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
```

```r
library(ROCR)
```

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

```
##
## Attaching package: 'gplots'
```

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

Reading in RDS objects

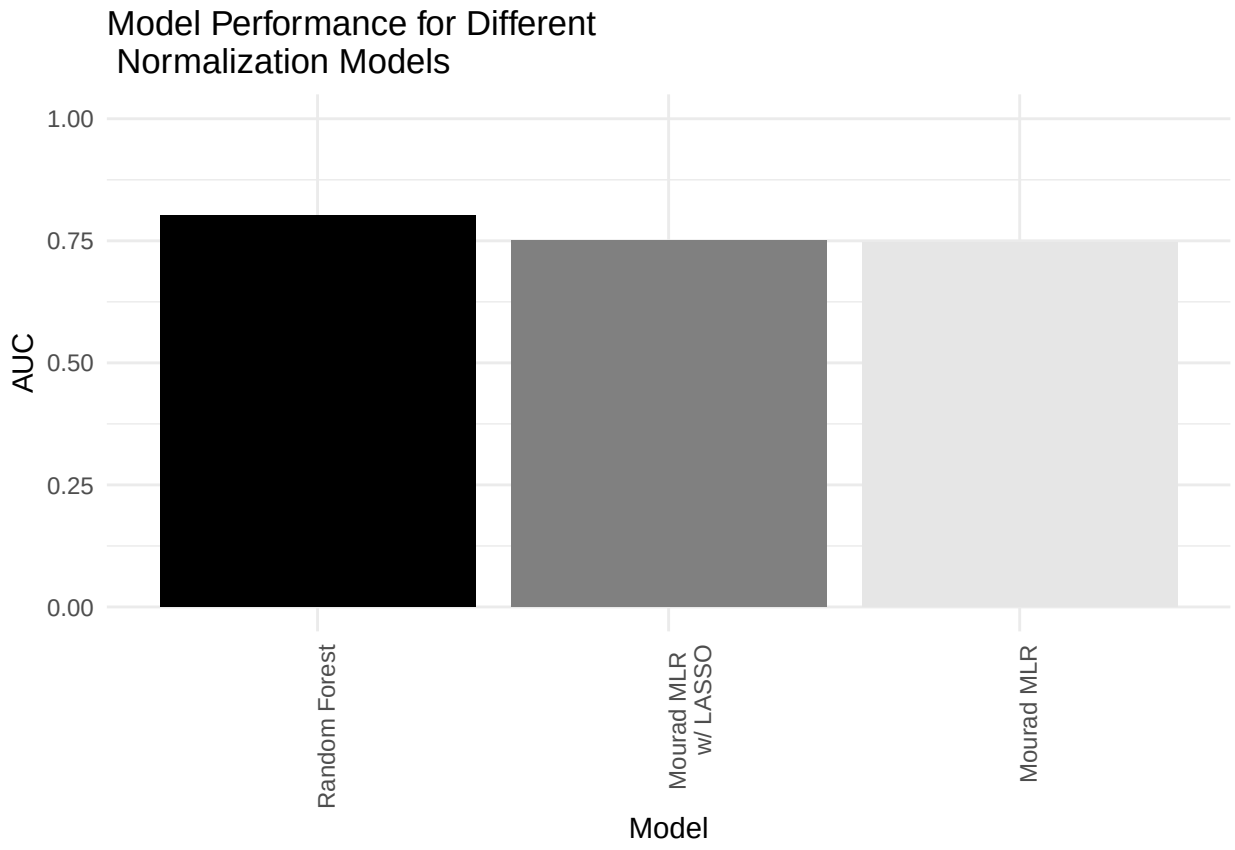
```r
#mourad model
mourad.auc <- readRDS("C:/Users/Spiro Stilianoudakis/Documents/TAD_data/RData/GM12878/mourad_model/moura

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

```r
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]])
```

```r
#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
```

