

CREDIT CARD

TÉCNICAS DE CLASIFICACIÓN

Jose López, Antonio Romero, Francisco del Val

2020-11-15

ABSTRACT

El objetivo del siguiente trabajo es buscar el mejor modelo que sea capaz de clasificar a los individuos que buscan una tarjeta de crédito. Utilizando la base de datos *CreditCard* del paquete *AER*. Para ello, hemos realizado un análisis exploratorio de los datos y los siguientes modelos de clasificación: lineal, logístico, LDA y QDA.

BASE DE DATOS

Tenemos un dataset con 1319 filas y 12 variables, estas son:

- **card**: ¿Se aceptó la solicitud de tarjeta de crédito?
- **reports**: Número de informes negativos importantes
- **age**: Edad en años y meses
- **income**: Renta anual (en 10.000 USD)
- **share**: Relación entre los gastos mensuales de la tarjeta de crédito e ingresos anuales
- **expenditure**: Gasto medio mensual con tarjeta de crédito
- **owner**: ¿El individuo es dueño de su casa?
- **selfemp**: ¿Es autoempleado?
- **dependents**: Número de dependientes
- **months**: Meses viviendo en la dirección actual
- **majorcards**: Número de tarjetas de crédito retenidas
- **active**: Número de cuentas de crédito activas

ANÁLISIS EXPLORATORIO

A continuación realizaremos un análisis exploratorio de nuestra base de datos, en el que analizaremos los estadísticos principales, correlaciones...

RESUMEN DE LOS ESTADÍSTICOS

En este apartado incluiremos un resumen con los estadísticos principales de cada variable:

Data summary

| | |
|-------------------|------|
| Name | data |
| Number of rows | 1319 |
| Number of columns | 12 |

Column type frequency:






| | |
|---------|---|
| factor | 3 |
| numeric | 9 |

| | |
|-----------------|------|
| Group variables | None |
|-----------------|------|

Variable type: factor

| skim_variable | n_missing | complete_rate | ordered | n_unique | top_counts |
|---------------|-----------|---------------|---------|----------|--------------------|
| card | 0 | 1 | FALSE | 2 | yes: 1023, no: 296 |
| owner | 0 | 1 | FALSE | 2 | no: 738, yes: 581 |
| selfemp | 0 | 1 | FALSE | 2 | no: 1228, yes: 91 |

Variable type: numeric

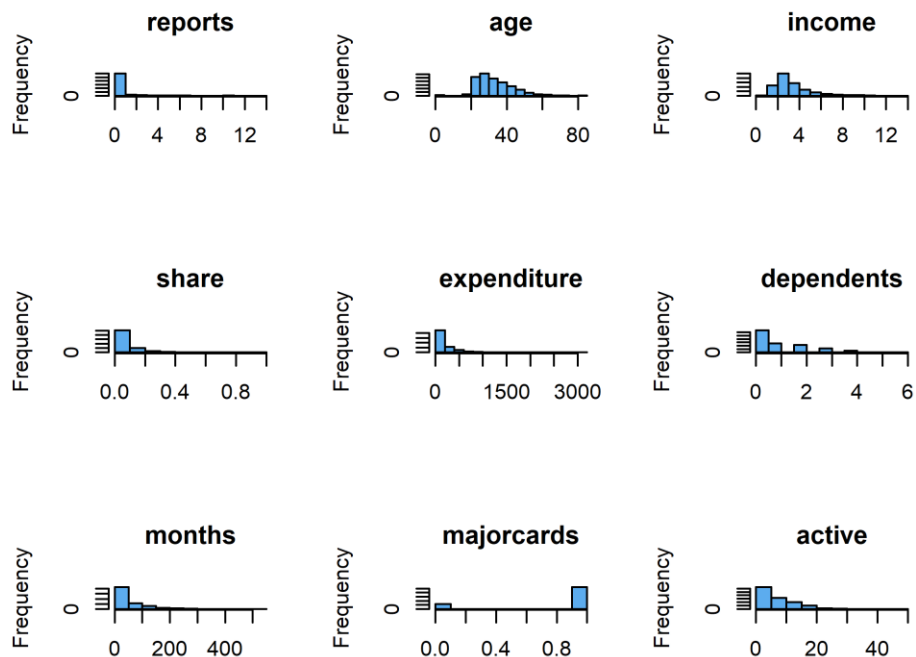
| skim_variable | n_missing | complete_rate | mean | sd | p0 | p25 | p50 | p75 | p100 | hist |
|---------------|-----------|---------------|--------|--------|------|-------|--------|--------|---------|---|
| reports | 0 | 1 | 0.46 | 1.35 | 0.00 | 0.00 | 0.00 | 0.00 | 14.00 |  |
| age | 0 | 1 | 33.21 | 10.14 | 0.17 | 25.42 | 31.25 | 39.42 | 83.50 |  |
| income | 0 | 1 | 3.37 | 1.69 | 0.21 | 2.24 | 2.90 | 4.00 | 13.50 |  |
| share | 0 | 1 | 0.07 | 0.09 | 0.00 | 0.00 | 0.04 | 0.09 | 0.91 |  |
| expenditure | 0 | 1 | 185.06 | 272.22 | 0.00 | 4.58 | 101.30 | 249.04 | 3099.50 |  |

