Tarea 2
Modelamiento Predictivo

Integrantes:

- Diego Gálvez
- Diego Rivera
- Cinthia Roa

Profesor: Rolando De La Cruz
Asignatura: Modelamiento Predictivo para la ciencia de datos
Fecha de entrega: Domingo 05 de Enero de 2020

Tabla de Contenidos

Contexto del problema

El conjunto de datos HMEQ entrega múltiple información sobre morosidad para 5960 préstamos con garantía hipotecaria otorgados por un banco. Un préstamo con garantía hipotecaria es un préstamo en el que el deudor utiliza el valor líquido (home equity) de su vivienda como garantía subyacente. El banco está interesado en determinar la fiabilidad de un cliente que solicita préstamos. Cuando un banco recibe una solicitud de préstamo, según el perfil del solicitante, el banco tiene que tomar una decisión sobre si continuar con la aprobación del préstamo o no. Dos tipos de riesgos están asociados con la decisión del banco:

I. Si el solicitante tiene un bajo riesgo, es decir, la persona es capaz de reembolsar el préstamo, entonces el no aprobar el préstamo, puede resultar en una pérdida de negocios para el banco.
II. Si el solicitante tiene un alto riesgo, es decir, no es probable que pague el préstamo, la aprobación del préstamo a la persona resulta en una pérdida financiera para el banco.

Los datos del archivo HMEQ contienen datos sobre 13 atributos y la clasificación que recoge la variable target BAD (1 = el solicitante incumplió con el préstamo o se encuentra seriamente en mora; 0 = préstamo pagado por el solicitante).

El conjunto de datos presenta las siguientes características:

Variable	Descripción	Tipo de dato	Datos nulos
BAD	1: solicitante incumplido en préstamo 0: préstamo pagado por el solicitante	Entero	0
LOAN	Monto de la solicitud de préstamo	Entero	0
MORTDUE	Monto adeudado de la hipoteca existente	Numérico	518
VALUE	Valor de la propiedad actual	Numérico	112
REASON	Razón del préstamo solicitado: - DebtCon: consolidación de la deuda - HomeImp: mejoras para el hogar	Categórico	252
JOB	Categorías ocupacionales	Categórico	279
YOJ	Años en el trabajo actual	Numérico	515
DEROG	Número de informes despectivos importantes	Entero	708
DELINQ	Número de líneas de crédito morosas	Entero	580
CLAGE	Edad de la línea de crédito más antigua (meses)	Numérico	308

NINQ	Número de consultas de crédito recientes	Entero	510
CLNO	Número de líneas de crédito	Entero	222
DEBTINC	Relación deuda-ingreso	Numérico	1267

Tabla 1: Descripción de variables

Objetivo e identificación de atributos

El objetivo del trabajo desarrollado consta en la automatización del proceso de toma de decisiones para la aprobación de líneas de crédito con garantía hipotecaria, siguiendo las recomendaciones de la ***Ley de Igualdad de Oportunidades de Crédito*** para crear un modelo de puntuación de crédito empíricamente derivado y estadísticamente sólido. El modelo se basará en los datos recopilados de solicitantes recientes que recibieron crédito a través del proceso actual de suscripción de préstamos. El modelo se construirá a partir de herramientas de modelado predictivo, pero el modelo creado debe ser lo suficientemente interpretable para proporcionar una razón para cualquier acción adversa (rechazos).

Para contextualizar, la Ley de Igualdad de Oportunidades de Crédito (ECOA) es una ley estadounidense promulgada en 1974 que hace ilegal que cualquier acreedor discrimine a un solicitante sobre la base de raza, color, religión, nacionalidad, sexo, estado civil o edad. Dado a esto, las entidades dentro del conjunto de datos HMEQ no presentan ninguna característica o atributo que puedan atribuirse a las propiedades mencionadas anteriormente.

Mencionado anteriormente, la variable target pertenece a la entidad BAD, que determina a partir de la información proporcionada, si el solicitante ha pagado el préstamo solicitado o no. Dado que se busca la morosidad del solicitante, los valores positivos (“1”) significan que si se presenta un impago del préstamo.

Análisis descriptivo de los atributos

Primero se averiguaron los estadísticos de resumen presentes en el conjunto de datos, dejando afuera las variables “REASON” y “JOB” debido a poseer una clasificación categórica.

	Min.	1st Qu.	Median	Mean	3st Qu.	Max.	S.D.	N.A.
LOAN	1100	11100	16300	18608	23300	89900	11207.48	0
MORTDUE	2063	46276	65019	73761	91488	399550	44457.61	518
VALUE	8000	66076	89236	101776	119824	855909	57385.78	112
YOJ	0.000	3.000	7.000	8.922	13.000	41.000	7.573982	515
DEROG	0.0000	0.0000	0.0000	0.2546	0.0000	10.0000	0.846047	708
DELINQ	0.0000	0.0000	0.0000	0.4494	0.0000	15.0000	1.127266	580
CLAGE	0.1	115.1	173.5	179.8	231.6	1168.2	85.81009	308
NINQ	0.000	0.000	1.000	1.186	2.000	17.000	1.728675	510
CLNO	0.0	15.0	20.0	21.3	26.0	71.0	10.13893	222
DEBTINC	0.5245	29.1400	34.8183	33.7799	39.0031	203.3121	8.601746	1267

Tabla 2: Estadísticos de resumen

A partir de la información presentada en la tabla “estadísticos de resumen”, es posible observar la presencia de datos “N.A”. en la gran parte de los atributos presentados. Para visualizar dicha información, se presenta el siguiente gráfico que muestra la cantidad de “N.A.” por atributo:

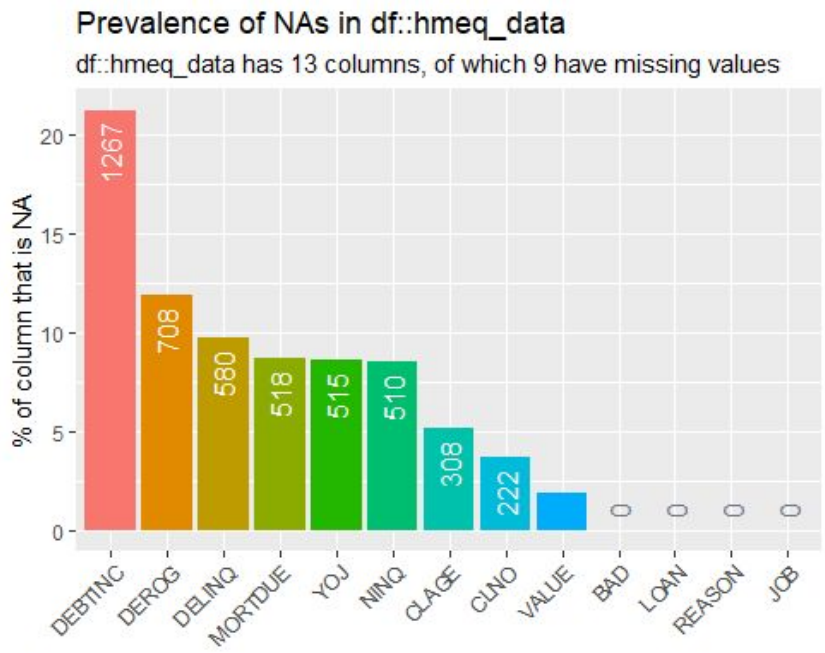


Figura 1: Prevalencia de N.A. dentro de HMEQ

Aunque en la figura 1 se puede observar que las variables “REASON” y “JOB” no presentan valores faltantes, esto no es cierto dada la información entregada dentro de la tabla 1, demostrando que sí poseen datos “N.A.”. La razón de porque no aparece es debido a la clase de tales entidades, ya que al no ser de una clase entera o numérica, el ggplot omite todos los valores dentro de ella, presentando problemas a la hora de limpieza o imputación de datos.

En función de esta información, y considerando los atributos que corresponden a variables categóricas mencionados anteriormente, se decide realizar un tratamiento de estos datos para trabajar el modelo predictivo posteriormente.

