Kyungjin Cho

Question 1

Building a Logistic Regression Model

```
#Data import
dat <- read.table("/Users/irenecho/Desktop/stat443/datafiles/rhcdata.txt", header=T, sep
= " ")
dat <- dat[-1]
```

```
# replace categorical variables into binary values
dat$meanbp1 <- replace(dat$meanbp1, dat$meanbp1 <68.5, 1)
dat$meanbp1 <- replace(dat$meanbp1, dat$meanbp1 >= 68.5, 0)
dat$cat1 <- replace(dat$cat1, dat$cat1 == c("CHF"), 1)
dat$cat1 <- replace(dat$cat1, dat$cat1 == c("MOSF w/Sepsis"), 1)
dat$cat1 <- replace(dat$cat1, dat$cat1 == "COPD", 0)
dat$cat1 <- replace(dat$cat1, dat$cat1 == "MOSF w/Malignancy", 0)
dat$cat1 <- replace(dat$cat1, dat$cat1 == "ARF", 0)
dat$cat1 <- replace(dat$cat1, dat$cat1 == "Coma", 0)
dat$cat1 <- replace(dat$cat1, dat$cat1 == "Cirrhosis", 0)
dat$cat1 <- replace(dat$cat1, dat$cat1 == "Colon Cancer", 0)
dat$cat1 <- replace(dat$cat1, dat$cat1 == "Lung Cancer", 0)

dat$cat2 <- replace(dat$cat2, dat$cat2 == c("CHF"), 1)
dat$cat2 <- replace(dat$cat2, dat$cat2 == c("MOSF w/Sepsis"), 1)
dat$cat2 <- replace(dat$cat2, dat$cat2 == "COPD", 0)
dat$cat2 <- replace(dat$cat2, dat$cat2 == "MOSF w/Malignancy", 0)
dat$cat2 <- replace(dat$cat2, dat$cat2 == "ARF", 0)
dat$cat2 <- replace(dat$cat2, dat$cat2 == "Coma", 0)
dat$cat2 <- replace(dat$cat2, dat$cat2 == "Cirrhosis", 0)
dat$cat2 <- replace(dat$cat2, dat$cat2 == "Colon Cancer", 0)
dat$cat2 <- replace(dat$cat2, dat$cat2 == "Lung Cancer", 0)
dat$pafi1 <- replace(dat$pafi1, dat$pafi1 <= 266.15625, 1)
dat$pafi1 <- replace(dat$pafi1, dat$pafi1 <= 266.15625, 0)
dat$ninsclas <- replace(dat$ninsclas, dat$ninsclas == "Private", 1)
dat$ninsclas <- replace(dat$ninsclas, dat$ninsclas == "Medicaid", 0)
dat$ninsclas <- replace(dat$ninsclas, dat$ninsclas == "Private & Medicare", 1)
dat$ninsclas <- replace(dat$ninsclas, dat$ninsclas == "Medicare & Medicaid", 0)

dat$swang1 <- as.factor(dat$swang1)
dat$death <- as.factor(dat$death)
dat$cardiohx <- as.factor(dat$cardiohx)
dat$chfhx <- as.factor(dat$chfhx)
dat$dementhx <- as.factor(dat$dementhx)
dat$psychhx <- as.factor(dat$psychhx)
dat$chrpulhx <- as.factor(dat$chrpulhx)
dat$renalhx <- as.factor(dat$renalhx)
dat$liverhx <- as.factor(dat$liverhx)
dat$gibledhx <- as.factor(dat$gibledhx)
dat$malighx <- as.factor(dat$malighx)
dat$transhx <- as.factor(dat$transhx)
dat$amihx <- as.factor(dat$amihx)
dat$dth30 <- as.factor(dat$dth30)
dat$cat1 <- as.factor(dat$cat1)
dat$cat2 <- as.factor(dat$cat2)
dat$ca <- as.factor(dat$ca)
dat$sex <- as.factor(dat$sex)
dat$dnr1 <- as.factor(dat$dnr1)
dat$ninsclas <- as.factor(dat$ninsclas)
dat$resp <- as.factor(dat$resp)
dat$card <- as.factor(dat$card)
dat$neuro <- as.factor(dat$neuro)
dat$gastr <- as.factor(dat$gastr)
```

```
dat$renal <- as.factor(dat$renal)
dat$meta <- as.factor(dat$meta)
dat$hema <- as.factor(dat$hema)
dat$seps <- as.factor(dat$seps)
dat$trauma <- as.factor(dat$trauma)
dat$ortho <- as.factor(dat$ortho)
dat$race <- as.factor(dat$race)
dat$income <- as.factor(dat$income)
dat$immunhx <- as.factor(dat$immunhx)
```

```
#convert NA to mean in integer strings
dat$dschdte[which(is.na(dat$dschdte))] = mean(dat$dschdte, na.rm = T)
dat$dthdte[which(is.na(dat$dthdte))] = mean(dat$dthdte, na.rm = T)
dat$hrt1[which(is.na(dat$hrt1))] = mean(dat$hrt1, na.rm = T)
dat$resp1[which(is.na(dat$resp1))] = mean(dat$resp1, na.rm = T)
dat$wtkilo1[which(is.na(dat$wtkilo1))] = mean(dat$wtkilo1, na.rm = T)
dat$urin1[which(is.na(dat$urin1))] = mean(dat$urin1, na.rm = T)
dat$adld3p[which(is.na(dat$adld3p))] = mean(dat$adld3p, na.rm = T)

#delete the NA rows of binary cat2 variable
dat = dat[!is.na(dat$cat2), ]
```

