Universidad Nacional Autónoma de México

Facultad de Ciencias

Score de crédito aplicando técnicas de minería de datos (caso de estudio)

T E S I S

que para obtener el grado de:
Actuario

presenta:
Mauricio Vallejo Castañón

director de tesis:
Act. Edgar Díaz Ordoñez

Ciudad Universitaria, México DF Noviembre, 2015

Score de crédito aplicando técnicas de minería de datos (caso de estudio)

por

Mauricio Vallejo Castañón

Tesis presentada para obtener el grado de

Actuario

en la

Facultad de Ciencias

Universidad Nacional Autónoma de México

Ciudad Universitaria, México DF. Noviembre, 2015

Índice general

Introducción

Uno de los principales problemas que enfrenta la industria del crédito es el deterioro de la cartera crediticia. La razones que llevan a este empeoramiento del negocio son diversas: mala elección de los clientes, determinación incorrecta de los montos a prestar, estrategias de cobranza deficientes, etc.

Para mitigar el riesgo que se presenta en cada uno de estos factores, se han desarrollado diversas metodologías. En particular, para la selección adecuada de clientes, se tienen *modelos de originación o credit scoring* que se construyen con técnicas de *minería de datos* y este tipo de modelos es el tema que motiva el desarrollo de esta tesis.

Esta tesis tiene como objetivo la desarrollo de una metodología de *originación de créditos*, que sirva como herramienta para evaluar el riesgo de incumplimiento asociado a cada cliente que hace una solicitud de crédito para que a partir de esta evaluación se determinen los casos en los cuales es viable autorizar el préstamo.

El desarrollo de este trabajo se hizo con el contexto de una financiera y las características de su operación, se describen en el Capítulo 1. En el Capítulo 2, se construye el criterio para determinar si un cliente es *bueno* o *malo*; además de que se describe el proceso para construir las tablas que sirven como insumo para el desarrollo del *modelo de originación*. Por otra parte, el Capítulo 3 contiene la teoría de las técnicas de *minería de datos* utilizadas para la construcción de dicho modelo. Y finalmente, el Capítulo 4 es la consecución del modelo, en el se describen los resultados y conclusiones.

Capítulo 1

Contexto del negocio

En la práctica, antes de iniciar con el desarrollo de un modelo se tiene que analizar el negocio y entender el entorno en el que este se encuentra operando. Este entendimiento, es sumamente importante para que se pueda identificar la necesidad del negocio y poder proponer soluciones que hagan sentido con el contexto.

Para el desarrollo de la metodología de *originación de créditos*, se consultó información de una financiera. Esta información abarca 7 años de operación de la empresa, del año 2006 hasta el año 2012. De esta consulta, se examinó información relacionada con las solicitudes de crédito y el comportamiento de las cuentas que integran la cartera de crédito.

En este capítulo, se describen algunas de las reglas que define la financiera para su operación y se muestran algunos indicadores que describen de forma general la eficiencia de esta empresa.

1.1. Operación de la empresa

Esta financiera, se enfoca en otorgar créditos para el consumo y maneja dos tipos de productos principalmente:

- **Consumo:** créditos para personas que perciben un salario.

- **Micro:** créditos para personas que cuentan con un micronegocio.

Los montos de los préstamos pueden ser de $ 1,500 a $ 70,000 pesos. Y se dividen en créditos semanales, quincenales y mensuales; según la frecuencia con la que el cliente percibe sus ingresos.

1.1.1. Gestión de solicitudes

La financiera establece las siguientes políticas para poder evaluar una solicitud:

- Nacionalidad mexicana.
- Solicitud del Buró de Crédito (BC).
- Edad mínima de 18 años y máxima de 75 años.
- Comprobante de domicilio.
- Comprobante de ingresos.
- Referencias familiares y personales.
- Aval o garantía.

Así mismo, la empresa establece que las solicitudes con las siguientes características deben rechazarse:

- Si la solicitud de crédito no cumple alguna de las políticas anteriormente descritas.
- Si el reporte de BC presenta un *score* bajo o mensajes negativos y de alerta.
- Si el reporte de BC presente cuentas con 60 días de mora o mayor, actual o en el histórico de pagos.
- Si el número de consultas en los tres meses más recientes es muy alto, ya que puede ser indicativo de algún problema.
- Si la capacidad de pago es muy baja.
- Si el ingreso mensual disponible para cada integrante de la familia es bajo.

Por otra parta, una vez que se ha evaluado la solicitud y se ha determinado que es viable otorgar un crédito, se lleva a cabo el cierre de operación; el cual, consiste en:

Establecimiento de pagos:

- Informar al solicitante la aprobación, el monto y el plazo de su préstamo.
- Establecer los días de pago.
- Requerir que para el cierre de préstamo estén presentes todas las personas involucradas en la solicitud.
- Fijar día y hora del cierre de préstamo.

Entrega del crédito:

- Explicar al solicitante las características del préstamo: monto del préstamo, comisión por apertura de crédito, número de pagos a realizar, pagos a capital, intereses, IVA y días de pago.
- Explicar las responsabilidades que se adquieren.
- Entregar una copia de los 10 puntos importantes para el manejo de su crédito.
- Firma de los documentos correspondientes.

1.1.2. Estructura de la cartera

La financiera estructura su cartera por categorías de morosidad, conocidas como *buckets*; los cuales, indican los días de atraso que tienen las cuentas. La cartera se integra con las cuentas que tienen de 0 a 119 días de atraso en sus pagos y se encuentran distribuidas del *bucket 0* al *bucket 4*. Si alguna cuenta llega a tener 120 días de atraso o más, se considera pérdida y forma parte del *bucket 5*. Esta clasificación de la cartera, se muestra en la Tabla 1-1.

Bucket	Días de atraso
0	0 días
1	1 a 29 días
2	30 a 59 días
3	60 a 89 días
4	90 a 119 días
5	120 o más días (pérdida)

Tabla 1-1: Categorías de morosidad.

La clasificación de las cuentas que conforman la cartera, se hace al término de cada mes y tiene la finalidad de conocer el deterioro que tiene la cartera; para que a partir de ahí, se tomen decisiones y se desarrollen diversas estrategias, como las de cobranza.

1.2. Eficiencia de la empresa

Para medir la eficiencia de la financiera, se tienen principalmente tres tipos de reportes; estos son: *through the door*, *vintage* y *coincidental*. Con este conjunto de indicadores, se puede conocer y analizar el negocio en cuanto a la captación de solicitudes, la colocación de créditos y el comportamiento de las cuentas. En esta sección, se muestran algunos de los resultados que se tienen del negocio al término del año 2012.

1.2.1. Through the door

Los indicadores que integran este reporte, tienen la finalidad de resumir algunos temas de interés sobre la llegada de solicitudes. Principalmente, en este reporte se muestra la cantidad de solicitudes de crédito que llegan a las sucursales mes a mes, la proporción de solicitudes aceptadas y la proporción de solicitudes rechazadas (Figura 1-1).

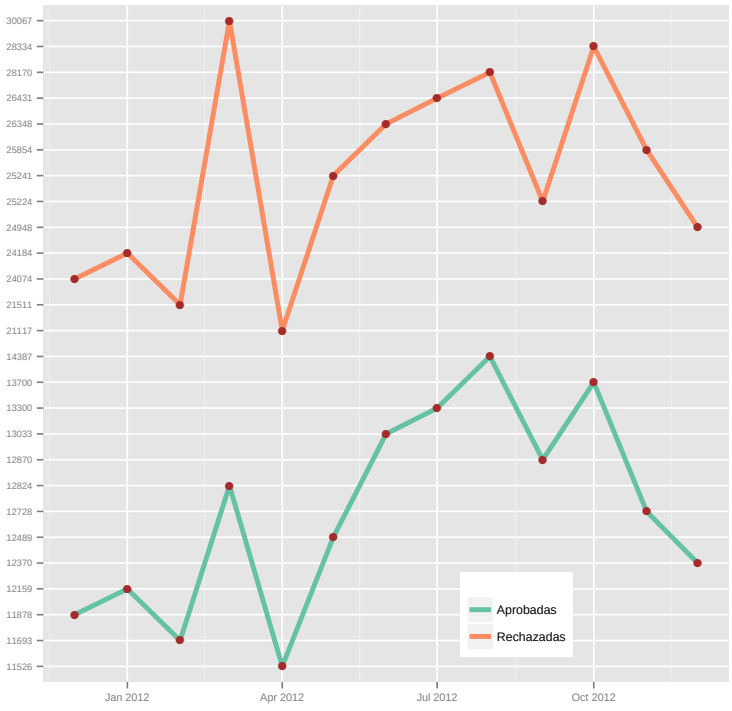

Figura 1-1: Solicitudes recibidas.

En la Figura 1-1, se muestra información desde diciembre de 2011 hasta diciembre de 2012; en donde se puede apreciar que la llegada de clientes tiene un comportamiento estacional que es debido a las necesidades económicas que presentan los clientes en algunos épocas del año. Por otra parte, se ve que se rechazan más solicitudes de las que se aceptan; esto puede ser debido a que las estrategias de mercadotecnia atraen a clientes con perfiles poco adecuados.

Por otra parte, en este reporte también se puede mostrar información de interés como el monto promedio de las solicitudes, la proporción de clientes aceptados con perfil de alto riesgo o las principales razones por las cuales se rechazan las solicitudes (Figura 1-2).

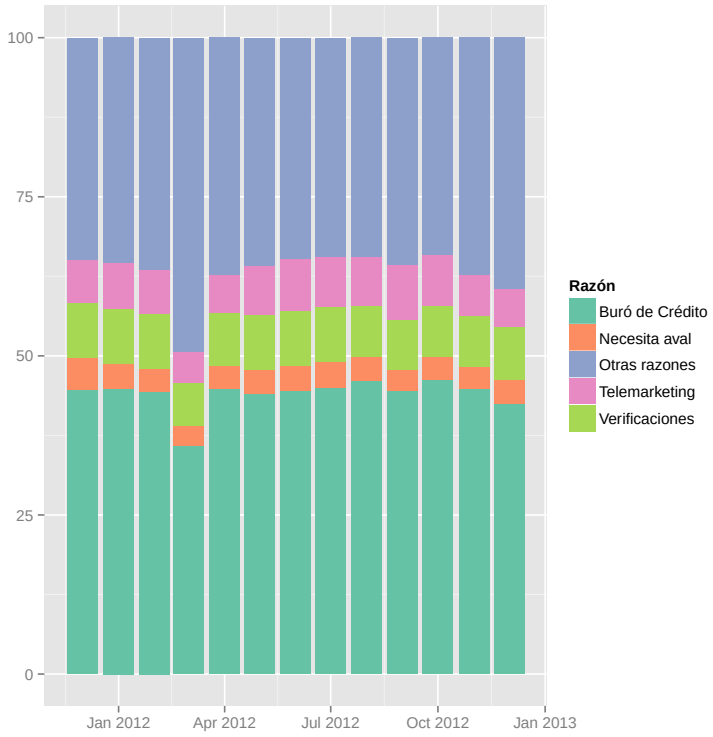

Figura 1-2: Razones de rechazo.

De la Figura 1-2, se puede observar que la principal razón de rechazo es la mala calificación en BC. Esta evaluación que hace Buró de Crédito a los clientes es confiable, ya que posee información de los clientes sobre otros sectores de crédito; con lo cual, es posible conocer y obtener algunas conclusiones de forma general acerca del comportamiento de los clientes. Por otra parte,

la segunda categoría que tiene mayor proporción de clientes rechazados es «Otras razones»; este tipo de información no es útil para negocio por lo que se tendría que hacer un esfuerzo por tener más detalle sobre estas solicitudes rechazadas.

En general, de este reporte se tiene que durante el año 2012 la financiera tuvo una abundante llegada de solicitudes de crédito respecto a otros; con lo cual, se ve que el negocio sigue atrayendo clientes.

1.2.2. Vintage

Este reporte es un conjunto de indicadores que muestran el comportamiento de las cuentas colocadas. Comúnmente en la industria del crédito es de interés medir qué proporción de las cuentas colocadas han llegado a tener 30, 60 o 90 días de mora en sus pagos en algún intervalo de tiempo; esto es para conocer la proporción de cuentas que se están deteriorando y tomar las medidas necesarias para controlar el riesgo. También, la finalidad de este reporte es conocer el porcentaje de cuentas que no se pudieron cobrar y se fueron directo a pérdida en menos de un año.

A continuación se describen tres indicadores de morosidad utilizados en la empresa:

- **E30 @ 7 mob:** este indicador muestra la proporción de cuentas colocadas que han llegado a tener 30 días o más de vencimiento en sus pagos, en un periodo de observación de 7 meses.
- **E90 @ 9 mob:** este indicador muestra la proporción de cuentas colocadas que han llegado a tener 90 días o más de vencimiento en sus pagos, en un periodo de observación 9 meses.
- **ANR @ 9 mob:** este indicador muestra la proporción de cuentas colocadas que se han convertido en pérdida, en un periodo de observación de 9 meses.

En la práctica, las ventanas de observación y los indicadores de morosidad se definen según las necesidades de negocio. Estos indicadores que se describieron anteriormente se muestran en la Figura 1-3 y pertenecen a la colocación de 2011 y el primer trimestre de 2012.

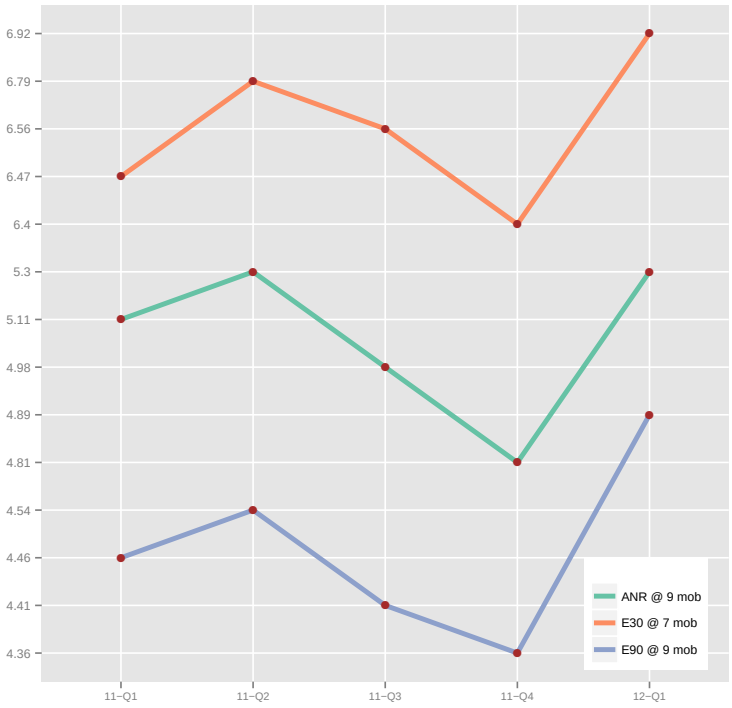

Figura 1-3: Indicadores de colocación.

De la Figura 1-3, se puede observar que el indicador *E30 @ 7 mob* está por debajo del 7 %; este resultado muestra que la proporción de cuentas colocadas que han llegado a tener 30 días de mora en un periodo de observación de 7 meses, representa una proporción muy pequeña del total de cuentas colocadas en el transcurso de 2011 e inicios de 2012. Interpretando de igual manera los indicadores *E90 @ 9 mob* y *ANR @ 9 mob*, se puede inferir de forma empírica que el riesgo en el que ha incurrido la financiera es bajo y controlable.

1.2.3. Coincidental

Este reporte tiene la finalidad de mostrar como está conformada la cartera en cuentas y en saldo (Figura 1-4), conocer la situación de mora que tienen las cuentas y mostrar la pérdida que ha tenido la empresa (Figura 1-5) para tener una visión general de como va el negocio.

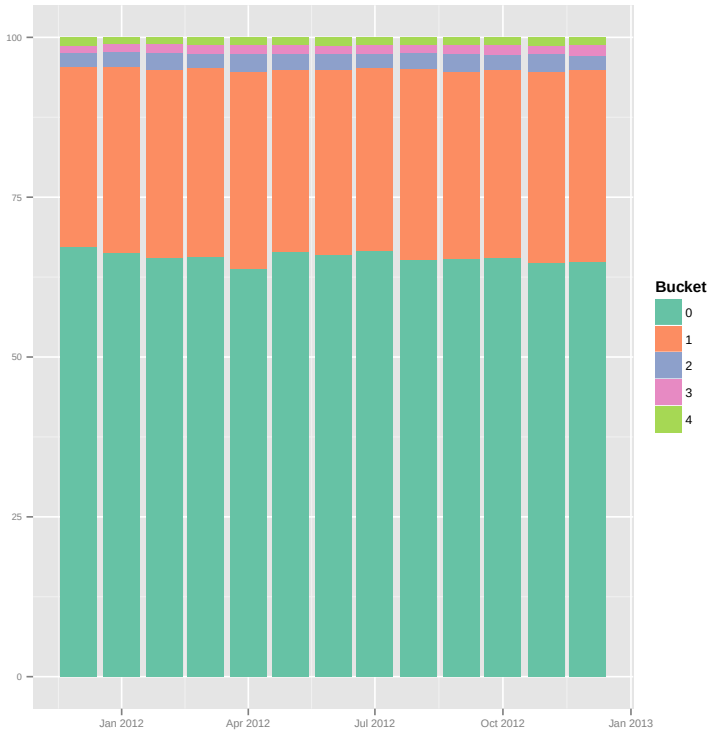

Figura 1-4: Distribución de la cartera.

