**Project – 4 – Cancer Prediction**

The raw data file Cancer Data has been piped into CanDat to save the raw data. Later the CanDat has been detected for missing data, and outliers. There was one foreign column with the title ‘X’, which did not have any observations and had just a NAs in it. So we have removed it. Since the data description does not have any mention of the same.

Further there are nine attributes to describe a particular patient’s record or scan of the tumour, which are as follows. The mean, standard deviation, and the worst values were populated in the data base.

a) Radius (mean of distances from center to points on the perimeter)

b) Texture (standard deviation of gray-scale values)

c) Perimeter

d) Area

e) Smoothness (local variation in radius lengths)

f) Compactness (perimeter^2 / area - 1.0)

g) Concavity (severity of concave portions of the contour)

h) Concave points (number of concave portions of the contour)

i) Symmetry

j) Fractal dimension ("coastline approximation" - 1)

There were no missing values observed in the dataset after removing the ‘X’ column. Then it was detected for outliers. There were outliers detected and they were handled by plugging the first quartile and 3rd quartile values in place of the outliers which were below and above the 1.5 \* IQR of each of the 30 variables which are expected to classify the tumour into malignant or benign better. The treated file is named “outCanDat”

Then training data set and testing data set was prepared with 75% and 25% of the data for further estimations.

train\_control <- trainControl(method="cv", number=10)

rfmodel3 <- train(diagnosis~., data = CanDattrain, trContol = train\_control,method ="rf")

> rfmodel3

Random Forest

426 samples

30 predictor

2 classes: 'B', 'M'

No pre-processing

Resampling: Bootstrapped (25 reps)

Summary of sample sizes: 426, 426, 426, 426, 426, 426, ...

Resampling results across tuning parameters:

mtry Accuracy Kappa

2 0.9527297 0.8960422

16 0.9516645 0.8947632

30 0.9482746 0.8875461

Accuracy was used to select the optimal model using the largest value.

The final value used for the model was mtry = 2.

control <- trainControl(method = "repeatedcv", number = 10, repeats = 3)

seed <- 7

metric <- 'Accuracy'

colnames(CanDattrain)

mtry <- sqrt(ncol(CanDattrain[,-1]))

tunegrid <- expand.grid(.mtry = mtry)

rfmodel4 <- train(diagnosis~.,data = CanDattrain, method ="rf", metric = metric, tuneGrid = tunegrid, trControl = control)

> rfmodel3$finalModel

OOB estimate of error rate: 3.05%

Confusion matrix:

B M class.error

B 262 3 0.01132075

M 10 151 0.06211180

> table(CanDattrain$diagnosis)

B M

265 161

> table(rfmodel3pred, CanDattest$diagnosis)

rfmodel3pred B M

B 90 3

M 2 48

> varImp(rfmodel3)

rf variable importance

only 20 most important variables shown (out of 30)

Overall

concave.points\_worst 100.000

perimeter\_worst 92.229

radius\_worst 91.416

area\_worst 90.927

concave.points\_mean 71.659

perimeter\_mean 68.331

area\_se 67.317

concavity\_mean 55.527

radius\_mean 54.184

area\_mean 50.820

concavity\_worst 37.998

radius\_se 30.322

compactness\_worst 28.366

texture\_worst 24.485

perimeter\_se 21.902

texture\_mean 20.608

compactness\_mean 15.221

smoothness\_worst 15.165

concavity\_se 10.881

symmetry\_worst 9.291

> rocmodel3

Data: model3data$rfmodel3pred in 92 controls (model3data$diagnosis 1) < 51 cases (model3data$diagnosis 2).

Area under the curve: 0.9597

> rfmodel4

Random Forest

426 samples

30 predictor

2 classes: 'B', 'M'

No pre-processing

Resampling: Cross-Validated (10 fold, repeated 3 times)

Summary of sample sizes: 383, 384, 383, 383, 383, 384, ...

Resampling results:

Accuracy Kappa

0.9616631 0.9182918

Tuning parameter 'mtry' was held constant at a value of 5.477226

> rfmodel4$finalModel

OOB estimate of error rate: 3.76%

Confusion matrix:

B M class.error

B 259 6 0.02264151

M 10 151 0.06211180

> table(CanDattrain$diagnosis)

B M

265 161

> table(rfmodel4pred, CanDattest$diagnosis)

rfmodel4pred B M

B 90 2

M 2 49

> varImp(rfmodel4)

rf variable importance

Overall

area\_worst 100.000

concave.points\_worst 91.350

perimeter\_worst 79.382

radius\_worst 67.548

concave.points\_mean 59.705

concavity\_mean 36.983

perimeter\_mean 35.573

area\_se 28.352

area\_mean 27.823

radius\_mean 26.159

concavity\_worst 24.387

texture\_worst 13.277

smoothness\_worst 8.405

perimeter\_se 8.261

radius\_se 7.660

texture\_mean 7.441

compactness\_worst 7.265

symmetry\_worst 4.987

compactness\_mean 4.187

fractal\_dimension\_worst 3.360

> rocmodel4

Data: model4data$rfmodel4pred in 92 controls (model4data$diagnosis 1) < 51 cases (model4data$diagnosis 2).

