GPH 2338 Project

Yi Yang & Sohaib Hasan & Guangyu Yang

2023-03-13

Data Upload

```
library(haven)
library(psych)
library(caret)
library(tidyverse)
library(ggplot2)
library(psych)
library(pander)
library(corrplot)
library(pander)
library(readr)
library(r02pro)
library(plyr)
library(tree)
library(gbm)
library(caret)
library(leaps)
library(readr)
```

Data Preparation

```
df<-readr::read_tsv("brca_metabric_clinical_data.tsv")
```

```
## Rows: 2509 Columns: 39
## -- Column specification --------------------------------------------------------
## Delimiter: "\t"
## chr (27): Study ID, Patient ID, Sample ID, Type of Breast Surgery, Cancer Ty...
## dbl (12): Age at Diagnosis, Cohort, Neoplasm Histologic Grade, Lymph nodes e...
##
## i Use `spec()` to retrieve the full column specification for this data.
## i Specify the column types or set `show_col_types = FALSE` to quiet this message.
```

```
colnames(df)
```

```
##  [1] "Study ID"                        "Patient ID"
##  [3] "Sample ID"                       "Age at Diagnosis"
##  [5] "Type of Breast Surgery"          "Cancer Type"
##  [7] "Cancer Type Detailed"            "Cellularity"
##  [9] "Chemotherapy"                    "Pam50 + Claudin-low subtype"
## [11] "Cohort"                         "ER status measured by IHC"
## [13] "ER Status"                      "Neoplasm Histologic Grade"
## [15] "HER2 status measured by SNP6"   "HER2 Status"
## [17] "Tumor Other Histologic Subtype" "Hormone Therapy"
## [19] "Inferred Menopausal State"      "Integrative Cluster"
## [21] "Primary Tumor Laterality"       "Lymph nodes examined positive"
## [23] "Mutation Count"                 "Nottingham prognostic index"
## [25] "Oncotree Code"                  "Overall Survival (Months)"
## [27] "Overall Survival Status"        "PR Status"
## [29] "Radio Therapy"                  "Relapse Free Status (Months)"
## [31] "Relapse Free Status"            "Number of Samples Per Patient"
## [33] "Sample Type"                    "Sex"
## [35] "3-Gene classifier subtype"      "TMB (nonsynonymous)"
## [37] "Tumor Size"                     "Tumor Stage"
## [39] "Patient’s Vital Status"
```

```
colSums(is.na(df))
```

```
##                       Study ID                     Patient ID
##                              0                              0
##                      Sample ID               Age at Diagnosis
##                              0                             11
##         Type of Breast Surgery                    Cancer Type
##                            554                              0
##           Cancer Type Detailed                    Cellularity
##                              0                            592
##                   Chemotherapy    Pam50 + Claudin-low subtype
##                            529                            529
##                         Cohort       ER status measured by IHC
##                             11                             83
##                      ER Status      Neoplasm Histologic Grade
##                             40                            121
##   HER2 status measured by SNP6                    HER2 Status
##                            529                            529
## Tumor Other Histologic Subtype                Hormone Therapy
##                            135                            529
##      Inferred Menopausal State            Integrative Cluster
##                            529                            529
##       Primary Tumor Laterality  Lymph nodes examined positive
##                            639                            266
##                 Mutation Count    Nottingham prognostic index
##                            151                            222
##                  Oncotree Code      Overall Survival (Months)
##                              0                            528
##        Overall Survival Status                      PR Status
##                            528                            529
##                  Radio Therapy   Relapse Free Status (Months)
##                            529                            121
##            Relapse Free Status  Number of Samples Per Patient
##                             21                              0
##                    Sample Type                            Sex
##                              0                              0
##      3-Gene classifier subtype            TMB (nonsynonymous)
##                            745                              0
##                    Tumor Size                    Tumor Stage
##                           149                            721
##        Patient’s Vital Status
##                           529
```

```
print(df)
```

```
## # A tibble: 2,509 x 39
##    ‘Study ID‘    Patie~1 Sampl~2 Age a~3 Type ~4 Cance~5 Cance~6 Cellu~7 Chemo~8
##    <chr>         <chr>   <chr>     <dbl> <chr>   <chr>   <chr>   <chr>   <chr>
##  1 brca_metabric MB-0000 MB-0000    75.6 MASTEC~ Breast~ Breast~ <NA>    NO
##  2 brca_metabric MB-0002 MB-0002    43.2 BREAST~ Breast~ Breast~ High    NO
##  3 brca_metabric MB-0005 MB-0005    48.9 MASTEC~ Breast~ Breast~ High    YES
##  4 brca_metabric MB-0006 MB-0006    47.7 MASTEC~ Breast~ Breast~ Modera~ YES
##  5 brca_metabric MB-0008 MB-0008    77.0 MASTEC~ Breast~ Breast~ High    YES
##  6 brca_metabric MB-0010 MB-0010    78.8 MASTEC~ Breast~ Breast~ Modera~ NO
##  7 brca_metabric MB-0014 MB-0014    56.4 BREAST~ Breast~ Breast~ Modera~ YES
##  8 brca_metabric MB-0020 MB-0020    70   MASTEC~ Breast~ Breast~ High    YES
##  9 brca_metabric MB-0022 MB-0022    89.1 BREAST~ Breast~ Breast~ Modera~ NO
## 10 brca_metabric MB-0025 MB-0025    76.2 <NA>    Breast~ Breast~ <NA>    <NA>
## # ... with 2,499 more rows, 30 more variables:
## #   ‘Pam50 + Claudin-low subtype‘ <chr>, Cohort <dbl>,
## #   ‘ER status measured by IHC‘ <chr>, ‘ER Status‘ <chr>,
## #   ‘Neoplasm Histologic Grade‘ <dbl>, ‘HER2 status measured by SNP6‘ <chr>,
## #   ‘HER2 Status‘ <chr>, ‘Tumor Other Histologic Subtype‘ <chr>,
## #   ‘Hormone Therapy‘ <chr>, ‘Inferred Menopausal State‘ <chr>,
## #   ‘Integrative Cluster‘ <chr>, ‘Primary Tumor Laterality‘ <chr>, ...
```

```
df1 <- df %>% na.omit()
colSums(is.na(df1))
```

```
##                      Study ID                     Patient ID
##                             0                              0
##                     Sample ID               Age at Diagnosis
##                             0                              0
##        Type of Breast Surgery                    Cancer Type
##                             0                              0
##          Cancer Type Detailed                    Cellularity
##                             0                              0
##                  Chemotherapy    Pam50 + Claudin-low subtype
##                             0                              0
##                        Cohort      ER status measured by IHC
##                             0                              0
##                     ER Status      Neoplasm Histologic Grade
##                             0                              0
##  HER2 status measured by SNP6                    HER2 Status
##                             0                              0
## Tumor Other Histologic Subtype                Hormone Therapy
##                             0                              0
##     Inferred Menopausal State            Integrative Cluster
##                             0                              0
##      Primary Tumor Laterality  Lymph nodes examined positive
##                             0                              0
##                 Mutation Count    Nottingham prognostic index
##                              0                              0
##                  Oncotree Code      Overall Survival (Months)
##                              0                              0
##        Overall Survival Status                      PR Status
##                              0                              0
##                  Radio Therapy   Relapse Free Status (Months)
##                              0                              0
##            Relapse Free Status  Number of Samples Per Patient
##                              0                              0
##                    Sample Type                            Sex
##                              0                              0
##      3-Gene classifier subtype            TMB (nonsynonymous)
##                              0                              0
##                     Tumor Size                    Tumor Stage
##                              0                              0
##         Patient’s Vital Status
##                              0
```

```
df2<-df1[,-c(1, 2, 3, 6, 27)] # remove StudyID, Patient ID, Sample ID, Cancer Type
print(df2)
```

```
## # A tibble: 1,092 x 34
##    Age at Diagn~1 Type ~2 Cance~3 Cellu~4 Chemo~5 Pam50~6 Cohort ER st~7 ER St~8
##             <dbl> <chr>   <chr>   <chr>   <chr>   <chr>    <dbl> <chr>   <chr>
##  1           43.2 BREAST~ Breast~ High    NO      LumA         1 Positve Positi~
##  2           77.0 MASTEC~ Breast~ High    YES     LumB         1 Positve Positi~
##  3           78.8 MASTEC~ Breast~ Modera~ NO      LumB         1 Positve Positi~
##  4           86.4 BREAST~ Breast~ Modera~ NO      LumB         1 Positve Positi~
##  5           84.2 MASTEC~ Breast~ High    NO      Her2         1 Negati~ Positi~
##  6           85.5 MASTEC~ Breast~ Modera~ NO      LumA         1 Positve Positi~
##  7           45.4 BREAST~ Breast~ High    YES     LumB         1 Positve Positi~
##  8           61.5 BREAST~ Breast~ High    NO      LumB         1 Positve Positi~
##  9           68.7 MASTEC~ Breast~ Low     YES     Basal        1 Negati~ Negati~
## 10           46.9 MASTEC~ Breast~ Modera~ NO      Normal       1 Positve Positi~
## # ... with 1,082 more rows, 25 more variables:
## #   ‘Neoplasm Histologic Grade‘ <dbl>, ‘HER2 status measured by SNP6‘ <chr>,
## #   ‘HER2 Status‘ <chr>, ‘Tumor Other Histologic Subtype‘ <chr>,
## #   ‘Hormone Therapy‘ <chr>, ‘Inferred Menopausal State‘ <chr>,
## #   ‘Integrative Cluster‘ <chr>, ‘Primary Tumor Laterality‘ <chr>,
## #   ‘Lymph nodes examined positive‘ <dbl>, ‘Mutation Count‘ <dbl>,
## #   ‘Nottingham prognostic index‘ <dbl>, ‘Oncotree Code‘ <chr>, ...
```

```
dim(df2)
```

```
## [1] 1092   34
```

```
is_categorical <- sapply(df2, is.character)
is_categorical
```

```
##               Age at Diagnosis         Type of Breast Surgery
##                          FALSE                           TRUE
##           Cancer Type Detailed                    Cellularity
##                           TRUE                           TRUE
##                   Chemotherapy    Pam50 + Claudin-low subtype
##                           TRUE                           TRUE
##                         Cohort       ER status measured by IHC
##                          FALSE                           TRUE
##                      ER Status      Neoplasm Histologic Grade
##                           TRUE                          FALSE
##   HER2 status measured by SNP6                    HER2 Status
##                           TRUE                           TRUE
## Tumor Other Histologic Subtype                Hormone Therapy
##                           TRUE                           TRUE
##      Inferred Menopausal State            Integrative Cluster
##                           TRUE                           TRUE
##       Primary Tumor Laterality  Lymph nodes examined positive
##                           TRUE                          FALSE
##                 Mutation Count    Nottingham prognostic index
##                          FALSE                          FALSE
##                  Oncotree Code      Overall Survival (Months)
##                           TRUE                          FALSE
##                      PR Status                  Radio Therapy
##                           TRUE                           TRUE
##   Relapse Free Status (Months)            Relapse Free Status
##                          FALSE                           TRUE
##  Number of Samples Per Patient                    Sample Type
##                          FALSE                           TRUE
##                            Sex      3-Gene classifier subtype
##                           TRUE                           TRUE
##            TMB (nonsynonymous)                     Tumor Size
##                          FALSE                          FALSE
##                    Tumor Stage         Patient’s Vital Status
##                          FALSE                           TRUE
```

