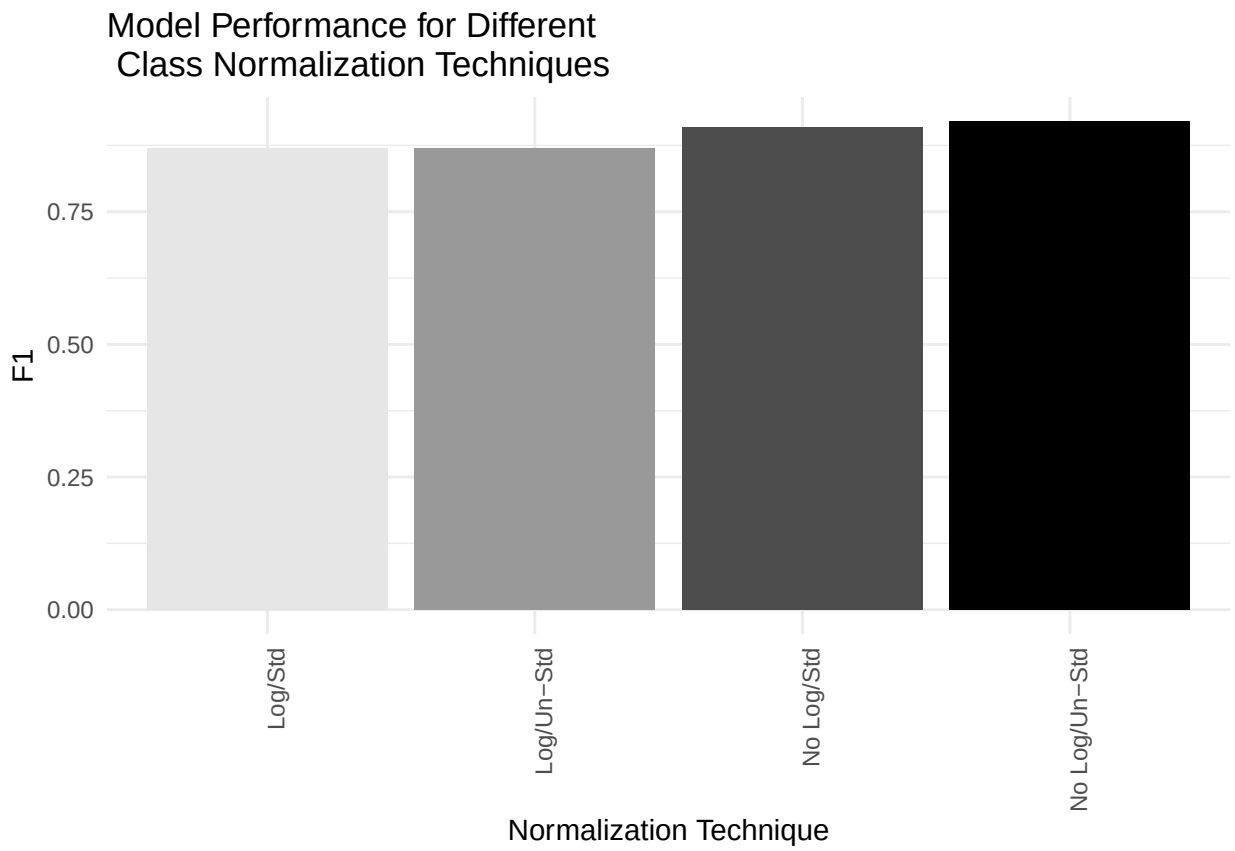
```
  xlab("Normalization Technique") +
  theme_minimal() +
  theme(axis.text.x = element_text(angle = 90, hjust = 1))
```




```
MCCplot<-ggplot(data=mccf1[1:4,], aes(x=Technique, y=Value, fill=Technique)) +
  xlab("Normalization Technique") + ylab("MCC") +
  geom_bar(stat="identity") +
  scale_fill_manual(values=gray(rev(c(0,.3,.6,.9))), guide=FALSE) +
  theme_minimal() +
  theme(axis.text.x = element_text(angle = 90, hjust = 1)) +
  ggtitle("Model Performance for Different \n Class Normalization Techniques")
MCCplot
```



