GPH 2338 Project

Yi 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) # For model selection
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)] # remove StudyID, Patient ID, Sample ID, Cancer Type
print(df2)
```

```
## # A tibble: 1,092 x 35
##    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, 26 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   35
```

```
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
##        Overall Survival Status                      PR Status
##                           TRUE                           TRUE
##                  Radio Therapy   Relapse Free Status (Months)
##                           TRUE                          FALSE
##            Relapse Free Status  Number of Samples Per Patient
##                           TRUE                          FALSE
##                    Sample Type                            Sex
##                           TRUE                           TRUE
##      3-Gene classifier subtype            TMB (nonsynonymous)
##                           TRUE                          FALSE
##                     Tumor Size                    Tumor Stage
##                          FALSE                          FALSE
##         Patient’s Vital Status
##                           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
  }
}
```

```
## Warning: Unknown or uninitialised column: ‘new_Cellularity‘.
```

```r
for (i in 1:nrow(df2)) {
  if (df2[i,]$`Pam50 + Claudin-low subtype` == "LumA") {
    df2$Pam50_Claudin_low_subtype[i] <- 1
  }
  else if (df2[i,]$`Pam50 + Claudin-low subtype` == "LumB") {
    df2$Pam50_Claudin_low_subtype[i] <- 2
  }
  else if (df2[i,]$`Pam50 + Claudin-low subtype` == "Her2"){
    df2$Pam50_Claudin_low_subtype[i] <- 3
  }
  else if (df2[i,]$`Pam50 + Claudin-low subtype` == "Basal"){
    df2$Pam50_Claudin_low_subtype[i] <- 4
  }
  else if (df2[i,]$`Pam50 + Claudin-low subtype` == "Normal"){
    df2$Pam50_Claudin_low_subtype[i] <- 5
  }
  else if (df2[i,]$`Pam50 + Claudin-low subtype` == "claudin-low"){
    df2$Pam50_Claudin_low_subtype[i] <- 6
  }
  else if (df2[i,]$`Pam50 + Claudin-low subtype` == "NC"){
    df2$Pam50_Claudin_low_subtype[i] <- 7
  }
}
```

```
## Warning: Unknown or uninitialised column: ‘Pam50_Claudin_low_subtype‘.
```

```r
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
  }
}
```

```
## Warning: Unknown or uninitialised column: ‘HER2_status_measured_by_SNP6‘.
```

```r
df2$HER2_Status_Positive = ifelse(df2$`HER2 Status` =="Positive",1,0)
for (i in 1:nrow(df2)) {
  if (df2[i,]$`Tumor Other Histologic Subtype` == "Ductal/NST") {
    df2$Tumor_Other_Histologic_Subtype[i] <- 1
  }
  else if (df2[i,]$`Tumor Other Histologic Subtype` == "Mixed") {
    df2$Tumor_Other_Histologic_Subtype[i] <- 2
  }
  else if (df2[i,]$`Tumor Other Histologic Subtype` == "Lobular"){
    df2$Tumor_Other_Histologic_Subtype[i] <- 3
  }
  else if (df2[i,]$`Tumor Other Histologic Subtype` == "Tubular/ cribriform"){
    df2$Tumor_Other_Histologic_Subtype[i] <- 4
  }
  else if (df2[i,]$`Tumor Other Histologic Subtype` == "Mucinous"){
    df2$Tumor_Other_Histologic_Subtype[i] <- 5
  }
  else if (df2[i,]$`Tumor Other Histologic Subtype` == "Medullary"){
    df2$Tumor_Other_Histologic_Subtype[i] <- 6
  }
  else if (df2[i,]$`Tumor Other Histologic Subtype` == "Other"){
    df2$Tumor_Other_Histologic_Subtype[i] <- 7
  }
}
```

```
## Warning: Unknown or uninitialised column: ‘Tumor_Other_Histologic_Subtype‘.
```

```
for (i in 1:nrow(df2)) {
  if (df2[i,]$`Oncotree Code` == "IDC") {
    df2$Oncotree_Code[i] <- 4
  }
  else if (df2[i,]$`Oncotree Code` == "MDLC") {
    df2$Oncotree_Code[i] <- 3
  }
  else if (df2[i,]$`Oncotree Code` == "ILC"){
    df2$Oncotree_Code[i] <- 2
  }
  else if (df2[i,]$`Oncotree Code` == "IMMC"){
    df2$Oncotree_Code[i] <- 1
  }
  else if (df2[i,]$`Oncotree Code` == "BREAST"){
    df2$Oncotree_Code[i] <- 5
  }
}
```