Ingeniería de variables

Tratamiento de datos categóricos (variables “REASON” y “JOB”)

Para darle mayor valor a la variable JOB dentro modelo a desarrollar, se realizó una puntuación a los distintos tipos de trabajo, dejando como “Other” todos aquellos que no tengan dato alguno, y con una mejor puntuación a quienes en base a trabajo tengan mejor chance de recibir una aprobación de líneas con garantía hipotecaria. Para esto, se tomó el siguiente criterio:

Para la variable JOB, se asignaron valores entre 1 y 6 a las distintas categorías ocupacionales, donde los valores en blanco, fueron asignados al valor 1 (“Other”).

JOB	Propuesta puntuación variable JOB	CASOS
Mgr	6	767
Office	5	948
Sales	4	109
ProfExe	3	1276
Self	2	2667
Other = Blank	1	193

Tabla 3: Asignación de puntajes para la variable JOB

Por su parte, la variable "REASON" que considera la razón de solicitud del préstamo, fue clasificada en valores de 1 a 3, y al igual que en la variable "JOB" a los valores en blanco se les asignó valor igual a 1 ("Other").

REASON	Propuesta puntuación variable REASON	CASOS
DebtCon	3	3928
HomeImp	2	1780
Other = Blank	1	252

Tabla 4: Asignación de puntajes para la variable REASON

Tratamiento de valores "N.A." en los datos

Para el tratamiento de datos "N.A." se decidió ocupar el método de ***listwise deletion***, el cual es una técnica que omite entidades completas si poseen tan solo un valor faltante. Antes del tratamiento de datos, el "Mapa de datos nulos" se comportaba de la siguiente manera:

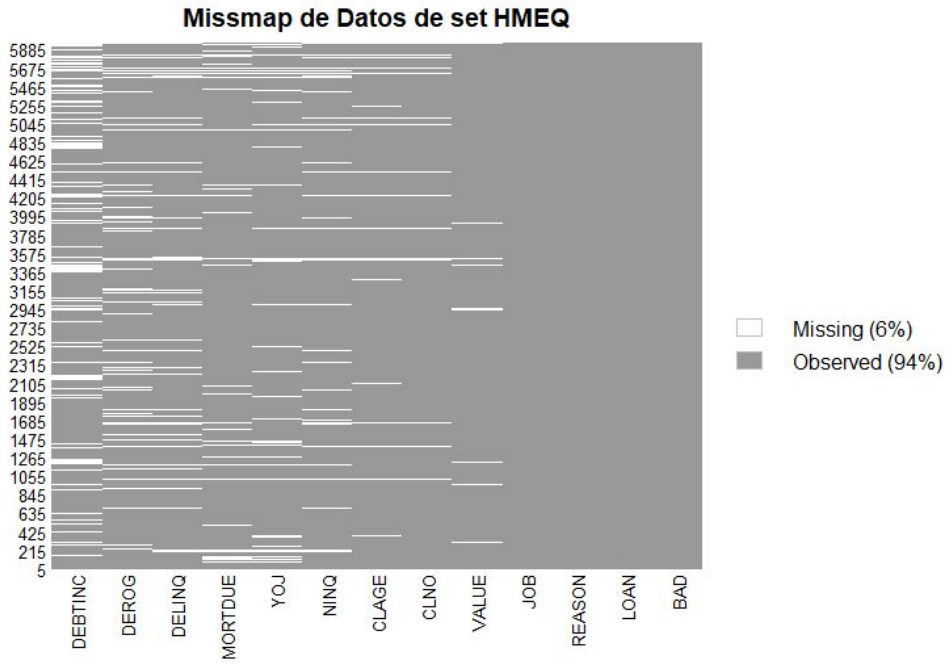


Figura 2: Mapa de datos nulos para el set HMEQ previo a la limpieza

Observando la leyenda perteneciente a la figura 2, un 6% de los datos son faltantes, siendo la mayoría de la variable DEBTINC (esto también se puede comprobar en las tabla 1 y 2, y predominantemente en la figura 1).

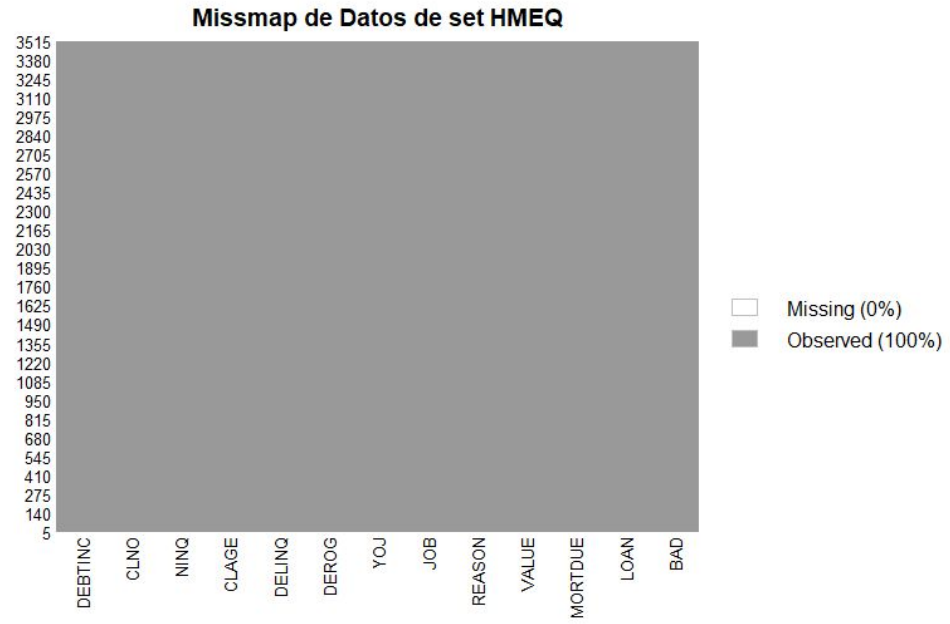


Figura 3: Mapa de datos nulos para el set HMEQ posterior a la limpieza

Posterior a la limpieza de datos “N.A.” realizada, se obtuvo una reducción en la cantidad de observaciones de nuestro data frame, pasando de un total de 5.960 observaciones a un total de 3.515 casos, manteniendo un 58,98% de los registros.

Análisis gráfico de los datos

Luego de haber realizado el tratamiento de los datos, buscaremos entender el comportamiento de cada una de las variables que pueden ayudarnos a construir el modelo. Para esto se presentan las distribuciones de las distintas variables numéricas contenidas en los datos.

