101C Project

Alison Wilbur

2023-12-05

Setting up the data

```
# Reading in project data
SA.Train <- read.csv("TrainSAData2.csv")
SA.Test <- read.csv("TestSAData2NoY.csv")

# Seeing NAs
sum(is.na(SA.Train))
```

```
## [1] 130776
```

```
sum(is.na(SA.Test))
```

```
## [1] 56035
```

```
# Remove ID column
SA.Train <- SA.Train[,-1]
SA.Test <- SA.Test[,-1]

# Character to factor
SA.Train <- SA.Train %>% mutate_if(is.character, as.factor)
SA.Test <- SA.Test %>% mutate_if(is.character, as.factor)

head(SA.Train)
```

```
##      sex age height weight waistline sight_left sight_right hear_left
## 1   Male  75    160     NA        NA         NA         0.7    Normal
## 2 Female  50    160     60      74.0        1.0         1.2    Normal
## 3   Male  65    170     80      95.0        1.0         1.5    Normal
## 4   <NA>  65    155     55      81.0        0.3         0.4  Abnormal
## 5   Male  35    160     60      85.0        1.0         1.0    Normal
## 6 Female  50    160     70      73.2        0.3         0.4    Normal
##   hear_right SBP DBP BLDS tot_chole HDL_chole LDL_chole triglyceride hemoglobin
## 1     Normal  NA  76  136       215        33       143          193       15.0
## 2     Normal 118  70  125       207        85        NA          110       13.3
## 3     Normal 149  83  130       115        48        33          170       16.4
## 4   Abnormal 118  67   97       171        65        67          195       13.9
## 5     Normal  96  62   78       114        42        58           72       16.0
## 6     Normal 119  79  220       178        61        80          181       10.5
```

```
##   urine_protein serum_creatinine SGOT_AST SGOT_ALT gamma_GTP      BMI
## 1             3              0.9       28       23        36 23.43750
## 2             1              0.6       28       19        22 23.43750
## 3             1              1.4       41       64        53       NA
## 4             1              0.8       26       25        NA 22.89282
## 5             1              1.0       17       24        34       NA
## 6             1              0.5       36       NA        20 27.34375
##   BMI.Category AGE.Category Smoking.Status Alcoholic.Status
## 1      Healthy     Very Old  Still Smoking                Y
## 2         <NA>     Mid-aged   Never Smoked                Y
## 3   Overweight          Old  Still Smoking                Y
## 4         <NA>          Old   Never Smoked                N
## 5      Healthy     Mid-aged  Still Smoking                N
## 6   Overweight     Mid-aged   Never Smoked                N
```

```
head(SA.Test)
```

```
##      sex age height weight waistline sight_left sight_right hear_left
## 1   <NA>  40    175     NA        76        1.5         1.2    Normal
## 2 Female  55    150     55        81        1.0         0.9    Normal
## 3 Female  35    155     50        73        0.2         0.2    Normal
## 4 Female  60    155     50        79        1.0         1.0    Normal
## 5   Male  55    165     65        84         NA         0.9    Normal
## 6   Male  45    170     55        73        1.5         1.2    Normal
##   hear_right SBP DBP BLDS tot_chole HDL_chole LDL_chole triglyceride hemoglobin
## 1     Normal 118  78   89       160        49        75          181         NA
## 2     Normal  89  52  109       240        67       154           95       12.6
## 3     Normal 102  63   86        NA        48       120           63       12.0
## 4     Normal  NA  76   97       222        61       140          101       12.9
## 5     Normal 102  63   NA       198        46       112          200       17.1
## 6     Normal 120  80   98       152        NA        55          283       14.5
##   urine_protein serum_creatinine SGOT_AST SGOT_ALT gamma_GTP      BMI
## 1             1              1.1       18       13        15 22.85714
## 2             1              0.7       47       32        27 24.44444
## 3             1              0.8       14       10        10 20.81165
## 4             1              1.0       33       NA        64 20.81165
## 5             2              0.7       21       33        78 23.87511
## 6             1              1.0       17       25        26       NA
##   BMI.Category AGE.Category Smoking.Status
## 1      Healthy     Mid-aged  Still Smoking
## 2         <NA>          Old           <NA>
## 3      Healthy     Mid-aged           <NA>
## 4      Healthy          Old   Never Smoked
## 5      Healthy          Old   Never Smoked
## 6      Healthy     Mid-aged  Still Smoking
```

Best Models: Random Forest and GBM

Trying Random Forest, imputing NAs with randomForest package (column medians for numerical predictors and modes for categorical predictors)

```
set.seed(127)

SA.rf.impute.tr <- na.roughfix(SA.Train)
SA.rf.impute.ts <- na.roughfix(SA.Test)
#
SA.RF.orig <- randomForest(Alcoholic.Status~., data=SA.rf.impute.tr, importance = TRUE, ntree = 1000)
SA.RF.orig
```

```
##
## Call:
##  randomForest(formula = Alcoholic.Status ~ ., data = SA.rf.impute.tr,      importance = TRUE, ntree =
##                Type of random forest: classification
##                      Number of trees: 1000
## No. of variables tried at each split: 5
##
##         OOB estimate of  error rate: 27.4%
## Confusion matrix:
##       N     Y class.error
## N 24998 10115   0.2880699
## Y  9063 25824   0.2597816
```

