Universidad Adolfo Ibáñez
Facultad de Ingeniería y Ciencias

# Modelamiento Predictivo
# Regresión Logística: Créditos con garantía Hipotecaria

Profesor: Rolando de la Cruz
Ayudante: Ismael Valdivia
Alumna: Carolina Iturriaga Bernal
Fecha: 05/01/2020

**ÍNDICE**

## Objetivo, atributos y variable target

El objetivo de este trabajo es poder desarrollar un modelo que ayude en la detección de clientes bancarios que pudiesen caer en morosidad al solicitar un crédito con garantía hipotecaria.
Para esto, se cuenta con un data set del historial d e 5960 créditos otorgados, los cuales ya se encuentran clasificados, según tuvieron incumplimiento o no.

La importancia de este tipo de decisión radica en que el banco necesita poder minimizar 2 tipos de riesgos. El primero, no aprobar un crédito a alguien que, si tiene alta probabilidad de pagar y segundo, dar el crédito a alguien que probablemente caerá en incumplimiento.

El data set cuenta con 13 variables descritas a continuación:

| | Variable | Descripción |
|---|---|---|
| *1* | BAD | Variable binaria, 1 = incumplimiento crédito, 0 = pagó el crédito |
| *2* | LOAN | Numérica. Monto de crédito solicitado |
| *3* | MORTDUE | Numérica. Monto adeudado de la hipoteca existente |
| *4* | VALUE | Numérica. Valor de la propiedad |
| *5* | REASON | Categórica. Motivo del crédito |
| *6* | JOB | Categórica. Ocupación del cliente |
| *7* | YOJ | Numérica. Años en trabajo actual |
| *8* | DEROG | Numérica. Cantidad de reportes de créditos no pagados |
| *9* | DELINQ | Numérica. Cantidad de líneas de crédito que no ha pagado |
| *10* | CLAGE | Numérica. Antigüedad de línea crédito más antigua en meses |
| *11* | NINQ | Numérica. Numero de consultas por créditos recientes |
| *12* | CLNO | Numérica. Cantidad de líneas de crédito |
| *13* | DEBTINC | Numérica. Ratio Deuda/Ingreso |

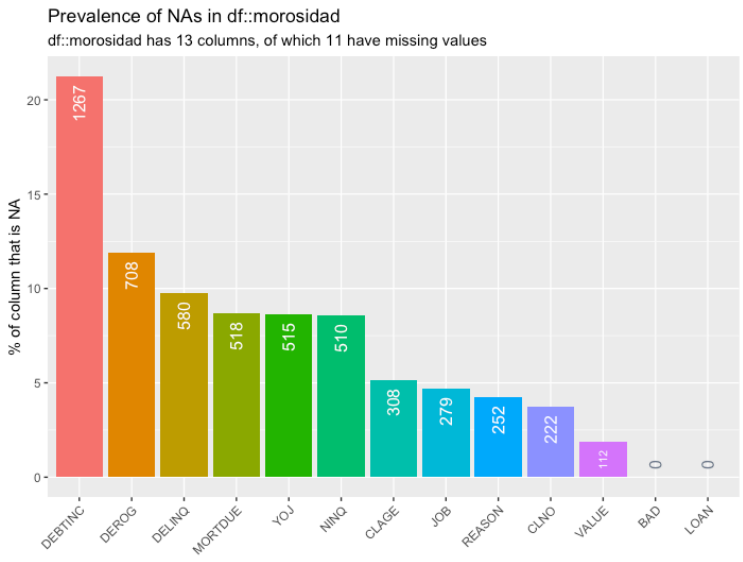
A continuación, un resumen de las variables y sus principales valores.

```
      BAD              LOAN          MORTDUE          VALUE          REASON          JOB           YOJ
 Min.   :0.0000   Min.   : 1100   Min.   :  2063   Min.   :  8000   DebtCon:3928   Mgr    : 767   Min.   : 0.000
 1st Qu.:0.0000   1st Qu.:11100   1st Qu.: 46276   1st Qu.: 66076   HomeImp:1780   Office : 948   1st Qu.: 3.000
 Median :0.0000   Median :16300   Median : 65019   Median : 89236   NA's   : 252   Other  :2388   Median : 7.000
 Mean   :0.1995   Mean   :18608   Mean   : 73761   Mean   :101776                  ProfExe:1276   Mean   : 8.922
 3rd Qu.:0.0000   3rd Qu.:23300   3rd Qu.: 91488   3rd Qu.:119824                  Sales  : 109   3rd Qu.:13.000
 Max.   :1.0000   Max.   :89900   Max.   :399550   Max.   :855909                  Self   : 193   Max.   :41.000
                                  NA's   :518      NA's   :112                     NA's   : 279   NA's   :515
     DEROG            DELINQ           CLAGE            NINQ            CLNO          DEBTINC
 Min.   : 0.0000   Min.   : 0.0000   Min.   :   0.0   Min.   : 0.000   Min.   : 0.0   Min.   :  0.5245
 1st Qu.: 0.0000   1st Qu.: 0.0000   1st Qu.: 115.1   1st Qu.: 0.000   1st Qu.:15.0   1st Qu.: 29.1400
 Median : 0.0000   Median : 0.0000   Median : 173.5   Median : 1.000   Median :20.0   Median : 34.8183
 Mean   : 0.2546   Mean   : 0.4494   Mean   : 179.8   Mean   : 1.186   Mean   :21.3   Mean   : 33.7799
 3rd Qu.: 0.0000   3rd Qu.: 0.0000   3rd Qu.: 231.6   3rd Qu.: 2.000   3rd Qu.:26.0   3rd Qu.: 39.0031
 Max.   :10.0000   Max.   :15.0000   Max.   :1168.2   Max.   :17.000   Max.   :71.0   Max.   :203.3121
 NA's   :708       NA's   :580       NA's   :308      NA's   :510      NA's   :222    NA's   :1267
```

## Análisis Descriptivo

NA

Al cargar la data, lo primero que se hace es revisar la cantidad de valores NA que contiene.
Además se observa que la data tiene celdad vacías, las cuales también son reemplazadas por NA.

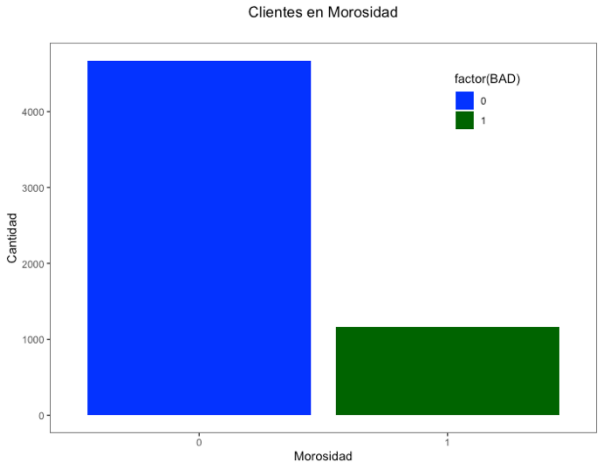



| | Variable | Cant. NA | % NA |
|---|---|---|---|
| **13** | DEBTINC | 1267 | 21.2% |
| **8** | DEROG | 708 | 11.9% |
| **9** | DELINQ | 580 | 9.7% |
| **3** | MORTDUE | 518 | 8.7% |
| **7** | YOJ | 515 | 8.6% |
| **11** | NINQ | 510 | 8.5% |
| **10** | CLAGE | 308 | 5.1% |
| **6** | JOB | 279 | 4.7% |
| **5** | REASON | 252 | 4.2% |
| **12** | CLNO | 222 | 3.7% |
| **4** | VALUE | 112 | 1.9% |
| **1** | BAD | 0 | 0% |
| **2** | LOAN | 0 | 0% |

Dado estos altos valores de Na, se decide hacer dos cosas.
1.- Eliminar aquellas filas que tienen más de un 50% de sus valores NA, ya que se considera que no aportan información.
2.- Para aquellas que tienen menos de un 50% con Na, se hará imputación de datos mediantes la librería MICE.

Luego de esto, el nuevo data set sin valores de Na tiene un total de 5834 observaciones.

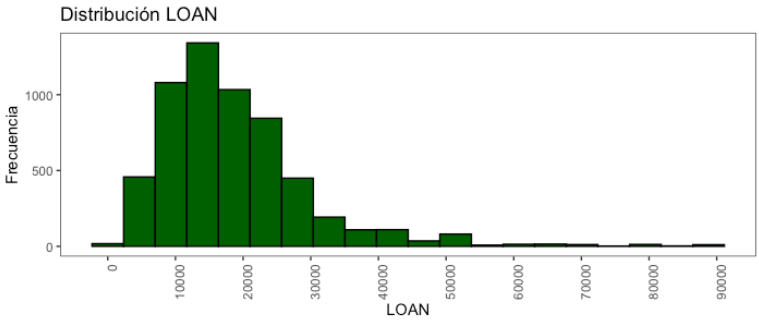
El primer análisis de los valores en el data set, es ver que proporción de estos están clasificados con BAD = 1 y BAD = 0. En total hay 4673 BAD = 1 y 1161 BAD = 0, lo cual representa un 80% y 20% respectivamente.

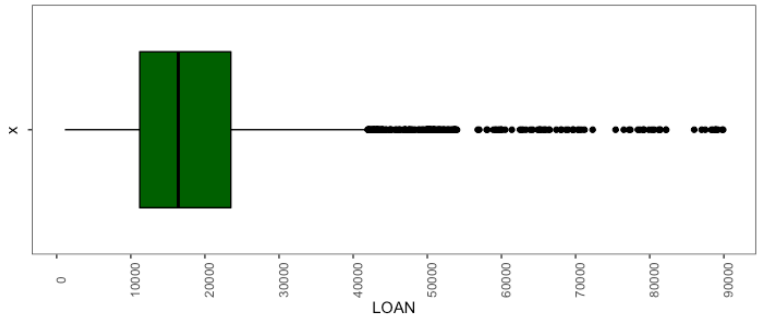


Correlaciones Variables numéricas


De todas las variables numérica, solo MORTDUE y VALUE, se encuentran altamente correlacionadas.

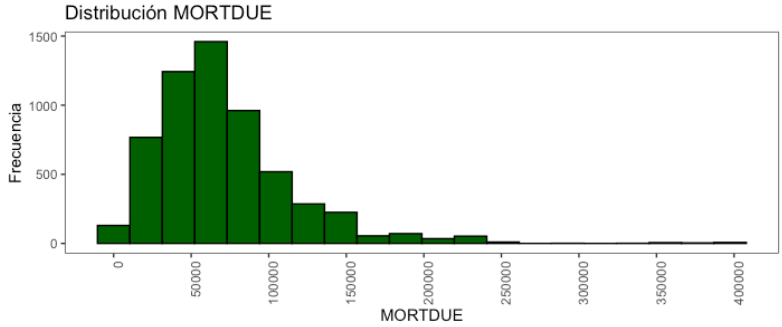
Gráficas Univariadas

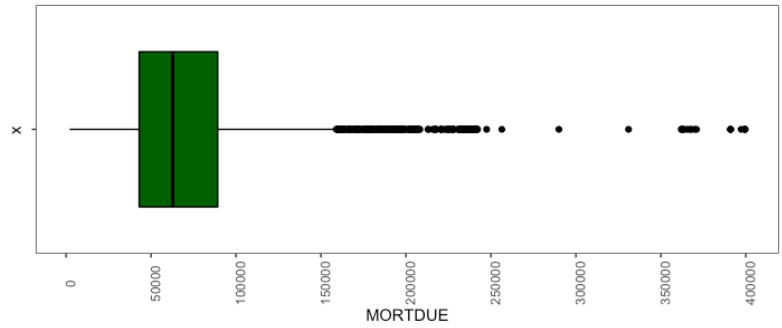



```
Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
1100   11200   16400   18724   23500   89900
```

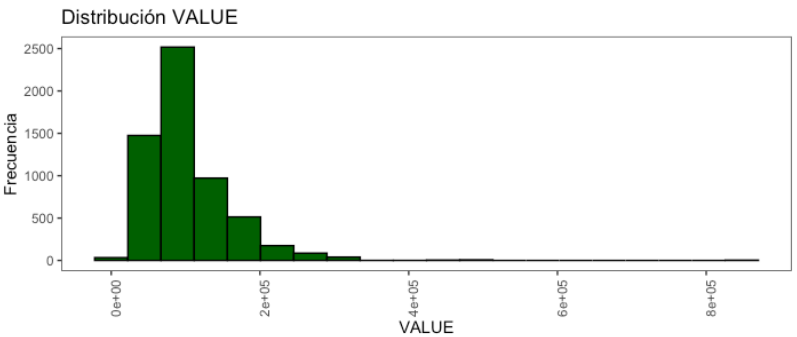
Se observa que LOAN distribuye de forma asimétrica hacia la derecha, y estamos en presencia de valores outliers.
El 50% de los datos se encuentra entre 11200 y 23500, por lo que valores de casi 90 mil, está bastante alejado.

