Técnicas de clasificación: CreditCard

JOSE LÓPEZ GALDÓN
ANTONIO ROMERO MARTÍNEZ-EIROA
FRANCISCO DEL VAL YAGÜE

17/10/2020

ÍNDICE

1. INTRODUCCIÓN

Objetivo del trabajo

Realizar un análisis de clasificación con la finalidad de categorizar a los individuos en función de si le concedemos la tarjeta de crédito o no.

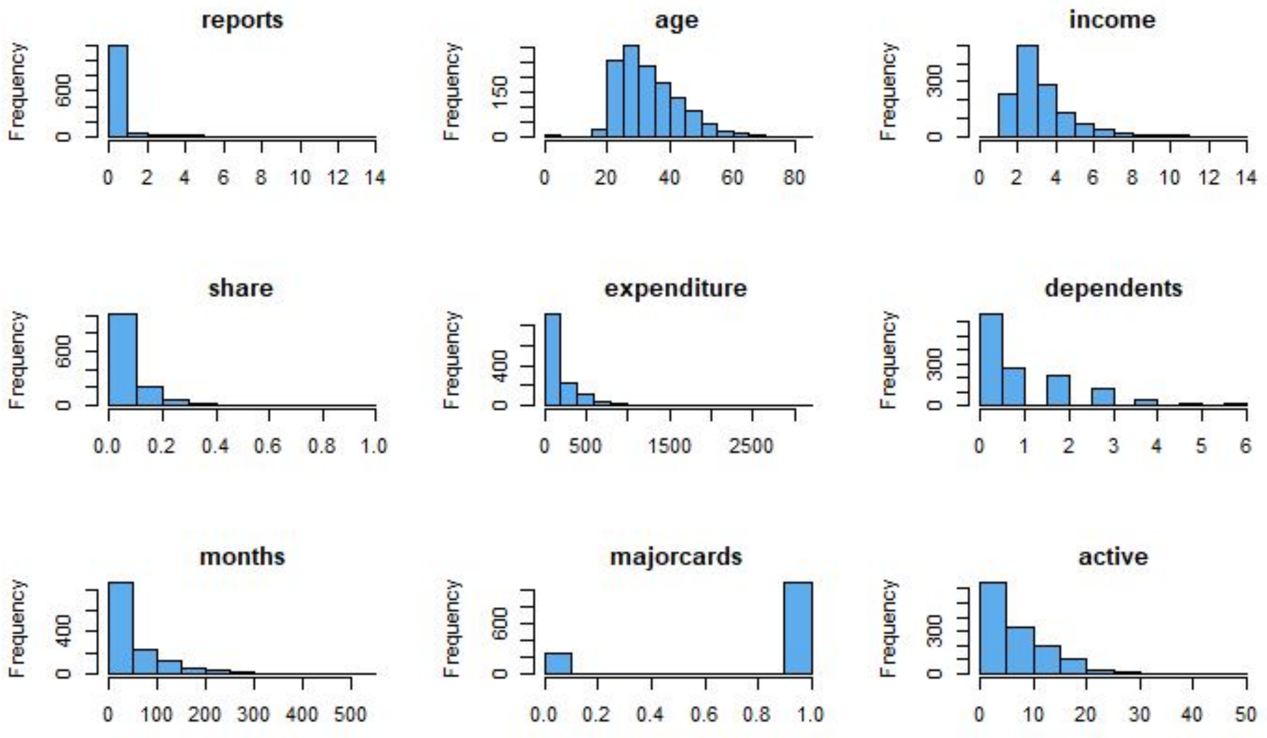
Base de datos

Para nuestro análisis hemos seleccionado la base de datos `CreditCard` del paquete `AER`.

Nuestra variables son:

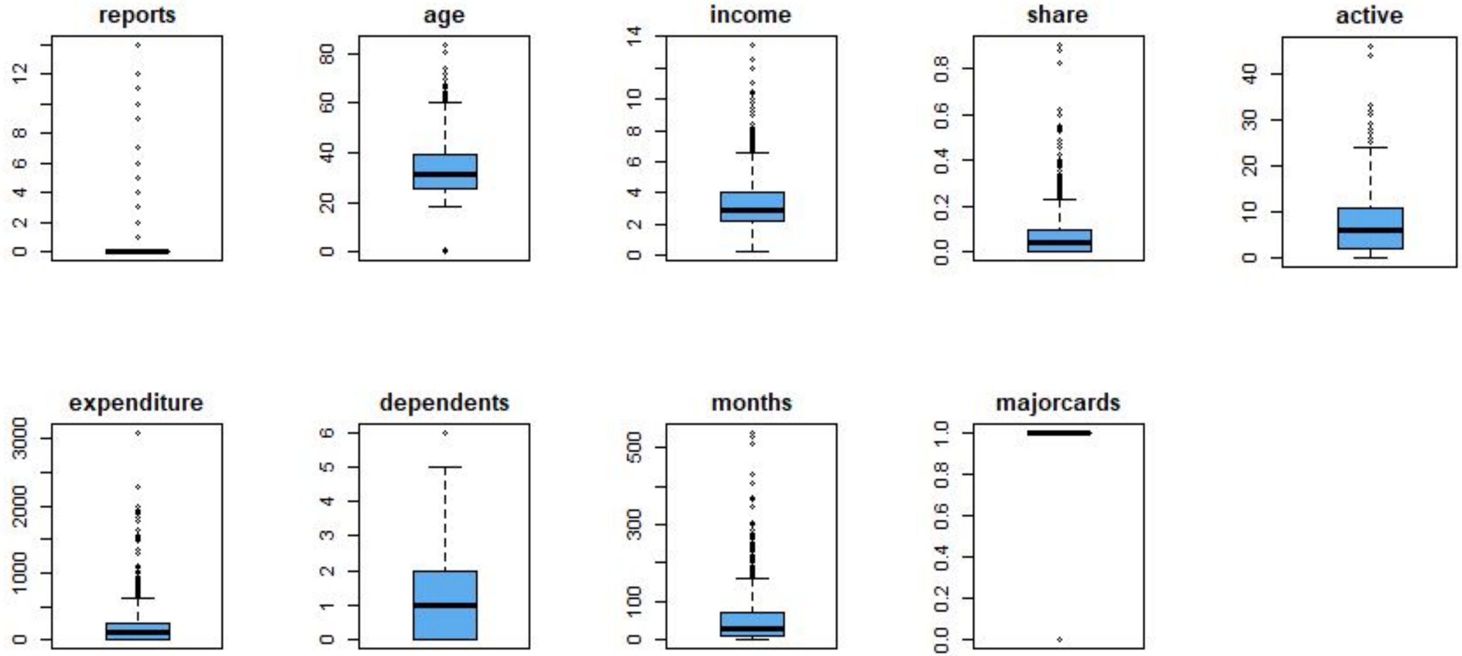
- `card`, `reports`, `age`, `income`, `share`, `expenditure`, `owner`, `selfemp`, `dependents`, `months`, `majorcards`, `active`

EDA



Histograma de las variables numéricas

EDA



Diagramas de caja de las variables numéricas

EDA

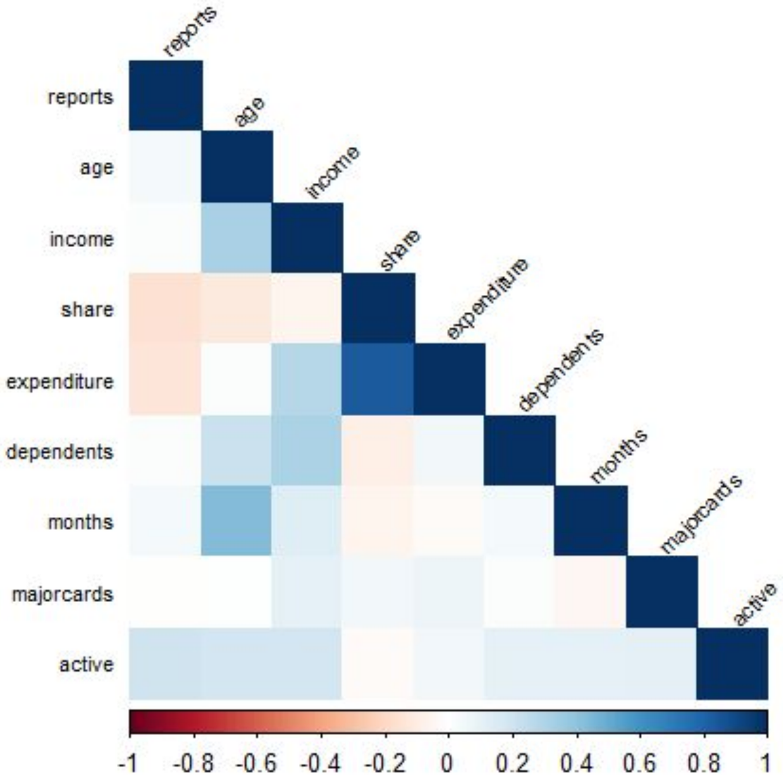


Gráfico de correlaciones

2. MODELOS DE CLASIFICACIÓN

MODELO LINEAL

Modelo Lineal

```
lm (formula = card ~ ., data = CreditCard)
```

```
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
```

