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]]
```

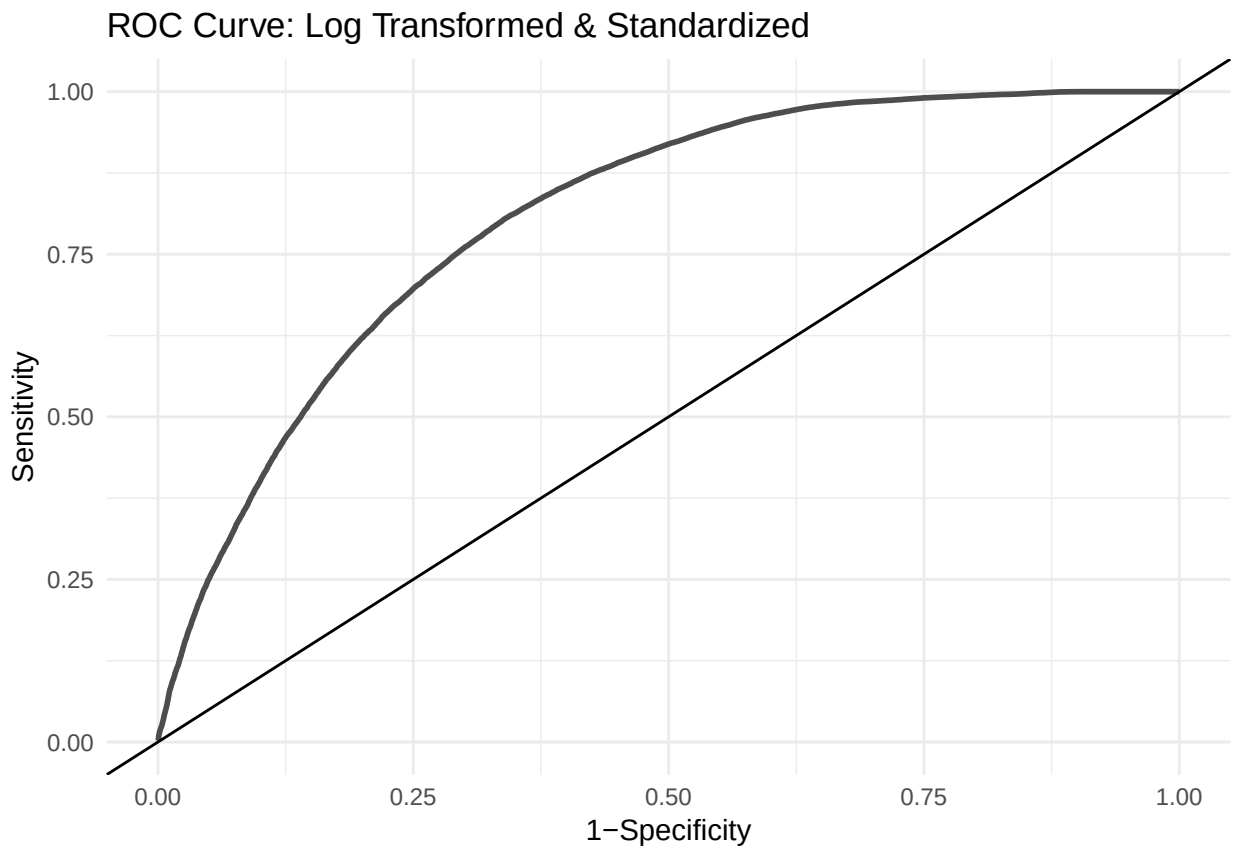
```
##   [1] 0.8156114 0.8038223 0.8070759 0.8127467 0.7929366 0.7925100 0.8066452
##   [8] 0.7845684 0.7922675 0.8042071 0.8146077 0.8074649 0.7965875 0.7863249
```

```
##  [15] 0.8045166 0.8150928 0.8010162 0.8185263 0.7745274 0.7991176 0.7951782
##  [22] 0.8109527 0.7999331 0.7928613 0.8322725 0.8254935 0.8135455 0.8034920
##  [29] 0.8259493 0.8159627 0.7788809 0.8074983 0.8106599 0.8062019 0.8169455
##  [36] 0.8225410 0.8022834 0.7901012 0.8191410 0.8081800 0.8009242 0.7865214
##  [43] 0.8324105 0.8110865 0.7958013 0.7961526 0.7953496 0.8163558 0.8056081
##  [50] 0.7966209 0.7863541 0.8021454 0.8087153 0.8310472 0.8273210 0.7970559
##  [57] 0.8277601 0.7932628 0.7872951 0.8035212 0.7990423 0.7992891 0.8127969
##  [64] 0.7978463 0.8037053 0.7897290 0.8204918 0.8221019 0.8151263 0.8207176
##  [71] 