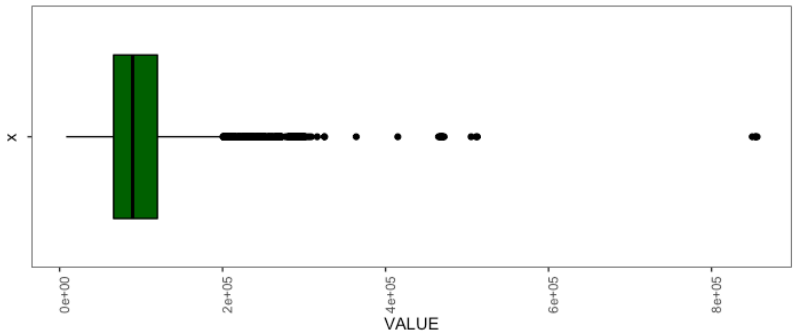



```
Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
2063   43039   62754   71189   89258  399550
```

MORTDUE, tiene una distribución que se asemeja mucho al la de LOAN, lo cual hace sentido, ya que el monto del crédito tiene directa relación con el valor de la hipoteca y valor de la propiedad.
El 50% de los valores están concentrados entre 43mil y 89 mil.

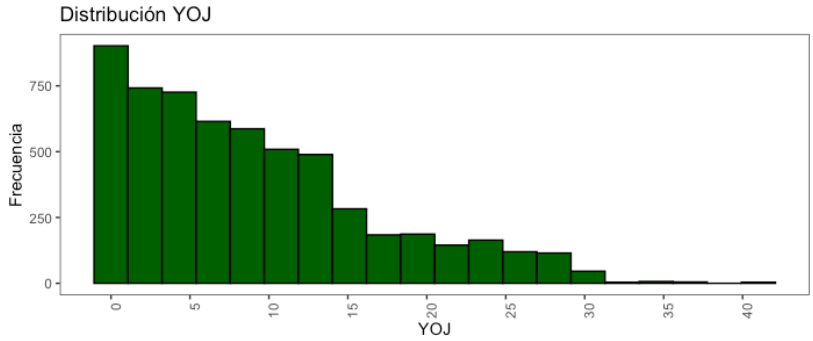


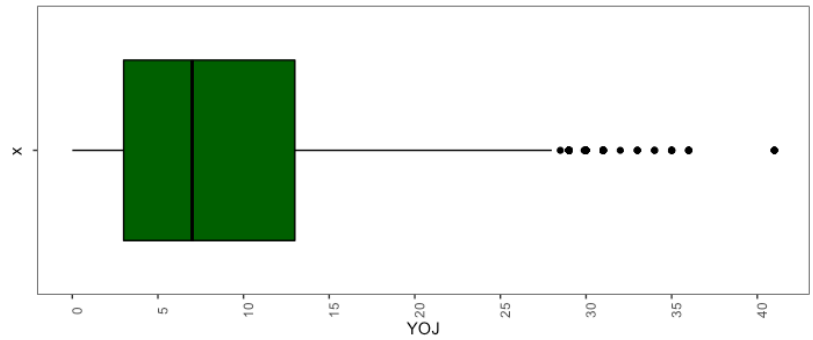




```
Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
8000   66258   89654  102217  120000  855909
```

Se confirma lo observado en las dos variables anteriores, ya que el valor de la propiedad es básicamente el que dicta que tan grande puede ser el monto solicitado, y por ende valor de la hipoteca.

Se observan un par de valores muy aislados de la distribución, cercanos a 850 mil, pero el 75% de los casos totales son menores o iguales que 120 mil.

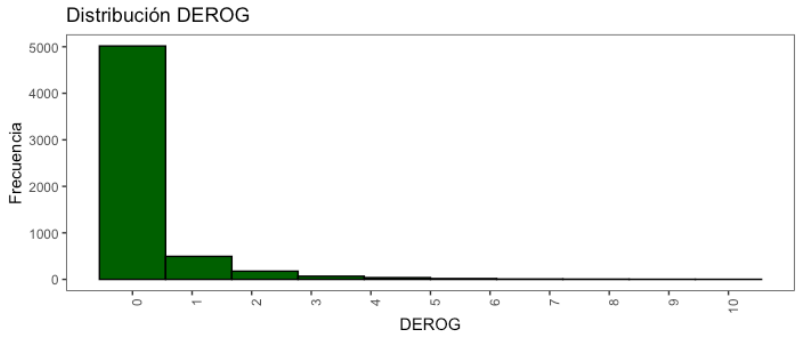


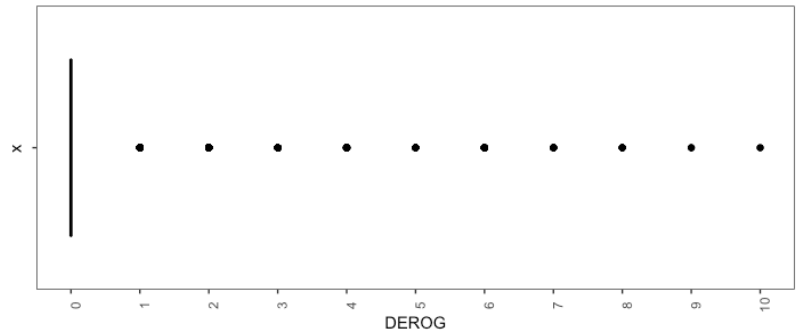




```
 Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
0.000   3.000   7.000   9.025  13.000  41.000
```

Los años en el trabajo actual de los solicitantes de crédito, tienen una distribución que es bastante lógica, partiendo del cero y disminuyendo a medida que avanzan los años.

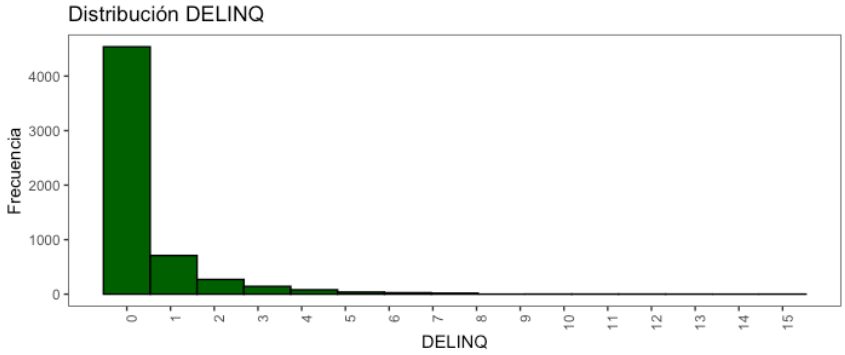
El 50% de los casos, está entre 0 y 7 años, ya después de esos valores se comienza a ver una mayor dispersión, y algunos outliers por sobre los 13 años.

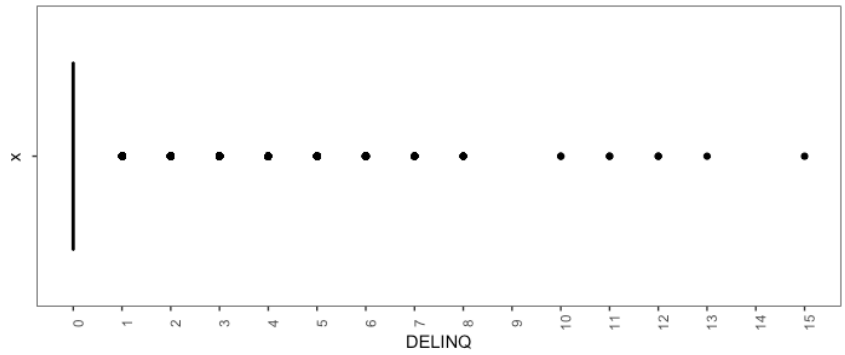



```
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
0.0000  0.0000  0.0000  0.2556  0.0000 10.0000
```

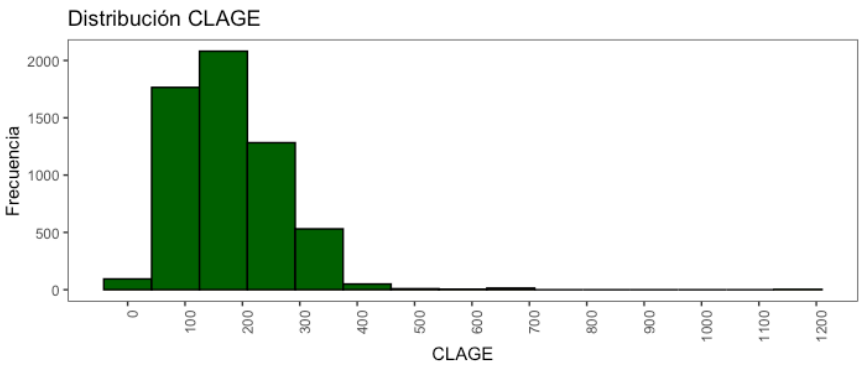
Los reportes de créditos no pagados son más bien casos aislados, ya que al menos un 75% de las personas tiene 0 reportes.
Luego hay personas que tienen entre 1 y 10 reportes.

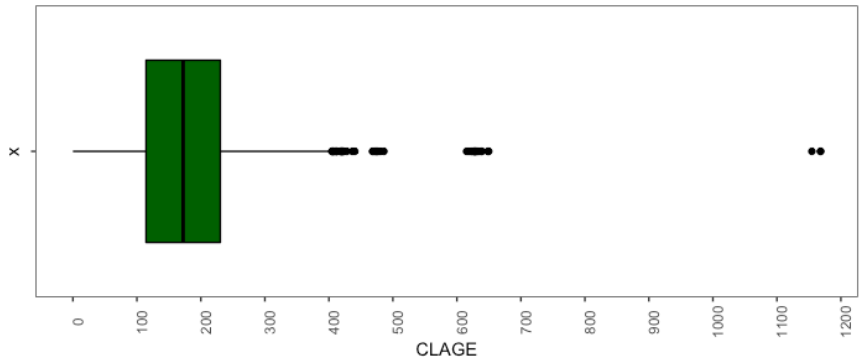



```
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
0.0000  0.0000  0.0000  0.4434  0.0000 15.0000
```

Con DELINQ, pasa exactamente lo mismo que con DEROG, ya que el tercer cuartil sigue siendo cero.
Luego hay personas que tienen entre 1 y 15 líneas de créditos no pagadas.