| | | | | | | | | | | |
|------------|---|---|-------|-------|-----|------|-------|-------|--------|----|
| dependents | 0 | 1 | 0.99 | 1.25 | 0.0 | 0.00 | 1.00 | 2.00 | 6.00 | |
| | | | | | 0 | | | | | -- |
| | | | | | | | | | | -- |
| months | 0 | 1 | 55.27 | 66.27 | 0.0 | 12.0 | 30.00 | 72.00 | 540.00 | |
| | | | | | 0 | 0 | | | | -- |
| | | | | | | | | | | -- |
| majorcards | 0 | 1 | 0.82 | 0.39 | 0.0 | 1.00 | 1.00 | 1.00 | 1.00 | |
| | | | | | 0 | | | | | -- |
| | | | | | | | | | | -- |
| active | 0 | 1 | 7.00 | 6.31 | 0.0 | 2.00 | 6.00 | 11.00 | 46.00 | |
| | | | | | 0 | | | | | -- |
| | | | | | | | | | | -- |

Tenemos 3 variables binarias, estas son **card**, **owner** y **selfemp**. El resto de variables son numéricas.

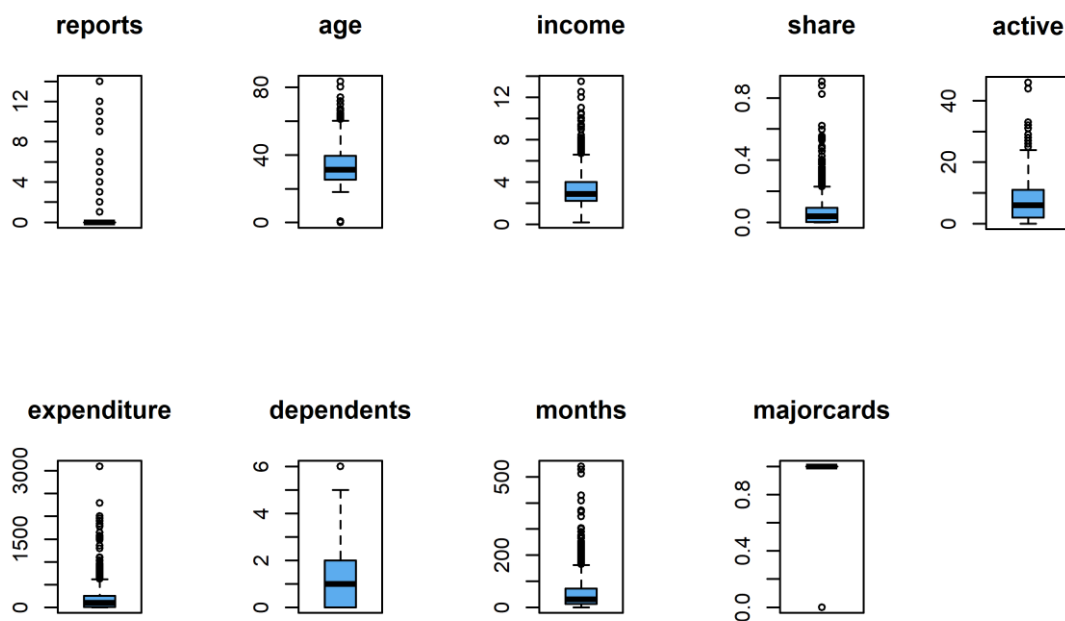
Como podemos observar, el coeficiente de variación de todas las variables es bastante elevado lo que nos indica que existe una alta desviación respecto a su media.