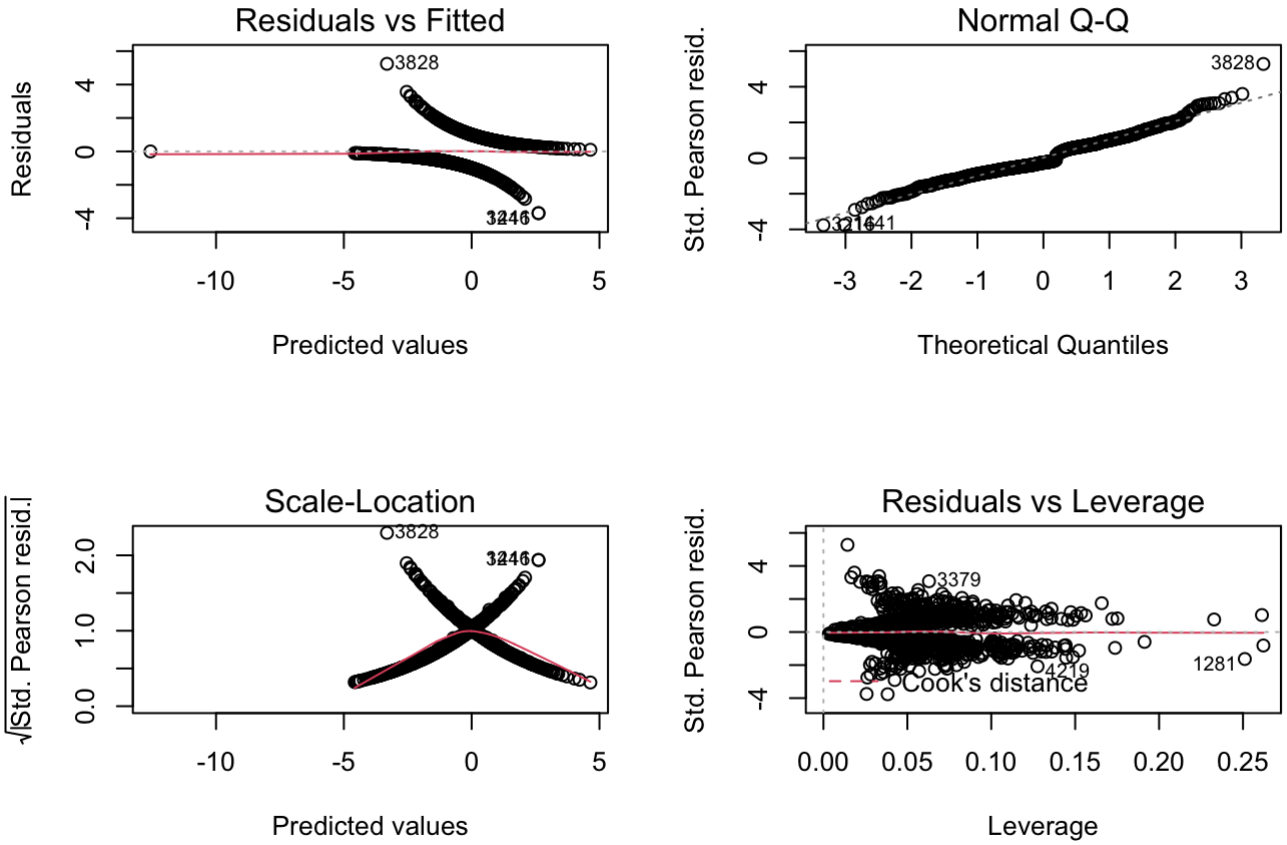
```
dat.lm <- glm(swang1 ~ . , data= dat, family=binomial)
summary(dat.lm)
```

```
##
## Call:
## glm(formula = swang1 ~ ., family = binomial, data = dat)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -2.3182  -0.8365  -0.3536   0.8788   2.5880
##
## Coefficients:
##                       Estimate Std. Error z value Pr(>|z|)
## (Intercept)          2.259e+01  6.682e+00   3.381 0.000721 ***
## cat11                1.071e+00  6.446e-01   1.662 0.096486 .
## cat21                1.141e+00  1.979e-01   5.764 8.21e-09 ***
## caNo                 9.137e-01  6.685e-01   1.367 0.171684
## caYes                1.558e-01  2.487e-01   0.627 0.530969
## sadmdte             -6.108e-03  3.737e-03  -1.635 0.102142
## dschdte              6.604e-03  3.595e-03   1.837 0.066171 .
## dthdte              -5.396e-04  3.442e-04  -1.568 0.116956
## lstctdte            -7.747e-05  9.215e-04  -0.084 0.933002
## death1              -1.538e-02  2.429e-01  -0.063 0.949511
## cardiohx1            3.852e-02  2.634e-01   0.146 0.883716
## chfhx1              -1.816e-02  2.680e-01  -0.068 0.945988
## dementhx1           -2.309e-01  2.655e-01  -0.870 0.384480
## psychhx1            -1.478e-01  2.872e-01  -0.515 0.606841
## chrpulhx1           -1.510e-01  2.253e-01  -0.670 0.502725
## renalhx1             4.082e-01  4.306e-01   0.948 0.343166
## liverhx1            -5.283e-02  3.772e-01  -0.140 0.888639
## gibledhx1           -4.188e-01  5.189e-01  -0.807 0.419635
## malighx1             3.988e-01  5.873e-01   0.679 0.497126
## immunhx1            -1.526e-01  1.672e-01  -0.913 0.361348
## transhx1             3.411e-01  2.368e-01   1.440 0.149807
## amihx1               7.916e-01  4.117e-01   1.923 0.054513 .
## age                 -1.300e-04  5.780e-03  -0.022 0.982056
## sexMale              1.343e-01  1.588e-01   0.845 0.397858
## edu                  4.190e-02  2.661e-02   1.575 0.115352
## surv2md1            -1.162e+00  6.575e-01  -1.768 0.077099 .
## das2d3pc             2.023e-02  1.582e-02   1.279 0.200991
## t3d30                5.853e-04  1.470e-02   0.040 0.968245
## dth301               4.642e-01  3.414e-01   1.360 0.173837
## aps1                 1.917e-02  6.912e-03   2.773 0.005547 **
## scoma1              -2.493e-03  3.152e-03  -0.791 0.428924
## meanbp1              3.961e-01  1.826e-01   2.170 0.030035 *
## wblc1                5.047e-03  4.834e-03   1.044 0.296474
## hrt1                 4.029e-03  2.498e-03   1.613 0.106789
## resp1               -2.251e-02  6.316e-03  -3.564 0.000365 ***
## temp1               -3.602e-02  4.047e-02  -0.890 0.373402
## pafi1               -2.581e-03  5.124e-04  -5.037 4.73e-07 ***
## alb1                 2.070e-01  1.120e-01   1.849 0.064489 .
## hema1               -1.458e-02  1.046e-02  -1.394 0.163332
## bili1               -1.805e-02  1.982e-02  -0.910 0.362649
## crea1               -7.661e-02  4.993e-02  -1.534 0.124970
## sod1                -1.489e-02  9.208e-03  -1.617 0.105943
## pot1                -2.822e-01  7.190e-02  -3.924 8.70e-05 ***
```

```
## paco21                -2.723e-02  7.166e-03  -3.800 0.000145 ***
## ph1                   -2.967e+00  8.024e-01  -3.697 0.000218 ***
## wtkilo1                2.596e-02  4.606e-03   5.636 1.74e-08 ***
## dnr1Yes               -7.394e-01  2.378e-01  -3.109 0.001877 **
## ninsclas1              1.510e-01  2.230e-01   0.677 0.498240
## ninsclasMedicare      -3.595e-01  2.615e-01  -1.375 0.169279
## ninsclasNo insurance   9.163e-02  3.457e-01   0.265 0.790933
## respYes                3.638e-02  1.629e-01   0.223 0.823243
## cardYes                6.063e-01  1.969e-01   3.079 0.002079 **
## neuroYes              -1.823e-01  2.809e-01  -0.649 0.516376
## gastrYes               2.597e-01  2.340e-01   1.110 0.267118
## renalYes               2.853e-01  2.969e-01   0.961 0.336588
## metaYes                3.598e-02  3.653e-01   0.098 0.921537
## hemaYes               -4.664e-01  2.368e-01  -1.970 0.048872 *
## sepsYes                4.273e-01  1.667e-01   2.564 0.010357 *
## traumaYes              1.814e+00  9.858e-01   1.840 0.065796 .
## orthoYes              -8.597e+00  3.247e+02  -0.026 0.978879
## adld3p                -9.529e-02  1.067e-01  -0.893 0.371655
## urin1                 -1.505e-06  6.567e-05  -0.023 0.981715
## raceother              3.433e-02  3.386e-01   0.101 0.919241
## racewhite             -6.369e-02  2.154e-01  -0.296 0.767460
## income$11-$25k         5.215e-01  3.044e-01   1.713 0.086671 .
## income$25-$50k         2.541e-01  2.959e-01   0.859 0.390381
## incomeUnder $11k       5.934e-01  3.033e-01   1.956 0.050418 .
## ptid                   2.030e-06  2.440e-05   0.083 0.933705
## survtime               9.888e-04  5.684e-04   1.740 0.081908 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 1594.8  on 1169  degrees of freedom
## Residual deviance: 1222.4  on 1101  degrees of freedom
##   (30 observations deleted due to missingness)
## AIC: 1360.4
##
## Number of Fisher Scoring iterations: 11
```

```
par(mfrow= c(2,2))
plot(dat.lm)
```

```
## Warning: not plotting observations with leverage one:
##   529
```



