Feature Selection Comparison

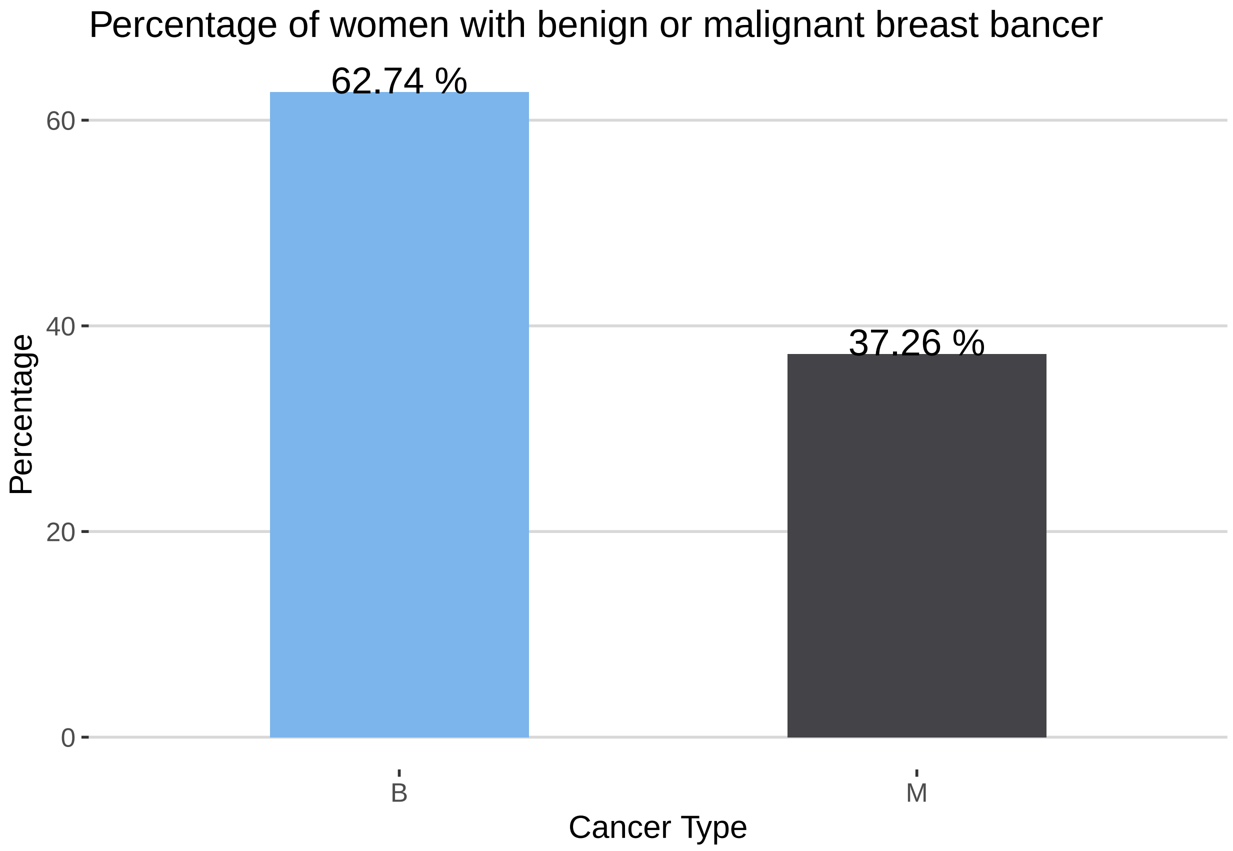
9/22/2020

```r
library(tidyverse)
library(data.table)
library(knitr)
library(caret)
library(glmnet)
library(ggthemes)

cancer <- fread("data.csv")

cancer[, V33 := NULL]
cancer[, diagnosis := factor(diagnosis)]
nms <- names(cancer)
nms <- gsub(" ", "_", nms)
names(cancer) <- nms
str(cancer)
```

```
## Classes 'data.table' and 'data.frame':   569 obs. of  32 variables:
##  $ id                     : int  842302 842517 84300903 84348301 84358402 843786 844359 84458202 8449
##  $ diagnosis              : Factor w/ 2 levels "B","M": 2 2 2 2 2 2 2 2 2 2 ...
##  $ radius_mean            : num  18 20.6 19.7 11.4 20.3 ...
##  $ texture_mean           : num  10.4 17.8 21.2 20.4 14.3 ...
##  $ perimeter_mean         : num  122.8 132.9 130 77.6 135.1 ...
##  $ area_mean              : num  1001 1326 1203 386 1297 ...
##  $ smoothness_mean        : num  0.1184 0.0847 0.1096 0.1425 0.1003 ...
##  $ compactness_mean       : num  0.2776 0.0786 0.1599 0.2839 0.1328 ...
##  $ concavity_mean         : num  0.3001 0.0869 0.1974 0.2414 0.198 ...
##  $ concave_points_mean    : num  0.1471 0.0702 0.1279 0.1052 0.1043 ...
##  $ symmetry_mean          : num  0.242 0.181 0.207 0.26 0.181 ...
##  $ fractal_dimension_mean : num  0.0787 0.0567 0.06 0.0974 0.0588 ...
##  $ radius_se              : num  1.095 0.543 0.746 0.496 0.757 ...
##  $ texture_se             : num  0.905 0.734 0.787 1.156 0.781 ...
##  $ perimeter_se           : num  8.59 3.4 4.58 3.44 5.44 ...
##  $ area_se                : num  153.4 74.1 94 27.2 94.4 ...
##  $ smoothness_se          : num  0.0064 0.00522 0.00615 0.00911 0.01149 ...
##  $ compactness_se         : num  0.049 0.0131 0.0401 0.0746 0.0246 ...
##  $ concavity_se           : num  0.0537 0.0186 0.0383 0.0566 0.0569 ...
##  $ concave_points_se      : num  0.0159 0.0134 0.0206 0.0187 0.0188 ...
##  $ symmetry_se            : num  0.03 0.0139 0.0225 0.0596 0.0176 ...
##  $ fractal_dimension_se   : num  0.00619 0.00353 0.00457 0.00921 0.00511 ...
##  $ radius_worst           : num  25.4 25 23.6 14.9 22.5 ...
##  $ texture_worst          : num  17.3 23.4 25.5 26.5 16.7 ...
##  $ perimeter_worst        : num  184.6 158.8 152.5 98.9 152.2 ...
##  $ area_worst             : num  2019 1956 1709 568 1575 ...
##  $ smoothness_worst       : num  0.162 0.124 0.144 0.21 0.137 ...
##  $ compactness_worst      : num  0.666 0.187 0.424 0.866 0.205 ...
##  $ concavity_worst        : num  0.712 0.242 0.45 0.687 0.4 ...
##  $ concave_points_worst   : num  0.265 0.186 0.243 0.258 0.163 ...
##  $ symmetry_worst         : num  0.46 0.275 0.361 0.664 0.236 ...
##  $ fractal_dimension_worst: num  0.1189 0.089 0.0876 0.173 0.0768 ...
##  - attr(*, ".internal.selfref")=<externalptr>
```

```
cancer[, id := NULL]
```

```
cancer[, .(freq = .N),
       by = diagnosis] %>%
    .[, perc := round(100 * freq/sum(freq), 2)] %>%

ggplot(aes(x=diagnosis, y=perc, fill = diagnosis)) +
    geom_bar(stat = "identity", width  = 0.5)+ theme_hc() +
    geom_text(aes(x=diagnosis, y=perc, label = paste(perc, "%")),
              position =  position_dodge(width = 0.5),
              vjust = 0.05, hjust = 0.5, size = 5)+
    scale_fill_hc(name = "")+
    labs(x = "Cancer Type",
         y = "Percentage",
         title = "Percentage of women with benign or malignant breast bancer")+
    theme(legend.position = "none",
          axis.title = element_text(size =12))
```



