hr_logistic_regression_using_caret_package

June 9, 2018

0.1 In this exercise, we will use the HR dataset and understand the following using caret package:

1. Building the logistic regression model
2. What is marked as the positive class by the model when using caret package
3. Writing the model equation and interpreting the model summary
4. Creating the Confusion Matrix and ROC plot on train data
5. Using mis-classification cost as a criteria to select the best cut-off
6. Using Younden Index as the criteria to select the best cut-off
7. Creating the Confusion Matrix and ROC plot on test data
8. Compare and discuss the result of logistic regression using caret vis-a-via stats package
9. Changing the base or reference category and evaluate the impact on the model (This is self work/assignment)
10. Change the cut-off value for train data in caret package (This is self work/assignment)

There are bugs/missing code in the entire exercise. The participants are expected to work upon them.

0.2 Here are some useful links:

1. **Read** about interaction variable coding
2. Refer **link** to know about adding lables to factors
3. Refer **link** to relvel factor variables
4. **Read** about the issues in stepwise regression
5. **Read** about the modelling activity via caret package
6. The **complete** list of tuning parameter for different models in caret package

---

1 Code starts here

We are going to use below mentioned libraries for demonstrating logistic regression:

```
In [1]: library(caret)     #for data partition. Model building
        #library(Deducer)  #for ROC plot
        library(ROCR)      #for ROC plot (other way)
```

```
Loading required package: lattice
Loading required package: ggplot2
Loading required package: gplots

Attaching package: gplots

The following object is masked from package:stats:

    lowess
```

1.1 Data Import and Manipulation

1.1.1 1. Importing a data set

Give the correct path to the data

```
In [2]: raw_df <- read.csv("/Users/Rahul/Documents/Datasets/IMB533_HR_Data_No_Missing_Value.csv
```

Note that `echo = FALSE` parameter prevents printing the R code that generated the plot.

1.1.2 2. Structure and Summary of the dataset

```
In [3]: str(raw_df)
        summary(raw_df)
```

```
'data.frame':        8995 obs. of  18 variables:
 $ SLNO                      : int  1 2 3 4 5 6 7 9 11 12 ...
 $ Candidate.Ref             : int  2110407 2112635 2112838 2115021 2115125 2117167 2119124 2
 $ DOJ.Extended              : Factor w/ 2 levels "No","Yes": 2 1 1 1 2 2 2 2 1 1 ...
 $ Duration.to.accept.offer  : int  14 18 3 26 1 17 37 16 1 6 ...
 $ Notice.period             : int  30 30 45 30 120 30 30 0 30 30 ...
 $ Offered.band              : Factor w/ 4 levels "E0","E1","E2",..: 3 3 3 3 3 2 3 2 2 2 ...
 $ Pecent.hike.expected.in.CTC: num  -20.8 50 42.8 42.8 42.6 ...
 $ Percent.hike.offered.in.CTC: num  13.2 320 42.8 42.8 42.6 ...
 $ Percent.difference.CTC    : num  42.9 180 0 0 0 ...
 $ Joining.Bonus             : Factor w/ 2 levels "No","Yes": 1 1 1 1 1 1 1 1 1 1 ...
 $ Candidate.relocate.actual : Factor w/ 2 levels "No","Yes": 1 1 1 1 2 1 1 1 1 1 ...
 $ Gender                    : Factor w/ 2 levels "Female","Male": 1 2 2 2 2 2 2 1 1 2 ...
 $ Candidate.Source          : Factor w/ 3 levels "Agency","Direct",..: 1 3 1 3 3 3 3 2 3 3 .
 $ Rex.in.Yrs                : int  7 8 4 4 6 2 7 8 3 3 ...
 $ LOB                       : Factor w/ 9 levels "AXON","BFSI",..: 5 8 8 8 8 8 8 7 2 3 ...
 $ Location                  : Factor w/ 11 levels "Ahmedabad","Bangalore",..: 9 3 9 9 9 9 9
 $ Age                       : int  34 34 27 34 34 34 32 34 26 34 ...
 $ Status                    : Factor w/ 2 levels "Joined","Not Joined": 1 1 1 1 1 1 1 1 1 1


      SLNO       Candidate.Ref      DOJ.Extended Duration.to.accept.offer
```

```
Min.   :    1   Min.   :2109586   No :4788     Min.   :  0.00
1st Qu.: 3208   1st Qu.:2386476   Yes:4207     1st Qu.:  3.00
Median : 5976   Median :2807482                Median : 10.00
Mean   : 5971   Mean   :2843647                Mean   : 21.43
3rd Qu.: 8739   3rd Qu.:3300060                3rd Qu.: 33.00
Max.   :12333   Max.   :3836076                Max.   :224.00

Notice.period    Offered.band Pecent.hike.expected.in.CTC
Min.   :  0.00   E0: 211      Min.   :-68.83
1st Qu.: 30.00   E1:5568      1st Qu.: 27.27
Median : 30.00   E2:2711      Median : 40.00
Mean   : 39.29   E3: 505      Mean   : 43.86
3rd Qu.: 60.00                3rd Qu.: 53.85
Max.   :120.00                Max.   :359.77

Percent.hike.offered.in.CTC Percent.difference.CTC Joining.Bonus
Min.   :-60.53              Min.   :-67.270        No :8578
1st Qu.: 22.09              1st Qu.: -8.330        Yes: 417
Median : 36.00              Median :  0.000
Mean   : 40.66              Mean   : -1.574
3rd Qu.: 50.00              3rd Qu.:  0.000
Max.   :471.43              Max.   :300.000

Candidate.relocate.actual    Gender             Candidate.Source
No :7705                  Female:1551   Agency           :2585
Yes:1290                  Male  :7444   Direct           :4801
                                        Employee Referral:1609

  Rex.in.Yrs          LOB           Location          Age
Min.   : 0.000   INFRA  :2850   Chennai  :3150   Min.   :20.00
1st Qu.: 3.000   ERS    :2426   Noida    :2727   1st Qu.:27.00
Median : 4.000   BFSI   :1396   Bangalore:2230   Median :29.00
Mean   : 4.239   ETS    : 691   Hyderabad: 341   Mean   :29.91
3rd Qu.: 6.000   CSMP   : 579   Mumbai   : 197   3rd Qu.:34.00
Max.   :24.000   AXON   : 568   Gurgaon  : 146   Max.   :60.00
                 (Other): 485   (Other)  : 204
       Status
Joined    :7313
Not Joined:1682
```