In order to reproduce the general logistic regression model, first, I have implemented “which” and “is.na” function to the columns with integer variables in order to get the mean of each columns and replace them with NA values.

Although I got the mean and replaced NA with the mean, I have omitted the missing values from cat1. I used this method because the data sets were too huge to use the Mice package in my computer. Also, getting a mean value of binary value and replace it with missing values is not an efficient way to deal with them especially with the columns with two categories. Therefore, I decided to remove them.

However, in the future, I do not think that this is a good idea to deal with the missing values because lot of datasets were gone by using this method.

According to the summary of my logistic model, I would build a logistic regression model of

$$y_i = Intercept + cat1 * b_1 + cat2 * b_2 + trauma * b_3$$

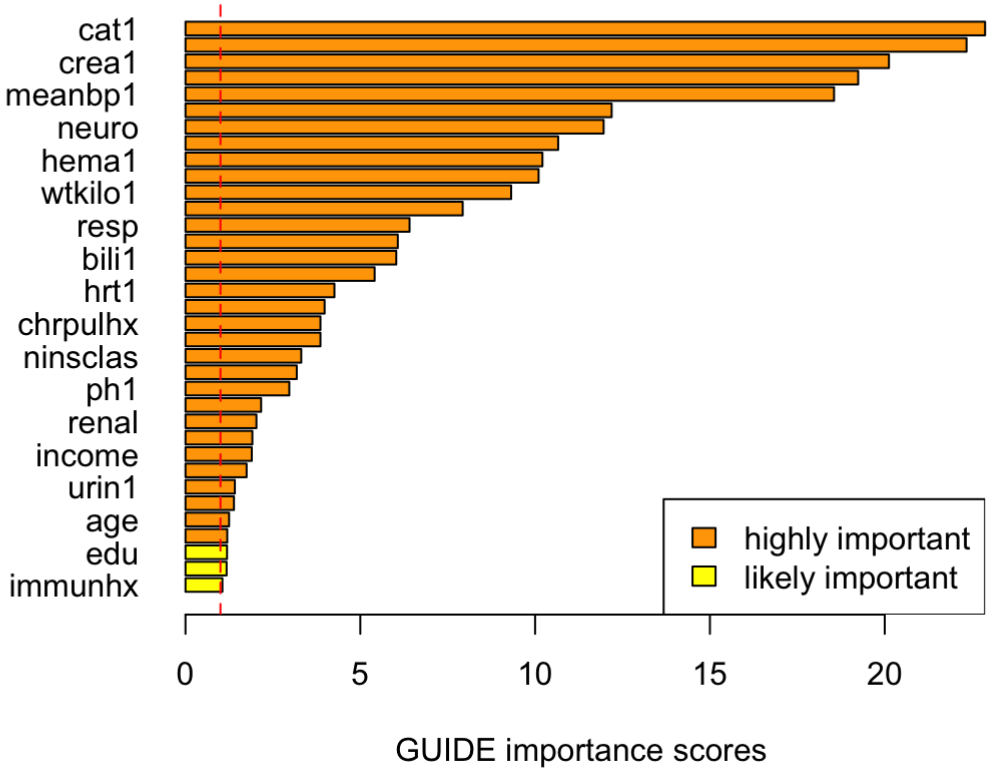
$$y_i = 22.59 + b_1 * 1.071 + b_2 * 1.141 + b_3 * 1.814$$

because cat1, cat2 and trauma have the highest coefficient. Also, their p-values are very small that we reject the null hypothesis concluding that they are in fact, one of the most important factors measuring RHC data.

