IDENTIFY GOOD CUSTOMER PROJECT

Tram Ngo

May 29,2022

Understanding the data

```
> summary(banka$age)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
  18.00   32.00   38.00   40.11   47.00   88.00
```

The variable “age” has mean of 40.11 and median of 38.00. In this case, since the mean is bigger than the median the data is skewed to the right.

```
> summary(banka$duration)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
    0.0   103.0   181.0   256.8   317.0  3643.0
```

The variable “duration” has mean of 256.8 and median of 181.0. In this case, since the mean is bigger than the median the data is skewed to the right.

```
> summary(banka$pdays)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
    0.0   999.0   999.0   960.4   999.0   999.0
```

The variable “pdays” has mean of 960 and median of 999.0. In this case, since the mean is smaller than the median the data is skewed to the left.

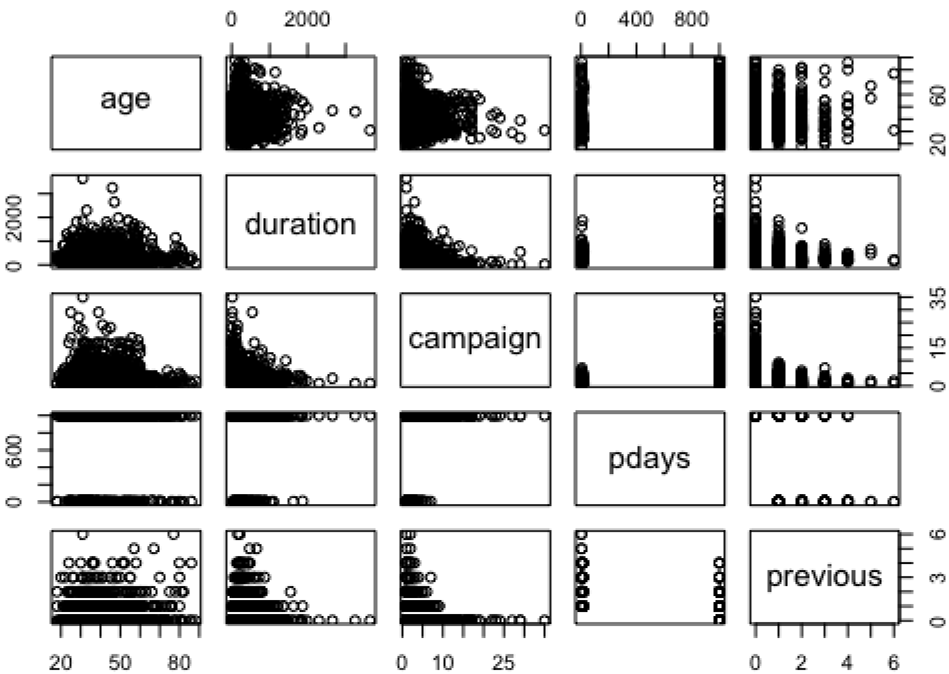
```
> summary(banka$campaign)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
  1.000   1.000   2.000   2.537   3.000  35.000
```

The variable “campaign” has mean of 2.537 and median of 2.000. In this case, since the mean is bigger than the median the data is skewed to the right.

```
> summary(banka$previous)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
 0.0000  0.0000  0.0000  0.1903  0.0000  6.0000
```

The variable “previous” has mean of 0.1903 and median of 0.0000. In this case, since the mean is bigger than the median the data is skewed to the right.

- **Relationship of numerical variables**.



As with other variables, it shows the strong correlations. One of these is the relationship between the variables duration and campaign. It is clear that as campaign duration increases, campaign duration decreases. However, as duration increases, previous decreases. Finally, the variables campaign and previous are examples of this because as campaign increases, so does previous.

Analyzing Data

Using Logistics Regression.