Test train

```r
set.seed(100)
train_sample <- sample(1:nrow(cancer), round(0.7*nrow(cancer)))
train_set <- cancer[train_sample,]
test_set <- cancer[-train_sample,]
```

Fit model

```r
library(broom)
glm_mod <- glm(diagnosis ~ .,
               data = train_set,
               family = binomial())
```

```r
tidy(glm_mod) %>% kable
```

term	estimate	std.error	statistic	p.value
(Intercept)	1.051554e+02	4.009638e+05	0.0002623	0.9997907
radius_mean	-9.903775e+02	1.560410e+05	-0.0063469	0.9949359
texture_mean	1.145513e+01	3.530283e+03	0.0032448	0.9974110
perimeter_mean	9.743985e+01	2.559560e+04	0.0038069	0.9969625
area_mean	2.585920e+00	5.745781e+02	0.0045006	0.9964091
smoothness_mean	2.947579e+03	1.035082e+06	0.0028477	0.9977279
compactness_mean	-8.526539e+03	8.904460e+05	-0.0095756	0.9923599
concavity_mean	2.219474e+03	4.943034e+05	0.0044901	0.9964174
concave_points_mean	1.138650e+04	9.282659e+05	0.0122664	0.9902131
symmetry_mean	-2.836263e+03	2.343713e+05	-0.0121016	0.9903446
fractal_dimension_mean	-3.493666e+03	1.652806e+06	-0.0021138	0.9983135
radius_se	-1.635119e+03	5.034541e+05	-0.0032478	0.9974086
texture_se	-1.992183e+01	2.387207e+04	-0.0008345	0.9993341
perimeter_se	5.366880e+01	2.473653e+04	0.0021696	0.9982689
area_se	2.032495e+01	3.870656e+03	0.0052510	0.9958103
smoothness_se	-2.378366e+04	2.990972e+06	-0.0079518	0.9936554
compactness_se	1.631593e+04	3.105827e+06	0.0052533	0.9958085
concavity_se	-6.128921e+03	5.893287e+05	-0.0103998	0.9917023
concave_points_se	3.931166e+04	2.950471e+06	0.0133239	0.9893694
symmetry_se	-2.073166e+04	2.814956e+06	-0.0073648	0.9941238
fractal_dimension_se	-1.088166e+05	1.888605e+07	-0.0057617	0.9954028
radius_worst	3.877885e+02	4.104385e+04	0.0094482	0.9924616
texture_worst	1.384641e+00	3.685304e+03	0.0003757	0.9997002
perimeter_worst	-2.947268e+01	5.744130e+03	-0.0051309	0.9959061
area_worst	-1.043073e+00	3.184540e+02	-0.0032754	0.9973866
smoothness_worst	-1.177138e+03	4.179253e+05	-0.0028166	0.9977527
compactness_worst	-1.178895e+03	2.761519e+05	-0.0042690	0.9965938
concavity_worst	3.930469e+02	1.677193e+05	0.0023435	0.9981302
concave_points_worst	-1.419982e+03	5.219040e+05	-0.0027208	0.9978291
symmetry_worst	3.534347e+03	2.750049e+05	0.0128519	0.9897459
fractal_dimension_worst	1.190558e+04	1.485167e+06	0.0080163	0.9936040

Forward

```
forward_select <- step(glm_mod, direction = "forward")
```

```
## Start:  AIC=62
## diagnosis ~ radius_mean + texture_mean + perimeter_mean + area_mean +
##     smoothness_mean + compactness_mean + concavity_mean + concave_points_mean +
##     symmetry_mean + fractal_dimension_mean + radius_se + texture_se +
##     perimeter_se + area_se + smoothness_se + compactness_se +
##     concavity_se + concave_points_se + symmetry_se + fractal_dimension_se +
##     radius_worst + texture_worst + perimeter_worst + area_worst +
##     smoothness_worst + compactness_worst + concavity_worst +
##     concave_points_worst + symmetry_worst + fractal_dimension_worst
```