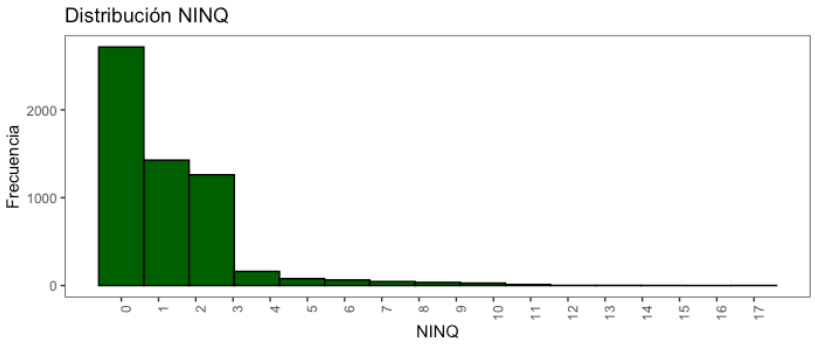


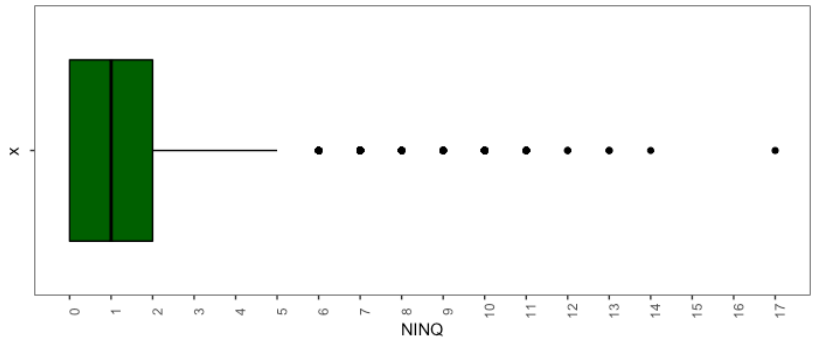




```
Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
 0.0   114.3   172.2   178.9   230.3  1168.2
```

La antigüedad de línea de crédito se concentra entre 114 y 230 meses.

Existen muy pocos valores que se encuentran bastante alejados, sobretodo si se compara la mediana = 172.2 y el máximo valor de 1168.2



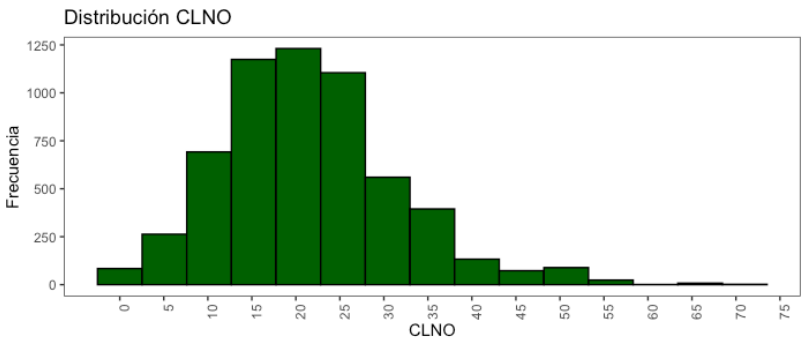



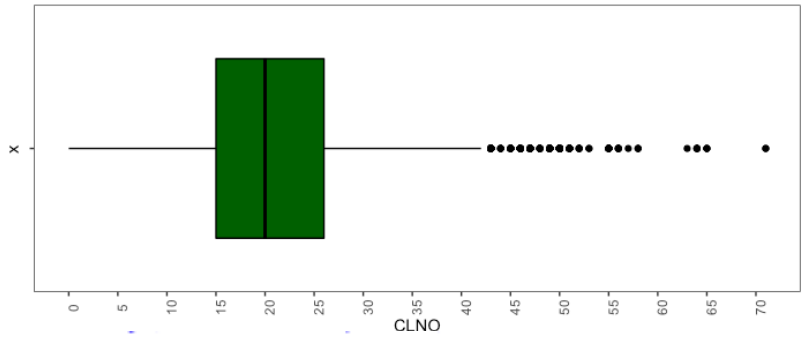

```
Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
0.00    0.00    1.00    1.18    2.00   17.00
```

El 75% de las personas tiene a lo más 2 consultas por crédito recientes.

De hecho, un 25% de los casos, tiene 0 consultas.

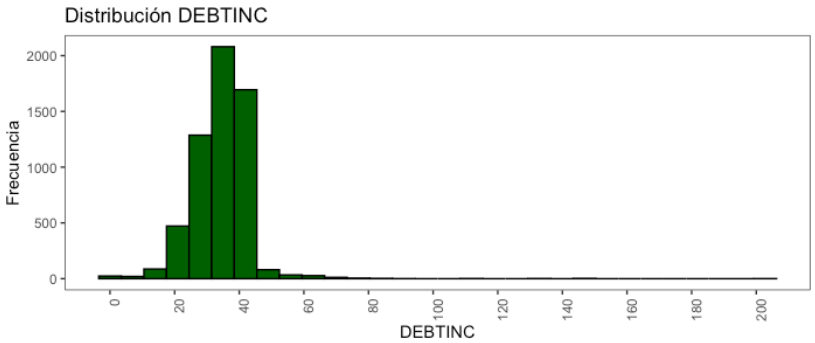
Luego el 25% superior en de distribución se distribuye en valores de entre 2 y 17

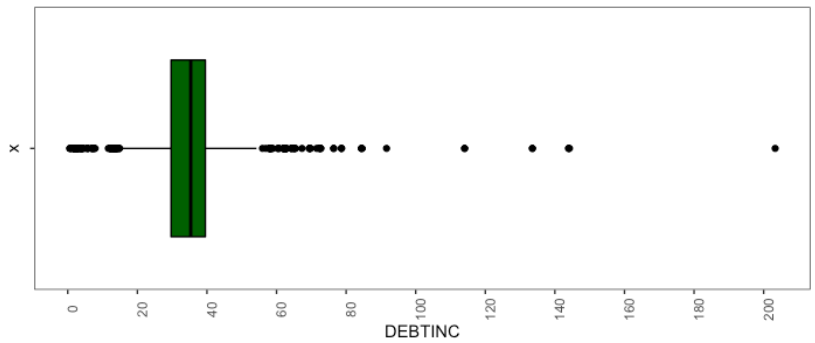



```
Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
 0.0    15.0    20.0    21.3    26.0    71.0
```

La distribución de CLNO, se ve muy similar a una Normal.
La media y mediana son muy cercanas con valores de 20 y 21.3 respectivamente.
El 50% de las observaciones están entre 15 y 26, por ende, se ve que hay una alta concentración.
Nuevamente hay presencia de outliers, los cuales van desde 40 a 71 líneas de crédito aprox.

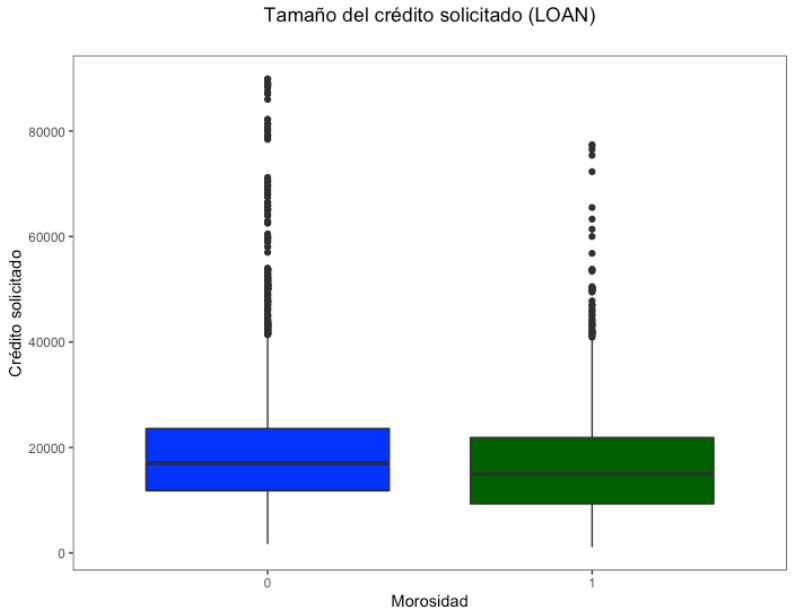



```
   Min.   1st Qu.    Median      Mean   3rd Qu.      Max.
 0.5245   29.6722   35.3376   34.4674   39.5418 203.3122
```

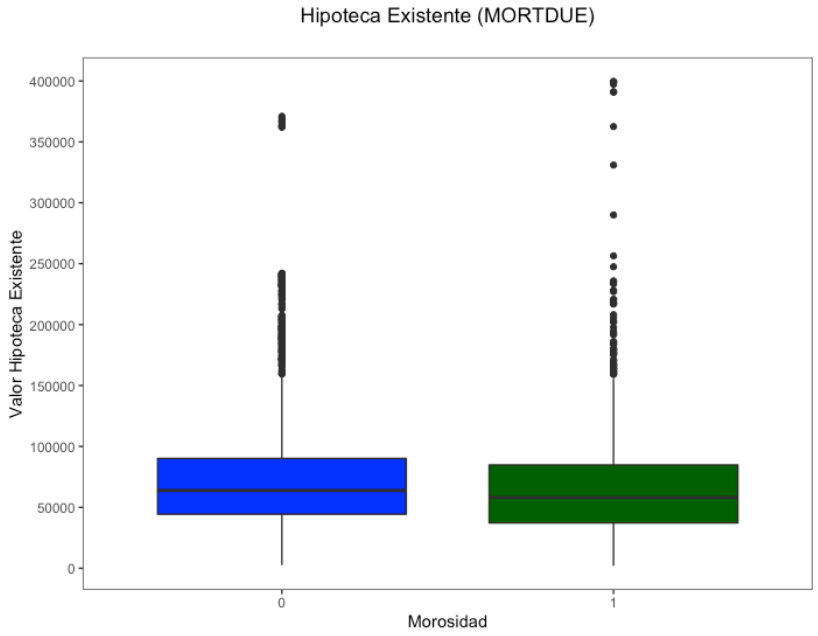
El ratio de deuda/ingreso, está muy concentrado en valores entre 29 y 39.
En este caso, se observa valores outliers que están hacia la izquierda de la distribución, con valores incluso menores a 1, lo cual puede ser indicador de personas que tiene un record de deuda sana respecto de sus ingresos.
Al igual que con las demás variables, hay outliers hacia la derecha con personas que superan incluso un ratio de 100.

Visualización de la relación de BAD con la demás variables




| Min. | 1st Qu. | Median | Mean | 3rd Qu. | Max. |
|---|---|---|---|---|---|
| 1700 | 11800 | 17000 | 19144 | 23600 | 89900 |

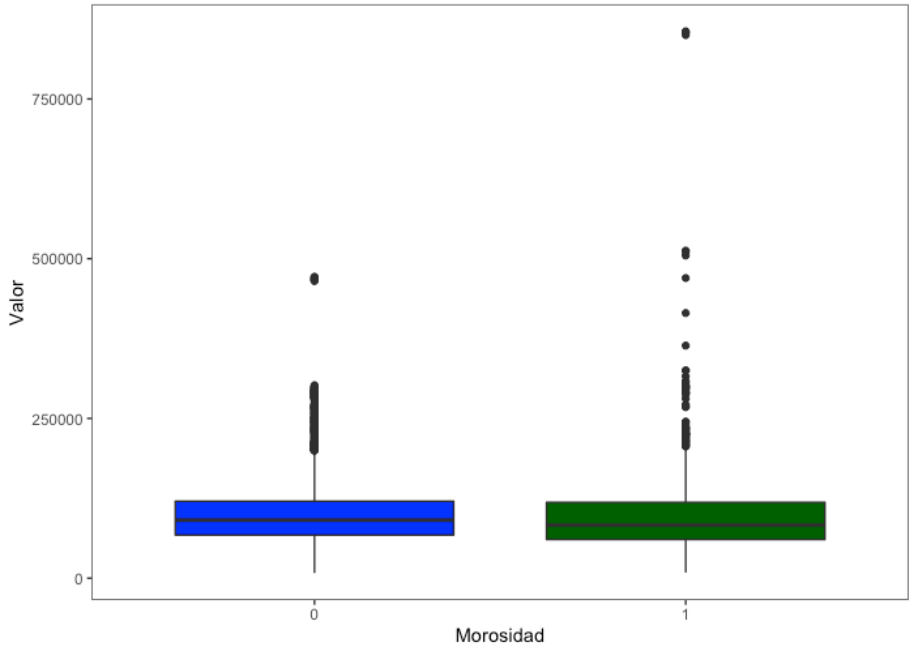
| Min. | 1st Qu. | Median | Mean | 3rd Qu. | Max. |
|---|---|---|---|---|---|
| 1100 | 9300 | 15000 | 17036 | 21900 | 77400 |




| Min. | 1st Qu. | Median | Mean | 3rd Qu. | Max. |
|---|---|---|---|---|---|
| 2619 | 44355 | 63833 | 71867 | 90154 | 371003 |

| Min. | 1st Qu. | Median | Mean | 3rd Qu. | Max. |
|---|---|---|---|---|---|
| 2063 | 37200 | 58204 | 68459 | 84843 | 399550 |