Create Dummy Variable for Categorical Variable

```
df2$Type_of_Breast_Surgery = ifelse(df2$`Type of Breast Surgery` =="BREAST CONSERVING",1,0)
df2$new_Chemotherapy = ifelse(df2$Chemotherapy =="NO",0,1)
for (i in 1:nrow(df2)) {
  if (df2[i,]$Cellularity == "High") {
    df2$new_Cellularity[i] <- 3
  }
  else if (df2[i,]$Cellularity == "Moderate") {
    df2$new_Cellularity[i] <- 2
  }
  else {
    df2$new_Cellularity[i] <- 1
  }
}

df2$Pam50_Claudin_low_subtype_Luma = ifelse(df2$`Pam50 + Claudin-low subtype` == "LumA",1,0)
df2$Pam50_Claudin_low_subtype_LumB = ifelse(df2$`Pam50 + Claudin-low subtype` == "LumB",1,0)
df2$Pam50_Claudin_low_subtype_Her2 = ifelse(df2$`Pam50 + Claudin-low subtype` == "Her2",1,0)
df2$Pam50_Claudin_low_subtype_Basal = ifelse(df2$`Pam50 + Claudin-low subtype` == "Basal",1,0)
df2$Pam50_Claudin_low_subtype_Normal = ifelse(df2$`Pam50 + Claudin-low subtype` == "Normal",1,0)
```

```
df2$Pam50_Claudin_low_subtype_claudin = ifelse(df2$`Pam50 + Claudin-low subtype` == "claudin-low",1,0)
df2$Pam50_Claudin_low_subtype_NC = ifelse(df2$`Pam50 + Claudin-low subtype` == "NC",1,0)


df2$ER_status_measured_by_IHC = ifelse(df2$`ER status measured by IHC` =="Positve",1,0)
df2$ER_Status_Positive = ifelse(df2$`ER Status` =="Positive",1,0)

for (i in 1:nrow(df2)) {
  if (df2[i,]$`HER2 status measured by SNP6` == "NEUTRAL") {
    df2$HER2_status_measured_by_SNP6[i] <- 4
  }
  else if (df2[i,]$`HER2 status measured by SNP6` == "GAIN") {
    df2$HER2_status_measured_by_SNP6[i] <- 3
  }
  else if (df2[i,]$`HER2 status measured by SNP6` == "LOSS"){
    df2$HER2_status_measured_by_SNP6[i] <- 2
  }
  else if (df2[i,]$`HER2 status measured by SNP6` == "UNDEF"){
    df2$HER2_status_measured_by_SNP6[i] <- 1
  }
}
df2$HER2_Status_Positive = ifelse(df2$`HER2 Status` =="Positive",1,0)
df2$Tumor_Other_Histologic_Subtype_Ductal = ifelse(df2$`Tumor Other Histologic Subtype` == "Ductal/NST"
df2$Tumor_Other_Histologic_Subtype_Tubular = ifelse(df2$`Tumor Other Histologic Subtype` == "Tubular/ c
df2$Tumor_Other_Histologic_Subtype_Medullary = ifelse(df2$`Tumor Other Histologic Subtype` == "Medullar

df2$Hormone_Therapy = ifelse(df2$`Hormone Therapy` =="NO",0,1)
df2$Inferred_Menopausal_State = ifelse(df2$`Inferred Menopausal State` =="Pre",0,1)
df2$Primary_Tumor_Laterality = ifelse(df2$`Primary Tumor Laterality` =="Right",0,1)
df2$PR_Status = ifelse(df2$`PR Status` =="Positive",1,0)
df2$Radio_Therapy = ifelse(df2$`Radio Therapy` =="NO",0,1)
df2$Relapse_Free_Status = ifelse(df2$`Relapse Free Status` =="0:Not Recurred",0,1)

df2$Gene_classifier_subtype_ERH = ifelse(df2$`3-Gene classifier subtype` == "ER+/HER2- High Prolif",1,0
df2$Gene_classifier_subtype_ERL = ifelse(df2$`3-Gene classifier subtype` == "ER+/HER2- Low Prolif",1,0)
df2$Gene_classifier_subtype_ERM = ifelse(df2$`3-Gene classifier subtype` == "ER-/HER2-",1,0)
df2$Gene_classifier_subtype_ERP = ifelse(df2$`3-Gene classifier subtype` == "HER2+",1,0)

df2$Overall_Survival_Status = ifelse(df2$`Patient's Vital Status` == "Living",1,0)

names(df2)[names(df2) == "Cancer Type Detailed"] <- "Cancer_Type_Detailed"
df2<- df2[df2$Cancer_Type_Detailed == "Breast Invasive Ductal Carcinoma",]


#df2
```

