# Measuring Performance: Ours vs Mourad

*Spiro Stilianoudakis*

## Contents



## Loading Libraries

```
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)
```

```
## Warning: package 'data.table' was built under R version 3.4.4
```

```
library(ggplot2)
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:data.table':
##
##     between, first, last

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

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

```
library(DMwR)
```

```
## Loading required package: grid

##
## Attaching package: 'DMwR'

## The following object is masked from 'package:plyr':
##
##     join
```

```
library(gridExtra)
```

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

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

```
library(knitr)
```

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

```
library(ROCR)
```

```
## Loading required package: gplots
```

```
##
## Attaching package: 'gplots'
## The following object is masked from 'package:stats':
##
##     lowess
```

## Reading in RDS objects

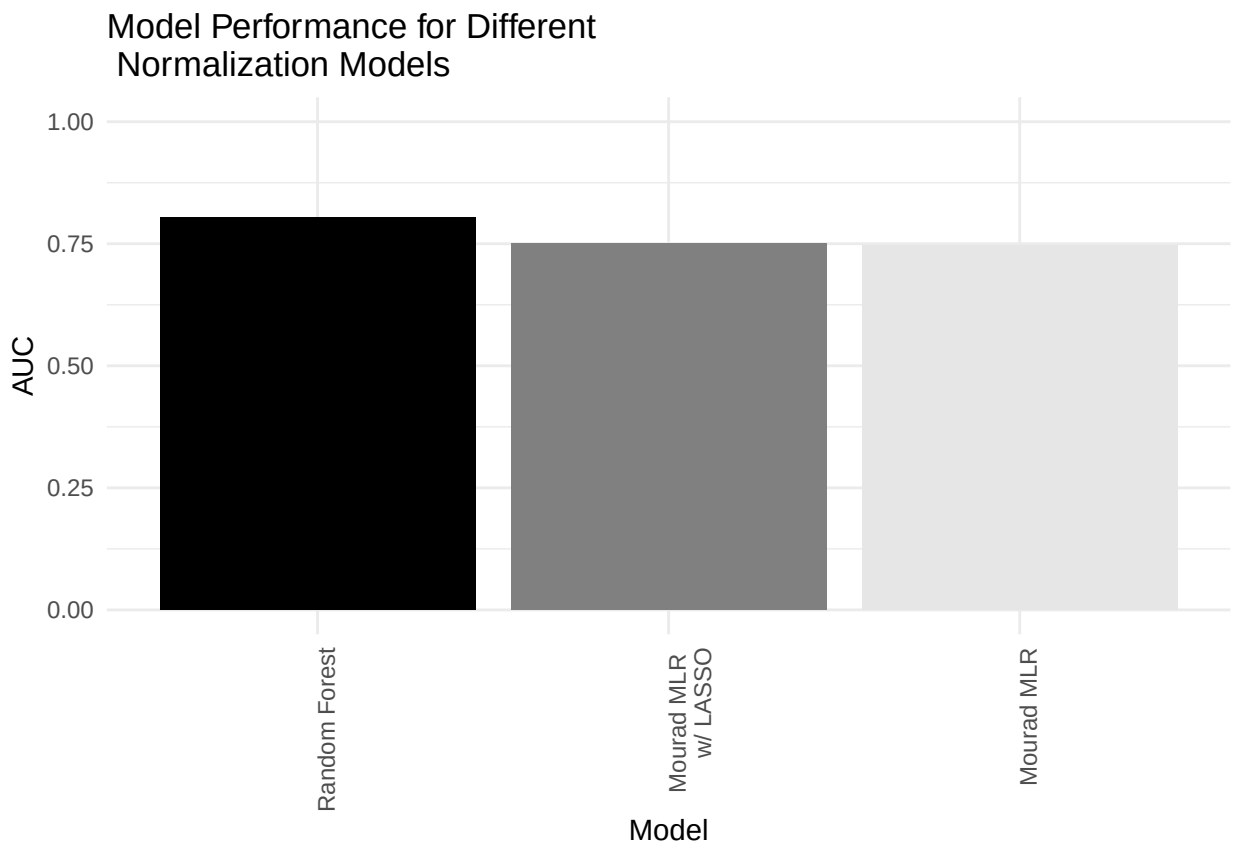
```
#mourad model
mourad.auc <- readRDS("C:/Users/Spiro Stilianoudakis/Documents/TAD_data/RData/GM12878/mourad_model/moura

mourad.roc <- readRDS("C:/Users/Spiro Stilianoudakis/Documents/TAD_data/RData/GM12878/mourad_model/moura

mourad.summary <- readRDS("C:/Users/Spiro Stilianoudakis/Documents/TAD_data/RData/GM12878/mourad_model/m