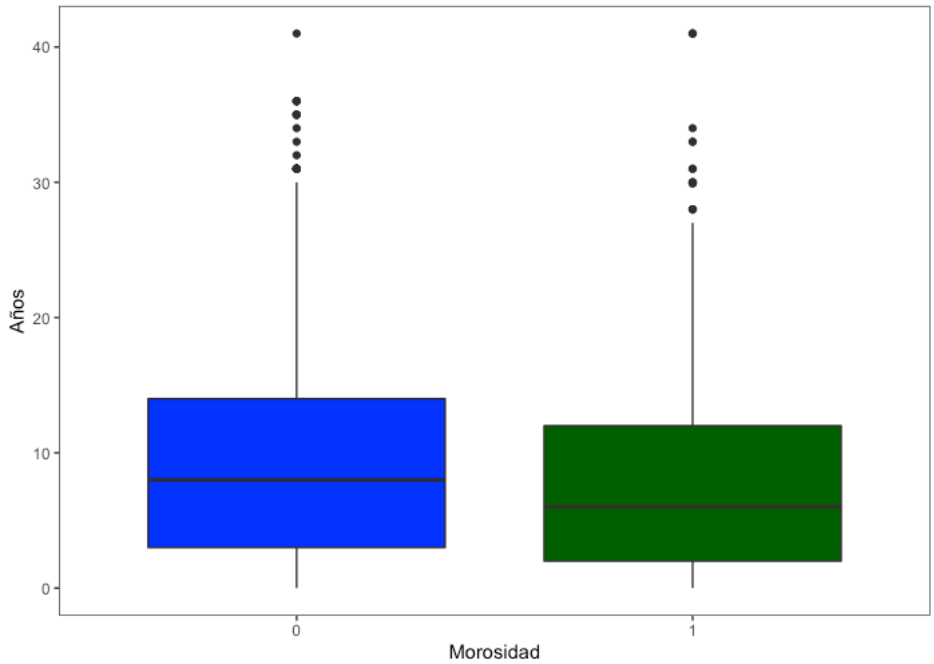
Valor propuedad actual (VALUE)




| Min. | 1st Qu. | Median | Mean | 3rd Qu. | Max. |
|---|---|---|---|---|---|
| 8000 | 67446 | 90883 | 102896 | 120451 | 471827 |

| Min. | 1st Qu. | Median | Mean | 3rd Qu. | Max. |
|---|---|---|---|---|---|
| 8800 | 60493 | 83000 | 99487 | 118734 | 855909 |

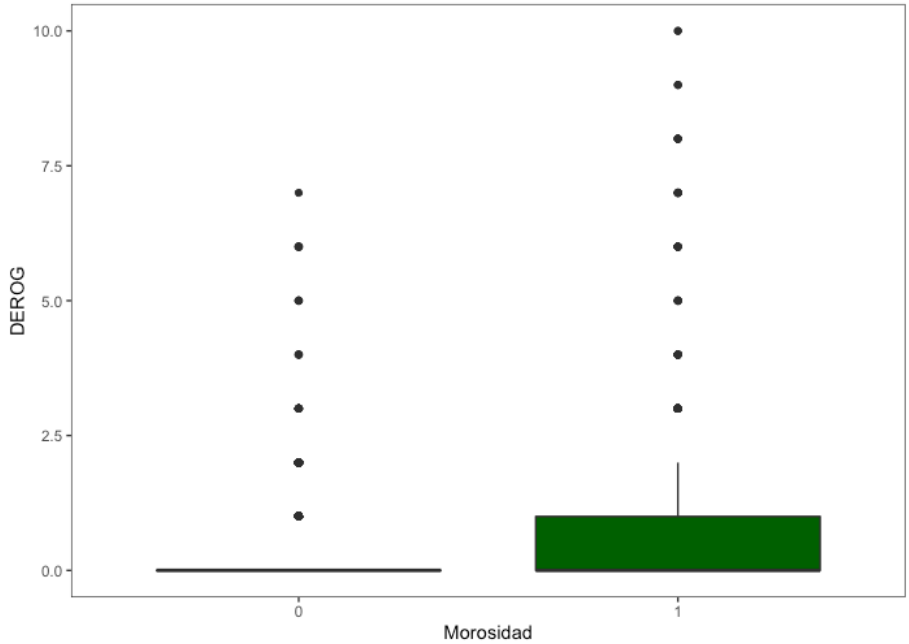
Años en trabajo actual (YOJ)




| Min. | 1st Qu. | Median | Mean | 3rd Qu. | Max. |
|---|---|---|---|---|---|
| 0.000 | 3.000 | 8.000 | 9.265 | 14.000 | 41.000 |

| Min. | 1st Qu. | Median | Mean | 3rd Qu. | Max. |
|---|---|---|---|---|---|
| 0.000 | 2.000 | 6.000 | 8.056 | 12.000 | 41.000 |

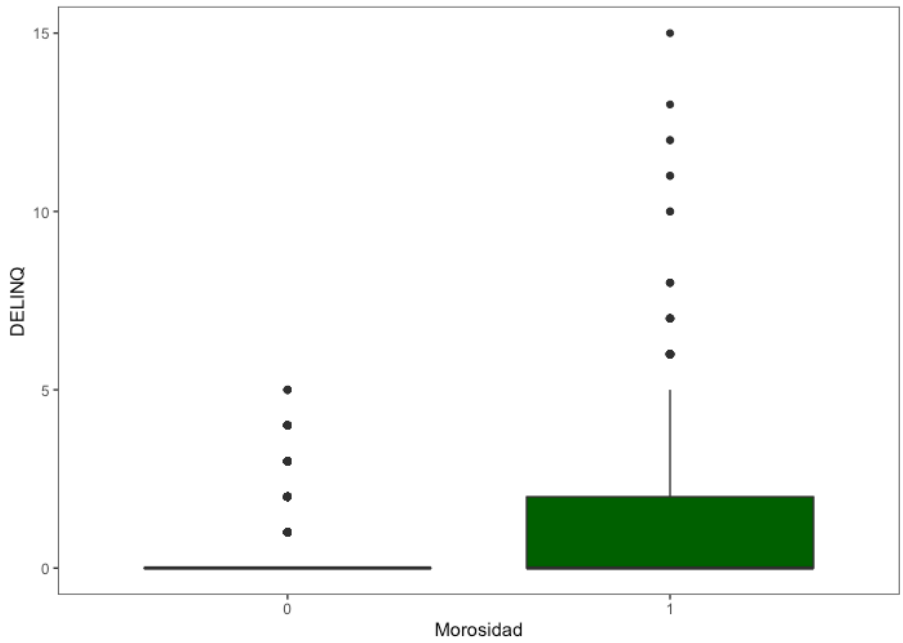
Reportes de créditos no pagados (DEROG)




| Min. | 1st Qu. | Median | Mean | 3rd Qu. | Max. |
|---|---|---|---|---|---|
| 0.0000 | 0.0000 | 0.0000 | 0.1457 | 0.0000 | 7.0000 |

| Min. | 1st Qu. | Median | Mean | 3rd Qu. | Max. |
|---|---|---|---|---|---|
| 0.0000 | 0.0000 | 0.0000 | 0.6977 | 1.0000 | 10.0000 |

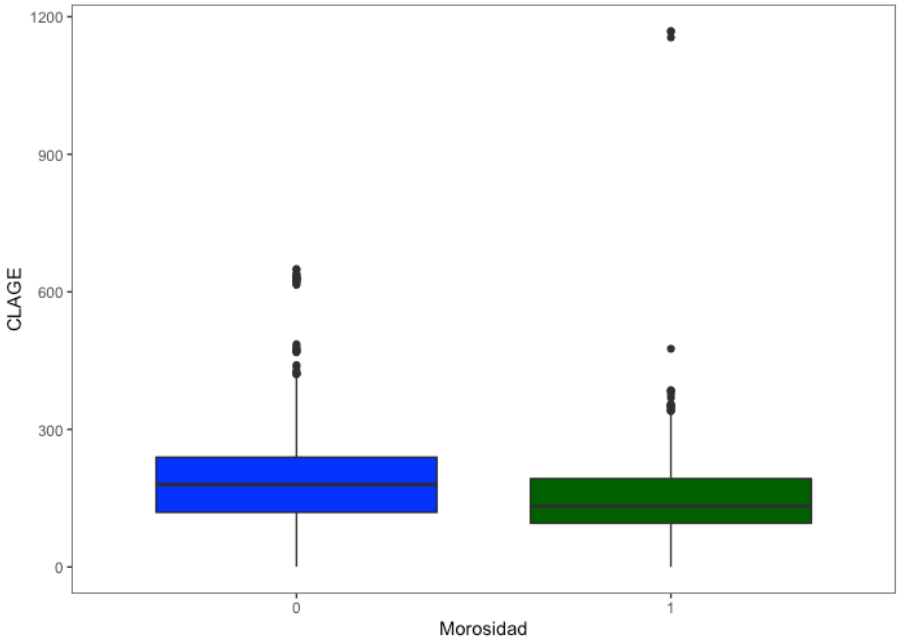
Líneas de créditos no pagados (DELINQ)




| Min. | 1st Qu. | Median | Mean | 3rd Qu. | Max. |
|---|---|---|---|---|---|
| 0.000 | 0.000 | 0.000 | 0.247 | 0.000 | 5.000 |

| Min. | 1st Qu. | Median | Mean | 3rd Qu. | Max. |
|---|---|---|---|---|---|
| 0.000 | 0.000 | 0.000 | 1.234 | 2.000 | 15.000 |

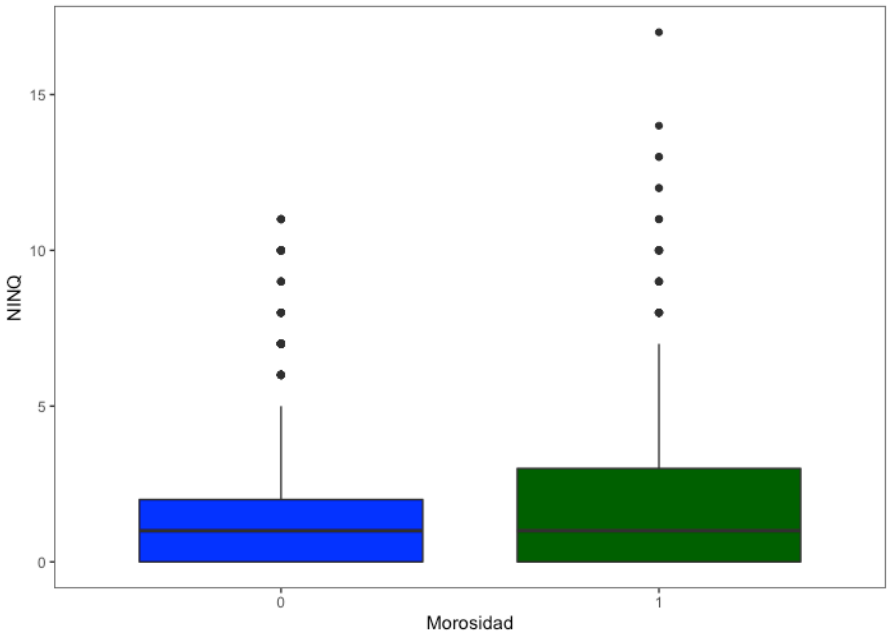
Antiguedad en meses de la
línea de crédito más antigua (CLAGE)




| Min. | 1st Qu. | Median | Mean | 3rd Qu. | Max. |
|---|---|---|---|---|---|
| 0.4867 | 119.0831 | 179.8333 | 186.0801 | 239.4333 | 649.7471 |

| Min. | 1st Qu. | Median | Mean | 3rd Qu. | Max. |
|---|---|---|---|---|---|
| 0.00 | 95.37 | 132.16 | 149.98 | 192.67 | 1168.23 |