Question 2

Reproducing GUIDE Importance scores plot

```
par(las=1,mar=c(5,12,4,2),cex=1)
leg.col <- c("orange","yellow")
leg.txt <- c("highly important","likely important")
x <- read.table("/Users/irenecho/Desktop/stat443/datafiles/imp.scr",header=TRUE)
score <- x$Score
vars <- x$Variable
type <- x$Type
barcol <- rep("orange",length(vars))
barcol[type == "L"] <- "yellow"
barcol[type == "U"] <- "cyan"
n <- sum(x$Type != "U")
barplot(rev(score[1:n]),names.arg=rev(vars[1:n]),col=rev(barcol[1:n]),horiz=TRUE, xlab=
"GUIDE importance scores")
abline(v=1,col="red",lty=2)
legend("bottomright",legend=leg.txt,fill=leg.col)
```



The guide importance score shows that cat1 is the most important variable followed by crea1, meanbp1 and neuro. It also provides the less likely important features such as education and immunehx.

Question 3

Effects of Logistic model

The logistic model I have created highlights the importance of cat1, cat2 and trauma while GUIDE tree and importance score signifies cat1, aps1, and crea1. From my equation, aps1 and crea1 are also considered as highly important; however, not as much as the GUIDE importance score.

My regression model considers education, immune history, sod1 as less likely important factors which GUIDE tree and importance score also agree. GUIDE and my logistic model concludes that cat1, cat2, trauma, aps1 and crea1 are more important than education, immune history and sod1.

However, what my logistic model and GUIDE's importance score do not agree is the order of importance. I had cat2 as the most important variable while GUIDE's importance score highlights the fact that cat1 is the most significant variable.

By looking through the data, I believe that there are some interactions that are different from the GUIDE importance scoring and tree when I work manually by hand. Also, I found it difficult to hand code the replace() function for each categorical variables and assigning the binary values.

Question 4

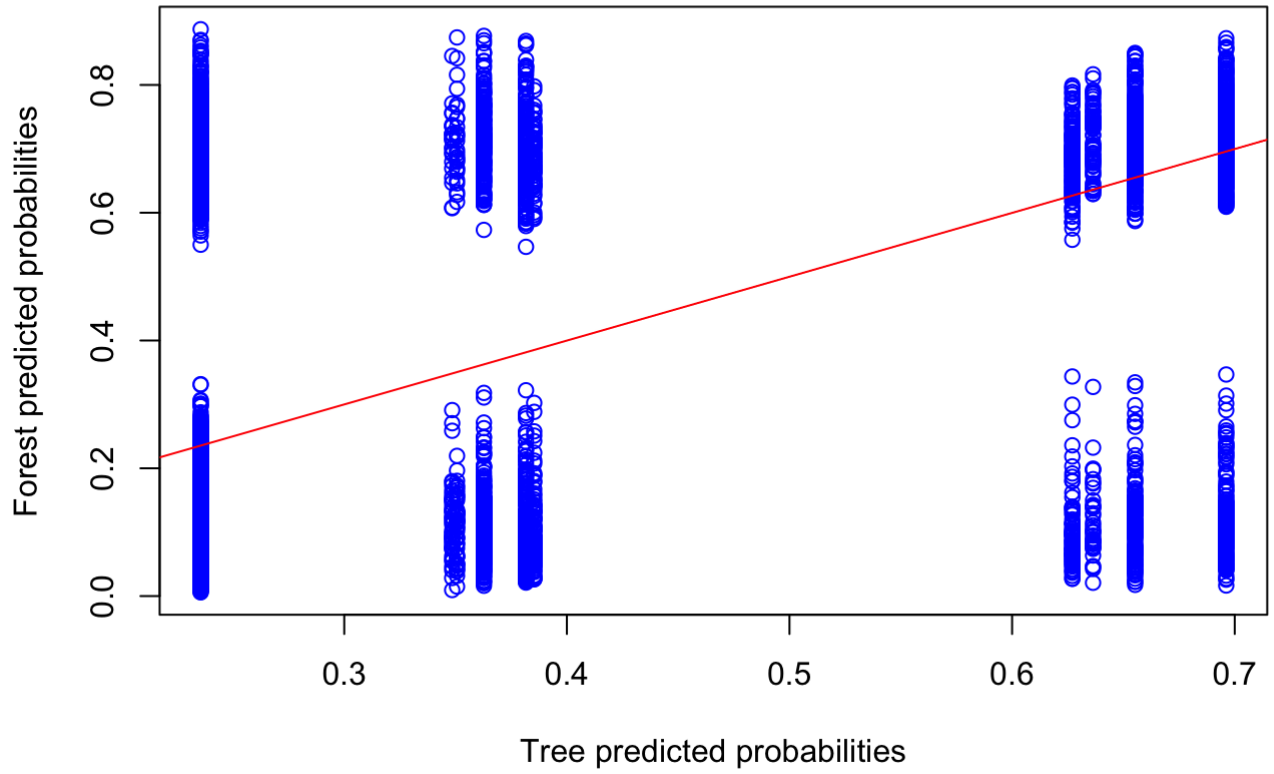
Building a GUIDE Forest Model

```
#creating guide description file in R
dat <- read.table("/Users/irenecho/Desktop/stat443/datafiles/rhcdata.txt",header=TRUE)
nvar <- ncol(dat)
varnames <- names(dat)
varnames
```

```
##  [1] "X"         "cat1"     "cat2"     "ca"        "sadmdte"  "dschdte"
##  [7] "dthdte"    "lstctdte" "death"    "cardiohx"  "chfhx"    "dementhx"
## [13] "psychhx"   "chrpulhx" "renalhx"  "liverhx"   "gibledhx" "malighx"
## [19] "immunhx"   "transhx"  "amihx"    "age"       "sex"      "edu"
## [25] "surv2md1"  "das2d3pc" "t3d30"    "dth30"     "aps1"     "scoma1"
## [31] "meanbp1"   "wblc1"    "hrt1"     "resp1"     "temp1"    "pafi1"
## [37] "alb1"      "hema1"    "bili1"    "crea1"     "sod1"     "pot1"
## [43] "paco21"    "ph1"      "swang1"   "wtkilo1"   "dnr1"     "ninsclas"
## [49] "resp"      "card"     "neuro"    "gastr"     "renal"    "meta"
## [55] "hema"      "seps"     "trauma"   "ortho"     "adld3p"   "urin1"
## [61] "race"      "income"   "ptid"     "survtime"
```

```
roles <- rep("s",nvar)
c.vars <- c("cat1", "cat2", "ca", "sadmdte", "dschdte", "dthdte", "lstctdte", "death",
"cardiohx", "chfhx", "dementhx", "psychhx", "chrpulhx", "renalhx", "liverhx", "gibledhx"
, "malighx", "immunhx", "transhx", "amihx",  "age", "sex",  "edu", "surv2md1", "das2d3p
c", "t3d30" ,"dth30", "aps1", "scoma1", "meanbp1", "wblc1", "hrt1",  "resp1", "temp1",
"pafi1", "alb1", "hema1", "bili1", "crea1", "sod1",  "pot1",  "paco21","ph1", "wtkilo1",
"dnr1", "ninsclas", "resp", "card", "neuro", "gastr",  "renal","meta", "hema", "seps",
"trauma", "ortho","adld3p",  "urin1",   "race",  "income")
roles[varnames %in% c.vars] <- "c"
x.vars <- c("X", "survtime", "ptid")
roles[varnames %in% x.vars] <- "x"
d.var <- "swang1"
roles[varnames %in% d.var] <- "d"
write("rhcdata.txt",file="desc.txt")
write("NA",file="desc.txt",append=TRUE)
write("2",file="desc.txt",append=TRUE)
write.table(cbind(1:nvar,varnames,roles),file="desc.txt",
            row.names=FALSE,col.names=FALSE,quote=FALSE,append=TRUE)
```

```
#R code for plotting predicted probabilities
tree <- read.table("/Users/irenecho/Desktop/stat443/datafiles/classpred.txt",header=TRUE
)
forest <- read.table("/Users/irenecho/Desktop/stat443/datafiles/forestpred.txt", header=
T)
tree.p <- tree[,6]
forest.p <- forest[,3]
plot(forest.p ~ tree.p, xlab="Tree predicted probabilities",ylab="Forest predicted proba
bilities",col="blue")
abline(c(0,1),col="red")
```