```
## Warning: Unknown or uninitialised column: ‘Oncotree_Code‘.
```

```
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$Overall_Survival_Status = ifelse(df2$`Overall Survival Status` =="1:DECEASED",1,0)
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)
for (i in 1:nrow(df2)) {
  if (df2[i,]$`3-Gene classifier subtype` == "ER+/HER2- High Prolif") {
    df2$Gene_classifier_subtype[i] <- 4
  }
  else if (df2[i,]$`3-Gene classifier subtype` == "ER+/HER2- Low Prolif") {
    df2$Gene_classifier_subtype[i] <- 3
  }
  else if (df2[i,]$`3-Gene classifier subtype` == "ER-/HER2-"){
    df2$Gene_classifier_subtype[i] <- 2
  }
  else if (df2[i,]$`3-Gene classifier subtype` == "HER2+"){
    df2$Gene_classifier_subtype[i] <- 1
  }
}
```

```
## Warning: Unknown or uninitialised column: ‘Gene_classifier_subtype‘.
```

```
for (i in 1:nrow(df2)) {
  if (df2[i,]$`Patient's Vital Status` == "Living") {
    df2$Patients_Vital_Status[i] <- 3
  }
  else if (df2[i,]$`Patient's Vital Status` == "Died of Disease") {
    df2$Patients_Vital_Status[i] <- 2
  }
  else if (df2[i,]$`Patient's Vital Status` == "Died of Other Causes"){
    df2$Patients_Vital_Status[i] <- 1
  }
}
```

```
## Warning: Unknown or uninitialised column: ‘Patients_Vital_Status‘.
```

```
for (i in 1:nrow(df2)) {
  if (df2[i,]$`Integrative Cluster` == "4ER+") {
    df2$Integrative_Cluster[i] <- 4.5
  }
  else if (df2[i,]$`Integrative Cluster` == "9") {
    df2$Integrative_Cluster[i] <- 9
  }
  else if (df2[i,]$`Integrative Cluster` == "7"){
    df2$Integrative_Cluster[i] <- 7
  }
  else if (df2[i,]$`Integrative Cluster` == "3"){
    df2$Integrative_Cluster[i] <- 3
  }
  else if (df2[i,]$`Integrative Cluster` == "10"){
    df2$Integrative_Cluster[i] <- 10
  }
  else if (df2[i,]$`Integrative Cluster` == "8"){
    df2$Integrative_Cluster[i] <- 8
  }
  else if (df2[i,]$`Integrative Cluster` == "6"){
    df2$Integrative_Cluster[i] <- 6
  }
  else if (df2[i,]$`Integrative Cluster` == "1"){
    df2$Integrative_Cluster[i] <- 1
  }
  else if (df2[i,]$`Integrative Cluster` == "2"){
    df2$Integrative_Cluster[i] <- 2
  }
  else if (df2[i,]$`Integrative Cluster` == "5"){
    df2$Integrative_Cluster[i] <- 5
  }
  else if (df2[i,]$`Integrative Cluster` == "4ER-"){
    df2$Integrative_Cluster[i] <- 3.5
  }
}
```

```
## Warning: Unknown or uninitialised column: ‘Integrative_Cluster‘.
```

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


df2
```

```
## # A tibble: 859 x 55
##    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           78.8 MASTEC~ Breast~ Modera~ NO      LumB         1 Positve Positi~
##  3           86.4 BREAST~ Breast~ Modera~ NO      LumB         1 Positve Positi~
##  4           85.5 MASTEC~ Breast~ Modera~ NO      LumA         1 Positve Positi~
##  5           45.4 BREAST~ Breast~ High    YES     LumB         1 Positve Positi~
##  6           61.5 BREAST~ Breast~ High    NO      LumB         1 Positve Positi~
##  7           68.7 MASTEC~ Breast~ Low     YES     Basal        1 Negati~ Negati~
##  8           49.9 MASTEC~ Breast~ Modera~ YES     LumA         1 Positve Positi~
##  9           54.2 MASTEC~ Breast~ High    NO      LumA         1 Positve Positi~
## 10           48.6 MASTEC~ Breast~ Low     NO      LumA         1 Positve Positi~
## # ... with 849 more rows, 46 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>, ‘Overall Survival (Months)‘ <dbl>, ...
```