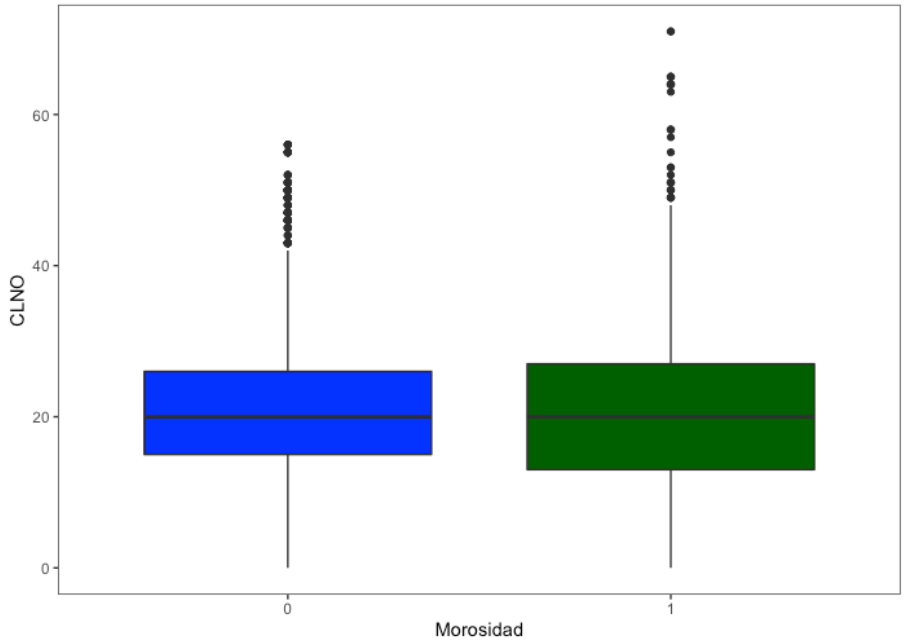
Número de consultas por créditos recientes (NINQ)




| Min. | 1st Qu. | Median | Mean | 3rd Qu. | Max. |
|---|---|---|---|---|---|
| 0.000 | 0.000 | 1.000 | 1.034 | 2.000 | 11.000 |

| Min. | 1st Qu. | Median | Mean | 3rd Qu. | Max. |
|---|---|---|---|---|---|
| 0.000 | 0.000 | 1.000 | 1.765 | 3.000 | 17.000 |

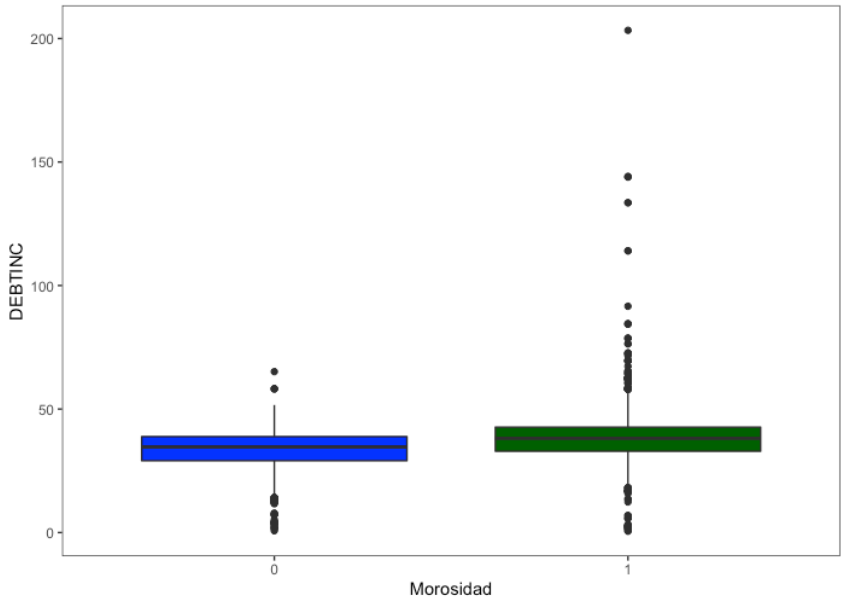
Número de líneas de crédito (CLNO)




| Min. | 1st Qu. | Median | Mean | 3rd Qu. | Max. |
|---|---|---|---|---|---|
| 0.00 | 15.00 | 20.00 | 21.32 | 26.00 | 56.00 |

| Min. | 1st Qu. | Median | Mean | 3rd Qu. | Max. |
|---|---|---|---|---|---|
| 0.00 | 13.00 | 20.00 | 21.19 | 27.00 | 71.00 |

Ratio Deuda/Ingreso (DEBTINC)




| Min. | 1st Qu. | Median | Mean | 3rd Qu. | Max. |
|---|---|---|---|---|---|
| 0.7203 | 29.0641 | 34.6449 | 33.4063 | 38.8476 | 65.1436 |

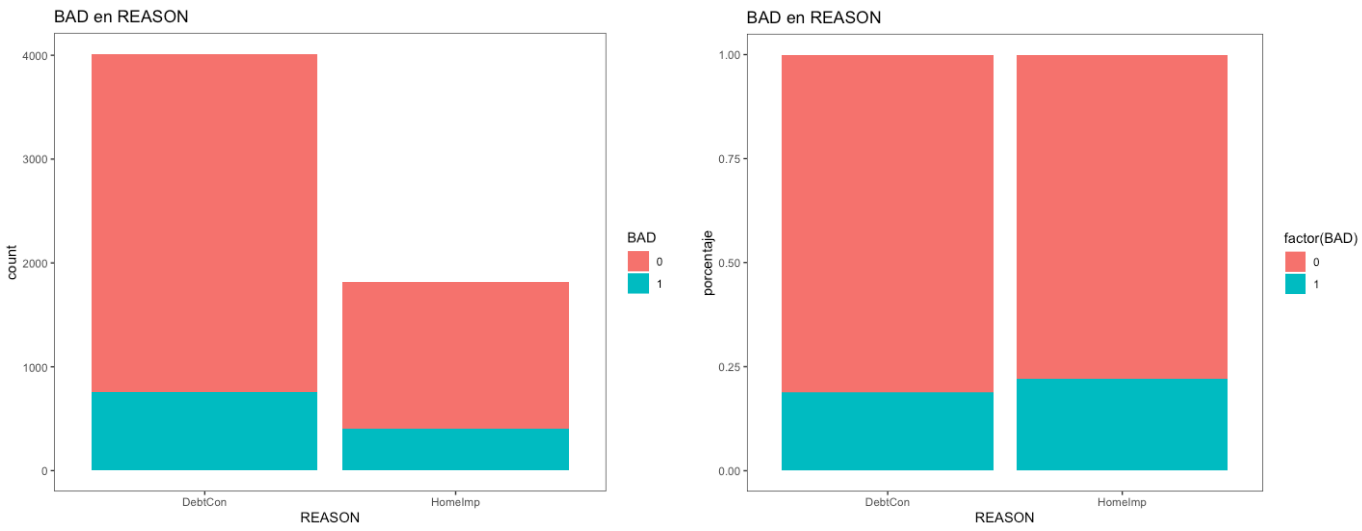
| Min. | 1st Qu. | Median | Mean | 3rd Qu. | Max. |
|---|---|---|---|---|---|
| 0.5245 | 32.9088 | 38.1390 | 38.7382 | 42.6667 | 203.3122 |

Con respecto a las variables numéricas, la mayoría no tiene al menos visualmente una diferencia que sea muy significativa a la hora de explicar a la variable dependiente BAD.

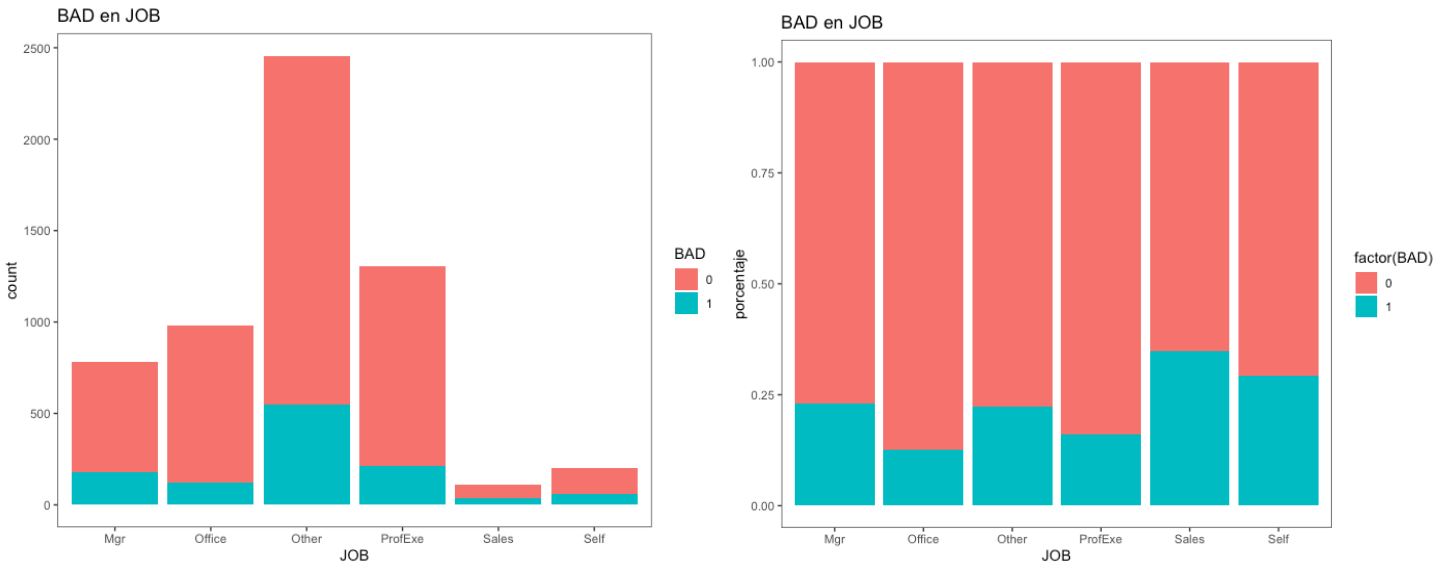
Las que destacan son:

**DEROG**, donde se ve que para quienes incurrieron en falta (BAD=1), hay mayor cantidad de observaciones de Reportes de créditos no pagados y con valores más altos.
**DELINQ**, es mismo caso que DEROG, pero relacionado a líneas de créditos no pagadas, lo cual hace sentido ya que una persona que no ha pagado sus obligaciones, debiese tender a una mayor probabilidad de no pago en el crédito.
**CLAGE**, la antiguedad en meses de línea de crédito más antigua, se observa que tiene un efecto inverso, ya que posiblemente quienes tengas líneas hace más tiempo, tiendan a tener una mayor estabilidad económica.
**NINQ**, se observa que el 75% de las personas, están concentradas en mayor número de solicitudes de crédito, en el caso de quienes no pagaron crédito, lo cual debe tender a quienes poseen mayor número de deudas caen en no pagar ciertas obligaciones.
**YOJ**, se observa que quienes no pagaron, tiene menor cantidad de años en su trabajo actual, lo cual se relaciona igualmente con la estabilidad económica que pueda tener ese individuo.