Create a new data frame and store the raw data copy. This is being done to have a copy of the raw data intact for further manipulation if needed.

```
In [4]: filter_df <- na.omit(raw_df) # listwise deletion of missing
```

1.1.3 3. Create train and test dataset

Reserve 80% for *training* and 20% of *test* *Correct the error in the below code chunk*

```
In [5]: set.seed(2341)
        trainIndex <- createDataPartition(filter_df$Status, p = 0.80, list = FALSE)
        train_df <- filter_df[trainIndex,]
        test_df <- filter_df[-trainIndex,]
```

We can pull the specific attribute needed to build the model is another data frame. This agian is more of a hygine practice to not touch the **train** and **test** data set directly.
Correct the error in the below code chunk

```
In [6]: lg_train_df <- as.data.frame(train_df[,c("DOJ.Extended",
                                                 "Duration.to.accept.offer",
                                                 "Notice.period",
                                                 "Offered.band",
                                                 "Percent.difference.CTC",
                                                 "Joining.Bonus",
                                                 "Gender",
                                                 "Candidate.Source",
                                                 "Rex.in.Yrs",
                                                 "LOB",
                                                 "Location",
                                                 "Age",
                                                 "Status"
        )])
```

Correct the error in the below code chunk

```
In [7]: lg_test_df <- as.data.frame(test_df[,c("DOJ.Extended",
                                               "Duration.to.accept.offer",
                                               "Notice.period",
                                               "Offered.band",
                                               "Percent.difference.CTC",
                                               "Joining.Bonus",
                                               "Gender",
                                               "Candidate.Source",
                                               "Rex.in.Yrs",
                                               "LOB",
                                               "Location",
                                               "Age",
                                               "Status"
        )])
```

1.2 Model Building: Using the caret() package

There are a number of models which can be built using caret package. To get the names of all the models possible.

```
In [8]: names(getModelInfo())
```