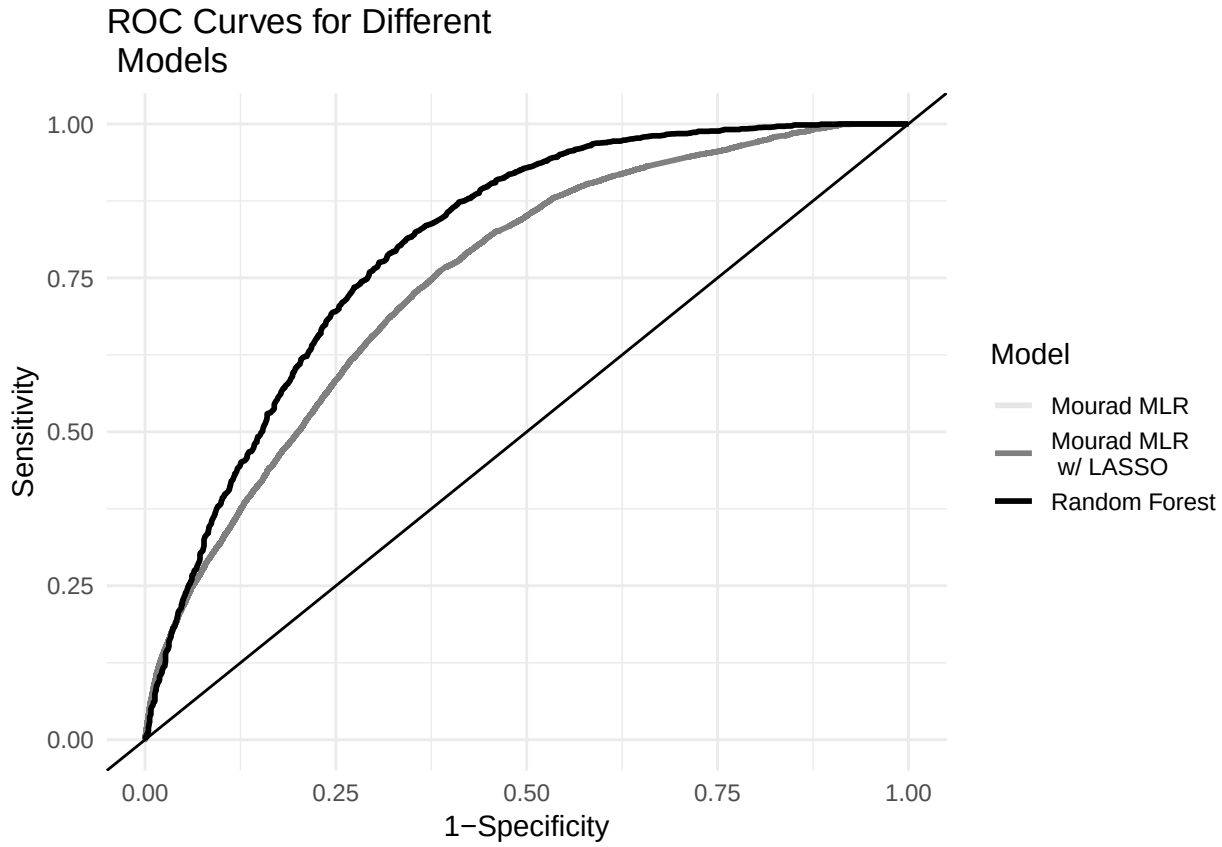


```
kable(auc.plot)
```

	Model	auc
3	Random Forest	0.8023156
2	Mourad MLR w/ LASSO	0.7503714
1	Mourad MLR	0.7473133