De la Figura 1-4, se puede ver que en general la cartera es saludable ya que el *bucket 3* y *4* representan una proporción pequeña del dinero prestado, mientras que los clientes cumplidos representan más del 60 % de la cartera.

Por otra parte, para medir la morosidad de los clientes que integran la cartera, a continuación se describen tres indicadores utilizados en la empresa:

- **30+ %:** este indicador muestra la proporción de cuentas que tienen saldos con 30 días o más de vencimiento en sus pagos.
- **90+ %:** este indicador muestra la proporción de cuentas que tienen saldos con 90 días o más de vencimiento en sus pagos.
- **ANR %:** este indicador muestra la pérdida anualizada en la que ha incurrido la empresa.

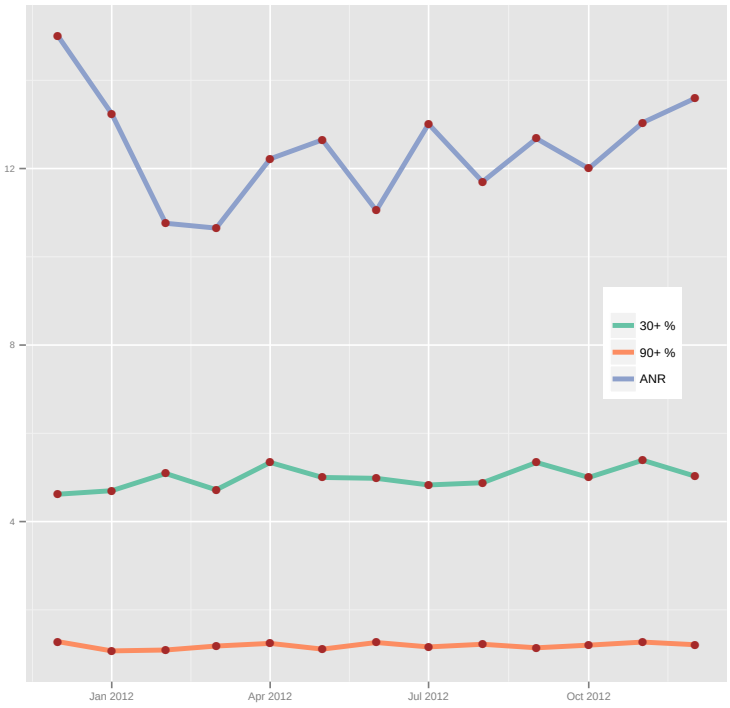

Figura 1-5: Indicadores de la cartera.

De la Figura 1-5, se puede observar que el indicador *30+ %* está por debajo del 5 %; este resultado muestra que la proporción de cuentas que han integrado la cartera durante el 2012 y que han llegado a tener 30 días de mora, representa una proporción muy pequeña del total de la cartera. Interpretando de igual manera los indicadores *90+ %* y *ANR %*, se infiere empíricamente que el riesgo en el que ha incurrido la financiera es mínimo y controlable.

1.3. Conclusiones

De los resultados observados en los tres reportes, se tiene que durante el 2012 la financiera tuvo un crecimiento promedio mensual de 0.49 %. Y se sabe que el negocio es rentable por la distribución que tiene la cartera en los diferentes niveles de morosidad, los indicadores en la colocación que muestran una tasa baja de riesgo y la pérdida anualizada que también se ha mantenido baja.

A pesar de que la empresa no presenta malos resultados, se busca tener una selección más eficiente de clientes y mantener o reducir el riesgo en el que se incurre, para esto es el desarrollo de la metodología de *originación de créditos* que se presenta en este trabajo.

Capítulo 2

Análisis exploratorio de datos

Como se menciona anteriormente, se pretende desarrollar una herramienta para tomar la decisión de aprobar o rechazar una solicitud de crédito. Esta herramienta es un *score de originación*; el cual, es un modelo de riesgo que clasifica a los solicitantes en clientes potencialmente *buenos* o *malos*.

Para el desarrollo de este modelo, se necesita definir el concepto de *malo* y de *bueno*; es decir, se tiene que construir una *variable objetivo* para el modelo. Por otra parte, se tienen que seleccionar ciertos atributos sociales y económicos que contengan información relevante para identificar a aquellos clientes que incurren en impago. Con estos atributos seleccionados, se construye una *tabla base analítica* (*ABT* por sus siglas en inglés); la cual, servirá de insumo para desarrollar el *modelo de clasificación*.

En este Capítulo 2, se presentan los análisis que se desarrollaron para la construcción de la *variable objetivo* y la construcción de la *ABT*. El desarrollo se hizo con el *software libre* R versión 3.1.2.

2.1. Construcción de la variable objetivo

Se realiza un análisis del comportamiento de las cuentas que han integrado la cartera. La finalidad es identificar el momento en el que las cuentas que presentan adeudos ya no se recuperan y pueden ser catalogadas como *malas*. Con esto, se establece la base para la definición de la *variable objetivo* del modelo.

En este análisis, se utiliza de insumo la tabla «Portafolio»; la cual, contiene información financiera (monto de los pagos, plazo, *bucket*, etc.) de las cuentas que han entrado a la cartera. De esta tabla, se consideran 555,202 cuentas colocadas desde enero de 2006 hasta diciembre de 2012 con un valor por arriba de los $ 5,400 millones de pesos.

2.1.1. Análisis de transición

Se realiza un análisis a través del tiempo; en el cual, se observa el paso de las cuentas por las diferentes categorías de morosidad que integran la cartera, desde el momento en que entran a la cartera hasta el momento en que salen de ella, ya sea porque la cuenta se liquidó o porque se convirtió en pérdida.

Con un enfoque de *cadenas de Markov*, se generaron las *probabilidades de transición* que tienen las cuentas de moverse entre cada uno de los *buckets* a través de los meses que llevan en la cartera (mob[1]).

Se comienza analizando los *buckets*; en los cuales, se encontraban distribuidas las cuentas en su primer y segundo *mob*. Con esto, se pueden observar las cuentas que fueron liquidadas y las que fueron enviadas a pérdida (*bucket* 5) tempranamente. Los resultados se muestran en la Figura 2-1.

[1]En la industria del crédito, a los meses que llevan las cuentas en la cartera se les conoce como *mob* (del acrónimo *months of books*).

Mob 1	
Bucket	**Cuentas**
Liquidan	2,904
0	529,655
1	22,642
2	-
3	-
4	-
5	1

Mob 2	
Bucket	**Cuentas**
Liquidan	25,929
0	463,049
1	65,744
2	466
3	-
4	-
5	14

Figura 2-1: Distribución de las cuentas en los *mob* 1 y 2.

Observando el recorrido que estas cuentas hicieron del *mob* 1 al *mob* 2, se construye la Tabla 2-1 que detalla estas transiciones. En esta tabla, las filas representan los *buckets* de las cuentas en el *mob* 1 y las columnas representan los *buckets* de las cuentas observadas en el *mob* 2.

Bucket	**Liq**	**0**	**1**	**2**	**3**	**4**	**5**	**Total**
Liq	2,904	-	-	-	-	-	-	2,904
0	21,792	454,139	53,714	1	-	-	9	529,655
1	1,234	8,910	12,030	465	-	-	3	22,642
2	-	-	-	-	-	-	-	-
3	-	-	-	-	-	-	-	-
4	-	-	-	-	-	-	-	-
5	1	-	-	-	-	-	1	1
Total	25,930	463,049	65,744	466	-	-	13	555,202

Tabla 2-1: Transición de las cuentas entre *mob* 1 y *mob* 2.

A partir de la Tabla 2-1, se estiman las probabilidades de transición que tiene una cuenta al moverse de un *bucket* a otro, pasando del *mob* 1 al *mob* 2.

Por ejemplo, en *mob* 1 se tienen 22,642 cuentas que están clasificadas como *bucket* 1. Estas cuentas al siguiente mes (*mob* 2) se encontraban distribuidas de la siguiente manera:

- 1,234 cuentas fueron liquidadas (*bucket* Liq).
- 8,910 regresaron al *bucket* 0.
- 12,030 se mantuvieron en el *bucket* 1.
- 465 avanzaron al *bucket* 2.
- 3 se fueron directo a pérdida (*bucket* 5).

Si se calcula la proporción de cuentas que hay en cada *bucket* del *mob 2*, respecto al total de cuentas que se encontraban en *bucket* 1 en el mes anterior (22,642 cuentas), se puede estimar la probabilidad de transición a cada uno de estos *buckets* (Tabla 2-2).

Liq	0	1	2	3	4	5	Total
1,234	8,910	12,030	465	-	-	3	22,642
5.45 %	39.35 %	53.13 %	2.05 %	-	-	0.01 %	100 %

Tabla 2-2: Transición de las cuentas clasificadas en *bucket* 1 y *mob* 1.

De lo anterior, se tiene un diagrama (Figura 2-2) que ilustra los movimientos entre *mob* 1 y *mob* 2, que tuvieron las cuentas del *bucket* 1.

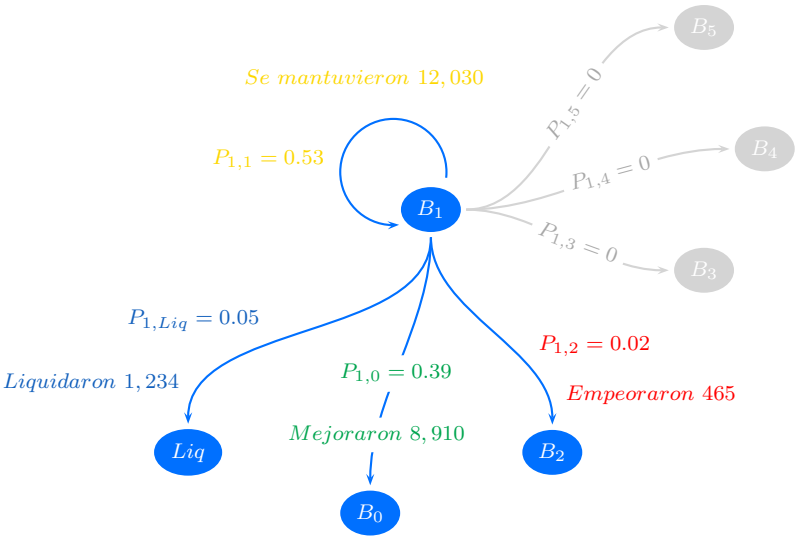

Figura 2-2: Transición de las cuentas que están en *bucket* 1 y *mob* 1.

Para las cuentas que en *mob* 1 están clasificadas en el *bucket* 1, este diagrama se puede interpretar de la siguiente forma:

- la probabilidad de liquidar el crédito al siguiente mes es de 0.05.
- la probabilidad de ponerse al corriente al siguiente mes es de 0.39.
- la probabilidad de mantenerse igual al siguiente mes es de 0.53.
- la probabilidad de empeorar al siguiente mes es de de 0.02.
- la probabilidad de convertirse en pérdida al siguiente mes es 0.

Estas proporciones se calculan de igual manera para los demás *buckets* obteniendo así la matriz de transiciones del *mob* 1 al *mob* 2 que se muestra en la Figura 2-3.

	Liq	B_0	B_1	B_2	$\cdots$	B_5
Liq	1	0	0	0	$\cdots$	0
B_0	0.04	0.86	0.10	0	$\cdots$	0
B_1	0.05	0.39	0.53	0.02	$\cdots$	0
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\ddots$	0
B_5	0	0	0	0	$\cdots$	1

Figura 2-3: Probabilidades de transición de cuentas entre *mob* 1 y *mob* 2.

Este proceso se realiza para los diferentes *mob* que tenían las cuentas y también se hace para los saldos de la cuentas. Los resultados se resumen en la Tabla 2-3 y Tabla 2-4, en donde se presenta la probabilidad que tienen las cuentas y sus saldos de mejorar, de mantenerse y de deteriorarse; pasando del mes de observación n al mes $n+1$.

En las *probabilidades de mejorar* y *mantenerse*, se marca con verde la probabilidad más grande y de ahí se degrada el color hasta llegar al rojo; el cual, representa la probabilidad más baja. En el caso de la *probabilidad de empeorar*, se pinta de verde la probabilidad más baja y con rojo la más alta.

Lo anterior, se hace con la finalidad de hacer más visibles las características que tiene cada *bucket*. Entonces, en el resumen de la Tabla 2-3, se tiene que los escenarios favorables para el negocio son los que tienen un verde más intenso; por el contrario, los escenarios rojos son los que son los menos favorables.

Mes	Mejora					Se mantiene					Empeora				
$m_{n,n+1}$	0	1	2	3	4	0	1	2	3	4	0	1	2	3	4
$m_{1,2}$	0.04	0.44	-	-	-	0.86	0.53	-	-	-	0.10	0.02	-	-	-
$m_{2,3}$	0.06	0.38	0.12	-	-	0.84	0.57	0.23	-	-	0.11	0.05	0.65	-	-
$m_{3,4}$	0.08	0.36	0.05	0.05	-	0.81	0.58	0.10	0.02	-	0.11	0.06	0.84	0.93	-
$m_{4,5}$	0.12	0.36	0.06	0.06	0.03	0.77	0.58	0.11	0.01	0.00	0.11	0.07	0.83	0.93	0.97
$m_{5,6}$	0.14	0.36	0.07	0.07	0.04	0.74	0.57	0.13	0.02	0.01	0.12	0.08	0.80	0.91	0.95
$m_{6,7}$	0.14	0.36	0.08	0.09	0.05	0.73	0.58	0.14	0.03	0.02	0.13	0.07	0.78	0.88	0.93
$m_{7,8}$	0.14	0.35	0.11	0.10	0.06	0.73	0.59	0.13	0.03	0.02	0.13	0.07	0.76	0.87	0.93
$m_{8,9}$	0.14	0.33	0.14	0.12	0.07	0.72	0.60	0.12	0.04	0.03	0.14	0.07	0.74	0.84	0.91
$m_{9,10}$	0.14	0.33	0.14	0.12	0.07	0.72	0.60	0.13	0.04	0.03	0.14	0.07	0.73	0.84	0.89
$m_{10,11}$	0.11	0.33	0.16	0.13	0.09	0.21	0.60	0.13	0.04	0.04	0.67	0.07	0.71	0.83	0.87
$m_{11,12}$	0.16	0.32	0.16	0.14	0.09	0.69	0.60	0.14	0.05	0.04	0.15	0.08	0.69	0.81	0.87
$m_{12,13}$	0.22	0.35	0.19	0.15	0.11	0.64	0.56	0.11	0.04	0.05	0.15	0.09	0.70	0.81	0.84
$m_{13,14}$	0.15	0.34	0.24	0.15	0.11	0.69	0.57	0.08	0.06	0.05	0.15	0.09	0.67	0.80	0.84
$m_{14,15}$	0.15	0.31	0.26	0.19	0.11	0.70	0.61	0.09	0.05	0.05	0.16	0.07	0.65	0.76	0.83
$m_{15,16}$	0.17	0.33	0.21	0.21	0.14	0.68	0.60	0.11	0.04	0.06	0.15	0.08	0.68	0.76	0.80
$m_{16,17}$	0.15	0.31	0.23	0.18	0.16	0.69	0.60	0.11	0.06	0.10	0.16	0.09	0.66	0.77	0.74
$m_{17,18}$	0.18	0.32	0.24	0.21	0.15	0.67	0.59	0.12	0.05	0.14	0.15	0.09	0.64	0.75	0.72
$m_{18,19}$	0.28	0.37	0.23	0.22	0.15	0.58	0.51	0.09	0.05	0.15	0.14	0.12	0.67	0.73	0.70
$m_{19,20}$	0.16	0.38	0.30	0.21	0.16	0.68	0.51	0.07	0.08	0.18	0.15	0.12	0.64	0.71	0.66
$m_{20,21}$	0.15	0.32	0.32	0.26	0.16	0.69	0.59	0.06	0.06	0.20	0.15	0.09	0.62	0.67	0.65
$m_{21,22}$	0.19	0.32	0.24	0.28	0.16	0.67	0.59	0.11	0.04	0.17	0.14	0.09	0.64	0.69	0.66
$m_{22,23}$	0.19	0.33	0.36	0.22	0.23	0.66	0.56	0.23	0.06	0.23	0.14	0.11	0.42	0.72	0.54
$m_{23,24}$	0.28	0.32	0.35	0.22	0.18	0.58	0.56	0.21	0.08	0.39	0.14	0.12	0.44	0.70	0.44
$m_{24,25}$	0.79	0.46	0.29	0.23	0.18	0.12	0.35	0.30	0.07	0.37	0.10	0.19	0.42	0.70	0.45

Tabla 2-3: Probabilidad de transición de las cuentas.

En la Tabla 2-3 se puede ver que las cuentas que se encuentran en *bucket* 3 en su tercer mes tienen un probabilidad de 0.93 de que al siguiente mes empeoren su situación (*bucket* 4 o 5), la probabilidad de que se mantengan en esa clasificación al siguiente mes es de 0.02 y la probabilidad de que mejoren su situación (*bucket* Liq, 1 ó 2) es de 0.06.

Observando de esta manera los demás meses, se puede concluir que una cuenta que llega al *bucket* 3 en sus primeros 12 meses tiene una probabilidad de recuperarse (inmediatamente 1 mes después) menor a 0.16.

De forma análoga, una cuenta que alcanza el *bucket* 4 en su primer año tiene una probabilidad de recuperarse en el siguiente mes inmediato de 0.12.