Create Dummy Variable for Numeric Variable

```
names(df2)[names(df2) == "Age at Diagnosis"] <- "Age"
names(df2)[names(df2) == "Neoplasm Histologic Grade"] <- "Neo_Grade"
names(df2)[names(df2) == "Lymph nodes examined positive"] <- "Lymph"
names(df2)[names(df2) == "Mutation Count"] <- "Mutation"
names(df2)[names(df2) == "Nottingham prognostic index"] <- "Nottingham"
```

```r
names(df2)[names(df2) == "Overall Survival (Months)"] <- "Overall_Month"
names(df2)[names(df2) == "Relapse Free Status (Months)"] <- "Relapse_Month"
names(df2)[names(df2) == "TMB (nonsynonymous)"] <- "TMB"
names(df2)[names(df2) == "Tumor Size"] <- "Tumor_Size"
names(df2)[names(df2) == "Tumor Stage"] <- "Tumor_Stage"
df2
```

```
## # A tibble: 859 x 62
##      Age Type o~1 Cance~2 Cellu~3 Chemo~4 Pam50~5 Cohort ER st~6 ER St~7 Neo_G~8
##    <dbl> <chr>    <chr>   <chr>   <chr>   <chr>    <dbl> <chr>   <chr>     <dbl>
##  1  43.2 BREAST ~ Breast~ High    NO      LumA         1 Positve Positi~       3
##  2  78.8 MASTECT~ Breast~ Modera~ NO      LumB         1 Positve Positi~       3
##  3  86.4 BREAST ~ Breast~ Modera~ NO      LumB         1 Positve Positi~       3
##  4  85.5 MASTECT~ Breast~ Modera~ NO      LumA         1 Positve Positi~       2
##  5  45.4 BREAST ~ Breast~ High    YES     LumB         1 Positve Positi~       3
##  6  61.5 BREAST ~ Breast~ High    NO      LumB         1 Positve Positi~       2
##  7  68.7 MASTECT~ Breast~ Low     YES     Basal        1 Negati~ Negati~       3
##  8  49.9 MASTECT~ Breast~ Modera~ YES     LumA         1 Positve Positi~       1
##  9  54.2 MASTECT~ Breast~ High    NO      LumA         1 Positve Positi~       1
## 10  48.6 MASTECT~ Breast~ Low     NO      LumA         1 Positve Positi~       2
## # ... with 849 more rows, 52 more variables:
## #   ‘HER2 status measured by SNP6‘ <chr>, ‘HER2 Status‘ <chr>,
## #   ‘Tumor Other Histologic Subtype‘ <chr>, ‘Hormone Therapy‘ <chr>,
## #   ‘Inferred Menopausal State‘ <chr>, ‘Integrative Cluster‘ <chr>,
## #   ‘Primary Tumor Laterality‘ <chr>, Lymph <dbl>, Mutation <dbl>,
## #   Nottingham <dbl>, ‘Oncotree Code‘ <chr>, Overall_Month <dbl>,
## #   ‘PR Status‘ <chr>, ‘Radio Therapy‘ <chr>, Relapse_Month <dbl>, ...
```

```r
df3 <- df2[,-c(2,3,4:6,8,9,11:17,21,23, 24, 26,27:30, 34)]
df3
```

```
## # A tibble: 859 x 39
##      Age Cohort Neo_Grade Lymph Mutation Notting~1 Overa~2 Relap~3   TMB Tumor~4
##    <dbl>  <dbl>     <dbl> <dbl>    <dbl>     <dbl>   <dbl>   <dbl> <dbl>   <dbl>
##  1  43.2      1         3     0        2      4.02    84.6    83.5   2.62      10
##  2  78.8      1         3     0        4      4.06     7.8     2.89  5.23      31
##  3  86.4      1         3     1        4      5.03    36.6    36.1   5.23      16
##  4  85.5      1         2     0        1      3.04   132.    123.    1.31      22
##  5  45.4      1         3     0        5      4.05   141.    139.    6.54      23
##  6  61.5      1         2     1        3      4.03   157.    155.    3.92      16
##  7  68.7      1         3     0        1      4.08     8.07    7.83  1.31      39
##  8  49.9      1         1     5        4      4.14    85.3    84.2   5.23      70
##  9  54.2      1         1     0        4      2.05   127.    125.    5.23      27
## 10  48.6      1         2     0        3      3.06    13.4    13.2   3.92      30
## # ... with 849 more rows, 29 more variables: Tumor_Stage <dbl>,
## #   Type_of_Breast_Surgery <dbl>, new_Chemotherapy <dbl>,
## #   new_Cellularity <dbl>, Pam50_Claudin_low_subtype_Luma <dbl>,
## #   Pam50_Claudin_low_subtype_LumB <dbl>, Pam50_Claudin_low_subtype_Her2 <dbl>,
## #   Pam50_Claudin_low_subtype_Basal <dbl>,
## #   Pam50_Claudin_low_subtype_Normal <dbl>,
## #   Pam50_Claudin_low_subtype_claudin <dbl>, ...
```

```
pander(summary(df3),caption='Descriptive Statistics of The Data')
```

Table 1: Descriptive Statistics of The Data (continued below)

Age	Cohort	Neo_Grade	Lymph
Min. :26.36	Min. :1.000	Min. :1.000	Min. : 0.000
1st Qu.:49.99	1st Qu.:1.000	1st Qu.:2.000	1st Qu.: 0.000
Median :60.62	Median :2.000	Median :3.000	Median : 0.000
Mean :60.05	Mean :2.191	Mean :2.517	Mean : 1.916
3rd Qu.:69.75	3rd Qu.:3.000	3rd Qu.:3.000	3rd Qu.: 2.000
Max. :96.29	Max. :5.000	Max. :3.000	Max. :41.000

Table 2: Table continues below

Mutation	Nottingham	Overall_Month	Relapse_Month
Min. : 1.000	Min. :2.018	Min. : 0.10	Min. : 0.10
1st Qu.: 3.000	1st Qu.:3.080	1st Qu.: 58.05	1st Qu.: 40.10
Median : 5.000	Median :4.050	Median :115.30	Median : 98.42
Mean : 5.423	Mean :4.216	Mean :124.05	Mean :109.35
3rd Qu.: 7.000	3rd Qu.:5.050	3rd Qu.:186.32	3rd Qu.:172.12
Max. :46.000	Max. :6.360	Max. :337.03	Max. :296.91

Table 3: Table continues below

TMB	Tumor_Size	Tumor_Stage	Type_of_Breast_Surgery
Min. : 1.308	Min. : 1.0	Min. :1.000	Min. :0.0000
1st Qu.: 3.923	1st Qu.: 17.0	1st Qu.:1.000	1st Qu.:0.0000
Median : 6.538	Median : 22.0	Median :2.000	Median :0.0000
Mean : 7.072	Mean : 25.7	Mean :1.767	Mean :0.4319
3rd Qu.: 9.153	3rd Qu.: 30.0	3rd Qu.:2.000	3rd Qu.:1.0000
Max. :60.146	Max. :180.0	Max. :4.000	Max. :1.0000

Table 4: Table continues below

new_Chemotherapy	new_Cellularity	Pam50_Claudin_low_subtype_Luma
Min. :0.0000	Min. :1.000	Min. :0.0000
1st Qu.:0.0000	1st Qu.:2.000	1st Qu.:0.0000
Median :0.0000	Median :3.000	Median :0.0000
Mean :0.2538	Mean :2.421	Mean :0.3667
3rd Qu.:1.0000	3rd Qu.:3.000	3rd Qu.:1.0000
Max. :1.0000	Max. :3.000	Max. :1.0000

Table 5: Table continues below

Pam50_Claudin_low_subtype_LumB	Pam50_Claudin_low_subtype_Her2
Min. :0.0000	Min. :0.0000
1st Qu.:0.0000	1st Qu.:0.0000
Median :0.0000	Median :0.0000
Mean :0.2608	Mean :0.1036
3rd Qu.:1.0000	3rd Qu.:0.0000
Max. :1.0000	Max. :1.0000

Table 6: Table continues below

Pam50_Claudin_low_subtype_Basal	Pam50_Claudin_low_subtype_Normal
Min. :0.0000	Min. :0.00000
1st Qu.:0.0000	1st Qu.:0.00000
Median :0.0000	Median :0.00000
Mean :0.1141	Mean :0.05122
3rd Qu.:0.0000	3rd Qu.:0.00000
Max. :1.0000	Max. :1.00000