- Divide the data randomly into a training and testing data set of appropriate sizes. The train/test ratio could be divided with the ratio of 50/50, 70/30 or 80/20. However, for the best result, we should use the 80/20 ratio. (4119*80% = 3295)

  ```
  > training_sample <- sample(4119,3295)
  > bank_train <- bankdata[training_sample,]
  > bank_test <- bankdata[-training_sample,]
  ```

- Analyzing data by using a Logistic Regression model.

```
> age <- banka$age
> job <- banka$job
> marital <-banka$marital
> education <- banka$education
> default <- banka$default
> housing <- banka$housing
> loan <- banka$loan
> contact <- banka$contact
> month <- banka$month
> day_of_week <- banka$day_of_week
> duration <- banka$duration
> pdays <- banka$pdays
> previous <- banka$previous
> campaign <- banka$campaign
> poutcome <- banka$poutcome
```

Because the variable “y” of original data is a binary variable, we produce a new variable which is called deposit.

```
> deposit <- ifelse(banka$y == "yes",1,0)
```

```
> deposit <- banka$deposit
```

- Now, let's organize data variable

```
> bankdata <- data.frame(deposit, age, job, marital, education, default, housing, loan,
contact, month, month, day_of_week, duration, campaign, pdays, previous,poutcome)
```

After storing all the variables from data, we produce a Logistic Regression model.

```
> model1 <- glm(deposit ~ age + job + marital  + education + default+ housing + loan+
contact+ month+ month+ day_of_week+ duration+ campaign+ pdays+
previous+poutcome , data = bank_train, family = binomial(link = logit))

> summary(model1)

Call:

glm(formula = deposit ~ age + job + marital + education + default +

    housing + loan + contact + month + month + day_of_week +

    duration + campaign + pdays + previous + poutcome, family = binomial(link = logit),

    data = bank_train)


Deviance Residuals:

    Min       1Q   Median       3Q      Max

-4.8833  -0.3386  -0.2234  -0.1290   2.9562


Coefficients: (1 not defined because of singularities)

                     Estimate Std. Error z value Pr(>|z|)

(Intercept)          -4.519e+00  9.656e-01  -4.680 2.87e-06 ***

age                   1.827e-02  7.881e-03   2.319 0.020421 *

jobblue-collar       -2.541e-01  2.599e-01  -0.978 0.328113

jobentrepreneur      -8.288e-01  4.829e-01  -1.716 0.086111 .

jobhousemaid          4.412e-01  4.075e-01   1.083 0.278875

jobmanagement        -3.544e-01  2.712e-01  -1.307 0.191255

jobretired           -2.484e-03  3.344e-01  -0.007 0.994073

jobself-employed     -7.320e-01  4.026e-01  -1.818 0.068999 .
```

jobservices	1.978e-01	2.697e-01	0.734	0.463201	
jobstudent	5.950e-01	3.877e-01	1.535	0.124869	
jobtechnician	1.335e-01	2.116e-01	0.631	0.528037	
jobunemployed	5.074e-01	3.665e-01	1.385	0.166159	
jobunknown	-5.514e-01	7.631e-01	-0.723	0.469946	
maritalmarried	2.228e-01	2.307e-01	0.966	0.334023	
maritalsingle	3.965e-01	2.608e-01	1.520	0.128420	
maritalunknown	-3.517e-02	1.120e+00	-0.031	0.974956	
educationbasic.6y	2.593e-01	3.889e-01	0.667	0.504966	
educationbasic.9y	1.074e-01	3.128e-01	0.343	0.731258	
educationhigh.school	1.650e-01	2.948e-01	0.560	0.575743	
educationilliterate	-9.870e+00	5.354e+02	-0.018	0.985292	
educationprofessional.course	2.292e-01	3.215e-01	0.713	0.475872	
educationuniversity.degree	4.506e-01	2.959e-01	1.523	0.127797	
educationunknown	5.668e-01	3.703e-01	1.531	0.125830	
defaultunknown	-1.944e-01	2.002e-01	-0.971	0.331544	
defaultyes	-1.012e+01	5.354e+02	-0.019	0.984926	
housingunknown	-5.755e-01	5.112e-01	-1.126	0.260342	
housingyes	-6.258e-02	1.322e-01	-0.473	0.635929	
loanunknown	NA	NA	NA	NA	
loanyes	-1.071e-01	1.797e-01	-0.596	0.551333	
contacttelephone	-1.461e+00	2.040e-01	-7.165	7.77e-13	***
monthaug	-4.504e-01	2.821e-01	-1.596	0.110381	
monthdec	1.668e+00	5.979e-01	2.790	0.005269	**
monthjul	-9.069e-01	2.928e-01	-3.098	0.001950	**
monthjun	9.158e-01	3.050e-01	3.002	0.002679	**
monthmar	2.221e+00	4.195e-01	5.295	1.19e-07	***

```
monthmay              -5.320e-01  2.695e-01  -1.974 0.048350 *
monthnov              -9.120e-01  3.086e-01  -2.956 0.003119 **
monthoct               1.274e+00  3.904e-01   3.263 0.001102 **
monthsep               8.766e-01  4.163e-01   2.106 0.035233 *
day_of_weekmon         1.051e-01  2.049e-01   0.513 0.607877
day_of_weekthu         3.736e-02  2.049e-01   0.182 0.855349
day_of_weektue         2.655e-02  2.103e-01   0.126 0.899538
day_of_weekwed         2.288e-01  2.121e-01   1.078 0.280874
duration               4.828e-03  2.393e-04  20.175  < 2e-16 ***
campaign              -1.358e-01  4.381e-02  -3.099 0.001943 **
pdays                 -1.266e-04  6.743e-04  -0.188 0.851076
previous               3.547e-01  1.740e-01   2.038 0.041519 *
poutcomenonexistent    2.448e-01  2.923e-01   0.838 0.402282
poutcomesuccess        2.300e+00  6.644e-01   3.461 0.000538 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 2845.8  on 4118  degrees of freedom
Residual deviance: 1749.3  on 4071  degrees of freedom
```

