***Eric Agyemang***

**REGRESSION MODELLING PROBLEM USING CARET**

**library**(caret)

## Loading required package: lattice ## Loading required package: ggplot2

**library**(lattice) **library**(ggplot2) **library**(mlbench) **library**(survival)

##

## Attaching package: 'survival'

## The following object is masked from 'package:caret':

##

## cluster

**library**(parallel) **library**(splines) **library**(plyr) **library**(gbm)

## Loaded gbm 2.1.4

**library**(glmnet)

## Loading required package: Matrix

## Loading required package: foreach ## Loaded glmnet 2.0-16

**library**(MASS) **library**(C50) **library**(naivebayes) **library**(klaR) **library**(kernlab)

##

## Attaching package: 'kernlab'

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

##

## alpha

**library**(Matrix) **library**(foreach)

Loaind data set.

**data**("iris")

iris**$**Species = **as.factor**(iris**$**Species) **head**(iris)

## Sepal.Length Sepal.Width Petal.Length Petal.Width Species

## 1 5.1 3.5 1.4 0.2 setosa ## 2 4.9 3.0 1.4 0.2 setosa ## 3 4.7 3.2 1.3 0.2 setosa ## 4 4.6 3.1 1.5 0.2 setosa ## 5 5.0 3.6 1.4 0.2 setosa

## 6 5.4 3.9 1.7 0.4 setosa

Preprocessing

**sum**(**is.na**(iris))

## [1] 0

Getting all list of algorithm that CARET supports.