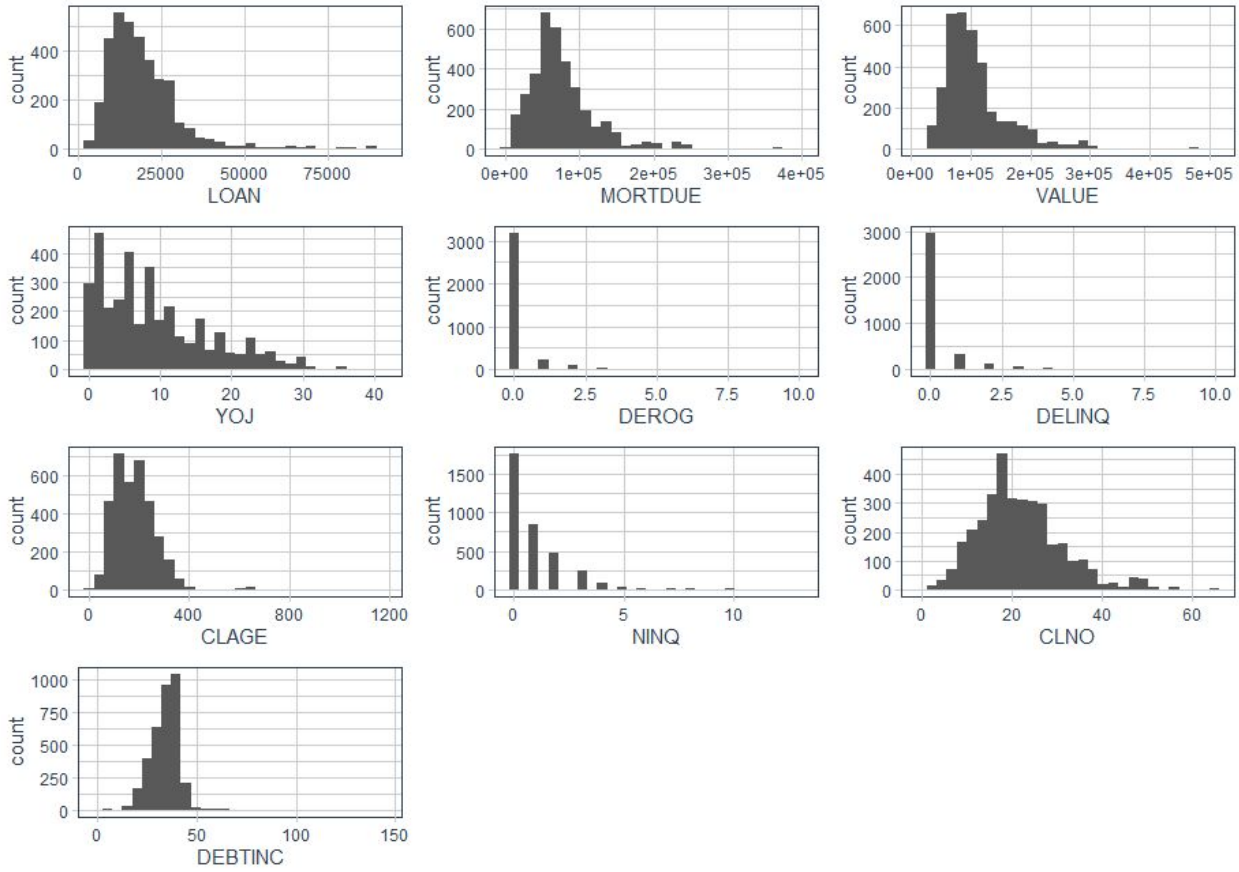


Figura 4: Distribución de variables

Visualmente es posible obtener algunas conclusiones rápida de acuerdo a la distribución de los datos. Cómo por ejemplo, es interesante observar en los tres primeros gráficos (LOAN, MORTDUE, VALUE) una forma gráfica similar entre ella, sin embargo, esto tiene cierta lógica, ya que corresponden a valores monetarios que podrían tener cierta relación, ya que, LOAN corresponde al monto de la solicitud de préstamo, MORTDUE al monto adeudado de la hipoteca existente, y VALUE corresponde al valor de la

propiedad actual. Además podemos pensar que a mayor valor de la propiedad actual podría permitir acceder a un monto superior de préstamo.

También es posible observar en las variables DEROG y DELINQ un gran cantidad de valores igual a cero, y una muy baja frecuencia en valores mayores a 0. Sin embargo, al contextualizar los resultados, tiene mucho sentido, ya que corresponden al número de informes respectivos importantes (DEROG), y al número de líneas de crédito morosas (DELINQ). Tomando en cuenta la información que estamos manejando, corresponde a observaciones sobre personas que han obtenido ya un crédito, por lo que probablemente hayan sido clasificados con estas variables, lo cual conlleva a concluir que han tenido un buen comportamiento financiero, lo que les permitió acceder al préstamo.

Del mismo modo, también analizamos las variables categóricas que transformamos en un comienzo (REASON y JOB), sin embargo las graficamos su relación con la variable target (BAD).

En el siguiente gráfico se observa la proporcionalidad de las variables REASON y JOB, con la variable BAD que corresponde al incumplimiento de pago del préstamo solicitado, donde 0 corresponde al pago del préstamo por parte del solicitante, y 1 corresponde a que el solicitante ha incumplido con el pago del préstamo solicitado. Para el atributo REASON no existe mucha variación entre los porcentajes de cada categoría, pero al llegar a la variable BAD existe un incremento sustancial para la categoría “Sales” en comparación a sus contrapartes.

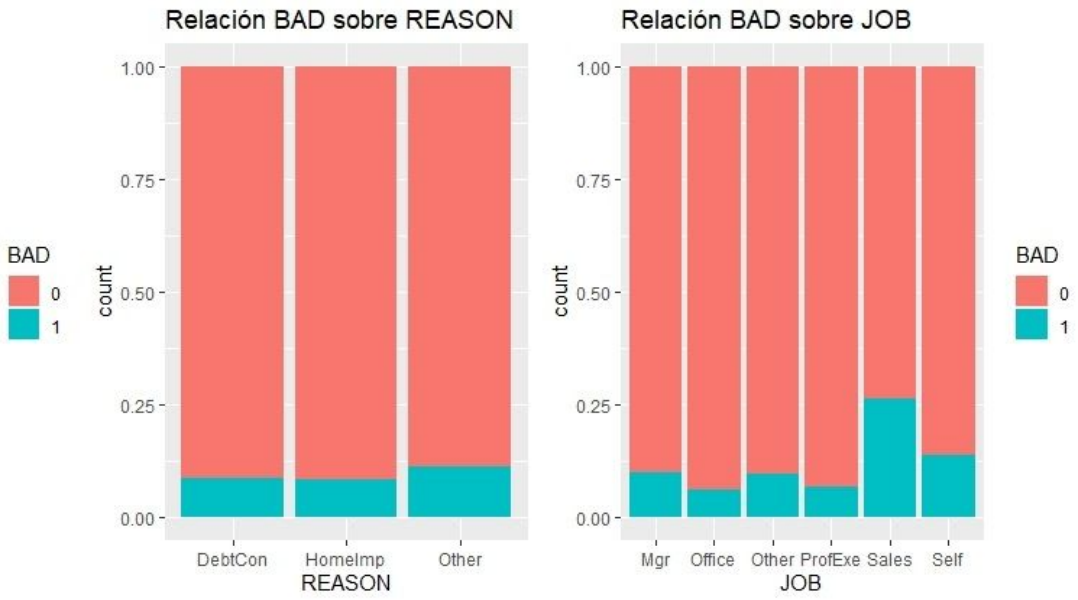


Figura 5: Proporcionalidad de variable BAD respecto a variables REASON y JOB

A partir de la figura 5, se es posible realizar una inferencia respecto al peso que ejerce el trabajo “Sales” sobre la morosidad de los solicitantes.

Otra variable analizada fue la entidad “NINQ”, que menciona el número de consultas de crédito recientes. Tal como se puede observar en la figura 4, los datos presentes dentro del set exhiben una alta concentración entre los valores 0 al 3, mientras que la cantidad para los valores 4 hacia adelante son considerablemente reducidos. Esto también se puede apreciar en el estadístico de resumen en la tabla 2, donde el primer cuartil, mediana y tercer cuartil se disocia prominentemente del máximo valor. Los datos de la variable, junto al número de registros, son los siguientes:

```
> table(hmeq_data$NINQ)

   0    1    2    3    4    5    6    7    8    9   10   11   13
1763  853  470  241   81   31   22   15   12    7   17    2    1
```

Debido a la baja frecuencia de los valores crecientes de la variable “NINQ”, los datos pueden interpretarse incorrectamente, representado dentro del modelo a desarrollar. Para combatir esto, se decide agrupar los valores mayores o iguales a 4, y ser transformados como factores, agrupándolos dentro de una nueva variable “NINQ2”.

```
> table(hmeq_data$NINQ2)

   0    1    2    3    4
1763  853  470  241  188
```

El problema de representatividad puede verse claramente en el costado izquierdo de la figura 6, donde la variable target es irregular a lo largo del número de consultas de crédito, llegando incluso a generar una barra completa de color celeste, separada del resto de la distribución, demostrando ser inconsistente para el desarrollo del modelo predictivo. Pero luego de realizar la agrupación de datos mencionada anteriormente, “NINQ2” entrega un efecto significativo en comparación con “NINQ”.