Backward

```
back_select <- step(glm_mod, direction = "backward")
```

```
## Start:  AIC=62
## diagnosis ~ radius_mean + texture_mean + perimeter_mean + area_mean +
##     smoothness_mean + compactness_mean + concavity_mean + concave_points_mean +
##     symmetry_mean + fractal_dimension_mean + radius_se + texture_se +
##     perimeter_se + area_se + smoothness_se + compactness_se +
##     concavity_se + concave_points_se + symmetry_se + fractal_dimension_se +
##     radius_worst + texture_worst + perimeter_worst + area_worst +
##     smoothness_worst + compactness_worst + concavity_worst +
##     concave_points_worst + symmetry_worst + fractal_dimension_worst
##
##                            Df Deviance    AIC
## - texture_worst             1     0.00  60.00
## - radius_se                 1     0.00  60.00
## - perimeter_se              1     0.00  60.00
## - texture_se                1     0.00  60.00
## - concave_points_worst      1     0.00  60.00
## - concavity_worst           1     0.00  60.00
## - smoothness_worst          1     0.00  60.00
## - smoothness_mean           1     0.00  60.00
## - fractal_dimension_mean    1     0.00  60.00
## - area_worst                1     0.00  60.00
## - area_se                   1     0.00  60.00
## - texture_mean              1     0.00  60.00
## - compactness_worst         1     0.00  60.00
## - area_mean                 1     0.00  60.00
## - radius_worst              1     0.00  60.00
## - perimeter_worst           1     0.00  60.00
## - concavity_mean            1     0.00  60.00
## - smoothness_se             1     0.00  60.00
## - perimeter_mean            1     0.00  60.00
## - radius_mean               1     0.00  60.00
## - compactness_se            1     0.00  60.00
## - concavity_se             1     0.00  60.00
## - symmetry_mean            1     0.00  60.00
## - concave_points_se        1     0.00  60.00
## - symmetry_se              1     0.00  60.00
## - compactness_mean         1     0.00  60.00
## - fractal_dimension_se     1     0.00  60.00
## - symmetry_worst           1     0.00  60.00
## <none>                         0.00  62.00
## - concave_points_mean      1   432.52 492.52
## - fractal_dimension_worst  1   865.05 925.05
## 
## Step:  AIC=60
## diagnosis ~ radius_mean + texture_mean + perimeter_mean + area_mean + 
##     smoothness_mean + compactness_mean + concavity_mean + concave_points_mean + 
##     symmetry_mean + fractal_dimension_mean + radius_se + texture_se + 
##     perimeter_se + area_se + smoothness_se + compactness_se + 
##     concavity_se + concave_points_se + symmetry_se + fractal_dimension_se + 
##     radius_worst + perimeter_worst + area_worst + smoothness_worst + 
##     compactness_worst + concavity_worst + concave_points_worst + 
##     symmetry_worst + fractal_dimension_worst
## 
##                           Df Deviance    AIC
## - texture_se               1     0.00  58.00
## - area_worst               1     0.00  58.00
## - radius_se                1     0.00  58.00
## - perimeter_se             1     0.00  58.00
## - concavity_worst          1     0.00  58.00
## - smoothness_worst         1     0.00  58.00
## - fractal_dimension_mean   1     0.00  58.00
## - concave_points_worst     1     0.00  58.00
## - smoothness_mean          1     0.00  58.00
## - compactness_worst        1     0.00  58.00
## - area_mean                1     0.00  58.00
## - concavity_mean           1     0.00  58.00
## - perimeter_worst          1     0.00  58.00
## - area_se                  1     0.00  58.00
## - perimeter_mean           1     0.00  58.00
## - radius_worst             1     0.00  58.00
## - radius_mean              1     0.00  58.00
## - compactness_se           1     0.00  58.00
## - concavity_se             1     0.00  58.00
## - fractal_dimension_worst  1     0.00  58.00
## - concave_points_se        1     0.00  58.00
## - smoothness_se            1     0.00  58.00
## - symmetry_mean            1     0.00  58.00
## - texture_mean             1     0.00  58.00
## - symmetry_worst           1     0.00  58.00
## - fractal_dimension_se     1     0.00  58.00
## - compactness_mean         1     1.51  59.51
## <none>                         0.00  60.00
## - concave_points_mean      1   792.96 850.96
## - symmetry_se              1   865.05 923.05
## 
## Step:  AIC=58
## diagnosis ~ radius_mean + texture_mean + perimeter_mean + area_mean +
##     smoothness_mean + compactness_mean + concavity_mean + concave_points_mean +
##     symmetry_mean + fractal_dimension_mean + radius_se + perimeter_se +
##     area_se + smoothness_se + compactness_se + concavity_se +
##     concave_points_se + symmetry_se + fractal_dimension_se +
##     radius_worst + perimeter_worst + area_worst + smoothness_worst +
##     compactness_worst + concavity_worst + concave_points_worst +
##     symmetry_worst + fractal_dimension_worst
##
##                           Df Deviance    AIC
## - radius_se                1     0.00  56.00
## - area_worst               1     0.00  56.00
## - concavity_worst          1     0.00  56.00
## - perimeter_se             1     0.00  56.00
## - smoothness_worst         1     0.00  56.00
## - fractal_dimension_mean   1     0.00  56.00
## - concave_points_worst     1     0.00  56.00
## - smoothness_mean          1     0.00  56.00
## - compactness_worst        1     0.00  56.00
## - concavity_mean           1     0.00  56.00
## - perimeter_worst          1     0.00  56.00
## - area_mean                1     0.00  56.00
## - radius_worst             1     0.00  56.00
## - area_se                  1     0.00  56.00
## - perimeter_mean           1     0.00  56.00
## - compactness_se           1     0.00  56.00
## - radius_mean              1     0.00  56.00
## - smoothness_se            1     0.00  56.00
## - concavity_se             1     0.00  56.00
## - concave_points_se        1     0.00  56.00
## - fractal_dimension_worst  1     0.00  56.00
## - compactness_mean         1     0.00  56.00
## - symmetry_worst           1     0.00  56.00
## <none>                           0.00  58.00
## - texture_mean             1    27.05  83.05
## - symmetry_mean            1   648.79 704.79
## - concave_points_mean      1   792.96 848.96
## - fractal_dimension_se     1   792.96 848.96
## - symmetry_se              1   937.13 993.13
##
## Step:  AIC=56
## diagnosis ~ radius_mean + texture_mean + perimeter_mean + area_mean +
##     smoothness_mean + compactness_mean + concavity_mean + concave_points_mean +
##     symmetry_mean + fractal_dimension_mean + perimeter_se + area_se +
##     smoothness_se + compactness_se + concavity_se + concave_points_se +
##     symmetry_se + fractal_dimension_se + radius_worst + perimeter_worst +
##     area_worst + smoothness_worst + compactness_worst + concavity_worst +
##     concave_points_worst + symmetry_worst + fractal_dimension_worst
##
##                           Df Deviance    AIC
## - area_worst               1     0.00  54.00
## - smoothness_mean          1     0.00  54.00
## - smoothness_worst         1     0.00  54.00
## - concave_points_worst     1     0.00  54.00
## - fractal_dimension_mean   1     0.00  54.00
## - perimeter_se             1     0.00  54.00
## - concavity_worst          1     0.00  54.00
## - area_mean                1     0.00  54.00
## - concavity_mean           1     0.00  54.00
## - perimeter_worst          1     0.00  54.00
## - radius_worst             1     0.00  54.00
## - area_se                  1     0.00  54.00
## - perimeter_mean           1     0.00  54.00
## - compactness_worst        1     0.00  54.00
## - radius_mean              