De forma general para cualquier mes, se puede ver que el *bucket* 3 y 4 presentan un rojo más intenso en las *probabilidades de mantenerse* y *empeorar*; con lo cual, se concluye que una cuenta que llega a estos estados de morosidad difícilmente se recupera y es muy probable que se convierta en pérdida.

En el resumen de la Tabla 2-4, se aprecia que los saldos tienen los mismos comportamientos que las cuentas. Y se concluye lo mismo, una vez que el saldo está en *bucket* 3 o 4 difícilmente se puede recuperar el crédito.

De las conclusiones obtenidas en la Tabla 2-3 y Tabla 2-4, se construye la base para la definición de la *variable objetivo*; la cual, se presenta en la sección 2.1.2.

Mes	Mejora					Se mantiene					Empeora				
$m_{n,n+1}$	0	1	2	3	4	0	1	2	3	4	0	1	2	3	4
$m_{1,2}$	0.08	0.47	-	-	-	0.82	0.51	-	-	-	0.09	0.02	-	-	-
$m_{2,3}$	0.09	0.40	0.12	-	-	0.81	0.55	0.22	-	-	0.10	0.05	0.67	-	-
$m_{3,4}$	0.11	0.38	0.07	0.05	-	0.78	0.57	0.10	0.01	-	0.11	0.06	0.83	0.94	-
$m_{4,5}$	0.15	0.37	0.07	0.05	0.02	0.74	0.56	0.10	0.01	0.00	0.11	0.06	0.83	0.94	0.98
$m_{5,6}$	0.16	0.38	0.09	0.08	0.05	0.72	0.55	0.12	0.02	0.01	0.11	0.07	0.79	0.91	0.95
$m_{6,7}$	0.17	0.38	0.11	0.09	0.05	0.71	0.56	0.12	0.02	0.01	0.12	0.06	0.77	0.88	0.93
$m_{7,8}$	0.17	0.37	0.13	0.10	0.06	0.70	0.57	0.12	0.03	0.02	0.12	0.06	0.75	0.87	0.92
$m_{8,9}$	0.18	0.36	0.16	0.13	0.07	0.70	0.57	0.11	0.03	0.03	0.13	0.06	0.73	0.84	0.90
$m_{9,10}$	0.18	0.37	0.17	0.12	0.07	0.69	0.57	0.11	0.03	0.03	0.13	0.06	0.72	0.85	0.90
$m_{10,11}$	0.17	0.37	0.18	0.13	0.09	0.20	0.57	0.11	0.04	0.03	0.63	0.06	0.70	0.83	0.88
$m_{11,12}$	0.20	0.37	0.21	0.14	0.09	0.67	0.57	0.12	0.04	0.03	0.13	0.06	0.67	0.82	0.88
$m_{12,13}$	0.21	0.37	0.23	0.15	0.11	0.66	0.57	0.10	0.03	0.04	0.13	0.06	0.67	0.81	0.85
$m_{13,14}$	0.20	0.37	0.26	0.15	0.11	0.66	0.57	0.10	0.03	0.03	0.14	0.06	0.64	0.81	0.86
$m_{14,15}$	0.21	0.38	0.26	0.17	0.11	0.66	0.56	0.11	0.04	0.03	0.14	0.06	0.63	0.79	0.87
$m_{15,16}$	0.22	0.38	0.25	0.17	0.12	0.65	0.56	0.10	0.04	0.03	0.13	0.06	0.65	0.79	0.84
$m_{16,17}$	0.23	0.38	0.28	0.16	0.13	0.64	0.55	0.10	0.04	0.03	0.14	0.07	0.61	0.80	0.84
$m_{17,18}$	0.24	0.40	0.29	0.20	0.13	0.62	0.53	0.10	0.03	0.03	0.13	0.07	0.60	0.76	0.84
$m_{18,19}$	0.25	0.41	0.31	0.19	0.15	0.62	0.52	0.09	0.04	0.03	0.13	0.07	0.60	0.77	0.82
$m_{19,20}$	0.25	0.41	0.33	0.22	0.14	0.61	0.52	0.10	0.06	0.04	0.14	0.07	0.57	0.72	0.81
$m_{20,21}$	0.28	0.41	0.32	0.27	0.13	0.58	0.52	0.09	0.04	0.04	0.13	0.07	0.58	0.69	0.83
$m_{21,22}$	0.33	0.43	0.31	0.20	0.15	0.55	0.50	0.10	0.05	0.03	0.12	0.08	0.58	0.76	0.82
$m_{22,23}$	0.41	0.46	0.41	0.22	0.20	0.47	0.45	0.20	0.06	0.02	0.12	0.09	0.39	0.72	0.78
$m_{23,24}$	0.56	0.51	0.41	0.23	0.13	0.33	0.37	0.19	0.06	0.07	0.10	0.12	0.40	0.72	0.81
$m_{24,25}$	0.82	0.62	0.41	0.22	0.16	0.10	0.20	0.22	0.06	0.08	0.09	0.18	0.37	0.72	0.76

Tabla 2-4: Probabilidad de transición los saldos.

2.1.2. Definición de la variable objetivo

De los análisis presentados anteriormente, se tiene que los estados en dónde se presenta una mayor *probabilidad de deterioro* son los *buckets* 3 y 4. Por lo tanto, estos son los estados candidatos para formar la definición de la *variable objetivo*.

Puesto que el objetivo de todo este análisis es identificar de forma oportuna cuando una cuenta ya no mejora; entonces, se propone el *buckets* 3 como base de la definición, debido a que con esta categoría de morosidad se logra identificar de forma anticipada a una cuenta que ya no mejora.

Entonces, la *variable objetivo* se define de la siguiente manera:

$$\textbf{Objetivo} := \begin{cases} \textit{Malo}: \ \textit{si la cuenta llega a} \text{ bucket } 3. \\ \textit{Bueno}: \ \textit{en otro caso.} \end{cases}$$

Después de que se construye la *variable objetivo*, se clasifican las cuentas analizadas en *malo* o *bueno*, pero en el momento en que todavía eran solicitudes de crédito. Teniendo esta clasificación en las solicitudes, se puede analizar de forma conjunta la *variable objetivo* con los atributos económicos y sociales que presenten los clientes. Lo anterior, se describe en la sección 2.2.

2.2. Análisis descriptivo y temporal

En esta sección se presenta el *análisis descriptivo* y *temporal* que se desarrolla con la finalidad de construir una *ABT* para las solicitudes de crédito analizadas. Esta tabla se construye con las variables que presentan una mejor distribución entre sus valores, un mayor efecto en el conjunto de datos y una buena estabilidad a través del tiempo.

Para el desarrollo de la *ABT*, se utiliza de insumo la tabla «Adquisicion»; de la cual, se analizaron 627,779 solicitudes aprobadas desde 2006 hasta 2012.

La tabla «Adquisicion» contiene 21 atributos; los cuales, fueron tomados de las solicitudes de crédito de los clientes en el momento en que estos la realizaron. Además, se incluye la *variable objetivo* que previamente se le asigna a las solicitudes aprobadas. Por otra parte, se incluyen 8 variables propuestas por negocio (son variables utilizadas en la operación) y que se generaron con los atributos originales.

2.2.1. Diccionario de datos de la tabla «Adquisicion»

A continuación se presenta un listado de las variables que contiene la tabla insumo, en este listado se muestra: el nombre, el tipo de dato[2], la descripción y los valores que presenta cada una de las variables.

- **I1** (id): número de cliente único asignado al solicitante, este valor se mantiene aún cuando el solicitante tenga más de una solicitud.
- **I2** (id): número de solicitud en la base de datos.
- **I3** (id): identificador de la sucursal donde se realizó la solicitud.
- **F1** (fecha): fecha en la que se realizó la solicitud del crédito.
- **F2** (fecha): fecha en la que se decidió la aprobación o el rechazo del crédito.
- **C1** (categórica nominal): variable que indica el género del solicitante. Los valores que tiene son los siguientes:
  - H: género masculino.
  - M: género femenino.
- **C2** (categórica nominal): frecuencia de pago del crédito solicitado. Los valores que tiene son los siguientes:
  - M: representa las solicitudes con frecuencia mensual.
  - Q: representa las solicitudes con frecuencia quincenal.
  - S: representa las solicitudes con frecuencia semanal.
- **C3** (categórica nominal): variable que indica el tipo de producto solicitado. Los valores que tiene son los siguientes:

[2]Los tipos de dato pueden ser: id, fecha, categórico nominal, categórico ordinal o numérico.

- C: créditos tipo consumo.
- M: créditos tipo micronegocio.

- **C4** (categórica ordinal): número de cuentas en las que el cliente tiene un atraso reportado en Buró de Crédito (BC). Esta variable contiene valores numéricos discretos; así que, se trata como variable categórica ordinal.
- **C5** (categórica nominal): variable que indica si la solicitud fue revisada y decidida por el Área Especial de Crédito (AEC). Los valores que tiene son los siguientes:
  - 0: no fue revisada por AEC.
  - 1: fue revisada por AEC.
- **C6** (categórica nominal): variable que indica si la solicitud fue gestionada por $Telemarketing$. Los valores que tiene son los siguientes:
  - 0: no se gestionó por $Telemarketing$.
  - 1: se gestionó por $Telemarketing$.
- **Objetivo** (categórica nominal): variable objetivo. Los valores que tiene son los siguientes:
  - $Malo$: si la cuenta llega a *bucket* 3.
  - $Bueno$: en otro caso.
- **T1** (categórica nominal): tipo de cliente que solicita el crédito, se usa como variable transversal. Los valores que tiene son los siguientes:
  - Nuevos: clientes que adquieren su primer crédito.
  - No nuevos: clientes que ya han adquirido al menos un crédito.
- **N1** (numérica): monto del préstamo total a financiar.
- **N2** (numérica): ingresos principales que declara el cliente.
- **N3** (numérica): plazo para pagar el total financiado.
- **N4** (numérica): años de vida que tiene el solicitante.
- **N5** (numérica): los meses de antigüedad en el trabajo que tiene el cliente.
- **N6** (numérica): los meses de antigüedad en el hogar que tiene el cliente.
- **N7** (numérica): remanente de ingresos del cliente.
- **N8** (numérica): saldo anterior al momento de solicitar renovación.
- **N9** (numérica): monto total a financiar del préstamo anterior.

Variables propuestas por negocio, construidas a partir de las originales:

- **C7** (categórica ordinal): número de créditos que tiene el cliente con la empresa. Esta variable contiene valores numéricos discretos; así que, se trata como variable categórica ordinal.

- **N10** (numérica): monto del pago que el cliente tendría que hacer en caso de aprobarse su crédito. Se calcula de la siguiente forma:

$$N10 = \frac{N1}{N3}$$

- **N11** (numérica): proporción del ingreso que el cliente tendría que destinar al pago de su crédito. Se calcula de la siguiente forma:

$$N11 = \frac{N10}{N2}$$

- **N12** (numérica): proporción del remanente de ingreso que tendrá el cliente en caso de aprobarse su crédito. Se calcula de la siguiente forma:

$$N12 = \frac{N7}{N2}$$

- **N13** (numérica): proporción de lo que le falta pagar al cliente al momento de pedir la renovación de su crédito. Se calcula de la siguiente forma:

$$N13 = \frac{N8}{N9}$$

- **N14** (numérica): variación de ciertos atributos que tiene el cliente entre su solicitud anterior y la actual. Los atributos considerados para medir la variación son las variables $N1$, $N3$, $N4$ y $N5$. Entonces, la variable $N14$ se calcula de la siguiente forma:

$$N14 = var_{N1} + var_{N3} + var_{N4} + var_{N5}$$

  donde:

$$var_{Ni} = \frac{\mid Ni_{anterior} - Ni_{actual} \mid}{Ni_{anterior}} \quad , \quad i \in 1, 3, 4, 5$$

- **N15** (numérica): esta variable es conocida en el negocio como *new money* e indica el monto que se le agrega al saldo del cliente al momento de renovar su crédito. Se calcula de la siguiente manera:

$$N15 = N1 \text{ - } N8$$

- **N16** (numérica): días que pasan entre la solicitud actual y la solicitud anterior. Se calcula de la siguiente forma:

$$N16 = F1_{nueva} \text{ - } F1_{anterior}$$

Cabe mencionar que las variables $C7$, $N13$, $N14$, $N15$ y $N16$ solo se construyen para las solicitudes de crédito de clientes que ya tenían aprobado al menos un crédito anteriormente. De esto se habla más a detalle cuando se presente el análisis de la *variable transversal*, $T1$ (sección 2.2.4).

En resumen los tipos de datos que contiene la tabla «Adquisicion» son los siguientes:

Tipo de variable	No. de variables
Identificador	3
Fecha	2
Categórica nominal	7
Categórica ordinal	2
Numérica	16

Tabla 2-5: Resumen de los tipos de datos.

La Tabla 2-5 muestra la naturaleza de las variables que se tienen disponibles, esto es importante para determinar el tipo de *análisis descriptivo* que le corresponde a cada una de ellas y también para determinar si se incluyen o se excluyen en los análisis que se describen a en la sección 2.2.2.

2.2.2. Método del análisis

En esta sección, se presenta el método del *análisis descriptivo* y *temporal* que se le aplica a las variables que describen los atributos de las solicitudes de crédito. Se excluyen 2 variables tipo identificador y 3 tipo fecha; por otra parte, se analizan 23 variables de atributo, 1 *variable transversal* y 1 *variable objetivo*.

En cuanto al *análisis descriptivo* que se presenta, se busca evaluar una perspectiva *bivariada*, se contrasta cada variable con la *variable objetivo* y se busca evaluar la concentración de los valores en cada una de las variables; ya que, se sabe que una gran concentración en un número reducido de valores resulta en un desempeño pobre de la variable para discriminar las observaciones.

Por otra parte, con el *análisis temporal*, se busca evaluar la estabilidad de las variables en el

tiempo; ya que a mayor estabilidad, mayor será la vigencia del *modelo de originación*.

Entonces, el método consiste en analizar la distribución de valores de la variables de atributo y su estabilidad en el tiempo. Para después clasificar a estas variables mediante un semáforo (Tabla 2-6); el cual, para el *análisis descriptivo bivariado* indica el grado de concentración de los valores y para el *análisis temporal* muestra el grado de estabilidad en el tiempo de las distribuciones.

Color	Distribución	Estabilidad
🟢	Correcta	Correcta
🟡	Aceptable	Aceptable
🔴	Deficiente	Deficiente

Tabla 2-6: Semáforo de clasificación de las variables.

A partir de lo que se observa en cada variable, se decide si la variable puede ser elegida o rechazada para integrar la ABT. En resumen, se tiene que si una variable muestra una fuerte concentración en la distribución de sus valores, un bajo efecto sobre el conjunto de datos y una deficiente estabilidad en el tiempo, se rechaza.

Por otra parte, como complemento y validación para la selección de variables numéricas se hace un *análisis descriptivo multivariado*, que también sirve como criterio cuantitativo para seleccionar variables. Con este análisis, se busca evaluar el efecto que tienen las variables numéricas en forma conjunta (contraria a la perspectiva *bivariada* que lo hace de forma aislada). Este análisis se basa en *análisis de componentes principales* (ver Apéndice A).

En la siguiente sección se muestra el análisis de la *variable objetivo* (sección 2.2.3) y la *variable transversal* (sección 2.2.4); las cuales, sirven de entrada al *análisis descriptivo* y *temporal* de las demás variables. Después, se presenta el análisis de las variables de atributo; el cual, se divide en variables categóricas (sección 2.2.5) y variables numéricas (sección 2.2.6 y 2.2.7).

2.2.3. Resumen del análisis de la variable objetivo

Objetivo: variable que clasifica a los créditos en *buenos* o *malos* según el comportamiento que hayan presentado.

Objetivo	N	%
Bueno	503,269	80.17
Malo	124,510	19.83

Tabla 2-7: Resumen descriptivo de *Objetivo*.

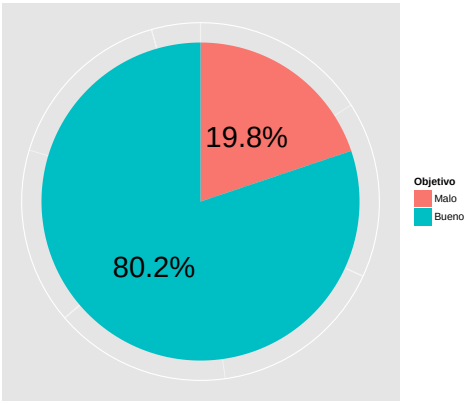

Figura 2-4: Distribución.

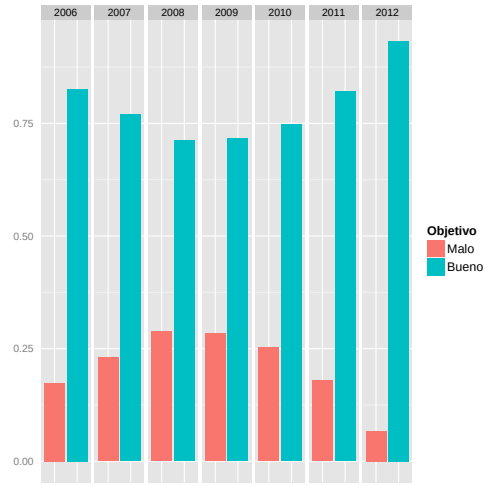

Figura 2-5: Estabilidad.

En la Figura 2-4, se puede observar que de todas las solicitudes analizadas, las que están catalogadas como *malas* representan una proporción de 19.83 % y además, se puede ver en la Figura 2-5 que la proporción de solicitudes *malas* va en decremento a partir del año 2009.