1. 'ada' 2. 'AdaBag' 3. 'AdaBoost.M1' 4. 'adaboost' 5. 'amdai' 6. 'ANFIS' 7. 'avNNet' 8. 'awnb' 9. 'awtan' 10. 'bag' 11. 'bagEarth' 12. 'bagEarthGCV' 13. 'bagFDA' 14. 'bagFDAGCV' 15. 'bam' 16. 'bartMachine' 17. 'bayesglm' 18. 'binda' 19. 'blackboost' 20. 'blasso' 21. 'blassoAveraged' 22. 'bridge' 23. 'brnn' 24. 'BstLm' 25. 'bstSm' 26. 'bstTree' 27. 'C5.0' 28. 'C5.0Cost' 29. 'C5.0Rules' 30. 'C5.0Tree' 31. 'cforest' 32. 'chaid' 33. 'CSimca' 34. 'ctree' 35. 'ctree2' 36. 'cubist' 37. 'dda' 38. 'deepboost' 39. 'DENFIS' 40. 'dnn' 41. 'dwdLinear' 42. 'dwdPoly' 43. 'dwdRadial' 44. 'earth' 45. 'elm' 46. 'enet' 47. 'evtree' 48. 'extraTrees' 49. 'fda' 50. 'FH.GBML' 51. 'FIR.DM' 52. 'foba' 53. 'FRBCS.CHI' 54. 'FRBCS.W' 55. 'FS.HGD' 56. 'gam' 57. 'gamboost' 58. 'gamLoess' 59. 'gamSpline' 60. 'gaussprLinear' 61. 'gaussprPoly' 62. 'gaussprRadial' 63. 'gbm_h2o' 64. 'gbm' 65. 'gcvEarth' 66. 'GFS.FR.MOGUL' 67. 'GFS.LT.RS' 68. 'GFS.THRIFT' 69. 'glm.nb' 70. 'glm' 71. 'glmboost' 72. 'glmnet_h2o' 73. 'glmnet' 74. 'glmStepAIC' 75. 'gpls' 76. 'hda' 77. 'hdda' 78. 'hdrda' 79. 'HYFIS' 80. 'icr' 81. 'J48' 82. 'JRip' 83. 'kernelpls' 84. 'kknn' 85. 'knn' 86. 'krlsPoly' 87. 'krlsRadial' 88. 'lars' 89. 'lars2' 90. 'lasso' 91. 'lda' 92. 'lda2' 93. 'leapBackward' 94. 'leapForward' 95. 'leapSeq' 96. 'Linda' 97. 'lm' 98. 'lmStepAIC' 99. 'LMT' 100. 'loclda' 101. 'logicBag' 102. 'LogitBoost' 103. 'logreg' 104. 'lssvmLinear' 105. 'lssvmPoly' 106. 'lssvmRadial' 107. 'lvq' 108. 'M5' 109. 'M5Rules' 110. 'manb' 111. 'mda' 112. 'Mlda' 113. 'mlp' 114. 'mlpKerasDecay' 115. 'mlpKerasDecayCost' 116. 'mlpKerasDropout' 117. 'mlpKerasDropoutCost' 118. 'mlpML' 119. 'mlpSGD' 120. 'mlpWeightDecay' 121. 'mlpWeightDecayML' 122. 'monmlp' 123. 'msaenet' 124. 'multinom' 125. 'mxnet' 126. 'mxnetAdam' 127. 'naive_bayes' 128. 'nb' 129. 'nbDiscrete' 130. 'nbSearch' 131. 'neuralnet' 132. 'nnet' 133. 'nnls' 134. 'nodeHarvest' 135. 'null' 136. 'OneR' 137. 'ordinalNet' 138. 'ORFlog' 139. 'ORFpls' 140. 'ORFridge' 141. 'ORFsvm' 142. 'ownn' 143. 'pam' 144. 'parRF' 145. 'PART' 146. 'partDSA' 147. 'pcaNNet' 148. 'pcr' 149. 'pda' 150. 'pda2' 151. 'penalized' 152. 'PenalizedLDA' 153. 'plr' 154. 'pls' 155. 'plsRglm' 156. 'polr' 157. 'ppr' 158. 'PRIM' 159. 'protoclass' 160. 'pythonKnnReg' 161. 'qda' 162. 'QdaCov' 163. 'qrf' 164. 'qrnn' 165. 'randomGLM' 166. 'ranger' 167. 'rbf' 168. 'rbfDDA' 169. 'Rborist' 170. 'rda' 171. 'regLogistic' 172. 'relaxo' 173. 'rf' 174. 'rFerns' 175. 'RFlda' 176. 'rfRules' 177. 'ridge' 178. 'rlda' 179. 'rlm' 180. 'rmda' 181. 'rocc' 182. 'rotationForest' 183. 'rotationForestCp' 184. 'rpart' 185. 'rpart1SE' 186. 'rpart2' 187. 'rpartCost' 188. 'rpartScore' 189. 'rqlasso' 190. 'rqnc' 191. 'RRF' 192. 'RRFglobal' 193. 'rrlda' 194. 'RSimca' 195. 'rvmLinear' 196. 'rvmPoly' 197. 'rvmRadial' 198. 'SBC' 199. 'sda' 200. 'sdwd' 201. 'simpls' 202. 'SLAVE' 203. 'slda' 204. 'smda' 205. 'snn' 206. 'sparseLDA' 207. 'spikeslab' 208. 'spls' 209. 'stepLDA' 210. 'stepQDA' 211. 'superpc' 212. 'svmBoundrangeString' 213. 'svmExpoString' 214. 'svmLinear' 215. 'svmLinear2' 216. 'svmLinear3' 217. 'svmLinearWeights' 218. 'svmLinearWeights2' 219. 'svmPoly' 220. 'svmRadial' 221. 'svmRadialCost' 222. 'svmRadialSigma' 223. 'svmRadialWeights' 224. 'svmSpectrumString' 225. 'tan' 226. 'tanSearch' 227. 'treebag' 228. 'vbmpRadial' 229. 'vglmAdjCat' 230. 'vglmContRatio' 231. 'vglmCumulative' 232. 'widekernelpls' 233. 'WM' 234. 'wsrf' 235. 'xgbDART' 236. 'xgbLinear' 237. 'xgbTree' 238. 'xyf'

To get the info on specific model:

```
In [9]: getModelInfo()$glm$type
```

1. 'Regression' 2. 'Classification'

The below chunk of code is standarized way of building model using caret package. Setting in the control parameters for the model.

```
In [10]: set.seed(1234)
         objControl <- trainControl(method = "cv", number = 2, returnResamp = 'none',
                                    summaryFunction = twoClassSummary,
                                    #summaryFunction = twoClassSummary, defaultSummary
                                    classProbs = TRUE,
                                    savePredictions = TRUE)
```

The search grid is basically a model fine tuning option. The paramter inside the **expan.grid()** function varies according to model. The **complete** list of tuning paramter for different models.

```
In [11]: #This parameter is for glmnet. Need not be executed if method  is glmStepAIC
         #searchGrid <-  expand.grid(alpha = c(1:10)*0.1,
         #                           lambda = c(1:5)/10)
```

The model building starts here. > 1. **metric= "ROC"** uses ROC curve to select the best model.Accuracy, Kappa are other options. To use this change twoClassSummary to defaultSummary in **ObjControl** 2. **verbose = FALSE**: does not show the processing output on console