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"
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 55
##      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, 45 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>,
## #   ‘Overall Survival Status‘ <chr>, ‘PR Status‘ <chr>, ...
```

```
df3 <- df2[,-c(2,3,4:6,8,9,11:17,21,23:25,27:32,35)]
df3
```

```
## # A tibble: 859 x 30
##      Age Cohort Neo_Grade Lymph Mutation Notti~1 Overa~2 Relap~3 Tumor~4 Tumor~5
##    <dbl>  <dbl>     <dbl> <dbl>    <dbl>   <dbl>   <dbl>   <dbl>   <dbl>   <dbl>
##  1  43.2      1         3     0        2    4.02   84.6    83.5       10       1
##  2  78.8      1         3     0        4    4.06    7.8     2.89      31       4
##  3  86.4      1         3     1        4    5.03   36.6    36.1       16       2
##  4  85.5      1         2     0        1    3.04  132.    123.        22       4
##  5  45.4      1         3     0        5    4.05  141.    139.        23       2
##  6  61.5      1         2     1        3    4.03  157.    155.        16       2
##  7  68.7      1         3     0        1    4.08    8.07    7.83      39       2
##  8  49.9      1         1     5        4    4.14   85.3    84.2       70       3
##  9  54.2      1         1     0        4    2.05  127.    125.        27       2
## 10  48.6      1         2     0        3    3.06   13.4    13.2       30       2
## # ... with 849 more rows, 20 more variables: Type_of_Breast_Surgery <dbl>,
## #   new_Chemotherapy <dbl>, new_Cellularity <dbl>,
## #   Pam50_Claudin_low_subtype <dbl>, ER_status_measured_by_IHC <dbl>,
## #   ER_Status_Positive <dbl>, HER2_status_measured_by_SNP6 <dbl>,
## #   HER2_Status_Positive <dbl>, Tumor_Other_Histologic_Subtype <dbl>,
## #   Oncotree_Code <dbl>, Hormone_Therapy <dbl>,
## #   Inferred_Menopausal_State <dbl>, Primary_Tumor_Laterality <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 1st Qu.:49.99	Min. :1.000 1st Qu.:1.000	Min. :1.000 1st Qu.:2.000	Min. : 0.000 1st Qu.: 0.000

Age	Cohort	Neo_Grade	Lymph
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

Tumor_Size	Tumor_Stage	Type_of_Breast_Surgery	new_Chemotherapy
Min. : 1.0	Min. :1.000	Min. :0.0000	Min. :0.0000
1st Qu.: 17.0	1st Qu.:1.000	1st Qu.:0.0000	1st Qu.:0.0000
Median : 22.0	Median :2.000	Median :0.0000	Median :0.0000
Mean : 25.7	Mean :1.767	Mean :0.4319	Mean :0.2538
3rd Qu.: 30.0	3rd Qu.:2.000	3rd Qu.:1.0000	3rd Qu.:1.0000
Max. :180.0	Max. :4.000	Max. :1.0000	Max. :1.0000

Table 4: Table continues below

new_Cellularity	Pam50_Claudin_low_subtype	ER_status_measured_by_IHC
Min. :1.000	Min. :1.000	Min. :0.0000
1st Qu.:2.000	1st Qu.:1.000	1st Qu.:0.0000
Median :3.000	Median :2.000	Median :1.0000
Mean :2.421	Mean :2.536	Mean :0.7404
3rd Qu.:3.000	3rd Qu.:4.000	3rd Qu.:1.0000
Max. :3.000	Max. :7.000	Max. :1.0000

Table 5: Table continues below

ER_Status_Positive	HER2_status_measured_by_SNP6	HER2_Status_Positive
Min. :0.0000	Min. :1.000	Min. :0.0000
1st Qu.:0.0000	1st Qu.:3.000	1st Qu.:0.0000
Median :1.0000	Median :4.000	Median :0.0000
Mean :0.7404	Mean :3.631	Mean :0.1444
3rd Qu.:1.0000	3rd Qu.:4.000	3rd Qu.:0.0000
Max. :1.0000	Max. :4.000	Max. :1.0000