Modelo Lineal

Matriz de confusión y precisión

	NO	YES
NO	103	4
YES	193	1019

`accuracy(lm) = 85 %`

Modelo Lineal AIC

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

	NO	YES
NO	104	2
YES	192	1021

```
accuracy(lm) = 85,29 %
```

MODELO LOGÍSTICO

Modelo Logístico

```
glm(formula = card ~ ., data = CreditCard,
family = binomial(link = logit))
```

```
## 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
```

Modelo Logístico

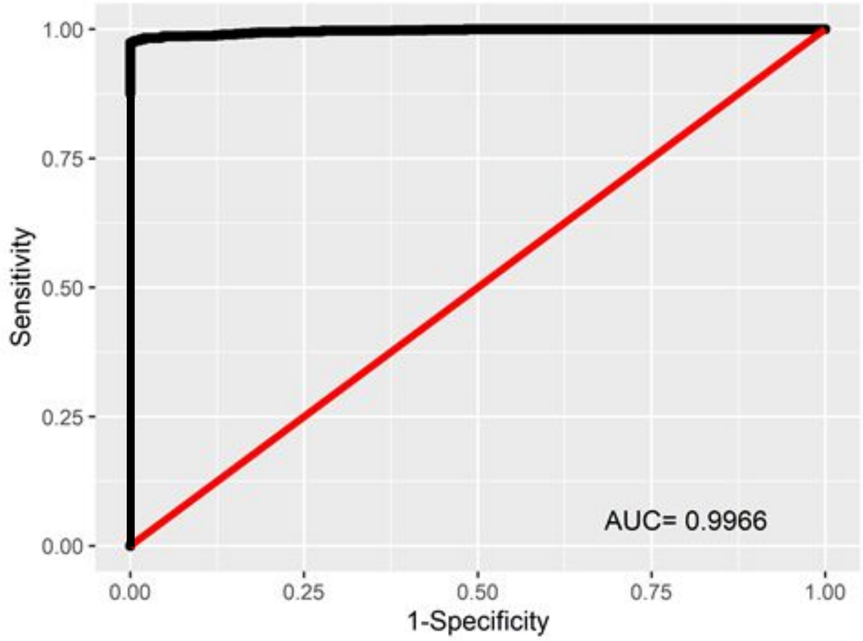
Matriz de confusión y precisión

	NO	YES
NO	294	23
YES	2	1000

`accuracy(glm) = 98,10 %`

Modelo Logístico

Curva ROC



MODELO LDA

Análisis Discriminante Lineal

```
lda(card ~ ., data = CreditCard)
```

	NO	YES
Probabilidades a priori	22,44 %	77,56 %

Análisis Discriminante Lineal

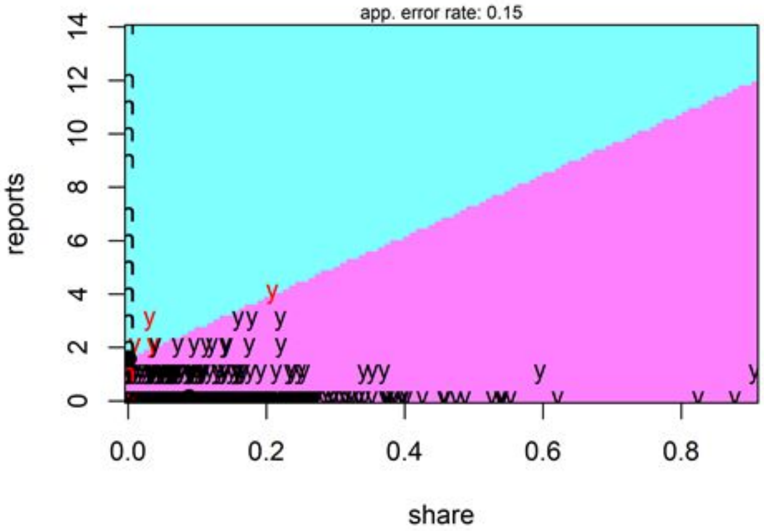
Matriz de confusión y precisión

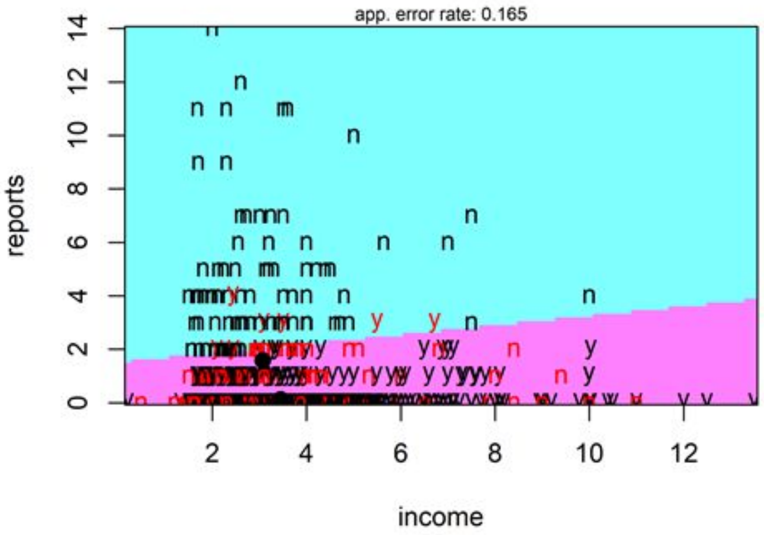