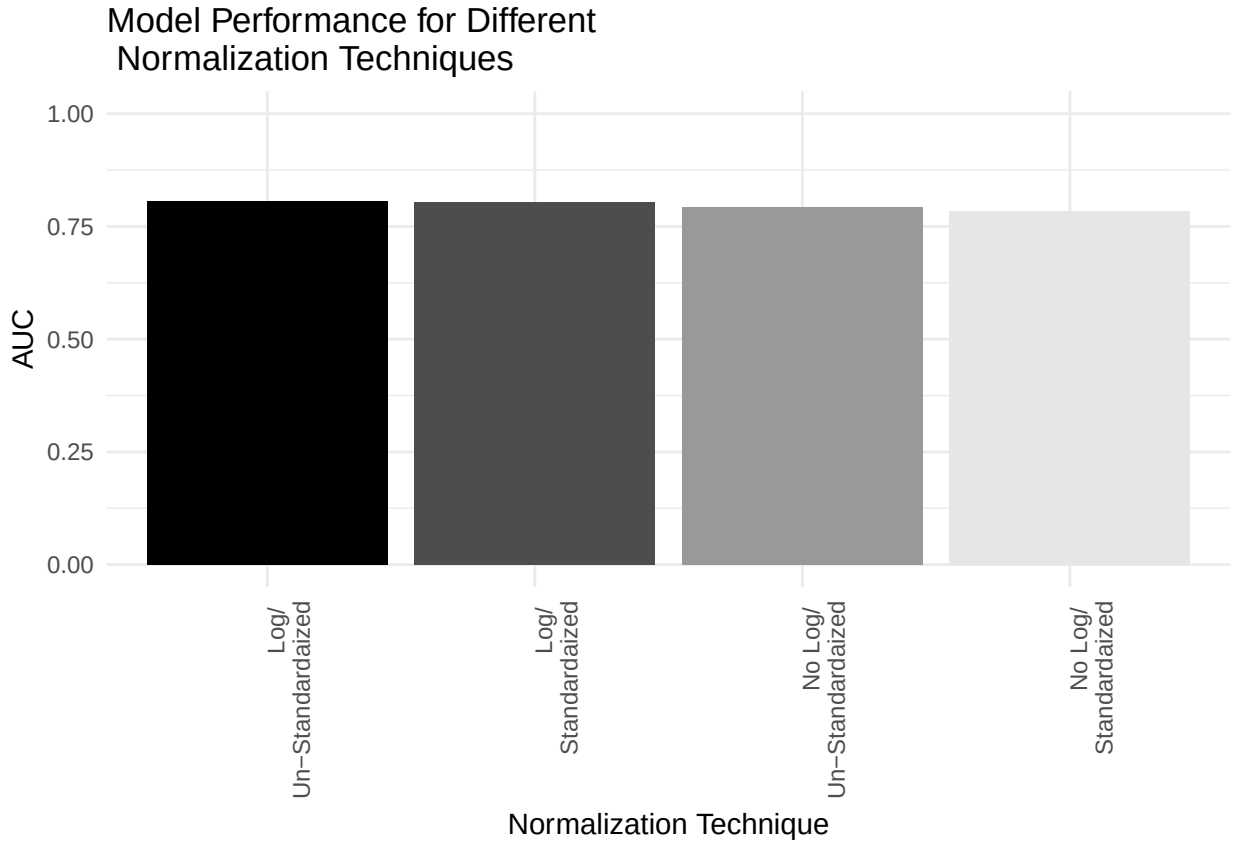

```
F1plot<-ggplot(data=mccf1[5:8,], aes(x=Technique, y=Value, fill=Technique)) +
  xlab("Normalization Technique") + ylab("F1") +
  geom_bar(stat="identity") +
  scale_fill_manual(values=gray(rev(c(0,.3,.6,.9))), guide=FALSE) +
  theme_minimal() +
  theme(axis.text.x = element_text(angle = 90, hjust = 1)) +
  ggtitle("Model Performance for Different \n Class Normalization Techniques")
F1plot
```



Comparing Models

```
auc.plot <- data.frame("Normalization Technique"=c("Log/Standardaized",
                                        "Log/Un-Standardaized",
                                        "No Log/Standardaized",
                                        "No Log/Un-Standardaized"),
                       auc=c(auc.ls,
                             auc.lns,
                             auc.nls,
                             auc.nlns))

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

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

p<-ggplot(data=auc.plot, aes(x=Normalization.Technique, y=auc, fill=Normalization.Technique)) +
  xlab("Normalization Technique") + ylab("AUC") +
  geom_bar(stat="identity") + ylim(0,1) +
  scale_fill_manual(values=grey(c(0,.3,.6,.9)), guide=FALSE) +
  scale_x_discrete(labels= c("Log/ \n Un-Standardaized",
              "Log/ \n Standardaized",
              "No Log/ \n Un-Standardaized",
```

```
              "No Log/ \n Standardaized")) +
  theme_minimal() +
  theme(axis.text.x = element_text(angle = 90, hjust = 1)) +
  ggtitle("Model Performance for Different \n Normalization Techniques")
p
```