OOB 27.4%
Kaggle: 72.953

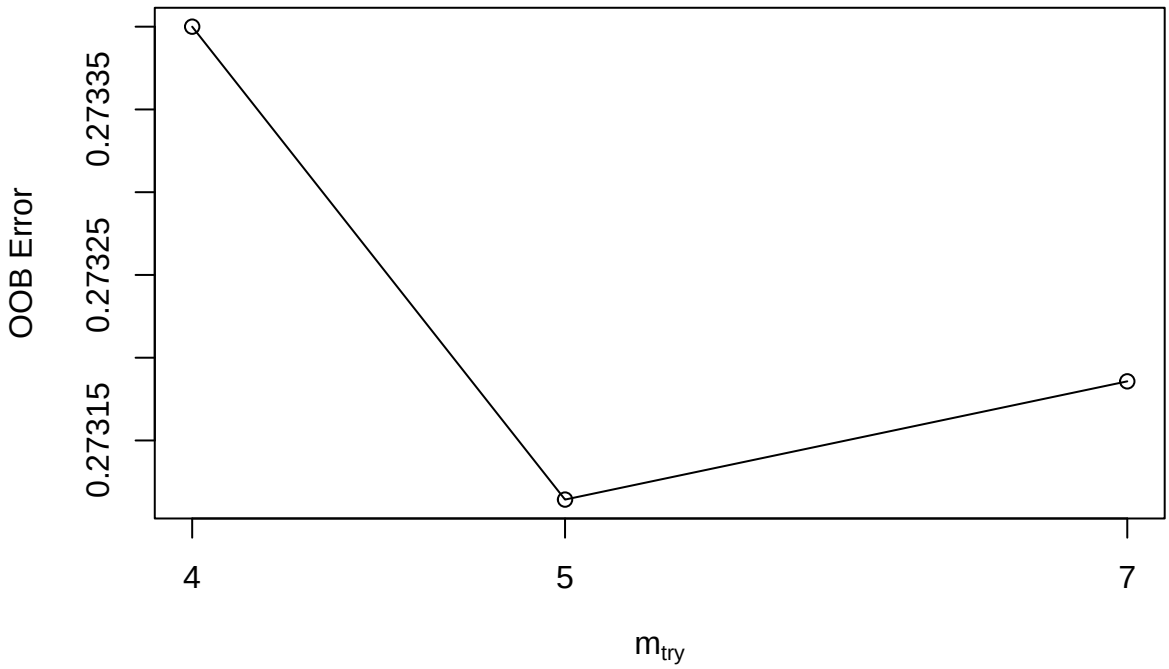
```
predRF <- predict(SA.RF.orig, newdata = SA.rf.impute.ts, type="class")
solution <- data.frame("ID" = c(1:30000), Alcoholic.Status=predRF)
write.csv(solution, row.names = FALSE, 'RFnewsolution2.csv')
```

Adjusting the Random Forest

```
y <- SA.rf.impute.tr[,27]
x <- SA.rf.impute.tr[,1:26]

bestmtry <- tuneRF(x, y, ntree=1000, stepFactor = 1.5, improve = 0.01, trace=TRUE)
```

```
## mtry = 5  OOB error = 27.31%
## Searching left ...
## mtry = 4     OOB error = 27.34%
## -0.001046135 0.01
## Searching right ...
## mtry = 7     OOB error = 27.32%
## -0.0002615336 0.01
```




```
bestmtry
```

```
##       mtry  OOBError
## 4.OOB    4 0.2734000
## 5.OOB    5 0.2731143
## 7.OOB    7 0.2731857
```

```
SA.RF.orig2 <- randomForest(Alcoholic.Status~., data=SA.rf.impute.tr, mtry=4, ntree=2000)
SA.RF.orig2
```

```
##
## Call:
##  randomForest(formula = Alcoholic.Status ~ ., data = SA.rf.impute.tr,      mtry = 4, ntree = 2000)
##                Type of random forest: classification
##                      Number of trees: 2000
## No. of variables tried at each split: 4
##
##         OOB estimate of  error rate: 27.26%
## Confusion matrix:
##       N     Y class.error
## N 24958 10155   0.2892091
## Y  8927 25960   0.2558833
```

```
predRF <- predict(SA.RF.orig2, newdata = SA.rf.impute.ts, type="class")
solution <- data.frame("ID" = c(1:30000), Alcoholic.Status=predRF)
write.csv(solution, row.names = FALSE, 'RFnewsolution3.csv')
```

OOB 27.26%
Kaggle: 72.816

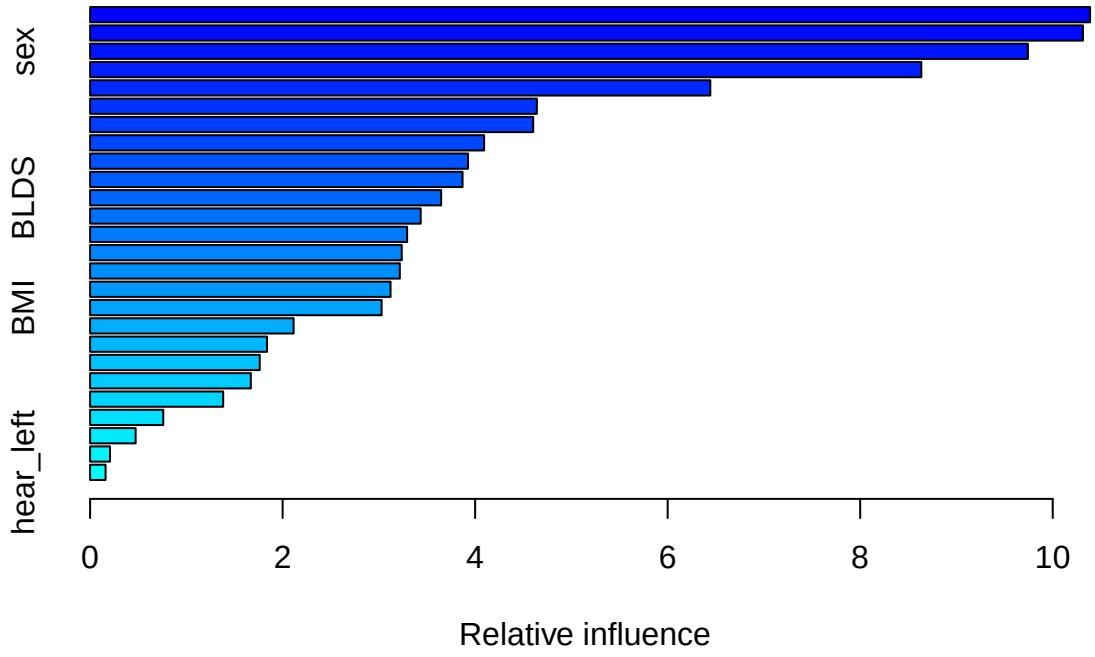
Highest Accuracy comes from Random Forest model with mtry = 5, ntrees = 1000.