AIC: 1845.3

```
Number of Fisher Scoring iterations: 12
```

→ In this regression model, most variables are not significant at level 0.05, even the AIC value is pretty high (which is1845.3). In other words, this model does not turn out the best regression for predicting. So, we should create a new model to determine which one we should use for predicting.

- We create the analysis of deviance for this model.

```
> anova (model1, test = "Chisq")
Analysis of Deviance Table

Model: binomial, link: logit

Response: deposit

Terms added sequentially (first to last)

 Df Deviance Resid. Df Resid. Dev  Pr(>Chi)
NULL                  4118     2845.8
age          1    14.53      4117     2831.3 0.0001377 ***
job         11    60.17      4106     2771.1 8.625e-09 ***
marital      3    15.08      4103     2756.1 0.0017485 **
education    7    10.44      4096     2745.6 0.1648036
default      2    22.46      4094     2723.2 1.326e-05 ***
housing      2     0.16      4092     2723.0 0.9243241
loan         1     0.24      4091     2722.8 0.6242710
contact      1    63.05      4090     2659.7 2.014e-15 ***
month        9   169.03      4081     2490.7 < 2.2e-16 ***
day_of_week  4     0.60      4077     2490.1 0.9627117
duration     1   569.02      4076     1921.1 < 2.2e-16 ***
campaign     1    15.63      4075     1905.4 7.718e-05 ***
pdays        1   141.35      4074     1764.1 < 2.2e-16 ***
previous     1     2.45      4073     1761.6 0.1178113
poutcome     2    12.32      4071     1749.3 0.0021118 **
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

- Now, let’s check a similar model without those insignificant variables based on the anova table (education , housing , loan, day_of_week, previous)

```
> model2 <- glm(deposit ~ age +job+ marital+ default+ contact+ month+ duration+
campaign+ pdays+ poutcome, data = bank_train, binomial (link=logit) )

> summary(model2)

Call:

glm(formula = deposit ~ age + job + marital + default + contact +

    month + duration + campaign + pdays + poutcome, family = binomial(link = logit),

    data = bank_train)

Deviance Residuals:

    Min       1Q   Median       3Q      Max

-4.9362  -0.3371  -0.2261  -0.1319   3.0060

Coefficients:

                      Estimate Std. Error z value Pr(>|z|)

(Intercept)        -3.229e+00  7.675e-01  -4.207 2.58e-05 ***

age                 1.797e-02  7.589e-03   2.369 0.017860 *

jobblue-collar     -4.235e-01  2.141e-01  -1.978 0.047880 *

jobentrepreneur    -8.620e-01  4.795e-01  -1.798 0.072216 .

jobhousemaid        3.035e-01  3.913e-01   0.776 0.438012

jobmanagement      -2.730e-01  2.656e-01  -1.028 0.304084

jobretired         -1.574e-01  3.262e-01  -0.482 0.629514

jobself-employed   -7.419e-01  3.991e-01  -1.859 0.063024 .