Ahora, se muestran gráficamente las variables categóricas




Estos gráficos muestran la variable REASON, la cual indica el motivo por el cual se solicitó el crédito, donde predominan quienes toman la deuda para consolidar otras deudas, versus quienes lo solicitan para hacer mejoras en el hogar.
Por otro lado, se observa que la proporción de quienes no pagaron el crédito en el grupo "HomeImp", es un poco mayor que en el otro, con respecto del total de la categoría.




Por último, la variable JOB, que indica la categoría de tipo de trabajo que tiene el solicitante del crédito, lo que más destaca, es que quienes tienen un trabajo en ventas, "sales", tienen mayor probabilidad de caer en imcumplimiento de pago, ya que versus las otras categorías, tienen mayor concentración de BAD=1. Tal vez esto se pudiese explicar, por que en ventas los sueldos por lo general son variables y eso puede afectar en que no todos los meses tendrá la misma capacidad de pago.

## Elección del Modelo a desarrollar

El modelo a desarrollar será el de Regresión Logística Binario, ya que lo que se quiere lograr es clasificar a los clientes entre dos clases, 1 o 0, siendo 1 un cliente altamente probable de caer en incumplimiento de pago y 0, un cliente que probablemente va a pagar la deuda y es buen negocio para el banco.

## Construcción del Modelo

Lo Primero, se hace la división de la data en entrenamiento y validación, en una proporción 80/20.

I.- Modelos Univariados

Se hace el desarrollo de Modelos Univariados, y para todos, la comparación de valores AUC y KS en muestras de entrenamiento y validación.

| Variable | AUC Train | AUC Test | KS Train | KS Test |
|---|---|---|---|---|
| **DELINQ** | 0.67943570965683 | 0.65015136971129 | 0.3738 | 0.3115 |
| **DEBTINC** | 0.654978176385935 | 0.671779240197889 | 0.2407 | 0.2822 |
| **CLAGE** | 0.637219590094342 | 0.624748486302887 | 0.2221 | 0.1916 |
| **DEROG** | 0.618089200873693 | 0.58775289817618 | 0.2864 | 0.2592 |
| **NINQ** | 0.597319413588075 | 0.608674684338773 | 0.1938 | 0.2122 |
| **LOAN** | 0.575790456454828 | 0.589421195451524 | 0.1238 | 0.1638 |
| **JOB** | 0.575370854614512 | 0.590048825961751 | 0.171 | 0.1823 |
| **VALUE** | 0.548204118518016 | 0.554017758251496 | 0.089 | 0.1046 |
| **YOJ** | 0.54233991290131 | 0.546760780476999 | 0.0944 | 0.1031 |
| **MORTDUE** | 0.541906924106909 | 0.553048622904823 | 0.0715 | 0.1046 |
| **REASON** | 0.514268477811196 | 0.550510411282581 | 0.0794 | 0.1662 |
| **CLNO** | 0.512125845429334 | 0.509737502768958 | 0.0753 | 0.0738 |

Observamos que las variables con mayor AUC son DELINQ y DEBTINC.
Además se se que todos los modelos univariables tienen un AUC superior a 0.5 lo cual es un buen indicador, los KS son bajos, ya que una sola variables está permitiendo hacer una buena separación entre ambas clases.

II.- Modelo Multivariado

Desarrollo del modelo logístico con todas las variables. A continuación resumen del modelo.

```
Coefficients:
                Estimate Std. Error z value Pr(>|z|)
(Intercept)   -3.379e+00  2.837e-01 -11.911  < 2e-16 ***
LOAN          -2.438e-05  5.019e-06  -4.857 1.19e-06 ***
MORTDUE       -8.165e-06  2.287e-06  -3.570 0.000357 ***
VALUE          6.830e-06  1.637e-06   4.173 3.01e-05 ***
REASONHomeImp  1.926e-01  9.867e-02   1.952 0.050933 .
JOBOffice     -6.355e-01  1.709e-01  -3.718 0.000201 ***
JOBOther       4.715e-02  1.334e-01   0.353 0.723737
JOBProfExe     5.539e-02  1.541e-01   0.359 0.719292
JOBSales       8.089e-01  3.098e-01   2.611 0.009028 **
JOBSelf        3.457e-01  2.552e-01   1.355 0.175573
YOJ           -9.227e-03  6.375e-03  -1.447 0.147769
DEROG          5.170e-01  5.217e-02   9.909  < 2e-16 ***
DELINQ         7.341e-01  4.247e-02  17.286  < 2e-16 ***
CLAGE         -5.748e-03  6.259e-04  -9.184  < 2e-16 ***
NINQ           1.424e-01  2.338e-02   6.093 1.11e-09 ***
CLNO          -1.637e-02  4.983e-03  -3.285 0.001021 **
DEBTINC        8.030e-02  6.439e-03  12.471  < 2e-16 ***
```

Se observa que la variable YOJ no es significativa, además hay algunas categorías dentro de REASON y JOB que no son significativas respecto de la categoría de referencia.

Con respecto a los coeficientes, los que tienen un valor positivo, indican que, a mayor valor, la probabilidad de que No se otorgue un crédito es mayor, y viceversa con quieres tienen signo negativo.

III.- Selección de Variables

Se utiliza el método Step para hacer selección de variables del modelo. En este caso, se hará, proceso Forward, Backward y Stepwise, y se hará comparación para ver que modelo entrega mejor resultado o en caso de ser similar, quien lo hace con menor número de variables.

Forward (AIC = 3467)

```
glm(formula = BAD ~ DELINQ + DEBTINC + CLAGE + DEROG + JOB +
   LOAN + NINQ + CLNO + REASON + VALUE + MORTDUE, family = binomial,
   data = data_train)
```

Backward (AIC = 3467)

```
glm(formula = BAD ~ LOAN + MORTDUE + VALUE + REASON + JOB + DEROG +
   DELINQ + CLAGE + NINQ + CLNO + DEBTINC, family = binomial,
   data = data_train)
```

Stepwise (AIC = 3467)

```
glm(formula = BAD ~ DELINQ + DEBTINC + CLAGE + DEROG + JOB +
   LOAN + NINQ + CLNO + REASON + VALUE + MORTDUE, family = binomial,
   data = data_train)
```

Los 3 métodos entregaron mismo resultado, dejando fuera solo la variable YOJ, similar a lo visto en el modelo multivariado que indicaba esa variable como no significativa. Más abajo el modelo final y el resumen donde se ven que todas las variables quedan como significativas.

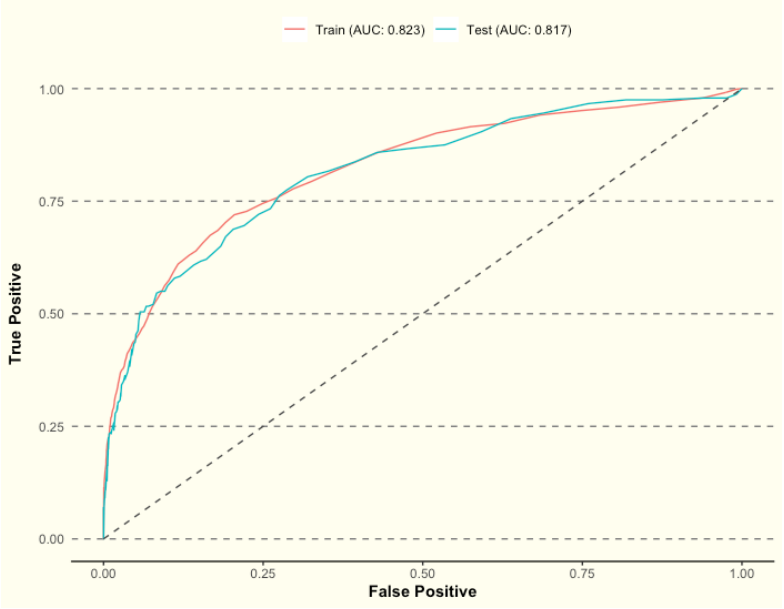
Modelo Final

glm(formula = BAD ~ DELINQ + DEBTINC + CLAGE + DEROG + JOB + LOAN + NINQ + CLNO + REASON + VALUE + MORTDUE , family = binomial, data = data_train)

```
Coefficients:
               Estimate Std. Error z value Pr(>|z|)
(Intercept)   -3.418e+00  2.808e-01 -12.175  < 2e-16 ***
DELINQ         7.520e-01  4.316e-02  17.424  < 2e-16 ***
DEBTINC        8.150e-02  6.342e-03  12.851  < 2e-16 ***
CLAGE         -6.461e-03  6.309e-04 -10.241  < 2e-16 ***
DEROG          5.103e-01  5.100e-02  10.006  < 2e-16 ***
JOBOffice     -5.365e-01  1.704e-01  -3.149 0.001637 **
JOBOther       1.603e-03  1.337e-01   0.012 0.990430
JOBProfExe     1.453e-01  1.533e-01   0.948 0.343237
JOBSales       1.019e+00  3.016e-01   3.377 0.000732 ***
JOBSelf        6.377e-01  2.488e-01   2.563 0.010364 *
LOAN          -2.681e-05  5.042e-06  -5.318 1.05e-07 ***
NINQ           1.445e-01  2.318e-02   6.234 4.54e-10 ***
CLNO          -1.847e-02  4.972e-03  -3.715 0.000203 ***
REASONHomeImp  2.389e-01  9.837e-02   2.428 0.015178 *
VALUE          6.158e-06  1.645e-06   3.743 0.000182 ***
MORTDUE       -6.302e-06  2.260e-06  -2.789 0.005294 **
```

| AUC Train | AUC Test | KS Train | KS Test |
|---|---|---|---|
| 0.8231358 | 0.8162257 | 0.5175 | 0.4866 |

Los valores de AUC y Ks son aceptables, el modelo estaría haciendo una buena clasificación, aunque idealmente sería tener un valor de KS superior a 0.5, ya que ahorá en la muestra de test está un poco por debajo, lo cual sigue siendo tolerable para este modelo.



Matriz de confusión Modelo Final, Entrenamiento y Test.

```
          Reference
Prediction    0    1
         0 3645  102
         1  578  343

               Accuracy : 0.8543
                 95% CI : (0.8439, 0.8643)
    No Information Rate : 0.9047
    P-Value [Acc > NIR] : 1

                  Kappa : 0.4288

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

            Sensitivity : 0.8631
            Specificity : 0.7708
         Pos Pred Value : 0.9728
         Neg Pred Value : 0.3724
             Prevalence : 0.9047
         Detection Rate : 0.7808
   Detection Prevalence : 0.8027
      Balanced Accuracy : 0.8170

       'Positive' Class : 0
```

```
          Reference
Prediction   0   1
         0 892  34
         1 153  87

               Accuracy : 0.8396
                 95% CI : (0.8173, 0.8602)
    No Information Rate : 0.8962
    P-Value [Acc > NIR] : 1

                  Kappa : 0.3991

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

            Sensitivity : 0.8536
            Specificity : 0.7190
         Pos Pred Value : 0.9633
         Neg Pred Value : 0.3625
             Prevalence : 0.8962
         Detection Rate : 0.7650
   Detection Prevalence : 0.7942
      Balanced Accuracy : 0.7863

       'Positive' Class : 0
```

El modelo muestra una buena precisión de 85% y 84% en cada muestra. La sensitividad muestra valores muy similares de 86% y 85%. El problema se observa en la especificidad, donde se ve que el modelo está teniendo algunos problemas para clasificar a los clientes con probabilidad de incumplimiento.