1     0.00  54.00
## - concave_points_se        1     0.00  54.00
## - concavity_se             1     0.00  54.00
## - compactness_se           1     0.00  54.00
## - symmetry_mean            1     0.00  54.00
## - fractal_dimension_worst  1     0.00  54.00
## - compactness_mean         1     0.00  54.00
## - symmetry_se              1     0.00  54.00
## <none>                         0.00  56.00
## - symmetry_worst           1    31.57  85.57
## - texture_mean             1    34.43  88.43
## - fractal_dimension_se     1   792.96 846.96
## - concave_points_mean      1   865.05 919.05
## - smoothness_se            1   865.05 919.05
##
## Step:  AIC=54
## diagnosis ~ radius_mean + texture_mean + perimeter_mean + area_mean +
##     smoothness_mean + compactness_mean + concavity_mean + concave_points_mean +
##     symmetry_mean + fractal_dimension_mean + perimeter_se + area_se +
##     smoothness_se + compactness_se + concavity_se + concave_points_se +
##     symmetry_se + fractal_dimension_se + radius_worst + perimeter_worst +
##     smoothness_worst + compactness_worst + concavity_worst +
##     concave_points_worst + symmetry_worst + fractal_dimension_worst
##
##                           Df Deviance    AIC
## - perimeter_se             1     0.00  52.00
## - smoothness_mean          1     0.00  52.00
## - smoothness_worst         1     0.00  52.00
## - concave_points_worst     1     0.00  52.00
## - fractal_dimension_mean   1     0.00  52.00
## - area_mean                1     0.00  52.00
## - concavity_mean           1     0.00  52.00
## - perimeter_worst          1     0.00  52.00
## - concavity_worst          1     0.00  52.00
## - area_se                  1     0.00  52.00
## - perimeter_mean           1     0.00  52.00
## - compactness_worst        1     0.00  52.00
## - radius_worst             1     0.00  52.00
## - radius_mean              1     0.00  52.00
## - concave_points_se        1     0.00  52.00
## - compactness_se           1     0.00  52.00
## - concavity_se             1     0.00  52.00
## - fractal_dimension_worst  1     0.00  52.00
## - symmetry_mean            1     0.00  52.00
## - fractal_dimension_se     1     0.00  52.00
## - compactness_mean         1     0.00  52.00
## - symmetry_se              1     0.00  52.00
## <none>                           0.00  54.00
## - symmetry_worst           1    31.84  83.84
## - texture_mean             1    39.46  91.46
## - concave_points_mean      1   792.96 844.96
## - smoothness_se            1   865.05 917.05
##
## Step:  AIC=52
## diagnosis ~ radius_mean + texture_mean + perimeter_mean + area_mean +
##     smoothness_mean + compactness_mean + concavity_mean + concave_points_mean +
##     symmetry_mean + fractal_dimension_mean + area_se + smoothness_se +
##     compactness_se + concavity_se + concave_points_se + symmetry_se +
##     fractal_dimension_se + radius_worst + perimeter_worst + smoothness_worst +
##     compactness_worst + concavity_worst + concave_points_worst +
##     symmetry_worst + fractal_dimension_worst
##
##                           Df Deviance    AIC
## - concave_points_worst     1     0.00  50.00
## - smoothness_worst         1     0.00  50.00
## - area_mean                1     0.00  50.00
## - fractal_dimension_mean   1     0.00  50.00
## - concavity_worst          1     0.00  50.00
## - smoothness_mean          1     0.00  50.00
## - concavity_mean           1     0.00  50.00
## - area_se                  1     0.00  50.00
## - smoothness_se            1     0.00  50.00
## - perimeter_mean           1     0.00  50.00
## - concave_points_se        1     0.00  50.00
## - perimeter_worst          1     0.00  50.00
## - compactness_worst        1     0.00  50.00
## - concavity_se             1     0.00  50.00
## - radius_mean              1     0.00  50.00
## - concave_points_mean      1     0.00  50.00
## <none>                           0.00  52.00
## - fractal_dimension_worst  1    28.26  78.26
## - compactness_se           1    28.98  78.98
## - compactness_mean         1    29.20  79.20
## - fractal_dimension_se     1    33.47  83.47
## - symmetry_se              1    33.96  83.96
## - symmetry_worst           1    35.47  85.47
## - radius_worst             1    36.30  86.30
## - texture_mean             1    39.89  89.89
## - symmetry_mean            1   648.79 698.79
##
## Step:  AIC=50
## diagnosis ~ radius_mean + texture_mean + perimeter_mean + area_mean +
##     smoothness_mean + compactness_mean + concavity_mean + concave_points_mean +
##     symmetry_mean + fractal_dimension_mean + area_se + smoothness_se +
##     compactness_se + concavity_se + concave_points_se + symmetry_se +
##     fractal_dimension_se + radius_worst + perimeter_worst + smoothness_worst +
##     compactness_worst + concavity_worst + symmetry_worst + fractal_dimension_worst
##
```

```
##                           Df Deviance    AIC
## - smoothness_worst         1     0.00  48.00
## - fractal_dimension_mean   1     0.00  48.00
## - area_mean                1     0.00  48.00
## - concavity_worst          1     0.00  48.00
## - smoothness_mean          1     0.00  48.00
## - concavity_mean           1     0.00  48.00
## - smoothness_se            1     0.00  48.00
## - perimeter_mean           1     0.00  48.00
## - perimeter_worst          1     0.00  48.00
## - area_se                  1     0.00  48.00
## - compactness_worst        1     0.00  48.00
## - concavity_se             1     0.00  48.00
## - radius_mean              1     0.00  48.00
## <none>                         0.00  50.00
## - symmetry_mean            1    21.77  69.77
## - fractal_dimension_worst  1    28.34  76.34
## - compactness_mean         1    30.82  78.82
## - compactness_se           1    31.00  79.00
## - concave_points_se        1    32.09  80.09
## - fractal_dimension_se     1    33.63  81.63
## - symmetry_se              1    34.73  82.73
## - symmetry_worst           1    35.59  83.59
## - radius_worst             1    36.48  84.48
## - texture_mean             1    40.72  88.72
## - concave_points_mean      1   720.87 768.87
## 
## Step:  AIC=48
## diagnosis ~ radius_mean + texture_mean + perimeter_mean + area_mean + 
##     smoothness_mean + compactness_mean + concavity_mean + concave_points_mean + 
##     symmetry_mean + fractal_dimension_mean + area_se + smoothness_se + 
##     compactness_se + concavity_se + concave_points_se + symmetry_se + 
##     fractal_dimension_se + radius_worst + perimeter_worst + compactness_worst + 
##     concavity_worst + symmetry_worst + fractal_dimension_worst
## 
##                           Df Deviance     AIC
## - area_mean                1     0.00   46.00
## - concavity_worst          1     0.00   46.00
## - smoothness_mean          1     0.00   46.00
## - fractal_dimension_mean   1     0.00   46.00
## - concavity_mean           1     0.00   46.00
## - perimeter_mean           1     0.00   46.00
## - compactness_worst        1     0.00   46.00
## - smoothness_se            1     0.00   46.00
## - concavity_se             1     0.00   46.00
## - perimeter_worst          1     0.00   46.00
## - radius_mean              1     0.00   46.00
## <none>                         0.00   48.00
## - symmetry_mean            1    23.14   69.14
## - fractal_dimension_worst  1    31.04   77.04
## - compactness_se           1    31.38   77.38
## - compactness_mean         1    31.60   77.60
## - concave_points_se        1    32.61   78.61
## - fractal_dimension_se     1    33.65   79.65
## - symmetry_se              1    34.81   80.81
## - symmetry_worst           1    35.63   