jobservices         5.054e-02  2.506e-01   0.202 0.840146

jobstudent          5.393e-01  3.655e-01   1.476 0.140076

jobtechnician       7.162e-02  1.925e-01   0.372 0.709929

jobunemployed       4.033e-01  3.601e-01   1.120 0.262822

jobunknown         -5.832e-01  7.566e-01  -0.771 0.440823

maritalmarried      2.306e-01  2.293e-01   1.006 0.314471

maritalsingle       4.493e-01  2.576e-01   1.744 0.081149 .

maritalunknown      1.070e-02  1.112e+00   0.010 0.992322
```

```
defaultunknown      -2.271e-01  1.969e-01  -1.153 0.248769
defaultyes          -9.237e+00  3.247e+02  -0.028 0.977309
contacttelephone    -1.450e+00  2.021e-01  -7.175 7.24e-13 ***
monthaug            -3.750e-01  2.776e-01  -1.351 0.176706
monthdec             1.734e+00  5.910e-01   2.934 0.003342 **
monthjul            -8.681e-01  2.892e-01  -3.002 0.002682 **
monthjun             9.605e-01  3.019e-01   3.181 0.001466 **
monthmar             2.260e+00  4.180e-01   5.407 6.40e-08 ***
monthmay            -5.020e-01  2.659e-01  -1.888 0.059030 .
monthnov            -8.699e-01  3.062e-01  -2.841 0.004498 **
monthoct             1.380e+00  3.857e-01   3.577 0.000348 ***
monthsep             9.795e-01  4.114e-01   2.381 0.017270 *
duration             4.810e-03  2.382e-04  20.188  < 2e-16 ***
campaign            -1.367e-01  4.373e-02  -3.126 0.001770 **
pdays               -6.826e-04  6.124e-04  -1.115 0.265011
poutcomenonexistent -1.714e-01  1.965e-01  -0.872 0.383080
poutcomesuccess      1.881e+00  6.212e-01   3.028 0.002459 **
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
(Dispersion parameter for binomial family taken to be 1)
  Null deviance: 2845.8  on 4118  degrees of freedom
Residual deviance: 1762.9  on 4086  degrees of freedom
AIC: 1828.9
Number of Fisher Scoring iterations: 11

> anova(model2, test = "Chisq")
Analysis of Deviance Table
```

Model: binomial, link: logit

Response: deposit

Terms added sequentially (first to last)

	Df	Deviance	Resid. Df	Resid. Dev	Pr(>Chi)	
NULL			4118	2845.8		
age	1	14.53	4117	2831.3	0.0001377	***
job	11	60.17	4106	2771.1	8.625e-09	***
marital	3	15.08	4103	2756.1	0.0017485	**
default	2	24.42	4101	2731.6	4.969e-06	***
contact	1	64.33	4100	2667.3	1.051e-15	***
month	9	171.47	4091	2495.8	< 2.2e-16	***
duration	1	566.30	4090	1929.5	< 2.2e-16	***
campaign	1	15.46	4089	1914.1	8.427e-05	***
pdays	1	140.90	4088	1773.2	< 2.2e-16	***
poutcome	2	10.30	4086	1762.9	0.0057898	**

Signif. codes: 0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

In the new model of logistic regression, the value of AIC reduces from 1845.3 to 1828.9, which is considered a large enough gain. Furthermore, there are 14/33 variables that are significant at level 0.05 (rather than a model which is only 14/49 variables that are significant). The analysis of the deviance table shows the major reductions in deviance for all of these variables, and the Pr(>Chi) is pretty small, which means it’s useful.

In conclusion, the model 2 (which the variables education , housing , loan, day_of_week, previous are being subtracted from model 1) will produce more accuracy results for predicting than the model 1 .

- Confusion Matrix

> predmodel <- predict(model2, newdata= bank_test, type = "response")

> pred_test <- ifelse(predmmodel > .01,1,0)

> CrossTable(deposit,pred_test)

```
   Cell Contents
|-------------------------|
|                       N |
| Chi-square contribution |
|           N / Row Total |
|           N / Col Total |
|         N / Table Total |
|-------------------------|