Table 7: Table continues below

Pam50_Claudin_low_subtype_claudin	Pam50_Claudin_low_subtype_NC
Min. :0.0000	Min. :0.000000
1st Qu.:0.0000	1st Qu.:0.000000
Median :0.0000	Median :0.000000
Mean :0.1013	Mean :0.002328
3rd Qu.:0.0000	3rd Qu.:0.000000
Max. :1.0000	Max. :1.000000

Table 8: Table continues below

ER_status_measured_by_IHC	ER_Status_Positive	HER2_status_measured_by_SNP6
Min. :0.0000	Min. :0.0000	Min. :1.000
1st Qu.:0.0000	1st Qu.:0.0000	1st Qu.:3.000
Median :1.0000	Median :1.0000	Median :4.000
Mean :0.7404	Mean :0.7404	Mean :3.631
3rd Qu.:1.0000	3rd Qu.:1.0000	3rd Qu.:4.000
Max. :1.0000	Max. :1.0000	Max. :4.000

Table 9: Table continues below

HER2_Status_Positive	Tumor_Other_Histologic_Subtype_Ductal
Min. :0.0000	Min. :0.0000
1st Qu.:0.0000	1st Qu.:1.0000
Median :0.0000	Median :1.0000
Mean :0.1444	Mean :0.9674

HER2_Status_Positive	Tumor_Other_Histologic_Subtype_Ductal
3rd Qu.:0.0000	3rd Qu.:1.0000
Max. :1.0000	Max. :1.0000

Table 10: Table continues below

Tumor_Other_Histologic_Subtype_Tubular
Min. :0.00000
1st Qu.:0.00000
Median :0.00000
Mean :0.01746
3rd Qu.:0.00000
Max. :1.00000

Table 11: Table continues below

Tumor_Other_Histologic_Subtype_Medullary	Hormone_Therapy
Min. :0.00000	Min. :0.0000
1st Qu.:0.00000	1st Qu.:0.0000
Median :0.00000	Median :1.0000
Mean :0.01513	Mean :0.6042
3rd Qu.:0.00000	3rd Qu.:1.0000
Max. :1.00000	Max. :1.0000

Table 12: Table continues below

Inferred_Menopausal_State	Primary_Tumor_Laterality	PR_Status
Min. :0.0000	Min. :0.000	Min. :0.0000
1st Qu.:0.5000	1st Qu.:0.000	1st Qu.:0.0000
Median :1.0000	Median :1.000	Median :0.0000
Mean :0.7497	Mean :0.525	Mean :0.4994
3rd Qu.:1.0000	3rd Qu.:1.000	3rd Qu.:1.0000
Max. :1.0000	Max. :1.000	Max. :1.0000

Table 13: Table continues below

Radio_Therapy	Relapse_Free_Status	Gene_classifier_subtype_ERH
Min. :0.0000	Min. :0.0000	Min. :0.000
1st Qu.:0.0000	1st Qu.:0.0000	1st Qu.:0.000
Median :1.0000	Median :0.0000	Median :0.000
Mean :0.6799	Mean :0.4214	Mean :0.362
3rd Qu.:1.0000	3rd Qu.:1.0000	3rd Qu.:1.000
Max. :1.0000	Max. :1.0000	Max. :1.000

Table 14: Table continues below

Gene_classifier_subtype_ERL	Gene_classifier_subtype_ERM	Gene_classifier_subtype_ERP	Overall_Survival_Status
Min. :0.0000	Min. :0.0000	Min. :0.0000	Min. :0.0000
1st Qu.:0.0000	1st Qu.:0.0000	1st Qu.:0.0000	1st Qu.:0.0000
Median :0.0000	Median :0.0000	Median :0.0000	Median :0.0000
Mean :0.3132	Mean :0.1921	Mean :0.1327	Mean :0.4435
3rd Qu.:1.0000	3rd Qu.:0.0000	3rd Qu.:0.0000	3rd Qu.:1.0000
Max. :1.0000	Max. :1.0000	Max. :1.0000	Max. :1.0000

```
pander(head(df3),caption='Head of data selection')
```

Table 16: Head of data selection (continued below)

Age	Cohort	Neo_Grade	Lymph	Mutation	Nottingham	Overall_Month
43.19	1	3	0	2	4.02	84.63
78.77	1	3	0	4	4.062	7.8
86.41	1	3	1	4	5.032	36.57
85.49	1	2	0	1	3.044	132
45.43	1	3	0	5	4.046	140.9
61.49	1	2	1	3	4.032	157.4

Table 17: Table continues below

Relapse_Month	TMB	Tumor_Size	Tumor_Stage	Type_of_Breast_Surgery
83.52	2.615	10	1	1
2.89	5.23	31	4	0
36.09	5.23	16	2	1
123.3	1.308	22	4	0
139	6.538	23	2	1
155.4	3.923	16	2	1

Table 18: Table continues below

new_Chemotherapy	new_Cellularity	Pam50_Claudin_low_subtype_Luma
0	3	1
0	2	0
0	2	0

new_Chemotherapy	new_Cellularity	Pam50_Claudin_low_subtype_Luma
0	2	1
1	3	0
0	3	0

Table 19: Table continues below

Pam50_Claudin_low_subtype_LumB	Pam50_Claudin_low_subtype_Her2
0	0
1	0
1	0
0	0
1	0
1	0

Table 20: Table continues below

Pam50_Claudin_low_subtype_Basal	Pam50_Claudin_low_subtype_Normal
0	0
0	0
0	0
0	0
0	0
0	0

Table 21: Table continues below

Pam50_Claudin_low_subtype_claudin	Pam50_Claudin_low_subtype_NC
0	0
0	0
0	0
0	0
0	0
0	0

Table 22: Table continues below

ER_status_measured_by_IHC	ER_Status_Positive	HER2_status_measured_by_SNP6
1	1	4
1	1	4
1	1	3
1	1	4
1	1	4
1	1	4

Table 23: Table continues below

HER2__Status__Positive	Tumor__Other__Histologic__Subtype__Ductal
0	1
0	1
0	1
0	1
0	1
0	1

Table 24: Table continues below

Tumor__Other__Histologic__Subtype__Tubular
0
0
0
0
0
0

Table 25: Table continues below

Tumor__Other__Histologic__Subtype__Medullary	Hormone__Therapy
0	1
0	1
0	1
0	1
0	1
0	1

Table 26: Table continues below

Inferred__Menopausal__State	Primary__Tumor__Laterality	PR__Status
0	0	1
1	1	1
1	0	0
1	1	1
0	0	1
1	1	1

Table 27: Table continues below

Radio__Therapy	Relapse__Free__Status	Gene__classifier__subtype__ERH
1	0	1
1	1	1
1	0	1
1	1	0