The factor names at times may not be consistent. R may expect **"Not.Joined"** but the actual level may be **"Not Joined"** This is corrected by using **make.names()** function to give syntactically valid names.

```
In [12]: #lg_train_df$StatusFactor <- as.factor(ifelse(lg_train_df$Status == "Joined", 1,0))
         set.seed(766)
         levels(lg_train_df$Status) <- make.names(levels(factor(lg_train_df$Status)))
         lg_caret_model <- train(lg_train_df[,1:12],
                                 lg_train_df[,13],
                                 method = 'glmStepAIC', #'glm', glmnet
                                 trControl = objControl,
                                 metric = "ROC",
                                 verbose = FALSE)
```

```
Start:  AIC=3281.64
.outcome ~ DOJ.Extended + Duration.to.accept.offer + Notice.period +
    Offered.band + Percent.difference.CTC + Joining.Bonus + Gender +
    Candidate.Source + Rex.in.Yrs + LOB + Location + Age

                           Df Deviance    AIC
- Location                 10   3231.4 3275.4
- Gender                    1   3217.7 3279.7
- Joining.Bonus             1   3218.2 3280.2
- Duration.to.accept.offer  1   3218.3 3280.3
<none>                          3217.6 3281.6
- DOJ.Extended              1   3221.8 3283.8
- Rex.in.Yrs                1   3223.2 3285.2
- LOB                       8   3238.5 3286.5
- Percent.difference.CTC    1   3226.2 3288.2
- Offered.band              3   3230.3 3288.3
- Age                       1   3230.0 3292.0
- Candidate.Source          2   3243.8 3303.8
```

```
- Notice.period             1   3313.3 3375.3

Step:  AIC=3275.43
.outcome ~ DOJ.Extended + Duration.to.accept.offer + Notice.period +
    Offered.band + Percent.difference.CTC + Joining.Bonus + Gender +
    Candidate.Source + Rex.in.Yrs + LOB + Age

                           Df Deviance    AIC
- Gender                    1   3231.5 3273.5
- Duration.to.accept.offer  1   3231.8 3273.8
- Joining.Bonus             1   3232.0 3274.0
<none>                          3231.4 3275.4
- DOJ.Extended              1   3236.2 3278.2
- Rex.in.Yrs                1   3236.8 3278.8
- Offered.band              3   3243.5 3281.5
- Percent.difference.CTC    1   3239.6 3281.6
- Age                       1   3243.8 3285.8
- LOB                       8   3261.9 3289.9
- Candidate.Source          2   3258.3 3298.3
- Notice.period             1   3322.9 3364.9

Step:  AIC=3273.48
.outcome ~ DOJ.Extended + Duration.to.accept.offer + Notice.period +
    Offered.band + Percent.difference.CTC + Joining.Bonus + Candidate.Source +
    Rex.in.Yrs + LOB + Age

                           Df Deviance    AIC
- Duration.to.accept.offer  1   3231.9 3271.9
- Joining.Bonus             1   3232.0 3272.0
<none>                          3231.5 3273.5
- DOJ.Extended              1   3236.3 3276.3
- Rex.in.Yrs                1   3236.8 3276.8
- Offered.band              3   3243.5 3279.5
- Percent.difference.CTC    1   3239.7 3279.7
- Age                       1   3243.8 3283.8
- LOB                       8   3261.9 3287.9
- Candidate.Source          2   3258.4 3296.4
- Notice.period             1   3323.0 3363.0

Step:  AIC=3271.89
.outcome ~ DOJ.Extended + Notice.period + Offered.band + Percent.difference.CTC +
    Joining.Bonus + Candidate.Source + Rex.in.Yrs + LOB + Age

                         Df Deviance    AIC
- Joining.Bonus           1   3232.4 3270.4
<none>                        3231.9 3271.9
- Rex.in.Yrs              1   3237.2 3275.2
- DOJ.Extended            1   3238.3 3276.3
```

```
- Offered.band            3   3244.0 3278.0
- Percent.difference.CTC  1   3240.1 3278.1
- Age                     1   3244.1 3282.1
- LOB                     8   3262.0 3286.0
- Candidate.Source        2   3258.7 3294.7
- Notice.period           1   3328.7 3366.7

Step:  AIC=3270.44
.outcome ~ DOJ.Extended + Notice.period + Offered.band + Percent.difference.CTC +
    Candidate.Source + Rex.in.Yrs + LOB + Age

                         Df Deviance    AIC
<none>                        3232.4 3270.4
- Rex.in.Yrs              1   3237.8 3273.8
- DOJ.Extended            1   3239.0 3275.0
- Offered.band            3   3244.4 3276.4
- Percent.difference.CTC  1   3240.6 3276.6
- Age                     1   3245.0 3281.0
- LOB                     8   3262.1 3284.1
- Candidate.Source        2   3259.6 3293.6
- Notice.period           1   3329.7 3365.7
Start:  AIC=3273.14
.outcome ~ DOJ.Extended + Duration.to.accept.offer + Notice.period +
    Offered.band + Percent.difference.CTC + Joining.Bonus + Gender +
    Candidate.Source + Rex.in.Yrs + LOB + Location + Age

                           Df Deviance    AIC
- Location                 10   3223.6 3267.6
- Joining.Bonus             1   3209.4 3271.4
- Percent.difference.CTC    1   3209.5 3271.5
- Duration.to.accept.offer  1   3209.8 3271.8
- Gender                    1   3209.8 3271.8
- Rex.in.Yrs                1   3210.1 3272.1
<none>                          3209.1 3273.1
- DOJ.Extended              1   3211.5 3273.5
- Age                       1   3216.3 3278.3
- LOB                       8   3234.0 3282.0
- Offered.band              3   3229.1 3287.1
- Candidate.Source          2   3237.2 3297.2
- Notice.period             1   3315.5 3377.5

Step:  AIC=3267.61
.outcome ~ DOJ.Extended + Duration.to.accept.offer + Notice.period +
    Offered.band + Percent.difference.CTC + Joining.Bonus + Gender +
    Candidate.Source + Rex.in.Yrs + LOB + Age

                           Df Deviance    AIC
- Joining.Bonus             1   3223.9 3265.9
```

```
- Duration.to.accept.offer  1   3223.9 3265.9
- Percent.difference.CTC    1   3223.9 3265.9
- Gender                    1   3224.5 3266.5
- Rex.in.Yrs                1   3224.8 3266.8
<none>                          3223.6 3267.6
- DOJ.Extended              1   3226.3 3268.3
- Age                       1   3230.9 3272.9
- Offered.band              3   3243.4 3281.4