	NO	YES
NO	114	6
YES	182	1017

`accuracy(lda) = 85,75 %`

Análisis Discriminante Lineal

Gráficos de partición





MODELO QDA

Análisis Discriminante Cuadrático

```
qda(card ~ ., data = CreditCard)
```

	NO	YES
Probabilidades a priori	22,44 %	77,56 %

Análisis Discriminante Cuadrático

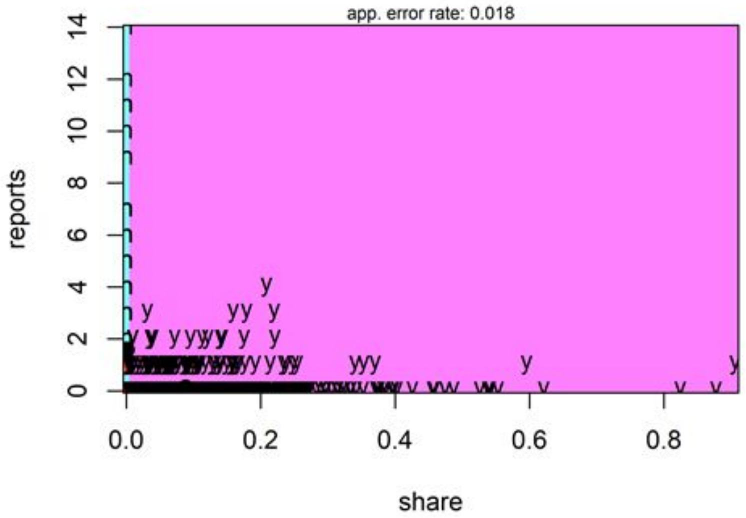
Matriz de confusión y precisión

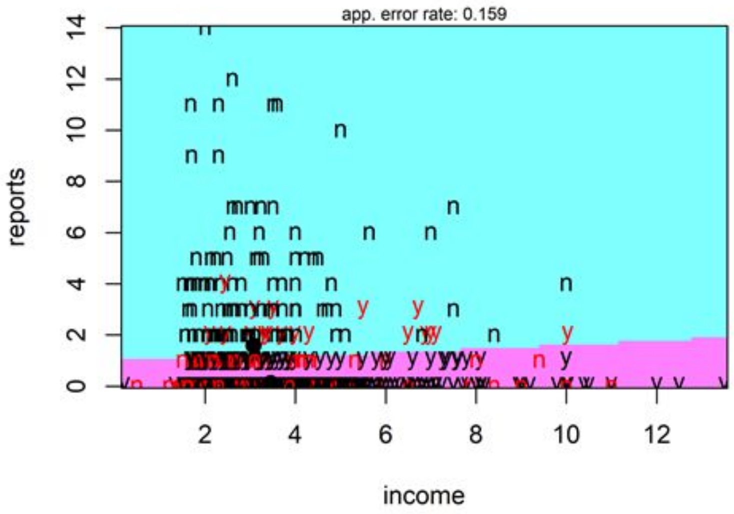
	NO	YES
NO	295	23
YES	1	1000

`accuracy(qda) = 98,18 %`

Análisis Discriminante Cuadrático

Gráficos de partición





3. CONCLUSIONES

Precisión de los modelos

Modelo Lineal	Modelo Lineal AIC	Modelo Logístico	LDA	QDA
85,06 %	85,29 %	98,10 %	85,75 %	98,18 %

Matriz de coste

	NO	YES
NO	0	10
YES	3	-3

Coste de los modelos

Modelo Lineal	Modelo Lineal AIC	Modelo Logístico	LDA	QDA
-2.438	-2.467	-2.764	-2.445	-2.767

* cantidades en 1.000 USD