#mourad model with lasso
mourad.lasso.auc <- readRDS("C:/Users/Spiro Stilianoudakis/Documents/TAD_data/RData/GM12878/mourad_model

mourad.lasso.roc <- readRDS("C:/Users/Spiro Stilianoudakis/Documents/TAD_data/RData/GM12878/mourad_model

mourad.lasso.summary <- readRDS("C:/Users/Spiro Stilianoudakis/Documents/TAD_data/RData/GM12878/mourad_m


#random forest
rflst <- readRDS("C:/Users/Spiro Stilianoudakis/Documents/TAD_data/RData/GM12878/our_pipeline/rflst.rds

#gbm
gbmlst <- readRDS("C:/Users/Spiro Stilianoudakis/Documents/TAD_data/RData/GM12878/our_pipeline/gbmlst.rd
```

## Model Performance

### AUCs

```
m.auc <- performance(mourad.auc,"auc")
m.auc <- m.auc@ y.values[[1]]

m.l.auc <- performance(mourad.lasso.auc,"auc")
m.l.auc <- m.l.auc@ y.values[[1]]

#random forest
rf.auc <- mean(rflst[[3]])

#random forest
gbm.auc <- mean(gbmlst[[3]])

#Plotting AUCs
auc.plot <- data.frame(Model=c("Mourad MLR",
                               "Mourad MLR w/ LASSO",
```

```
                                        "Random Forest"),
                      auc=c(m.auc,
                            m.l.auc,
                            rf.auc))

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

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

p<-ggplot(data=auc.plot, aes(x=Model, y=auc, fill=Model)) +
  xlab("Model") + ylab("AUC") +
  geom_bar(stat="identity") + ylim(0,1) +
  scale_fill_manual(values=grey(c(0,.5,.9)), guide=FALSE) +
  scale_x_discrete(labels= c("Random Forest",
                             "Mourad MLR \n w/ LASSO",
                             "Mourad MLR")) +
  theme_minimal() +
  theme(axis.text.x = element_text(angle = 90, hjust = 1)) +
  ggtitle("Model Performance for Different \n Normalization Models")
p
```



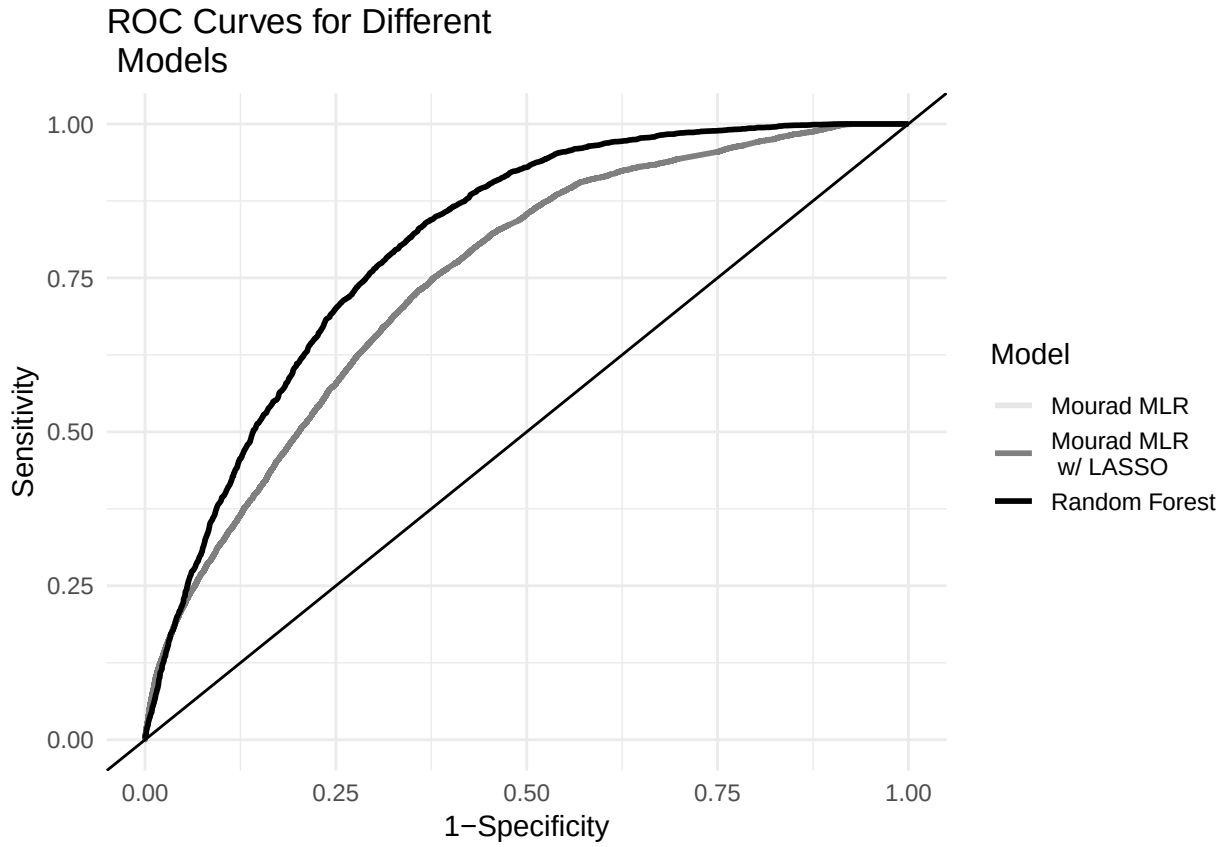