- LOB                       8   3257.3 3285.3
- Candidate.Source          2   3254.4 3294.4
- Notice.period             1   3325.6 3367.6

Step:  AIC=3265.86
.outcome ~ DOJ.Extended + Duration.to.accept.offer + Notice.period +
    Offered.band + Percent.difference.CTC + Gender + Candidate.Source +
    Rex.in.Yrs + LOB + Age

                           Df Deviance    AIC
- Duration.to.accept.offer  1   3224.1 3264.1
- Percent.difference.CTC    1   3224.2 3264.2
- Gender                    1   3224.7 3264.7
- Rex.in.Yrs                1   3225.1 3265.1
<none>                          3223.9 3265.9
- DOJ.Extended              1   3226.5 3266.5
- Age                       1   3231.0 3271.0
- Offered.band              3   3243.8 3279.8
- LOB                       8   3258.3 3284.3
- Candidate.Source          2   3254.6 3292.6
- Notice.period             1   3325.7 3365.7

Step:  AIC=3264.14
.outcome ~ DOJ.Extended + Notice.period + Offered.band + Percent.difference.CTC +
    Gender + Candidate.Source + Rex.in.Yrs + LOB + Age

                         Df Deviance    AIC
- Percent.difference.CTC  1   3224.4 3262.4
- Gender                  1   3224.9 3262.9
- Rex.in.Yrs              1   3225.3 3263.3
<none>                        3224.1 3264.1
- DOJ.Extended            1   3227.7 3265.7
- Age                     1   3231.2 3269.2
- Offered.band            3   3244.0 3278.0
- LOB                     8   3258.4 3282.4
- Candidate.Source        2   3255.0 3291.0
- Notice.period           1   3333.1 3371.1

Step:  AIC=3262.44
.outcome ~ DOJ.Extended + Notice.period + Offered.band + Gender +
```

```
    Candidate.Source + Rex.in.Yrs + LOB + Age

                   Df Deviance    AIC
- Gender            1   3225.3 3261.3
- Rex.in.Yrs        1   3225.6 3261.6
<none>                  3224.4 3262.4
- DOJ.Extended      1   3228.0 3264.0
- Age               1   3231.5 3267.5
- Offered.band      3   3244.3 3276.3
- LOB               8   3259.0 3281.0
- Candidate.Source  2   3255.4 3289.4
- Notice.period     1   3334.1 3370.1

Step:  AIC=3261.26
.outcome ~ DOJ.Extended + Notice.period + Offered.band + Candidate.Source +
    Rex.in.Yrs + LOB + Age

                   Df Deviance    AIC
- Rex.in.Yrs        1   3226.5 3260.5
<none>                  3225.3 3261.3
- DOJ.Extended      1   3228.9 3262.9
- Age               1   3232.2 3266.2
- Offered.band      3   3244.6 3274.6
- LOB               8   3259.8 3279.8
- Candidate.Source  2   3257.0 3289.0
- Notice.period     1   3335.2 3369.2

Step:  AIC=3260.49
.outcome ~ DOJ.Extended + Notice.period + Offered.band + Candidate.Source +
    LOB + Age

                   Df Deviance    AIC
<none>                  3226.5 3260.5
- DOJ.Extended      1   3230.1 3262.1
- Age               1   3232.2 3264.2
- Offered.band      3   3246.6 3274.6
- Candidate.Source  2   3257.6 3287.6
- LOB               8   3271.4 3289.4
- Notice.period     1   3338.8 3370.8
Start:  AIC=6503.95
.outcome ~ DOJ.Extended + Duration.to.accept.offer + Notice.period +
    Offered.band + Percent.difference.CTC + Joining.Bonus + Gender +
    Candidate.Source + Rex.in.Yrs + LOB + Location + Age

                           Df Deviance    AIC
- Joining.Bonus             1   6440.0 6502.0
- Gender                    1   6440.5 6502.5
- Duration.to.accept.offer  1   6441.3 6503.3
```

```
<none>                           6440.0 6504.0
- Location                  10   6462.9 6506.9
- Rex.in.Yrs                 1   6445.3 6507.3
- Percent.difference.CTC     1   6445.9 6507.9
- DOJ.Extended               1   6446.5 6508.5
- Age                        1   6459.2 6521.2
- Offered.band               3   6471.9 6529.9
- LOB                        8   6484.0 6532.0
- Candidate.Source           2   6493.7 6553.7
- Notice.period              1   6641.5 6703.5

Step:  AIC=6501.97
.outcome ~ DOJ.Extended + Duration.to.accept.offer + Notice.period +
    Offered.band + Percent.difference.CTC + Gender + Candidate.Source +
    Rex.in.Yrs + LOB + Location + Age

                           Df Deviance    AIC
- Gender                    1   6440.6 6500.6
- Duration.to.accept.offer  1   6441.3 6501.3
<none>                          6440.0 6502.0
- Location                 10   6463.0 6505.0
- Rex.in.Yrs                1   6445.3 6505.3
- Percent.difference.CTC    1   6445.9 6505.9
- DOJ.Extended              1   6446.6 6506.6
- Age                       1   6459.3 6519.3
- Offered.band              3   6472.0 6528.0
- LOB                       8   6484.2 6530.2
- Candidate.Source          2   6493.7 6551.7
- Notice.period             1   6641.7 6701.7

Step:  AIC=6500.55
.outcome ~ DOJ.Extended + Duration.to.accept.offer + Notice.period +
    Offered.band + Percent.difference.CTC + Candidate.Source +
    Rex.in.Yrs + LOB + Location + Age

                           Df Deviance    AIC
- Duration.to.accept.offer  1   6441.9 6499.9
<none>                          6440.6 6500.6
- Location                 10   6463.6 6503.6
- Rex.in.Yrs                1   6446.0 6504.0
- Percent.difference.CTC    1   6446.5 6504.5
- DOJ.Extended              1   6447.2 6505.2
- Age                       1   6459.7 6517.7
- Offered.band              3   6472.0 6526.0
- LOB                       8   6484.8 6528.8
- Candidate.Source          2   6494.9 6550.9
- Notice.period             1   6642.5 6700.5
```

```
Step:  AIC=6499.91
.outcome ~ DOJ.Extended + Notice.period + Offered.band + Percent.difference.CTC +
    Candidate.Source + Rex.in.Yrs + LOB + Location + Age

                         Df Deviance    AIC
<none>                        6441.9 6499.9
- Location               10   6464.3 6502.3
- Rex.in.Yrs              1   6447.2 6503.2
- Percent.difference.CTC  1   6447.9 6503.9
- DOJ.Extended            1   6451.7 6507.7
- Age                     1   6460.9 6516.9
- Offered.band            3   6473.2 6525.2
- LOB                     8   6485.7 6527.7
- Candidate.Source        2   6496.3 6550.3
- Notice.period           1   6654.1 6710.1
```