Trying GBM with original data (keeping NA values)

```
# New response column
SA.Train$Alcoholic.Status.Y <- ifelse(SA.Train$Alcoholic.Status == "Y", 1, 0)

set.seed(127)

SA.GBM.orig <- gbm(
  Alcoholic.Status.Y~.,
  data = SA.Train[,-27],
  distribution = "bernoulli",
  n.trees = 2500,
  interaction.depth = 6
  )
summary(SA.GBM.orig)
```




```
##                              var    rel.inf
## Smoking.Status     Smoking.Status 10.3882319
## age                         age 10.3143740
## sex                         sex  9.7421293
## gamma_GTP           gamma_GTP  8.6350295
## HDL_chole           HDL_chole  6.4431154
## SGOT_ALT             SGOT_ALT  4.6412304
## triglyceride     triglyceride  4.6029286
## hemoglobin         hemoglobin  4.0929161
## tot_chole           tot_chole  3.9253121
## LDL_chole           LDL_chole  3.8690525
## BLDS                     BLDS  3.6466532
## waistline           waistline  3.4345266
## SBP                       SBP  3.2931240
## SGOT_AST             SGOT_AST  3.2375485
```

```
## height                      height  3.2176477
## DBP                            DBP  3.1211379
## BMI                            BMI  3.0286727
## AGE.Category          AGE.Category  2.1142257
## sight_left              sight_left  1.8375809
## serum_creatinine serum_creatinine  1.7623299
## sight_right            sight_right  1.6697102
## weight                      weight  1.3824720
## BMI.Category          BMI.Category  0.7598201
## urine_protein        urine_protein  0.4729115
## hear_right              hear_right  0.2067034
## hear_left                hear_left  0.1606155
```

```
# Training data
pred.GB1 <- predict(SA.GBM.orig,data=SA.Train[,-27],n.trees = 2500, type="response")
P.bo <- ifelse(pred.GB1<0.5,0,1)

table(SA.Train$Alcoholic.Status.Y,P.bo)
```

```
##    P.bo
##         0     1
##   0 28551  6562
##   1  6297 28590
```

```
mean(SA.Train$Alcoholic.Status.Y!=P.bo)
```

```
## [1] 0.1837
```

```
# Testing data
pred.bo.t <- predict(SA.GBM.orig,newdata=SA.Test,n.trees = 2500, type="response")
P.bo.t <- ifelse(pred.bo.t<0.5,0,1)
P.bo.t <- ifelse(P.bo.t == 0, "N", "Y")

solution <- data.frame("ID" = c(1:30000), Alcoholic.Status=P.bo.t)
write.csv(solution, row.names = FALSE, 'newGBMsolution.csv')
```

Training data accuracy: 81.65 Kaggle test data accuracy: 72.27

Other Methods Tried

Imputing NA values using column means, Hmisc package

```
# Replace NAs with mean of each col
for (i in 1:27){
  SA.Train[,i] <- impute(SA.Train[,i], mean)
}
for(i in 1:26){
```

```
  SA.Test[,i] <- impute(SA.Test[,i], mean)
}

head(SA.Train)
```

```
##      sex age height   weight waistline sight_left sight_right hear_left
## 1   Male  75    160 63.23468  81.26761  0.9807226         0.7    Normal
## 2 Female  50    160 60.00000  74.00000  1.0000000         1.2    Normal
## 3   Male  65    170 80.00000  95.00000  1.0000000         1.5    Normal
## 4   Male  65    155 55.00000  81.00000  0.3000000         0.4  Abnormal
## 5   Male  35    160 60.00000  85.00000  1.0000000         1.0    Normal
## 6 Female  50    160 70.00000  73.20000  0.3000000         0.4    Normal
##   hear_right      SBP DBP BLDS tot_chole HDL_chole LDL_chole triglyceride
## 1     Normal 122.4681  76  136       215        33  143.0000          193
## 2     Normal 118.0000  70  125       207        85  113.2694          110
## 3     Normal 149.0000  83  130       115        48   33.0000          170
## 4   Abnormal 118.0000  67   97       171        65   67.0000          195
## 5     Normal  96.0000  62   78       114        42   58.0000           72
## 6     Normal 119.0000  79  220       178        61   80.0000          181
##   hemoglobin urine_protein serum_creatinine SGOT_AST SGOT_ALT gamma_GTP
## 1       15.0             3              0.9       28 23.00000  36.00000
## 2       13.3             1              0.6       28 19.00000  22.00000
## 3       16.4             1              1.4       41 64.00000  53.00000
## 4       13.9             1              0.8       26 25.00000  36.75733
## 5       16.0             1              1.0       17 24.00000  34.00000
## 6       10.5             1              0.5       36 25.67331  20.00000
##        BMI BMI.Category AGE.Category Smoking.Status Alcoholic.Status
## 1 23.43750      Healthy     Very Old  Still Smoking                Y
## 2 23.43750      Healthy     Mid-aged   Never Smoked                Y
## 3 23.91194   Overweight          Old  Still Smoking                Y
## 4 22.89282      Healthy          Old   Never Smoked                N
## 5 23.91194      Healthy     Mid-aged  Still Smoking                N
## 6 27.34375   Overweight     Mid-aged   Never Smoked                N
##   Alcoholic.Status.Y
## 1                  1
## 2                  1
## 3                  1
## 4                  0
## 5                  0
## 6                  0
```