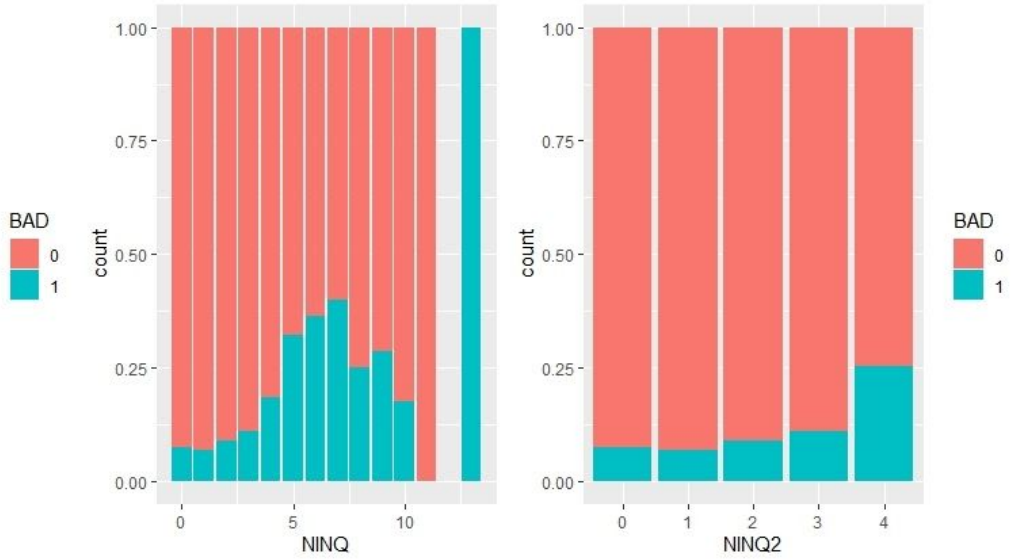


Figura 6: Proporcionalidad de variable BAD respecto a la variable NINQ

El mismo problema sucede con la variable DEROG y DELINQ, donde los registros se encuentran concentrados en los menores valores. Esto se puede comprobar dentro de la tabla 2, donde el primer cuartil, mediana y tercer cuartil presentan una separación considerable al valor máximo.

```
> table(hmeq_data$DEROG)

   0    1    2    3    4    5    6    7    8    9   10
3188  212   80   23    4    1    2    2    1    1    1
> table(hmeq_data$DELINQ)

   0    1    2    3    4    5    6    7    8   10
2964  333  126   50   21    6    7    6    1    1
```

De la misma manera que la variable “NINQ”, se agruparon los valores para evitar una incorrecta representación de ambas entidades dentro del modelo, por lo que se aglomeraron las instancias superiores e iguales a 2 en ambos casos.

```
> table(hmeq_data$DEROG2)

   0    1    2
3188  212  115
> table(hmeq_data$DELINQ2)

   0    1    2
2964  333  218
```

En las siguientes figuras, de manera similar a la figura 6, se aprecia claramente la estandarización de los datos en contraste a las cifras preliminares no agrupadas. “DEROG” y “DELINQ” exhiben solicitantes morosos casi inmediatos al momento de alcanzar ciertos rangos de cifras, como por ejemplo, cuando un solicitante posee 4 o más informes despectivos importantes (“DEROG”) entonces automáticamente es calificado como un cliente moroso. Lo mismo sucede para la variable “DELINQ”, donde al igual que el gráfico de la variable “NINQ” (figura 6), presenta una columna desunida con el resto de las barras acumulativas.

Se espera que la agrupación de valores presenten un modelo más estable y efectivo.

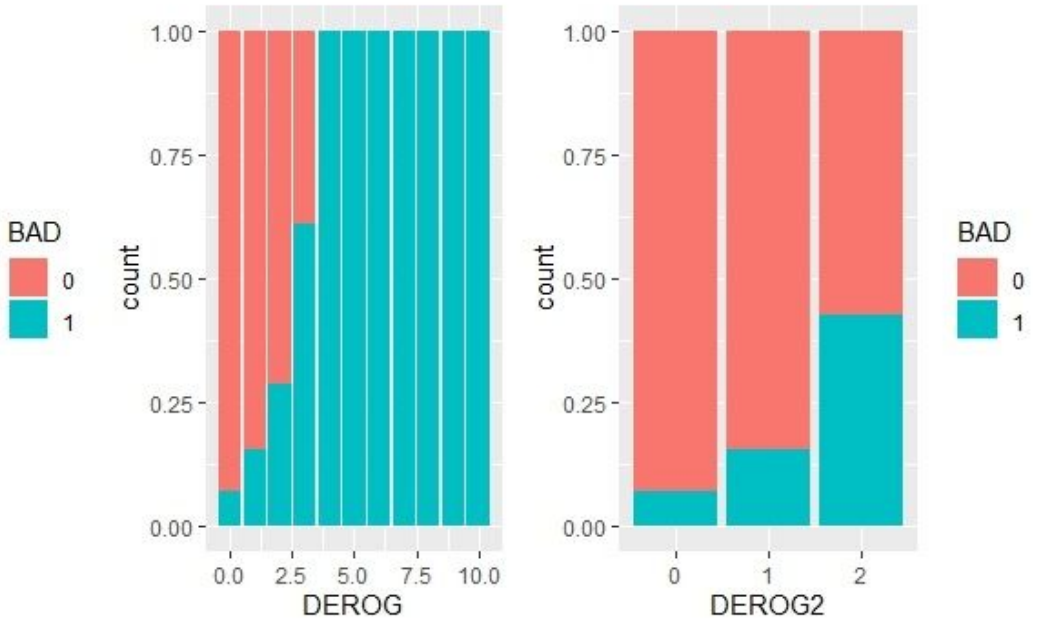


Figura 7: Proporcionalidad de variable BAD respecto a la variable DEROG

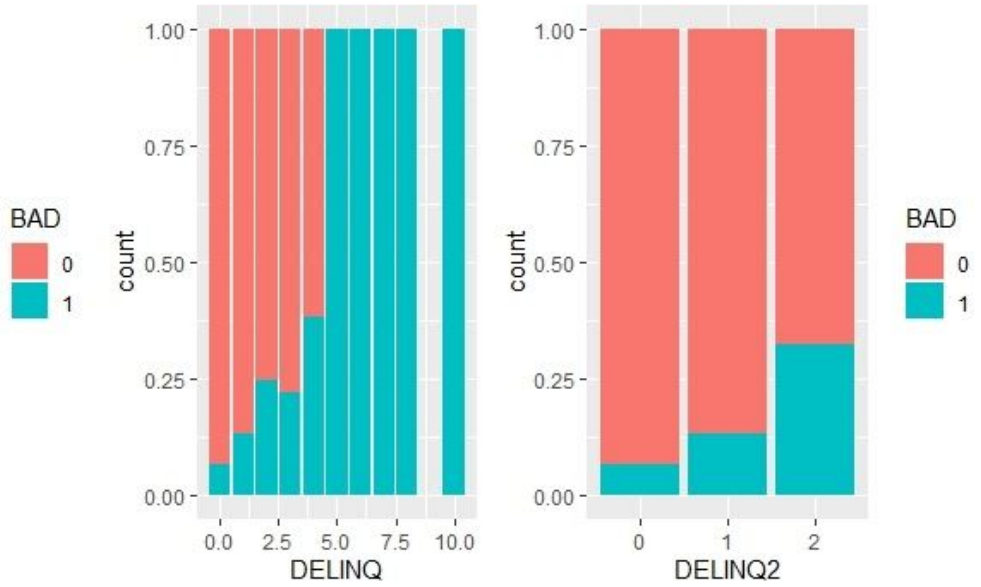


Figura 8: Proporcionalidad de variable BAD respecto a la variable DELINQ

El siguiente análisis realizado corresponde a la visualización de outliers para las variables continuas, es decir, dejando de lado las variables trabajadas anteriormente (“NINQ”, “DEROG” y “DELINQ”), esto debido a que fueron transformadas a factores. Dentro de la figura 9 se divisan los boxplot de las 7 variables restantes, donde a primera vista se nota que existe poca disparidad de las entidades a partir de

la variable target BAD. Junto a esto, no se es posible decidir que las variables tienen un efecto significativo en las variables de respuesta, por lo que no se continúa con una exploración de los datos.