1.3 Model Evaluation

1.3.1 1. One useful plot from caret package is the variable importance plot

In case you get an error "Invalid Graphic state", uncomment the line below

```
In [13]: lg_caret_model
         summary(lg_caret_model$finalModel)

         #dev.off()
         #plot(varImp(lg_caret_model, scale = TRUE))
```

```
Generalized Linear Model with Stepwise Feature Selection

7197 samples
  12 predictor
   2 classes: 'Joined', 'Not.Joined'

No pre-processing
Resampling: Cross-Validated (2 fold)
Summary of sample sizes: 3598, 3599
Resampling results:

  ROC        Sens       Spec
  0.6780952  0.9929929  0.03789004




Call:
NULL
```

```
Deviance Residuals:
    Min       1Q   Median       3Q      Max
-1.3563  -0.6804  -0.5317  -0.3576   2.7421

Coefficients:
                                    Estimate Std. Error z value Pr(>|z|)
(Intercept)                         2.259535   1.289022   1.753  0.07962 .
DOJ.ExtendedYes                    -0.207165   0.066462  -3.117  0.00183 **
Notice.period                       0.020616   0.001419  14.525  < 2e-16 ***
Offered.bandE1                     -1.192264   0.206380  -5.777 7.60e-09 ***
Offered.bandE2                     -1.065211   0.226899  -4.695 2.67e-06 ***
Offered.bandE3                     -1.223899   0.295230  -4.146 3.39e-05 ***
Percent.difference.CTC             -0.004578   0.001957  -2.339  0.01932 *
Candidate.SourceDirect             -0.368604   0.072463  -5.087 3.64e-07 ***
Candidate.SourceEmployee Referral  -0.736148   0.107263  -6.863 6.74e-12 ***
Rex.in.Yrs                          0.050985   0.021981   2.320  0.02036 *
LOBBFSI                            -0.173311   0.150020  -1.155  0.24799
LOBCSMP                            -0.122352   0.172973  -0.707  0.47935
LOBEAS                              0.227690   0.188959   1.205  0.22821
LOBERS                             -0.224907   0.141978  -1.584  0.11317
LOBETS                             -0.346498   0.170092  -2.037  0.04164 *
LOBHealthcare                      -0.050359   0.281765  -0.179  0.85815
LOBINFRA                           -0.661916   0.154049  -4.297 1.73e-05 ***
LOBMMS                            -13.541537 257.039396  -0.053  0.95798
LocationBangalore                  -1.600191   1.232193  -1.299  0.19406
LocationChennai                    -1.605304   1.231172  -1.304  0.19227
LocationCochin                    -13.966457 333.290302  -0.042  0.96657
LocationGurgaon                    -1.693583   1.255248  -1.349  0.17727
LocationHyderabad                  -1.726577   1.241172  -1.391  0.16420
LocationKolkata                    -1.959904   1.261936  -1.553  0.12040
LocationMumbai                     -1.874808   1.255072  -1.494  0.13523
LocationNoida                      -1.947670   1.230605  -1.583  0.11349
LocationOthers                    -14.051586 246.089350  -0.057  0.95447
LocationPune                       -1.664387   1.291366  -1.289  0.19745
Age                                -0.043467   0.010105  -4.302 1.70e-05 ***
---
Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1   1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 6936.1  on 7196  degrees of freedom
Residual deviance: 6441.9  on 7168  degrees of freedom
AIC: 6499.9

Number of Fisher Scoring iterations: 13
```

1.3.2 2. The prediction and confusion Matrix on train data.