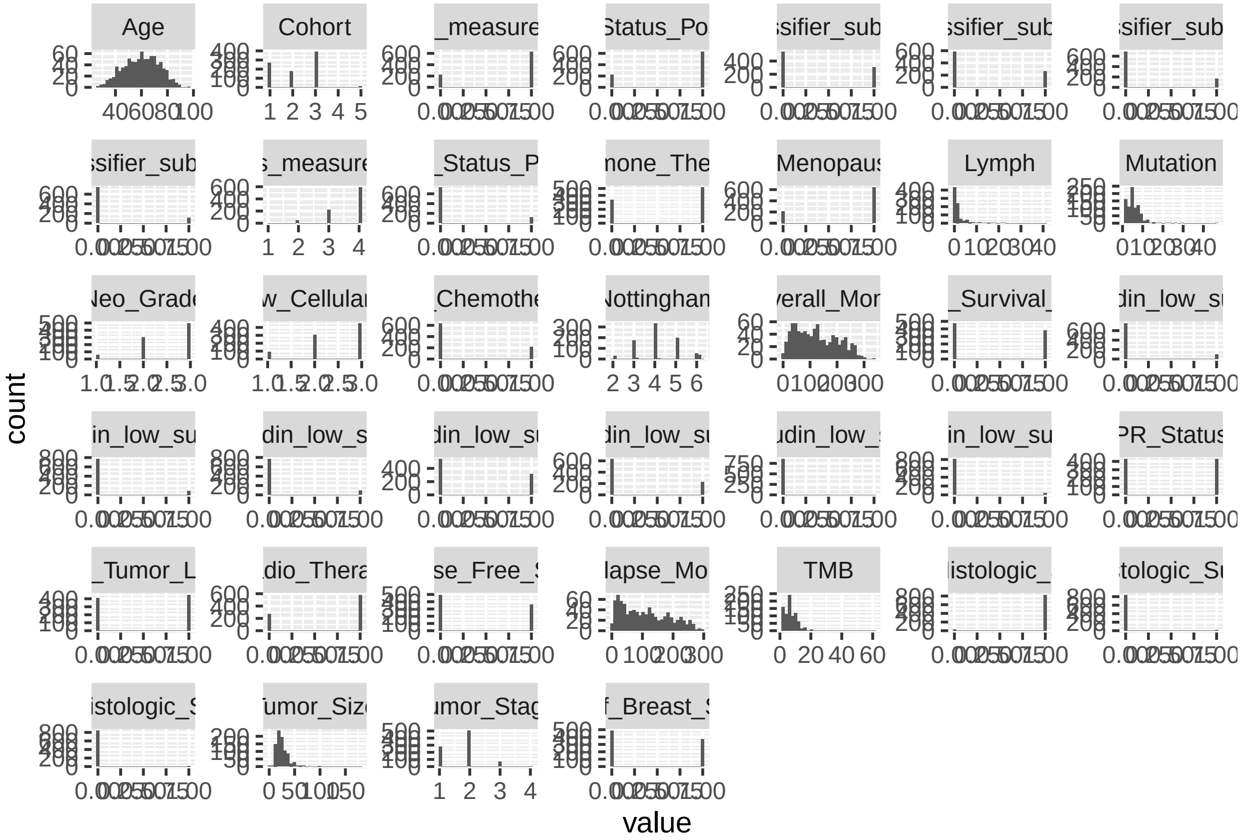
Radio_Therapy	Relapse_Free_Status	Gene_classifier_subtype_ERH
1	0	1
1	0	1

Table 28: Table continues below

Gene_classifier_subtype_ERL	Gene_classifier_subtype_ERM
0	0
0	0
0	0
1	0
0	0
0	0

Gene_classifier_subtype_ERP	Overall_Survival_Status
0	1
0	0
0	0
0	0
0	1
0	1

```r
df3 %>%
  keep(is.numeric) %>%
  gather() %>%
  ggplot(aes(value)) +
  facet_wrap(~ key, scales = "free") +
  geom_histogram()
```



```r
tr_ind <- 1:(nrow(df3) * 0.8)
df3_tall <- df3[tr_ind, ]
nrow(df3_tall)
```

```
## [1] 687
```

```r
df3_te <- df3[-tr_ind, ]
nrow(df3_te)
```

```
## [1] 172
```

```r
tr_ind2 <- 1:(nrow(df3_tall) * 0.8)
df3_tr <- df3_tall[tr_ind2, ]
nrow(df3_tr)
```

```
## [1] 549
```

```r
df3_va <- df3_tall[-tr_ind2, ]
nrow(df3_va)
```

```
## [1] 138
```

variable selection

```
set.seed(0)
fit_BIC <- regsubsets(Overall_Survival_Status ~ ., data = df3_tr)
```

```
## Reordering variables and trying again:
```

```
summary_BIC <- summary(fit_BIC)
min_BIC <- which.min(summary_BIC$bic)
min_BIC
```

```
## [1] 6
```

```
coef_BIC = coef(fit_BIC,min_BIC)
coef_BIC
```

```
##                 (Intercept)                         Age
##                 0.899883923                -0.008044691
##                      Cohort               Overall_Month
##                -0.226179447                 0.001480852
##               Relapse_Month      Type_of_Breast_Surgery
##                 0.001660879                 0.112900365
## Gene_classifier_subtype_ERL
##                 0.112044557
```

```
formula1 <- Overall_Survival_Status ~ Age + Cohort + Overall_Month + Relapse_Month + Type_of_Breast_Surg
```

Backward Stepwise Selection with Cp

```
fit_BACKWARD <- regsubsets(Overall_Survival_Status ~ ., data = df3_tr, method = "backward", nvmax = ncol
```

```
## Reordering variables and trying again:
```

```
summary_BACKWARD <- summary(fit_BACKWARD)
mix_BACKWARD <- which.min(summary_BACKWARD$cp)
mix_BACKWARD
```

```
## [1] 14
```

```
coef_BACKWARD = coef(fit_BACKWARD, mix_BACKWARD)
coef_BACKWARD
```

```
##                       (Intercept)                               Age
##                       0.945854103                      -0.009104305
##                            Cohort                             Lymph
##                      -0.245788977                      -0.011378779
##                     Overall_Month                     Relapse_Month
##                       0.001412848                       0.001588274
##                       Tumor_Stage            Type_of_Breast_Surgery
##                       0.063561136                       0.097766195
##                  new_Chemotherapy    Pam50_Claudin_low_subtype_Luma
##                      -0.088293305                       0.100704550
##   Pam50_Claudin_low_subtype_Basal Pam50_Claudin_low_subtype_claudin
##                       0.096864829                       0.085244210
##              HER2_Status_Positive       Gene_classifier_subtype_ERH
##                      -0.020653504                      -0.043335000
##       Gene_classifier_subtype_ERL
##                       0.025591333
```

```
formula2 <- Overall_Survival_Status ~ Age + Cohort + Lymph + Overall_Month + Relapse_Month + Tumor_Stag
```

```
x_tr<-as.matrix(df3_tr[,c(1:ncol(df3_tr) - 1)])
y_tr<-as.matrix(df3_tr[,ncol(df3_tr)])
x_te<-as.matrix(df3_te[,c(1:ncol(df3_tr) - 1)])
y_te<-as.matrix(df3_te[,ncol(df3_tr)])
```