HISTOGRAMAS



En la imagen superior tenemos los histogramas de las variables numéricas y como podemos observar ninguna sigue una distribución normal, en general, todas tienen colas pesadas, en el caso de la renta, la media se sitúa en torno a 30.000 USD pero hay individuos con rentas superiores a los 100.000 USD.

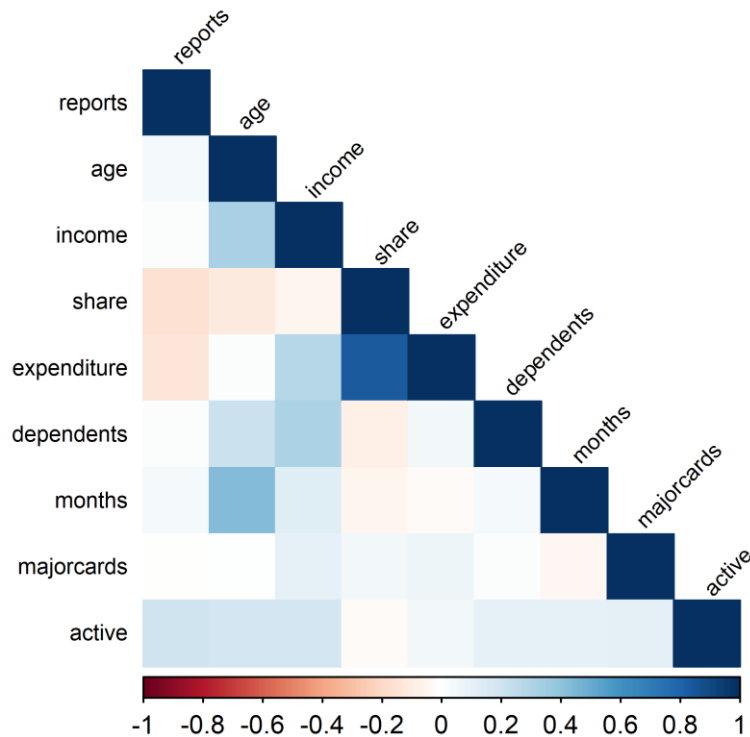
DIAGRAMAS DE CAJA



En este caso podemos observar los *boxplots* de las variables numéricas, como hemos comentado anteriormente presentan grandes desviaciones causadas por los *outliers*, podemos observar como la variable **age** tiene muchos *outliers* en la parte superior de la distribución, lo que nos indica una asimetría positiva o sesgada a la derecha.

CORRELACIONES

A continuación realizaremos un *heatmap* con las correlaciones.



Se puede observar que no existe una alta correlación entre las variables de nuestro dataset, salvo la relación entre **share** y **expenditure**, esto se debe a que la variable **share** incluye los gastos mensuales, ya explicados por la variable **expenditure**.

INGENIERÍA DE VARIABLES

Hemos generado unas variables dummy para las columnas categóricas, de tal manera que no tengamos problemas a la hora de calcular los modelos. Además como el gasto mensual (**expenditure**) está explicado en la proporción de gasto con la tarjeta (**share**) hemos eliminado esta variable.

MODELOS DE CLASIFICACIÓN

En este apartado vamos a realizar diversos modelos vistos durante la asignatura *Técnicas de Clasificación*. Comenzaremos con un modelo de regresión lineal, continuaremos con uno de regresión logística, para finalizar LDA y QDA.