0.8023879 0.7956465 0.8083055 0.7761208 0.7979508 0.8057252 0.8195550
##  [78] 0.8089997 0.7978463 0.7921880 0.8306122 0.7973068 0.7979090 0.8083138
##  [85] 0.8049306 0.8021203 0.8133406 0.8242765 0.8330085 0.8007988 0.8149757
##  [92] 0.8048595 0.8048428 0.7824063 0.8140432 0.8098486 0.7980428 0.8213031
##  [99] 0.8091293 0.8137546
```

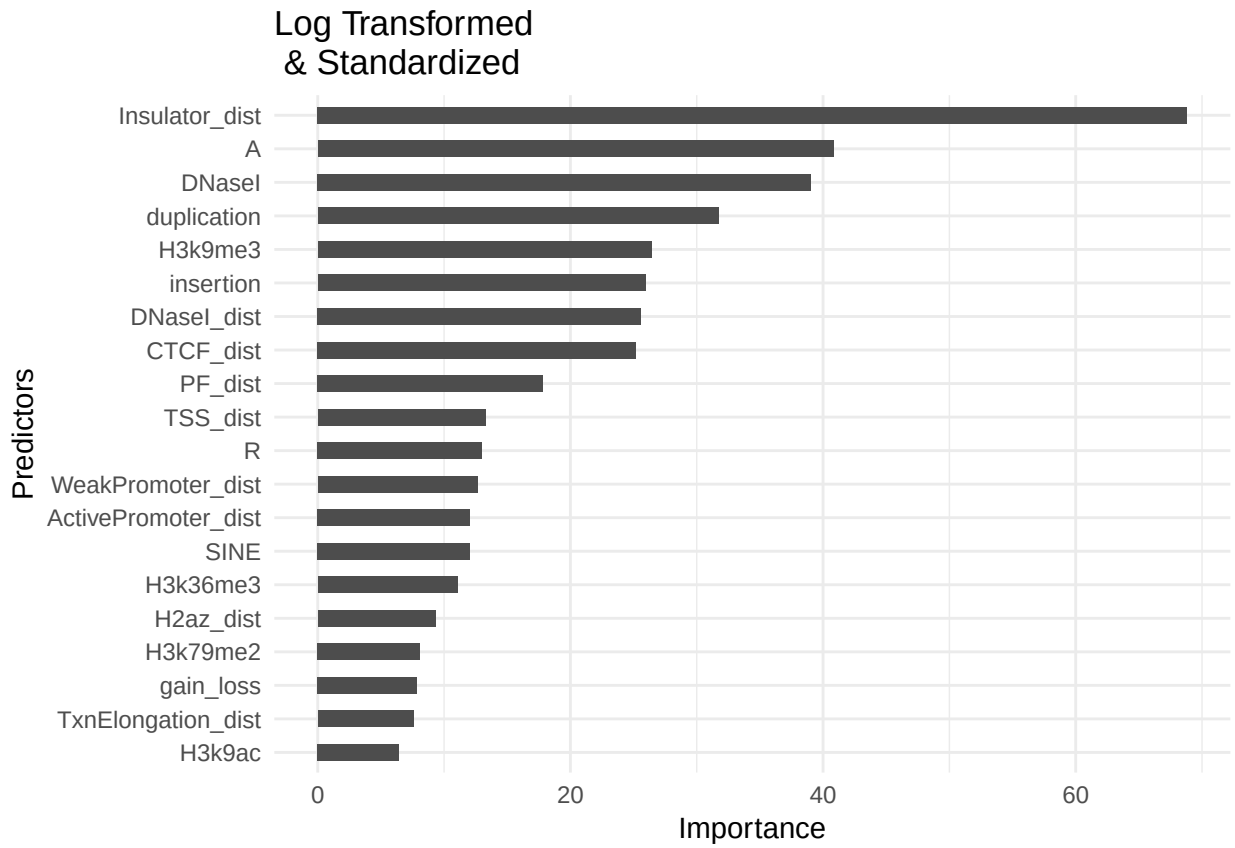
```
auc.ls <- round(mean(enetlst_ls[[3]]),3)
auc.ls
```

```
## [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$Importance),]
varimp.ls.df <- varimp.ls.df[(dim(varimp.ls.df)[1]-19):dim(varimp.ls.df)[1],]
varimp.ls.df$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

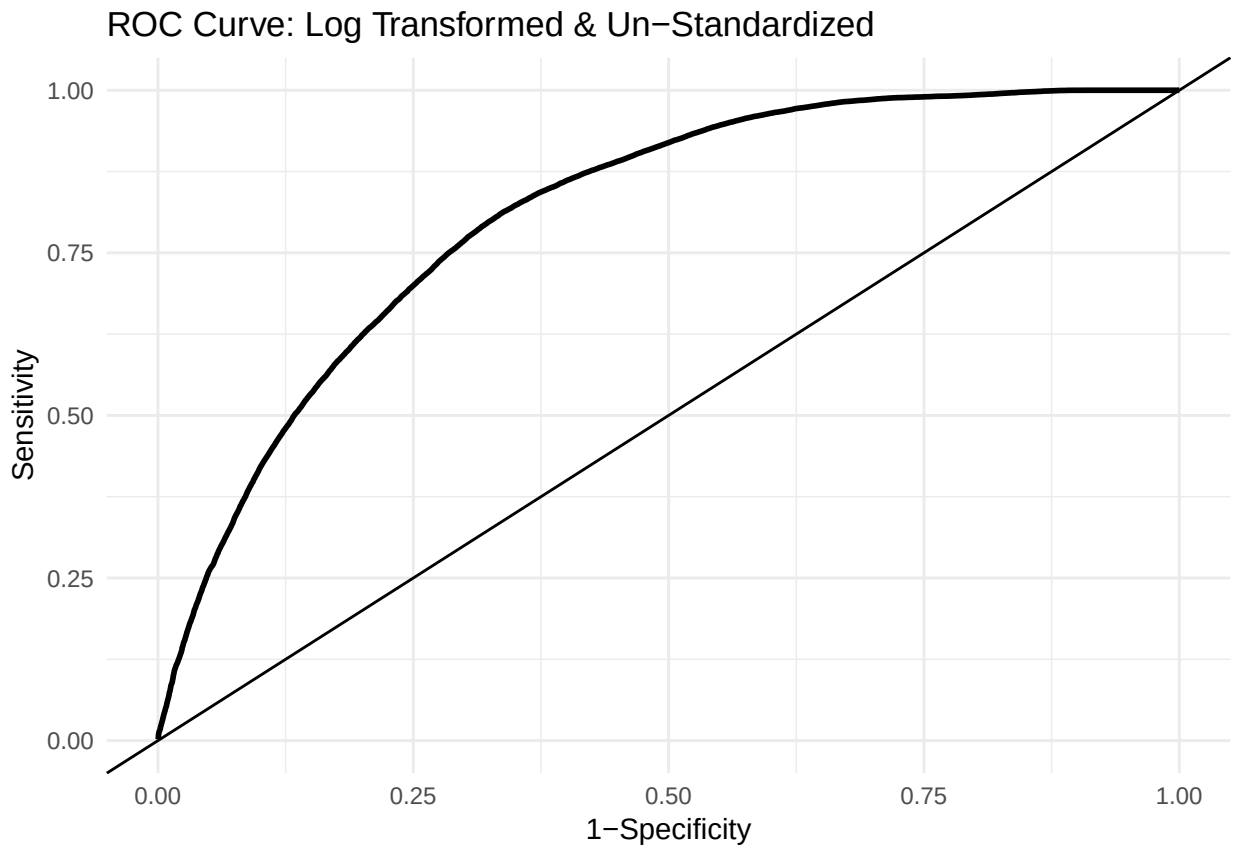
```
enetlst_lns <- readRDS("C:/Users/Spiro Stilianoudakis/Documents/TAD_data/RData/GM12878/comparing_normali

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

```
