ASA_Assignment_2 Shreyas Namjoshi 23/03/2022 R Markdown Loading the data and dividing it in Train and Test data<-read_excel("FinalDataSet_ASA.xlsx")</pre> ## 70% of the data for training train_size <- floor(0.70 * nrow(data))</pre> ## set the seed to make your partition reproducible set.seed(213) train_ind <- sample(seq_len(nrow(data)),</pre> size = train_size) train <- data[train_ind,]</pre> test <- data[-train_ind,]</pre> Q1. Build a logistic regression equation to predict whether the person is likely to accept the bank's offer for a personal loan. If necessary, create new variables to improve the model performance. Default.Model <- glm(train\$`Personal Loan` ~ ., data=train, family = binomial)</pre> summary(Default.Model) ## ## glm(formula = train\$`Personal Loan` ~ ., family = binomial, data = train) ## Deviance Residuals: Min 1Q Median 3Q Max ## -2.59426 -0.15757 -0.04659 -0.01468 2.89222 ## `Securities Account` 5.799e-02 8.125e-01 0.071 0.94310 ## `CD Account` 5.086e+00 9.696e-01 5.245 1.56e-07 *** ## Online -1.452e-01 5.454e-01 -0.266 0.79003 ## CreditCard -1.660e+00 6.751e-01 -2.458 0.01396 * ## ---## Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1 ## (Dispersion parameter for binomial family taken to be 1) Null deviance: 395.14 on 699 degrees of freedom ## Residual deviance: 137.04 on 687 degrees of freedom ## AIC: 163.04 ## Number of Fisher Scoring iterations: 8 Q2. Carry out the omnibus test to test whether the model as a whole is significant. Comment on the result of the omnibus test. library(lmtest) ## Loading required package: zoo ## Attaching package: 'zoo' ## The following objects are masked from 'package:base': ## as.Date, as.Date.numeric lrtest(Default.Model) ## Likelihood ratio test ## Model 1: train\$`Personal Loan` ~ Age + Experience + Income + `ZIP Code` + Family + CCAvg + Education + Mortgage + `Securities Account` + `CD Account` + Online + CreditCard ## Model 2: train\$`Personal Loan` ~ 1 ## #Df LogLik Df Chisq Pr(>Chisq) ## 1 13 -68.522 ## 2 1 -197.571 -12 258.1 < 2.2e-16 *** ## Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1 Here We reject the Model 2 as we have sufficient evidence that model 2(model with only intercept) is not good (assumed value of alpha is 0.05). This means that atleast one of the coeff is significant. Q3. Test the hypothesis that β = 0 for all β , where β indicates the coefficient corresponding to jth explanatory variable. Comment on the result of these hypothesis tests. summary(Default.Model) ## ## Call: ## glm(formula = train\$`Personal Loan` ~ ., family = binomial, data = train) ## Deviance Residuals: Min 1Q Median 3Q Max ## -2.59426 -0.15757 -0.04659 -0.01468 2.89222 ## ## Coefficients: ## `Securities Account` 5.799e-02 8.125e-01 0.071 0.94310 ## `CD Account` 5.086e+00 9.696e-01 