Total Observations in Table:  4119


             | pred_test
     deposit |         0 |         1 | Row Total |
-------------|-----------|-----------|-----------|
           0 |      3587 |        81 |      3668 |
             |     5.972 |    91.645 |           |
             |     0.978 |     0.022 |     0.891 |
             |     0.928 |     0.321 |           |
             |     0.871 |     0.020 |           |
-------------|-----------|-----------|-----------|
           1 |       280 |       171 |       451 |
             |    48.572 |   745.351 |           |
             |     0.621 |     0.379 |     0.109 |
             |     0.072 |     0.679 |           |
             |     0.068 |     0.042 |           |
-------------|-----------|-----------|-----------|
Column Total |      3867 |       252 |      4119 |
             |     0.939 |     0.061 |           |
-------------|-----------|-----------|-----------|
```

The test data for this Decision Tree model included 4119 observations. 3587 cases were correctly predicted, accounting for 87% of the total, and these are true negatives. Furthermore, 171 out of 824 observations were correctly predicted, representing a 4% accuracy rate, and these are true positives. There are also 280 false negatives, accounting for 7% percent, and 81 false positives, accounting for 2%. The total accuracy of the model is 91%, which is a high enough chance of accuracy and the percentage of error is just 9%. The Logistic Regression Cross Table is considered a good fit model.

Decision Tree Classification

- **DECISION TREE CROSS TABLE**

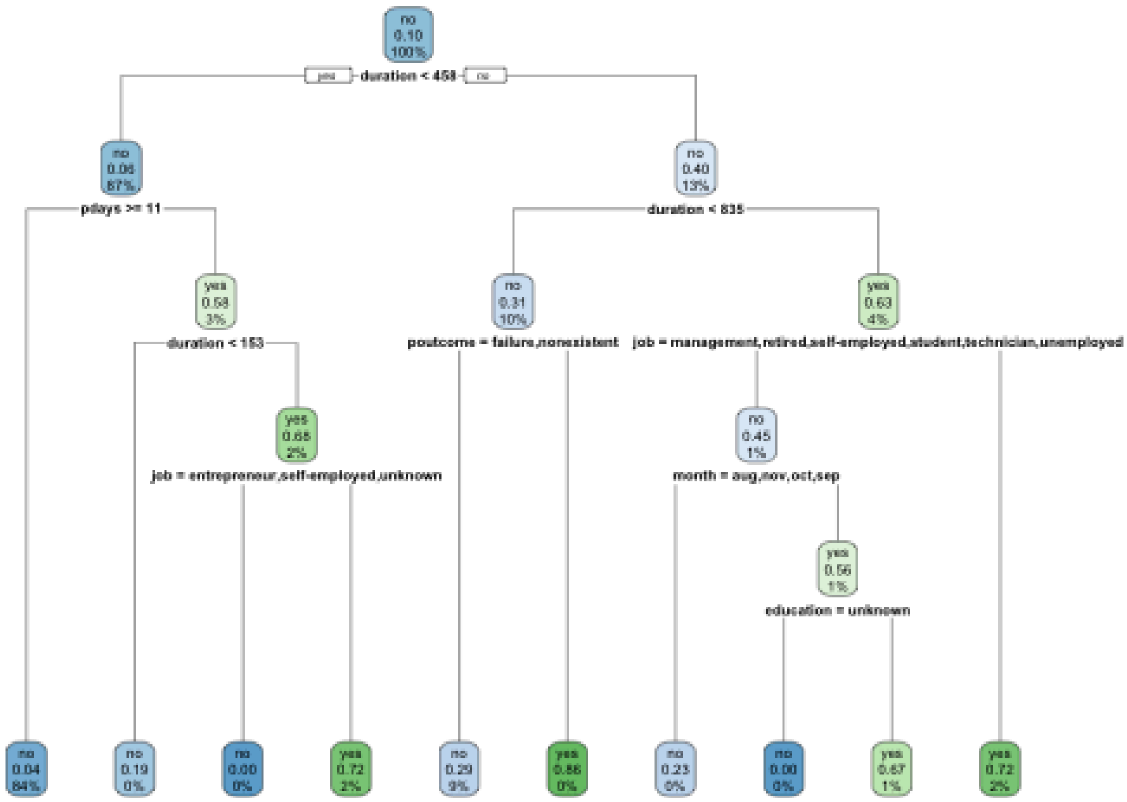
```
> install.packages("rpart.plot")

> library(rpart.plot)

> library(gmodels)

 > DTtrain<- sample(c(1:4119),3295)

>  bank_DTtrain <- banka[DTtrain, 1:16]

 > bank_DTtest <- banka[-DTtrain, 1:16]

>  DTmodel <- rpart(bank_DTtrain$y ~ ., method = "class", data =
bank_DTtrain,minsplit=10)

>  predict <- predict(DTmodel, newdata= bank_DTtest, type='class')