##   [1] 0.8178990 0.8037889 0.8082051 0.8145617 0.7967255 0.7963031 0.8069379
##   [8] 0.7884660 0.7931457 0.8075987 0.8188859 0.8066243 0.7964453 0.7947809
##  [15] 0.8081842 0.8216209 0.8056332 0.8248662 0.7772499 0.8016895 0.7996822
##  [22] 0.8139302 0.8035714 0.7896495 0.8346144 0.8260999 0.8171295 0.8032034
##  [29] 0.8277643 0.8142732 0.7823310 0.8095642 0.8184384 0.8112036 0.8189696
##  [36] 0.8305997 0.8044204 0.7910547 0.8203036 0.8082302 0.8016728 0.7911132
##  [43] 0.8332929 0.8131608 0.8021119 0.7978128 0.8007862 0.8264386 0.8119982
##  [50] 0.7998453 0.7913642 0.8055830 0.8084727 0.8333431 0.8341502 0.8001756
##  [57] 0.8320132 0.7959225 0.7894781 0.8086986 0.7999875 0.8010915 0.8203120
##  [64] 0.7969430 0.8049222 0.7933255 0.8216251 0.8245651 0.8163934 0.8235865
##  [71] 0.8051062 0.7971270 0.8107394 0.7769823 0.7978212 0.8132988 0.8222064
##  [78] 0.8113625 0.7988709 0.7956340 0.8339453 0.7983690 0.8020994 0.8151054
##  [85] 0.8183883 0.8063399 0.8142857 0.8291736 0.8347148 0.8022708 0.8177191
##  [92] 0.8102919 0.8128889 0.7857687 0.8208431 0.8158414 0.8010497 0.8257653
##  [99] 0.8115632 0.8185388
```

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

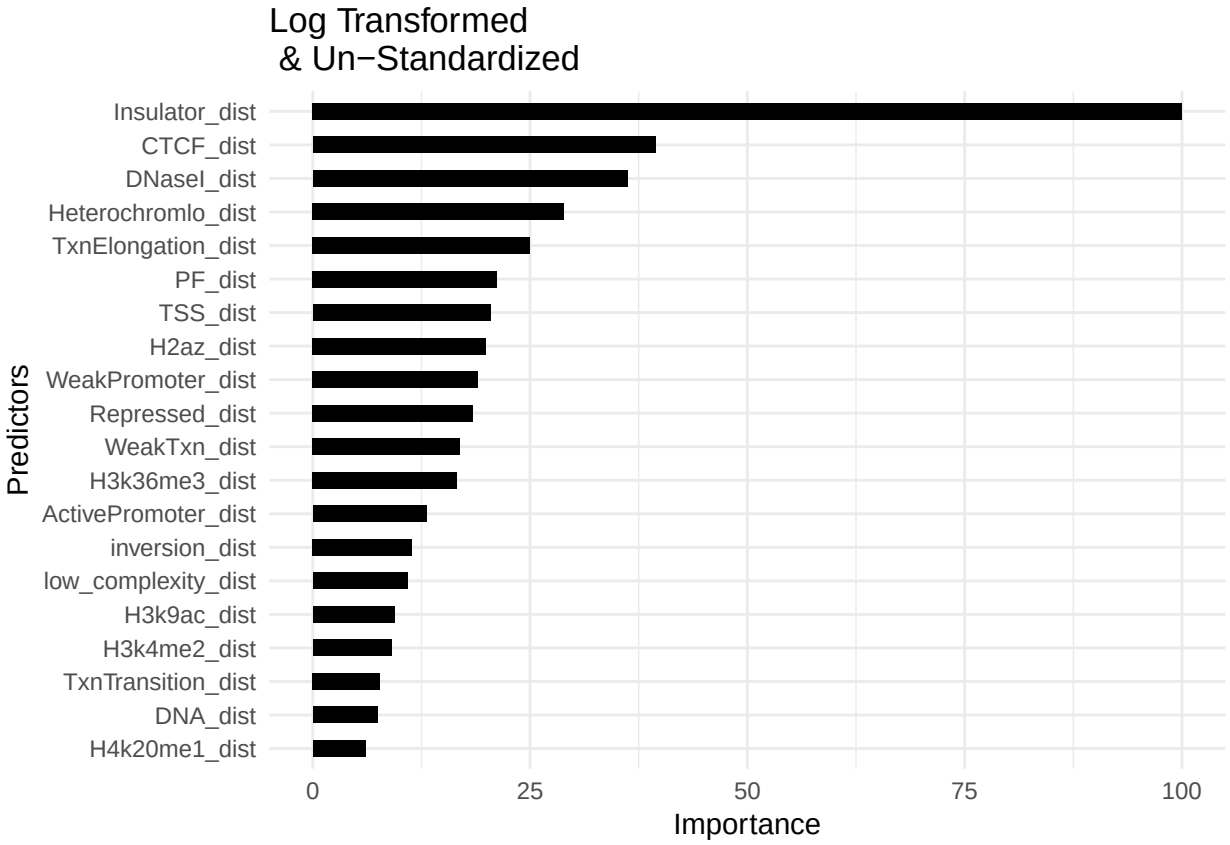
```
## [1] 0.809
```

```
#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$Importance),]
varimp.lns.df <- varimp.lns.df[(dim(varimp.lns.df)[1]-19):dim(varimp.lns.df)[1],]
varimp.lns.df$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]]
```

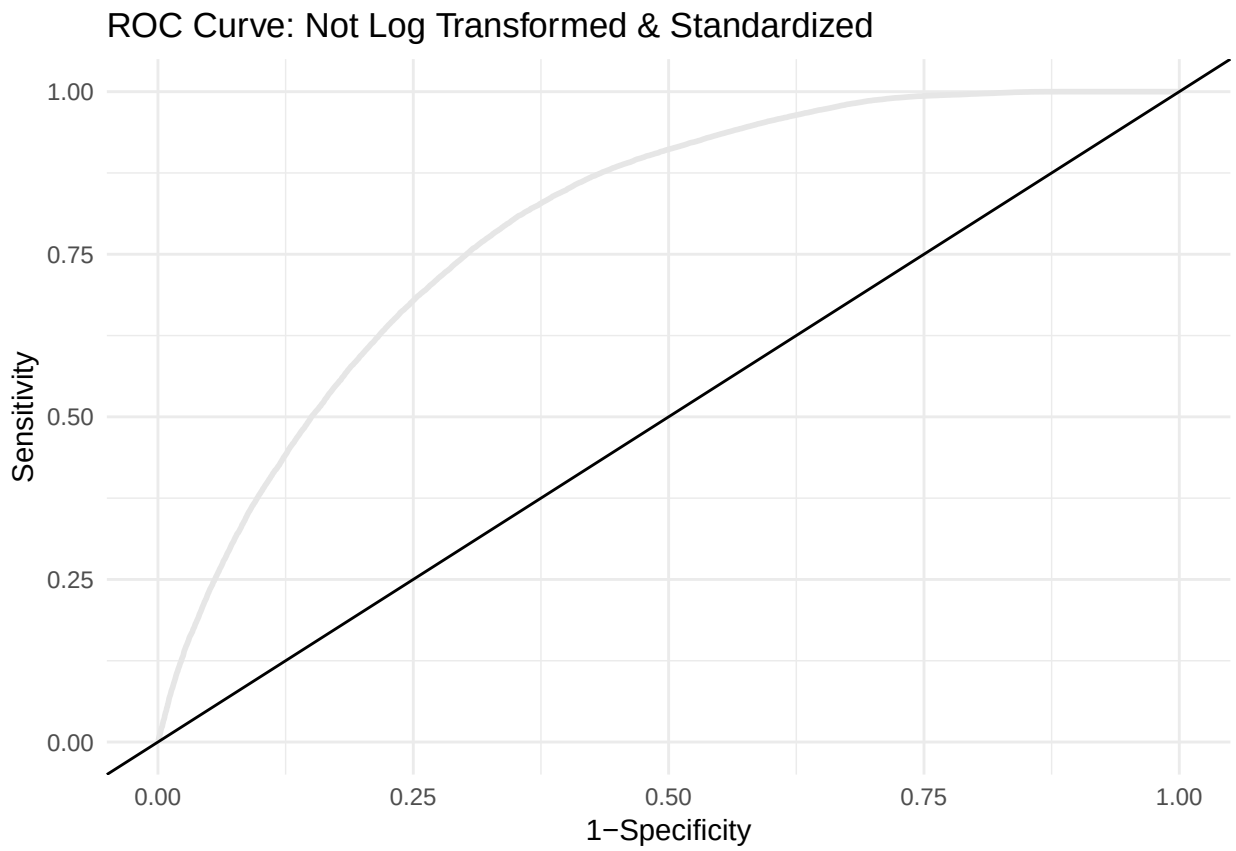
```
##   [1] 0.8111450 0.8066201 0.7845099 0.7946345 0.7872324 0.7747449 0.7904065
##   [8] 0.7863625 0.7799306 0.7993769 0.8043200 0.8093091 0.7882695 0.7756733
##  [15] 0.7935681 0.8058883 0.7966502 0.8228755 0.7758448 0.7946387 0.7938525
```

```