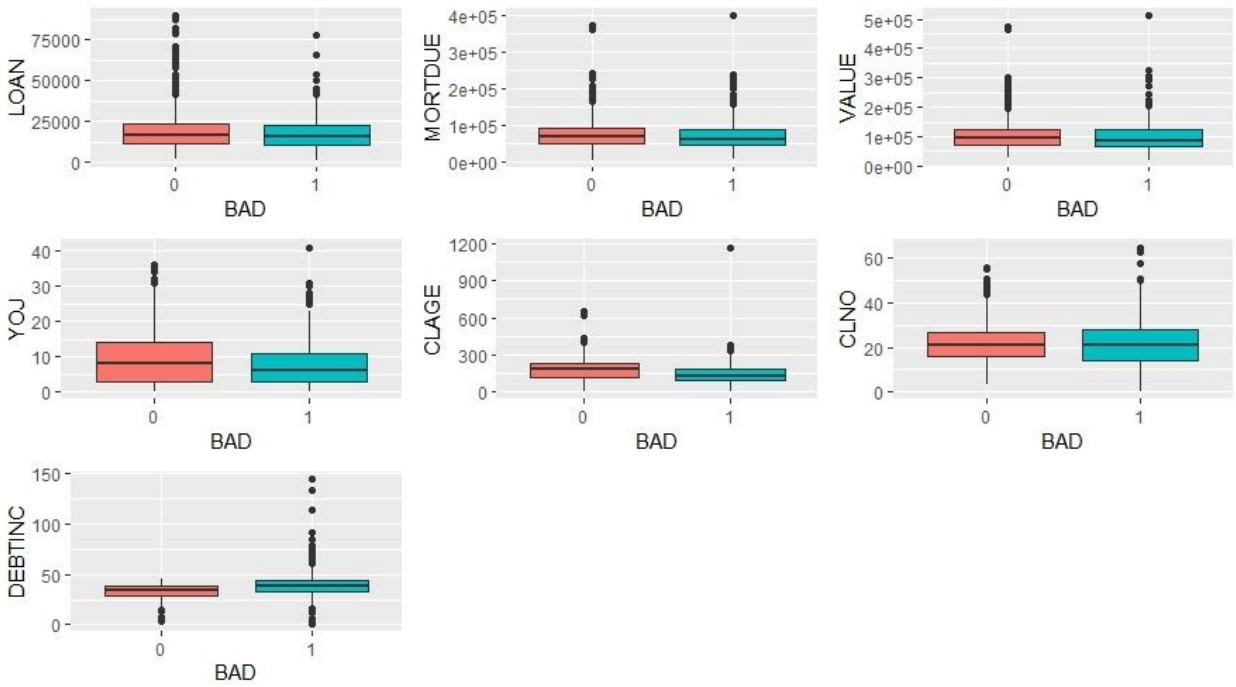


Figura 9: Boxplot de variables continuas

Análisis de correlación

Se realizó un análisis de correlación de las variables, con el objetivo de mitigar alguna multicolinealidad entre las variables que pudiese afectar al modelo predictivo. En la figura 10 se observa el mapa de correlación elaborado para las variables numéricas del set de datos, omitiendo todas aquellas que sean categóricas. Como se puede observar, existe una alta multicolinealidad entre las variables “MORTDUE” y “VALUE”, donde la explicación más cercana se retribuye a la definición de las variables, donde el monto adeudado de la hipoteca existente va de la mano con el valor de la propiedad, de ahí la alta correlación.

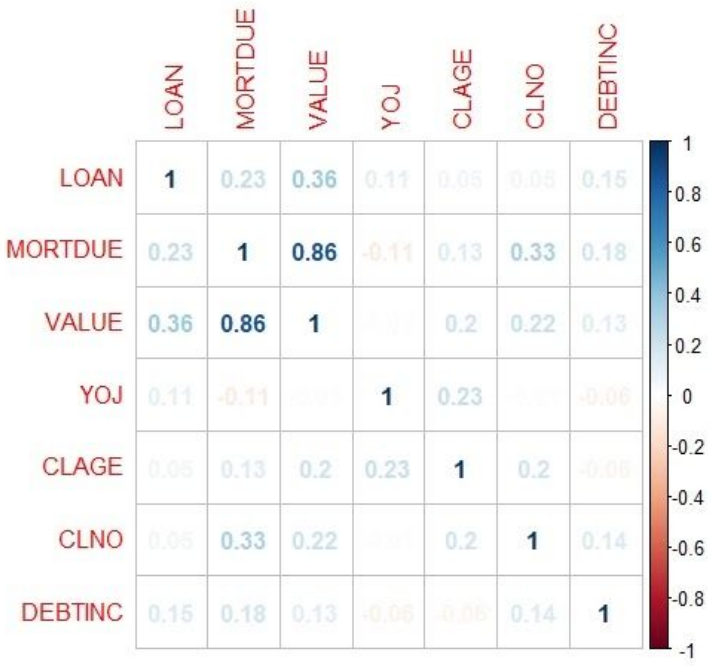


Figura 10: Mapa de correlación

Para ello se considera una nueva variable que contempla el patrimonio real de la propiedad a evaluar, basado en el diferencial entre el valor de la propiedad actual y el monto adeudado por la hipoteca existente lo que consideraría como garantía real del solicitante de crédito. Esta nueva variable ("GARANTIA") muestra el siguiente estadístico de resumen:

	Min.	1st Qu.	Median	Mean	3st Qu.	Max.	S.D.
GARANTIA	-205445	17226	26815	31253	39846	191958	27715.54

Tabla 5: estadístico de resumen para la variable "GARANTIA"

Para aquellos valores con resultados negativos en la nueva variable "GARANTIA" podrían considerarse como potenciales clientes no aptos para aprobación de crédito, debido a que la garantía estaría basada en un pasivo y no en un activo. En la figura 11 se observa la correlación con la nueva variable, apreciándose una correlación ideal para las variables continuas.

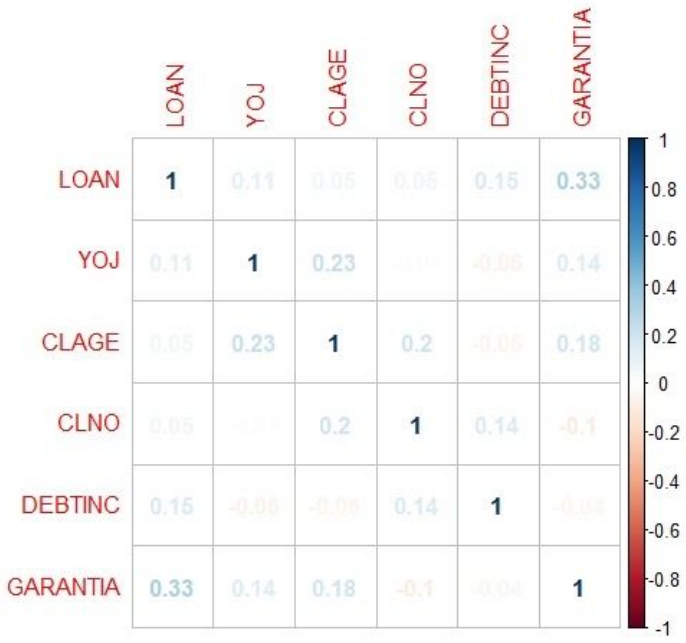


Figura 11: Mapa de correlación con variable “GARANTIA”

Elección y construcción del modelo predictivo

En una primera instancia se realizará un análisis con las nueve variable descritas anteriormente , además de ello se considerará el total de la base (exceptuando datos faltantes), los cuales serían 3515 registros, dentro de los cuales la variable a predecir (BAD) cuenta con 3206 registro “0” y 309 registros “1”, mostrando un desbalance en los datos. En una segunda instancia, se realizará el mismo análisis con una muestra de data balanceada.