I believe that the GUIDE tree model to estimate predicted probabilities are more accurate. Since I have manually created the GUIDE description file and classified c, x and s variables, I think that GUIDE logistic regression tree is a better predictor of RHC data.

Question 5

Input and Output files

(refer to the next page)

Input file for importance scoring (Imp.in)

```
GUIDE       (do not edit this file unless you know what you are doing)
 36.2       (version of GUIDE that generated this file)
 2          (1=model fitting, 2=importance or DIF scoring, 3=data conversion)
"imp.out"   (name of output file)
 1          (1=classification, 2=regression, 3=propensity score grouping)
 1          (1=univariate and interaction splits, 2=skip interactions)
"rhcdsc1.txt"  (name of data description file)
 1          (1=estimated priors, 2=equal priors, 3=other priors)
 1          (1=unit misclassification costs, 2=other)
 2          (1=split point from quantiles, 2=use exhaustive search)
 1          (1=default max. number of split levels, 2=specify no. in next line)
 1          (1=default min. node size, 2=specify min. value in next line)
 2          (0=no LaTeX code, 1=tree without node numbers, 2=tree with node numbers)
"imp.tex"  (latex file name)
 1          (1=color terminal nodes, 2=no colors)
 2          (0=#errors, 1=sample sizes, 2=sample proportions, 3=posterior probs, 4=nothing)
 1          (1=no storage, 2=store fit and split variables, 3=store split variables and values)
 1          (1=do not create description file for selected variables, 2=create the file)
 1          (1=create file for importance scores, 2=do not create)
"imp.scr"  (file name for importance scores)
  1     (rank of top variable to split root node)
```

Output file for importance scoring (Imp.out)

```
         GGG   U   U  I  DDDD   EEEE
        G   G  U   U  I  D   D  E
        G      U   U  I  D   D  E
        G  GG  U   U  I  D   D  EEE
        G   G  U   U  I  D   D  E
        G   G  U   U  I  D   D  E
         GGG    UUU   I  DDDD   EEEE

GUIDE Classification and Regression Trees and Forests
Version 36.2 (Build date: January 8, 2021)
Compiled with GFortran 10.2.0 on macOS Big Sur 11.1
Copyright (c) 1997-2020 Wei-Yin Loh. All rights reserved.
This software is based upon work supported by the U.S. Army Research Office,
the National Science Foundation and the National Institutes of Health.

This job was started on 02/23/21 at 17:09

Classification tree
No pruning
Data description file: rhcdsc1.txt
Training sample file: rhcdata.txt
Missing value code: NA
```

Records in data file start on line 2
20 N variables changed to S
D variable is swang1
Number of records in data file: 5735
Length of longest entry in data file: 19
Missing values found among categorical variables
Separate categories will be created for missing categorical variables
Missing values found among non-categorical variables
Number of classes: 2
Training sample class proportions of D variable swang1:

Class	#Cases	Proportion
NoRHC	3551	0.61918047
RHC	2184	0.38081953

Total #cases	#cases w/ miss. D	#missing ord. vals	#X-var	#N-var	#F-var	#S-var
5735	0	3443	13	0	0	20

#P-var	#M-var	#B-var	#C-var	#I-var
0	0	0	30	0

Number of cases used for training: 5735
Number of split variables: 50
Number of cases excluded due to 0 weight or missing D: 0

Importance scoring of variables
Simple node models
Estimated priors
Unit misclassification costs
Univariate split highest priority
Interaction splits 2nd priority; no linear splits
Split values for N and S variables based on exhaustive search
Maximum number of split levels: 4
Minimum node sample size: 57
Starting 300 permutations to standardize means of importance scores
Finished permutations to standardize means of importance scores
95 and 99% thresholds for unadjusted importance scores = 21.276 25.602

Note: final tree is shorter due to pruning of sibling nodes with same predicted values.
Structure of final tree. Each terminal node is marked with a T.

Node cost is node misclassification cost divided by number of training cases

Node label	Total cases	Train cases	Predicted class	Node cost	Split variables	Interacting variable
1	5735	5735	NoRHC	3.808E-01	cat1	
2	1683	1683	RHC	4.599E-01	meanbp1	
4	1117	1117	RHC	3.796E-01	pafi1	
8T	655	655	RHC	3.038E-01	resp1	
9	462	462	RHC	4.870E-01	ninsclas	
18T	244	244	RHC	3.730E-01	bili1	
19T	218	218	NoRHC	3.853E-01	card	
5	566	566	NoRHC	3.816E-01	alb1	
10	158	158	RHC	4.810E-01	pafi1	

20T	72	72	RHC	3.194E-01	-
21T	86	86	NoRHC	3.837E-01	-
11T	408	408	NoRHC	3.284E-01	bili1
3	4052	4052	NoRHC	3.147E-01	pafi1
6	1292	1292	NoRHC	4.837E-01	resp
12	581	581	RHC	4.200E-01	dnr1
24T	515	515	RHC	3.903E-01	cat1
25T	66	66	NoRHC	3.485E-01	-
13	711	711	NoRHC	4.051E-01	seps
26T	110	110	RHC	3.636E-01	-
27T	601	601	NoRHC	3.627E-01	aps1
7T	2760	2760	NoRHC	2.355E-01	aps1

Number of terminal nodes of final tree: 11
Total number of nodes of final tree: 21
Second best split variable (based on curvature test) at root node is aps1

Classification tree:
For categorical variable splits, values not in training data go to the right

Node 1: cat1 = "CHF", "MOSF w/Sepsis"
  Node 2: meanbp1 <= 68.500000 or NA
    Node 4: pafi1 <= 266.15625
      Node 8: RHC
    Node 4: pafi1 > 266.15625 or NA
      Node 9: ninsclas = "No insurance", "Private", "Private & Medicare"
        Node 18: RHC
      Node 9: ninsclas /= "No insurance", "Private", "Private & Medicare"
        Node 19: NoRHC
  Node 2: meanbp1 > 68.500000
    Node 5: alb1 <= 2.9499511
      Node 10: pafi1 <= 216.25000
        Node 20: RHC
      Node 10: pafi1 > 216.25000 or NA
        Node 21: NoRHC
    Node 5: alb1 > 2.9499511 or NA
      Node 11: NoRHC
Node 1: cat1 /= "CHF", "MOSF w/Sepsis"
  Node 3: pafi1 <= 142.35938
    Node 6: resp = "No"
      Node 12: dnr1 = "No"
        Node 24: RHC
      Node 12: dnr1 /= "No"
        Node 25: NoRHC
    Node 6: resp /= "No"
      Node 13: seps = "Yes"
        Node 26: RHC
      Node 13: seps /= "Yes"
        Node 27: NoRHC
  Node 3: pafi1 > 142.35938 or NA
    Node 7: NoRHC

```
************************************************************

 Predictor means below are means of cases with no missing values.

Node 1: Intermediate node
 A case goes into Node 2 if cat1 = "CHF", "MOSF w/Sepsis"
 cat1 mode = "ARF"
 Class      Number   Posterior
 NoRHC        3551  0.6192E+00
 RHC          2184  0.3808E+00
 Number of training cases misclassified = 2184
 Predicted class is NoRHC
 ----------------------------
Node 2: Intermediate node
 A case goes into Node 4 if meanbp1 <= 68.500000 or NA
 meanbp1 mean = 72.674985
 Class      Number   Posterior
 NoRHC         774  0.4599E+00
 RHC           909  0.5401E+00
 Number of training cases misclassified = 774
 Predicted class is RHC
 ----------------------------
Node 4: Intermediate node
 A case goes into Node 8 if pafi1 <= 266.15625
 pafi1 mean = 241.37331
 Class      Number   Posterior
 NoRHC         424  0.3796E+00
 RHC           693  0.6204E+00
 Number of training cases misclassified = 424
 Predicted class is RHC
 ----------------------------
Node 8: Terminal node
 Class      Number   Posterior
 NoRHC         199  0.3038E+00
 RHC           456  0.6962E+00
 Number of training cases misclassified = 199
 Predicted class is RHC
 ----------------------------
Node 9: Intermediate node
 A case goes into Node 18 if ninsclas = "No insurance", "Private",
 "Private & Medicare"
 ninsclas mode = "Private"
 Class      Number   Posterior
 NoRHC         225  0.4870E+00
 RHC           237  0.5130E+00
 Number of training cases misclassified = 225
 Predicted class is RHC
 ----------------------------
Node 18: Terminal node
 Class      Number   Posterior
```

```
 NoRHC          91  0.3730E+00
 RHC           153  0.6270E+00
 Number of training cases misclassified = 91
 Predicted class is RHC
 ----------------------------
Node 19: Terminal node
 Class      Number   Posterior
 NoRHC         134  0.6147E+00
 RHC            84  0.3853E+00
 Number of training cases misclassified = 84
 Predicted class is NoRHC
 ----------------------------
Node 5: Intermediate node
 A case goes into Node 10 if alb1 <= 2.9499511
 alb1 mean = 3.1795715
 Class      Number   Posterior
 NoRHC         350  0.6184E+00
 RHC           216  0.3816E+00
 Number of training cases misclassified = 216
 Predicted class is NoRHC
 ----------------------------
Node 10: Intermediate node
 A case goes into Node 20 if pafi1 <= 216.25000
 pafi1 mean = 239.66683
 Class      Number   Posterior
 NoRHC          76  0.4810E+00
 RHC            82  0.5190E+00
 Number of training cases misclassified = 76
 Predicted class is RHC
 ----------------------------
Node 20: Terminal node
 Class      Number   Posterior
 NoRHC          23  0.3194E+00
 RHC            49  0.6806E+00
 Number of training cases misclassified = 23
 Predicted class is RHC
 ----------------------------
Node 21: Terminal node
 Class      Number   Posterior
 NoRHC          53  0.6163E+00
 RHC            33  0.3837E+00
 Number of training cases misclassified = 33
 Predicted class is NoRHC
 ----------------------------
Node 11: Terminal node
 Class      Number   Posterior
 NoRHC         274  0.6716E+00
 RHC           134  0.3284E+00
 Number of training cases misclassified = 134
 Predicted class is NoRHC
 ----------------------------
```