#Elastic Net

```
library(glmnet)
set.seed(0)
error_lasso_array <- c()
lasso_cv_array <- c()
for (i in 3:10) {
  for (j in seq(0, 1, 0.05)){
    lasso_cv<-cv.glmnet(x=x_tr,y=y_tr,type.measure='mse',nfold=i,alpha=j,keep=T)
    result_la<-predict(lasso_cv,newx=x_te,interval='prediction')
    error_lasso <- (err_la<-mean((y_te-result_la)^2))
    lasso_cv_array[ (i-3)*20 + j * 20] <- lasso_cv
    error_lasso_array[ (i-3)*20 + j * 20] <- error_lasso
  }
}
minlasso <- which.min(error_lasso_array)
minlasso
```

```
## [1] 14
```

```
lasso_cv <- cv.glmnet(x=x_tr,y=y_tr,type.measure='mse',nfold=3,alpha=0.7,keep=T)
lasso_cv$lambda.min
```

```
## [1] 0.006111171
```

```
coef_lasso = coef(lasso_cv, s=lasso_cv$lambda.min)
coef_lasso
```

```
## 39 x 1 sparse Matrix of class "dgCMatrix"
##                                                    s1
## (Intercept)                               1.313318746
## Age                                      -0.009443496
## Cohort                                   -0.171038476
## Neo_Grade                                 .
## Lymph                                    -0.007365011
## Mutation                                  .
## Nottingham                                .
## Overall_Month                             0.003429122
## Relapse_Month                            -0.001817453
## TMB                                       .
## Tumor_Size                                .
## Tumor_Stage                               0.036633042
## Type_of_Breast_Surgery                    0.048615669
## new_Chemotherapy                         -0.029287268
## new_Cellularity                           .
## Pam50_Claudin_low_subtype_Luma            0.016304892
## Pam50_Claudin_low_subtype_LumB           -0.031452007
## Pam50_Claudin_low_subtype_Her2           -0.099321914
## Pam50_Claudin_low_subtype_Basal           .
## Pam50_Claudin_low_subtype_Normal         -0.043951812
## Pam50_Claudin_low_subtype_claudin         0.007147511
## Pam50_Claudin_low_subtype_NC             -0.134127074
## ER_status_measured_by_IHC                 .
## ER_Status_Positive                        .
## HER2_status_measured_by_SNP6             -0.028333771
## HER2_Status_Positive                      .
## Tumor_Other_Histologic_Subtype_Ductal     .
## Tumor_Other_Histologic_Subtype_Tubular    .
## Tumor_Other_Histologic_Subtype_Medullary -0.125312918
## Hormone_Therapy                           0.020037384
## Inferred_Menopausal_State                 0.016000287
## Primary_Tumor_Laterality                 -0.023472828
## PR_Status                                 0.002936933
## Radio_Therapy                             0.075721202
## Relapse_Free_Status                      -0.480242982
## Gene_classifier_subtype_ERH              -0.036680086
## Gene_classifier_subtype_ERL               .
## Gene_classifier_subtype_ERM               .
## Gene_classifier_subtype_ERP               .
```

```
formula3 <- Overall_Survival_Status ~ Age + Cohort + Lymph + Overall_Month + Relapse_Month + Tumor_Stag
```

```
library(pROC)
set.seed(0)
predict_BIC = glm(formula1,df3_tr,family = "binomial")
pred_BIC = round(predict(predict_BIC,df3_te,type = "response"))
roc_bic <- roc(df3_te$Overall_Survival_Status,pred_BIC,smooth=F)
auc(roc_bic)
```

```
## Area under the curve: 0.5896
```

```
predict_BACKWARD  = glm(formula2,df3_tr,family = "binomial")
pred_BACKWARD = round(predict(predict_BACKWARD,df3_te,type = "response"))
roc_back <- roc(df3_te$Overall_Survival_Status,pred_BACKWARD,smooth=F)
auc(roc_back)
```

```
## Area under the curve: 0.5597
```

```
predict_net = glm(formula3,df3_tr,family = "binomial")
pred_net = round(predict(predict_net,df3_te,type = "response"))
roc_net <- roc(df3_te$Overall_Survival_Status,pred_net,smooth=F)
auc(roc_net)
```

```
## Area under the curve: 0.6866
```

```
which.max(data.frame(auc(roc_bic),  auc(roc_back), auc(roc_net))) # All the errors listed
```

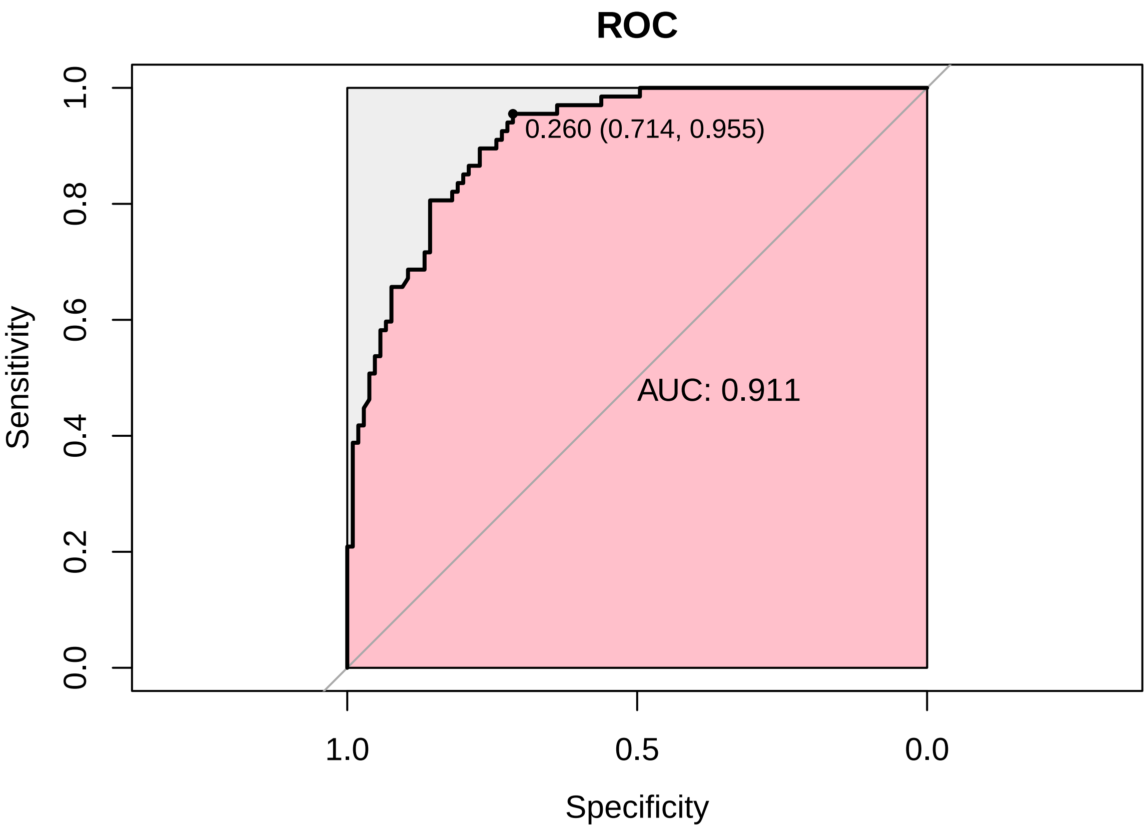
```
## auc.roc_net.
##            3
```

cross validation #Random Forest

```
library(randomForest)
set.seed(0)
K <- 10
n_all <- nrow(df3_tr)
n_all2 <- nrow(df3_va)
fold_auc_rf<-as.numeric()
auc_all <- c()
fold_ind <- sample(1:K, n_all, replace = TRUE)
fold_ind2 <- sample(1:K, n_all2, replace = TRUE)

for (i in c(10,100,10)) {
  for (j in 1:K) {
    rf_model <- randomForest(formula3, data = df3_tr[fold_ind != j, ], ntree = i, importance = T)
    pred_prob <- predict(rf_model, newdata = df3_va[fold_ind2 == j, ], type = "response")
    pred_label <- ifelse(pred_prob > 0.5, 1, 0)
    roc_rf <- roc(df3_va[fold_ind2 == j, ]$Overall_Survival_Status,pred_label,smooth=F)
    auc_all[((i/10 -1)*10) + j] <- auc(roc_rf)
  }
}
which.max(auc_all)
```