MODELO DE REGRESIÓN LINEAL

En estadística la regresión lineal o ajuste lineal es un modelo matemático usado para aproximar la relación de dependencia entre una variable dependiente Y , las variables independientes X_i y un término aleatorio ε . Este modelo puede ser expresado como:

$$Y_t = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \dots + \beta_p X_p + \varepsilon$$

donde: - Y_t : variable dependiente o explicada. - X_n : variables explicativas o independientes. - β_n : parámetros, miden la influencia que las variables explicativas tienen sobre el regrediendo.

```
# Definimos nuestro modelo
modelo_lineal <- lm(card ~ ., data = data)

# Observamos el modelo
summary(modelo_lineal)
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)  5.643e-01  3.986e-02  14.155  < 2e-16 ***
## reports      -1.318e-01  7.180e-03 -18.360  < 2e-16 ***
## age          -6.022e-04  1.103e-03  -0.546  0.585319
## income        2.115e-02  6.158e-03   3.434  0.000612 ***
## share         1.399e+00  1.002e-01  13.963  < 2e-16 ***
## owner         7.208e-02  2.181e-02   3.305  0.000975 ***
## selfemp      -5.780e-02  3.689e-02  -1.567  0.117351
## dependents   -2.230e-02  8.089e-03  -2.757  0.005910 **
## months        4.164e-05  1.567e-04   0.266  0.790513
## majorcards    6.279e-02  2.430e-02   2.584  0.009868 **
## active        9.287e-03  1.587e-03   5.854  6.07e-09 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Como podemos observar las variables más significativas son los reportes negativos, el ingreso anual, la proporción del gasto, ser propietario de la vivienda y el número de tarjetas activas.

MATRIZ DE CONFUSIÓN Y PRECISIÓN

A continuación, calcularemos la matriz de confusión del modelo y su precisión.

```
## fit.pred    0    1
##           0 103    4
##           1 193 1019

## [1] 0.8506444
```

En cuanto a la matriz de confusión, podemos observar que de 107 personas a las que NO deberíamos haberle dado crédito, se lo hemos concedido a 4. Mientras que de un total de 1212 individuos a los que deberíamos haberle concedido el crédito, a 193 se les ha denegado la petición. Por lo tanto, nuestro modelo es muy

restrictivo a la hora de conceder el crédito, ya que el coste de conceder un crédito a una persona con posibilidad de insolvencia es muy elevado.

Tras el cálculo de la precisión, obtenemos un resultado del 85.06 %

Mediante el método AIC vamos a seleccionar el mejor modelo y volver a calcular para ver si obtenemos una mejor precisión:

Por lo tanto, el mejor modelo sería:

```
lm(formula = card ~ reports + income + share + owner + selfemp + dependents  
+ majorcards + active, data = data)
```

```
## fit.pred    0    1  
##           0 104    2  
##           1 192 1021  
  
## [1] 0.8529189
```

Como podemos hemos mejorado ligeramente la precisión del modelo, pasando del 85.06 % a 85.29 %

MODELO DE REGRESIÓN LOGÍSTICA

En estadística, la regresión logística es un tipo de análisis de regresión utilizado para predecir el resultado de una variable categórica (una variable que puede adoptar un número limitado de categorías) en función de las variables independientes o predictoras. Es útil para modelar la probabilidad de un evento ocurriendo como función de otros factores. El análisis de regresión logística se enmarca en el conjunto de Modelos Lineales Generalizados (GLM por sus siglas en inglés) que usa como función de enlace la función logit. Las probabilidades que describen el posible resultado de un único ensayo se modelan, como una función de variables explicativas, utilizando una función logística.

$$Y_i = B(p_i, n_i), i = 1, \dots, m$$

donde los números de ensayos Bernoulli n_i son conocidos y las probabilidades de éxito p_i son desconocidas.

CÁLCULO DEL MODELO

```
# Definimos nuestro modelo  
modelo_logistico <- glm(card ~ ., data = data, family = binomial(link = logit))  
  
# Observamos el modelo  
## Coefficients:  
##           Estimate Std. Error z value Pr(>|z|)  
## (Intercept) -5.089e+00  9.601e-01  -5.300 1.16e-07 ***  
## reports     -2.504e+00  1.010e+00  -2.479 0.01318 *  
## age         1.631e-02  2.219e-02   0.735 0.46239
```

```
## income      3.459e-01  1.496e-01   2.312  0.02080 *
## share       3.020e+03  6.235e+02   4.844  1.27e-06 ***
## owner       2.531e-01  5.568e-01   0.454  0.64948
## selfemp     4.853e-01  6.816e-01   0.712  0.47646
## dependents -6.529e-01  2.630e-01  -2.482  0.01305 *
## months      -4.157e-03  4.119e-03  -1.009  0.31296
## majorcards   3.502e-01  5.527e-01   0.634  0.52634
## active      9.529e-02  3.455e-02   2.758  0.00581 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

El modelo logístico toma como variables más representativas la relación del gasto respecto al ingreso anual.