2.2.4. Resumen del análisis de la variable transversal

T1: tipo de cliente que realiza la solicitud.

T1	N	%
Nuevos	347,105	55.29
No nuevos	280,674	44.70

Tabla 2-8: Resumen descriptivo de $T1$.

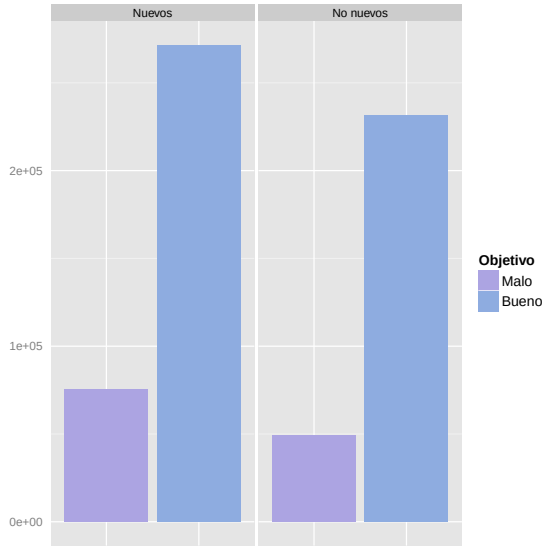

Figura 2-6: Distribución.

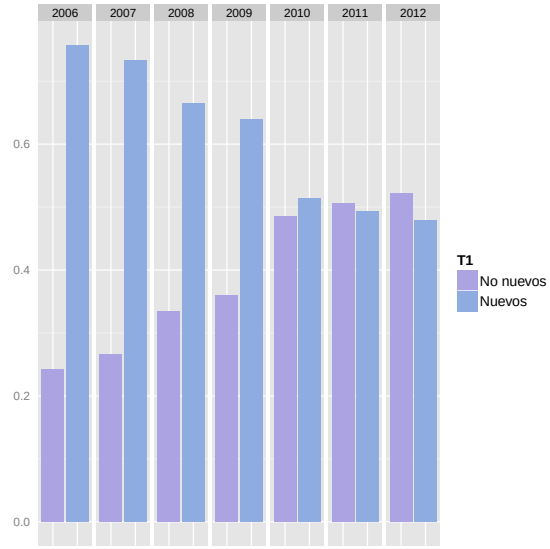

Figura 2-7: Estabilidad.

En la Tabla 2-8 se puede ver que los clientes **Nuevos** representan el 55.29 % de las solicitudes, mientras que los **No nuevos** el 44.70 %. También se puede ver en la Figura 2-7, que este último tipo de clientes va en aumento conforme pasa el tiempo.

Debido a que los clientes **No nuevos** ya tienen historia con la empresa y presentan más atributos (variables $C7$, $N13$, $N14$, $N15$ y $N16$); para el análisis de las variables de estudio, se decide segmentar las solicitudes por la variable **T1** y construir dos tipos de ABT.

2.2.5. Resumen del análisis de las variables categóricas

En esta sección se presenta un resumen gráfico y tabular, en donde se muestra la distribución de valores agrupados por las variables **Objetivo** y **T1**. Además, se presenta la conclusión a la que se llega para formar la *ABT*.

La nomenclatura que se presenta en el resumen, es la siguiente:

- B: Representa las solicitudes de clientes clasificados como *Bueno*.
- M: Representa las solicitudes de clientes clasificados como *Malo*.
- T: Representa el total de las solicitudes.

C1: género del solicitante de crédito.

T1	Valores	N_B	$\%_B$	N_M	$\%_M$	N_T	$\%_T$
Nuevos	H	94,213	34.7	49,431	65.5	143,644	41.4
	M	177,431	65.3	26,030	34.5	203,461	58.6
	Total	271,644	100.0	75,461	100.0	347,105	100.0
No nuevos	H	80,266	34.6	31,905	65.0	112,171	40.0
	M	151,359	65.3	17,144	35.0	168,503	60.0
	Total	231,625	100.0	49,049	100.0	280,674	100.0

Tabla 2-9: Resumen descriptivo de $C1$.

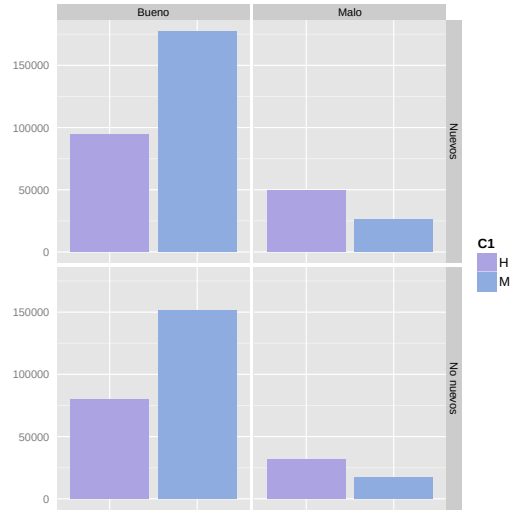

Figura 2-8: ● Distribución correcta.

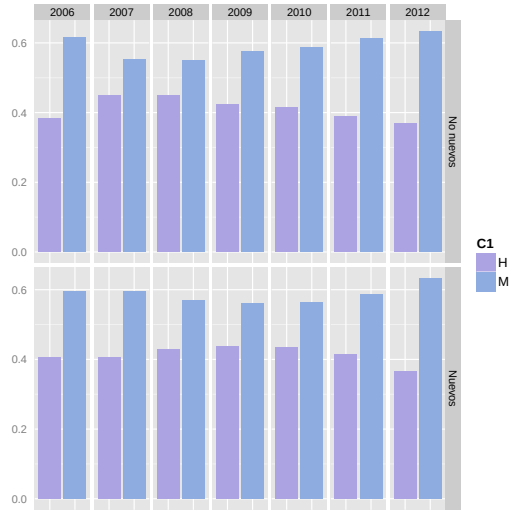

Figura 2-9: ● Estabilidad correcta.

✓ Esta variable se acepta para integrar la *ABT*.

C2: frecuencia de pago del crédito solicitado.

T1	Valores	N_B	$\%_B$	N_M	$\%_M$	N_T	$\%_T$
Nuevos	M	16,545	6.1	3,173	4.2	19,718	5.7
	Q	90,739	33.4	22,258	29.5	112,997	32.5
	S	164,360	60.5	50,030	66.3	214,390	61.8
	Total	271,644	100.0	75,461	100.0	347,105	100.0
No nuevos	M	16,776	7.2	2,496	5.1	19,272	6.9
	Q	71,075	30.7	11,848	24.2	82,923	29.5
	S	143,774	62.1	34,705	70.8	178,479	63.6
	Total	231,625	100.0	49,049	100.0	280,674	100.0

Tabla 2-10: Resumen descriptivo de $C2$.

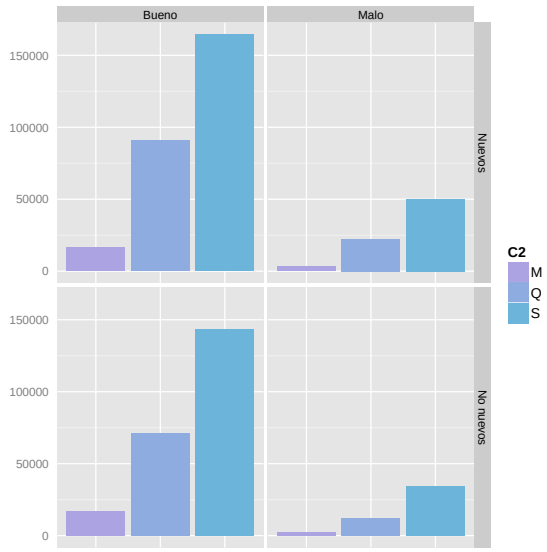

Figura 2-10: ● Distribución correcta.

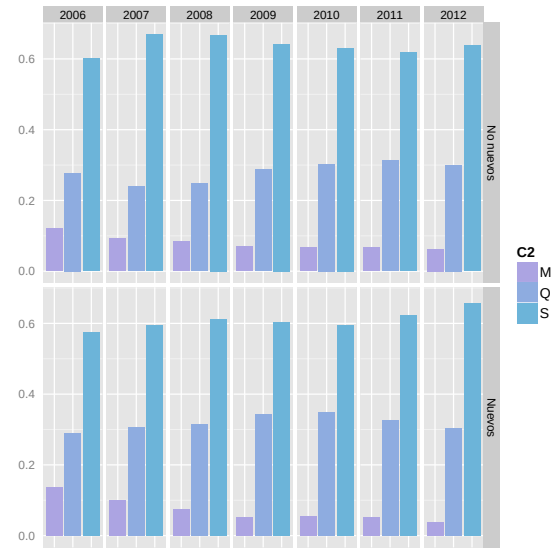

Figura 2-11: ● Estabilidad correcta.

✓ Esta variable se acepta para integrar la ABT.

C3: variable que describe el tipo de producto asignado a la solicitud.

T1	Valores	N_B	$\%_B$	N_M	$\%_M$	N_T	$\%_T$
Nuevos	C	153,742	56.6	39,107	51.8	192,849	55.6
	M	117,902	43.4	36,354	48.2	154,256	44.4
	Total	271,644	100.0	75,461	100.0	347,105	100.0
No nuevos	C	124,686	53.8	22,768	46.4	147,454	52.5
	M	106,939	46.2	26,281	53.6	133,220	47.5
	Total	231,625	100.0	49,049	100.0	280,674	100.0

Tabla 2-11: Resumen descriptivo de $C3$.

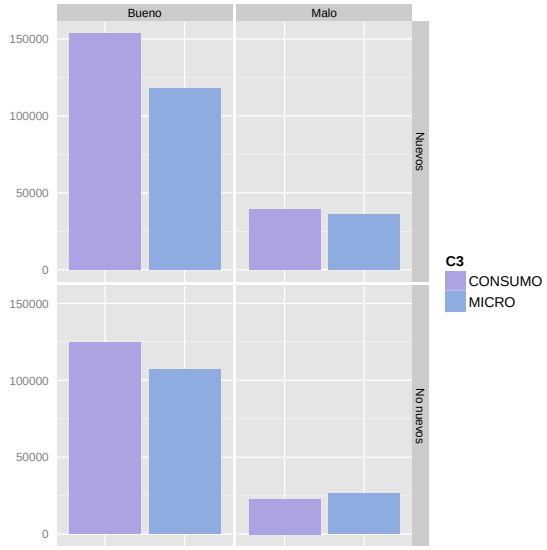

Figura 2-12: ● Distribución correcta.

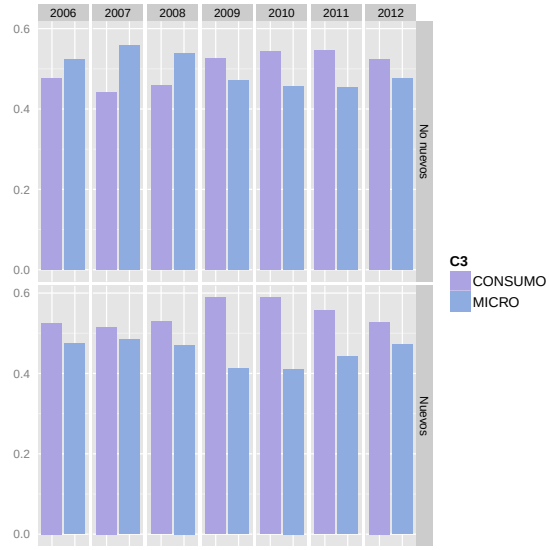

Figura 2-13: ● Estabilidad correcta.

✓ Esta variable se acepta para integrar la ABT.

C4: número de cuentas en las que el cliente tiene atrasos reportados en BC.

T1	Valores	N_B	$\%_B$	N_M	$\%_M$	N_T	$\%_T$
Nuevos	0	127,812	47.0	0	0.0	127,812	36.8
	1	69,370	25.5	0	0.0	69,370	20.0
	2	35,409	13.0	41,432	54.9	76,841	22.1
	3	18,155	6.7	17,730	23.5	35,885	10.3
	4	9,662	3.6	8,172	10.8	17,834	5.1
	5	5,123	1.9	3,902	5.2	9,025	2.6
	6 o más	6,113	2.2	4,225	5.6	10,338	3.0
	Total	271,644	100.0	75,461	100.0	347,105	100.0
No nuevos	0	96,619	41.7	0	0.0	96,619	34.4
	1	60,325	26.0	0	0.0	60,325	21.5
	2	33,590	14.5	19,851	40.5	53,441	19.0
	3	18,568	8.0	12,877	26.2	31,445	11.2
	4	10,083	4.3	6,946	14.2	17,029	6.1
	5	5,713	2.5	3,883	7.9	9,596	3.4
	6 o más	6,727	2.9	5,492	11.2	12,219	4.3
	Total	231,625	100.0	49,049	100.0	280,674	100.0

Tabla 2-12: Resumen descriptivo de $C4$.

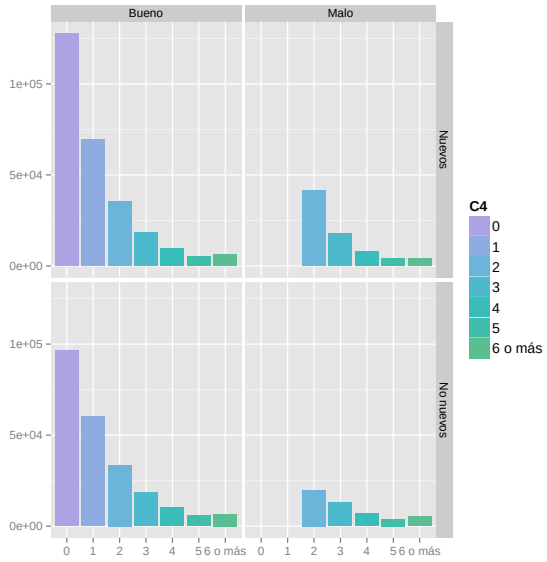

Figura 2-14: ● Distribución correcta.

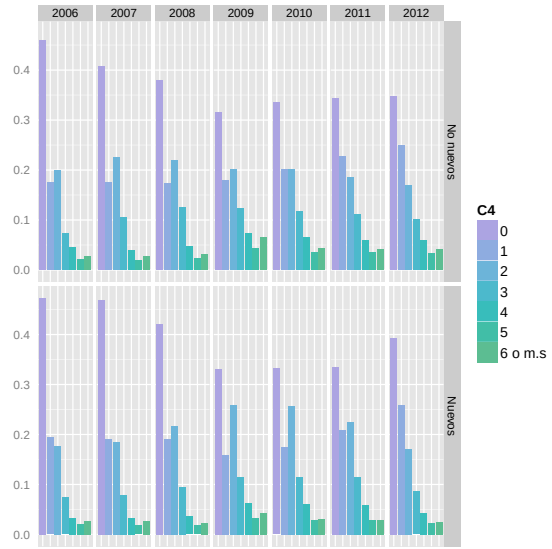

Figura 2-15: ● Estabilidad correcta.

✓ Esta variable se acepta para integrar la *ABT*.

C5: variable que indica si la solicitud fue revisada y decidida por AEC.

T1	Valores	N_B	$\%_B$	N_M	$\%_M$	N_T	$\%_T$
Nuevos	0	247,201	91.0	68,459	90.7	315,660	90.9
	1	24443	9.0	7002	9.3	31445	9.1
	Total	271,644	100.0	75,461	100.0	347,105	100.0
No nuevos	0	198,284	85.6	29,967	61.1	228,251	81.3
	1	33,341	14.4	19,082	38.9	52,423	18.7
	Total	231,625	100.0	49,049	100.0	280,674	100.0

Tabla 2-13: Resumen descriptivo de $C5$.

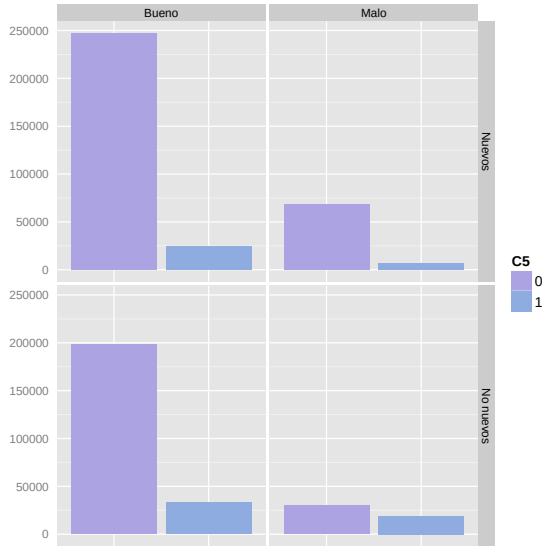

Figura 2-16: ● Distribución concentrada.

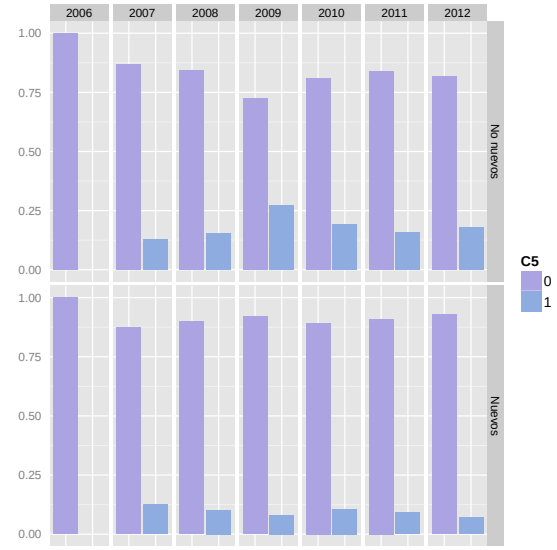

Figura 2-17: ● Estabilidad correcta.