Creating Dummy Variables

```
SA.Train$sex.Male <- ifelse(SA.Train$sex == "Male", 1, 0)
SA.Train$hear_left.Normal <- ifelse(SA.Train$hear_left == "Normal", 1, 0)
SA.Train$hear_right.Normal <- ifelse(SA.Train$hear_right == "Normal", 1, 0)
SA.Train$BMI.Healthy <- ifelse(SA.Train$BMI.Category == "Healthy", 1, 0)
SA.Train$AGE.YoungToMid <- ifelse(SA.Train$AGE.Category == "Young" | SA.Train$AGE.Category == "Mid-aged
SA.Train$NeverSmoked <- ifelse(SA.Train$Smoking.Status =="Never Smoked", 1, 0)

SA.Test$sex.Male <- ifelse(SA.Test$sex == "Male", 1, 0)
```

```
SA.Test$hear_left.Normal <- ifelse(SA.Test$hear_left == "Normal", 1, 0)
SA.Test$hear_right.Normal <- ifelse(SA.Test$hear_right == "Normal", 1, 0)
SA.Test$BMI.Healthy <- ifelse(SA.Test$BMI.Category == "Healthy", 1, 0)
SA.Test$AGE.YoungToMid <- ifelse(SA.Test$AGE.Category == "Young" | SA.Test$AGE.Category == "Mid-aged", 
SA.Test$NeverSmoked <- ifelse(SA.Test$Smoking.Status =="Never Smoked", 1, 0)

SA.Train <- SA.Train[,-c(1, 8, 9, 24, 25, 26, 27)]
SA.Test <- SA.Test[,-c(1, 8, 9, 24, 25, 26)]

head(SA.Train)
```

```
##   age height   weight waistline sight_left sight_right      SBP DBP BLDS
## 1  75    160 63.23468  81.26761  0.9807226         0.7 122.4681  76  136
## 2  50    160 60.00000  74.00000  1.0000000         1.2 118.0000  70  125
## 3  65    170 80.00000  95.00000  1.0000000         1.5 149.0000  83  130
## 4  65    155 55.00000  81.00000  0.3000000         0.4 118.0000  67   97
## 5  35    160 60.00000  85.00000  1.0000000         1.0  96.0000  62   78
## 6  50    160 70.00000  73.20000  0.3000000         0.4 119.0000  79  220
##   tot_chole HDL_chole LDL_chole triglyceride hemoglobin urine_protein
## 1       215        33  143.0000          193       15.0             3
## 2       207        85  113.2694          110       13.3             1
## 3       115        48   33.0000          170       16.4             1
## 4       171        65   67.0000          195       13.9             1
## 5       114        42   58.0000           72       16.0             1
## 6       178        61   80.0000          181       10.5             1
##   serum_creatinine SGOT_AST SGOT_ALT gamma_GTP      BMI Alcoholic.Status.Y
## 1              0.9       28 23.00000  36.00000 23.43750                  1
## 2              0.6       28 19.00000  22.00000 23.43750                  1
## 3              1.4       41 64.00000  53.00000 23.91194                  1
## 4              0.8       26 25.00000  36.75733 22.89282                  0
## 5              1.0       17 24.00000  34.00000 23.91194                  0
## 6              0.5       36 25.67331  20.00000 27.34375                  0
##   sex.Male hear_left.Normal hear_right.Normal BMI.Healthy AGE.YoungToMid
## 1        1                1                 1           1              0
## 2        0                1                 1           1              1
## 3        1                1                 1           0              0
## 4        1                0                 0           1              0
## 5        1                1                 1           1              1
## 6        0                1                 1           0              1
##   NeverSmoked
## 1           0
## 2           1
## 3           0
## 4           1
## 5           0
## 6           1
```