Variables a incluir en el modelo :

BAD	LOAN	REASON	JOB	YOJ	CLAGE
CLNO	DEBTINC	NINQ2	DEROW2	DELINQ2	GARANTIA

Data No Balanceada

El primer modelo desarrollado se utiliza el ajuste de modelos lineales generalizados con todas las variables del modelo en la data de entrenamiento, bajo esta regresión no se utiliza ningún análisis de selección de variables, resultando 6 de ellas significativas.

```
Coefficients:
             Estimate Std. Error z value Pr(>|z|)
(Intercept) -3.455e+00  7.747e-01  -4.460 8.18e-06 ***
LOAN        -2.197e-05  7.688e-06  -2.858 0.004268 **
REASON2     -8.726e-01  5.353e-01  -1.630 0.103029
REASON3     -6.782e-01  5.245e-01  -1.293 0.195988
JOB2        -8.056e-01  3.941e-01  -2.044 0.040948 *
JOB3        -7.553e-01  4.009e-01  -1.884 0.059574 .
JOB4         4.305e-01  5.535e-01   0.778 0.436645
JOB5        -1.393e+00  4.226e-01  -3.295 0.000983 ***
JOB6        -8.183e-01  4.196e-01  -1.950 0.051156 .
YOJ         -1.156e-02  9.941e-03  -1.162 0.245078
CLAGE       -5.618e-03  1.036e-03  -5.422 5.90e-08 ***
CLNO        -1.195e-02  7.791e-03  -1.533 0.125163
DEBTINC      1.035e-01  1.023e-02  10.111  < 2e-16 ***
NINQ21      -3.610e-01  1.834e-01  -1.968 0.049015 *
NINQ22      -2.737e-01  2.181e-01  -1.255 0.209521
NINQ23      -2.161e-02  2.564e-01  -0.084 0.932833
NINQ24       1.007e+00  2.344e-01   4.297 1.73e-05 ***
DEROG21      3.601e-01  2.286e-01   1.576 0.115080
DEROG22      1.917e+00  2.525e-01   7.591 3.17e-14 ***
DELINQ21     1.000e+00  2.019e-01   4.952 7.34e-07 ***
DELINQ22     2.043e+00  2.006e-01  10.186  < 2e-16 ***
GARANTIA     5.444e-06  2.923e-06   1.862 0.062569 .
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

Resultado de la selección de variables con stepwise.

Se decide realizar como método de selección de variables stepwise, ya que dicho método considera tanto el método backward como forward. Dicho análisis concluye que consideremos las 9 variables en el modelo, no considerando las variables REASON y YOJ

BAD	LOAN	DELINQ2	JOB	GARANTIA	CLAGE
CLNO	DEBTINC	NINQ2	DEROW2		

En base al modelo entregado por el método stepwise, se decide realizar método de regularización Lasso bajo validación cruzada (nfolds= 5), con el objetivo de reducir más variables y lograr un modelo parsimonioso, sin embargo bajo este tipo de regresión sólo deja fuera 3 categorías de JOB y una de NINQ2.

(Intercept)	-3.0159707
X.Intercept.	.
JOB2	.
JOB3	.
JOB4	1.1515111
JOB5	-0.4676829
JOB6	.
NINQ21	-0.2684564
NINQ22	-0.1137683
NINQ23	.
NINQ24	0.9642930
DEROG21	0.3043816
DEROG22	1.8215972
DELINQ21	0.8576135
DELINQ22	1.8966391
LOAN	-0.1749765
CLAGE	-0.4339788
CLNO	-0.0777276
DEBTINC	0.7870699
GARANTIA	0.0946518

Evaluación del modelo

Si sólo analizamos el Accuracy de dicho modelo , este se consideraría un modelo idóneo ya que el estadístico accuracy es de 0,9218, es decir, si miramos esto en porcentaje diríamos que el 92% de las predicciones coinciden con las observaciones. Sin embargo, al desglosar este estadístico, nos damos cuenta que presenta una sensibilidad de un 0,99 , es decir, el modelo está clasificando bien a los verdaderos positivos , no así a los falsos negativos , ya que presenta una especificidad de 0,1877, valor muy por debajo de del 0,8. Esto quiere decir , que modelo no logra predecir bien los valores 1, que en este caso serían los clientes que no pagaron los créditos, esto puede significar un gran problema para la empresa ya que se le puede otorgar crédito a alguien que no está siendo capaz de reembolsar los montos prestados.Lo anterior puede suceder debido a problema de desbalanceo de la data, es decir puede que el modelo se sobreentrene y pueda lograr predecir bien solo una categoría, en ese caso serían los casos cero.

Accuracy: 0,9218
Sensitivity: 0,9925
Specificity: 0,1877

Data Balanceada

Bajo el método de stepwise , se consideran 9 variables , dejando fuera las variables REASON y GARANTIA

BAD	LOAN	DELINQ2	JOB	YOJ	CLAGE
CLNO	DEBTINC	NINQ2	DEROG2		

Luego, considerando los resultados del método de stepwise, se aplica regresión Lasso, solo deja una categoría de la variable JOB fuera:

```
(Intercept)  -0.43127193
X.Intercept.  .
JOB2          0.04239021
JOB3         -0.04824268
JOB4          0.93397503
JOB5         -0.61176364
JOB6          .
NINQ21       -0.13155732
NINQ22       -0.09647873
NINQ23        0.10208308
NINQ24        0.82204356
DEROG21       0.46415346
DEROG22       1.35639996
DELINQ21      0.89391707
DELINQ22      1.89628796
LOAN         -0.21986995
YOJ          -0.13332527
CLAGE        -0.29775344
CLNO         -0.13494972
DEBTINC       0.67785201
```

Evaluación del modelo

Se evalúa el modelo con las variables recomendadas por el método de stepwise.
Al balancear la muestra, logramos una matriz de confusión con resultados mucho mejor que con la data no balanceada, en cuanto a sensibilidad presenta un nivel por sobre 0,8 lo que logra predecir de buena manera los valores ceros, sin embargo, el modelo sigue siendo débil en cuanto a predecir los valores uno, ya que presenta una especificidad de 0,66

Accuracy: 0,7314
Sensitivity: 0,8091
Specificity: 0,6602

Otro método: Bajo division data entrenamiento y testeo

Se realiza la partición de datos entre la base de entrenamiento y la base de testeo, utilizando como tamaño de entrenamiento un 80% de los datos, estableciendo 2813 observaciones para esta, y 702 para el testeo. Se aplica método de stepwise:

Opción Modelo 1 (Modelos lineales generalizados) sin balanceo

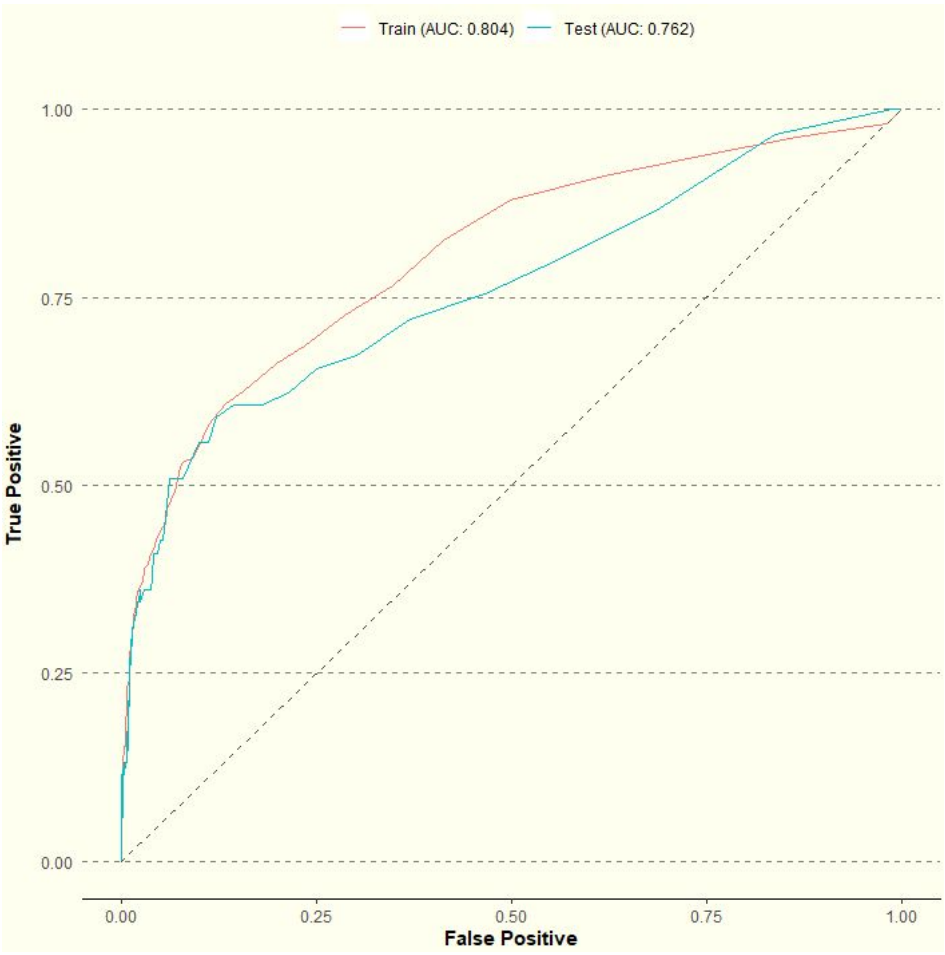
```
Coefficients:
             Estimate Std. Error z value Pr(>|z|)
(Intercept) -5.479e+00  4.778e-01 -11.466  < 2e-16 ***
LOAN        -2.058e-05  7.870e-06  -2.615 0.008913 **
CLAGE       -5.039e-03  1.125e-03  -4.478 7.55e-06 ***
CLNO        -1.420e-02  8.472e-03  -1.676 0.093694 .
DEBTINC      1.163e-01  1.177e-02   9.887  < 2e-16 ***
NINQ21      -2.273e-01  1.977e-01  -1.150 0.250255
NINQ22      -2.116e-01  2.385e-01  -0.887 0.374847
NINQ23      -1.694e-01  2.871e-01  -0.590 0.555151
NINQ24       1.001e+00  2.578e-01   3.883 0.000103 ***
DEROG21      2.178e-01  2.631e-01   0.828 0.407850
DEROG22      1.940e+00  2.662e-01   7.286 3.20e-13 ***
DELINQ21     7.720e-01  2.253e-01   3.427 0.000610 ***
DELINQ22     1.913e+00  2.182e-01   8.765  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