✗ Esta variable se rechaza para integrar la ABT.

C6: variable que indica si la solicitud fue gestionada por *Telemarketing* desde el inicio.

T1	Valores	N_B	$\%_B$	N_M	$\%_M$	N_T	$\%_T$
Nuevos	0	259,993	95.7	73,226	97.0	333,219	96.0
	1	11,651	4.3	2,235	3.0	13,886	4.0
	Total	271,644	100.0	75,461	100.0	347,105	100.0
No nuevos	0	231,101	99.8	49,009	99.9	280,110	99.8
	1	524	0.2	40	0.1	564	0.2
	Total	231,625	100.0	49,049	100.0	280,674	100.0

Tabla 2-14: Resumen descriptivo de *C*6.

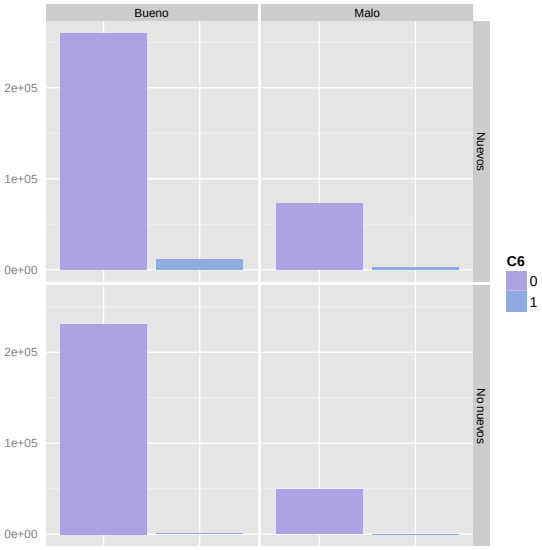

Figura 2-18: ● Distribución concentrada.

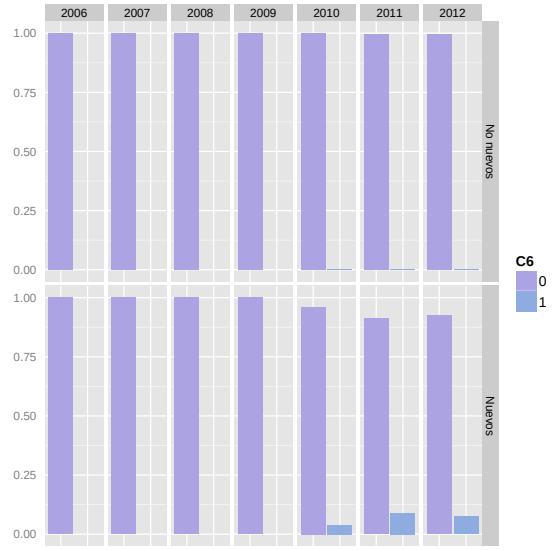

Figura 2-19: ● Estabilidad aceptable.

✗ Esta variable se rechaza para integrar la *ABT*.

C7: número de solicitudes que ha realizado el cliente.

T1	Valores	N_B	$\%_B$	N_M	$\%_M$	N_T	$\%_T$
No nuevos	1	120,934	52.2	31,129	63.5	152,063	54.2
	2	57,216	24.7	11,193	22.8	68,409	24.4
	3	28,266	12.2	4,196	8.6	32,462	11.6
	4	13,481	5.8	1,665	3.4	15,146	5.4
	5	6,313	2.7	552	1.1	6,865	2.5
	6 o más	5,415	2.3	314	0.6	5,729	2.0
	Total	231,625	100.0	49,049	100.0	280,674	100.0

Tabla 2-15: Resumen descriptivo de $C7$.

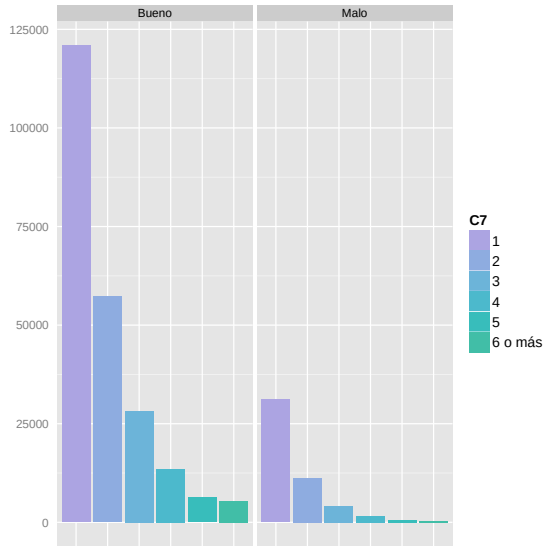

Figura 2-20: ● Distribución correcta.

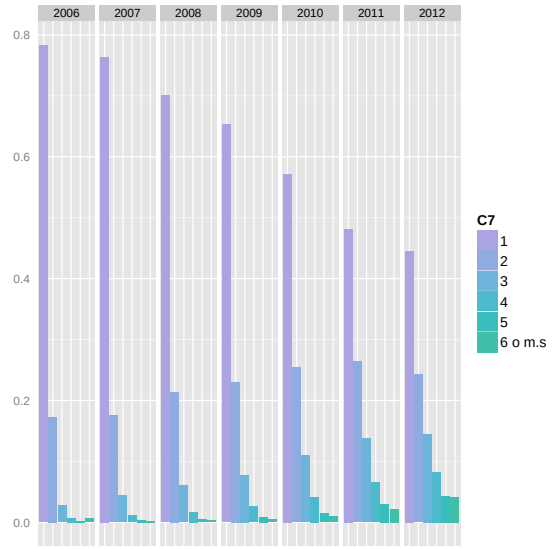

Figura 2-21: ● Estabilidad aceptable.

✓ Esta variable se acepta para integrar la ABT.

2.2.6. Resumen del análisis de las variables numéricas

En el caso de las variables numéricas, antes de ser analizadas se les aplican las *transformaciones de Box-Cox*[3], con la finalidad de mitigar la variabilidad y de esta forma corregir las distribuciones.

En esta sección se presenta el resumen del análisis con un gráfico que muestra la distribución de valores y una tabla que presenta sus principales estadísticos, segmentados por las variables **Objetivo** y **T1**. Además, se presenta la conclusión a la que se llega para formar la *ABT*.

La nomenclatura que se presenta en el resumen del análisis, es la siguiente:

- B: representa las solicitudes de clientes clasificados como *Bueno*.
- M: representa las solicitudes de clientes clasificados como *Malo*.
- T: representa el total de las solicitudes.
- Min: valor mínimo.
- Max: valor máximo.
- x: media.
- s: desviación estándar.
- q_1: primer cuartil.
- q_2: segundo cuartil.
- q_3: tercer cuartil.
- Iqr: rango intercuartil.
- Out: número de observaciones atípicos.
- Ext: número de observaciones extremos.
- Na: número de observaciones faltantes.

[3]Son una familia de transformaciones, utilizadas para arreglar problemas de *normalidad* y *heterocedasticidad* en las variables. Ver Apéndice B.

N1: monto del préstamo total a financiar.

T1	Estadísticos	Buenos	Malos	Total
Nuevos	N	271,644	75,461	347,105
	Min	0	16	0
	Max	75.29	64.65	75.29
	$\bar{x}$	32.49	32.73	32.55
	s	4.21	3.9	4.15
	q_1	30.23	30.23	30.23
	q_2	32.67	33.34	32.89
	q_3	35.04	35.04	35.04
	Iqr	4.81	4.81	4.81
	Out	3,242	831	4,073
	Ext	573	135	708
	Na	0	0	0
No nuevos	N	231,625	49,049	280,674
	Min	0	7	0
	Max	63.76	64.5	64.5
	$\bar{x}$	33.89	32.89	33.72
	s	5.22	5.19	5.23
	q_1	30.78	29.94	30.67
	q_2	34.17	33.29	34
	q_3	36.97	35.88	36.83
	Iqr	6.2	5.94	6.17
	Out	3,370	624	4,002
	Ext	356	116	457
	Na	0	0	0

Tabla 2-16: Resumen descriptivo de $N1$.

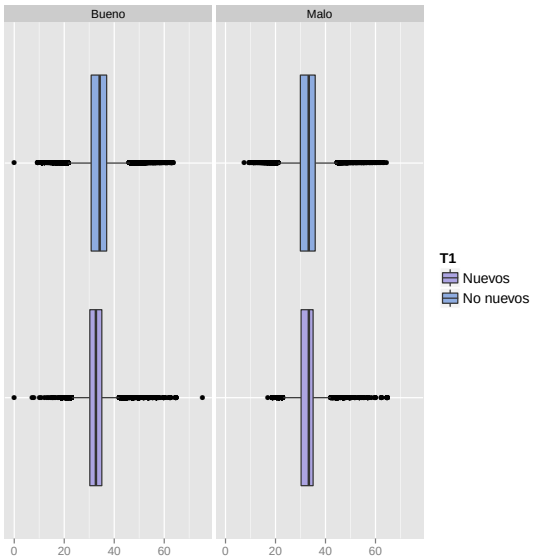

Figura 2-22: ● Distribución correcta.

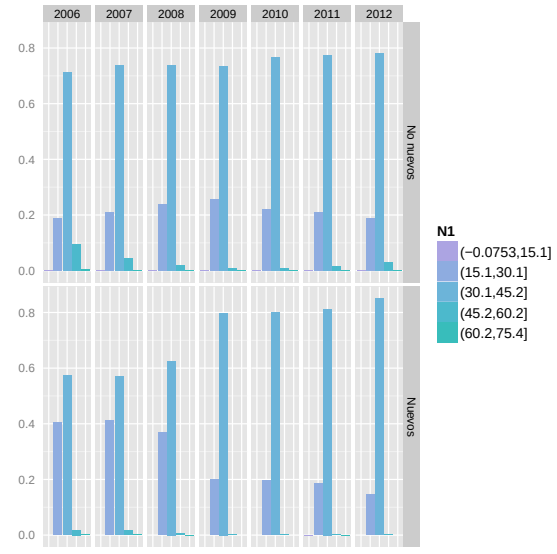

Figura 2-23: ● Estabilidad aceptable.

✓ Esta variable se acepta para integrar la *ABT*.

N2: ingresos principales del cliente.

T1	Estadísticos	Buenos	Malos	Total
Nuevos	N	271,644	75,461	347,105
	Min	0	0	0
	Max	190.29	126.18	190.29
	$\bar{x}$	32.06	31.6	31.96
	s	4.48	4.33	4.45
	q_1	29.23	28.76	29.12
	q_2	31.64	31.21	31.48
	q_3	34.74	33.94	34.59
	Iqr	5.51	5.18	5.47
	Out	3,106	911	4,024
	Ext	753	271	1,008
	Na	672	157	829
No nuevos	N	231,625	49,049	280,674
	Min	0	0	0
	Max	190.29	125.96	190.29
	$\bar{x}$	32.64	31.08	32.37
	s	4.46	4.69	4.54
	q_1	29.74	28.25	29.64
	q_2	32.46	31.21	32.05
	q_3	35.55	33.83	35.18
	Iqr	5.8	5.58	5.54
	Out	2,350	563	3,241
	Ext	483	280	877
	Na	327	98	425

Tabla 2-17: Resumen descriptivo de $N2$.

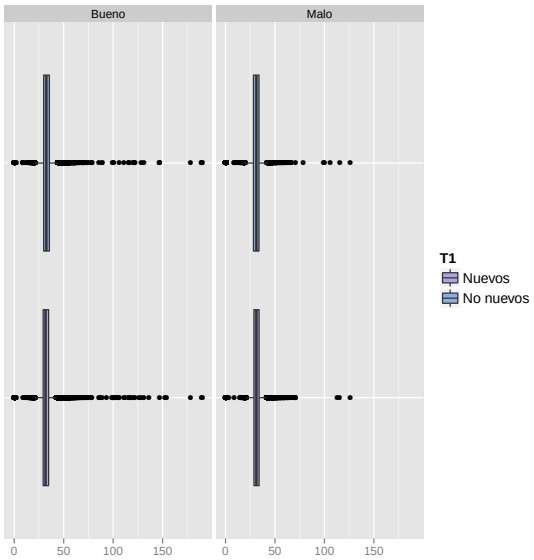

Figura 2-24: ● Distribución aceptable.

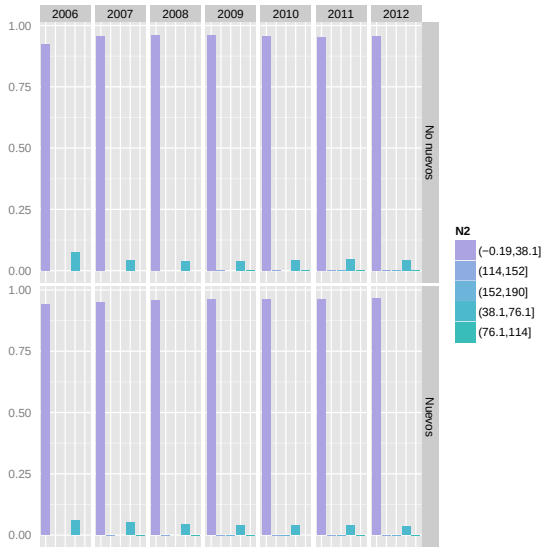

Figura 2-25: ● Estabilidad aceptable.

✗ Esta variable se rechaza para integrar la ABT.

N3: plazo para pagar el total financiado.

T1	Estadísticos	Buenos	Malos	Total
Nuevos	N	271,644	75,461	347,105
	Min	0	2	0
	Max	22.9	20.72	22.9
	$\bar{x}$	12.31	12.89	12.44
	s	3.49	3.29	3.45
	q_1	9.83	10.33	9.83
	q_2	12.28	12.28	12.28
	q_3	15.55	15.55	15.55
	Iqr	5.72	5.22	5.72
	Out	0	0	0
	Ext	0	0	0
	Na	0	0	0
No nuevos	N	231,625	49,049	280,674
	Min	0	0	0
	Max	56.95	56.95	56.95
	$\bar{x}$	28.09	28.01	28.08
	s	10.57	9.98	10.47
	q_1	22.22	22.22	22.22
	q_2	28.48	28.48	28.48
	q_3	37.3	37.3	37.3
	Iqr	15.08	15.08	15.08
	Out	0	0	0
	Ext	0	0	0
	Na	0	0	0

Tabla 2-18: Resumen descriptivo de $N3$.

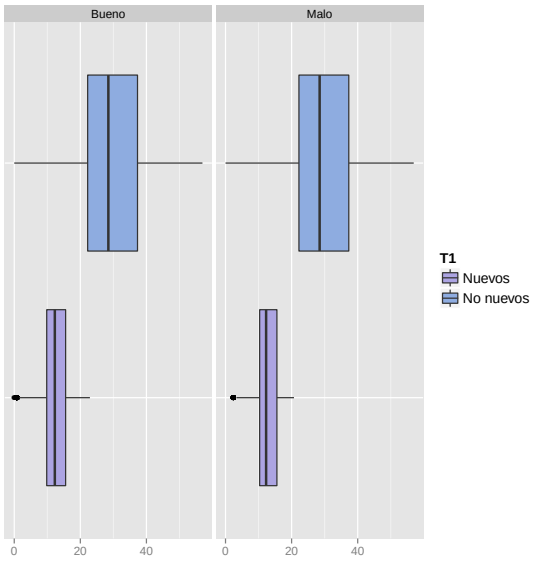

Figura 2-26: ● Distribución correcta.

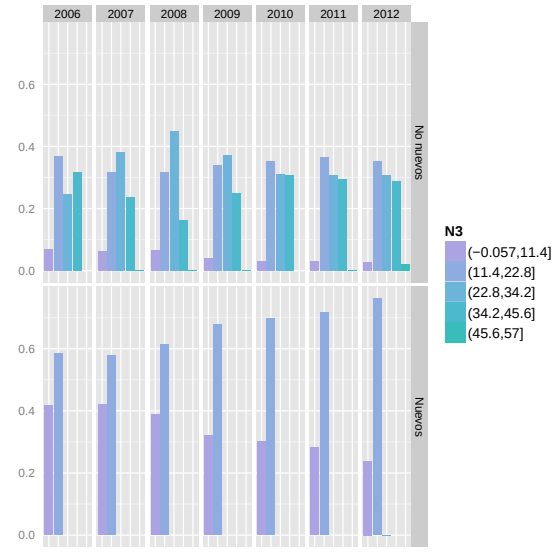

Figura 2-27: ● Estabilidad correcta.

✓ Esta variable se acepta para integrar la ABT.

N4: años de vida que tiene el cliente.

T1	Estadísticos	Buenos	Malos	Total
Nuevos	N	271,644	75,461	347,105
	Min	2	0	0
	Max	26.69	22.98	26.69
	$\bar{x}$	14	8.78	12.87
	s	5.24	5.27	5.67
	q_1	9.83	4.4	8.32
	q_2	14.02	8.32	12.67
	q_3	17.85	12.67	17.02
	Iqr	8.02	8.27	8.71
	Out	0	0	0
	Ext	0	0	0
	Na	19	4	23
No nuevos	N	231,625	49,049	280,674
	Min	2	0	0
	Max	26.69	24.11	26.69
	$\bar{x}$	15.08	9.98	14.19
	s	5.06	5.26	5.45
	q_1	11.28	6.16	10.32
	q_2	15.33	9.83	14.46
	q_3	19.07	14.02	18.26
	Iqr	7.8	7.85	7.94
	Out	0	0	0
	Ext	0	0	0
	Na	13	0	13

Tabla 2-19: Resumen descriptivo de $N4$.