**names**(**getModelInfo**())

|  |  |  |  |
| --- | --- | --- | --- |
| ## [1] "ada" | "AdaBag" | "AdaBoost.M1" | |
| ## [4] "adaboost" | "amdai" | "ANFIS" | |
| ## [7] "avNNet" | "awnb" | "awtan" | |
| ## [10] "bag" | "bagEarth" | "bagEarthGCV" | |
| ## [13] "bagFDA" | "bagFDAGCV" | "bam" | |
| ## [16] "bartMachine" | "bayesglm" | "binda" | |
| ## [19] "blackboost" | "blasso" | "blassoAveraged" | |
| ## [22] "bridge" | "brnn" | "BstLm" | |
| ## [25] "bstSm" | "bstTree" | "C5.0" | |
| ## [28] "C5.0Cost" | "C5.0Rules" | "C5.0Tree" | |
| ## [31] "cforest" | "chaid" | "CSimca" | |
| ## [34] "ctree" | "ctree2" | "cubist" | |
| ## [37] "dda" | "deepboost" | "DENFIS" | |
| ## [40] "dnn" | "dwdLinear" | "dwdPoly" | |
| ## [43] "dwdRadial" | "earth" | "elm" | |
| ## [46] "enet" | "evtree" | "extraTrees" | |
| ## [49] "fda" | "FH.GBML" | "FIR.DM" | |
| ## [52] "foba" | "FRBCS.CHI" | "FRBCS.W" | |
| ## [55] "FS.HGD" | "gam" | "gamboost" | |
| ## [58] "gamLoess" | "gamSpline" | "gaussprLinear" | |
| ## [61] "gaussprPoly" | "gaussprRadial" | "gbm\_h2o" | |
| ## [64] "gbm" | "gcvEarth" | "GFS.FR.MOGUL" | |
| ## [67] "GFS.LT.RS" | "GFS.THRIFT" | "glm.nb" | |
| ## [70] "glm" | "glmboost" | "glmnet\_h2o" | |
| ## [73] "glmnet" | "glmStepAIC" | "gpls" | |
| ## [76] "hda" | "hdda" | "hdrda" | |
| ## [79] "HYFIS" | "icr" | "J48" | |
| ## [82] "JRip" | "kernelpls" | "kknn" | |
| ## [85] "knn" | "krlsPoly" | "krlsRadial" | |
| ## [88] "lars" | "lars2" | "lasso" | |
| ## [91] "lda" | "lda2" | "leapBackward" | |
| ## [94] "leapForward" | "leapSeq" | "Linda" | |
| ## [97] "lm" | "lmStepAIC" | "LMT" | |
| ## [100] "loclda" | "logicBag" | "LogitBoost" | |
| ## [103] "logreg" | "lssvmLinear" | | "lssvmPoly" | |
| ## [106] "lssvmRadial" | "lvq" | | "M5" | |
| ## [109] "M5Rules" | "manb" | | "mda" | |
| ## [112] "Mlda" | "mlp" | | "mlpKerasDecay" | |
| ## [115] "mlpKerasDecayCost" | "mlpKerasDropout" | | "mlpKerasDropoutCost" | |
| ## [118] "mlpML" | "mlpSGD" | | "mlpWeightDecay" | |
| ## [121] "mlpWeightDecayML" | "monmlp" | | "msaenet" | |
| ## [124] "multinom" | "mxnet" | | "mxnetAdam" | |
| ## [127] "naive\_bayes" | "nb" | | "nbDiscrete" | |
| ## [130] "nbSearch" | "neuralnet" | | "nnet" | |
| ## [133] "nnls" | "nodeHarvest" | | "null" | |
| ## [136] "OneR" | "ordinalNet" | | "ORFlog" | |
| ## [139] "ORFpls" | "ORFridge" | | "ORFsvm" | |
| ## [142] "ownn" | "pam" | | "parRF" | |
| ## [145] "PART" | "partDSA" | | "pcaNNet" | |
| ## [148] "pcr" | "pda" | | "pda2" | |
| ## [151] "penalized" | "PenalizedLDA" | | "plr" | |
| ## [154] "pls" | "plsRglm" | | "polr" | |
| ## [157] "ppr" | "PRIM" | | "protoclass" | |
| ## [160] "qda" | "QdaCov" | | "qrf" | |
| ## [163] "qrnn" | "randomGLM" | | "ranger" | |
| ## [166] "rbf" | "rbfDDA" | | "Rborist" | |
| ## [169] "rda" | "regLogistic" | | "relaxo" | |
| ## [172] "rf" | "rFerns" | | "RFlda" | |
| ## [175] "rfRules" | "ridge" | | "rlda" | |
| ## [178] "rlm" | "rmda" | | "rocc" | |
| ## [181] "rotationForest" | "rotationForestCp" | | "rpart" | |
| ## [184] "rpart1SE" | "rpart2" | | "rpartCost" | |
| ## [187] "rpartScore" | "rqlasso" | | "rqnc" | |
| ## [190] "RRF" | "RRFglobal" | | "rrlda" | |
| ## [193] "RSimca" | "rvmLinear" | | "rvmPoly" | |
| ## [196] "rvmRadial" | "SBC" | | "sda" | |
| ## [199] "sdwd" | "simpls" | | "SLAVE" | |
| ## [202] "slda" | "smda" | | "snn" | |
| ## [205] "sparseLDA" | "spikeslab" | | "spls" | |
| ## [208] "stepLDA" "stepQDA" | | | "superpc" | |
| ## [211] "svmBoundrangeString" "svmExpoString" | | | "svmLinear" | |
| ## [214] "svmLinear2" "svmLinear3" | | | "svmLinearWeights" | |
| ## [217] "svmLinearWeights2" "svmPoly" | | | "svmRadial" | |
| ## [220] "svmRadialCost" "svmRadialSigma" | | | "svmRadialWeights" | |
| ## [223] "svmSpectrumString" "tan" | | | "tanSearch" | |
| ## [226] "treebag" "vbmpRadial" | | | "vglmAdjCat" | |
| ## [229] "vglmContRatio" "vglmCumulative" | | | "widekernelpls" | |
| ## [232] "WM" "wsrf" | | | "xgbDART" | |
| ## [235] "xgbLinear" "xgbTree" | | | "xyf" | |

Splitting with CARET,we create testing and training data set using CARET.This will enable us develop and evaluate our model.

inTrain = **createDataPartition**(y = iris**$**Species, p = .75, list = FALSE) training = iris[inTrain,] testing = iris[**-**inTrain,]

Next,we create the cross validationg method that will be used by CARET to create the training sets. fitControl <- **trainControl**(

method = "repeatedcv", number = 10, repeats = 10)

Now we ready to develop the model.we make use of the train function in CARET to regress the dependant variable onto all of other covariates.

|  |
| --- |
| **set.seed**(12345)  gbmFit1 <- **train**(Species **~** ., data = training,  method = "gbm", trControl = fitControl, verbose = FALSE) |

We use CARET to choose our final model thus interaction depth of 1,number trees at 50, accuracay of 94% and a kappa of 90%

gbmFit1

## Stochastic Gradient Boosting

##

## 114 samples

## 4 predictor

## 3 classes: 'setosa', 'versicolor', 'virginica'

##

## No pre-processing

## Resampling: Cross-Validated (10 fold, repeated 10 times) ## Summary of sample sizes: 102, 104, 102, 103, 102, 103, ...