81.63
## - radius_worst             1    37.03   83.03
## - texture_mean             1    40.73   86.73
## - concave_points_mean      1   792.96  838.96
## - area_se                  1  2667.23 2713.23
##
## Step:  AIC=46
## diagnosis ~ radius_mean + texture_mean + perimeter_mean + smoothness_mean +
##     compactness_mean + concavity_mean + concave_points_mean +
##     symmetry_mean + fractal_dimension_mean + area_se + smoothness_se +
##     compactness_se + concavity_se + concave_points_se + symmetry_se +
##     fractal_dimension_se + radius_worst + perimeter_worst + compactness_worst +
##     concavity_worst + symmetry_worst + fractal_dimension_worst
##
##                           Df Deviance    AIC
## - concavity_worst          1    0.000 44.000
## - smoothness_mean          1    0.000 44.000
## - fractal_dimension_mean   1    0.000 44.000
## - concavity_mean           1    0.000 44.000
## - smoothness_se            1    0.000 44.000
## - compactness_worst        1    0.000 44.000
## - perimeter_mean           1    0.000 44.000
## - perimeter_worst          1    0.000 44.000
## - radius_mean              1    0.000 44.000
## <none>                          0.000 46.000
## - symmetry_mean            1   23.136 67.136
## - concavity_se             1   23.587 67.587
## - concave_points_mean      1   24.739 68.739
## - area_se                  1   28.439 72.439
## - fractal_dimension_worst  1   31.461 75.461
## - compactness_se           1   31.977 75.977
## - concave_points_se        1   32.711 76.711
## - symmetry_se              1   35.155 79.155
## - fractal_dimension_se     1   35.470 79.470
## - compactness_mean         1   36.154 80.154
## - symmetry_worst           1   36.228 80.228
## - radius_worst             1   37.031 81.031
## - texture_mean             1   40.765 84.765
##
## Step:  AIC=44
## diagnosis ~ radius_mean + texture_mean + perimeter_mean + smoothness_mean +
##     compactness_mean + concavity_mean + concave_points_mean +
##     symmetry_mean + fractal_dimension_mean + area_se + smoothness_se +
##     compactness_se + concavity_se + concave_points_se + symmetry_se +
##     fractal_dimension_se + radius_worst + perimeter_worst + compactness_worst +
##     symmetry_worst + fractal_dimension_worst
##
##                           Df Deviance    AIC
## - fractal_dimension_mean   1    0.000 42.000
## - smoothness_mean          1    0.000 42.000
## - smoothness_se            1    0.000 42.000
## - concavity_mean           1    0.000 42.000
## - perimeter_mean           1    0.000 42.000
## - perimeter_worst          1    0.000 42.000
## <none>                          0.000 44.000
## - compactness_worst        1   18.355 60.355
## - radius_mean              1   20.575 62.575
## - symmetry_mean            1   23.267 65.267
## - concavity_se             1   23.938 65.938
## - concave_points_mean      1   25.335 67.335
## - area_se                  1   28.560 70.560
## - fractal_dimension_worst  1   31.525 73.525
## - compactness_se           1   32.295 74.295
## - concave_points_se        1   33.771 75.771
## - symmetry_se              1   35.317 77.317
## - fractal_dimension_se     1   35.482 77.482
## - symmetry_worst           1   36.256 78.256
## - radius_worst             1   37.031 79.031
## - compactness_mean         1   37.440 79.440
## - texture_mean             1   42.140 84.140
##
## Step:  AIC=42
## diagnosis ~ radius_mean + texture_mean + perimeter_mean + smoothness_mean +
##     compactness_mean + concavity_mean + concave_points_mean +
##     symmetry_mean + area_se + smoothness_se + compactness_se +
##     concavity_se + concave_points_se + symmetry_se + fractal_dimension_se +
##     radius_worst + perimeter_worst + compactness_worst + symmetry_worst +
##     fractal_dimension_worst
##
##                           Df Deviance    AIC
## - smoothness_mean          1    0.000 40.000
## - smoothness_se            1    0.000 40.000
## - concavity_mean           1    0.000 40.000
## - perimeter_mean           1    0.000 40.000
## - perimeter_worst          1    0.000 40.000
## <none>                          0.000 42.000
## - compactness_worst        1   19.408 59.408
## - radius_mean              1   20.603 60.603
## - concave_points_mean      1   25.404 65.404
## - symmetry_mean            1   26.370 66.370
## - concavity_se             1   26.380 66.380
## - area_se                  1   29.967 69.967
## - fractal_dimension_worst  1   32.004 72.004
## - compactness_se           1   32.505 72.505
## - concave_points_se        1   33.882 73.882
## - symmetry_se              1   35.439 75.439
## - fractal_dimension_se     1   36.176 76.176
## - symmetry_worst           1   36.796 76.796
## - radius_worst             1   37.234 77.234
## - compactness_mean         1   39.671 79.671
## - texture_mean             1   42.329 82.329
##
## Step:  AIC=40
## diagnosis ~ radius_mean + texture_mean + perimeter_mean + compactness_mean +
##     concavity_mean + concave_points_mean + symmetry_mean + area_se +
##     smoothness_se + compactness_se + concavity_se + concave_points_se +
##     symmetry_se + fractal_dimension_se + radius_worst + perimeter_worst +
##     compactness_worst + symmetry_worst + fractal_dimension_worst
##
##                           Df Deviance    AIC
## - smoothness_se            1    0.000 38.000
## - concavity_mean           1    0.000 38.000
## - perimeter_worst          1    0.000 38.000
## <none>                          0.000 40.000
## - compactness_worst        1   21.007 59.007
## - perimeter_mean           1   21.594 59.594
## - radius_mean              1   26.202 64.202
## - concavity_se             1   26.386 64.386
## - symmetry_mean            1   27.731 65.731
## - compactness_se           1   33.258 71.258
## - fractal_dimension_worst  1   33.536 71.536
## - concave_points_se        1   33.946 71.946
## - symmetry_se              1   36.586 74.586
## - fractal_dimension_se     1   36.826 74.826
## - concave_points_mean      1   36.994 74.994
## - radius_worst             1   38.359 76.359
## - symmetry_worst           1   38.385 76.385
## - compactness_mean         1   39.777 77.777
## - area_se                  1   41.245 79.245
## - texture_mean             1   42.635 80.635
##
## Step:  AIC=38
## diagnosis ~ radius_mean + texture_mean + perimeter_mean + compactness_mean +
##     concavity_mean + concave_points_mean + symmetry_mean + area_se +
##     compactness_se + concavity_se + concave_points_se + symmetry_se +
##     fractal_dimension_se + radius_worst + perimeter_worst + compactness_worst +
##     symmetry_worst + fractal_dimension_worst
##
##                           Df Deviance     AIC
## <none>                           0.00   38.00
## - perimeter_mean           1    22.05   58.05
## - compactness_worst        1    23.08   59.08
## - concavity_mean           1    25.78   61.78
## - radius_mean              1    26.20   62.20
## - concavity_se             1    28.18   64.18
## - symmetry_mean            1    28.24   64.24
## - compactness_se           1    33.27   69.27
## - concave_points_se        1    34.41   70.41
## - fractal_dimension_worst  1    34.77   70.77
## - symmetry_se              1    36.59   72.59
## - concave_points_mean      1    37.00   73.00
## - fractal_dimension_se     1    38.01   74.01
## - symmetry_worst           1    38.94   74.94
## - radius_worst             1    39.52   75.52
## - compactness_mean         1    41.26   77.26
## - area_se                  1    42.74   78.74
## - texture_mean             1    44.61   80.61
## - perimeter_worst          1  1081.31 1117.31
```