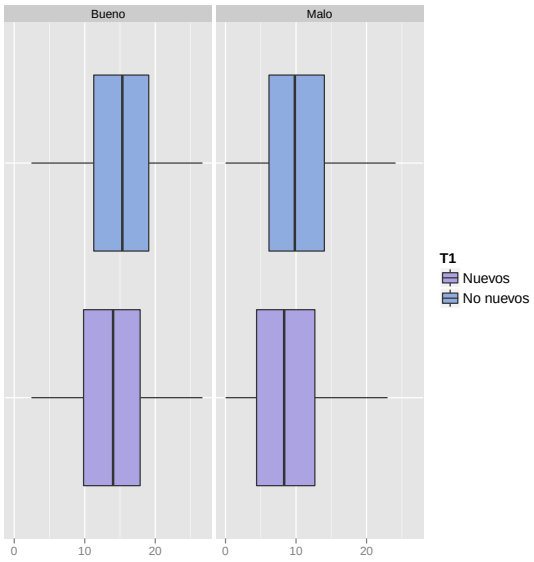

Figura 2-28: ● Distribución correcta.

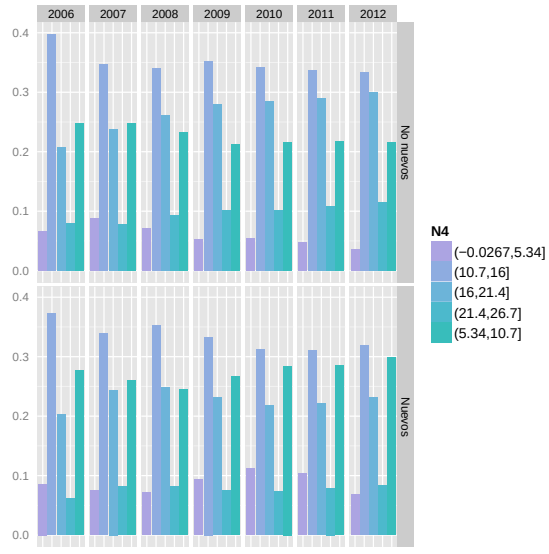

Figura 2-29: ● Estabilidad correcta.

✓ Esta variable se acepta para integrar la ABT.

N5: los meses de antigüedad en el trabajo que tiene el cliente.

T1	Estadísticos	Buenos	Malos	Total
Nuevos	N	271,644	75,461	347,105
	Min	8	0	0
	Max	51.4	47.32	51.4
	$\bar{x}$	19.41	13.13	18.04
	s	7.04	8.73	7.88
	q_1	13.75	6	12.97
	q_2	17.7	12.28	16.65
	q_3	23.06	19.35	22.33
	Iqr	9.31	13.35	9.36
	Out	6,464	114	7,755
	Ext	3	0	3
	Na	1,300	406	1,706
No nuevos	N	231,625	49,049	280,674
	Min	8	0	0
	Max	51.4	45.92	51.4
	$\bar{x}$	20.89	14.32	19.74
	s	7.16	8.95	7.91
	q_1	15.32	7.59	14.37
	q_2	19.63	13.62	18.88
	q_3	25.13	20.27	24.46
	Iqr	9.81	12.68	10.09
	Out	2,565	216	2,961
	Ext	0	0	0
	Na	692	338	1,030

Tabla 2-20: Resumen descriptivo de $N5$.

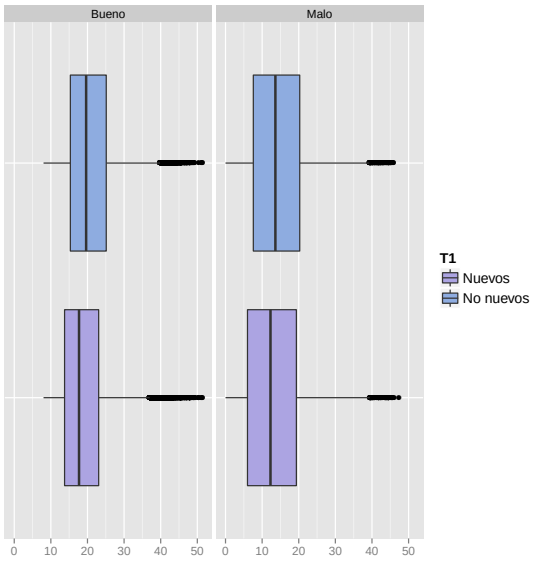

Figura 2-30: ● Distribución correcta.

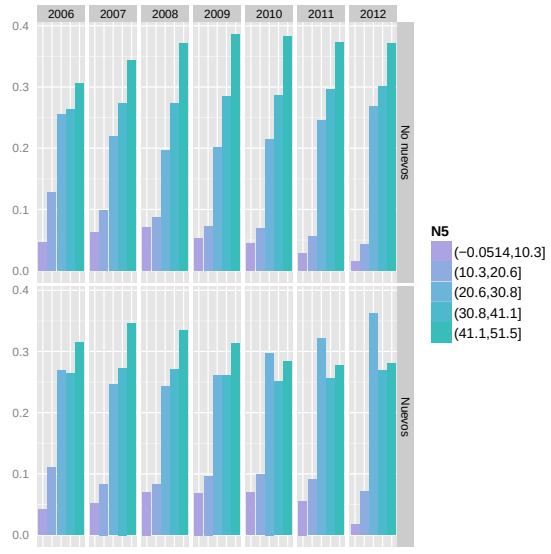

Figura 2-31: ● Estabilidad correcta.

✓ Esta variable se acepta para integrar la ABT.

N6: los meses de antigüedad que tiene el cliente en su actual domicilio.

T1	Estadísticos	Buenos	Malos	Total
Nuevos	N	271,644	75,461	347,105
	Min	10	0	0
	Max	51.55	48.12	51.55
	$\bar{x}$	22.53	14.09	20.69
	s	5.58	7.87	7.07
	q_1	18.49	8.2	17.18
	q_2	21.07	12.7	20.18
	q_3	25.35	18.78	24.53
	Iqr	6.86	10.59	7.35
	Out	8,650	1,163	9,629
	Ext	133	0	137
	Na	0	0	0
No nuevos	N	231,625	49,049	280,674
	Min	8	0	0
	Max	51.81	48.44	51.81
	$\bar{x}$	23.75	15.97	22.39
	s	6.19	8.35	7.25
	q_1	19.26	9.83	18.2
	q_2	22.33	14.61	21.58
	q_3	26.98	21.07	26.35
	Iqr	7.72	11.23	8.16
	Out	6,296	499	6,780
	Ext	113	0	42
	Na	0	0	0

Tabla 2-21: Resumen descriptivo de $N6$.

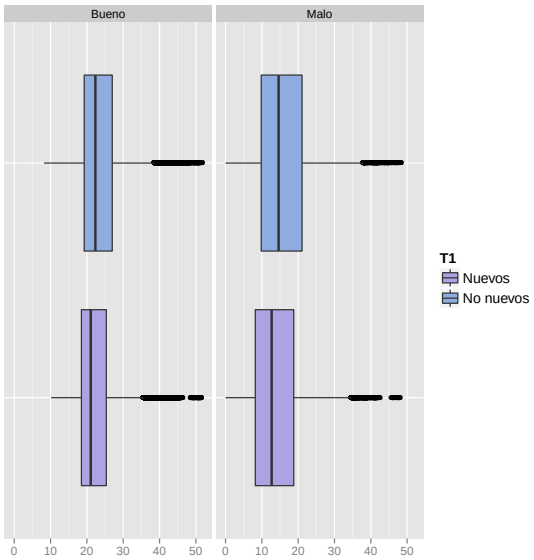

Figura 2-32: ● Distribución correcta.

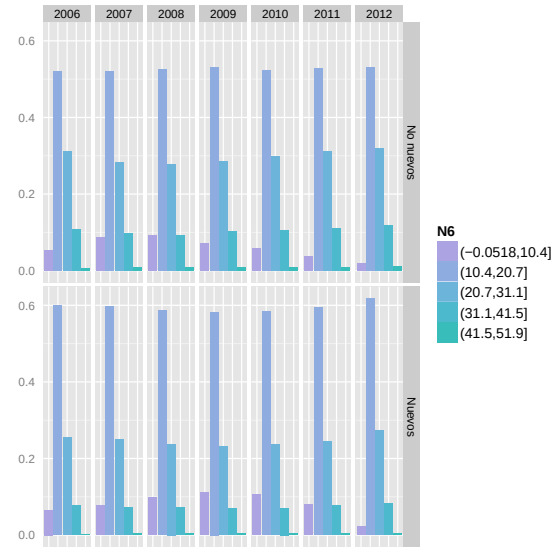

Figura 2-33: ● Estabilidad correcta.

✓ Esta variable se acepta para integrar la *ABT*.

N7: remanente de ingresos del cliente.

T1	Estadísticos	Buenos	Malos	Total
Nuevos	N	271,644	75,461	347,105
	Min	0	0	0
	Max	12.22	11.32	12.22
	$\bar{x}$	7.33	7.29	7.32
	s	0.71	0.71	0.71
	q_1	6.83	6.79	6.82
	q_2	7.3	7.25	7.29
	q_3	7.79	7.73	7.78
	Iqr	0.96	0.94	0.96
	Out	1,885	672	2,526
	Ext	114	45	159
	Na	23	3	26
No nuevos	N	231,625	49,049	280,674
	Min	0	0	0
	Max	83.74	71.17	83.74
	$\bar{x}$	21.7	18.31	21.11
	s	4.57	7.7	5.41
	q_1	18.41	16.26	18.07
	q_2	21.11	19.41	20.84
	q_3	24.3	22.7	24.04
	Iqr	5.89	6.44	5.97
	Out	4420	609	10283
	Ext	305	27	5515
	Na	18	2	20

Tabla 2-22: Resumen descriptivo de $N7$.

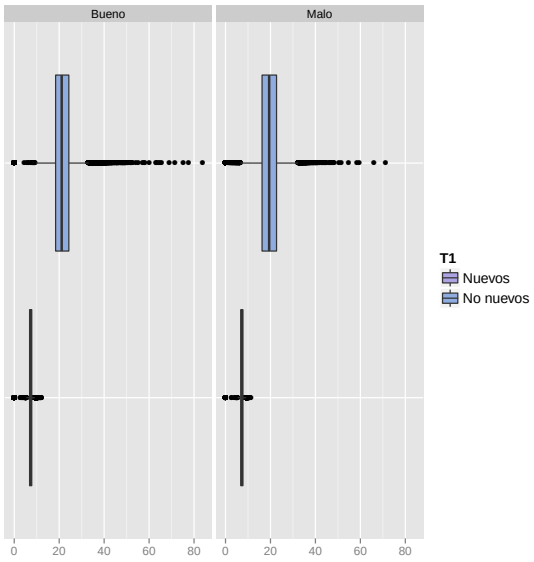

Figura 2-34: ● Distribución concentrada.

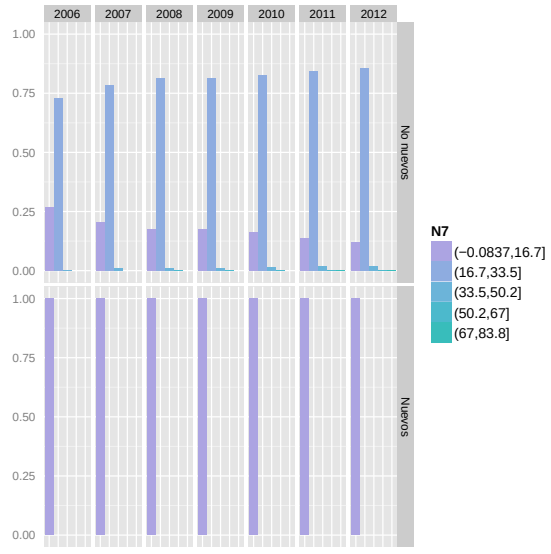

Figura 2-35: ● Estabilidad correcta.

✗ Esta variable se rechaza para integrar la ABT.

N8: saldo anterior al momento de solicitar renovación.

T1	Estadísticos	Buenos	Malos	Total
No nuevos	N	231,625	49,049	280,674
	Min	0	0	0
	Max	58.73	57.15	58.73
	$\bar{x}$	20.17	21.15	20.34
	s	14.49	14.73	14.54
	q_1	0	0	0
	q_2	26.46	28.09	26.78
	q_3	31.82	32.44	31.95
	Iqr	31.82	32.44	31.95
	Out	0	0	0
	Ext	0	0	0
	Na	0	0	0

Tabla 2-23: Resumen descriptivo de $N8$.

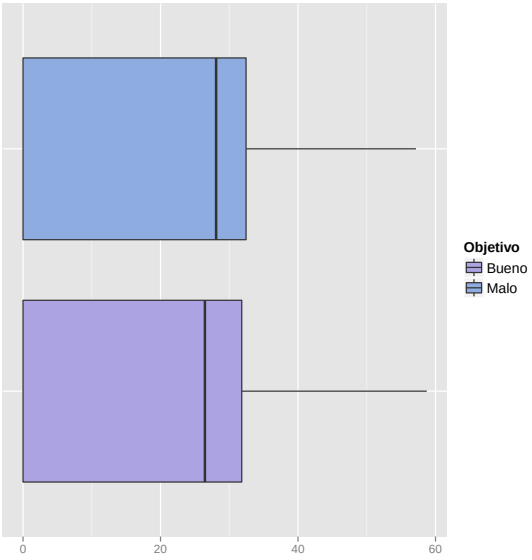

Figura 2-36: 🟡 Distribución aceptable.

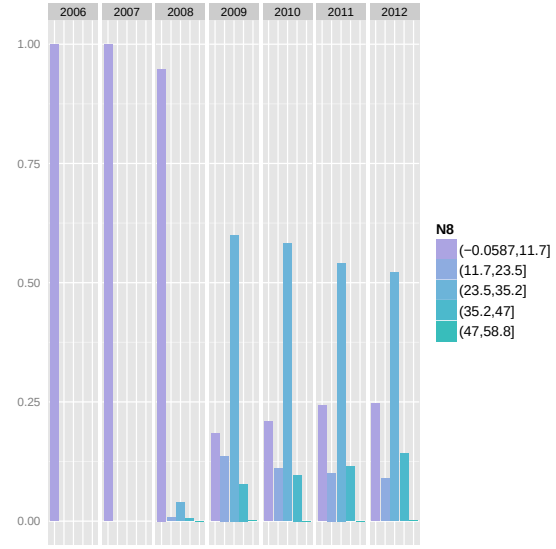

Figura 2-37: 🟡 Estabilidad aceptable.

✗ Esta variable se rechaza para integrar la *ABT*.

N9: monto del préstamo anterior.

T1	Estadísticos	Buenos	Malos	Total
No nuevos	N	231,625	49,049	280,674
	Min	0	0	0
	Max	64.77	64.77	64.77
	$\bar{x}$	23.65	23.98	23.71
	s	16.31	16.3	16.31
	q_1	0	0	0
	q_2	31.65	32.23	31.65
	q_3	35.65	35.65	35.65
	Iqr	35.65	35.65	35.65
	Out	0	0	0
	Ext	0	0	0
	Na	0	0	0

Tabla 2-24: Resumen descriptivo de $N9$.

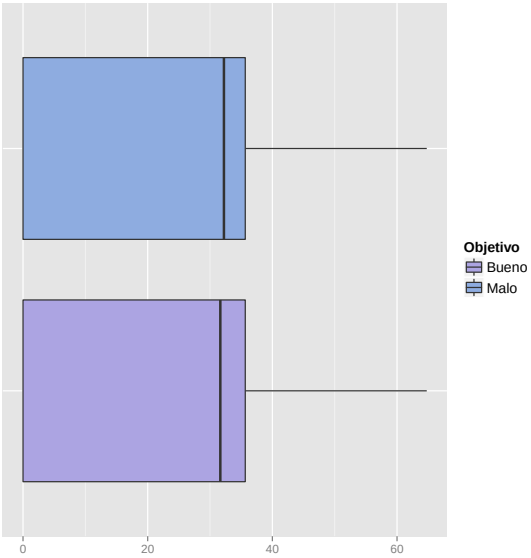

Figura 2-38: 🟡 Distribución aceptable.

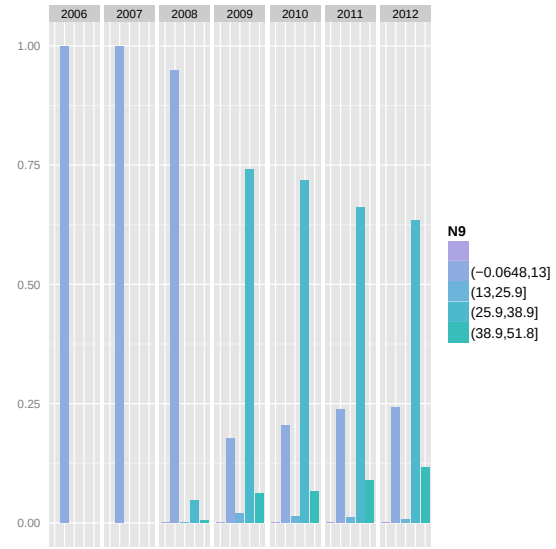

Figura 2-39: 🟡 Estabilidad aceptable.

✗ Esta variable se rechaza para integrar la ABT.