MATRIZ DE CONFUSIÓN Y PRECISIÓN

A continuación, calcularemos la matriz de confusión del modelo y su precisión.

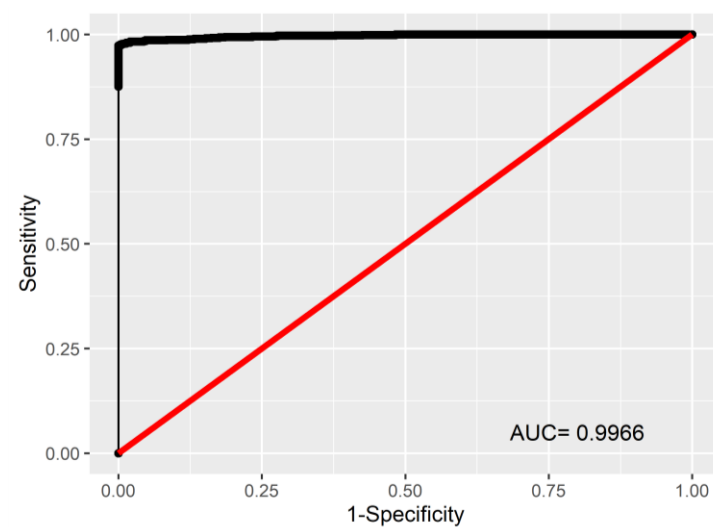
```
## fit.pred    0    1
##            0 294  23
##            1   2 1000

## [1] 0.9810462
```

En cuanto a la matriz de confusión, podemos observar que de 317 personas a las que NO deberíamos haberle dado crédito, se lo hemos concedido a 23. Mientras que de un total de 1002 individuos a los que deberíamos haberle concedido el crédito, a 2 se les ha denegado la petición. Por lo tanto, nuestro modelo ya no es tan restrictivo a la hora de conceder el crédito, porque solo denagamos el 0.02 % de los créditos a las personas que si deberíamos concederselo.

Tras el cálculo de la precisión, obtenemos un resultado del 98.1 %

CURVA ROC



Recordemos que cuánto más se aleje de la línea a 45° mejor modelo tendremos. En nuestro se trata de un modelo muy preciso.

ANÁLISIS DISCRIMINANTE LINEAL (LDA)

El Análisis Discriminante Lineal o *Linear Discriminant Analysis* (LDA) es un método de clasificación supervisado de variables cualitativas en el que dos o más grupos son conocidos a priori y nuevas observaciones se clasifican en uno de ellos en función de sus características. Haciendo uso del teorema de Bayes, LDA estima la probabilidad de que una observación, dado un determinado valor de los predictores, pertenezca a cada una de las clases de la variable cualitativa, $P(Y = k|X = x)$. Finalmente se asigna la observación a la clase k para la que la probabilidad predicha es mayor.

CÁLCULO DEL MODELO

```
# Definimos nuestro modelo
modelo_LDA <- lda(card ~ . -expenditure, data = data)
modelo_LDA

## Prior probabilities of groups:
##           0           1
## 0.2244124 0.7755876

##
## Coefficients of linear discriminants:
##                               LD1
## reports      -0.6585986025
## age          -0.0030087301
## income       0.1056653309
## share        6.9867960643
## owner        0.3601156087
## selfemp     -0.2887638263
## dependents  -0.1114229697
## months       0.0002080268
## majorcards   0.3136758394
## active       0.0463993281
```

La probabilidad a priori de no conceder el crédito es del 22.44 %, mientras que la probabilidad de concederlo es 77.55 %.