Numeric Variable Selection: PCA

```
# Principle components of the data
SA.comp <- princomp(SA.Train, scale=TRUE)
```

```
## Warning: In princomp.default(SA.Train, scale = TRUE) :
##  extra argument 'scale' will be disregarded
```

```
summary(SA.comp)
```

```
## Importance of components:
##                             Comp.1     Comp.2     Comp.3      Comp.4      Comp.5
## Standard deviation     99.2289921 48.0553810 46.5673150 27.20216420 23.67216388
## Proportion of Variance  0.5814477  0.1363696  0.1280548  0.04369587  0.03309096
## Cumulative Proportion   0.5814477  0.7178173  0.8458721  0.88956801  0.92265897
##                             Comp.6      Comp.7      Comp.8       Comp.9
## Standard deviation     18.0277314 15.84319021 15.32145787 12.257761962
## Proportion of Variance  0.0191918  0.01482241  0.01386225  0.008872702
## Cumulative Proportion   0.9418508  0.95667318  0.97053543  0.979408131
##                            Comp.10    Comp.11     Comp.12     Comp.13
## Standard deviation     11.670589661 8.65594226 7.941536642 5.994596454
## Proportion of Variance  0.008043019 0.00442448 0.003724282 0.002122039
## Cumulative Proportion   0.987451151 0.99187563 0.995599913 0.997721952
##                            Comp.14     Comp.15      Comp.16      Comp.17
## Standard deviation     5.741967161 1.450101980 1.260183e+00 6.368807e-01
## Proportion of Variance 0.001946951 0.000124174 9.377791e-05 2.395244e-05
## Cumulative Proportion  0.999668903 0.999793077 9.998869e-01 9.999108e-01
##                             Comp.18      Comp.19      Comp.20      Comp.21
## Standard deviation     5.344055e-01 4.972797e-01 4.796607e-01 4.157800e-01
## Proportion of Variance 1.686457e-05 1.460276e-05 1.358632e-05 1.020847e-05
## Cumulative Proportion  9.999277e-01 9.999423e-01 9.999559e-01 9.999661e-01
##                             Comp.22      Comp.23      Comp.24      Comp.25
## Standard deviation     3.985526e-01 3.851105e-01 3.597775e-01 2.885831e-01
## Proportion of Variance 9.380040e-06 8.757987e-06 7.643664e-06 4.917850e-06
## Cumulative Proportion  9.999754e-01 9.999842e-01 9.999919e-01 9.999968e-01
##                             Comp.26      Comp.27
## Standard deviation     2.015620e-01 1.187037e-01
## Proportion of Variance 2.399113e-06 8.320736e-07
## Cumulative Proportion  9.999992e-01 1.000000e+00
```

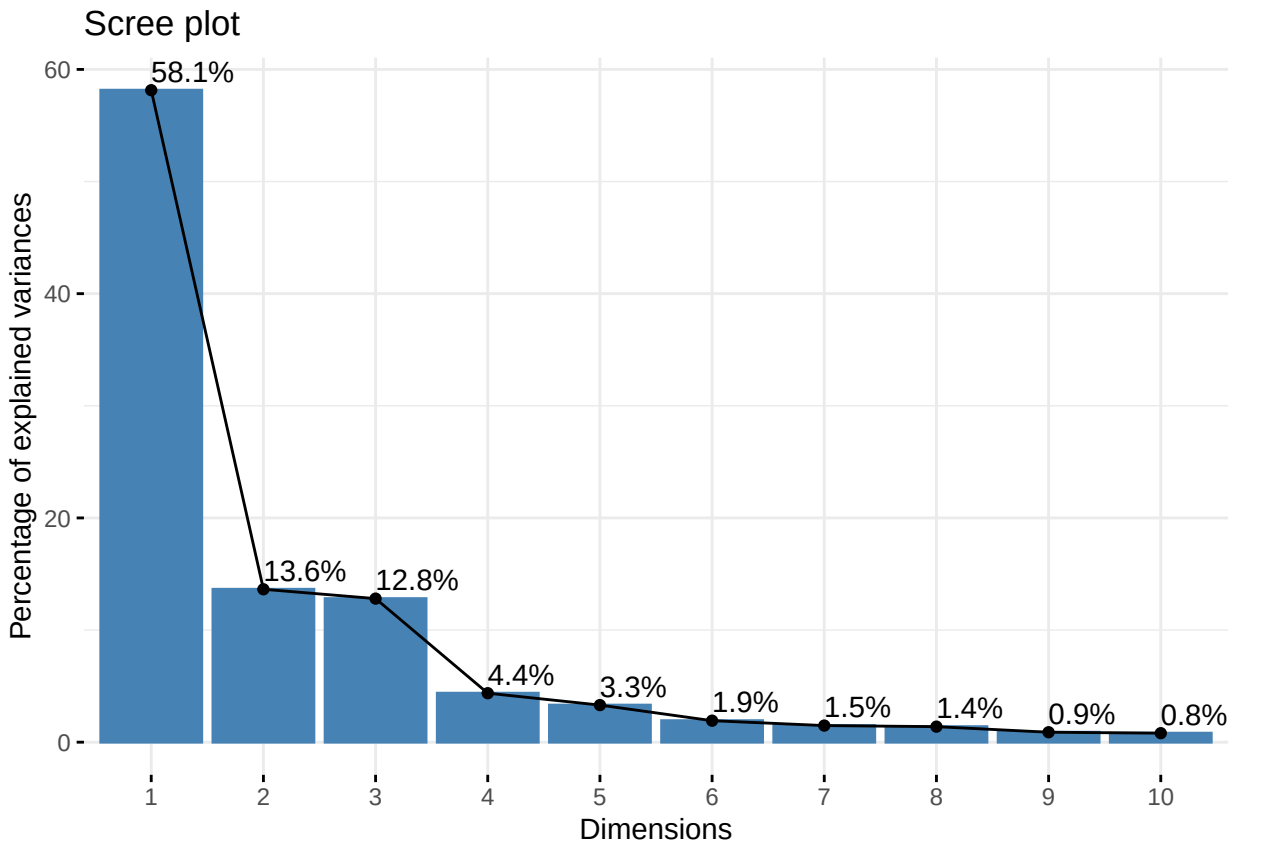
```
SA.comp$loadings[,1:4]
```

```
##                              Comp.1        Comp.2        Comp.3        Comp.4
## age                   6.158182e-03  2.266958e-03  3.002744e-03  7.087705e-03
## height                1.288611e-02 -1.135391e-02  2.293877e-02  1.970203e-02
## weight                3.524750e-02 -1.024007e-03  4.103603e-02  7.313678e-02
## waistline             2.914734e-02  5.744847e-05  3.212815e-02  5.867858e-02
## sight_left            8.851378e-05  5.603371e-05  7.855691e-05  1.070611e-04
## sight_right           4.383762e-05  2.546500e-05  1.053732e-04  1.371807e-04
## SBP                   2.699084e-02  2.244556e-03  3.626238e-02  3.759024e-02
## DBP                   1.929676e-02  8.980266e-03  2.612180e-02  2.245965e-02
## BLDS                  5.246649e-02 -3.509433e-02  5.833909e-02  4.857229e-02
## tot_chole             1.079953e-01  7.188102e-01  9.034542e-02 -4.385704e-02
## HDL_chole            -4.843089e-02  3.929699e-02  2.140790e-02 -7.157037e-02
## LDL_chole             1.661021e-02  6.765324e-01  8.074964e-02  2.605012e-02
## triglyceride          9.731269e-01 -6.282319e-02 -1.968142e-01 -1.134529e-02
## hemoglobin            3.789699e-03  1.379209e-03  5.449412e-03  5.364971e-03
## urine_protein         1.766008e-04 -1.448129e-04  3.384119e-04  2.619710e-04
```