The syntax for prediction in caret is almost similar expect the the **type** attribute expects input as **'raw'** or **'prob'**. In case of prob, the predicted value holds the probability of both positive and negative class.

```
In [14]: #Missing code. May result in error
         levels(lg_train_df$Status) <- make.names(levels(factor(lg_train_df$Status)))
         caretPredictedClass <- predict(object = lg_caret_model, lg_train_df[,1:12], type = 'ra
         confusionMatrix(caretPredictedClass,lg_train_df$Status)
```

```
Confusion Matrix and Statistics

            Reference
Prediction   Joined Not.Joined
  Joined       5807       1294
  Not.Joined     44         52

               Accuracy : 0.8141
                 95% CI : (0.8049, 0.823)
    No Information Rate : 0.813
    P-Value [Acc > NIR] : 0.4115

                  Kappa : 0.0484
 Mcnemar's Test P-Value : <2e-16

            Sensitivity : 0.99248
            Specificity : 0.03863
         Pos Pred Value : 0.81777
         Neg Pred Value : 0.54167
             Prevalence : 0.81298
         Detection Rate : 0.80686
   Detection Prevalence : 0.98666
      Balanced Accuracy : 0.51556

       'Positive' Class : Joined
```

1.3.3 3. The optimal cut-off

Creating empty vectors to store the results.

```
In [15]: msclaf_cost <- c()
         youden_index <- c()
         cutoff <- c()
         P11 <- c() #correct classification of positive as positive
         P00 <- c() #correct classification of negative as negative
         P10 <- c() #misclassification of positive class to negative class
         P01 <- c() #misclassification of negative class to positive class
```

Select the optimal cut-off value, if:

1. cost of misclassifying Not Joined as Joined is twice as costly as cost of micalssifying Joined as Not Joined
2. both sensitivity and specificity are equally important

The best cut-off is the one which minimizes the misclassification cost (in case of ***option 1***) or which maximizes the Youden's Index (in case of ***Option 2***).
fix the bug here: clue is in the above **two options**

```
In [16]: train_predicted_prob = predict(object = lg_caret_model, lg_train_df[,1:12], type = 'p
         #variable with all the values as joined
         n <- length(lg_train_df$Status)

         costs = matrix(c(0,2,1, 0), ncol = 2)
         colnames(costs) = rownames(costs) = c("Joined", "Non Joined")
         as.table(costs)
```

```
           Joined Non Joined
Joined          0          1
Non Joined      2          0
```

The misclassification cost table is:

```
In [17]: # defining log odds in favor of Joined
         for (i in seq(0.05, 1, .05)) {
           predicted_y = rep("Not Joined", n)
           predicted_y[train_predicted_prob[1] > i] = "Joined"
           tbl <- table(lg_train_df$Status, predicted_y)
           if ( i <= 1) {
             #Classifying Not Joined as Joined
             P10[20*i] <- tbl[2]/(tbl[2] + tbl[4])

             P11[20*i] <- tbl[4]/(tbl[2] + tbl[4])

             #Classifying Joined as Not Joined
             P01[20*i] <- tbl[3]/(tbl[1] + tbl[3])

             P00[20*i] <- tbl[1]/(tbl[1] + tbl[3])

             cutoff[20*i] <- i
             msclaf_cost[20*i] <- P10[20*i]*costs[2] + P01[20*i]*costs[3]
             youden_index[20*i] <- P11[20*i] + P00[20*i] - 1
           }
         }
         df_cost_table <- cbind(cutoff,P10,P01,msclaf_cost, P11, P00, youden_index)
```

The table summarizing the optimal cut-off value:
write the cost.table into a csv file

```
In [18]: df_cost_table
         #write.csv(df_cost_table, "Optimal_Cutoff_caret.csv")
```

cutoff	P10	P01	msclaf_cost	P11	P00	youden_index
0.05	NA	NA	NA	NA	NA	NA
0.10	NA	NA	NA	NA	NA	NA
0.15	NA	NA	NA	NA	NA	NA
0.20	NA	NA	NA	NA	NA	NA
0.25	NA	NA	NA	NA	NA	NA
0.30	NA	NA	NA	NA	NA	NA
0.35	NA	NA	NA	NA	NA	NA
0.40	1.00000000	0.0005127329	2.0005127	0.00000000	0.99948727	-0.0005127329
0.45	0.98662704	0.0034182191	1.9766723	0.01337296	0.99658178	0.0099547378
0.50	0.96136701	0.0075200820	1.9302541	0.03863299	0.99247992	0.0311129046
0.55	0.92793462	0.0153819860	1.8712512	0.07206538	0.98461801	0.0566833929
0.60	0.88484398	0.0263202871	1.7960083	0.11515602	0.97367971	0.0888357307
0.65	0.82540862	0.0512732866	1.7020905	0.17459138	0.94872671	0.1233180953
0.70	0.72362556	0.0974192446	1.5446704	0.27637444	0.90258076	0.1789551982
0.75	0.57726597	0.1791146813	1.3336466	0.42273403	0.82088532	0.2436193455
0.80	0.39895988	0.3331054521	1.1310252	0.60104012	0.66689455	0.2679346668
0.85	0.22659733	0.5344385575	0.9876332	0.77340267	0.46556144	0.2389641171
0.90	0.08692422	0.7602119296	0.9340604	0.91307578	0.23978807	0.1528638505
0.95	0.01040119	0.9586395488	0.9794419	0.98959881	0.04136045	0.0309592625
1.00	NA	NA	NA	NA	NA	NA