ROC Curves

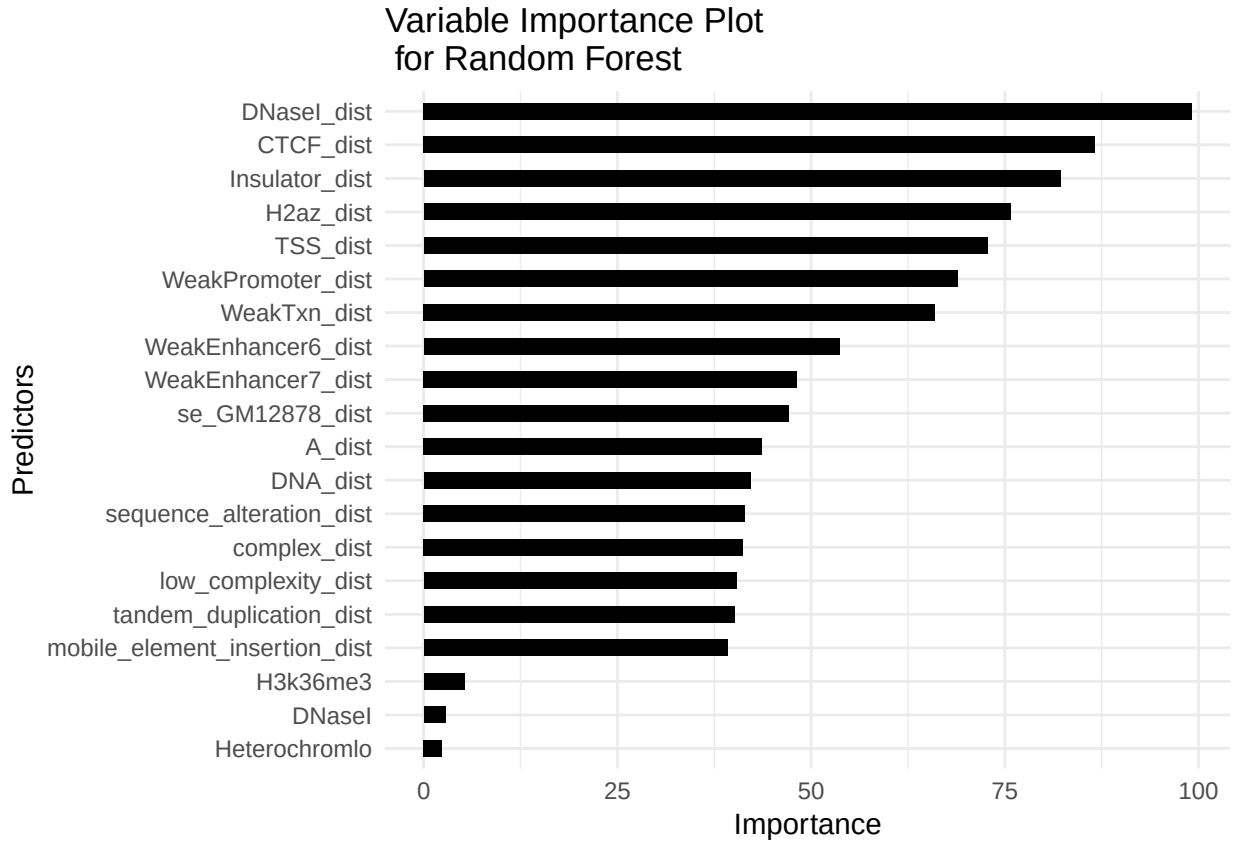
```
#mourad model
mourad.roc.fpr <- mourad.roc@ x.values[[1]]
mourad.roc.tpr <- mourad.roc@ y.values[[1]]
mourad.roc.df <- cbind.data.frame(fpr=mourad.roc.fpr,
                                  tpr=mourad.roc.tpr,
                                  Model = rep("M", length(mourad.roc.tpr)))

#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$Estimate <- as.numeric(as.character(mourad.lasso.summary$Estimate))