##  [22] 0.7972232 0.7803906 0.7878262 0.8292071 0.8171420 0.8057084 0.7836986
##  [29] 0.8114838 0.7980596 0.7740507 0.8011877 0.7989545 0.7947767 0.8018484
##  [36] 0.8272959 0.8075025 0.7874833 0.8053906 0.7978003 0.7745567 0.7832929
##  [43] 0.8199440 0.8176230 0.7919873 0.7995818 0.7817247 0.8042656 0.7906741
##  [50] 0.7769028 0.7708264 0.7974573 0.8072558 0.8237579 0.8256440 0.7882277
##  [57] 0.8198561 0.7858188 0.7936726 0.8046546 0.7838784 0.7836358 0.8112956
##  [64] 0.7883406 0.7946136 0.7967046 0.8071303 0.8099072 0.8024925 0.8194045
##  [71] 0.7871654 0.7822432 0.8029065 0.7763801 0.7935723 0.7963031 0.8085647
##  [78] 0.8128680 0.7788600 0.7852794 0.8077576 0.7919957 0.7925979 0.8050226
##  [85] 0.7848235 0.7906030 0.8000962 0.8162429 0.8227208 0.7860865 0.8046713
##  [92] 0.7969137 0.8093510 0.7780654 0.8008406 0.8004684 0.7989252 0.8087195
##  [99] 0.8016017 0.7850953
```

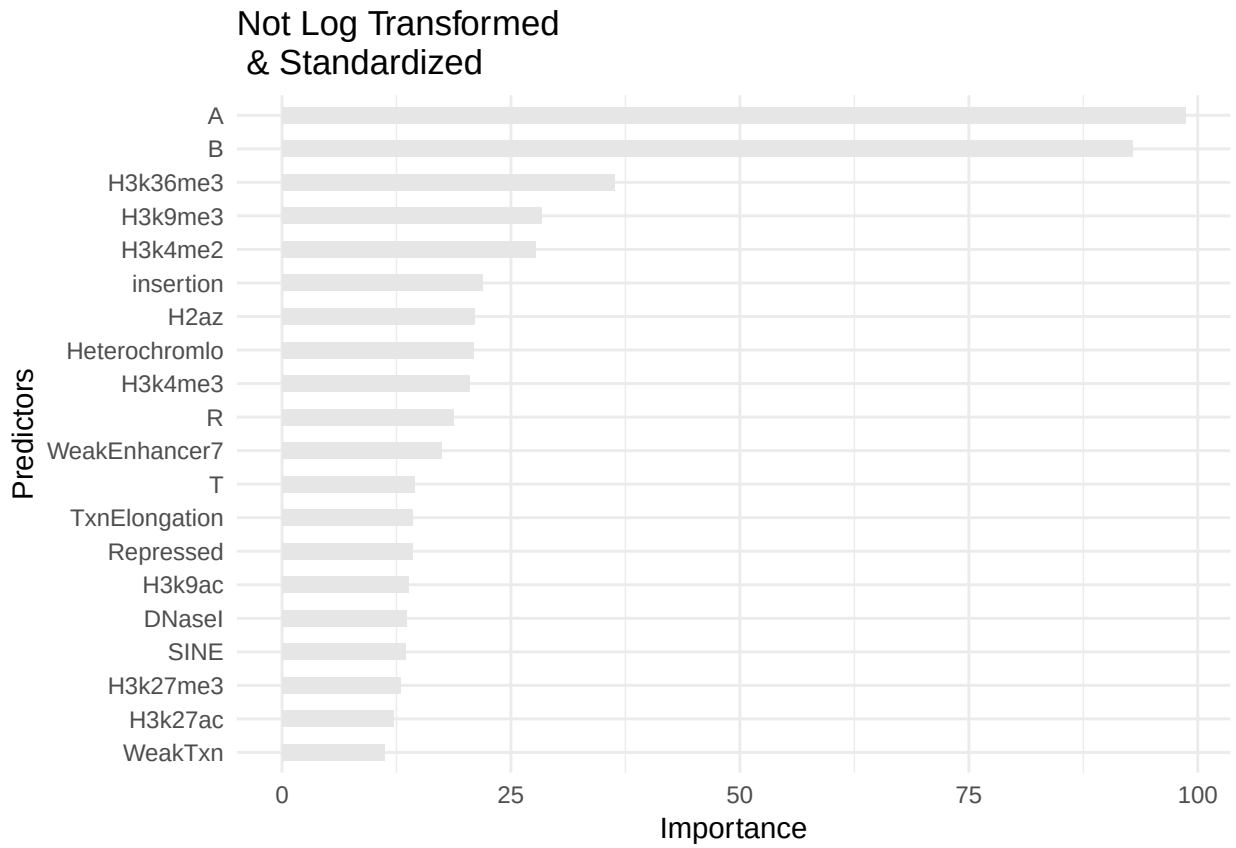
```
auc.nls <- round(mean(enetlst_nls[[3]]),3)
auc.nls