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

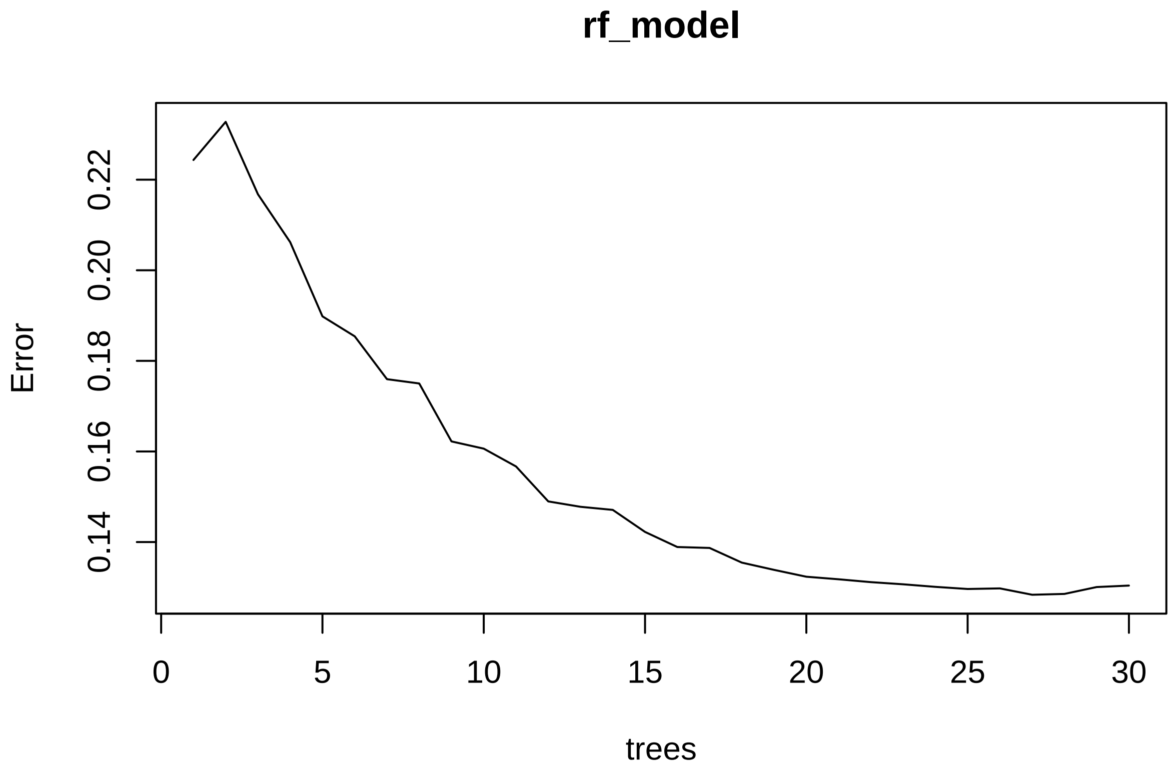
```
rf_model <- randomForest(formula3, data = df3_tr, ntree = 30, importance = T)
pred_prob <- predict(rf_model, newdata = df3_te, type = "response")
rf_pred<-as.character(pred_prob)
rf_pred<-as.numeric(pred_prob)
rf_roc<-roc(df3_te$Overall_Survival_Status,rf_pred,smooth=F)

plot(rf_roc, auc.polygon=T, auc.polygon.col='pink', smooth=F,print.auc=T, max.auc.polygon=T,print.thres
```



```
rf_auc <- auc(rf_roc)
rf_auc
```

```
## Area under the curve: 0.9112
```

```
plot(rf_model)
```



#KNN

```
set.seed(0)
K <- 10
n_all <- nrow(df3_tr)
n_all2 <- nrow(df3_va)
fold_auc_rf<-as.numeric()
auc_all3 <- c()
fold_ind <- sample(1:K, n_all, replace = TRUE)
fold_ind2 <- sample(1:K, n_all2, replace = TRUE)

table(df3$Overall_Survival_Status)
```

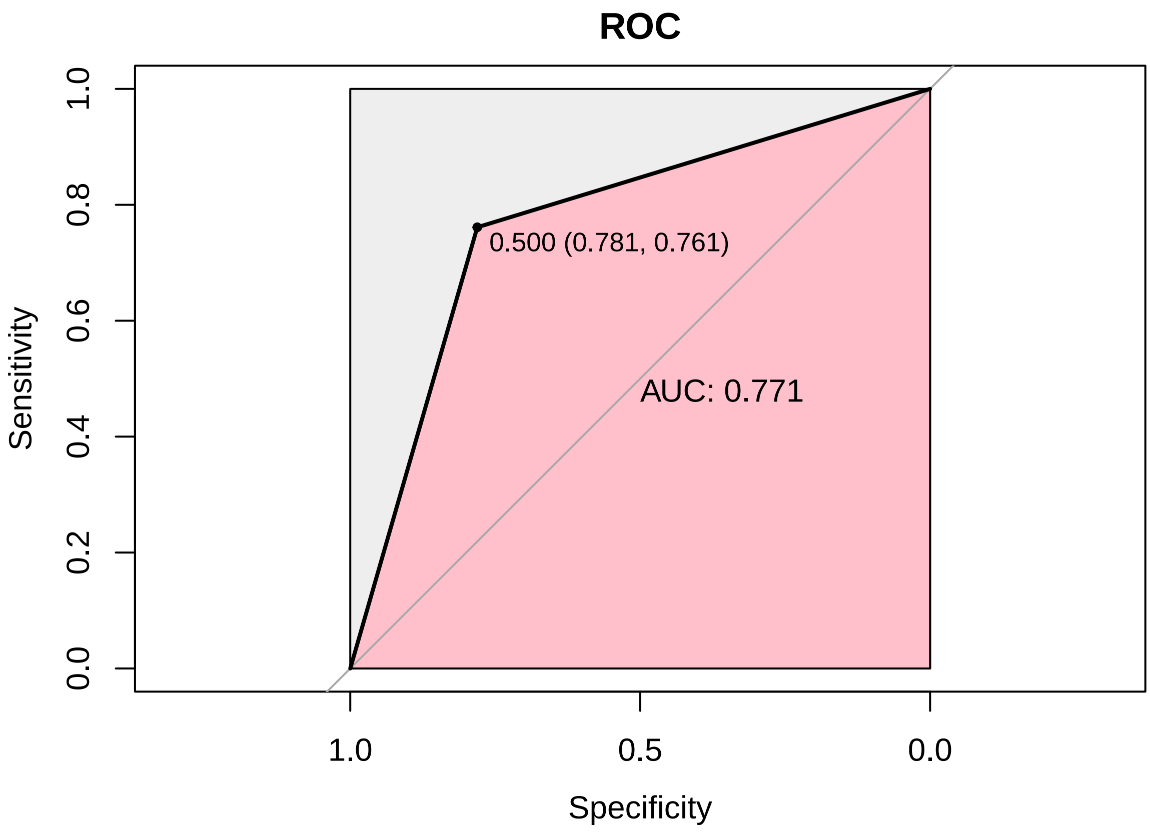
```
##
##   0   1
## 478 381
```

```
for (j in 2:K) {
  for (i in 5:20) {
     knn_model <- knn3(formula3 <- Overall_Survival_Status ~ Age + Cohort + Lymph + Overall_Month + Rela
    knn_prob <- predict(knn_model, newdata = df3_va[fold_ind2 == j, ])
    pred_label <- ifelse(knn_prob > 0.5, 1, 0)
    roc_rf <- roc(df3_va[fold_ind2 == j, ]$Overall_Survival_Status,pred_label[,2],smooth=F)
    auc_all3[((j-2)*15) + i - 4] <- auc(roc_rf)
  }
}
```

```
which.max(auc_all3)
```

```
## [1] 126
```

```
knn_model <- knn3(formula3 <- Overall_Survival_Status ~ Age + Cohort + Lymph + Overall_Month + Relapse_
knn_prob <- predict(knn_model, newdata = df3_te)
pred_label <- ifelse(knn_prob > 0.5, 1, 0)
knn_roc <- roc(df3_te$Overall_Survival_Status,pred_label[,2],smooth=F)
plot(knn_roc, auc.polygon=T, auc.polygon.col='pink', smooth=F,print.auc=T, max.auc.polygon=T,print.thres
```



```
knn_auc <- auc(knn_roc)
knn_auc
```

```
## Area under the curve: 0.7711
```