N10: monto de los pagos que el cliente tendría que hacer en caso de aprobarse su crédito.

T1	Estadísticos	Buenos	Malos	Total
Nuevos	N	271,644	75,461	347,105
	Min	0	1	0
	Max	3.53	3.51	3.53
	$\bar{x}$	2.81	2.79	2.8
	s	0.19	0.18	0.18
	q_1	2.7	2.68	2.7
	q_2	2.78	2.75	2.78
	q_3	2.95	2.92	2.94
	Iqr	0.25	0.23	0.24
	Out	993	682	1577
	Ext	67	18	110
	Na	0	0	0
No nuevos	N	231,625	49,049	280,674
	Min	0	1	0
	Max	3.55	3.49	3.55
	$\bar{x}$	2.83	2.8	2.83
	s	0.18	0.18	0.18
	q_1	2.72	2.69	2.71
	q_2	2.82	2.78	2.81
	q_3	2.95	2.91	2.94
	Iqr	0.23	0.23	0.23
	Out	1,336	618	1,780
	Ext	92	6	70
	Na	0	0	0

Tabla 2-25: Resumen descriptivo de $N10$.

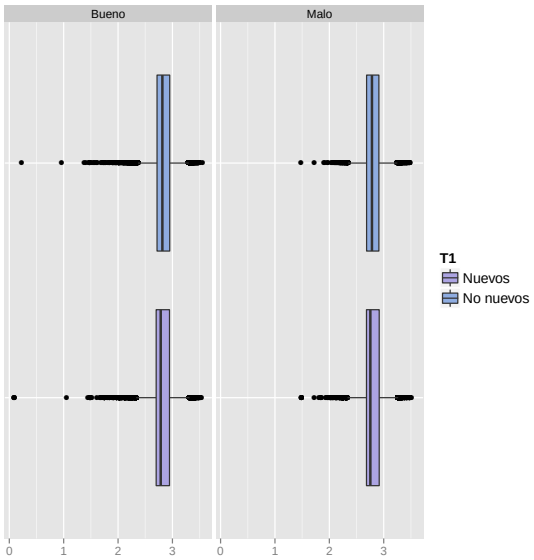

Figura 2-40: ● Distribución correcta.

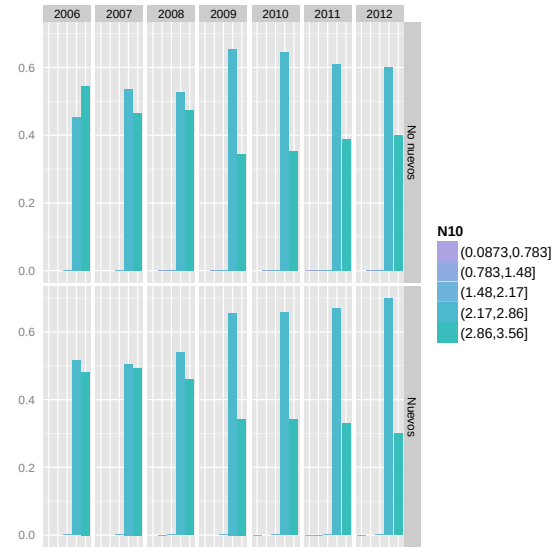

Figura 2-41: ● Estabilidad correcta.

✓ Esta variable se acepta para integrar la *ABT*.

N11: proporción del ingreso que el cliente tendría que destinar al pago de su crédito.

T1	Estadísticos	Buenos	Malos	Total
Nuevos	N	271,644	75,461	347,105
	Min	0	0	0
	Max	0.33	0.33	0.33
	$\bar{x}$	0.03	0.03	0.03
	s	0.02	0.02	0.02
	q_1	0.01	0.01	0.01
	q_2	0.02	0.02	0.02
	q_3	0.03	0.03	0.03
	Iqr	0.02	0.02	0.02
	Out	10,241	3,046	13,213
	Ext	4,567	1,345	5,937
	Na	672	157	829
No nuevos	N	231,625	49,049	280,674
	Min	0	0	0
	Max	0.33	0.33	0.33
	$\bar{x}$	0.03	0.03	0.03
	s	0.02	0.03	0.02
	q_1	0.01	0.02	0.01
	q_2	0.02	0.02	0.02
	q_3	0.03	0.03	0.03
	Iqr	0.02	0.02	0.02
	Out	9,374	1,980	11,248
	Ext	4,947	1,316	6,243
	Na	327	98	425

Tabla 2-26: Resumen descriptivo de $N11$.

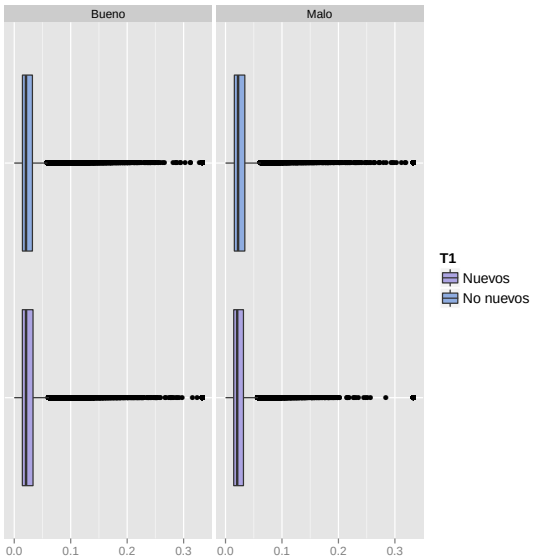

Figura 2-42: ● Distribución aceptable.

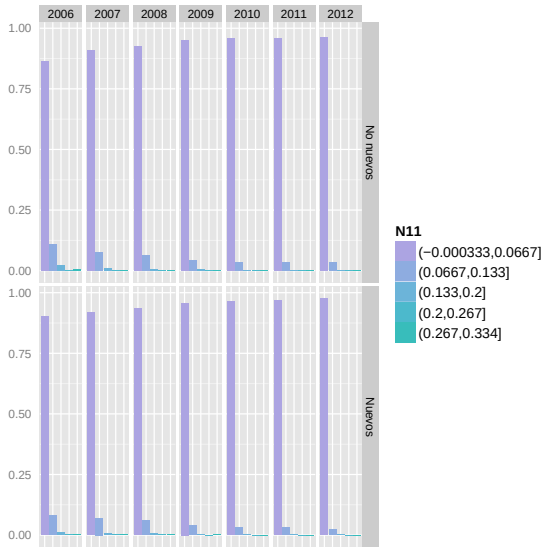

Figura 2-43: ● Estabilidad correcta.

✓ Esta variable se acepta para integrar la ABT.

N12: proporción del remanente de ingreso que tendrá el cliente en caso de aprobarse su crédito.

T1	Estadísticos	Buenos	Malos	Total
Nuevos	N	271,644	75,461	347,105
	Min	0	0	0
	Max	0.33	0.33	0.33
	$\bar{x}$	0.16	0.16	0.16
	s	0.06	0.05	0.06
	q_1	0.12	0.12	0.12
	q_2	0.15	0.15	0.15
	q_3	0.2	0.19	0.2
	Iqr	0.08	0.07	0.08
	Out	1,103	474	1,555
	Ext	0	0	0
	Na	23	3	26
No nuevos	N	231,625	49,049	280,674
	Min	0	0	0
	Max	0.33	0.33	0.33
	$\bar{x}$	0.16	0.14	0.15
	s	0.05	0.07	0.06
	q_1	0.12	0.1	0.12
	q_2	0.15	0.14	0.15
	q_3	0.19	0.18	0.19
	Iqr	0.08	0.08	0.08
	Out	1,493	525	2,053
	Ext	0	0	0
	Na	18	2	20

Tabla 2-27: Resumen descriptivo de $N12$.

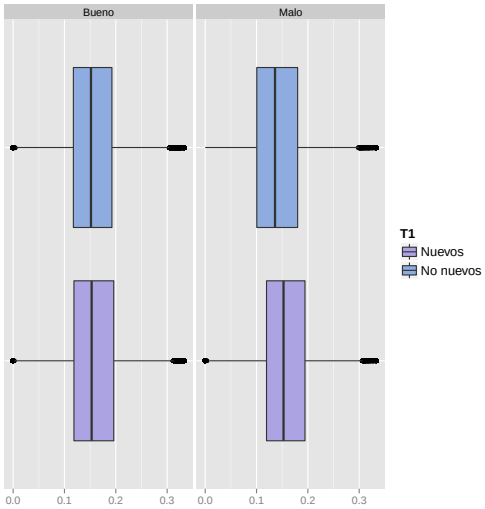

Figura 2-44: ● Distribución correcta.

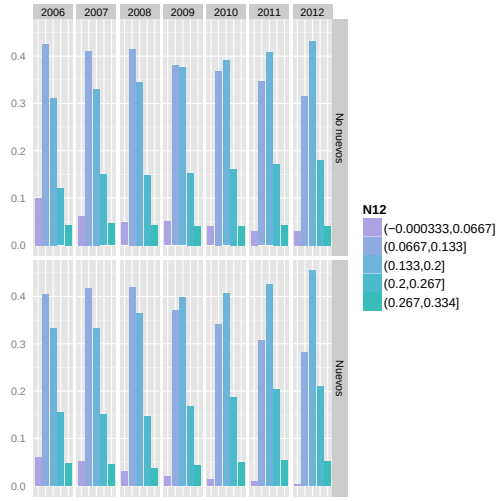

Figura 2-45: ● Estabilidad correcta.

✓ Esta variable se acepta para integrar la ABT.

N13: proporción de lo que le falta pagar al cliente al momento de pedir la renovación de su crédito

T1	Estadísticos	Buenos	Malos	Total
No nuevos	N	231,625	49,049	280,674
	Min	0	0	0
	Max	1.09	1.13	1.13
	$\bar{x}$	0.4	0.43	0.4
	s	0.32	0.33	0.33
	q_1	0	0	0
	q_2	0.47	0.54	0.48
	q_3	0.7	0.72	0.7
	Iqr	0.7	0.72	0.7
	Out	0	0	0
	Ext	0	0	0
	Na	0	0	0

Tabla 2-28: Resumen descriptivo de $N13$.

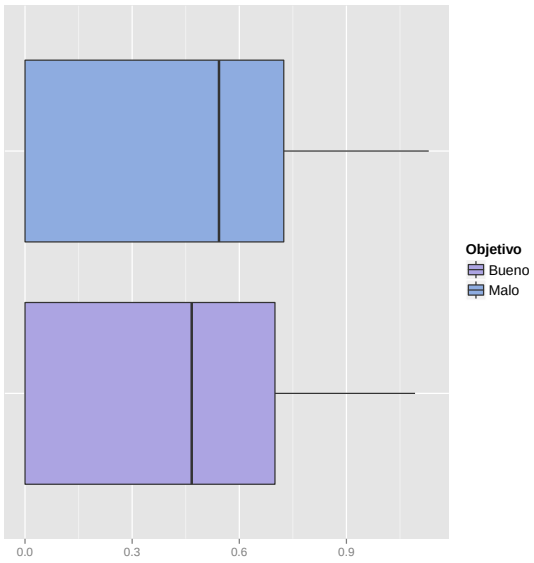

Figura 2-46: ● Distribución aceptable.

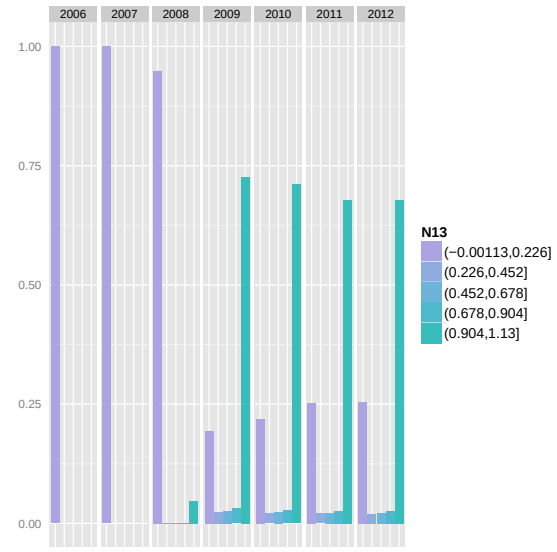

Figura 2-47: ● Estabilidad aceptable.

✗ Esta variable se rechaza para integrar la ABT.

N14: variación de ciertos atributos que tiene el cliente entre su solicitud anterior y la actual.

T1	Estadísticos	Buenos	Malos	Total
No nuevos	N	231,625	49,049	280,674
	Min	0	0	0
	Max	3.29	2.11	3.29
	$\bar{x}$	0.89	0.85	0.89
	s	0.12	0.15	0.13
	q_1	0.84	0.76	0.83
	q_2	0.9	0.88	0.89
	q_3	0.94	0.93	0.94
	Iqr	0.1	0.17	0.11
	Out	11,339	731	12,785
	Ext	4,239	90	4,720
	Na	707	338	1,045

Tabla 2-29: Resumen descriptivo de $N14$.

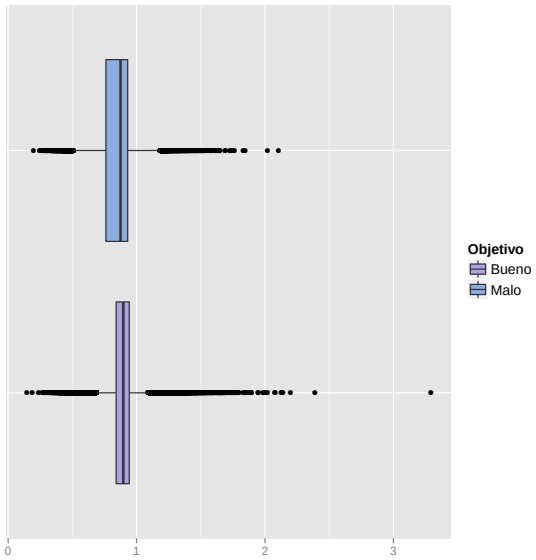

Figura 2-48: ● Distribución aceptable.

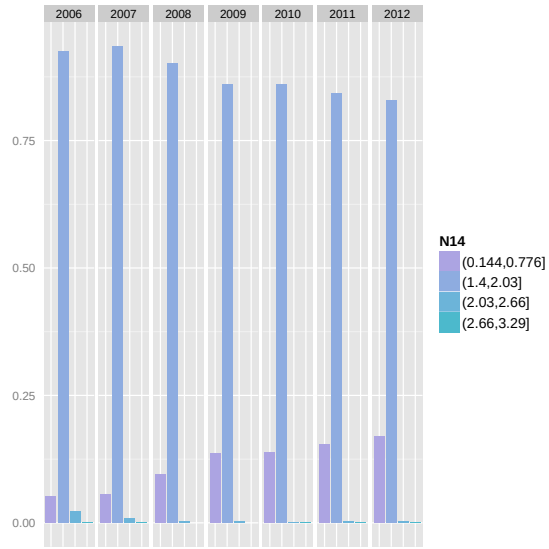

Figura 2-49: ● Estabilidad correcta.

✓ Esta variable se acepta para integrar la ABT.

N15: monto que se le agrega al saldo del cliente al momento de renovar su crédito.

T1	Estadísticos	Buenos	Malos	Total
No nuevos	N	231,625	49,049	280,674
	Min	23	0	0
	Max	64.3	65.02	65.02
	$\bar{x}$	32.58	29.77	32.09
	s	4.54	6.94	5.15
	q_1	29.16	25.07	28.9
	q_2	32.11	29.96	32.01
	q_3	35.41	34.83	35.27
	Iqr	6.25	9.76	6.37
	Out	3,345	169	3,898
	Ext	159	1	220
	Na	0	0	0

Tabla 2-30: Resumen descriptivo de $N15$.

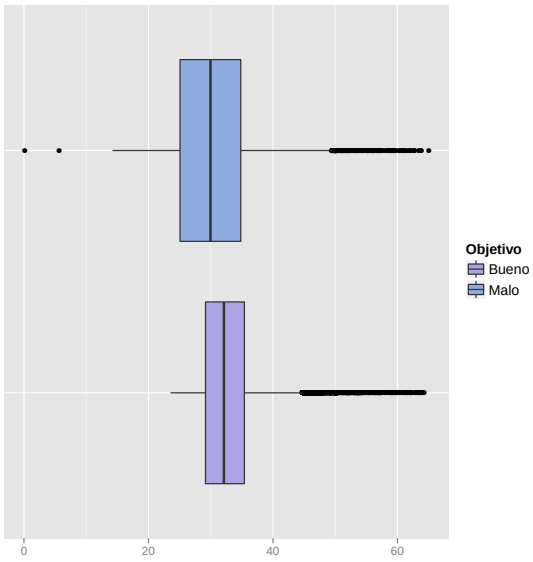

Figura 2-50: ● Distribución correcta.

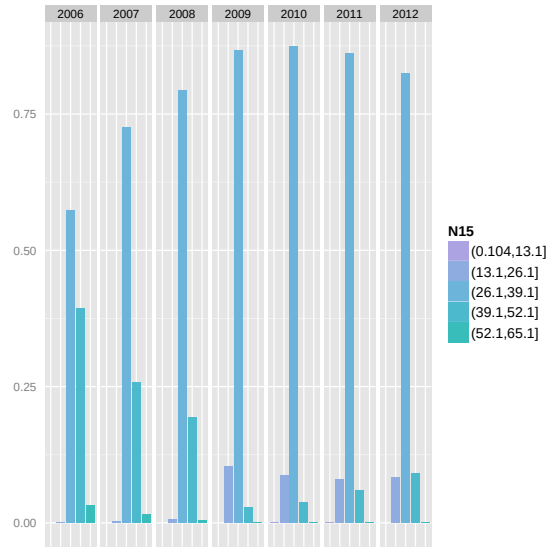

Figura 2-51: ● Estabilidad aceptable.