```

```
## [1] 0.797
```

```
#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$Importance),]
varimp.nls.df <- varimp.nls.df[(dim(varimp.nls.df)[1]-19):dim(varimp.nls.df)[1],]
varimp.nls.df$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

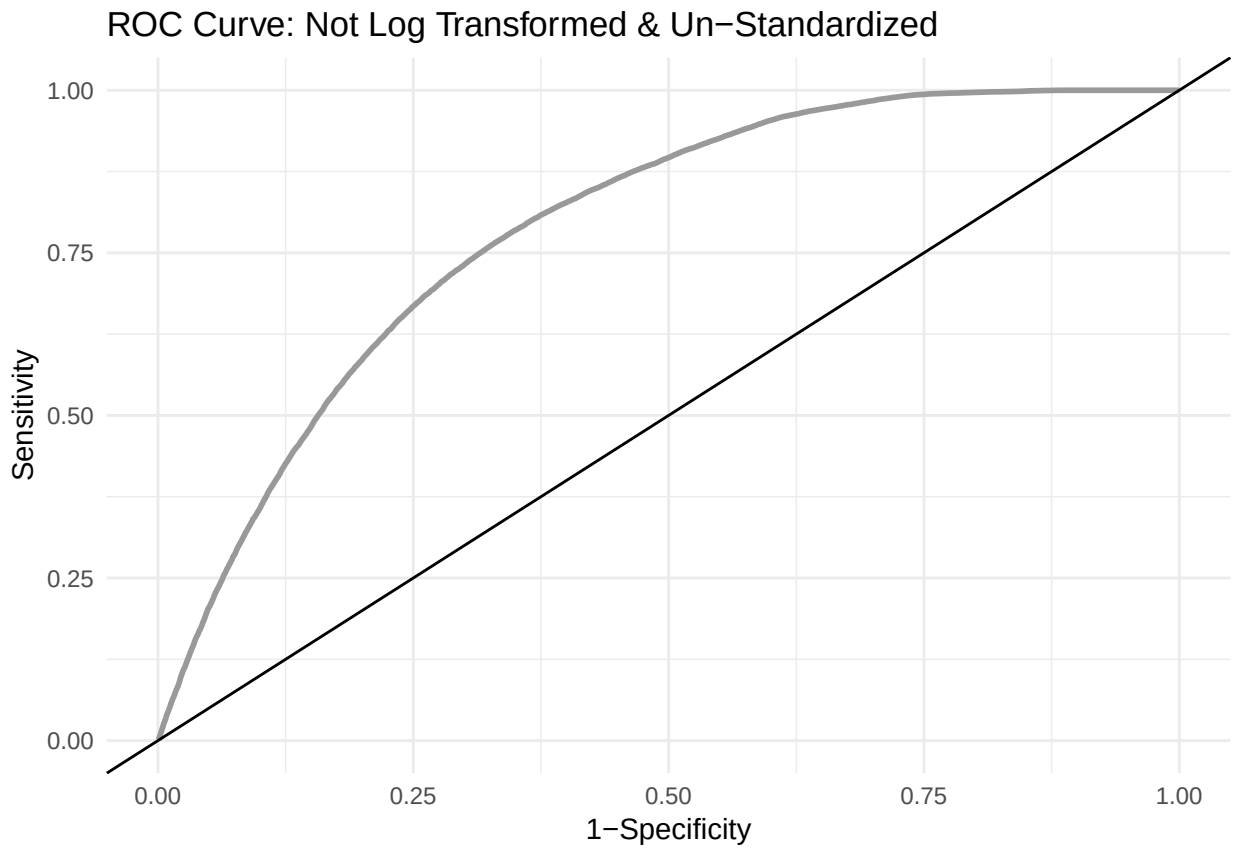
```
enetlst_nlns <- readRDS("C:/Users/Spiro Stilianoudakis/Documents/TAD_data/RData/GM12878/comparing_normal

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

```
##   [1] 0.7986994 0.7787596 0.7937521 0.7895408 0.7808381 0.7500585 0.7766728
##   [8] 0.7701572 0.7644614 0.7897792 0.7813148 0.7954040 0.7779316 0.7691327
##  [15] 0.7891895 0.7938734 0.7951698 0.8094973 0.7717380 0.7804282 0.7919329
##  [22] 0.7887504 0.7655110 0.7810346 0.8149423 0.7938106 0.7930830 0.7800769
##  [29] 0.7986576 0.7913558 0.7699481 0.7951112 0.7891561 0.7786258 0.7987203
##  [36] 0.8086609 0.7977250 0.7874080 0.7983147 0.7973319 