```
kable(auc.plot)
```

| | Model | auc |
|---|---|---|
| 3 | Random Forest | 0.8045185 |

| | Model | auc |
|---|---|---|
| 2 | Mourad MLR w/ LASSO | 0.7503714 |
| 1 | Mourad MLR | 0.7474979 |

## ROC Curves

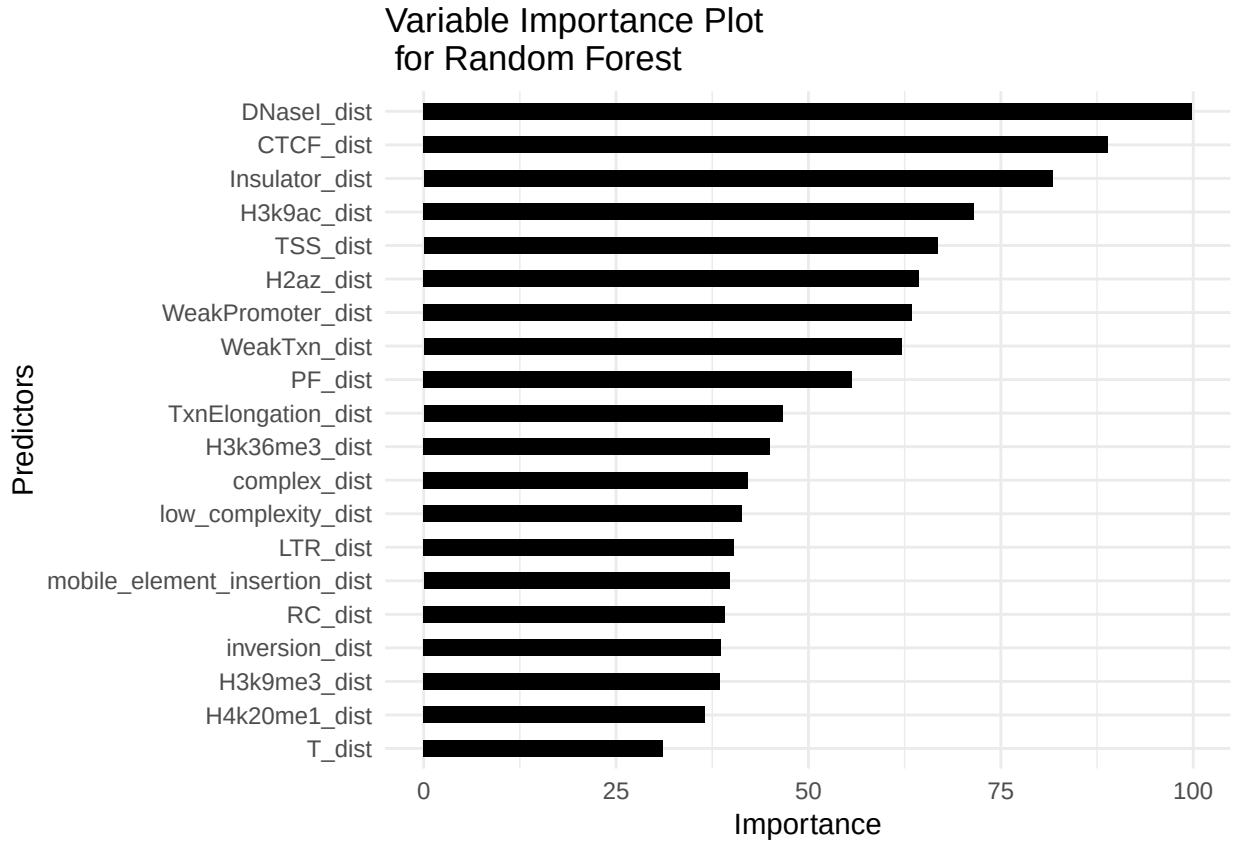
```
#mourad model
mourad.roc.fpr <- mourad.roc@ x.values[[1]]
mourad.roc.tpr <- mourad.roc@ y.values[[1]]
mourad.roc.df <- cbind.data.frame(fpr=mourad.roc.fpr,
                                  tpr=mourad.roc.tpr,
                                  Model = rep("M", length(mourad.roc.tpr)))

