hw4

Qianhui Yang

10/10/2020

Q1

Accuracy=0.553

```
## -- Attaching packages ---------------------------------------------------------------------------- tid
## v ggplot2 3.3.2     v purrr   0.3.4
## v tibble  3.0.3     v dplyr   1.0.2
## v tidyr   1.1.2     v stringr 1.4.0
## v readr   1.4.0     v forcats 0.5.0
## -- Conflicts ------------------------------------------------------------------------------- tidyverse
## x dplyr::filter() masks stats::filter()
## x dplyr::lag()    masks stats::lag()
## Loaded gbm 2.1.8
##
## Attaching package: 'MLmetrics'
## The following object is masked from 'package:base':
##
##     Recall
##
## -- Column specification --------------------------------------------------------------------------
## cols(
##   Gender = col_character(),
##   Height = col_double(),
##   Weight = col_double(),
##   Index = col_double()
## )
## # A tibble: 6 x 3
## # Groups:   Gender [2]
##   Gender exp_group     n
##   <fct>  <chr>     <int>
## 1 Female test         85
## 2 Female train        85
## 3 Female validate     85
## 4 Male   test         81
## 5 Male   train        82
## 6 Male   validate     82
## [1] 0.5329341
```

Q2

accuracy=0.443

```
## Using 100 trees...
```

```
## [1] 0.4431138
```

Q3

F1_score=0.89

```
##
## -- Column specification --------------------------------------------------------
## cols(
##   Gender = col_character(),
##   Height = col_double(),
##   Weight = col_double(),
##   Index = col_double()
## )
```

```
## # A tibble: 6 x 3
## # Groups:   Gender [2]
##   Gender exp_group     n
##   <chr>  <chr>     <int>
## 1 Female test         85
## 2 Female train        85
## 3 Female validate     85
## 4 Male   test         16
## 5 Male   train        17
## 6 Male   validate     17
```

```
## Using 100 trees...
```

```
## [1] 0.8950276
```

Q4

ROC curve has limited area (AUC close to 0.5) under the curve. The ROC curve is plotted with TPR against the FPR where TPR is on y-axis and FPR is on the x-axis. ROC curve is a performance measurement for classification problem at various thresholds settings. ROC is a probability curve . It tells how much model is capable of distinguishing between classes. Higher the area under, better the model is at predicting 0s as 0s and 1s as 1s.

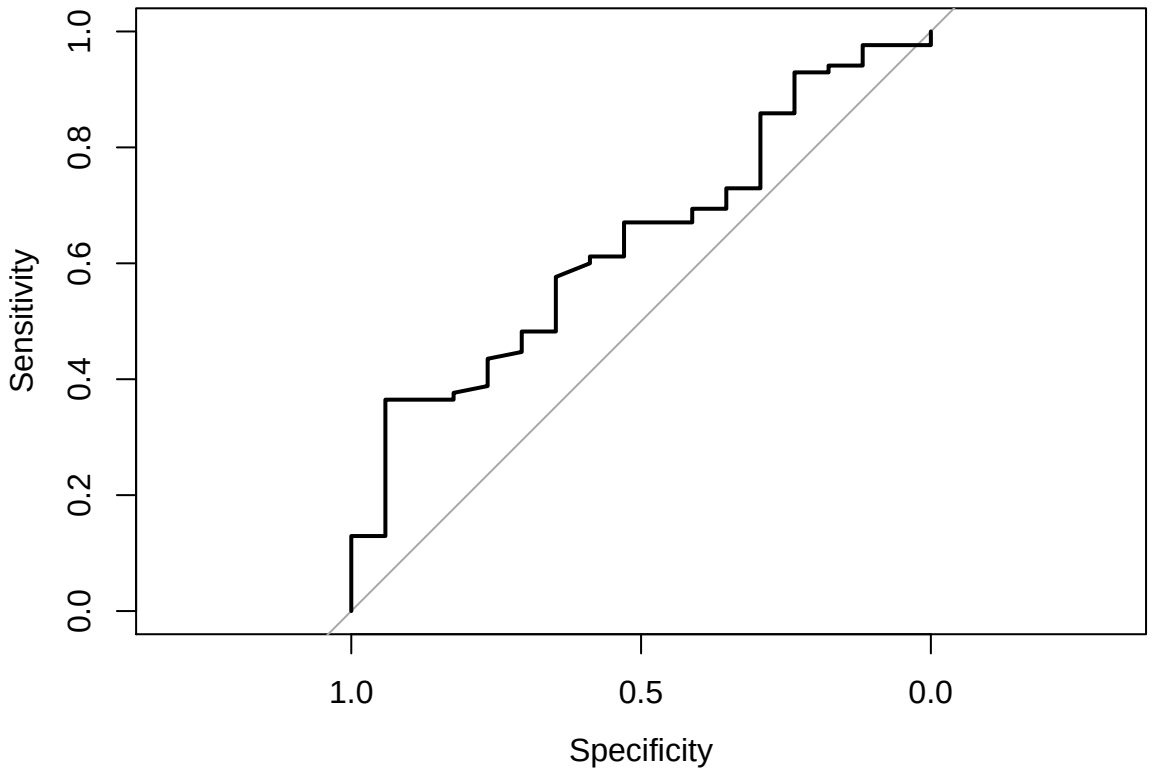
```
## Type 'citation("pROC")' for a citation.
```

```
##
## Attaching package: 'pROC'
```

```
## The following objects are masked from 'package:stats':
##
##     cov, smooth, var
```

```
## Setting levels: control = 0, case = 1
```

```
## Setting direction: controls < cases
```



#Q5 The first cluster is male , the second is female. Because the first cluster has overall heavier in weight, taller in height, and larger index. K Means gives assignments for each cluster as well as the N cluster centers and optimizes the sum of squared distances to the closest cluster center.The center can represent the character of the cluster in Kmeans.

```
##     Weight   Height    Index
## 1  78.4664 169.8221 2.857708
## 2 134.2024 170.0688 4.659919
```