0.7802400 0.7714829
##  [43] 0.8126296 0.8040231 0.7852584 0.8003722 0.7825109 0.7914604 0.7913851
##  [50] 0.7755018 0.7776138 0.7857143 0.7909334 0.8157452 0.8164604 0.7828663
##  [57] 0.8087655 0.7743602 0.7833222 0.7878931 0.7655989 0.7629391 0.7931582
##  [64] 0.7815741 0.7810054 0.7872658 0.7939696 0.7970266 0.7960898 0.8071136
##  [71] 0.7806792 0.7841753 0.7883071 0.7655069 0.7748285 0.7896161 0.8080420
##  [78] 0.7963449 0.7716711 0.7690365 0.8011250 0.7872365 0.7801062 0.7926815
##  [85] 0.7838909 0.7877300 0.7828413 0.8075025 0.7960815 0.7771830 0.7779734
##  [92] 0.7717506 0.8050100 0.7736283 0.7929450 0.7937437 0.7926899 0.7924557
##  [99] 0.7949649 0.7751673
```

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

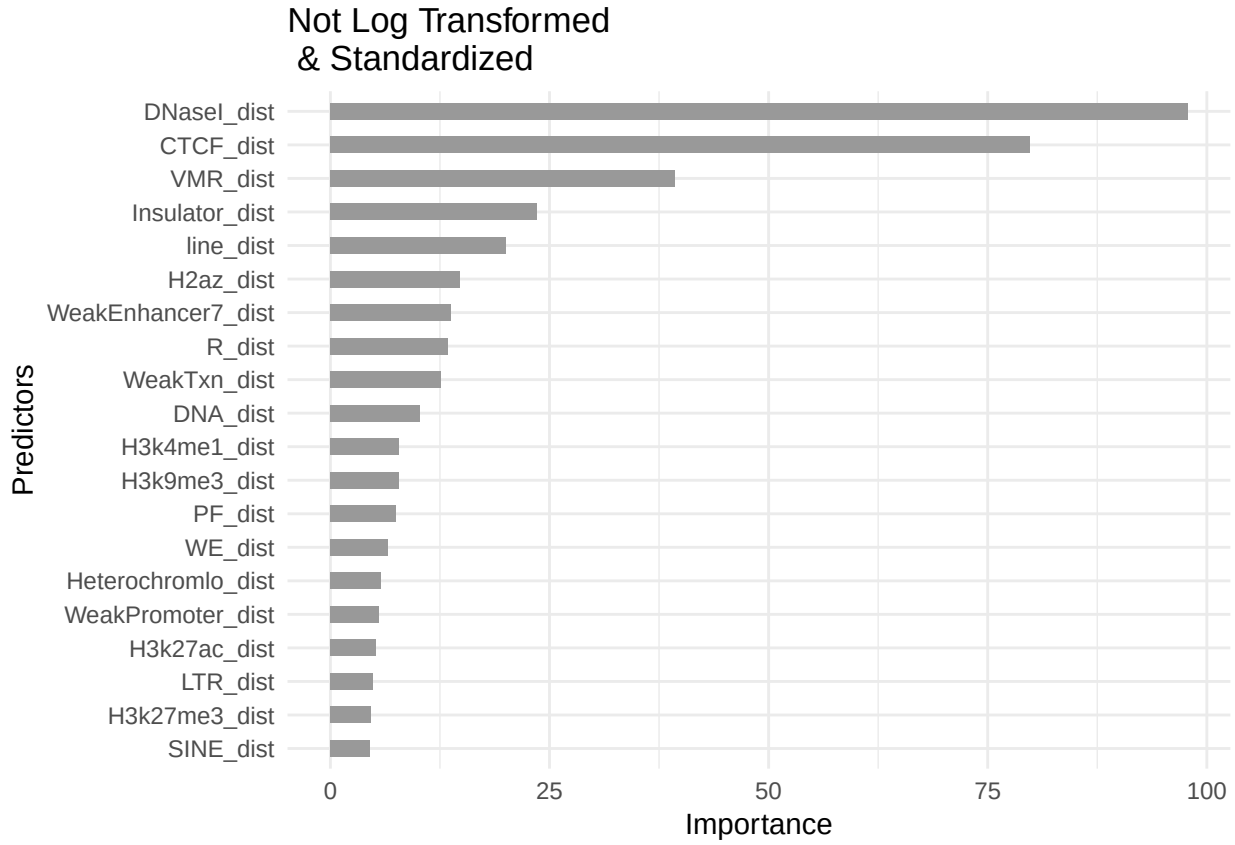
```
## [1] 0.788
```

```
#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$Importance),]