#mourad model w/ lasso
mourad.lasso.roc.fpr <- mourad.roc@ x.values[[1]]
mourad.lasso.roc.tpr <- mourad.roc@ y.values[[1]]
mourad.lasso.roc.df <- cbind.data.frame(fpr=mourad.lasso.roc.fpr,
                                  tpr=mourad.lasso.roc.tpr,
                                  Model = rep("MwL", length(mourad.lasso.roc.fpr)))

#random forest
rf.fpr <- rowMeans(rflst[[2]])
rf.tpr <- rowMeans(rflst[[1]])
rf.roc.df <- cbind.data.frame(fpr=rf.fpr,
                                  tpr=rf.tpr,
                                  Model = rep("RF", length(rf.fpr)))

#concatenating data frames
allrocdat <- rbind.data.frame(mourad.roc.df, mourad.lasso.roc.df, rf.roc.df)

ggplot(data=allrocdat, aes(x=fpr, y=tpr, color=Model)) +
  geom_line(size=1) +
  scale_colour_manual(name="Model",
    labels=c("Mourad MLR",
             "Mourad MLR \n w/ LASSO",
             "Random Forest"),
    values=grey(c(.9,.5,0))) +
  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 Models")
```




## Variable Importance Plot

```
#RF
varimp.rf <- as.vector(rowMeans(rflst[[4]]))

rownames(rflst[[4]])[grep("Gm12878_", rownames(rflst[[4]]))] <- gsub("Gm12878_", "", rownames(rflst[[4]]

#rownames(rflst[[4]])[grep("_dist", rownames(rflst[[4]]))] <- gsub("_dist", "", rownames(rflst[[4]])[gr

varimp.rf.df <- data.frame(Feature=rownames(rflst[[4]]),
                           Importance=varimp.rf)
varimp.rf.df <- varimp.rf.df[order(varimp.rf.df$Importance),]
numvarrf <- dim(varimp.rf.df)[1]
varimp.rf.df <- varimp.rf.df[(numvarrf-19):numvarrf,]
varimp.rf.df$Feature <- factor(varimp.rf.df$Feature,levels=varimp.rf.df$Feature)

rfp <- ggplot(varimp.rf.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="black") +
  coord_flip() +
  theme_minimal() +
  ggtitle("Variable Importance Plot \n for Random Forest")
rfp
```




## Estimates from Mourad Models

```
#mourad model
dim(mourad.summary)
```

```
## [1] 60  9
```

```
sig.vars <- mourad.summary[mourad.summary$`Pr(>|z|)` < 0.05,]
dim(sig.vars)
```