Area under the curve: 0.9695

> rfmodel5

Random Forest

426 samples

30 predictor

2 classes: 'B', 'M'

No pre-processing

Resampling: Cross-Validated (10 fold, repeated 3 times)

Summary of sample sizes: 383, 384, 383, 384, 384, 383, ...

Resampling results across tuning parameters:

mtry Accuracy Kappa

3 0.9578040 0.9094862

4 0.9547023 0.9031803

5 0.9562167 0.9064137

6 0.9577855 0.9097805

7 0.9594097 0.9129269

8 0.9608871 0.9168029

11 0.9585791 0.9115270

12 0.9562527 0.9064542

14 0.9585607 0.9115547

24 0.9562536 0.9068398

30 0.9554968 0.9051949

Accuracy was used to select the optimal model using the largest value.

The final value used for the model was mtry = 8.

> rfmodel3$finalModel

OOB estimate of error rate: 3.05%

Confusion matrix:

B M class.error

B 262 3 0.01132075

M 10 151 0.06211180

> table(CanDattrain$diagnosis)

B M

265 161

> varImp(rfmodel5)

rf variable importance

Overall

perimeter\_worst 100.000

area\_worst 97.580

concave.points\_worst 87.889

radius\_worst 83.114

concave.points\_mean 57.440

area\_mean 24.251

concavity\_mean 19.554

perimeter\_mean 19.467

area\_se 17.883

radius\_mean 14.428

texture\_worst 13.879

concavity\_worst 12.724

texture\_mean 8.593

smoothness\_worst 6.428

compactness\_worst 4.696

radius\_se 4.491

symmetry\_worst 4.239

perimeter\_se 4.238

smoothness\_mean 2.365

fractal\_dimension\_worst 2.272

> table(rfmodel5pred, CanDattest$diagnosis)

rfmodel5pred B M

B 90 3

M 2 48

> rocmodel5

Data: model5data$rfmodel5pred in 92 controls (model5data$diagnosis 1) < 51 cases (model5data$diagnosis 2).

Area under the curve: 0.9597

> rfmodel6

Stochastic Gradient Boosting

426 samples

30 predictor

2 classes: 'B', 'M'

No pre-processing

Resampling: Cross-Validated (5 fold, repeated 3 times)

Summary of sample sizes: 340, 341, 341, 341, 341, 341, ...

Resampling results across tuning parameters:

interaction.depth n.trees Accuracy Kappa

1 50 0.9514729 0.8957032

1 100 0.9624259 0.9196834

1 150 0.9640036 0.9231846

2 50 0.9585317 0.9114331

2 100 0.9679343 0.9314876

2 150 0.9663566 0.9281817

3 50 0.9546284 0.9026636

3 100 0.9577474 0.9094997

3 150 0.9608755 0.9160948

Tuning parameter 'shrinkage' was held constant at a value of 0.1

Tuning

parameter 'n.minobsinnode' was held constant at a value of 10

Accuracy was used to select the optimal model using the largest value.

The final values used for the model were n.trees = 100, interaction.depth =

2, shrinkage = 0.1 and n.minobsinnode = 10.

> rfmodel6$finalModel

A gradient boosted model with bernoulli loss function.

100 iterations were performed.

There were 30 predictors of which 30 had non-zero influence.

> table(CanDattrain$diagnosis)

B M

265 161

> varImp(rfmodel6)

Error in relative.influence(object, n.trees = numTrees) :

could not find function "relative.influence"

> table(rfmodel6pred, CanDattest$diagnosis)

rfmodel6pred B M

B 90 2

M 2 49

> rocmodel6

Data: model6data$rfmodel6pred in 92 controls (model6data$diagnosis 1) < 51 cases (model6data$diagnosis 2).

Area under the curve: 0.9695

> Accuracy(CM3)

[1] 0.965035

> RecallorSensitivity(CM3)

[1] 0.9782609

> Specificity(CM3)

[1] 0.9411765

> Precision(CM3)

[1] 0.9677419

> Accuracy(CM4)

[1] 0.972028

> RecallorSensitivity(CM4)

[1] 0.9782609

> Specificity(CM4)

[1] 0.9607843

> Precision(CM4)

[1] 0.9782609

> Accuracy(CM5)

[1] 0.965035

> RecallorSensitivity(CM5)

[1] 0.9782609

> Specificity(CM5)

[1] 0.9411765

> Precision(CM5)

[1] 0.9677419

> Accuracy(CM6)

[1] 0.972028

> RecallorSensitivity(CM6)

[1] 0.9782609

> Specificity(CM6)

[1] 0.9607843

> Precision(CM6)

[1] 0.9782609