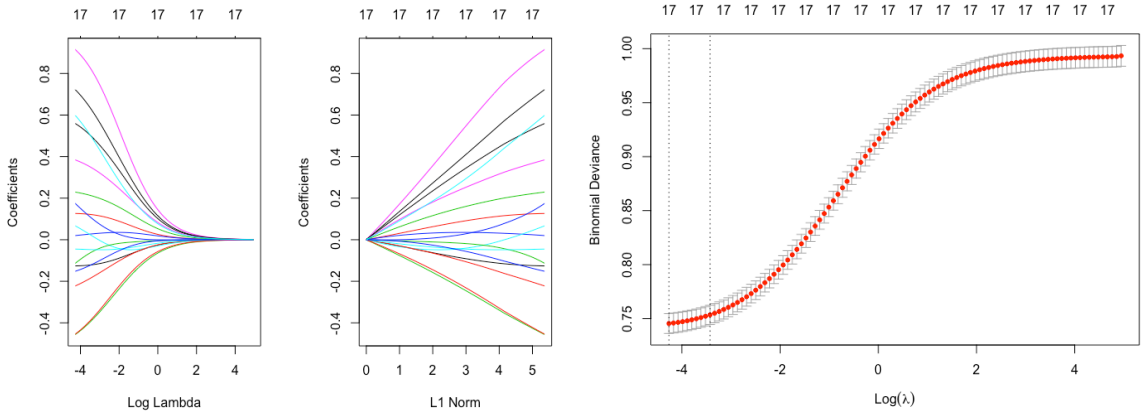
Interpretación de algunos coeficientes del modelo:

DELIQ
OR = 2.121238, por cada línea de crédito no pagada que tenga el cliente, este tendrá 2.12 veces más probabilidad de que NO le den el crédito
DEBTINC
OR = 1.084913, por cada unidad que aumente en su ratio de deuda/ingreso, el cliente tiene 1 vez más de probabilidad de que NO le den el crédito
CLAGE
OR = 0.9935598, este or al ser un valor entre 0 y 1 se calcula en 1-0.9935598 = 0.0064402. lo cual indica que a mayor antiguedad de la línea de crédito más antigua en meses, tiene un 0.6% más de propabilidad de que si le den el crédito
DEROG
OR = 1.665791, por cada reporte de crédito no pagado, tiene 1.66 veces más opciones de que que NO le otorguen el crédito

IV.- Regularización

Siguiendo con el análisis, se utilizó también modelos de regularización, para ver si es posible llegar a un modelo que ajuste mejor y tenga mayor poder de clasificación mediante la penalización de los coeficientes.
Lo primero acá es transformar las variables categóricas a variables dummy, escalar las que son numéricas y llevarlas a un formato de matriz.

- Regresión Logística con Ridge (alpha = 0)
Visualización de los coeficientes de todas las variables del modelo, para los distintos valores de lambda. Luego mediante validación cruzada, se busca el valor de lambda que minimiza el error de testeo.




Coeficientes del Modelo Ridge

| | |
|---|---|
| (Intercept) | -1.64665875 |
| REASONDebtCon | -0.12607541 |
| REASONHomeImp | 0.12622134 |
| JOBOffice | -0.45606374 |
| JOBOther | 0.01994174 |
| JOBProfExe | 0.06658683 |
| JOBSales | 0.91485874 |
| JOBSelf | 0.55903712 |
| LOAN | -0.22244738 |
| MORTDUE | -0.11335099 |
| VALUE | 0.17414364 |
| YOJ | -0.04507260 |
| DEROG | 0.38392832 |
| DELINQ | 0.72209879 |
| CLAGE | -0.45406078 |
| NINQ | 0.22918997 |
| CLNO | -0.15181387 |
| DEBTINC | 0.59828786 |

Este modelo, solo penaliza los coeficientes, pero no lleva a cero sus valores, por lo que este conserva todas las variables y aquellas que tienen menor poder predictivo, aparecen con coeficientes mucho menores. Por ejemplo, YOJ que anteriormente se encontró como no significativa, ahora tiene un coeficiente de -0.04, lo cual indica que los años en el trabajo actual en muy baja medida afectan en la probabilidad de que un cliente tenga incumplimiento

Matriz de confusión Modelo Ridge, Entrenamiento y Test

```
          Reference
Prediction    0    1
         0 3664  617
         1   83  304

               Accuracy : 0.85
                 95% CI : (0.8395, 0.8602)
    No Information Rate : 0.8027
    P-Value [Acc > NIR] : < 2.2e-16

                  Kappa : 0.3941

 Mcnemar's Test P-Value : < 2.2e-16

            Sensitivity : 0.9778
            Specificity : 0.3301
         Pos Pred Value : 0.8559
         Neg Pred Value : 0.7855
             Prevalence : 0.8027
         Detection Rate : 0.7849
   Detection Prevalence : 0.9171
      Balanced Accuracy : 0.6540

       'Positive' Class : 0
```

```
          Reference
Prediction   0   1
         0 900 160
         1  26  80

               Accuracy : 0.8405
                 95% CI : (0.8182, 0.861)
    No Information Rate : 0.7942
    P-Value [Acc > NIR] : 3.341e-05

                  Kappa : 0.3848

 Mcnemar's Test P-Value : < 2.2e-16

            Sensitivity : 0.9719
            Specificity : 0.3333
         Pos Pred Value : 0.8491
         Neg Pred Value : 0.7547
             Prevalence : 0.7942
         Detection Rate : 0.7719
   Detection Prevalence : 0.9091
      Balanced Accuracy : 0.6526

       'Positive' Class : 0
```

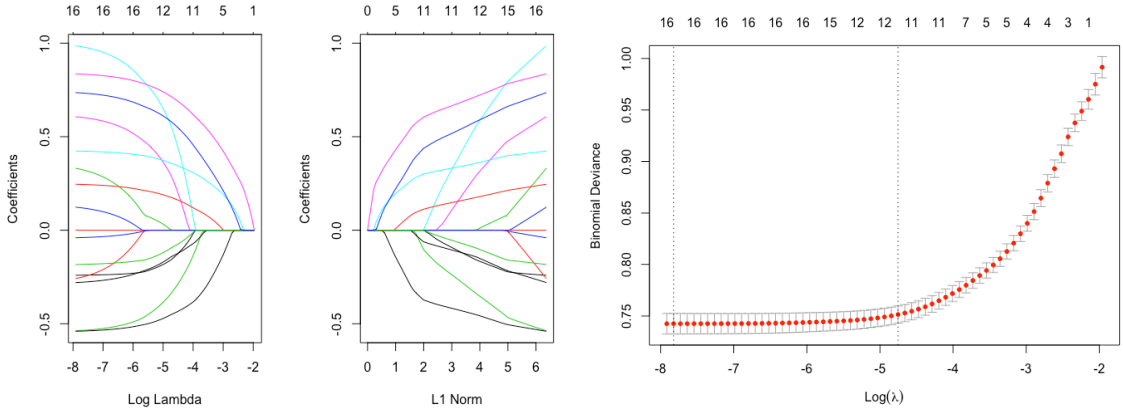
Con respecto al modelo anterior,la sensitividad aumenta mucho, teniendo para ambas muestras un valor de 97%, pero la especificidad se ve muy disminuída, con un valor de 33%, por lo cual el modelo no está pudiendo diferenciar entre clases. El. accuracy se mantiene, pero es por la capacidad de clasificar correctamente los valores positivos solamente.

| AUC Train | AUC Test | KS Train | KS Test |
|---|---|---|---|
| 0.8242045 | 0.8159017 | 0.5094 | 0.4866 |

Los valores de AUC y KS para ambas muestras, prácticamente no varían comparado con el modelo logístico analizado antes.

- Regresión Logística con Lasso (alpha = 1)

Visualización de los coeficientes de todas las variables del modelo, para los distintos valores de lambda. Luego mediante validación cruzada, se busca el valor de lambda que minimiza el error de testeo.




Coeficiente del modelo Lasso

| | |
|---|---|
| (Intercept) | -1.596356e+00 |
| REASONDebtCon | -2.405868e-01 |
| REASONHomeImp | 9.072084e-15 |
| JOBOffice | -5.369122e-01 |
| JOBOther | . |
| JOBProfExe | 1.216378e-01 |
| JOBSales | 9.837215e-01 |
| JOBSelf | 6.052147e-01 |
| LOAN | -2.791312e-01 |
| MORTDUE | -2.551587e-01 |
| VALUE | 3.282788e-01 |
| YOJ | -4.004850e-02 |
| DEROG | 4.231463e-01 |
| DELINQ | 8.352454e-01 |
| CLAGE | -5.398405e-01 |
| NINQ | 2.452999e-01 |
| CLNO | -1.832123e-01 |
| DEBTINC | 7.348997e-01 |

A diferencia de Ridge, acá los coeficientes si se pueden llevar a cero, de hecho, se tiene que la variable JOBOther, ya no forma parte del modelo, y el valor de lambda óptimo, lleva a tener 16 en vez de 17 variables.

A demás la, variable REASONHomeImp, queda dentro del modelo, pero con un coeficiente muy penalizado, llegando casi a ser cero.

```
          Reference
Prediction    0    1
         0 3646  580
         1  101  341

               Accuracy : 0.8541
                 95% CI : (0.8437, 0.8641)
    No Information Rate : 0.8027
    P-Value [Acc > NIR] : < 2.2e-16

                  Kappa : 0.4271

 Mcnemar's Test P-Value : < 2.2e-16

            Sensitivity : 0.9730
            Specificity : 0.3702
         Pos Pred Value : 0.8628
         Neg Pred Value : 0.7715
             Prevalence : 0.8027
         Detection Rate : 0.7811
   Detection Prevalence : 0.9053
      Balanced Accuracy : 0.6716

       'Positive' Class : 0
```

```
          Reference
Prediction   0   1
         0 894 150
         1  32  90

               Accuracy : 0.8439
                 95% CI : (0.8218, 0.8643)
    No Information Rate : 0.7942
    P-Value [Acc > NIR] : 8.66e-06

                  Kappa : 0.4162

 Mcnemar's Test P-Value : < 2.2e-16

            Sensitivity : 0.9654
            Specificity : 0.3750
         Pos Pred Value : 0.8563
         Neg Pred Value : 0.7377
             Prevalence : 0.7942
         Detection Rate : 0.7667
   Detection Prevalence : 0.8954
      Balanced Accuracy : 0.6702

       'Positive' Class : 0
```

| AUC Train | AUC Test | KS Train | KS Test |
|---|---|---|---|
| 0.8235803 | 0.8153843 | 0.5134 | 0.4818 |

Con respecto a Ridge, la única diferencia es que aumenta levemente la especificidad, pasando de 33% a 37%, lo cual sigue siendo muy bajo, respecto de lo que se espera en el modelo, ya que tampoco estaría diferenciando bien entre las clases.

V.- WOE e IV

Por último se hizo calculo de WOE para las variables que obtuvimos como resultado en stepwise, así como los IV para las mismas.
Acá se observa que la variables Reason tiene un valor por debajo de 0.02, lo cual indicaría bajo poder predictivo.

Para verificar, se hace el calculo incluyendo todas las variables.

```
1  LOAN      0.821
2  DELINQ    0.666
3  DEROG     0.396
4  CLNO      0.26
5  NINQ      0.176
6  DEBTINC   0.145
7  JOB       0.089
8  MORTDUE   0.067
9  CLAGE     0.042
10 VALUE     0.035
11 REASON    0.008
```

| | | |
|---|---|---|
| 1 | LOAN | 0.821 |
| 2 | DELINQ | 0.666 |
| 3 | DEROG | 0.396 |
| 4 | YOJ | 0.319 |
| 5 | CLNO | 0.26 |
| 6 | NINQ | 0.176 |
| 7 | DEBTINC | 0.145 |
| 8 | JOB | 0.089 |
| 9 | MORTDUE | 0.067 |
| 10 | CLAGE | 0.042 |
| 11 | VALUE | 0.035 |
| 12 | REASON | 0.008 |

Ahora se observa como, la variable YOJ tiene un IV que indicaría que tiene fuerte poder predictivo y que REASON sigue estando muy por debajo de 0.02. Esto se puede deber a que este método tiene la capacidad de capturar relaciones no lineales.