Table 6: Table continues below

Tumor_Other_Histologic_Subtype	Oncotree_Code	Hormone_Therapy
Min. :1.000	Min. :4	Min. :0.0000
1st Qu.:1.000	1st Qu.:4	1st Qu.:0.0000
Median :1.000	Median :4	Median :1.0000
Mean :1.128	Mean :4	Mean :0.6042
3rd Qu.:1.000	3rd Qu.:4	3rd Qu.:1.0000
Max. :6.000	Max. :4	Max. :1.0000

Table 7: Table continues below

Inferred_Menopausal_State	Primary_Tumor_Laterality	Overall_Survival_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 :1.0000
Mean :0.7497	Mean :0.525	Mean :0.5565
3rd Qu.:1.0000	3rd Qu.:1.000	3rd Qu.:1.0000
Max. :1.0000	Max. :1.000	Max. :1.0000

Table 8: Table continues below

PR_Status	Radio_Therapy	Relapse_Free_Status
Min. :0.0000	Min. :0.0000	Min. :0.0000
1st Qu.:0.0000	1st Qu.:0.0000	1st Qu.:0.0000
Median :0.0000	Median :1.0000	Median :0.0000
Mean :0.4994	Mean :0.6799	Mean :0.4214
3rd Qu.:1.0000	3rd Qu.:1.0000	3rd Qu.:1.0000
Max. :1.0000	Max. :1.0000	Max. :1.0000

Gene_classifier_subtype	Patients_Vital_Status	Integrative_Cluster
Min. :1.000	Min. :1.000	Min. : 1.000
1st Qu.:2.000	1st Qu.:2.000	1st Qu.: 3.500
Median :3.000	Median :2.000	Median : 5.000
Mean :2.905	Mean :2.235	Mean : 5.861
3rd Qu.:4.000	3rd Qu.:3.000	3rd Qu.: 8.000
Max. :4.000	Max. :3.000	Max. :10.000

```
table(df3$cancer)
```

```
## Warning: Unknown or uninitialised column: ‘cancer‘.
```

```
## < table of extent 0 >
```

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

Table 10: 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 11: Table continues below

Relapse_Month	Tumor_Size	Tumor_Stage	Type_of_Breast_Surgery
83.52	10	1	1
2.89	31	4	0
36.09	16	2	1
123.3	22	4	0
139	23	2	1
155.4	16	2	1

Table 12: Table continues below

new_Chemotherapy	new_Cellularity	Pam50_Claudin_low_subtype
0	3	1
0	2	2
0	2	2
0	2	1
1	3	2
0	3	2

Table 13: 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 14: Table continues below

HER2_Status_Positive	Tumor_Other_Histologic_Subtype	Oncotree_Code
0	1	4
0	1	4
0	1	4
0	1	4

HER2_Status_Positive	Tumor_Other_Histologic_Subtype	Oncotree_Code
0	1	4
0	1	4

Table 15: Table continues below

Hormone_Therapy	Inferred_Menopausal_State	Primary_Tumor_Laterality
1	0	0
1	1	1
1	1	0
1	1	1
1	0	0
1	1	1