sig.vars.lasso <- mourad.lasso.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
srpRNA	-4.24869
A	3.43288
B	2.58396
RC	-1.29609
novel_sequence_insertion	-1.26161
CTCF	1.14327
Insulator	1.01586
ActivePromoter	0.92045
WeakPromoter	0.66831
DNaseI	0.57307
WeakEnhancer6	0.53279
satellite	0.44559
RepetitiveCNV15	-0.40150
H3k27me3	0.36100
H2az	0.32995
sequence_alteration	-0.31331

GenomicFeature	Estimate
WE	-0.30484
H3k9me3	-0.26855
TxnElongation	0.25109
se	-0.24898

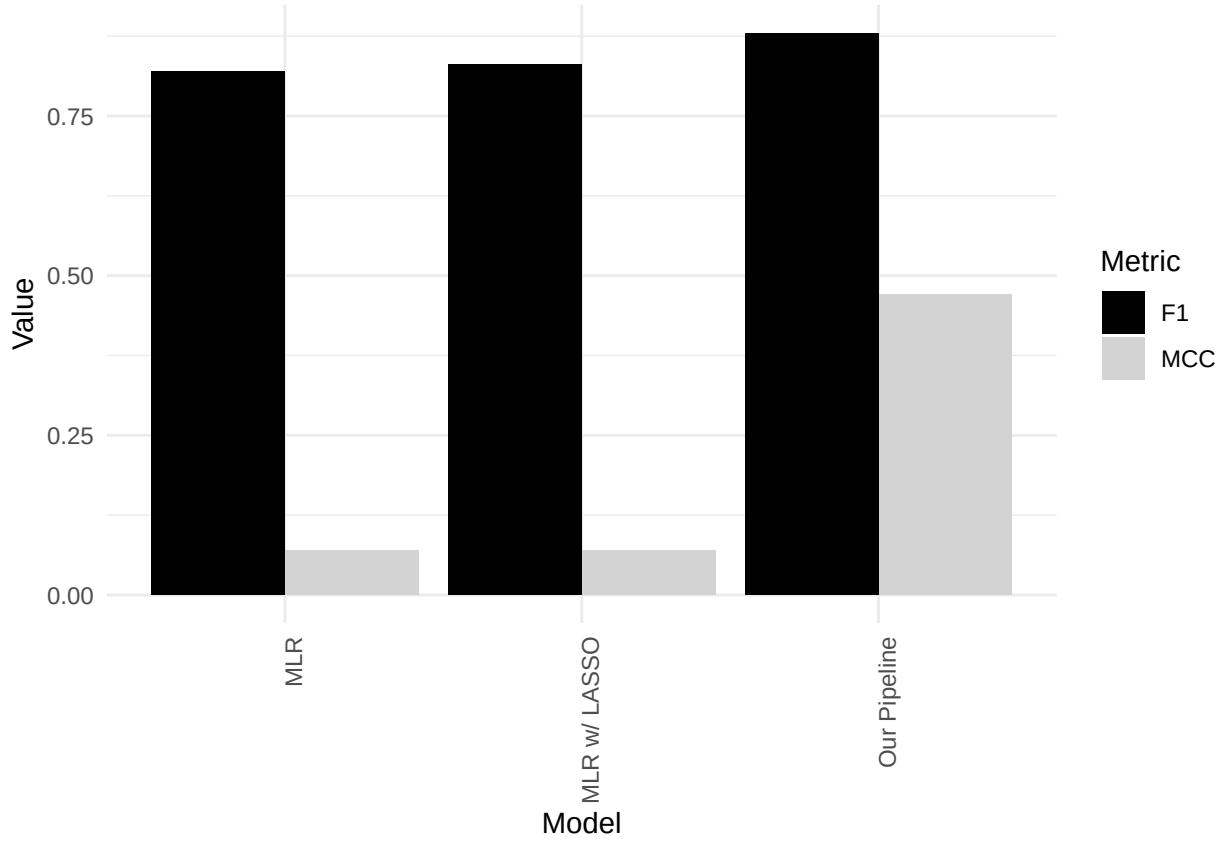
Comparing Performance Metrics