5.245 1.56e-07 *** ## Online -1.452e-01 5.454e-01 -0.266 0.79003 ## CreditCard -1.660e+00 6.751e-01 -2.458 0.01396 * ## ---## Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1 ## (Dispersion parameter for binomial family taken to be 1) ## Null deviance: 395.14 on 699 degrees of freedom ## Residual deviance: 137.04 on 687 degrees of freedom ## AIC: 163.04 ## Number of Fisher Scoring iterations: 8 Comment: In the above table, P value represents the Walds Test P Value. Thus only significant variables contrib uting to the model are Income, Family, CCAvg, Education, CD Account, Credit Card. Rest of the coefficients are non si gnificant. Q4. Carry out the hypothesis test that the model fits the data. Comment on the results. library(ResourceSelection) ## Warning: package 'ResourceSelection' was built under R version 4.1.3 ## ResourceSelection 0.3-5 2019-07-22 hoslem.test(Default.Model\$y, fitted(Default.Model)) ## Hosmer and Lemeshow goodness of fit (GOF) test ## ## data: Default.Model\$y, fitted(Default.Model) ## X-squared = 1.345, df = 8, p-value = 0.995 From the above P value, we do not have sufficient evidence to reject Null hypothesis. This implies that model f its the data well. Q5. The bank would like to address the top 30 persons with an offer for personal loan based on the probability (propensity). Create a table displaying all the details of the "top" 30 persons who are most likely to accept the bank's offer. Make sure to include the probability of accepting the offer along with all the other details. Here we are including only test set to find the top 30 person who are most likely to accept bank offer based o n probability pred<-as.data.frame(predict(Default.Model, test[, -9], type = "response"))</pre> pred ## predict(Default.Model, test[, -9], type = "response") ## 1 8.418288e-04 ## 2 9.956183e-01 ## 3 1.609878e-01 ## 4 2.644283e-03 ## 5 1.508703e-02 ## 6 4.307246e-04 ## 7 1.418949e-03 ## 8 6.176101e-04 ## 9 6.516075e-02 ## 10 7.047398e-02 ## 11 1.280599e-03 ## 12 1.221742e-03 ## 13 2.809465e-02 ## 14 1.727570e-04 ## 15 6.176452e-03 ## 16 1.979536e-05 ## 17 2.081221e-04 ## 18 1.724965e-02 ## 19 7.774295e-03 ## 20 2.030980e-05 ## 21 6.434005e-03 ## 22 8.839954e-01 ## 23 7.218382e-04 ## 24 6.911800e-04 ## 25 6.010217e-04 ## 26 2.810038e-02 ## 27 1.796101e-04 ## 28 1.874957e-05 ## 29 2.038808e-04 ## 30 2.115832e-04 ## 31 4.931409e-02 ## 32 6.391681e-02 ## 33 1.689934e-04 ## 34 2.349675e-01 ## 35 1.160709e-02 ## 36 9.990597e-01 ## 37 1.198259e-01 ## 38 1.241903e-02 ## 39 1.014960e-02 ## 40 1.734388e-03 ## 41 1.099951e-02 ## 42 1.020230e-03 ## 43 9.414143e-01 ## 44 9.994166e-01 ## 45 1.531989e-01 1.981792e-02 ## 46 ## 47 2.182466e-03 ## 48 3.290679e-04 ## 49 1.878818e-04 ## 50 4.119771e-03 ## 51 3.185136e-04 ## 52 9.920441e-01 ## 