varimp.nlns.df <- varimp.nlns.df[(dim(varimp.nlns.df)[1]-19):dim(varimp.nlns.df)[1],]
varimp.nlns.df$Feature <- factor(varimp.nlns.df$Feature,
```

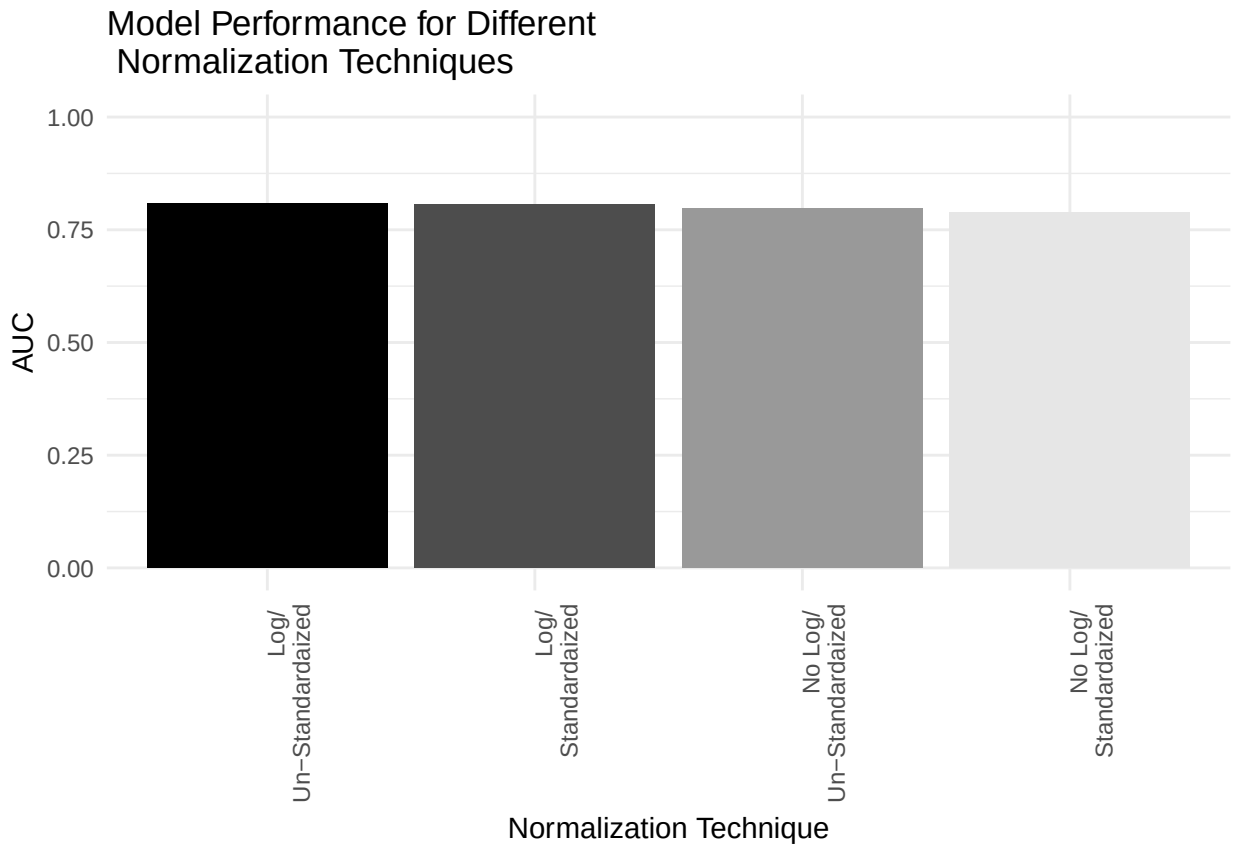
```
                                    levels=varimp.nlns.df$Feature)
p.nlns <- ggplot(varimp.nlns.df, aes(x=Feature, y=Importance)) +
  xlab("Predictors") +
  ylab("Importance") +
  #ggtitle("Importance Plot for Gradient Boosting Machine") +
  geom_bar(stat="identity",
           width=.5,
       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



Comparing 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
```

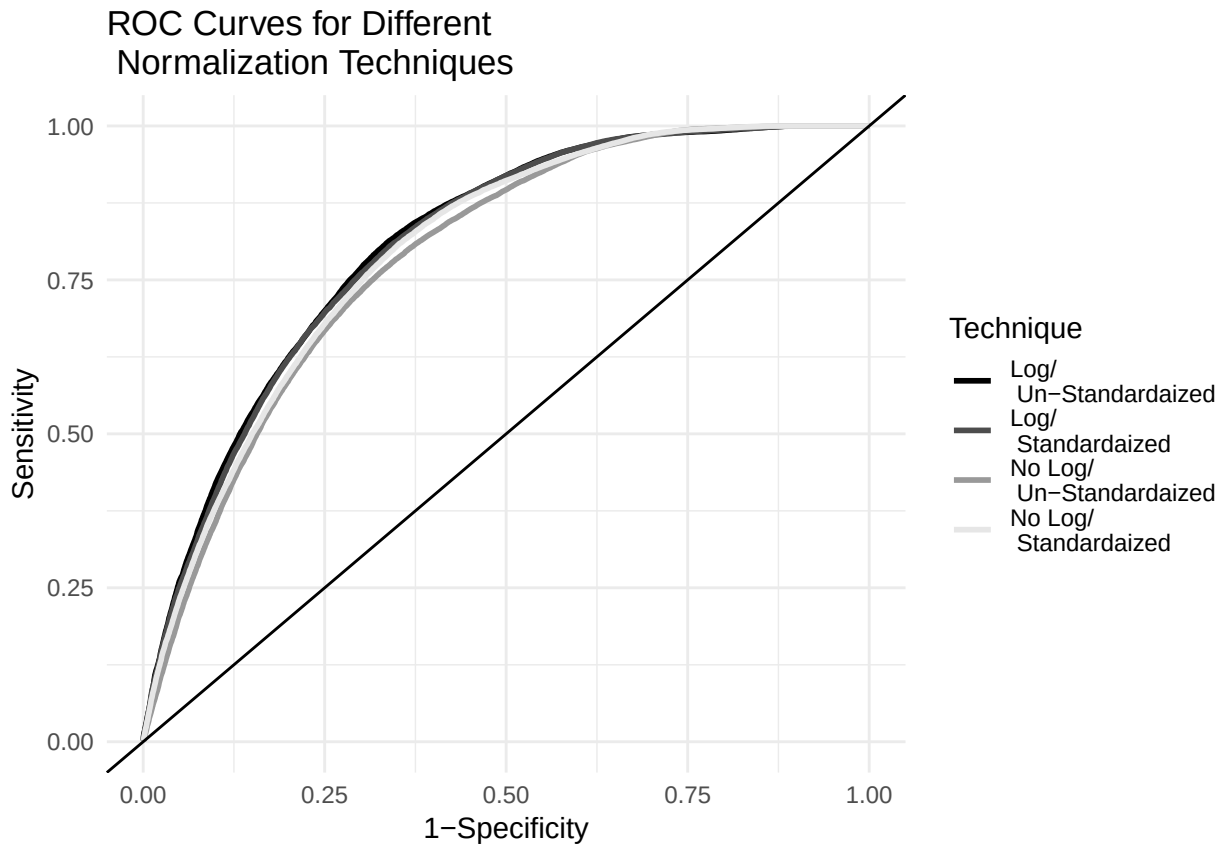



```
#datatable(auc.plot)
kable(auc.plot)
```

	Normalization.Technique	auc
2	Log/Un-Standardaized	0.809
1	Log/Standardaized	0.806

	Normalization.Technique	auc
3	No Log/Standardaized	0.797
4	No Log/Un-Standardaized	0.788

```
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)
```