Todos los valores VIF, están por debajo el nivel 5, por tanto no presenta problemas de multicolinealidad.

Evaluación de Calidad Predictiva

K-S y ROC Modelo Modelo 1

Si analizamos KS y la curva de ROC (AUC), estas presentan un nivel de 0,49 y 0,76 respectivamente, estos valores son aceptables, ya que un ks por sobre 0,4 podría llegar a considerarse y una curva de ROC por sobre los niveles 0,5 es bueno. Por tanto bajo estos estadísticos el modelo 1 , bajo data de entrenamiento y testeo presentan buenos niveles.



Modelo 1	Entrenamiento	Testeo
Accuracy	0,925	0,9217
Sensibilidad	0,9296	0,9284
Especificidad	0,7857	0,6667

Se aprecia que tanto a nivel de entrenamiento como testeo presentan niveles similares, asimismo, estos presentan un accuracy por sobre 0,9, es decir, el modelo está logrando su objetivo, sin embargo a nivel de especificidad el modelo sigue débil, ya que no logra predecir de buena manera los valores 1.

Opción Modelo 2 (Todas Las Variables) con balanceo

```
Coefficients:
             Estimate Std. Error z value Pr(>|z|)
(Intercept) -1.378e+00  5.424e-01  -2.541  0.01105 *
LOAN        -2.169e-05  9.005e-06  -2.409  0.01598 *
YOJ         -2.560e-02  1.610e-02  -1.590  0.11177
CLAGE       -3.294e-03  1.435e-03  -2.295  0.02175 *
CLNO        -1.086e-02  1.089e-02  -0.998  0.31850
DEBTINC      6.320e-02  1.257e-02   5.027 4.99e-07 ***
NINQ21      -2.262e-01  2.757e-01  -0.821  0.41187
NINQ22      -2.346e-01  3.145e-01  -0.746  0.45564
NINQ23       1.062e-01  3.942e-01   0.269  0.78768
NINQ24       9.397e-01  3.904e-01   2.407  0.01609 *
DEROG21      9.571e-01  4.209e-01   2.274  0.02298 *
DEROG22      1.535e+00  4.955e-01   3.097  0.00195 **
DELINQ21     7.073e-01  3.128e-01   2.261  0.02376 *
DELINQ22     1.797e+00  3.786e-01   4.746 2.07e-06 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

K-S y ROC Modelo Modelo 2

KS= 0,45

AUC curva de ROC= 0,7989

Modelo 2	Entrenamiento	Testeo
Accuracy	0,7198	0,7295
Sensibilidad	0,6926	0,7000
Especificidad	0,7559	0,7692

Peso de atributos

Ahora continuamos con el cálculo del WOE para las variables que quedaron de stepwise
LOAN + JOB + YOJ + CLAGE + CLNO + DEBTINC + NINQ2 + DEROG2 + DELINQ2 + GARANTIA

Eligiendo todas las variables del modelo, variables con valor de informacion menor a 0.02 deberían ser descartadas (en este caso se descartan las variables CLAGE , DEBTINC y GARANTIA).

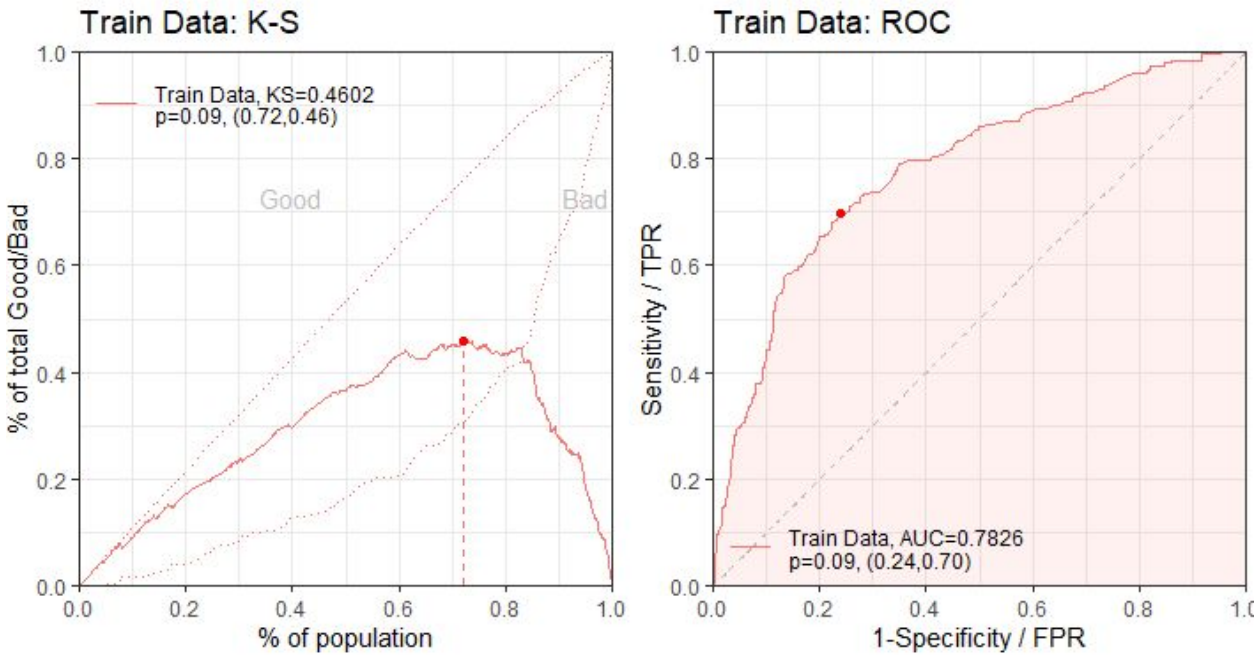
```
1  LOAN          0.716
2  DELINQ2       0.395
3  DEROG2        0.339
4  CLNO          0.337
5  YOJ           0.191
6  NINQ2         0.137
7  JOB           0.058
8  GARANTIA      0.013
9  CLAGE         0
10 DEBTINC       0
```

Posterior a ellos, se realiza una transformación de las variables restantes a woe, con el fin de que pasen a ser continuas y ver si logran predecir de mejor manera, sin embargo, este tipo de transformación pierde interpretabilidad en el modelo.

Scorecard de entrenamiento

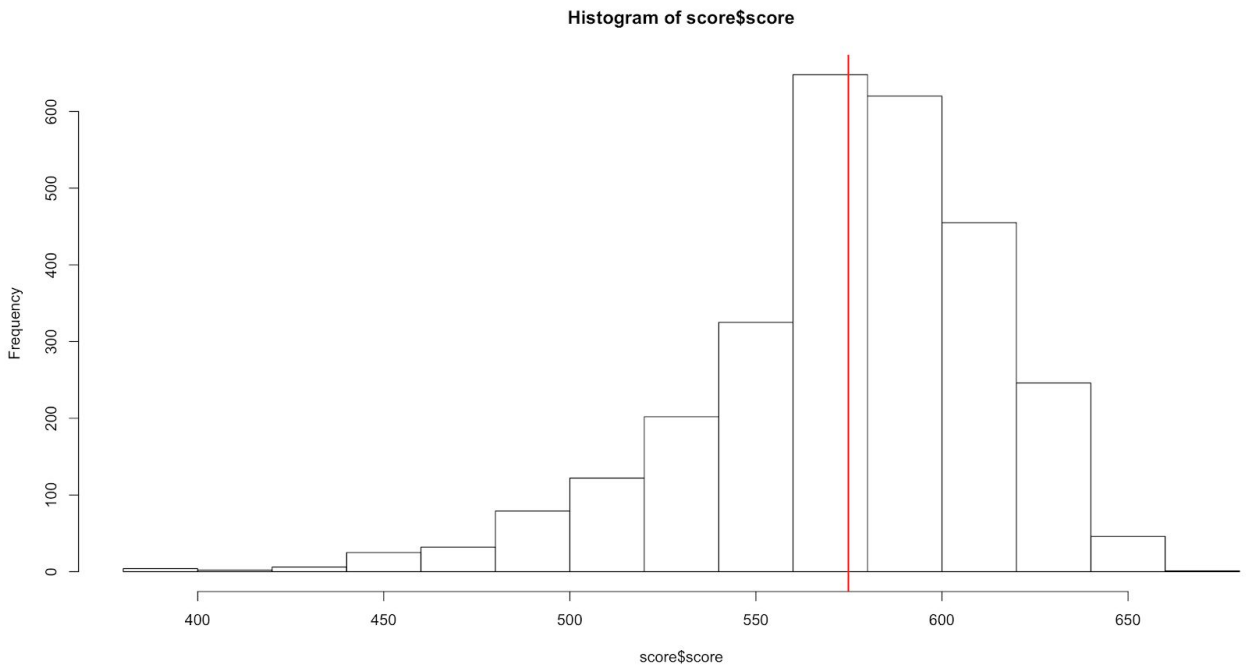
```
Coefficients:
             Estimate Std. Error z value Pr(>|z|)
(Intercept)  -2.3218     0.1524 -15.232  < 2e-16 ***
LOAN_woe      0.2412     0.4323   0.558 0.576899
YOJ_woe       1.0161     0.5000   2.032 0.042148 *
CLNO_woe      0.7787     0.4223   1.844 0.065173 .
NINQ2_woe     1.1394     0.3261   3.494 0.000475 ***
DEROG2_woe    0.8140     0.2659   3.061 0.002205 **
DELINQ2_woe   0.9480     0.2091   4.534 5.78e-06 ***
---
```

Logran con un nivel de 0,01 de significancia las variables NINQ2, DEROG2, DELIQ2, con transformación woe sean significativas, a continuación mediremos su calidad predictiva:



El modelo bajo metodología scorecard y transformación de variables WOE, presenta un KS: de 9,4602 y un AUC de 0,7826, ambos valores son considerados buenos en el modelo. Sin embargo al analizar la especificidad del modelo , estos no logran llegar a un nivel superior de 0,8 en ninguna de las dos muestras (entrenamiento y testeo)

Modelo 3: SCORECARD	Entrenamiento	Testeo
Accuracy	0,9157	0,9188
Sensibilidad	0,9184	0,9232
Especificidad	0,6774	0,6667



El scorecard promedio es de 564,2, por lo que se podría diseñar un modelo en donde los nuevos clientes que tengan un scorecard superior a este número se le otorgue el crédito solicitado.