## Resampling results across tuning parameters:

##

## interaction.depth n.trees Accuracy Kappa

## 1 50 0.9359394 0.9037010 ## 1 100 0.9358788 0.9036017 ## 1 150 0.9367879 0.9049767 ## 2 50 0.9358788 0.9035686 ## 2 100 0.9351212 0.9024266 ## 2 150 0.9358636 0.9035591 ## 3 50 0.9348788 0.9020417 ## 3 100 0.9350455 0.9023012

## 3 150 0.9332121 0.8995586

##

## 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 = 150,

## interaction.depth = 1, shrinkage = 0.1 and n.minobsinnode = 10.

**plot**(gbmFit1)

Accuracy (Repeated Cross−Validation)

0.934

0.935

0.936

60

80

100

120

140

Max Tree Depth

1

2

3

# # Boosting Iterations

Using the training model to make our predictions.As we can see there is 99% probability that the first flow in the data set is a setosa species.

**predict**(gbmFit1, newdata = **head**(testing), type = "prob")

## setosa versicolor virginica

## 1 0.9999978 2.172458e-06 2.797920e-10 ## 2 0.9999988 1.153803e-06 5.862131e-11 ## 3 0.9999929 7.142686e-06 2.984878e-11 ## 4 0.9999985 1.523831e-06 7.742135e-11 ## 5 0.9999978 2.172458e-06 3.524350e-10 ## 6 0.9999716 2.836760e-05 2.094037e-10

**summary**(education)

## gender NationalITy PlaceofBirth StageID

## F:175 KW :179 KuwaIT :180 HighSchool : 33 ## M:305 Jordan :172 Jordan :176 lowerlevel :199 ## Palestine: 28 Iraq : 22 MiddleSchool:248

## Iraq : 22 lebanon : 19

## lebanon : 17 SaudiArabia: 16

## Tunis : 12 USA : 16

## (Other) : 50 (Other) : 51

## GradeID SectionID Topic Semester Relation ## G-02 :147 A:283 IT : 95 F:245 Father:283 ## G-08 :116 B:167 French : 65 S:235 Mum :197

## G-07 :101 C: 30 Arabic : 59 ## G-04 : 48 Science: 51 ## G-06 : 32 English: 45 ## G-11 : 13 Biology: 30

|  |  |
| --- | --- |
| ## (Other): 23 | (Other):135 |
| ## raisedhands | VisITedResources AnnouncementsView Discussion |
| ## Min. : 0.00 | Min. : 0.0 Min. : 0.00 Min. : 1.00 |
| ## 1st Qu.: 15.75 | 1st Qu.:20.0 1st Qu.:14.00 1st Qu.:20.00 |
| ## Median : 50.00 | Median :65.0 Median :33.00 Median :39.00 |
| ## Mean : 46.77 | Mean :54.8 Mean :37.92 Mean :43.28 |
| ## 3rd Qu.: 75.00 | 3rd Qu.:84.0 3rd Qu.:58.00 3rd Qu.:70.00 |
| ## Max. :100.00 | Max. :99.0 Max. :98.00 Max. :99.00 |

##

## ParentAnsweringSurvey ParentschoolSatisfaction StudentAbsenceDays Class

## No :210 Bad :188 Above-7:191 H:142 ## Yes:270 Good:292 Under-7:289 L:127

## M:211

## ## ## ##

|  |
| --- |
| **set.seed**(17)  *# Stratified sampling*  TrainingDataIndex <- **createDataPartition**(education**$**Class, p=0.75, list = FALSE)  *# Create Training Data*  trainingData <- education[TrainingDataIndex,] testData <- education[**-**TrainingDataIndex,]  TrainingParameters <- **trainControl**(method = "repeatedcv", number = 10, repeats=10) |
| *# Train a model with above parameters. We will use C5.0 algorithm*  DecTreeModel <- **train**(Class **~** ., data = trainingData, method = "C5.0",  preProcess=**c**("scale","center"), trControl= TrainingParameters, na.action = na.omit ) |