```
## [1] 37  9
```

```
sig.vars <- sig.vars[order(abs(sig.vars$Estimate), decreasing = TRUE),]
rownames(sig.vars) <- NULL
sig.vars <- sig.vars[1:20, which(colnames(sig.vars) %in% c("GenomicFeature","Estimate", "Pr(>|z|)"))]

kable(sig.vars)
```

| GenomicFeature | Estimate | Pr(>\|z\|) |
|---|---|---|
| A | 3.8435655 | 0.0000000 |
| B | 3.0004561 | 0.0000000 |
| Insulator | 1.8631419 | 0.0000000 |
| ActivePromoter | 1.8180323 | 0.0000000 |
| WeakPromoter | 1.5380977 | 0.0000000 |
| WeakEnhancer6 | 1.3605898 | 0.0000001 |
| TxnElongation | 1.2286235 | 0.0000006 |
| WeakTxn | 1.1905517 | 0.0000011 |
| Repressed | 1.1758725 | 0.0000019 |
| PoisedPromoter | 1.1510972 | 0.0000458 |
| TxnTransition | 1.1367741 | 0.0000071 |
| CTCF | 1.0690906 | 0.0000000 |
| WeakEnhancer7 | 1.0399866 | 0.0000254 |
| Heterochromlo | 1.0286227 | 0.0000242 |
| StrongEnhancer5 | 1.0038337 | 0.0001420 |
| StrongEnhancer4 | 0.9625891 | 0.0002621 |
| satellite | 0.7478179 | 0.0034025 |
| WE | -0.4638145 | 0.0056090 |
| DNaseI | 0.4612106 | 0.0016588 |
| sequence_alteration | -0.4475249 | 0.0036293 |

```
#mourad model w/ lasso
dim(mourad.lasso.summary)
```

```
## [1] 60  3
```

```
mourad.lasso.summary <- as.data.frame(mourad.lasso.summary)
mourad.lasso.summary$Estimate <- as.numeric(mourad.lasso.summary$Estimate)

sig.vars.lasso <- mourad.summary[order(abs(mourad.lasso.summary$Estimate), decreasing = TRUE),]
rownames(sig.vars.lasso) <- NULL
sig.vars.lasso <- sig.vars.lasso[1:20, which(colnames(sig.vars.lasso) %in% c("GenomicFeature","Estimate

kable(sig.vars.lasso)
```

| GenomicFeature | Estimate |
|---|---|
| A | 3.8435655 |
| B | 3.0004561 |
| CTCF | 1.0690906 |
| Insulator | 1.8631419 |
| ActivePromoter | 1.8180323 |
| WeakPromoter | 1.5380977 |
| DNaseI | 0.4612106 |
| WeakEnhancer6 | 1.3605898 |
| satellite | 0.7478179 |
| H3k27me3 | 0.3433862 |
| H2az | 0.3275703 |
| TxnElongation | 1.2286235 |
| Repressed | 1.1758725 |
| WeakTxn | 1.1905517 |
| H3k36me3 | 0.2228918 |
| PoisedPromoter | 1.1510972 |

| GenomicFeature | Estimate |
|---|---|
| H3k4me3 | 0.1821350 |
| tandem_duplication | 0.2050713 |
| SINE | 0.2394262 |
| H4k20me1 | 0.1957359 |

## Comparing Results

```
#finding common features between the models

#remove "_dist" from feature list of random forest
rffeat <- varimp.rf.df$Feature[order(varimp.rf.df$Importance, decreasing = TRUE)]
rffeat <- gsub("_dist", "", rffeat)
rffeat <- factor(rffeat)
rfrank <- 1:20

mrank <- match(rffeat, sig.vars$GenomicFeature)

mwlrank <- match(rffeat, sig.vars.lasso$GenomicFeature)

rankdf <- cbind.data.frame(Feature=rffeat,
                           "Random Forest" = rfrank <- 1:20,
                           "Mourad" = mrank,
                           "Mourad w/ LASSO" = mwlrank)

kable(rankdf)
```

| Feature | Random Forest | Mourad | Mourad w/ LASSO |
|---|---|---|---|
| DNaseI | 1 | 19 | 7 |
| CTCF | 2 | 12 | 3 |
| Insulator | 3 | 3 | 4 |
| H3k9ac | 4 | NA | NA |
| TSS | 5 | NA | NA |
| H2az | 6 | NA | 11 |
| WeakPromoter | 7 | 5 | 6 |
| WeakTxn | 8 | 8 | 14 |
| PF | 9 | NA | NA |
| TxnElongation | 10 | 7 | 12 |
| H3k36me3 | 11 | NA | 15 |
| complex | 12 | NA | NA |
| low_complexity | 13 | NA | NA |
| LTR | 14 | NA | NA |
| mobile_element_insertion | 15 | NA | NA |
| RC | 16 | NA | NA |
| inversion | 17 | NA | NA |
| H3k9me3 | 18 | NA | NA |
| H4k20me1 | 19 | NA | 20 |
| T | 20 | NA | NA |