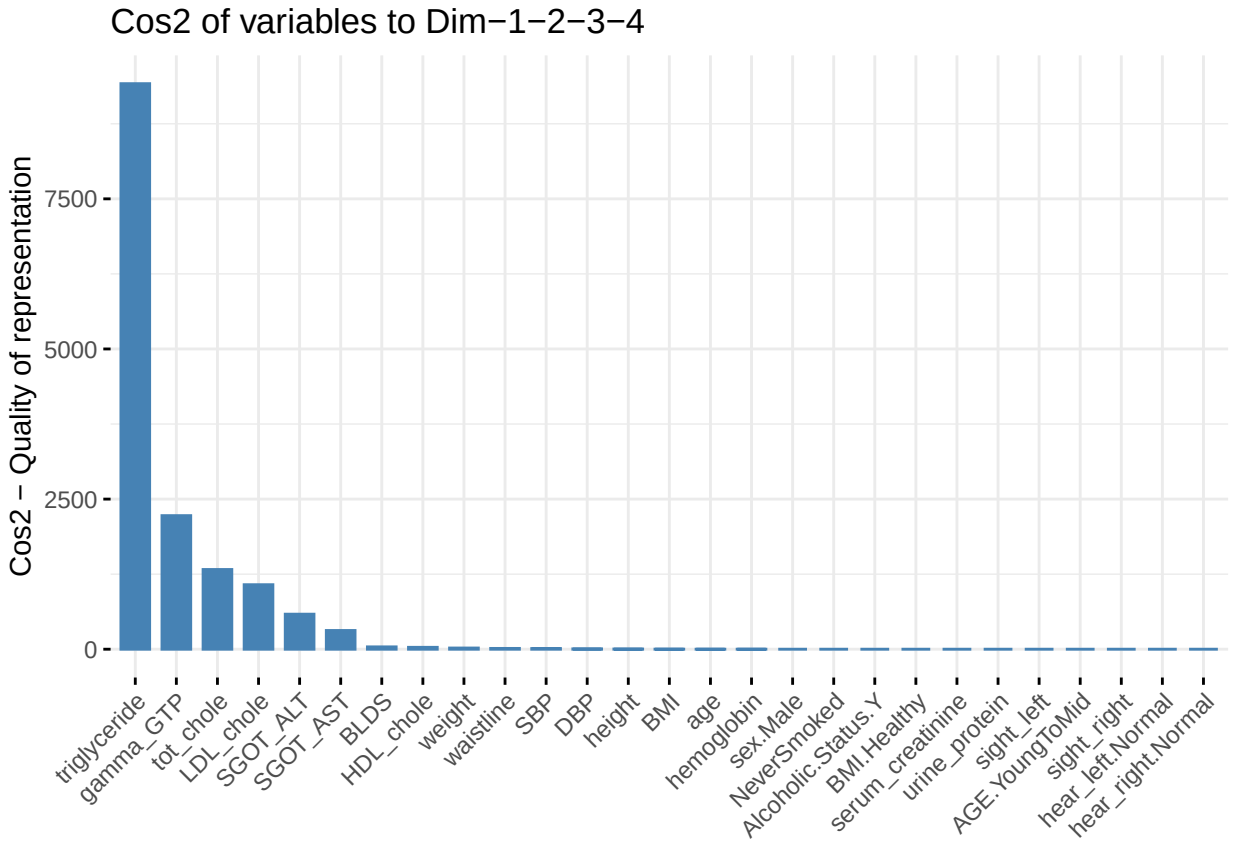
```
## serum_creatinine    2.791863e-04 -2.172934e-04  4.790365e-04  4.360275e-04
## SGOT_AST            3.006845e-02 -2.786679e-02  1.956375e-01  5.426932e-01
## SGOT_ALT            5.496297e-02 -1.385976e-02  2.409396e-01  7.600797e-01
## gamma_GTP           1.690093e-01 -1.330164e-01  9.170791e-01 -3.254853e-01
## BMI                 9.119033e-03  2.661211e-03  8.412317e-03  2.066638e-02
## Alcoholic.Status.Y  5.879820e-04 -5.253028e-04  1.734070e-03 -9.271271e-04
## sex.Male            1.093227e-03 -7.868124e-04  1.930052e-03  1.324263e-03
## hear_left.Normal    4.902147e-06  1.105544e-04  2.430522e-07  3.493657e-06
## hear_right.Normal  -3.590701e-06  8.398128e-05 -2.636062e-05 -3.588350e-05
## BMI.Healthy        -8.469240e-04  1.235014e-04 -8.673623e-04 -1.731089e-03
## AGE.YoungToMid     -5.610091e-05  5.660319e-05 -4.482975e-05 -2.424116e-04
## NeverSmoked        -9.854110e-04  6.476944e-04 -1.605028e-03 -2.135743e-04
```

After computing the PCA for the training data, the results show that that first 4 principal components are the most significant since they explain almost 90% of the total variance.

```
# Scree plot
fviz_eig(SA.comp, addlabels = TRUE)
```




```
# Viz to see how much each variable contributes to the 4 components
fviz_cos2(SA.comp, choice="var", axes=1:4)
```



The scree plot visualizes the importance of each principal component, and the cos2 visualization shows how much each variable contributes to the 4 most significant principal components.