MATRIZ DE CONFUSIÓN Y PRECISIÓN

```
##      0      1
##  0  114      6
##  1  182 1017

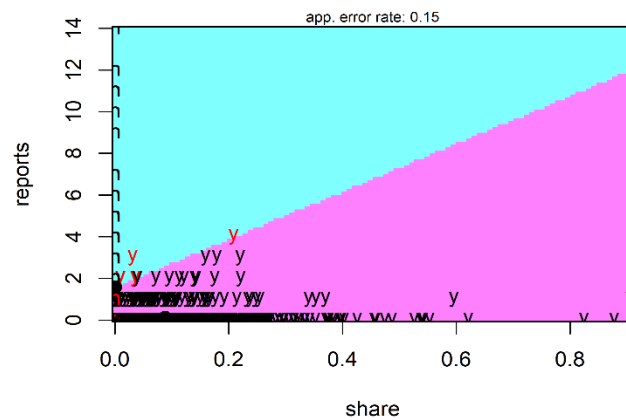
## [1] 0.8574678
```

En este caso, en la matriz de confusión podemos observar que de 120 personas a las que NO deberíamos haberle dado crédito, se lo hemos concedido a 6. Mientras que de un total de 1219 individuos a los que deberíamos haberle concedido el crédito, a 182 se les ha denegado la petición. Por lo tanto, nuestro modelo vuelve a ser muy restrictivo a la hora de conceder el crédito..

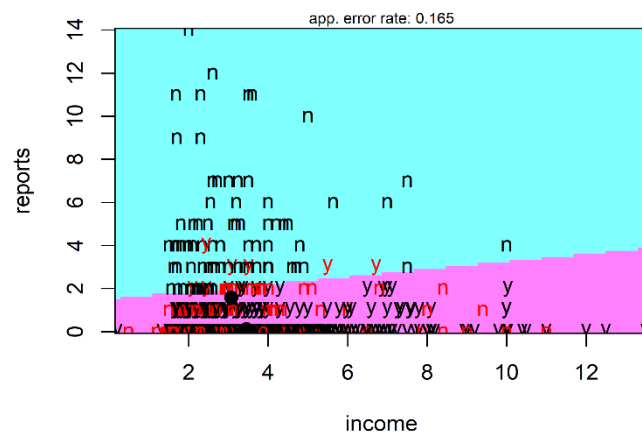
Tras el cálculo de la precisión, obtenemos un resultado del 85.74 %

GRÁFICOS DE PARTICIÓN

Gráficos de partición



Gráficos de partición



ANÁLISIS DISCRIMINANTE CUADRÁTICO (QDA)

El clasificador cuadrático o *Quadratic Discriminant Analysis* QDA se asemeja en gran medida al LDA, con la única diferencia de que el QDA considera que cada clase k tiene su propia matriz de covarianza (Σ^k) y, como consecuencia, la función discriminante toma forma cuadrática.

CÁLCULO DEL MODELO

```
# Definimos nuestro modelo
modelo_QDA <- qda(card ~ . - expenditure, data = data)
modelo_QDA

## Prior probabilities of groups:
##          0          1
## 0.2244124 0.7755876
```

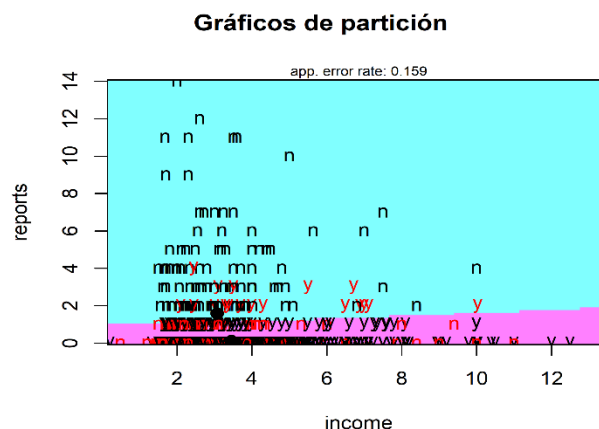
La probabilidad a priori de no conceder el crédito es del 22.44 %, mientras que la probabilidad de concederlo es 77.55 %.

```
##          0          1
##    0  295    23
##    1    1 1000
## [1] 0.9818044
```