53 8.090021e-02 7.836039e-04 ## 54 ## 55 8.079000e-02 ## 56 4.511085e-03 ## 57 1.206786e-03 ## 58 1.310382e-04 ## 59 9.994056e-01 ## 60 6.333927e-04 ## 61 1.810752e-06 ## 62 1.384147e-02 ## 63 4.366570e-03 ## 64 2.981498e-02 ## 65 1.473681e-01 ## 66 6.082317e-03 ## 67 2.404840e-02 ## 68 1.521723e-05 ## 69 4.157563e-04 ## 70 6.944990e-01 ## 71 2.377572e-04 2.810703e-01 ## 72 ## 73 1.668855e-04 ## 74 1.424295e-02 ## 75 3.785533e-03 ## 76 1.941953e-02 ## 77 1.010398e-01 ## 78 9.193497e-01 ## 79 8.551990e-05 ## 80 6.297426e-04 ## 81 5.910134e-03 ## 82 9.167394e-06 ## 83 5.046492e-05 ## 84 3.463007e-04 9.260120e-02 ## 85 ## 86 1.679016e-03 ## 87 3.399187e-04 1.075429e-04 ## 88 ## 89 1.626299e-03 ## 90 2.075343e-02 6.724374e-03 ## 91 ## 92 3.307499e-01 ## 93 7.058023e-03 ## 94 5.065513e-02 ## 95 1.863478e-03 ## 96 7.680659e-03 ## 97 2.593477e-01 ## 98 7.659541e-04 ## 99 1.185320e-01 ## 100 1.701242e-05 ## 101 9.957945e-01 ## 102 8.525656e-05 ## 103 1.886401e-03 ## 104 1.312633e-04 ## 105 7.894698e-05 ## 106 6.400107e-02 ## 107 4.969747e-02 ## 108 9.979879e-04 ## 109 5.435707e-03 ## 110 6.807250e-04 ## 111 8.906576e-05 ## 112 1.024571e-01 ## 113 6.512523e-04 ## 114 9.212519e-02 ## 115 4.432837e-03 ## 116 6.333927e-04 ## 117 1.875136e-03 ## 118 7.722858e-01 ## 119 9.984689e-01 ## 120 1.053345e-03 ## 121 8.985052e-07 ## 122 8.878795e-03 ## 123 1.243645e-04 ## 124 1.400756e-01 8.155254e-04 ## 125 ## 126 2.809666e-04 ## 127 1.397535e-05 ## 128 1.995949e-04 ## 129 6.179207e-03 ## 130 1.747876e-03 ## 131 2.083369e-06 ## 132 2.262351e-01 ## 133 1.131766e-03 ## 134 1.299417e-02 2.201776e-02 ## 135 ## 136 7.576123e-04 ## 137 1.251158e-04 ## 138 3.202975e-01 ## 139 8.878795e-03 ## 140 4.187920e-03 8.092606e-05 ## 141 ## 142 2.964706e-04 ## 143 3.247234e-03 ## 144 1.247799e-02 ## 145 1.110786e-01 ## 146 3.246906e-03 ## 147 1.236736e-01 ## 148 3.007968e-04 ## 149 4.357979e-02 ## 150 1.817360e-01 ## 151 7.540736e-03 ## 152 3.426520e-04 ## 153 1.434662e-05 ## 154 7.521551e-05 ## 155 8.167043e-03 ## 156 5.100362e-03 ## 157 1.628382e-06 ## 158 2.332584e-03 ## 159 5.131524e-02 ## 160 1.783679e-03 ## 161 3.309950e-03 ## 162 7.069907e-05 ## 163 3.037958e-06 ## 164 9.887576e-01 ## 165 6.786549e-03 ## 166 1.075930e-02 ## 167 4.026487e-02 ## 168 5.977844e-04 ## 169 7.935928e-02 ## 170 1.676245e-02 ## 171 8.155254e-04 ## 172 5.755503e-04 ## 173 1.626299e-03 ## 174 7.099767e-04 ## 175 1.020230e-03 ## 176 7.143871e-01 ## 177 3.784165e-04 ## 178 3.938185e-03 ## 179 1.146013e-03 ## 180 4.961856e-04 ## 181 7.097066e-04 ## 182 3.291051e-02 ## 183 2.151720e-03 ## 184 2.225574e-02 ## 185 3.800647e-05 ## 186 3.507943e-01 ## 187 1.178220e-05 ## 188 2.252488e-02 ## 189 2.133239e-03 ## 190 3.530077e-04 ## 191 3.728312e-03 ## 192 2.085172e-01 ## 193 3.242152e-02 ## 194 1.420799e-03 ## 195 2.324767e-03 ## 196 7.875045e-03 ## 197 6.076913e-02 ## 198 1.703572e-03 ## 199 1.221098e-03 ## 200 7.707230e-03 ## 201 2.586517e-02 ## 202 1.241095e-02 ## 203 2.634041e-03 ## 204 3.232225e-03 ## 205 1.088847e-02 ## 206 4.903447e-06 ## 207 3.048216e-05 ## 208 5.423379e-03 ## 209 2.718748e-04 ## 210 4.186268e-02 ## 211 1.556368e-03 ## 212 2.229575e-03 ## 213 3.744882e-03 ## 214 8.193708e-04 ## 215 1.587839e-05 ## 216 1.978235e-04 ## 217 1.487868e-03 ## 218 9.082568e-02 ## 219 1.236736e-01 ## 220 1.986081e-02 ## 221 1.852487e-03 ## 222 1.594309e-02 ## 223 9.995467e-01 ## 224 1.089372e-01 ## 225 1.757160e-03 ## 226 1.520498e-02 ## 227 9.994671e-01 ## 228 3.547396e-03 ## 229 1.300550e-03 ## 230 7.330400e-03 ## 231 9.973563e-03 ## 232 1.459615e-04 ## 233 5.407511e-04 ## 234 9.894620e-04 ## 235 1.077114e-03 ## 236 4.117900e-04 ## 237 5.002124e-01 ## 238 1.529561e-03 ## 239 1.093506e-03 ## 240 6.885982e-02 ## 241 9.680831e-04 ## 242 1.052941e-04 ## 243 1.531989e-01 ## 244 1.791850e-02 ## 245 3.099173e-02 ## 246 3.322952e-04 ## 247 1.551698e-04 ## 248 5.401939e-01 ## 249 5.325785e-03 ## 250 3.082967e-02 ## 251 1.383643e-02 ## 252 6.694256e-05 ## 253 2.372403e-03 ## 254 2.516951e-02 ## 255 1.031277e-02 ## 256 1.419912e-02 ## 257 1.151150e-04 ## 258 5.568901e-03 ## 259 1.813606e-01 ## 260 9.973167e-01 ## 261 2.647909e-02 ## 262 1.545944e-02 ## 263 4.305358e-02 ## 264 1.891579e-02 ## 265 1.307254e-02 4.779950e-04 ## 266 ## 267 2.081903e-03 ## 268 6.966180e-05 ## 269 5.553065e-05 ## 270 6.197915e-01 ## 271 1.576769e-03 5.937167e-05 ## 272 ## 273 1.138956e-03 ## 274 2.173796e-03 ## 275 1.958756e-05 ## 276 9.995925e-01 8.058295e-06 ## 277 5.327426e-03 ## 278 ## 279 5.020133e-06 ## 280 4.177345e-01 ## 281 6.990968e-05 ## 282 7.429783e-04 ## 283 3.163868e-02 ## 284 2.261694e-04 ## 285 4.065897e-05 ## 286 1.698471e-01 ## 287 2.324907e-04 ## 288 4.117900e-04 ## 289 2.750433e-03 ## 290 1.599429e-03 ## 291 1.790936e-03 ## 292 1.595665e-03 ## 293 1.429076e-04 7.509834e-05 ## 294 ## 295 6.166997e-02 ## 296 3.052183e-02 ## 297 1.007460e-03 ## 298 1.318357e-04 ## 299 1.194185e-03 ## 300 7.281801e-02 test['Prob of Availing']<-pred</pre> test1<-test[order(test\$`Prob of Availing`,decreasing = TRUE),]</pre> testLG<-head(test1, n=30) testLG ## # A tibble: 30 x 14 Age Experience Income `ZIP Code` Family CCAvg Education Mortgage <dbl> <dbl> <dbl> <dbl> <dbl> <dbl> < <dbl> <dbl> ## 38 143 95039 2 6.4 0 64 3 ## 50 26 192 90245 2 1.8 301 2 3 ## 41 15 159 90057 1 5.5 3 ## 53 28 175 95060 4 3 3.6 3 15 202 ## 45 91380 3 10 3 0 5 19 174 92028 3 1.7 231 2 3.8 422 ## 57 31 164 94607 3 ## 8 56 32 173 94022 1 4.6 88 26 110 204 ## 9 52 94501 2 5.4 23 170 90254 ## 10 2 6.5 ... with 20 more rows, and 6 more variables: Personal Loan <dbl>, Securities Account <dbl>, CD Account <dbl>, Online <dbl>, CreditCard <dbl>, Prob of Availing <dbl> Q6. Compare the above list of 30 persons against the 30 persons obtained from Discriminant Analysis (Assignment 1). Comment on the similarities and dissimilarities lda.fit<-lda(data\$`Personal Loan`~.,data=data,subset = train_ind)</pre> lda.pred <- predict(lda.fit, test)</pre> probdf<-lda.pred\$posterior</pre> test['Prob of Availing LDA']<-probdf[,"1"]</pre> test['class']<-lda.pred\$class</pre> testLDA<-test[order(test\$`Prob of Availing LDA`, decreasing = TRUE),]</pre> testLDA30<-head(testLDA, n=30) summary(testLDA30) ZIP Code Age Experience Income ## Min. :27.00 Min. : 49.0 :90057 Min. : 0.00 Min. 1st Qu.:36.50 1st Qu.:12.25 1st Qu.:119.0 1st Qu.:91463 Median :46.50 Median :22.00 Median :153.5 Median :92352 Mean :143.9 Mean :45.73 Mean :20.40 Mean :92885 3rd Qu.:56.00 3rd Qu.:30.75 3rd Qu.:174.8 3rd Qu.:94452 ## :64.00 Max. :38.00 Max. :202.0 :95827 Max. Max. ## Family CCAvg Education Mortgage Min. :1.000 Min. : 0.300 Min. :1.000 Min. : 0.0 1st Qu.:1.000 1st Qu.: 1.925 1st Qu.:2.000 1st Qu.: 0.0 Median :2.000 Median : 3.550 Median :3.000 Median : 0.0 Mean :2.267 Mean : 3.960 Mean :2.367 Mean :115.3 3rd Qu.:3.000 3rd Qu.: 6.000 3rd Qu.:3.000 3rd Qu.:203.8 :4.000 ## Max. Max. :10.000 Max. :3.000 Max. :422.0 Personal Loan Securities Account CD Account **Online** :0.0000 Min. :0.0000 Min. Min. :0.0000 Min. :0.0000 1st Qu.:0.0000 1st Qu.:0.0000 1st Qu.:0.0000 1st Qu.:1.0000 Median :1.0000 Median :0.0000 Median :1.0000 Median :1.0000 Mean :0.6667 Mean :0.1667 Mean :0.5333 Mean :0.7667 3rd Qu.:1.0000 3rd Qu.:0.0000 3rd Qu.:1.0000 3rd Qu.:1.0000 ## :1.0000 Max. :1.0000 Max. :1.0000 :1.0000 Max. CreditCard Prob of Availing Prob of Availing LDA class ## :0.0 :0.007681 Min. :0.1178 0:10 1:20 1st Qu.:0.0 1st Qu.:0.335761 1st Qu.:0.2913 Median :0.5 Median :0.828141 Median :0.8766 :0.5 :0.662669 Mean :0.6847 3rd Qu.:1.0 3rd Qu.:0.9994 3rd Qu.:0.996936 Max. :1.0 Max. :0.999592 :0.9999 Max. summary(testLG) ZIP Code Age Experience Income :27.00 Min. : 0.00 Min. : 49.0 :90057 Min. 1st Qu.:41.00 1st Qu.:115.0 1st Qu.:15.00 1st Qu.:91521 Median :48.00 Median :23.00 Median :147.5 Median :92352 ## Mean :48.40 Mean :22.77 Mean :143.7 Mean :92788 3rd Qu.:57.75 3rd Qu.:31.75 3rd Qu.:173.8 3rd Qu.:94452 :65.00 :39.00 Max. :202.0 :95819 ## Family Education Personal Loan CCAvg Mortgage Min. :1.000 Min. : 0.300 Min. :1.000 Min. : 0.0 Min. :0.0 1st Qu.:1.000 1st Qu.: 1.725 1st Qu.:3.000 1st Qu.: 0.0 1st Qu.:0.0 ## Median :2.000 Median : 3.250 Median :3.000 Median : 0.0 Median :1.0 Mean :2.367 Mean : 3.860 Mean :2.667 Mean :106.4 Mean 3rd Qu.:3.750 3rd Qu.: 5.875 3rd Qu.:3.000 3rd Qu.:198.2 3rd Qu.:1.0 :4.000 :10.000 :3.000 :422.0 Max. Max. Max. :1.0 Securities Account CD Account Online CreditCard ## Min. :0.0000 :0.0000 Min. :0.0000 Min. :0.0000 1st Qu.:0.0000 ## 1st Qu.:0.0000 1st Qu.:1.0000 1st Qu.:0.0000 Median :0.0000 Median :0.0000 Median :0.0000 Median :1.0000 ## Mean :0.1667 :0.4333 :0.8333 Mean :0.4333 Mean Mean 3rd Qu.:0.0000 3rd Qu.:1.0000 3rd Qu.:1.0000 3rd Qu.:1.0000 :1.0000 :1.0000 Max. :1.0000 Max. :1.0000 Max. Prob of Availing ## Min. :0.2085 1st Qu.:0.3675 Median :0.8281 :0.7060 ## 3rd Qu.:0.9969 ## Max. :0.9996 print("Education Count") ## [1] "Education Count" table(testLG\$Education) ## 1 2 3 ## 3 4 23 table(testLDA30\$Education) ## 1 2 3 ## 7 5 18 print("CD Count") ## [1] "CD Count" table(testLG\$`CD Account`) ## 0 1 ## 17 13 table(testLDA30\$`CD Account`) ## 0 1 ## 14 16 print("Credit Card Count") ## [1] "Credit Card Count" table(testLG\$CreditCard) ## 0 1 ## 17 13 table(testLDA30\$CreditCard) ## 0 **1** ## 15 15 print("CM for LDA for test set") ## [1] "CM for LDA for test set" resultsLDA <- confusionMatrix(data=testLDA30\$class, reference=as.factor(testLDA30\$`Personal Loan`)) print(resultsLDA) ## Confusion Matrix and Statistics ## Reference ## Prediction 0 1 0 5 5 1 5 15 ## ## Accuracy : 0.6667 ## 95% CI : (0.4719, 0.8271) ## No Information Rate: 0.6667 P-Value [Acc > NIR] : 0.5848 ## ## ## Kappa : 0.25 ## Mcnemar's Test P-Value : 1.0000 ## ## ## Sensitivity: 0.5000 ## Specificity: 0.7500 ## Pos Pred Value : 0.5000 ## Neg Pred Value : 0.7500 Prevalence : 0.3333 ## Detection Rate: 0.1667 Detection Prevalence : 0.3333 ## Balanced Accuracy : 0.6250 ## ## 'Positive' Class : 0 ## print("CM for LR for test set") ## [1] "CM for LR for test set" predicted<-as.numeric(testLG\$`Prob of Availing`>0.5) resultsLG<-confusionMatrix(data = as.factor(predicted), reference = as.factor(testLG\^Personal Loan\^))</pre> print(resultsLG) ## Confusion Matrix and Statistics ## Reference ## Prediction 0 1 0 6 3

1 3 18

No Information Rate : 0.7

Mcnemar's Test P-Value : 1.0000

P-Value [Acc > NIR] : 0.1595

Accuracy : 0.8

Sensitivity: 0.6667 Specificity: 0.8571

Pos Pred Value : 0.6667 Neg Pred Value : 0.8571 Prevalence : 0.3000

Detection Rate : 0.2000

Detection Prevalence : 0.3000

'Positive' Class : 0

Balanced Accuracy : 0.7619

Kappa : 0.5238

95% CI : (0.6143, 0.9229)

1. we see similarities in both the output's with respect to age,income,family, and CCAvg,securities Account

5. We also see improved level of accuracy, specificity and sensitivity for Logistic Regression as compared to

2. We can see that there are more people with education as UG and Graduate in LDA than in Logistic.

4. We also see that number of people who use credit are more in LDA than in Logistic Regression

3. We see that in LDA, more number of people are there who have Certificate of Deposit

##

##

##

##

##

##

##

##

##

##

##

##

LDA for test data.