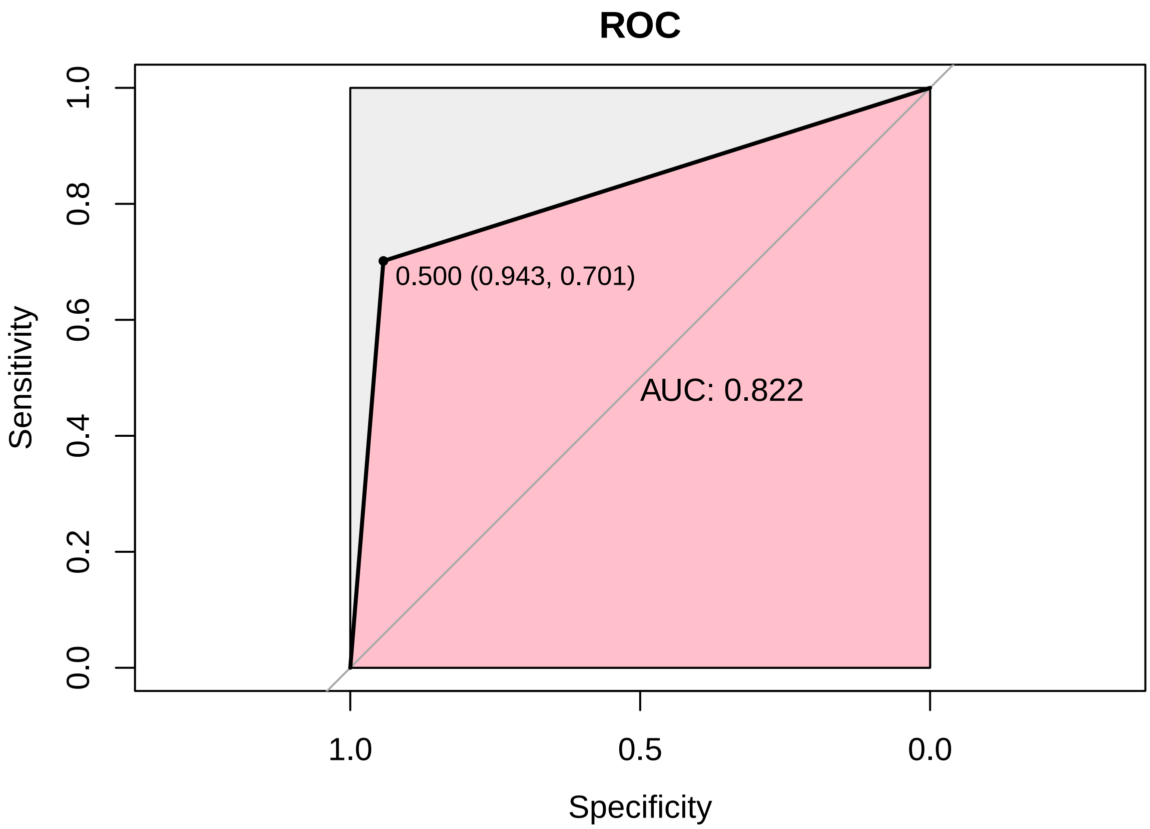
#Gradient Boosting Tree

```
set.seed(0)

K <- 10
n_all <- nrow(df3_tr)
n_all2 <- nrow(df3_va)
fold_auc_rf<-as.numeric()
```

```r
auc_all2 <- c()
fold_ind <- sample(1:K, n_all, replace = TRUE)
fold_ind2 <- sample(1:K, n_all2, replace = TRUE)

for (i in 3:10) {
  for (j in seq(0.01, 0.1, 0.005)) {
  gbm_cv<-gbm(formula3,data = df3_tr[fold_ind != i, ], distribution = "multinomial", shrinkage = j )
  result_gbm<-predict(gbm_cv,df3_va[fold_ind2 == i, ],type="response")
  pred_label <- ifelse(result_gbm > 0.5, 1, 0)
  lengthforpredict <- nrow(as.matrix(pred_label))
  lengthforpredict2 <- lengthforpredict / 2 + 1
  vector <- as.vector(pred_label)
  vector <- vector[lengthforpredict2 : lengthforpredict]
  roc_rf <- roc(df3_va[fold_ind2 == i, ]$Overall_Survival_Status, vector,smooth=F)
  auc_all2[((i - 3)*18) + j * 200] <- auc(roc_rf)
  }
}
which.max(auc_all2)
```

```
## [1] 8
```

```r
gbm_cv<-gbm(formula3,data = df3_tr, distribution = "multinomial", shrinkage =  0.04)
result_gbm<-predict(gbm_cv,df3_te,type="response")
pred_label <- ifelse(result_gbm > 0.5, 1, 0)
lengthforpredict <- nrow(as.matrix(pred_label))
lengthforpredict2 <- lengthforpredict / 2 + 1
vector <- as.vector(pred_label)
vector <- vector[lengthforpredict2 : lengthforpredict]
roc_gbt <- roc(df3_te$Overall_Survival_Status, vector,smooth=F)
plot(roc_gbt, auc.polygon=T, auc.polygon.col='pink', smooth=F,print.auc=T, max.auc.polygon=T,print.thres
```



```
gbt_auc <- auc(roc_gbt)
gbt_auc
```

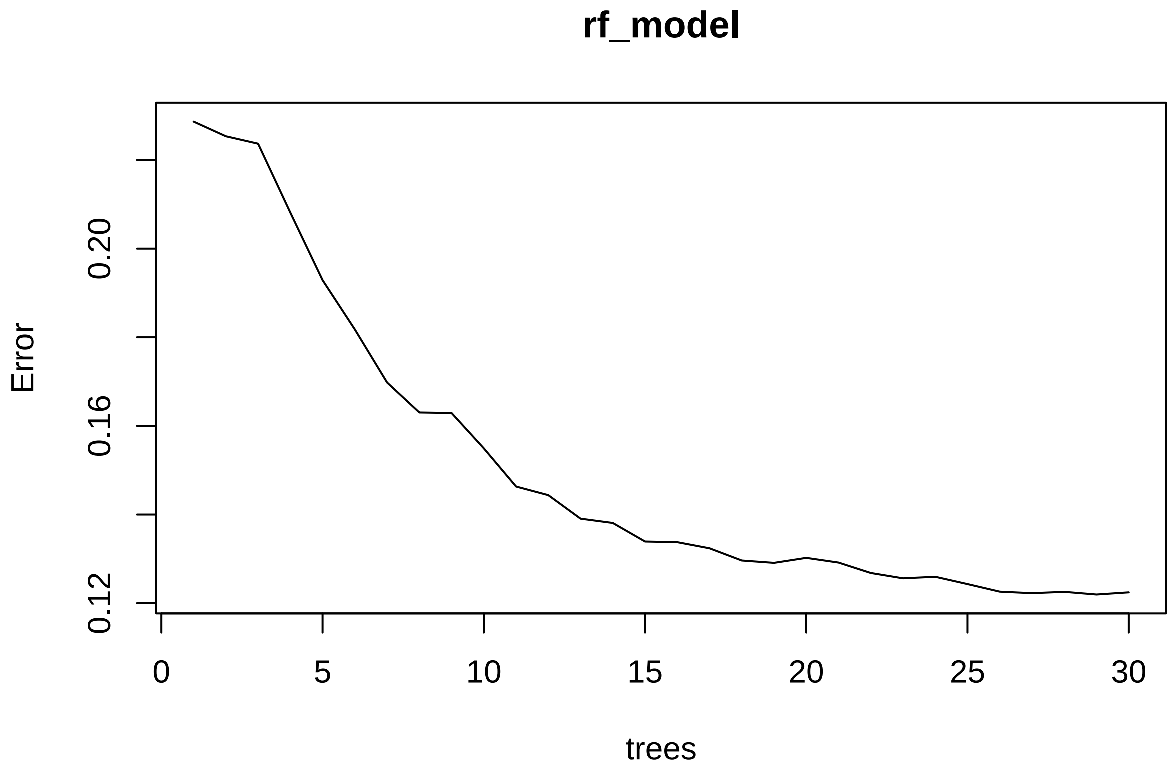
```
## Area under the curve: 0.8222
```

#Comparison

```
which.max(data.frame(rf_auc, knn_auc, gbt_auc))
```

```
## rf_auc
##      1
```

```
rf_model <- randomForest(formula3, data = df3_tr, ntree = 30, importance = T)
pred_prob <- predict(rf_model, newdata = df3_te, type = "response")
pred_label <- ifelse(pred_prob > 0.5, 1, 0)

plot(rf_model)
```



Model Present

```
#Accuracy
gbm_confusion<-table(df3_te$Overall_Survival_Status,pred_label,dnn=c('Actual','Predicted'))
gbm_confusion
```

```
##       Predicted
## Actual  0  1
##      0 97  8
##      1 17 50
```

```
lr_accuracy <- (gbm_confusion[1,1] + gbm_confusion[2,2]) / (gbm_confusion[1,1] + gbm_confusion[1,2] + g

lr_precision <- gbm_confusion[2,2] / (gbm_confusion[2,2] + gbm_confusion[1,2])

lr_recall <- gbm_confusion[2,2] / (gbm_confusion[2,2] + gbm_confusion[2,1])

lr_f1 <- 2/(1/lr_precision + 1/lr_recall)

lr_accuracy
```

```
## [1] 0.8546512
```

```
lr_precision
```

```
## [1] 0.862069
```

```
lr_recall
```

```
## [1] 0.7462687
```

```
lr_f1
```

```
## [1] 0.8
```