Model Filtering

Spiro Stilianoudakis

August 16, 2018

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(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
```

```r
library(DMwR)
```

```
## Loading required package: grid

##
## Attaching package: 'DMwR'

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

```r
library(gridExtra)
```

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

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

```r
#library(DT)
library(ggplot2)
```

Reading in data

```r
gm12878_10kb <- readRDS("C:/Users/Spiro Stilianoudakis/Documents/TAD_data/RData/GM12878/10kb_bins/chr22/
```

Taking log2 transform of continous data

```r
#Taking log2 transform of continous data
cols <- c(grep("dist",colnames(gm12878_10kb)))
gm12878_10kb[,cols] <- apply(gm12878_10kb[,cols], 2, function(x){log(x + 1, base=2)})
```

Changing binary variables to factors

```
cols <- c(intersect(grep("score",colnames(gm12878_10kb), invert = TRUE),
          grep("dist",colnames(gm12878_10kb), invert = TRUE)))
gm12878_10kb[,cols] <- lapply(gm12878_10kb[,cols], factor)
```

Changing levels of response (y) to yes no

```
levels(gm12878_10kb$y) <- c("No", "Yes")
```

Removing zero variance predictors

```
nzv <- nearZeroVar(gm12878_10kb[,-1], saveMetrics= TRUE)
nzvar <- rownames(nzv[nzv$nzv,])

nzvar
```

```
##  [1] "complex"                  "mobile_element_insertion"
##  [3] "novel_sequence_insertion" "sequence_alteration"
##  [5] "low_complexity"           "other"
##  [7] "RC"                       "satellite"
##  [9] "Gm12878_RepetitiveCNV14"  "Gm12878_RepetitiveCNV15"
## [11] "Gm12878_PoisedPromoter"   "CHR"
```

```
gm12878_10kb_f <- gm12878_10kb[, -which(colnames(gm12878_10kb) %in% nzvar)]

saveRDS(gm12878_10kb_f, "C:/Users/Spiro Stilianoudakis/Documents/TAD_data/RData/GM12878/10kb_bins/chr22/
```