Using Entropy-Based Feature Selection Algorithms

```r
library(FSelectorRcpp)
x <- information_gain(diagnosis ~ ., train_set)
x %>% arrange(desc(importance)) %>%
  kable()
```

attributes	importance
perimeter_worst	0.4850561
area_worst	0.4675581
concave_points_worst	0.4538449
radius_worst	0.4478213
concave_points_mean	0.4155797
perimeter_mean	0.4087355
area_mean	0.3881128
radius_mean	0.3814810
area_se	0.3664849
concavity_mean	0.3499271
concavity_worst	0.3458024
radius_se	0.2562297
perimeter_se	0.2523637
compactness_worst	0.2145325
compactness_mean	0.2142234
concavity_se	0.1483622
concave_points_se	0.1402913
texture_mean	0.1265121
texture_worst	0.1217746
symmetry_worst	0.1008219
smoothness_worst	0.0941130
compactness_se	0.0691604
symmetry_mean	0.0669995
smoothness_mean	0.0641805
fractal_dimension_worst	0.0596582
symmetry_se	0.0272433
fractal_dimension_se	0.0257642
fractal_dimension_mean	0.0231045
texture_se	0.0000000
smoothness_se	0.0000000

Recursive Feature Elimination (RFE)

```r
ctrl <- rfeControl(functions = rfFuncs,
                   method = "repeatedcv",
                   repeats = 5,
                   verbose = FALSE)

lmProfile <- rfe(diagnosis ~ .,
                 data = train_set,
                 rfeControl = ctrl)
```

```
lmProfile
```

```
##
## Recursive feature selection
##
## Outer resampling method: Cross-Validated (10 fold, repeated 5 times)
##
## Resampling performance over subset size:
##
##  Variables Accuracy  Kappa AccuracySD KappaSD Selected
##          4   0.9080 0.8044    0.03563 0.07491
##          8   0.9397 0.8715    0.03133 0.06639
##         16   0.9482 0.8900    0.03104 0.06576        *
##         30   0.9402 0.8725    0.03295 0.07047
##
## The top 5 variables (out of 16):
##    perimeter_worst, concave_points_worst, area_worst, radius_worst, concave_points_mean
```

```
lmProfile$optVariables
```

```
##  [1] "perimeter_worst"      "concave_points_worst" "area_worst"
##  [4] "radius_worst"         "concave_points_mean"  "area_se"
##  [7] "texture_worst"        "concavity_worst"      "texture_mean"
## [10] "concavity_mean"       "area_mean"            "radius_se"
## [13] "smoothness_worst"     "perimeter_mean"       "perimeter_se"
## [16] "radius_mean"
```

```
var
```

```
## function (x, y = NULL, na.rm = FALSE, use)
## {
##     if (missing(use))
##         use <- if (na.rm)
##             "na.or.complete"
##         else "everything"
##     na.method <- pmatch(use, c("all.obs", "complete.obs", "pairwise.complete.obs",
##         "everything", "na.or.complete"))
##     if (is.na(na.method))
##         stop("invalid 'use' argument")
##     if (is.data.frame(x))
##         x <- as.matrix(x)
##     else stopifnot(is.atomic(x))
##     if (is.data.frame(y))
##         y <- as.matrix(y)
##     else stopifnot(is.atomic(y))
##     .Call(C_cov, x, y, na.method, FALSE)
## }
## <bytecode: 0x0000000029447278>
## <environment: namespace:stats>
```

Model

```r
cv_fold <- createFolds(train_set$diagnosis, k = 5)

train_ctrl <- trainControl(method = "cv",
                           number = 5,
                           summaryFunction = twoClassSummary,
                           classProbs = TRUE,
                           allowParallel=T,
                           index = cv_fold,
                           verboseIter = FALSE,
                           savePredictions = TRUE,
                           search = "grid")
glm_grid <- expand.grid(
                       alpha = 0:1,
                       lambda = seq(0.0001, 1, length = 10)
                      )
```

```r
full_model <- train(
    diagnosis~.,
    data = train_set,
    method = "glmnet",
    metric = "ROC",
    trControl = train_ctrl,
    tuneGrid = glm_grid
)

full_model
```

```
## glmnet
##
## 398 samples
##  30 predictor
##   2 classes: 'B', 'M'
##
## No pre-processing
## Resampling: Cross-Validated (5 fold)
## Summary of sample sizes: 79, 79, 80, 81, 79
## Resampling results across tuning parameters:
##
##   alpha  lambda  ROC        Sens       Spec
##   0      0.0001  0.9883920  0.9867969  0.9127354
##   0      0.1112  0.9882908  0.9898425  0.8897304
##   0      0.2223  0.9872652  0.9888325  0.8633518
##   0      0.3334  0.9865297  0.9898477  0.8501558
##   0      0.4445  0.9860878  0.9898477  0.8369869
##   0      0.5556  0.9857958  0.9898477  0.8238315
##   0      0.6667  0.9854119  0.9898477  0.8156076
##   0      0.7778  0.9850109  0.9908629  0.8024522
##   0      0.8889  0.9847022  0.9908629  0.7975207
##   0      1.0000  0.9843605  0.9908629  0.7942284
##   1      0.0001  0.9751883  0.9644411  0.8996206
##   1      0.1112  0.9744598  0.9888325  0.7744615
##   1      0.2223  0.9658903  0.9959391  0.6462133
##   1      0.3334  0.9650977  1.0000000  0.1712505
##   1      0.4445  0.5000000  1.0000000  0.0000000
##   1      0.5556  0.5000000  1.0000000  0.0000000
##   1      0.6667  0.5000000  1.0000000  0.0000000
##   1      0.7778  0.5000000  1.0000000  0.0000000
##   1      0.8889  0.5000000  1.0000000  0.0000000
##   1      1.0000  0.5000000  1.0000000  0.0000000
##
## ROC was used to select the optimal model using the largest value.
## The final values used for the model were alpha = 0 and lambda = 1e-04.
```