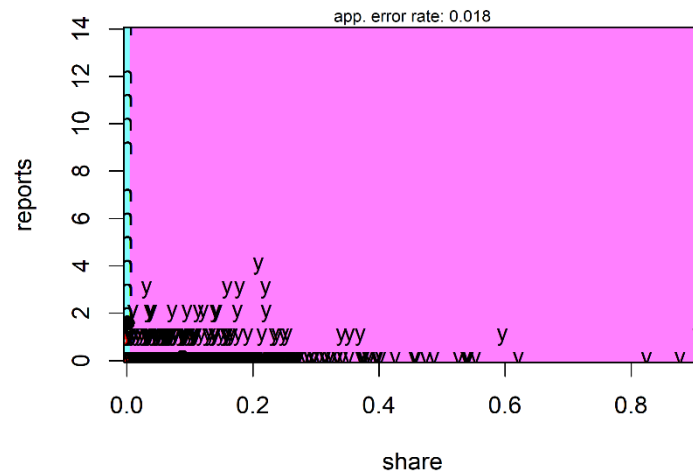
En este caso, en la matriz de confusión podemos observar que de 318 personas a las que NO deberíamos haberle dado crédito, se lo hemos concedido a 23. Mientras que de un total de 1001 individuos a los que deberíamos haberle concedido el crédito, a tan solo 1 se le ha denegado la petición. Por lo tanto, no es restrictivo a la hora de conceder el préstamo lo que hace que tengamos mayor precisión.

Tras el cálculo de la precisión, obtenemos un resultado del 98.18 %

GRÁFICOS DE PARTICIÓN



Gráficos de partición



CONCLUSIONES:

| ## | precision_lineal | precision_lineal_AIC | precision_logistico | precision_LDA | precision_QDA |
|----|------------------|----------------------|---------------------|---------------|---------------|
| ## | 0.8506444 | 0.8529189 | 0.9810462 | 0.8574678V | 0.9818044 |

Por lo tanto, elegimos el modelo QDA puesto que tiene mayor precisión más del 98%. Sin embargo, conociendo el negocio y el coste asociado a conceder una tarjeta de crédito a un individuo incapaz de responder a las deudas contraídas no seleccionaríamos un modelo que conceda préstamos a personas que van a impagar, como es el caso del QDA. Si somos un banco adverso al riesgo y no nos interesa casi ningún cliente moroso, sacrificando a posibles clientes por tener un modelo más restrictivo, seleccionaríamos el modelo lineal AIC. Sin embargo, si no tenemos tanta aversión al riesgo y no queremos perder clientes seleccionaríamos el LDA, que manteniendo un comportamiento de clasificación conservador no es tan restrictivo el modelo lineal AIC.

REFERENCIAS

Fellows, I. (2015, diciembre). Package 'Deducer'. Recuperado de <https://cran.r-project.org/web/packages/Deducer/Deducer.pdf>

Generalized linear model. (s. f.). En Wikipedia. Recuperado 15 de noviembre de 2020, de https://en.wikipedia.org/wiki/Generalized_linear_model

Kleiber, C., & Zeileis, A. (2020, junio). Package 'AER'. Recuperado de <https://cran.r-project.org/web/packages/AER/AER.pdf>

Ligges, U. (2020, febrero). Package 'klaR'. Recuperado de <https://cran.r-project.org/web/packages/klaR/klaR.pdf>

Linear discriminant analysis. (s. f.). En Wikipedia. Recuperado 15 de noviembre de 2020, de https://en.wikipedia.org/wiki/Linear_discriminant_analysis

Logistic regression. (s. f.). En Wikipedia. Recuperado 15 de noviembre de 2020, de https://en.wikipedia.org/wiki/Logistic_regression

Quadratic classifier. (s. f.). En Wikipedia. Recuperado 15 de noviembre de 2020, de https://en.wikipedia.org/wiki/Quadratic_classifier#Quadratic_discriminant_analysis

Ripley, B. (2020, septiembre). Package 'MASS'. Recuperado de <https://cran.r-project.org/web/packages/MASS/MASS.pdf>