Rango de Puntos - Scorecard

	score
1:	523
2:	548
3:	544
4:	548
5:	548

2809:	583
2810:	583
2811:	583
2812:	559
2813:	583

Utilización del modelo predictivo y conclusiones

El modelo para evaluar la probabilidad de cada cliente es el modelo 1 testeado , que consta finalmente de 7 variables.

$$BAD = \beta 0 + \beta 1 * LOAN + \beta 2 * CLAGE + \beta 3 * CLNO + \beta 4 * DEBTINC + \beta 5 * NINQ2 + \beta 6 * DEROG2 + \beta 7 * DELINQ2$$

El modelo final presenta un accuracy y sensibilidad por sobre 0,9 por lo que señala que logra predecir bien a los clientes que si se les debe otorgar un crédito, sin embargo, el modelo aún presenta algunas falencias para lograr predecir a los malos clientes, es decir, aquellos quienes no se les debiese otorgar el crédito, por lo que se hace necesario en un futuro evaluar nuevas variables a considerar para poder lograr un modelo con una mejor calidad predictiva.

Modelo 1	Entrenamiento	Testeo
Accuracy	0,925	0,9217
Sensibilidad	0,9296	0,9284
Especificidad	0,7857	0,6667

Medida de asociación (OR) e interpretaciones:

Variable	Descripción	coeficiente	OR (medida de asociación)
BAD	1: solicitante incumplido en préstamo 0: préstamo pagado por el solicitante	variable dependiente	-
LOAN**	Monto de la solicitud de préstamo	-0,0000205	0,99
DEROG2	Número de informes despectivos importantes	categoría referencia	-
DEROG21	1 informe despectivo importante	0,217	1,24
DEROG22***	mayor o igual a 2 informes	1,94	6,95
DELINQ2	Número de líneas de crédito morosas (cero línea de crédito morosa)	categoría referencia	-
DELINQ21***	una línea de crédito morosa	0,772	2,16
DELINQ22***	mayor o igual a 2 lineas de credito morosas	1,91	6,77

CLAGE***	Edad de la línea de crédito más antigua (meses)	-0,00503	0,99
NINQ2	Número de consultas de crédito recientes (0 consultas)	categoría referencia	
NINQ21	1 consulta	-0,227	0,79
NINQ22	2 consultas	-0,211	0,80
NINQ23	3 consultas	-0,169	0,84
NINQ24***	4 o más consultas	1,00	2,72
CLNO*	Número de líneas de crédito	-0,0142	0,98
DEBTINC***	Relación deuda-ingreso	0,116	1,12

LOAN: Ante la variación de una unidad en el monto de solicitud del préstamo, la persona tiene una menor probabilidad de que incumpla el préstamo del 1% (1-0,99), siendo una variable significativa con un nivel de significancia de 0,01.

DEROG22: Las personas que cuentan con 2 o más informes respectivos importantes presentan 6,95 veces más de probabilidad de que incumpla el préstamo en relación a los que tienen cero informe despectivo importante, siendo una variable estadísticamente significativa, a un nivel de significancia del 0,001.

CLAGE ante la variación de una unidad en la edad de la cuenta más antigua , disminuye la probabilidad de que incumpla el crédito en un 1% (1-0,99) , siendo una variable estadísticamente significativa, a un nivel de significancia del 0,001.

CLNO:Ante la variación de una unidad en el número de líneas de crédito , el cliente disminuye la probabilidad de que incumpla el crédito en un 2% (1 - 0,98), siendo una variable estadísticamente significativa, a un nivel de significancia del 0,1

DEBTINC: Ante la variación de una unidad en la relación deuda-ingreso , el cliente presenta 1,12 veces mas de probabilidad de que incumpla el crédito , siendo una variable estadísticamente significativa, a un nivel de significancia del 0,001.

NINQ24: Los clientes que cuentan con 4 o más números de consultas de crédito recientes presentan 2,72 veces más de probabilidad de que incumpla el crédito con respecto a los clientes que tienen cero número de consultas de crédito recientes, siendo una variable estadísticamente significativa, a un nivel de significancia de 0,001

DELINQ21: Los clientes que tienen una línea de crédito morosa tienen 2,26 veces más de probabilidad de que incumpla el crédito en relación a los clientes que tienen cero línea de crédito morosa, siendo una variable estadísticamente significativa, a un nivel de significancia del 0,001.

DELINQ22: Los clientes que tienen dos o más líneas de crédito morosas tienen 6,77 más de probabilidad de que incumplan el crédito en relación a los clientes que tienen cero línea de crédito morosa, siendo una variable estadísticamente significativa, a un nivel de significancia del 0,001.