From this visualization, triglyceride, gamma_GTP, tot_chole, LDL_chole, SGOT_ALT, and SGOT_AST appear to be the most significant predictors in the data set.

New Data Frame with Selected Predictors

```
# New SA.Train
SA.Train.sel <- data.frame(
  # Selected numeric predictors
  Triglyceride = SA.Train$triglyceride,
  Gamma.GTP = SA.Train$gamma_GTP,
  Tot.Chole = SA.Train$tot_chole,
  LDL.Chole = SA.Train$LDL_chole,
  SGOT.ALT = SA.Train$SGOT_ALT,
  SGOT.AST = SA.Train$SGOT_AST,
  Alcoholic.Status.Y= SA.Train$Alcoholic.Status.Y,
  # All categorical predictors
  sex.Male = SA.Train$sex.Male,
  hear_left.Normal =SA.Train$hear_left.Normal,
  hear_right.Normal =SA.Train$hear_right.Normal,
  BMI.Healthy = SA.Train$BMI.Healthy,
  AGE.YoungToMid = SA.Train$AGE.YoungToMid,
  NeverSmoked = SA.Train$NeverSmoked
)
```

```r
# New SA.Test
SA.Test.sel <- data.frame(
  # Selected numeric predictors
  Triglyceride = SA.Test$triglyceride,
  Gamma.GTP = SA.Test$gamma_GTP,
  Tot.Chole = SA.Test$tot_chole,
  LDL.Chole = SA.Test$LDL_chole,
  SGOT.ALT = SA.Test$SGOT_ALT,
  SGOT.AST = SA.Test$SGOT_AST,
  # All categorical predictors
  sex.Male = SA.Test$sex.Male,
  hear_left.Normal =SA.Test$hear_left.Normal,
  hear_right.Normal =SA.Test$hear_right.Normal,
  BMI.Healthy = SA.Test$BMI.Healthy,
  AGE.YoungToMid = SA.Test$AGE.YoungToMid,
  NeverSmoked = SA.Test$NeverSmoked
)
```

Categorical Variable Selection: Random Forest

```r
SA.Train.sel$Alcoholic.Status.Y <- as.factor(SA.Train.sel$Alcoholic.Status.Y)

set.seed(124)

SA.RF.VS <- randomForest(Alcoholic.Status.Y~., data = SA.Train.sel, mtry=3, importance = TRUE, ntree = 5
SA.RF.VS
```

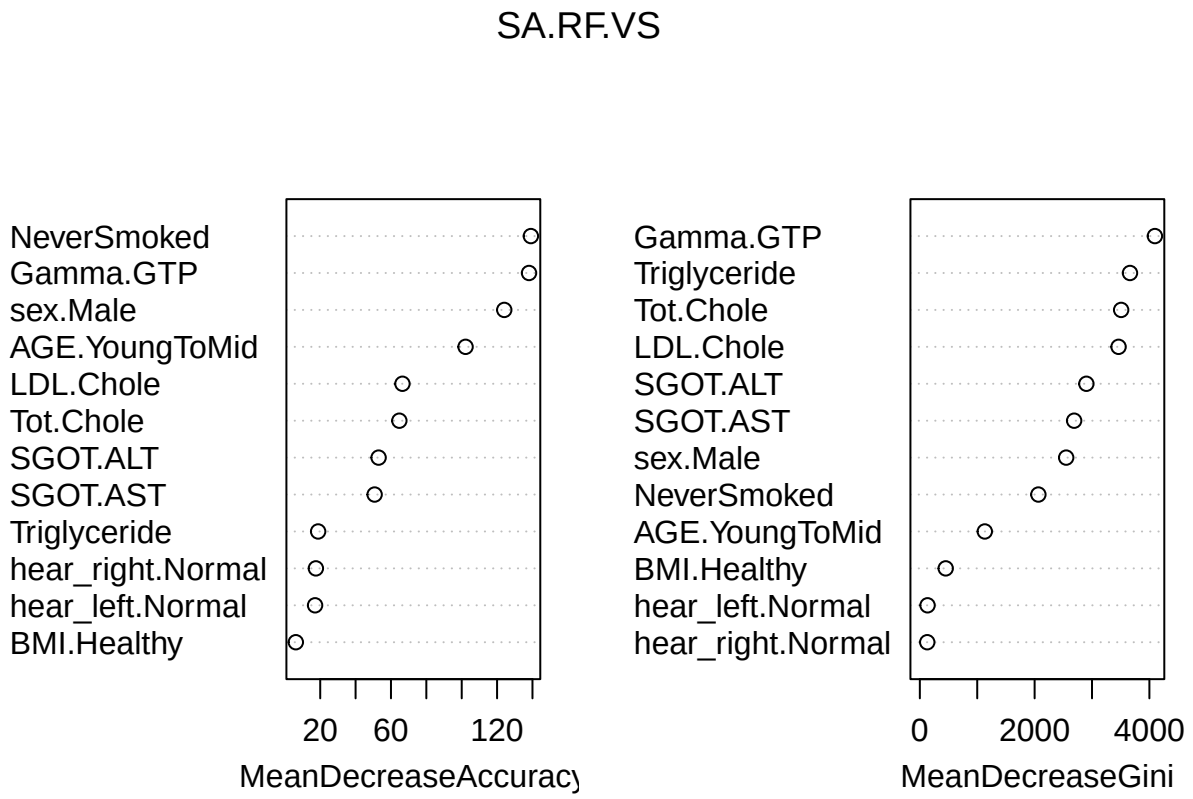
```
##
## Call:
##  randomForest(formula = Alcoholic.Status.Y ~ ., data = SA.Train.sel,      mtry = 3, importance = TRU
##                Type of random forest: classification
##                      Number of trees: 500
## No. of variables tried at each split: 3
##
##         OOB estimate of  error rate: 29.74%
## Confusion matrix:
##       0     1 class.error
## 0 24920 10193   0.2902913
## 1 10625 24262   0.3045547
```

```r
importance(SA.RF.VS)
```

```
##                             0          1 MeanDecreaseAccuracy MeanDecreaseGini
## Triglyceride      -25.647831  44.580824            18.834958        3661.6032
## Gamma.GTP          42.465780 105.174623           138.065662        4095.7381
## Tot.Chole          41.256462  19.146261            64.741118        3507.9083
## LDL.Chole          35.462686  30.464741            66.489999        3462.8814
## SGOT.ALT           18.816186  26.074867            53.015872        2901.8425
## SGOT.AST            8.589031  39.057027            50.760788        2688.6919
## sex.Male           63.661608  51.545494           124.057294        2549.8530
```

```
## hear_left.Normal   -7.287819  27.919754          17.093439        134.3764
## hear_right.Normal  -1.070163  23.311091          17.545909        129.2939
## BMI.Healthy        -1.360733   9.304299           6.209884        450.7875
## AGE.YoungToMid     41.270135 105.944464         102.175516       1131.9180
## NeverSmoked       111.084265  23.251231         139.107332       2065.4335
```

```
varImpPlot(SA.RF.VS)
```