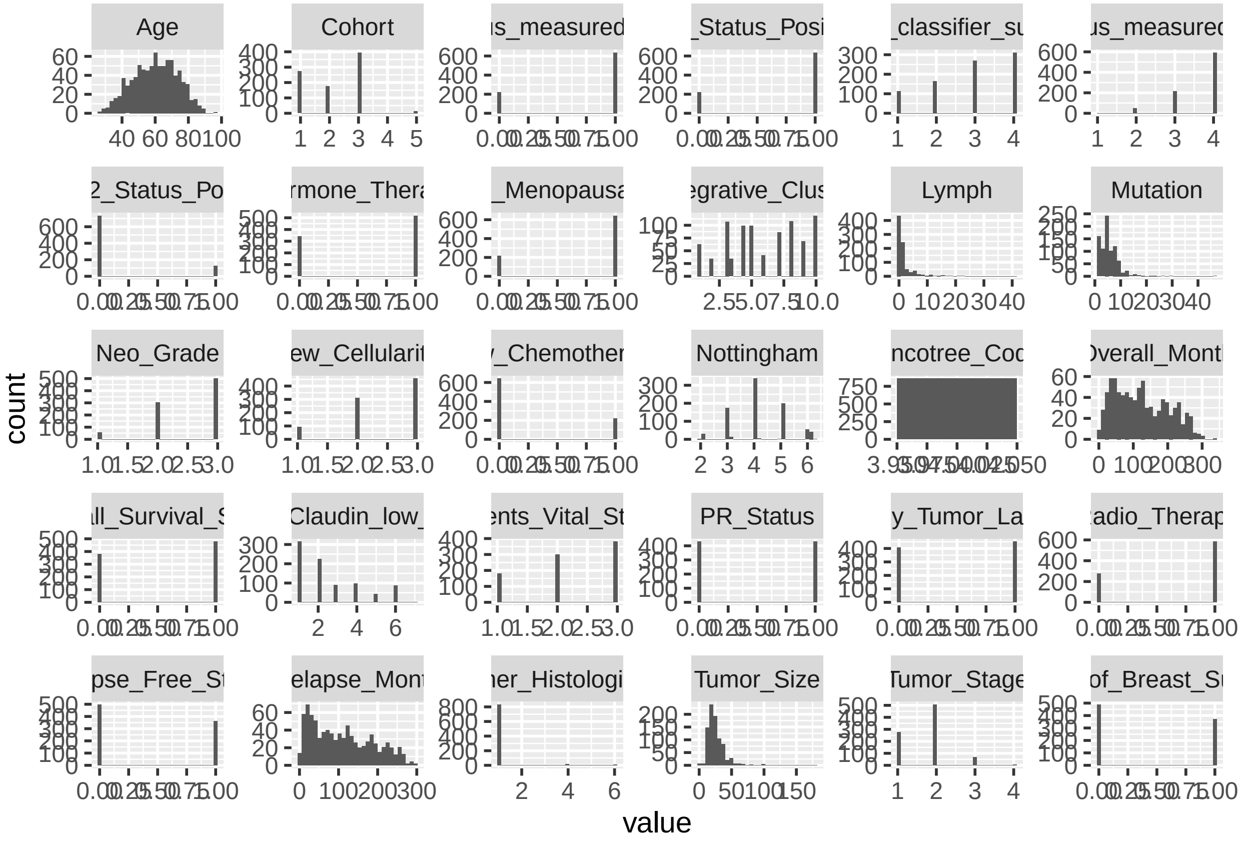
Table 16: Table continues below

Overall_Survival_Status	PR_Status	Radio_Therapy	Relapse_Free_Status
0	1	1	0
1	1	1	1
1	0	1	0
1	1	1	1
0	1	1	0
0	1	1	0

Gene_classifier_subtype	Patients_Vital_Status	Integrative_Cluster
4	3	4.5
4	2	7
4	1	9
3	2	3
4	3	10
4	3	7

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

```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```



```r
tr_ind <- 1:(nrow(df3) * 0.7)
df3_tr <- df3[tr_ind, ]
nrow(df3_tr)
```

```
## [1] 601
```

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

```
## [1] 258
```

variable selection

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

```
## Warning in leaps.setup(x, y, wt = wt, nbest = nbest, nvmax = nvmax, force.in =
## force.in, : 1 linear dependencies found
```

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