## Warning in preProcess.default(thresh = 0.95, k = 5, freqCut = 19,

## uniqueCut = 10, : These variables have zero variances: NationalITyvenzuela, ## PlaceofBirthvenzuela

## Warning in preProcess.default(thresh = 0.95, k = 5, freqCut = 19,

## uniqueCut = 10, : These variables have zero variances: NationalITyvenzuela, ## PlaceofBirthvenzuela

## Warning in preProcess.default(thresh = 0.95, k = 5, freqCut = 19,

## uniqueCut = 10, : These variables have zero variances: NationalITyvenzuela,

## PlaceofBirthvenzuela

## Warning in preProcess.default(thresh = 0.95, k = 5, freqCut = 19,

## uniqueCut = 10, : These variables have zero variances: NationalITyvenzuela,

## PlaceofBirthvenzuela

## Warning in preProcess.default(thresh = 0.95, k = 5, freqCut = 19,

## uniqueCut = 10, : These variables have zero variances: NationalITyvenzuela, ## PlaceofBirthvenzuela

## Warning in preProcess.default(thresh = 0.95, k = 5, freqCut = 19,

## uniqueCut = 10, : These variables have zero variances: NationalITyvenzuela, ## PlaceofBirthvenzuela

## Warning in preProcess.default(thresh = 0.95, k = 5, freqCut = 19,

## uniqueCut = 10, : These variables have zero variances: NationalITyvenzuela, ## PlaceofBirthvenzuela

## Warning in preProcess.default(thresh = 0.95, k = 5, freqCut = 19,

## uniqueCut = 10, : These variables have zero variances: NationalITyvenzuela, ## PlaceofBirthvenzuela

## Warning: 'trials' should be <= 8 for this object. Predictions generated ## using 8 trials

## Warning in preProcess.default(thresh = 0.95, k = 5, freqCut = 19,

## uniqueCut = 10, : These variables have zero variances: NationalITyvenzuela, ## PlaceofBirthvenzuela

## Warning in preProcess.default(thresh = 0.95, k = 5, freqCut = 19,

## uniqueCut = 10, : These variables have zero variances: NationalITyvenzuela, ## PlaceofBirthvenzuela

## Warning in preProcess.default(thresh = 0.95, k = 5, freqCut = 19,

## uniqueCut = 10, : These variables have zero variances: NationalITyvenzuela, ## PlaceofBirthvenzuela

## Warning in preProcess.default(thresh = 0.95, k = 5, freqCut = 19,

## uniqueCut = 10, : These variables have zero variances: NationalITyvenzuela,

## PlaceofBirthvenzuela

## Warning in preProcess.default(thresh = 0.95, k = 5, freqCut = 19,

## uniqueCut = 10, : These variables have zero variances: NationalITyvenzuela, ## PlaceofBirthvenzuela

## Warning in preProcess.default(thresh = 0.95, k = 5, freqCut = 19,

## uniqueCut = 10, : These variables have zero variances: NationalITyvenzuela, ## PlaceofBirthvenzuela

## Warning in preProcess.default(thresh = 0.95, k = 5, freqCut = 19,

## uniqueCut = 10, : These variables have zero variances: NationalITyvenzuela,

## PlaceofBirthvenzuela

## Warning in preProcess.default(thresh = 0.95, k = 5, freqCut = 19,

## uniqueCut = 10, : These variables have zero variances: NationalITyvenzuela,

## PlaceofBirthvenzuela

## Warning in preProcess.default(thresh = 0.95, k = 5, freqCut = 19,

## uniqueCut = 10, : These variables have zero variances: NationalITyvenzuela,

## PlaceofBirthvenzuela

## Warning in preProcess.default(thresh = 0.95, k = 5, freqCut = 19,

## uniqueCut = 10, : These variables have zero variances: NationalITyvenzuela,

## PlaceofBirthvenzuela

## Warning in preProcess.default(thresh = 0.95, k = 5, freqCut = 19,

## uniqueCut = 10, : These variables have zero variances: NationalITyvenzuela, ## PlaceofBirthvenzuela

## Warning in preProcess.default(thresh = 0.95, k = 5, freqCut = 19,

## uniqueCut = 10, : These variables have zero variances: NationalITyvenzuela, ## PlaceofBirthvenzuela

## Warning: 'trials' should be <= 3 for this object. Predictions generated ## using 3 trials

## Warning in preProcess.default(thresh = 0.95, k = 5, freqCut = 19,

## uniqueCut = 10, : These variables have zero variances: NationalITyvenzuela, ## PlaceofBirthvenzuela

## Warning in preProcess.default(thresh = 0.95, k = 5, freqCut = 19,

## uniqueCut = 10, : These variables have zero variances: NationalITyvenzuela, ## PlaceofBirthvenzuela

## Warning in preProcess.default(thresh = 0.95, k = 