```
options(scipen = 999)

#mourad MLR
mouradperf <- readRDS("C:/Users/Spiro Stilianoudakis/Documents/TAD_data/RData/GM12878/mourad_model/moura

mouradperf <- round(mouradperf,2)
mouradperf[1:5,1] <- round(mouradperf[1:5,1], 0)

#mourad LASSO
mourad.lasso.perf <- readRDS("C:/Users/Spiro Stilianoudakis/Documents/TAD_data/RData/GM12878/mourad_mod

mourad.lasso.perf <- round(mourad.lasso.perf,2)
mourad.lasso.perf[1:5,1] <- round(mourad.lasso.perf[1:5,1],0)

#Our pipeline
rfperf <- readRDS("C:/Users/Spiro Stilianoudakis/Documents/TAD_data/RData/GM12878/our_pipeline/rfperf.r

rfperf <- round(as.matrix(rowMeans(rfperf)),2)
rfperf[1:5,1] <- round(rfperf[1:5,1],0)


perfdat <- cbind.data.frame(rownames(mouradperf),
                            mouradperf,
                            mourad.lasso.perf,
                            rfperf)
rownames(perfdat) <- NULL
colnames(perfdat) <- c("Metric", "MLR", "MLR w/ LASSO Regularization", "Our Pipeline")

kable(perfdat)
```

Metric	MLR	MLR w/ LASSO Regularization	Our Pipeline
TN	572916.00	571162.00	330.00
FN	2371.00	2314.00	101.00
FP	283532.00	285285.00	162.00
TP	5495.00	5553.00	386.00
Total	864314.00	864314.00	978.00
Sensitivity	0.70	0.71	0.79
Specificity	0.67	0.67	0.67
Kappa	0.02	0.02	0.46
Accuracy	0.67	0.67	0.73
Precision	0.02	0.02	0.70
FPR	0.33	0.33	0.33

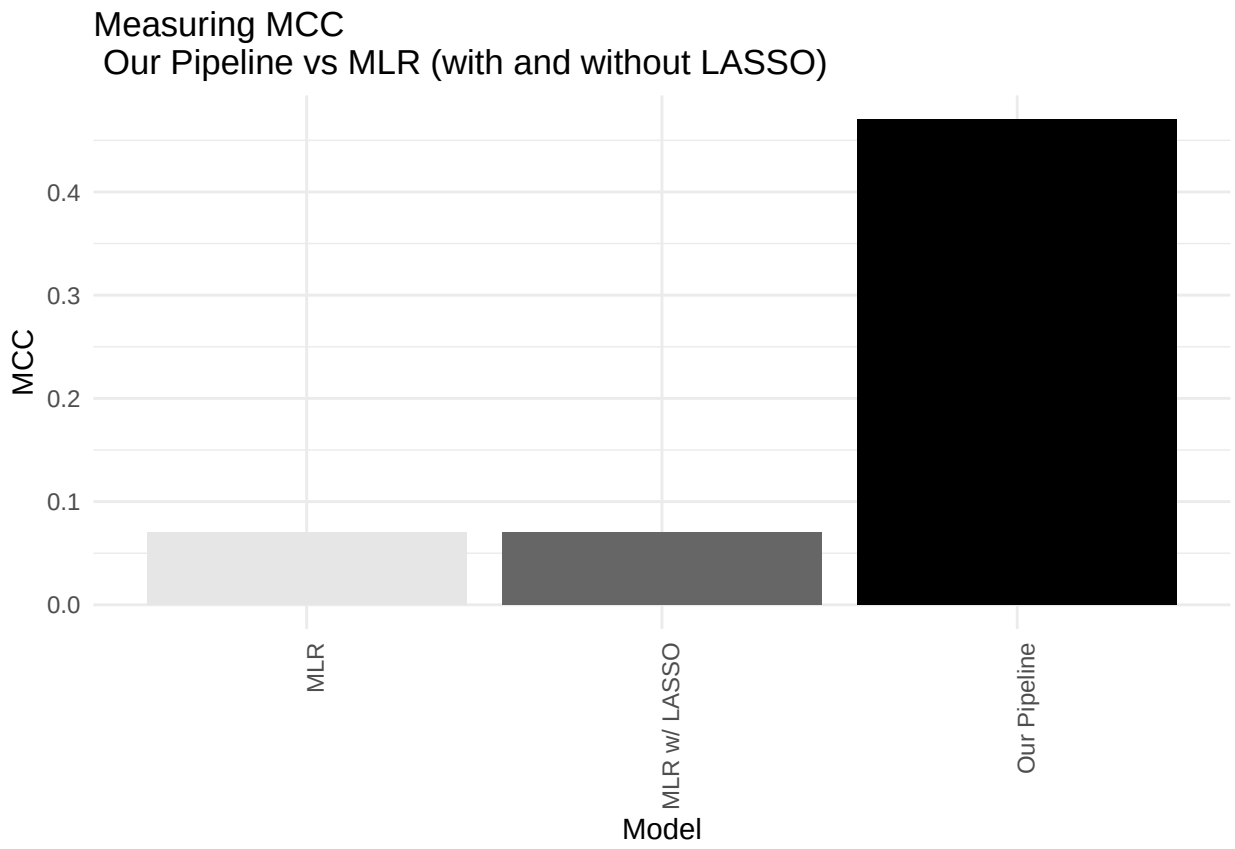
Metric	MLR	MLR w/ LASSO Regularization	Our Pipeline
FNR	0.30	0.29	0.21
FOR	0.00	0.00	0.23
NPV	1.00	1.00	0.77
MCC	0.07	0.07	0.47
F1	0.82	0.83	0.88