```
summary_BIC <- summary(fit_BIC)
summary_BIC
```

```
## Subset selection object
## Call: regsubsets.formula(Overall_Survival_Status ~ ., data = df3_tr)
## 29 Variables  (and intercept)
##                                Forced in Forced out
## Age                                FALSE      FALSE
## Cohort                             FALSE      FALSE
## Neo_Grade                          FALSE      FALSE
## Lymph                              FALSE      FALSE
## Mutation                           FALSE      FALSE
## Nottingham                         FALSE      FALSE
## Overall_Month                      FALSE      FALSE
## Relapse_Month                      FALSE      FALSE
## Tumor_Size                         FALSE      FALSE
## Tumor_Stage                        FALSE      FALSE
## Type_of_Breast_Surgery             FALSE      FALSE
## new_Chemotherapy                   FALSE      FALSE
## new_Cellularity                    FALSE      FALSE
## Pam50_Claudin_low_subtype          FALSE      FALSE
## ER_status_measured_by_IHC          FALSE      FALSE
## ER_Status_Positive                 FALSE      FALSE
## HER2_status_measured_by_SNP6       FALSE      FALSE
## HER2_Status_Positive               FALSE      FALSE
## Tumor_Other_Histologic_Subtype     FALSE      FALSE
## Hormone_Therapy                    FALSE      FALSE
## Inferred_Menopausal_State          FALSE      FALSE
## Primary_Tumor_Laterality           FALSE      FALSE
## PR_Status                          FALSE      FALSE
## Radio_Therapy                      FALSE      FALSE
## Relapse_Free_Status                FALSE      FALSE
## Gene_classifier_subtype            FALSE      FALSE
## Patients_Vital_Status              FALSE      FALSE
## Integrative_Cluster                FALSE      FALSE
## Oncotree_Code                      FALSE      FALSE
## 1 subsets of each size up to 9
## Selection Algorithm: exhaustive
##          Age Cohort Neo_Grade Lymph Mutation Nottingham Overall_Month
## 1  ( 1 ) " " " "    " "       " "   " "      " "        " "          
## 2  ( 1 ) " " " "    " "       " "   " "      " "        " "          
## 3  ( 1 ) " " " "    " "       " "   " "      " "        "*"          
## 4  ( 1 ) " " " "    " "       " "   " "      " "        "*"          
## 5  ( 1 ) " " "*"    " "       " "   " "      " "        "*"          
## 6  ( 1 ) " " "*"    "*"       " "   " "      " "        "*"          
## 7  ( 1 ) " " "*"    "*"       " "   " "      " "        "*"          
## 8  ( 1 ) " " "*"    "*"       " "   " "      " "        "*"          
## 9  ( 1 ) " " "*"    " "       " "   " "      " "        "*"          
##          Relapse_Month Tumor_Size Tumor_Stage Type_of_Breast_Surgery
## 1  ( 1 ) " "           " "        " "         " "                   
## 2  ( 1 ) " "           " "        " "         " "                   
## 3  ( 1 ) " "           " "        " "         " "                   
## 4  ( 1 ) "*"           " "        " "         " "                   
## 5  ( 1 ) "*"           " "        " "         " "
## 6  ( 1 ) "*"           " "        " "         " "
## 7  ( 1 ) "*"           " "        " "         " "
## 8  ( 1 ) "*"           " "        " "         " "
## 9  ( 1 ) "*"           " "        " "         " "
##          new_Chemotherapy new_Cellularity Pam50_Claudin_low_subtype
## 1  ( 1 ) " "              " "             " "
## 2  ( 1 ) " "              " "             " "
## 3  ( 1 ) " "              " "             " "
## 4  ( 1 ) " "              " "             " "
## 5  ( 1 ) " "              " "             " "
## 6  ( 1 ) " "              " "             " "
## 7  ( 1 ) " "              " "             " "
## 8  ( 1 ) " "              "*"             " "
## 9  ( 1 ) " "              "*"             "*"
##          ER_status_measured_by_IHC ER_Status_Positive
## 1  ( 1 ) " "                       " "
## 2  ( 1 ) " "                       " "
## 3  ( 1 ) " "                       " "
## 4  ( 1 ) " "                       " "
## 5  ( 1 ) " "                       " "
## 6  ( 1 ) " "                       " "
## 7  ( 1 ) " "                       " "
## 8  ( 1 ) " "                       " "
## 9  ( 1 ) " "                       " "
##          HER2_status_measured_by_SNP6 HER2_Status_Positive
## 1  ( 1 ) " "                          " "
## 2  ( 1 ) " "                          " "
## 3  ( 1 ) " "                          " "
## 4  ( 1 ) " "                          " "
## 5  ( 1 ) " "                          " "
## 6  ( 1 ) " "                          " "
## 7  ( 1 ) " "                          " "
## 8  ( 1 ) " "                          " "
## 9  ( 1 ) " "                          " "
##          Tumor_Other_Histologic_Subtype Oncotree_Code Hormone_Therapy
## 1  ( 1 ) " "                            " "           " "
## 2  ( 1 ) " "                            " "           " "
## 3  ( 1 ) " "                            " "           " "
## 4  ( 1 ) " "                            " "           " "
## 5  ( 1 ) " "                            " "           " "
## 6  ( 1 ) " "                            " "           " "
## 7  ( 1 ) " "                            " "           " "
## 8  ( 1 ) " "                            " "           " "
## 9  ( 1 ) " "                            " "           " "
##          Inferred_Menopausal_State Primary_Tumor_Laterality PR_Status
## 1  ( 1 ) " "                       " "                      " "
## 2  ( 1 ) " "                       " "                      " "
## 3  ( 1 ) " "                       " "                      " "
## 4  ( 1 ) " "                       " "                      " "
## 5  ( 1 ) " "                       " "                      " "
## 6  ( 1 ) " "                       " "                      " "
## 7  ( 1 ) " "                       " "                      " "
## 8  ( 1 ) " "                       " "                      " "
## 9  ( 1 ) " "                       " "                      " "
##          Radio_Therapy Relapse_Free_Status Gene_classifier_subtype
## 1  ( 1 ) " "           " "                 " "
## 2  ( 1 ) " "           "*"                 " "
## 3  ( 1 ) " "           "*"                 " "
## 4  ( 1 ) " "           "*"                 " "
## 5  ( 1 ) " "           "*"                 " "
## 6  ( 1 ) " "           "*"                 " "
## 7  ( 1 ) "*"           "*"                 " "
## 8  ( 1 ) "*"           "*"                 " "
## 9  ( 1 ) "*"           "*"                 "*"
##          Patients_Vital_Status Integrative_Cluster
## 1  ( 1 ) "*"                   " "
## 2  ( 1 ) "*"                   " "
## 3  ( 1 ) "*"                   " "
## 4  ( 1 ) "*"                   " "
## 5  ( 1 ) "*"                   " "
## 6  ( 1 ) "*"                   " "
## 7  ( 1 ) "*"                   " "
## 8  ( 1 ) "*"                   " "
## 9  ( 1 ) "*"                   " "
```

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