> CrossTable(bank_DTtest$y, predict)
```

```
   Cell Contents
|-------------------------|
|                       N |
| Chi-square contribution |
|           N / Row Total |
|           N / Col Total |
|         N / Table Total |
|-------------------------|


Total Observations in Table:  824


              | predict
bank_DTtest$y |        no |       yes | Row Total |
--------------|-----------|-----------|-----------|
           no |       700 |        15 |       715 |
              |     1.519 |    21.659 |           |
              |     0.979 |     0.021 |     0.868 |
              |     0.909 |     0.278 |           |
              |     0.850 |     0.018 |           |
--------------|-----------|-----------|-----------|
          yes |        70 |        39 |       109 |
              |     9.964 |   142.073 |           |
              |     0.642 |     0.358 |     0.132 |
              |     0.091 |     0.722 |           |
              |     0.085 |     0.047 |           |
--------------|-----------|-----------|-----------|
 Column Total |       770 |        54 |       824 |
              |     0.934 |     0.066 |           |
--------------|-----------|-----------|-----------|
```

The test data for this Decision Tree model included 824 observations. 700 cases were correctly predicted, accounting for 85% of the total, and these are true negatives. Furthermore, 39 out of 824 observations were correctly predicted, representing a 5% accuracy rate, and these are true positives. There are also 70 false negatives, accounting for 8.5% percent, and 39 false positives, accounting for 4.7%. The total accuracy of the model is 90%, indicating that this model has an accuracy rate in the 80-90 percent range, which is excellent.

- Now we produce the Decision Tree model

  > rpart.plot(DTmodel)



For the Root Node, it checks to see if the person is subscribed; if not, it checks to see if a duration (call) has been made. If it has not been made, it proceeds to the second node to the right, which asks again if the call has been made and then proceeds to the month or job based on the response. Otherwise, if the call was made, it goes to the second node to the left and asks if a number of days have passed since the client was contacted before returning to the duration question.

In conclusion, with the accuracy percent of 90% in Decision Tree Cross Table, and the percentage of “yes” if a person is subscribed is 87%. There is good evidence to consider that Decision Tree Classification is good for producing or predicting results.

Analysis Data by using Naive Bayesian.

```
> install.packages("e1071")
 > library(e1071)
> bankatrainbayes <- bankdata[training_sample, c(1:16)]
> train_labels <- bankdata[training_sample, c("deposit")]
> bankatestbayes <- bankdata[-training_sample, c(1:16)]
> test_labels <- bankdata[-training_sample, c("deposit")]
> modelBayes <- naiveBayes(bankatrainbayes, train_labels, laplace =0)
> bankapredict <- predict(modelBayes, bankatestbayes, type= "raw")
> predclass <- ifelse(bankapredict[,2] >= 0.01,1,0)
> CrossTable(test_labels, predclass)
```

```
   Cell Contents
|-------------------------|
|                       N |
| Chi-square contribution |
|           N / Row Total |
|           N / Col Total |
|         N / Table Total |
|-------------------------|


Total Observations in Table:  824


             | predclass 
 test_labels |         0 |         1 | Row Total | 
-------------|-----------|-----------|-----------|
           0 |       716 |        18 |       734 | 
             |     9.589 |    63.572 |           | 
             |     0.975 |     0.025 |     0.891 | 
             |     1.000 |     0.167 |           | 
             |     0.869 |     0.022 |           | 
-------------|-----------|-----------|-----------|
           1 |         0 |        90 |        90 | 
             |    78.204 |   518.463 |           | 
             |     0.000 |     1.000 |     0.109 | 
             |     0.000 |     0.833 |           | 
             |     0.000 |     0.109 |           | 
-------------|-----------|-----------|-----------|
Column Total |       716 |       108 |       824 | 
             |     0.869 |     0.131 |           | 
-------------|-----------|-----------|-----------|
```

The test data for this Decision Tree model included 824 observations. 716 cases were correctly predicted, accounting for 87% of the total, and these are true negatives. Furthermore, 90 out of 824 observations were correctly predicted, representing a 11% accuracy rate, and these are true positives. There are also 0 false negatives, accounting for 0% percent, and 18 false positives, accounting for 2%. The total accuracy of the model is 98%, which is a very high chance of accuracy and the percentage of error is just 2%. The Naive Bayesian Cross Table is considered a good fit model.