```
mccf1 <- data.frame(Metric = c(rep("MCC",3), rep("F1",3)),
                    Model = rep(c("MLR",
                                  "MLR w/ LASSO",
                                  "Our Pipeline"), 2),
                    Value = c(as.numeric(perfdat[15,2:4]), as.numeric(perfdat[16,2:4])))

ggplot(data=mccf1, aes(x=Model, y=Value, fill=Metric)) +
  geom_bar(stat="identity", position=position_dodge()) +
  scale_fill_manual(values=c('black','lightgray')) +
  xlab("Model") +
  theme_minimal() +
  theme(axis.text.x = element_text(angle = 90, hjust = 1))
```

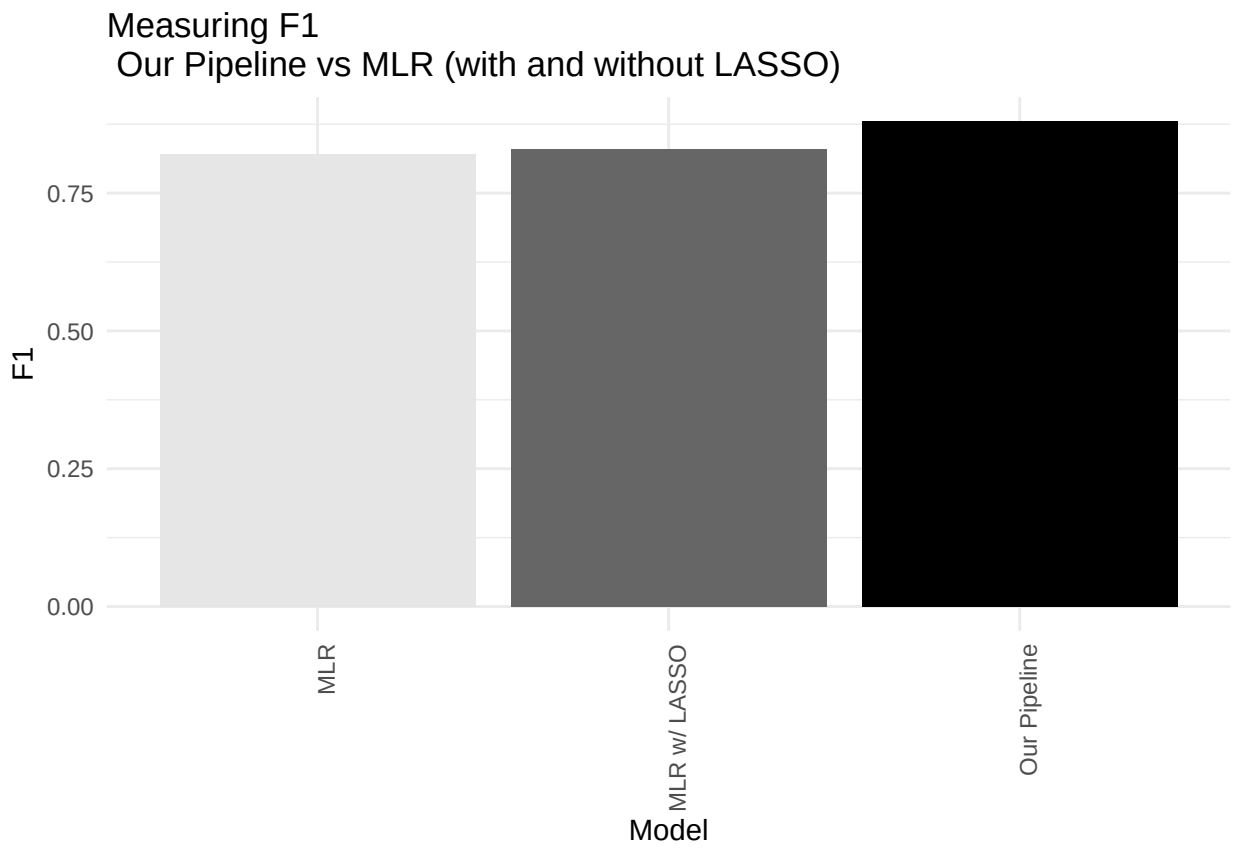



```
MCCplot<-ggplot(data=mccf1[1:3,], aes(x=Model, y=Value, fill=Model)) +
  xlab("Model") + ylab("MCC") +
  geom_bar(stat="identity") +
  scale_fill_manual(values=gray(rev(c(0,.4,.9))), guide=FALSE) +
  theme_minimal() +
```

```
  theme(axis.text.x = element_text(angle = 90, hjust = 1)) +