```
## [1] 7
```

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

```
##             (Intercept)                  Cohort               Neo_Grade
##             0.109958905             0.170746122             0.001105008
##           Overall_Month           Relapse_Month     Relapse_Free_Status
##            -0.005476658             0.003709453             0.542600588
## Gene_classifier_subtype     Integrative_Cluster
##             0.041350604             0.004721194
```

#Forward Stepwise Selection with Adjusted R squared

```
fit_FORWARD <- regsubsets(Overall_Survival_Status ~ ., data = df3_tr, method = "forward", nvmax = 10)
```

```
## Warning in leaps.setup(x, y, wt = wt, nbest = nbest, nvmax = nvmax, force.in =
## force.in, : 1 linear dependencies found

## Reordering variables and trying again:
```

```
as.factor(df3_tr$Cancer_Type_Detailed)
```

```
## Warning: Unknown or uninitialised column: `Cancer_Type_Detailed`.

## factor(0)
## Levels:
```

```
summary_FORWARD <- summary(fit_FORWARD)
max_FORWARD <- which.max(summary_FORWARD$adjr2)
max_FORWARD
```

```
## [1] 11
```

```
coef_FORWARD = coef(fit_FORWARD, max_FORWARD)
coef_FORWARD
```

```
##               (Intercept)                    Cohort                 Neo_Grade
##               1.393389377               0.028986465               0.007402980
##             Overall_Month             Relapse_Month          new_Cellularity
##              -0.002790297               0.002428075               0.011228038
## Pam50_Claudin_low_subtype       HER2_Status_Positive       Relapse_Free_Status
##               0.007805674               0.032912246               0.449010809
##   Gene_classifier_subtype     Patients_Vital_Status      Integrative_Cluster
##               0.016738115              -0.508268445               0.002911585
```

Backward Stepwise Selection with Cp

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

```
## Warning in leaps.setup(x, y, wt = wt, nbest = nbest, nvmax = nvmax, force.in =
## force.in, : 1 linear dependencies found
```

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

```
## Warning in rval$lopt[] <- rval$vorder[rval$lopt]: number of items to replace is
## not a multiple of replacement length
```

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

```
## [1] 15
```

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

```
##               (Intercept)                       Age                    Cohort
##              1.4597517531             -0.0004591716              0.0264038737
##                 Neo_Grade                Nottingham             Overall_Month
##              0.0187388732             -0.0120862346             -0.0028608747
##             Relapse_Month    Type_of_Breast_Surgery           new_Cellularity
##              0.0024906577             -0.0227149327              0.0126887146
## Pam50_Claudin_low_subtype       HER2_Status_Positive       Relapse_Free_Status
##              0.0072688859              0.0298195405              0.4509786320
##   Gene_classifier_subtype     Patients_Vital_Status      Integrative_Cluster
##              0.0168519260             -0.5100030489              0.0029181885
##             Oncotree_Code
##              0.0000000000
```