Con esto, se hace nuevamente calculo de los WOE pero dejando fuera solamente a la variable REASON.
Luego se guardan estos valores como variables, tanto para el set de entrenamiento como para el de validación.

Abajo el resumen del modelo WOE, primero con todas las variables guardadas, y luego dejando fuera "MORTDUE_woe", ya que no es significativa.
En el segundo modelo, se obtiene que todas las variables son significativas.

Coefficients:

| | Estimate | Std. Error | z value | Pr(>\|z\|) | |
|---|---|---|---|---|---|
| (Intercept) | -1.39987 | 0.04635 | -30.201 | < 2e-16 | *** |
| LOAN_woe | 0.91786 | 0.11308 | 8.117 | 4.77e-16 | *** |
| MORTDUE_woe | 0.10366 | 0.25497 | 0.407 | 0.68435 | |
| VALUE_woe | 0.92182 | 0.14356 | 6.421 | 1.35e-10 | *** |
| JOB_woe | 0.63792 | 0.15544 | 4.104 | 4.06e-05 | *** |
| YOJ_woe | 0.66579 | 0.20678 | 3.220 | 0.00128 | ** |
| DEROG_woe | 0.71435 | 0.07234 | 9.875 | < 2e-16 | *** |
| DELINQ_woe | 0.99272 | 0.05727 | 17.335 | < 2e-16 | *** |
| CLAGE_woe | 0.94183 | 0.09353 | 10.070 | < 2e-16 | *** |
| NINQ_woe | 0.54470 | 0.11019 | 4.943 | 7.68e-07 | *** |
| CLNO_woe | 0.79000 | 0.15035 | 5.254 | 1.49e-07 | *** |
| DEBTINC_woe | 1.03013 | 0.06036 | 17.066 | < 2e-16 | *** |

Coefficients:

| | Estimate | Std. Error | z value | Pr(>\|z\|) | |
|---|---|---|---|---|---|
| (Intercept) | -1.40040 | 0.04634 | -30.222 | < 2e-16 | *** |
| LOAN_woe | 0.91983 | 0.11294 | 8.144 | 3.82e-16 | *** |
| VALUE_woe | 0.94969 | 0.12615 | 7.528 | 5.15e-14 | *** |
| JOB_woe | 0.64031 | 0.15535 | 4.122 | 3.76e-05 | *** |
| YOJ_woe | 0.66075 | 0.20636 | 3.202 | 0.00137 | ** |
| DEROG_woe | 0.71535 | 0.07229 | 9.896 | < 2e-16 | *** |
| DELINQ_woe | 0.99335 | 0.05724 | 17.354 | < 2e-16 | *** |
| CLAGE_woe | 0.94236 | 0.09353 | 10.075 | < 2e-16 | *** |
| NINQ_woe | 0.54364 | 0.11013 | 4.936 | 7.97e-07 | *** |
| CLNO_woe | 0.79571 | 0.14975 | 5.314 | 1.08e-07 | *** |
| DEBTINC_woe | 1.03000 | 0.06038 | 17.058 | < 2e-16 | *** |

## Calidad Predictiva Modelo Elegido

```
          Reference
Prediction   0   1
         0 882  44
         1 135 105

               Accuracy : 0.8465
                 95% CI : (0.8245, 0.8667)
    No Information Rate : 0.8722
    P-Value [Acc > NIR] : 0.9955

                  Kappa : 0.4537

 Mcnemar's Test P-Value : 1.733e-11

            Sensitivity : 0.8673
            Specificity : 0.7047
         Pos Pred Value : 0.9525
         Neg Pred Value : 0.4375
             Prevalence : 0.8722
         Detection Rate : 0.7564
   Detection Prevalence : 0.7942
      Balanced Accuracy : 0.7860

       'Positive' Class : 0
```

```
          Reference
Prediction    0    1
         0 3597  150
         1  525  396

               Accuracy : 0.8554
                 95% CI : (0.845, 0.8654)
    No Information Rate : 0.883
    P-Value [Acc > NIR] : 1

                  Kappa : 0.4607

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

            Sensitivity : 0.8726
            Specificity : 0.7253
         Pos Pred Value : 0.9600
         Neg Pred Value : 0.4300
             Prevalence : 0.8830
         Detection Rate : 0.7706
   Detection Prevalence : 0.8027
      Balanced Accuracy : 0.7990

       'Positive' Class : 0
```

| AUC Train | AUC Test | KS Train | KS Test |
|---|---|---|---|
| 0.8561323 | 0.8486389 | 0.5486 | 0.5544 |

Finalmente el modelo elegido para predecir es el modeloWOE, ya que si bien comparado con el modelo Final visto al principio, tienen valores de accuracy muy similares, también para los valores de sensitividad y especificidad, los valores de AUC y KS son mejores, de hecho en la muestra de validación para el modelo Final, está por debajo de 0.5 y woe es un 0.55.

## Interpretación peso de los atributos

| Variables | Coeficiente |
|---|---|
| **Intercept** | -1.4003998 |
| **LOAN_woe** | 0.9198262 |
| **VALUE_woe** | 0.9496922 |
| **JOB_woe** | 0.6403076 |
| **YOJ_woe** | 0.6607527 |
| **DEROG_woe** | 0.7153499 |
| **DELINQ_woe** | 0.9933460 |
| **CLAGE_woe** | 0.9423604 |
| **NINQ_woe** | 0.5436387 |
| **CLNO_woe** | 0.7957126 |
| **DEBTINC_woe** | 1.0300047 |

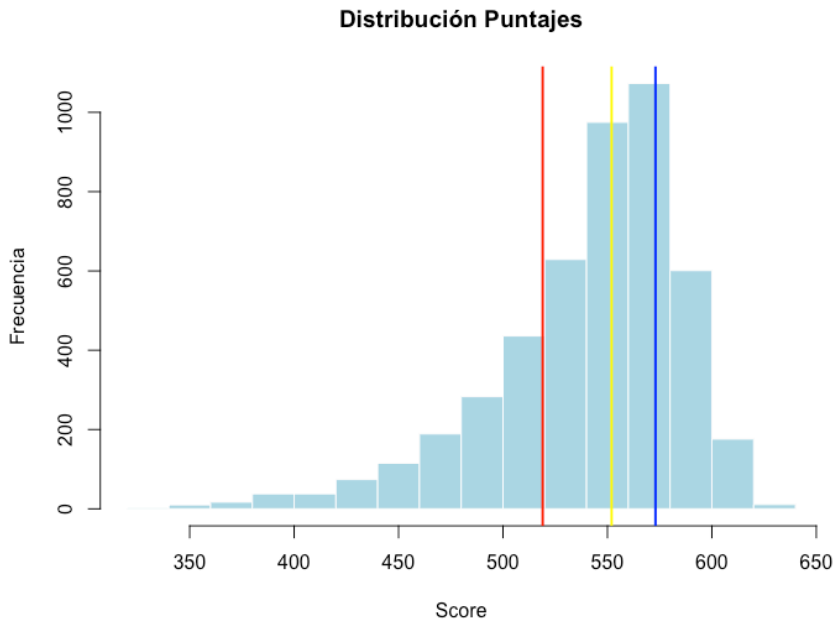
La interpretación de los atributos del modelo WOE, está dada por el impacto que genera el cambio de una categoría a la siguiente, ya que lo que se hace en un principio es categorizar las variables de manera que de una categoría a otra haya suficiente diferencia.
Más abajo, algunas de las variables con sus respectivas categorizaciones, donde se ve el corte de cada una y las diferencias entre la razón de malos vs buenos de cada una.
Por ejemplo DELINQ, donde se divide la variable en 3 categorías, siendo el grupo de 3 o más líneas de crédito no pagadas, el de mayor concentración de malos.

Lo mísmo se observa con DEBTINC, cuando la variable toma un valor de 43 o más.
Por otro lado y como ya se había mencionado anteriormente, la variable CLAGE, tiene un comportamiento inverso, lo cual también se evidencia en cada una de sus categorías, ya que a mayor antigüedad, disminuye la proporción de malos.
Con respecto a VALUE, esta relación no es tan directa ni tan clara, ya que primero se observa una disminución en la concentación de malos, y luego un aumento para dos categorías.
Si es importante notar que mientras menor es el valor de la propiedad, más es la probabilidad de incumplimiento.

```
$DELINQ
   variable      bin count count_distr good bad   badprob        woe    bin_iv total_iv breaks
1:   DELINQ [-Inf,1)  3645   0.7808483 3160 485 0.1330590 -0.4709277 0.1491615 0.611816      1
2:   DELINQ    [1,3)   774   0.1658098  506 268 0.3462532  0.7677011 0.1197204 0.611816      3
3:   DELINQ [3, Inf)   249   0.0533419   81 168 0.6746988  2.1327656 0.3429341 0.611816    Inf

$VALUE
   variable            bin count count_distr good bad   badprob         woe       bin_iv  total_iv breaks
1:    VALUE    [-Inf,40000)   240  0.05141388  152  88 0.3666667  0.85670706 0.0471039223 0.1352331  40000
2:    VALUE   [40000,50000)   246  0.05269923  176  70 0.2845528  0.48126201 0.0139726883 0.1352331  50000
3:    VALUE   [50000,90000)  1884  0.40359897 1517 367 0.1947983 -0.01587737 0.0001012548 0.1352331  90000
4:    VALUE  [90000,125000)  1246  0.26692374 1067 179 0.1436597 -0.38196968 0.0345328016 0.1352331 125000
5:    VALUE [125000,170000)   534  0.11439589  407 127 0.2378277  0.23862466 0.0069853467 0.1352331 170000
6:    VALUE [170000,200000)   254  0.05441302  229  25 0.0984252 -0.81159542 0.0275708303 0.1352331 200000
7:    VALUE   [200000, Inf)   264  0.05655527  199  65 0.2462121  0.28433321 0.0049662512 0.1352331    Inf

$CLAGE
   variable        bin count count_distr good bad   badprob        woe     bin_iv  total_iv breaks
1:    CLAGE  [-Inf,70)   268  0.05741217  162 106 0.3955224  0.9790935 0.07035541 0.2395325     70
2:    CLAGE   [70,170)  2026  0.43401885 1520 506 0.2497532  0.3033218 0.04360098 0.2395325    170
3:    CLAGE  [170,240)  1338  0.28663239 1137 201 0.1502242 -0.3295928 0.02808187 0.2395325    240
4:    CLAGE [240, Inf)  1036  0.22193659  928 108 0.1042471 -0.7476497 0.09749424 0.2395325    Inf

$DEBTINC
   variable       bin count count_distr good bad   badprob        woe      bin_iv  total_iv breaks
1:  DEBTINC [-Inf,30)  1237  0.26499572 1099 138 0.1115602 -0.6716515 0.096357929 0.5915139     30
2:  DEBTINC   [30,43)  3130  0.67052271 2562 568 0.1814696 -0.1031713 0.006915164 0.5915139     43
3:  DEBTINC [43, Inf)   301  0.06448158   86 215 0.7142857  2.3195415 0.488240791 0.5915139    Inf
```

## Uso Modelo Predictivo

Para el uso del modelo predictivo elegido se crea un scorecard, con el objetivo de poder tener un puntaje para cada cliente y poder definir en base a este si se le otorga o no un crédito con garantía hipotecaria.




La propuesta en este caso para el uso del modelo, es en base a 3 puntos de corte.
La línea roja representa el cuartil 25 = 519, y bajo ese umbral, se deben rechazar todos los clientes.
La línea azul, representa el cuartil 75 = 573 y sobre ese umbral, se debe dar el crédito a todos los clientes.
Luego en el rango intercuartil dependiendo del tipo de cliente y evaluación del banco, se puede tomar a todos quienes están entre la mediana = 552(linea amarilla) y el cuartil 75 y darles el crédito, y quienes caen entre la linea amarilla y roja deben pasar a revisión por parte del banco.