Based on the variable importance plot from the Random Forest fit to the training data, the most significant categorical variables are NeverSmoked, sex.Male, and AGE.YoungToMid.

Updated Data Frame with Selected Predictors

```
SA.Train.sel <- SA.Train.sel[,-c(9,10,11)]
SA.Test.sel <- SA.Test.sel[,-c(8,9,10)]

head(SA.Train.sel)
```

```
##   Triglyceride Gamma.GTP Tot.Chole LDL.Chole SGOT.ALT SGOT.AST
## 1          193  36.00000       215  143.0000 23.00000       28
## 2          110  22.00000       207  113.2694 19.00000       28
## 3          170  53.00000       115   33.0000 64.00000       41
## 4          195  36.75733       171   67.0000 25.00000       26
## 5           72  34.00000       114   58.0000 24.00000       17
## 6          181  20.00000       178   80.0000 25.67331       36
##   Alcoholic.Status.Y sex.Male AGE.YoungToMid NeverSmoked
```

```
## 1                  1         1              0           0
## 2                  1         0              1           1
## 3                  1         1              0           0
## 4                  0         1              0           1
## 5                  0         1              1           0
## 6                  0         0              1           1
```

```
head(SA.Test.sel)
```

```
##   Triglyceride Gamma.GTP Tot.Chole LDL.Chole SGOT.ALT SGOT.AST sex.Male
## 1          181        15  160.0000        75 13.00000       18        1
## 2           95        27  240.0000       154 32.00000       47        0
## 3           63        10  195.5427       120 10.00000       14        0
## 4          101        64  222.0000       140 25.72149       33        0
## 5          200        78  198.0000       112 33.00000       21        1
## 6          283        26  152.0000        55 25.00000       17        1
##   AGE.YoungToMid NeverSmoked
## 1              1           0
## 2              0           1
## 3              1           1
## 4              0           1
## 5              0           1
## 6              1           0
```

Building Models using Selected Predictors

Logstic Regression

```
SA.LR <- glm(Alcoholic.Status.Y~., data = SA.Train.sel, family="binomial")
```

```
# Training data
pred.LR <- predict(SA.LR,data=SA.Train.sel, type="response")
P.bo <- ifelse(pred.LR<0.5,0,1)

# Training Confusion Matrix
table(SA.Train.sel$Alcoholic.Status.Y,P.bo)
```

```
##    P.bo
##         0     1
##   0 24513 10600
##   1 10693 24194
```

```
# Training Error Rate
mean(SA.Train.sel$Alcoholic.Status.Y!=P.bo)
```

```
## [1] 0.3041857
```

Kaggle score: 0.69413

Random Forest

```
SA.RF.sel <- randomForest(
  x = SA.Train.sel[,c("Triglyceride", "Gamma.GTP", "Tot.Chole", "LDL.Chole", "SGOT.ALT", "SGOT.AST", "s
  y = SA.Train.sel$Alcoholic.Status.Y,
  mtry = 2,
  ntree=500,
  stepFactor=1.5,
  improve=0.01,
)
SA.RF.sel
```

```
##
## Call:
##  randomForest(x = SA.Train.sel[, c("Triglyceride", "Gamma.GTP",      "Tot.Chole", "LDL.Chole", "SGOT
##                Type of random forest: classification
##                      Number of trees: 500
## No. of variables tried at each split: 2
##
##         OOB estimate of  error rate: 29.74%
## Confusion matrix:
##       0     1 class.error
## 0 24847 10266   0.2923703
## 1 10550 24337   0.3024049
```

Neural Net

```
SA.NN <- neuralnet(
  Alcoholic.Status.Y ~ .,
  data = SA.Train.sel,
  hidden = 5,
  linear.output=FALSE,
  lifesign = "full",
  rep = 2,
  algorithm = "rprop+",
  stepmax = 100000
)
```