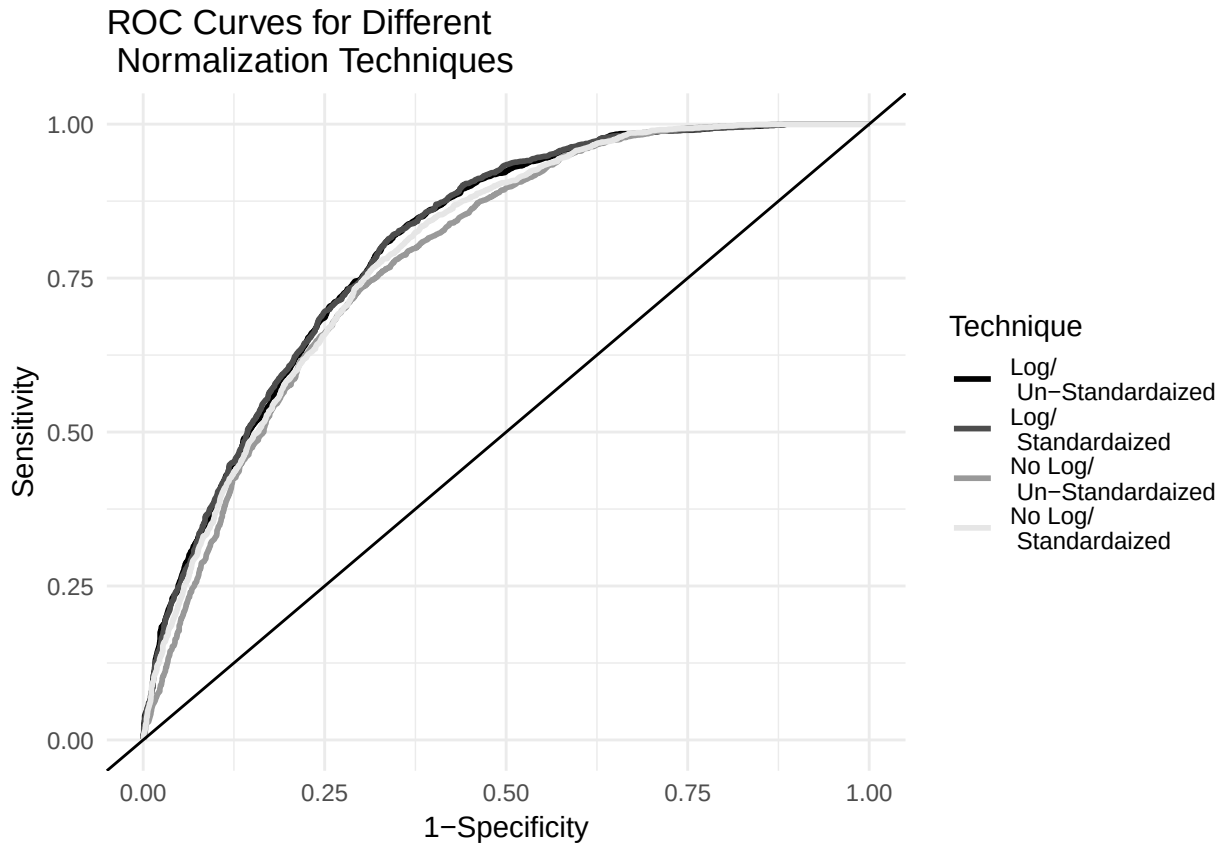



```
#datatable(auc.plot)
kable(auc.plot)
```

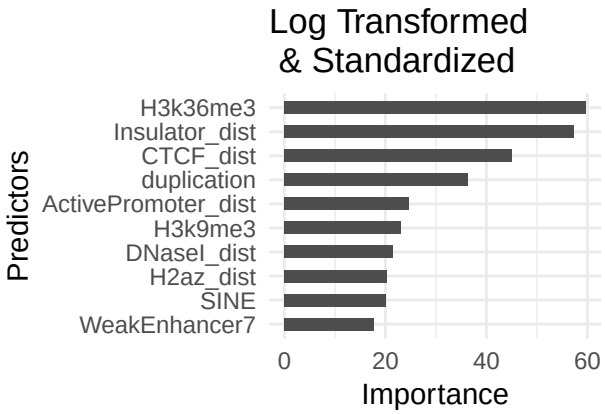
Normalization.Technique	auc
Log/Standardaized	0.806
Log/Un-Standardaized	0.805
No Log/Standardaized	0.794
No Log/Un-Standardaized	0.784