5, freqCut = 19,

## uniqueCut = 10, : These variables have zero variances: NationalITyvenzuela, ## PlaceofBirthvenzuela

## Warning in preProcess.default(thresh = 0.95, k = 5, freqCut = 19,

## uniqueCut = 10, : These variables have zero variances: NationalITyvenzuela,

## PlaceofBirthvenzuela

## Warning in preProcess.default(thresh = 0.95, k = 5, freqCut = 19,

## uniqueCut = 10, : These variables have zero variances: NationalITyvenzuela, ## PlaceofBirthvenzuela

## Warning in preProcess.default(thresh = 0.95, k = 5, freqCut = 19,

## uniqueCut = 10, : These variables have zero variances: NationalITyvenzuela, ## PlaceofBirthvenzuela

## Warning in preProcess.default(thresh = 0.95, k = 5, freqCut = 19,

## uniqueCut = 10, : These variables have zero variances: NationalITyvenzuela, ## PlaceofBirthvenzuela

## Warning in preProcess.default(thresh = 0.95, k = 5, freqCut = 19,

## uniqueCut = 10, : These variables have zero variances: NationalITyvenzuela,

## PlaceofBirthvenzuela

## Warning in preProcess.default(thresh = 0.95, k = 5, freqCut = 19,

## uniqueCut = 10, : These variables have zero variances: NationalITyvenzuela,

## PlaceofBirthvenzuela

## Warning in preProcess.default(thresh = 0.95, k = 5, freqCut = 19,

## uniqueCut = 10, : These variables have zero variances: NationalITyvenzuela,

## PlaceofBirthvenzuela

## Warning in preProcess.default(thresh = 0.95, k = 5, freqCut = 19,

## uniqueCut = 10, : These variables have zero variances: NationalITyvenzuela, ## PlaceofBirthvenzuela

## Warning in preProcess.default(thresh = 0.95, k = 5, freqCut = 19,

## uniqueCut = 10, : These variables have zero variances: NationalITyvenzuela, ## PlaceofBirthvenzuela

## Warning in preProcess.default(thresh = 0.95, k = 5, freqCut = 19,

## uniqueCut = 10, : These variables have zero variances: NationalITyvenzuela, ## PlaceofBirthvenzuela

## Warning in preProcess.default(thresh = 0.95, k = 5, freqCut = 19,

## uniqueCut = 10, : These variables have zero variances: NationalITyvenzuela, ## PlaceofBirthvenzuela

## Warning in preProcess.default(thresh = 0.95, k = 5, freqCut = 19,

## uniqueCut = 10, : These variables have zero variances: NationalITyvenzuela, ## PlaceofBirthvenzuela

## Warning in preProcess.default(thresh = 0.95, k = 5, freqCut = 19,

## uniqueCut = 10, : These variables have zero variances: NationalITyvenzuela,

## PlaceofBirthvenzuela

## Warning in preProcess.default(thresh = 0.95, k = 5, freqCut = 19,

## uniqueCut = 10, : These variables have zero variances: NationalITyvenzuela, ## PlaceofBirthvenzuela

## Warning in preProcess.default(thresh = 0.95, k = 5, freqCut = 19,

## uniqueCut = 10, : These variables have zero variances: NationalITyvenzuela, ## PlaceofBirthvenzuela

## Warning in preProcess.default(thresh = 0.95, k = 5, freqCut = 19,

## uniqueCut = 10, : These variables have zero variances: NationalITyvenzuela, ## PlaceofBirthvenzuela

## Warning in preProcess.default(thresh = 0.95, k = 5, freqCut = 19,

## uniqueCut = 10, : These variables have zero variances: NationalITyvenzuela, ## PlaceofBirthvenzuela

*#Predictions* DTPredictions <-**predict**(DecTreeModel, testData, na.action = na.pass)

*# Print confusion matrix and results*

cmTree <-**confusionMatrix**(DTPredictions, testData**$**Class) **print**(cmTree)

## Confusion Matrix and Statistics

##

## Reference

## Prediction H L M

## H 25 0 11 ## L 1 29 3

## M 9 2 38

##

## Overall Statistics

##

## Accuracy : 0.7797

## 95% CI : (0.6941, 0.8507)

## No Information Rate : 0.4407

## P-Value [Acc > NIR] : 5.958e-14

##

## Kappa : 0.6626 ## Mcnemar's Test P-Value : 0.7055

|  |  |
| --- | --- |
| ##  ## Statistics by Class:  ## |  |
| ## | Class: H Class: L Class: M |
| ## Sensitivity | 0.7143 0.9355 0.7308 |
| ## Specificity | 0.8675 0.9540 0.8333 |
| ## Pos Pred Value | 0.6944 0.8788 0.7755 |
| ## Neg Pred Value | 0.8780 0.9765 0.7971 |
| ## Prevalence | 0.2966 0.2627 0.4407 |
| ## Detection Rate | 0.2119 0.2458 0.3220 |
| ## Detection Prevalence | 0.3051 0.2797 0.4153 |
| ## Balanced Accuracy | 0.7909 0.9448 0.7821 |