  ggtitle("Measuring MCC \n Our Pipeline vs MLR (with and without LASSO)")
MCCplot
```




```
F1plot<-ggplot(data=mccf1[4:6,], aes(x=Model, y=Value, fill=Model)) +
  xlab("Model") + ylab("F1") +
  geom_bar(stat="identity") +
  scale_fill_manual(values=gray(rev(c(0,.4,.9))), guide=FALSE) +
  theme_minimal() +
  theme(axis.text.x = element_text(angle = 90, hjust = 1)) +
  ggtitle("Measuring F1 \n Our Pipeline vs MLR (with and without LASSO)")
F1plot
```



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	10

Feature	Random Forest	Mourad	Mourad w/ LASSO
CTCF	2	12	6
Insulator	3	3	7
H2az	4	NA	15
TSS	5	NA	NA
WeakPromoter	6	5	9
WeakTxn	7	8	NA
WeakEnhancer6	8	6	11
WeakEnhancer7	9	13	NA
se_GM12878	10	NA	NA
A	11	1	2
DNA	12	NA	NA
sequence_alteration	13	20	16
complex	14	NA	NA
low_complexity	15	NA	NA
tandem_duplication	16	NA	NA
mobile_element_insertion	17	NA	NA
H3k36me3	18	NA	NA
DNaseI	19	19	10
Heterochromlo	20	14	NA