```
set.seed(0)
predict_BIC = glm(Overall_Survival_Status~Relapse_Month + Tumor_Size + Integrative_Cluster  + Cohort + 
pred_BIC = round(predict(predict_BIC,df3_te,type = "response"))
error_BIC = mean((df3_te$Overall_Survival_Status - pred_BIC)^2)
error_BIC
```

```
## [1] 0.3255814
```

```
predict_FORWARD = glm(Overall_Survival_Status ~ Cohort + Neo_Grade + Overall_Month + Relapse_Month + ne
pred_FORWARD = round(predict(predict_FORWARD,df3_te,type = "response"))
#pred_FORWARD
error_FORWARD =  mean((df3_te$Overall_Survival_Status - pred_FORWARD)^2)
#error_FORWARD

predict_BACKWARD = glm(Overall_Survival_Status ~  Age+ Cohort  +  Mutation + Overall_Month + Relapse_Mo
pred_BACKWARD = round(predict(predict_BACKWARD,df3_tr,type = "response"))
error_BACKWARD = mean((df3_te$Overall_Survival_Status - pred_BACKWARD)^2)
error_BACKWARD
```

```
## [1] 0.5091514
```

```
which.min(data.frame(error_BIC, error_BACKWARD)) # All the errors listed
```

```
## error_BIC
##         1
```

```
formula = Overall_Survival_Status ~ Cohort + Neo_Grade + Overall_Month + Relapse_Month + new_Cellularit
```

Model 1

```
set.seed(0)
library(randomForest)
```

```
## randomForest 4.7-1.1

## Type rfNews() to see new features/changes/bug fixes.

##
## Attaching package: 'randomForest'

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

## The following object is masked from 'package:ggplot2':
##
##     margin

## The following object is masked from 'package:psych':
##
##     outlier
```

```
rf_model <- randomForest(formula, data = df3_tr,ntree = 10,importance = T)
predictions <- predict(rf_model, df3_te)
mean((df3_te$Overall_Survival_Status - predictions)^2)
```

```
## [1] 0.008002068
```

Model 2

```
set.seed(0)
knn_test_error <- vector()
new <- data.frame(k=numeric(), knn_test_error)

for(i in seq(from = 1, to = 100, by = 5)) {
  knmodel <- knn3(formula, df3_tr, k = i)
  knn_test <- predict(knmodel, newdata = df3_te)
  knn_test_error <- mean((knn_test - df3_te$Overall_Survival_Status)^2)
  new[i,] <- c(i, knn_test_error)
}
frame = data.frame(knn_test_error,1:24)
data.frame(knn_test_error,1:24)[which.min(frame$knn_test_error),]
```

```
##   knn_test_error X1.24
## 1      0.3156751     1
```

Model 3 Lasso

```
x_tr<-as.matrix(df3_tr[,c(1:23, 25:ncol(df3_tr))])
y_tr<-as.matrix(df3_tr[,24])
x_te<-as.matrix(df3_te[,c(1:23, 25:ncol(df3_te))])
y_te<-as.matrix(df3_te[,24])

library(glmnet)
```

```
## Warning: package 'glmnet' was built under R version 4.2.2
```

```
## Loading required package: Matrix
```

```
## Warning: package 'Matrix' was built under R version 4.2.2
```

```
##
## Attaching package: 'Matrix'
```

```
## The following objects are masked from 'package:tidyr':
##
##     expand, pack, unpack
```

```
## Loaded glmnet 4.1-6
```

```
set.seed(0)
ridge<-glmnet(x=x_tr,y=y_tr,alpha=0)
#plot(ridge,xvar='lambda')

ridge_cv<-cv.glmnet(x=x_tr,y=y_tr,type.measure='mse',nfold=10,alpha=0)

best_ridge<-coef(ridge_cv, s = ridge_cv$lambda.min)

result_ridge<-predict(ridge_cv,newx=x_te,interval='prediction')
(err_ridge<-mean((y_te-result_ridge)^2))
```

```
## [1] 0.01644492
```

Model 4 Lasso

```
set.seed(0)
lasso<-glmnet(x=x_tr,y=y_tr,alpha=1)
#plot(lasso,xvar='lambda')
```

```
lasso_cv<-cv.glmnet(x=x_tr,y=y_tr,type.measure='mse',nfold=10,alpha=1,keep=T)
#lasso_cv$lambda.min

result_la<-predict(lasso_cv,newx=x_te,interval='prediction')
(err_la<-mean((y_te-result_la)^2))
```

```
## [1] 0.01214962
```

Model Selection