Node 3: Intermediate node
 A case goes into Node 6 if pafi1 <= 142.35938
 pafi1 mean = 211.08630
 Class Number Posterior
 NoRHC 2777 0.6853E+00
 RHC 1275 0.3147E+00
 Number of training cases misclassified = 1275
 Predicted class is NoRHC

Node 6: Intermediate node
 A case goes into Node 12 if resp = "No"
 resp mode = "Yes"
 Class Number Posterior
 NoRHC 667 0.5163E+00
 RHC 625 0.4837E+00
 Number of training cases misclassified = 625
 Predicted class is NoRHC

Node 12: Intermediate node
 A case goes into Node 24 if dnr1 = "No"
 dnr1 mode = "No"
 Class Number Posterior
 NoRHC 244 0.4200E+00
 RHC 337 0.5800E+00
 Number of training cases misclassified = 244
 Predicted class is RHC

Node 24: Terminal node
 Class Number Posterior
 NoRHC 201 0.3903E+00
 RHC 314 0.6097E+00
 Number of training cases misclassified = 201
 Predicted class is RHC

Node 25: Terminal node
 Class Number Posterior
 NoRHC 43 0.6515E+00
 RHC 23 0.3485E+00
 Number of training cases misclassified = 23
 Predicted class is NoRHC

Node 13: Intermediate node
 A case goes into Node 26 if seps = "Yes"
 seps mode = "No"
 Class Number Posterior
 NoRHC 423 0.5949E+00
 RHC 288 0.4051E+00
 Number of training cases misclassified = 288
 Predicted class is NoRHC

Node 26: Terminal node

```
 Class      Number   Posterior
 NoRHC          40  0.3636E+00
 RHC            70  0.6364E+00
 Number of training cases misclassified = 40
 Predicted class is RHC
 ----------------------------
Node 27: Terminal node
 Class      Number   Posterior
 NoRHC         383  0.6373E+00
 RHC           218  0.3627E+00
 Number of training cases misclassified = 218
 Predicted class is NoRHC
 ----------------------------
Node 7: Terminal node
 Class      Number   Posterior
 NoRHC        2110  0.7645E+00
 RHC           650  0.2355E+00
 Number of training cases misclassified = 650
 Predicted class is NoRHC
 ----------------------------
 ****************************************************
 Variables used for splitting:
 alb1
 cat1
 dnr1
 meanbp1
 ninsclas
 pafi1
 resp
 seps

 Number of terminal nodes: 11

 Scaled importance scores of predictor variables
     Score     Rank  Variable
  2.287E+01     1.00  cat1
  2.234E+01     2.00  aps1
  2.012E+01     3.00  crea1
  1.924E+01     4.00  pafi1
  1.855E+01     5.00  meanbp1
  1.219E+01     6.00  alb1
  1.196E+01     7.00  neuro
  1.066E+01     8.00  card
  1.021E+01     9.00  hema1
  1.010E+01    10.00  cat2
  9.315E+00    11.00  wtkilo1
  7.929E+00    12.00  seps
  6.409E+00    13.00  resp
  6.073E+00    14.00  dnr1
  6.029E+00    15.00  bili1
  5.410E+00    16.00  paco21
```

```
 4.260E+00    17.00  hrt1
 3.979E+00    18.00  transhx
 3.860E+00    19.00  chrpulhx
 3.860E+00    20.00  resp1
 3.311E+00    21.00  ninsclas
 3.183E+00    22.00  dementhx
 2.969E+00    23.00  ph1
 2.163E+00    24.00  psychhx
 2.029E+00    25.00  renal
 1.913E+00    26.00  gastr
 1.896E+00    27.00  income
 1.748E+00    28.00  cardiohx
 1.412E+00    29.00  urin1
 1.386E+00    30.00  trauma
 1.247E+00    31.00  age
 1.194E+00    32.00  sex
--------- variables above this line are highly important -----
 1.188E+00    33.00  edu
 1.176E+00    34.00  sod1
 1.056E+00    35.00  immunhx
--------- variables below this line are unimportant ----------
 9.441E-01    36.00  malighx
 9.116E-01    37.00  wblc1
 8.732E-01    38.00  ca
 8.616E-01    39.00  amihx
 8.071E-01    40.00  scoma1
 6.766E-01    41.00  chfhx
 5.859E-01    42.00  gibledhx
 4.112E-01    43.00  renalhx
 4.095E-01    44.00  pot1
 3.971E-01    45.00  ortho
 3.431E-01    46.00  liverhx
 3.412E-01    47.00  hema
 3.280E-01    48.00  meta
 2.586E-01    49.00  temp1
 1.296E-01    50.00  race

Variables with scores above 1.19 are highly important
Variables with scores between 1.0 and 1.19 are likely important
Variables with scores below 1.0 are unimportant
No. highly important, likely important, and unimportant split variables: 32, 3, 15
LaTeX code for tree is in imp.tex
Importance scores are stored in imp.scr
Elapsed time in seconds: 183.51
```