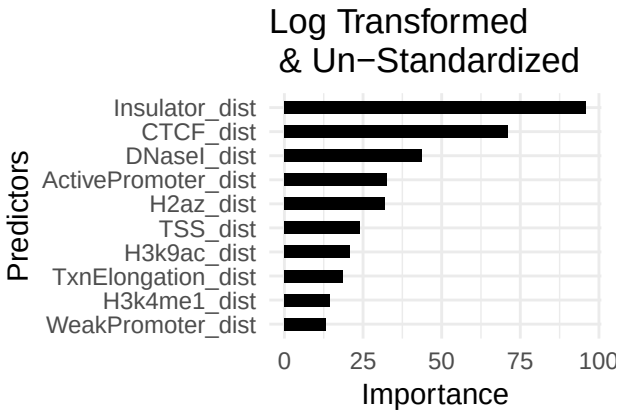
```
rocdat.ls$Technique <- "ls"
rocdat.lns$Technique <- "lns"
rocdat.nls$Technique <- "nls"
rocdat.nlns$Technique <- "nlns"
allrocdat <- rbind.data.frame(rocdat.ls, rocdat.lns, rocdat.nls, rocdat.nlns)

ggplot(data=allrocdat, aes(x=fpr, y=tpr, color=Technique)) +
  geom_line(size=1) +
  scale_colour_manual(name="Technique",
```

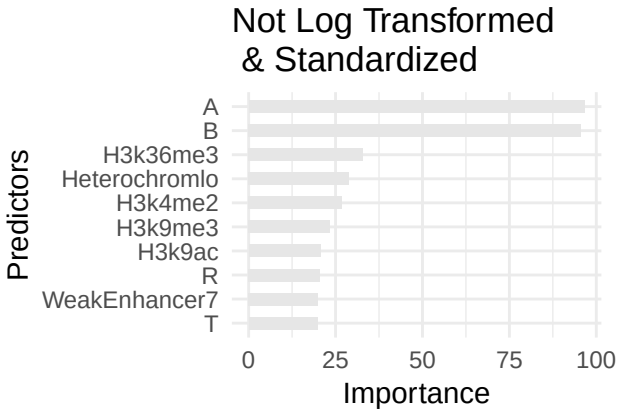
```
    labels=c("Log/ \n Un-Standardaized",
             "Log/ \n Standardaized",
             "No Log/ \n Un-Standardaized",
              "No Log/ \n Standardaized"),
    values=grey(c(0,.3,.6,.9))) +
  xlab("1-Specificity") +
  ylab("Sensitivity") +
  xlim(0, 1) +
  ylim(0, 1) +
  geom_abline(intercept=0, slope=1) +
  theme_minimal() +
  ggtitle("ROC Curves for Different \n Normalization Techniques")
```

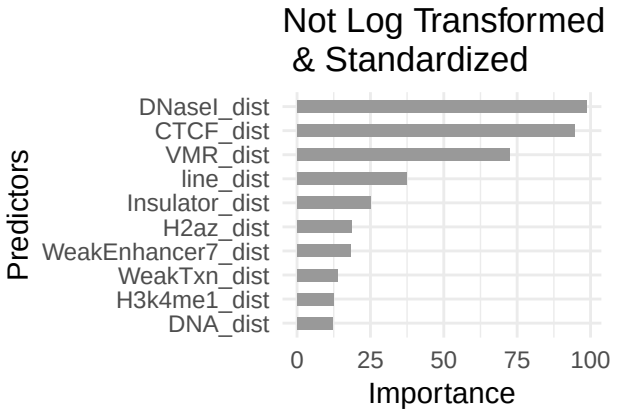



```
grid.arrange(p.ls,p.lns,p.nls,p.nlns,ncol=2)
```


Log Transformed
& Standardized
Predictors
H3k36me3
Insulator_dist
CTCF_dist
duplication
ActivePromoter_dist
H3k9me3
DNaseI_dist
H2az_dist
SINE
WeakEnhancer7
0
20
40
60
Importance


Log Transformed
& Un−Standardized
Predictors
Insulator_dist
CTCF_dist
DNaseI_dist
ActivePromoter_dist
H2az_dist
TSS_dist
H3k9ac_dist
TxnElongation_dist
H3k4me1_dist
WeakPromoter_dist
0
25
50
75
100
Importance


Not Log Transformed
& Standardized
Predictors
A
B
H3k36me3
Heterochromlo
H3k4me2
H3k9me3
H3k9ac
R
WeakEnhancer7
T
0
25
50
75
100
Importance


Not Log Transformed
& Standardized
Predictors
DNaseI_dist
CTCF_dist
VMR_dist
line_dist
Insulator_dist
H2az_dist
WeakEnhancer7_dist
WeakTxn_dist
H3k4me1_dist
DNA_dist
0
25
50
75
100
Importance