1.3.4 4. Confusion Matrix on the test data

The **predict** function is used to get the predicted probability on the new dataset. The probability value along with the optimal cut-off can be used to build confusion matrix

```
In [19]: test_predicted_prob = predict(lg_caret_model, lg_test_df, type = "prob")

         #variable with all the values as joined
         n <- length(lg_test_df$Status)
         predicted_y = rep("Not Joined", n)

         # defining log odds in favor of not joining
         predicted_y[test_predicted_prob[1] > 0.80] = "Joined"

         #add the model_precition in the data
         lg_test_df$predicted_y <- predicted_y

         ###Create the confusionmatrix###
         addmargins(table(lg_test_df$Status, lg_test_df$predicted_y))
         mean(lg_test_df$predicted_y == lg_test_df$Status)
```

	Joined	Not Joined	Sum
Joined	953	509	1462
Not Joined	123	213	336
Sum	1076	722	1798

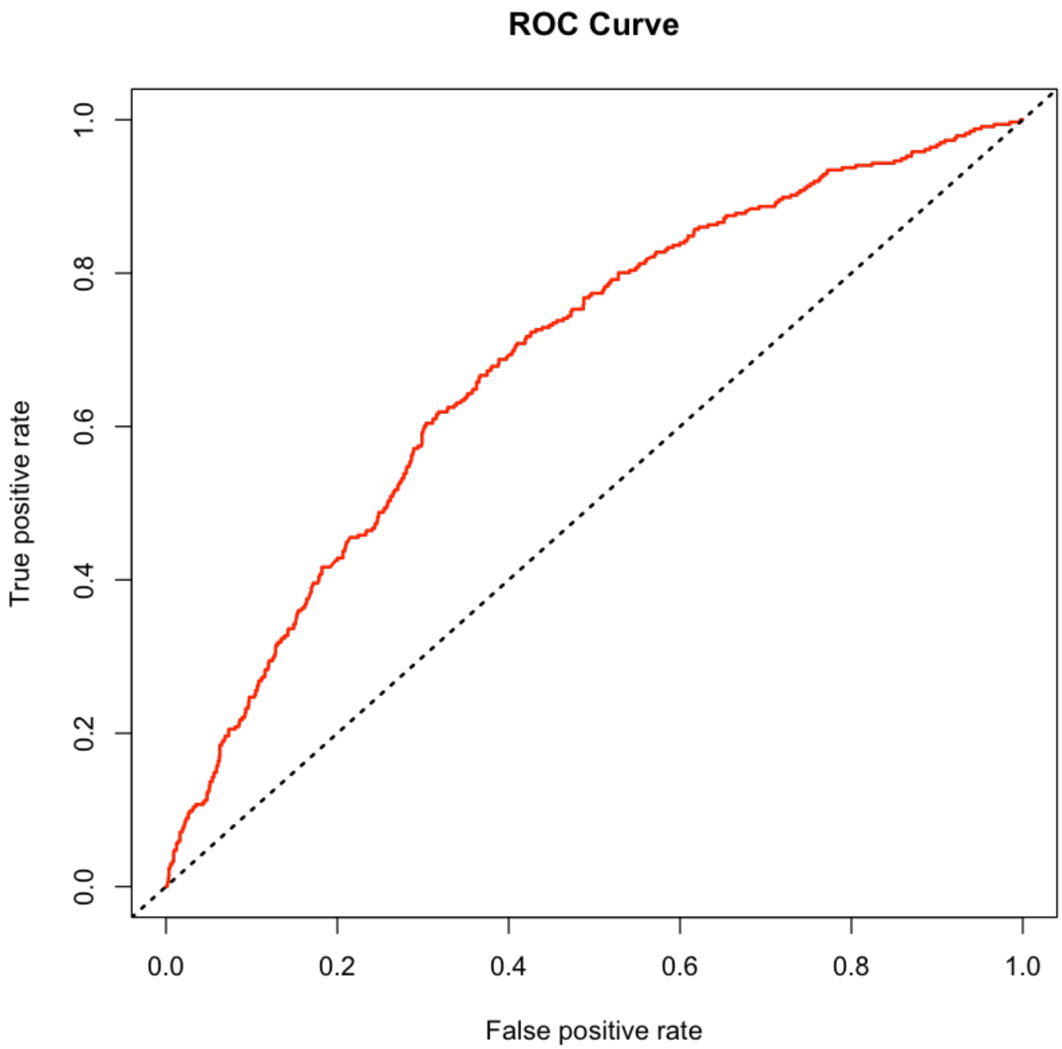
0.648498331479422

1.3.5 5. ROC Plot on the test data

ROCR package can be used to evaluate the model performace on the test data. The same package can also be used to get the model performace on the test data.

```
In [20]: #error in below line
         lgPredObj <- prediction(test_predicted_prob[2],lg_test_df$Status)
         lgPerfObj <- performance(lgPredObj, "tpr","fpr")
         plot(lgPerfObj,main = "ROC Curve",col = 2,lwd = 2)
         abline(a = 0,b = 1,lwd = 2,lty = 3,col = "black")
         performance(lgPredObj, "auc")
```

```
An object of class "performance"
Slot "x.name":
[1] "None"

Slot "y.name":
[1] "Area under the ROC curve"

Slot "alpha.name":
[1] "none"

Slot "x.values":
list()

Slot "y.values":
[[1]]
[1] 0.6877728


Slot "alpha.values":
list()
```



End of Document