Forward model

```
forward_model <- train(
    forward_select$formula,
    data = train_set,
    method = "glmnet",
    metric = "ROC",
    trControl = train_ctrl,
    tuneGrid = glm_grid
)

forward_model
```

```
## glmnet
##
## 398 samples
##  30 predictor
##   2 classes: 'B', 'M'
##
## No pre-processing
## Resampling: Cross-Validated (5 fold)
## Summary of sample sizes: 79, 79, 80, 81, 79
## Resampling results across tuning parameters:
##
##   alpha  lambda  ROC        Sens       Spec
##   0      0.0001  0.9883920  0.9867969  0.9127354
##   0      0.1112  0.9882908  0.9898425  0.8897304
##   0      0.2223  0.9872652  0.9888325  0.8633518
##   0      0.3334  0.9865297  0.9898477  0.8501558
##   0      0.4445  0.9860878  0.9898477  0.8369869
##   0      0.5556  0.9857958  0.9898477  0.8238315
##   0      0.6667  0.9854119  0.9898477  0.8156076
##   0      0.7778  0.9850109  0.9908629  0.8024522
##   0      0.8889  0.9847022  0.9908629  0.7975207
##   0      1.0000  0.9843605  0.9908629  0.7942284
##   1      0.0001  0.9751883  0.9644411  0.8996206
##   1      0.1112  0.9744598  0.9888325  0.7744615
##   1      0.2223  0.9658903  0.9959391  0.6462133
##   1      0.3334  0.9650977  1.0000000  0.1712505
##   1      0.4445  0.5000000  1.0000000  0.0000000
##   1      0.5556  0.5000000  1.0000000  0.0000000
##   1      0.6667  0.5000000  1.0000000  0.0000000
##   1      0.7778  0.5000000  1.0000000  0.0000000
##   1      0.8889  0.5000000  1.0000000  0.0000000
##   1      1.0000  0.5000000  1.0000000  0.0000000
##
## ROC was used to select the optimal model using the largest value.
## The final values used for the model were alpha = 0 and lambda = 1e-04.
```

Fit model with variables selected from backward selection

```
back_model <- train(
    back_select$formula,
    data = train_set,
    method = "glmnet",
    metric = "ROC",
    trControl = train_ctrl,
    tuneGrid = glm_grid
)

back_model
```

```
## glmnet
##
## 398 samples
##  18 predictor
##   2 classes: 'B', 'M'
##
## No pre-processing
## Resampling: Cross-Validated (5 fold)
## Summary of sample sizes: 79, 79, 80, 81, 79
## Resampling results across tuning parameters:
##
##   alpha  lambda  ROC        Sens       Spec
##   0      0.0001  0.9864075  0.9746141  0.8847175
##   0      0.1112  0.9844619  0.9766446  0.8583796
##   0      0.2223  0.9826679  0.9796903  0.8336946
##   0      0.3334  0.9813837  0.9817259  0.8172063
##   0      0.4445  0.9804572  0.9817259  0.8040509
##   0      0.5556  0.9797062  0.9817259  0.7794066
##   0      0.6667  0.9790299  0.9827411  0.7728492
##   0      0.7778  0.9785379  0.9837563  0.7629725
##   0      0.8889  0.9778948  0.9847716  0.7547622
##   0      1.0000  0.9774693  0.9847716  0.7481913
##   1      0.0001  0.9745458  0.9512431  0.9145102
##   1      0.1112  0.9723028  0.9857764  0.7447907
##   1      0.2223  0.9679694  0.9949239  0.5837014
##   1      0.3334  0.9647793  1.0000000  0.1317437
##   1      0.4445  0.5000000  1.0000000  0.0000000
##   1      0.5556  0.5000000  1.0000000  0.0000000
##   1      0.6667  0.5000000  1.0000000  0.0000000
##   1      0.7778  0.5000000  1.0000000  0.0000000
##   1      0.8889  0.5000000  1.0000000  0.0000000
##   1      1.0000  0.5000000  1.0000000  0.0000000
##
## ROC was used to select the optimal model using the largest value.
## The final values used for the model were alpha = 0 and lambda = 1e-04.
```

Fit model with variables selected from backward selection

```
back_model <- train(
    back_select$formula,
    data = train_set,
    method = "glmnet",
    metric = "ROC",
    trControl = train_ctrl,
    tuneGrid = glm_grid
)

back_model
```

```
## glmnet
##
## 398 samples
##  18 predictor
##   2 classes: 'B', 'M'
##
## No pre-processing
## Resampling: Cross-Validated (5 fold)
## Summary of sample sizes: 79, 79, 80, 81, 79
## Resampling results across tuning parameters:
##
##   alpha  lambda  ROC        Sens       Spec
##   0      0.0001  0.9864075  0.9746141  0.8847175
##   0      0.1112  0.9844619  0.9766446  0.8583796
##   0      0.2223  0.9826679  0.9796903  0.8336946
##   0      0.3334  0.9813837  0.9817259  0.8172063
##   0      0.4445  0.9804572  0.9817259  0.8040509
##   0      0.5556  0.9797062  0.9817259  0.7794066
##   0      0.6667  0.9790299  0.9827411  0.7728492
##   0      0.7778  0.9785379  0.9837563  0.7629725
##   0      0.8889  0.9778948  0.9847716  0.7547622
##   0      1.0000  0.9774693  0.9847716  0.7481913
##   1      0.0001  0.9745458  0.9512431  0.9145102
##   1      0.1112  0.9723028  0.9857764  0.7447907
##   1      0.2223  0.9679694  0.9949239  0.5837014
##   1      0.3334  0.9647793  1.0000000  0.1317437
##   1      0.4445  0.5000000  1.0000000  0.0000000
##   1      0.5556  0.5000000  1.0000000  0.0000000
##   1      0.6667  0.5000000  1.0000000  0.0000000
##   1      0.7778  0.5000000  1.0000000  0.0000000
##   1      0.8889  0.5000000  1.0000000  0.0000000
##   1      1.0000  0.5000000  1.0000000  0.0000000
```

```
##
## ROC was used to select the optimal model using the largest value.
## The final values used for the model were alpha = 0 and lambda = 1e-04.
```

Fit model with variables selected from entropy

```
setDT(x)
#selector predictors with importance of more than 0.05
predictors <- x[importance > 0.05, attributes]

entropy_predctors <- train_set[, ..predictors]
entropy_y <- train_set$diagnosis
entropy_model <- train(
    entropy_predctors,
    entropy_y,
    method = "glm",
    metric = "ROC",
    trControl = train_ctrl
)

entropy_model
```

```
## Generalized Linear Model
##
## 398 samples
##  25 predictor
##   2 classes: 'B', 'M'
##
## No pre-processing
## Resampling: Cross-Validated (5 fold)
## Summary of sample sizes: 79, 79, 80, 81, 79
## Resampling results:
##
##   ROC        Sens       Spec
##   0.9335298  0.9227701  0.8288714
```