Input file for GUIDE Forest (forestin.txt)

```
GUIDE       (do not edit this file unless you know what you are doing)
 36.2       (version of GUIDE that generated this file)
 1          (1=model fitting, 2=importance or DIF scoring, 3=data conversion)
```

"forestout.txt" (name of output file)
 2   (1=one tree, 2=ensemble)
 2   (1=bagging, 2=rforest)
 2   (1=random splits of missing values, 2=nonrandom)
 1   (1=classification, 2=regression)
 2   (1=interaction tests, 2=skip them)
"desc.txt" (name of data description file)
 1   (1=accept default number of trees, 2=change)
 1   (1=accept default number of variables for splitting, 2=change it)
 1   (1=estimated priors, 2=equal priors, 3=other priors)
 1   (1=unit misclassification costs, 2=other)
 1   (1=split point from quantiles, 2=use exhaustive search)
 1   (1=accept default splitting fraction, 2=change it)
 1   (1=default max. number of split levels, 2=specify no. in next line)
 1   (1=default min. node size, 2=specify min. value in next line)
"forestpred.txt" (file name for predicted class and probability estimates)
 1   (rank of top variable to split root node)

Output file for GUIDE Forest (forestout.txt)

```
 GGG   U   U  I  DDDD   EEEE
G   G U   U  I  D   D  E
G     U   U  I  D   D  E
G  GG U   U  I  D   D  EEE
G   G U   U  I  D   D  E
G   G U   U  I  D   D  E
 GGG    UUU   I  DDDD   EEEE
```

GUIDE Classification and Regression Trees and Forests
Version 36.2 (Build date: January 8, 2021)
Compiled with GFortran 10.2.0 on macOS Big Sur 11.1

This software is based upon work supported by the U.S. Army Research Office,
the National Science Foundation and the National Institutes of Health.

This job was started on 02/24/21 at 03:27

Random forest of classification trees
No pruning
Data description file: desc.txt
Training sample file: rhcdata.txt
Missing value code: NA
Records in data file start on line 2
D variable is swang1
Number of records in data file: 5735
Length of longest entry in data file: 19
Missing values found among categorical variables
Separate categories will be created for missing categorical variables
Number of classes: 2

Training sample class proportions of D variable swang1:

Class	#Cases	Proportion
NoRHC	3551	0.61918047
RHC	2184	0.38081953

Summary information for training sample of size 5735
d=dependent, b=split and fit cat variable using indicator variables,
c=split-only categorical, i=fit-only categorical (via indicators),
s=split-only numerical, n=split and fit numerical, f=fit-only numerical,
m=missing-value flag variable, p=periodic variable, w=weight

Column	Name		Minimum	Maximum	#Codes/ Levels/ Periods	#Missing
2	cat1	c		9		
3	cat2	c		7		
4	ca	c		3		
5	sadmdte	c		1401		
6	dschdte	c		1487		
7	dthdte	c		1593		
8	lstctdte	c		932		
9	death	c		2		
10	cardiohx	c		2		
11	chfhx	c		2		
12	dementhx	c		2		
13	psychhx	c		2		
14	chrpulhx	c		2		
15	renalhx	c		2		
16	liverhx	c		2		
17	gibledhx	c		2		
18	malighx	c		2		
19	immunhx	c		2		
20	transhx	c		2		
21	amihx	c		2		
22	age	c		5036		
23	sex	c		2		
24	edu	c		42		
25	surv2md1	c		1522		
26	das2d3pc	c		1023		
27	t3d30	c		29		
28	dth30	c		2		
29	aps1	c		123		
30	scoma1	c		11		
31	meanbp1	c		178		
32	wblc1	c		520		
33	hrt1	c		189		
34	resp1	c		72		
35	temp1	c		118		
36	pafi1	c		1342		
37	alb1	c		57		
38	hema1	c		450		
39	bili1	c		266		

40	crea1	c	148
41	sod1	c	73
42	pot1	c	81
43	paco21	c	266
44	ph1	c	96
45	swang1	d	2
46	wtkilo1	c	922
47	dnr1	c	2
48	ninsclas	c	6
49	resp	c	2
50	card	c	2
51	neuro	c	2
52	gastr	c	2
53	renal	c	2
54	meta	c	2
55	hema	c	2
56	seps	c	2
57	trauma	c	2
58	ortho	c	2
59	adld3p	c	9
60	urin1	c	1441
61	race	c	3
62	income	c	4

Total #cases	#cases w/ miss. D	#missing ord. vals	#X-var	#N-var	#F-var	#S-var
5735	0	0	3	0	0	0

#P-var	#M-var	#B-var	#C-var	#I-var
0	0	0	60	0

Number of cases used for training: 5735
Number of split variables: 60

Number of trees in ensemble: 500
Number of variables used for splitting: 21
Simple node models
Estimated priors
Unit misclassification costs
Univariate split highest priority
No interaction splits
Fraction of cases used for splitting each node: .0174
Maximum number of split levels: 20
Minimum node sample size: 28
Mean number of terminal nodes: 44.79

Classification matrix for training sample:

Predicted class	True class NoRHC	RHC
NoRHC	3551	0
RHC	0	2184
Total	3551	2184

Number of cases used for tree construction: 5735
Number misclassified: 0
Resubstitution estimate of mean misclassification cost: .0000

Number of OOB cases: 5735
Number OOB misclassified: 2194
OOB estimate of mean misclassification cost: .3826
Mean number of trees per OOB observation: 183.91

Predicted class probabilities are stored in forestpred.txt
Elapsed time in seconds: 84.133