Fit model with variables selected Recursive Feature Elimination

```
recu_pred <- lmProfile$optVariables
recursive_predctors <- train_set[, ..recu_pred]
recursive_y <- train_set$diagnosis
recu_model <- train(
    recursive_predctors,
    recursive_y,
    method = "glm",
    metric = "ROC",
    trControl = train_ctrl
)

recu_model
```

```
## Generalized Linear Model
##
## 398 samples
##  16 predictor
##   2 classes: 'B', 'M'
##
## No pre-processing
## Resampling: Cross-Validated (5 fold)
## Summary of sample sizes: 79, 79, 80, 81, 79
## Resampling results:
##
##   ROC        Sens       Spec
##   0.9275906  0.9268103  0.8437339
```

Full model test accuracy

```
for_glm <- predict(full_model, test_set, type = "prob")


for_glm1 <- ifelse(for_glm[, "M"] > 0.5, "M", "B")
for_glm1 <- factor(for_glm1, levels = levels(test_set$diagnosis))



confusionMatrix(for_glm1, test_set$diagnosis,positive = "M")
```

```
## Confusion Matrix and Statistics
##
##           Reference
## Prediction   B   M
##          B 110   4
##          M   1  56
##
##                Accuracy : 0.9708
##                  95% CI : (0.9331, 0.9904)
##     No Information Rate : 0.6491
##     P-Value [Acc > NIR] : <2e-16
##
##                   Kappa : 0.9351
##
##  Mcnemar's Test P-Value : 0.3711
##
##             Sensitivity : 0.9333
##             Specificity : 0.9910
##          Pos Pred Value : 0.9825
##          Neg Pred Value : 0.9649
##              Prevalence : 0.3509
##          Detection Rate : 0.3275
##    Detection Prevalence : 0.3333
##       Balanced Accuracy : 0.9622
##
##        'Positive' Class : M
```

```
##
```

Forward test accuracy

```
for_glm <- predict(forward_model, test_set, type = "prob")


for_glm1 <- ifelse(for_glm[, "M"] > 0.5, "M", "B")
for_glm1 <- factor(for_glm1, levels = levels(test_set$diagnosis))



confusionMatrix(for_glm1, test_set$diagnosis,positive = "M")
```

```
## Confusion Matrix and Statistics
##
##           Reference
## Prediction   B   M
##          B 110   4
##          M   1  56
##
##                 Accuracy : 0.9708
##                   95% CI : (0.9331, 0.9904)
##     No Information Rate : 0.6491
##     P-Value [Acc > NIR] : <2e-16
##
##                    Kappa : 0.9351
##
##  Mcnemar's Test P-Value : 0.3711
##
##              Sensitivity : 0.9333
##              Specificity : 0.9910
##           Pos Pred Value : 0.9825
##           Neg Pred Value : 0.9649
##               Prevalence : 0.3509
##           Detection Rate : 0.3275
##     Detection Prevalence : 0.3333
##        Balanced Accuracy : 0.9622
##
##         'Positive' Class : M
##
```

Backward test accuracy

```
for_glm <- predict(back_model, test_set, type = "prob")


for_glm1 <- ifelse(for_glm[, "M"] > 0.5, "M", "B")
for_glm1 <- factor(for_glm1, levels = levels(test_set$diagnosis))
```

```r
confusionMatrix(for_glm1, test_set$diagnosis,positive = "M")
```

```
## Confusion Matrix and Statistics
##
##           Reference
## Prediction   B   M
##          B 110   4
##          M   1  56
##
##                Accuracy : 0.9708
##                  95% CI : (0.9331, 0.9904)
##     No Information Rate : 0.6491
##     P-Value [Acc > NIR] : <2e-16
##
##                   Kappa : 0.9351
##
##  Mcnemar's Test P-Value : 0.3711
##
##             Sensitivity : 0.9333
##             Specificity : 0.9910
##          Pos Pred Value : 0.9825
##          Neg Pred Value : 0.9649
##              Prevalence : 0.3509
##          Detection Rate : 0.3275
##    Detection Prevalence : 0.3333
##       Balanced Accuracy : 0.9622
##
##        'Positive' Class : M
##
```

entropy method test accuracy

```r
for_glm <- predict(entropy_model, test_set, type = "prob")


for_glm1 <- ifelse(for_glm[, "M"] > 0.5, "M", "B")
for_glm1 <- factor(for_glm1, levels = levels(test_set$diagnosis))



confusionMatrix(for_glm1, test_set$diagnosis,positive = "M")
```

```
## Confusion Matrix and Statistics
##
##           Reference
## Prediction   B   M
##          B 106   2
##          M   5  58
```

```
##
##                Accuracy : 0.9591
##                  95% CI : (0.9175, 0.9834)
##     No Information Rate : 0.6491
##     P-Value [Acc > NIR] : <2e-16
##
##                   Kappa : 0.9112
##
##  Mcnemar's Test P-Value : 0.4497
##
##             Sensitivity : 0.9667
##             Specificity : 0.9550
##          Pos Pred Value : 0.9206
##          Neg Pred Value : 0.9815
##              Prevalence : 0.3509
##          Detection Rate : 0.3392
##    Detection Prevalence : 0.3684
##       Balanced Accuracy : 0.9608
##
##        'Positive' Class : M
##
```

Recursive Feature Elimination method test accuracy

```r
for_glm <- predict(recu_model, test_set, type = "prob")


for_glm1 <- ifelse(for_glm[, "M"] > 0.5, "M", "B")
for_glm1 <- factor(for_glm1, levels = levels(test_set$diagnosis))



confusionMatrix(for_glm1, test_set$diagnosis,positive = "M")
```

```
## Confusion Matrix and Statistics
##
##           Reference
## Prediction   B   M
##          B 104   0
##          M   7  60
##
##                Accuracy : 0.9591
##                  95% CI : (0.9175, 0.9834)
##     No Information Rate : 0.6491
##     P-Value [Acc > NIR] : < 2e-16
##
##                   Kappa : 0.9125
##
##  Mcnemar's Test P-Value : 0.02334
##
##             Sensitivity : 1.0000
##             Specificity : 0.9369
##          Pos Pred Value : 0.8955
##          Neg Pred Value : 1.0000
##              Prevalence : 0.3509
##          Detection Rate : 0.3509
##    Detection Prevalence : 0.3918
##       Balanced Accuracy : 0.9685
##
##        'Positive' Class : M
##
```