✓ Esta variable se acepta para integrar la *ABT*.

N16: días que pasan entre la solicitud actual y la solicitud anterior.

T1	Estadísticos	Buenos	Malos	Total
No nuevos	N	231,625	49,049	280,674
	Min	6	7	6
	Max	8.25	8.06	8.25
	$\bar{x}$	7.41	7.42	7.41
	s	0.1	0.1	0.1
	q_1	7.34	7.34	7.34
	q_2	7.39	7.4	7.39
	q_3	7.46	7.48	7.46
	Iqr	0.12	0.14	0.12
	Out	7,250	552	7,924
	Ext	1,816	31	1,722
	Na	0	0	0

Tabla 2-31: Resumen descriptivo de $N16$.

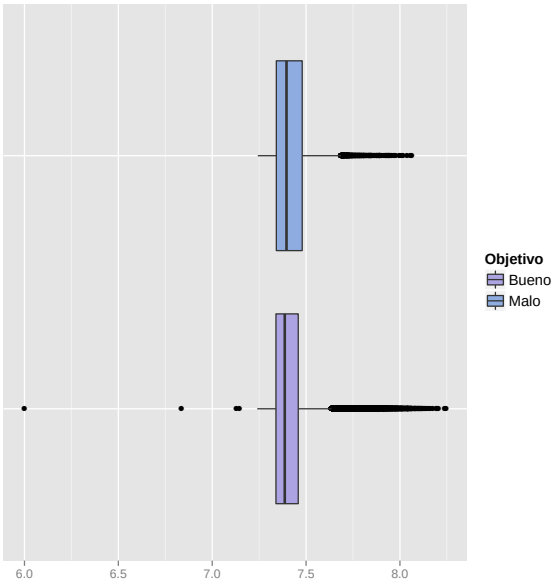

Figura 2-52: ● Distribución correcta.

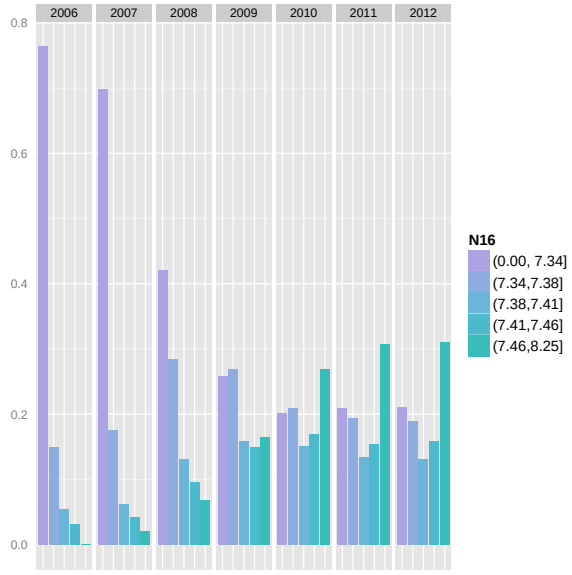

Figura 2-53: ● Estabilidad aceptable.

✓ Esta variable se acepta para integrar la ABT.

2.2.7. Análisis de componentes principales

En esta sección se presentan los resultados del *análisis de componentes principales* que se realizó sobre las variables numéricas.

La finalidad de este análisis es identificar a aquellas variables que influyan más sobre el conjunto de datos; para de esta manera determinar cuales son más relevantes para construir el modelo.

El *análisis de componentes principales* se llevó a cabo con SAS® Enterprise Miner™ 12.1.

Resultados del análisis para clientes nuevos

Previo a la ejecución de este algoritmo, se realizaron *transformaciones de rango*[4] sobre las variables para trasladarlas a un intervalo contenido en el [0,1]. A las variables transformadas se les agrega el sufijo «RANGE» para diferenciarlas de las originales. En la Figura 2-54, se muestran los estadísticos más comunes de las variables transformadas.

Variable	Mean	Standard Deviation	Minimum	Median	Maximum
RANGE_N1	0.047709	0.026998	0.006427	0.043923	0.566533
RANGE_N10	0.032969	0.034547	0.000982	0.02276	1
RANGE_N11	0.000059	0.005468	0	7.311E-7	1
RANGE_N12	0.000037	0.004072	0	3.798E-7	1
RANGE_N2	0.001305	0.003877	0	0.001092	0.983452
RANGE_N3	0.328841	0.159019	0.006494	0.324675	0.662338
RANGE_N4	0.414185	0.214519	0	0.403509	1
RANGE_N5	0.173991	0.133798	0	0.140351	1
RANGE_N6	0.189697	0.118991	0	0.166201	1
RANGE_N7	0.011798	0.011509	0	0.008649	0.560243

Figura 2-54: Principales estadísticos de las variables transformadas.

[4]Este tipo de transformaciones, son de la forma: $\frac{(x - min)}{(max - min)}$, donde x es el valor de la variable a transformar, min es el mínimo valor de la variable y max es el valor máximo de la variable.

De los resultados obtenidos se tiene la matriz de correlaciones (Figura 2-55); en la cual, se observa que son las 4 primeras componentes, las que tienen un mayor efecto sobre los datos, debido a que sus *valores propios (eigenvalues)* son mayores a 1.

Eigenvalues of Correlation Matrix

	Eigenvalue	Difference	Proportion	Cumulative
1	2.01867701	0.22035099	0.2019	0.2019
2	1.79832602	0.43195680	0.1798	0.3817
3	1.36636922	0.14032315	0.1366	0.5183
4	1.22604607	0.24767815	0.1226	0.6409
5	0.97836791	0.03784315	0.0978	0.7388
6	0.94052476	0.16024038	0.0941	0.8328
7	0.78028438	0.22254357	0.0780	0.9109
8	0.55774081	0.35601345	0.0558	0.9666
9	0.20172736	0.06979091	0.0202	0.9868
10	0.13193645		0.0132	1.0000

Figura 2-55: *Valores propios* de la matriz de correlaciones.

Además, cada una de estas 4 componentes por si sola explica al menos 10 % de la variabilidad de los datos, contrario a las demás componentes. En conjunto estas 4 componentes explican el 64.09 % de la variabilidad de los datos; por lo cual, se considera la componente 5 para acumular al menos 70 % de variabilidad explicada. Las demás componentes se descartan.

Para darle sentido a las 5 componentes principales elegidas, se tiene la Figura 2-56; en la cual, se muestran los coeficientes de las variables con los cuales se calcula cada una de las componentes principales.

Variable	Principal Component 1	Principal Component 2	Principal Component 3	Principal Component 4	Principal Component 5
RANGE_N10	0.586868	0.010223	-0.33454	-0.17839	-0.1084
RANGE_N1	0.473444	0.012499	-0.15624	0.477528	-0.25435
RANGE_N7	0.380746	0.007642	-0.07815	0.298261	0.091452
RANGE_N4	0.318902	0.003329	0.582858	-0.09024	-0.04734
RANGE_N5	0.295467	-0.00187	0.576345	-0.08957	-0.06672
RANGE_N3	-0.26639	.0004633	0.283101	0.742297	-0.13629
RANGE_N2	0.141653	-.000894	-0.0424	0.226901	0.914514
RANGE_N6	0.078933	-0.00668	0.321982	-0.18018	0.231414
RANGE_N12	-0.01058	0.706969	0.00582	-0.00378	0.004524
RANGE_N11	-0.01001	0.706976	0.004206	-0.00702	0.004032

Figura 2-56: Coeficientes de las componentes principales.

Los valores que se muestran en la Figura 2-56, se pueden interpretar de forma semejante a los valores de una correlación. Por ejemplo, se tiene que las variables $N10$ y $N1$ son las que mejor explican a la primera componente; ya que sus coeficientes (0.5868 y 0.4734 respectivamente), son los más grandes en valor absoluto para esa componente.

Analizando de igual manera para las otras componentes, se obtiene lo siguiente:

- La componente principal 1 está mejor explicada por $N10$ y $N1$; ya que tienen coeficientes de 0.5868 y 0.4734 respectivamente.
- La componente principal 2 está mejor explicada por $N12$ y $N11$; ya que tienen coeficientes de 0.7069 y 0.7069 respectivamente.
- La componente principal 3 está mejor explicada por $N4$ y $N5$; ya que tienen coeficientes de 0.5828 y 0.5763 respectivamente.
- La componente principal 4 está mejor explicada por $N3$ y $N1$; ya que tienen coeficientes de 0.7422 y 0.4775 respectivamente.
- La componente principal 5 está mejor explicada por $N2$; ya que tiene un coeficiente de 0.9145.

De lo anterior, se tiene que las variables más influyentes son: $N1$, $N2$, $N3$, $N4$, $N5$, $N10$, $N11$ y $N12$; por lo cual, utilizando estos resultados como criterio de selección, se concluye que estas variables se deben considerar para la generación del *modelo de originación*. Las demás de descartan.

Resultados del análisis para clientes no nuevos

En la Figura 2-57, se muestran los estadísticos de las variables transformadas para llevar a cabo la ejecución del algoritmo.

Variable	Mean	Standard Deviation	Minimum	Median	Maximum
RANGE_N1	0.103071	0.059592	0	0.096064	1
RANGE_N10	0.02782	0.02905	0	0.019306	1
RANGE_N11	0.000044	0.004051	0	2.653E-7	1
RANGE_N12	0.000058	0.005667	0	4.177E-7	1
RANGE_N13	0.341647	0.301466	0	0.378532	0.998783
RANGE_N14	0.160699	0.030954	0.018811	0.16086	1
RANGE_N15	0.090578	0.057478	0	0.078739	1
RANGE_N16	0.433985	0.041294	0	0.423057	0.854574
RANGE_N2	0.001331	0.00175	0	0.001167	0.365158
RANGE_N3	0.402459	0.184223	0	0.405229	0.830065
RANGE_N4	0.451112	0.213866	0	0.438596	1
RANGE_N5	0.18819	0.13567	0	0.153627	1
RANGE_N6	0.216403	0.13242	0	0.18724	0.998613
RANGE_N7	0.008102	0.007748	0	0.00609	0.479119

Figura 2-57: Principales estadísticos de las variables transformadas.

En cuanto a la matriz de correlaciones obtenida (Figura 2-58); se observa que son las 7 primeras componentes, las que tienen un mayor efecto sobre los datos, debido a que sus *valores propios (eigenvalues)* son mayores a 1.

De forma conjunta, las siete componentes explican el 74.97 % de la variabilidad de los datos. Las demás componentes se descartan, por si solas explican a lo más 5 % de la variabilidad de los datos.

Eigenvalues of Correlation Matrix

	Eigenvalue	Difference	Proportion	Cumulative
1	3.08992998	0.63335857	0.1931	0.1931
2	2.45657141	0.90296414	0.1535	0.3467
3	1.55360727	0.10602991	0.0971	0.4438
4	1.44757736	0.13368613	0.0905	0.5342
5	1.31389123	0.18583465	0.0821	0.6163
6	1.12805658	0.12328227	0.0705	0.6869
7	1.00477432	0.04629317	0.0628	0.7497
8	0.95848115	0.11257330	0.0599	0.8096
9	0.84590785	0.19467395	0.0529	0.8624
10	0.65123390	0.07967558	0.0407	0.9031
11	0.57155832	0.01950445	0.0357	0.9388
12	0.55205387	0.35030039	0.0345	0.9734
13	0.20175348	0.07079899	0.0126	0.9860
14	0.13095449	0.03730569	0.0082	0.9941
15	0.09364880	0.09364880	0.0059	1.0000
16	0.00000000		0.0000	1.0000

Figura 2-58: *Valores propios* de la matriz de correlaciones.

Los coeficientes de las variables; con los cuales, se calcula cada una de las componentes principales se muestran en la Figura 2-59.

Variable	Principal Component 1	Principal Component 2	Principal Component 3	Principal Component 4	Principal Component 5	Principal Component 6	Principal Component 7
RANGE_N8	0.544452	-0.08497	-0.08084	0.00966	-0.08038	0.058289	-0.03726
RANGE_N9	0.506362	-0.04625	-0.11259	0.014197	-0.08857	0.204007	-0.05509
RANGE_N13	0.473617	-0.24084	-0.03622	0.001889	-0.01067	-0.17404	0.009776
RANGE_N1	0.337523	0.448699	0.193241	-0.0082	-0.10486	0.162319	-0.07755
RANGE_N3	0.194835	-0.08824	0.546451	-0.05871	0.343336	0.360485	0.063726
RANGE_N16	-0.14794	0.041866	-0.16407	0.01615	-0.04175	0.725063	-0.18519
RANGE_N15	-0.11163	0.559537	0.279705	-0.01732	-0.04303	0.124486	-0.05124
RANGE_N10	0.102642	0.447375	-0.31098	0.042923	-0.36919	-0.17583	-0.11346
RANGE_N14	0.09163	0.215376	0.470456	-0.04194	0.099843	-0.40248	-0.01697
RANGE_N4	0.085042	0.19572	-0.30521	-0.01553	0.555034	-0.06174	-0.03458
RANGE_N7	0.083946	0.2797	-0.17396	0.008289	0.017666	0.025038	0.268868
RANGE_N5	0.052635	0.187982	-0.29425	-0.02105	0.546839	0.00387	-0.06691
RANGE_N2	0.023793	0.083175	-0.0645	-0.00366	-0.01346	0.107398	0.922287
RANGE_N6	0.008444	0.057702	-0.04145	-0.01531	0.307805	-0.13503	-0.0815
RANGE_N12	0.002572	0.002818	0.050256	0.703261	0.054343	-0.00688	0.010152
RANGE_N11	0.001273	0.004484	0.043975	0.704591	0.039377	0.003634	0.001855

Figura 2-59: Coeficientes de las componentes principales.

Analizando los coeficientes de las componentes principales, se obtiene lo siguiente:

- La componente principal 1 está mejor explicada por $N8$ y $N9$; ya que tienen coeficientes de 0.5444 y

0.5063 respectivamente.

- La componente principal 2 está mejor explicada por $N15$, $N1$ y $N10$; ya que tienen coeficientes de 0.5595, 0.4486 y 0.4473 respectivamente.
- La componente principal 3 está mejor explicada por $N3$ y $N14$; ya que tienen coeficientes de 0.5464 y 0.4704 respectivamente.
- La componente principal 4 está mejor explicada por $N12$ y $N11$; ya que tienen coeficientes de 0.7032 y 0.7045 respectivamente.
- La componente principal 5 está mejor explicada por $N4$ y $N5$; ya que tienen coeficientes de 0.5550 y 0.5468 respectivamente.
- La componente principal 6 está mejor explicada por $N16$ y $N14$; ya que tienen coeficientes de 0.7250 y -0.4024 respectivamente.
- La componente principal 7 está mejor explicada por $N2$; ya que tiene un coeficiente de 0.9222.

De lo anterior, se tiene que las variables más influyentes son: $N1$, $N2$, $N3$, $N4$, $N5$, $N8$, $N9$, $N10$, $N11$, $N12$, $N14$, $N15$ y $N16$; por lo cual, se deben de considerar para la generación del *modelo de originación*. Las demás de descartan.

2.3. Conclusiones

El grado de concentración de las distribuciones y el grado de estabilidad en el tiempo que presentan las variables es adecuado en términos generales.

Aunque se tiene algunas variables con deficiencias en distribución y estabilidad, debido probablemente a los cambios en las estrategias de selección de clientes que ha experimentado la empresa, además de que se almacenan características de los clientes que no aportan mucho valor.

En la Tabla 2-32, se resume la evaluación de las 23 variables de estudio, la decisión tomada acerca de su selección (tomando en cuenta solo el *análisis bivariado* y *temporal*) y el tipo de tabla *ABT* para la cual están siendo contempladas las variables:

Variables	Distribución	Estabilidad	Decisión	*ABT*
$C1$	🟢	🟢	✓	Nuevos, no nuevos
$C2$	🟢	🟢	✓	Nuevos, no nuevos
$C3$	🟢	🟢	✓	Nuevos, no nuevos
$C4$	🟢	🟢	✓	Nuevos, no nuevos
$C5$	🔴	🟢	✗	Nuevos, no nuevos
$C6$	🔴	🔴	✗	Nuevos, no nuevos
$C7$	🟢	🟡	✓	No nuevos
$N1$	🟢	🟡	✓	Nuevos, no nuevos
$N2$	🟡	🟡	✗	Nuevos, no nuevos
$N3$	🟢	🟢	✓	Nuevos, no nuevos
$N4$	🟢	🟢	✓	Nuevos, no nuevos
$N5$	🟢	🟢	✓	Nuevos, no nuevos
$N6$	🟢	🟢	✓	Nuevos, no nuevos
$N7$	🔴	🟢	✗	Nuevos, no nuevos
$N8$	🟡	🟡	✗	No nuevos
$N9$	🟡	🟡	✗	No nuevos
$N10$	🟢	🟢	✓	Nuevos, no nuevos
$N11$	🟡	🟢	✓	Nuevos, no nuevos
$N12$	🟢	🟢	✓	Nuevos, no nuevos
$N13$	🟡	🟡	✗	No nuevos
$N14$	🟡	🟢	✓	No nuevos
$N15$	🟢	🟡	✓	No nuevos
$N16$	🟢	🟡	✓	No nuevos

Tabla 2-32: Lista de variables evaluadas.

Por otra parte, contrastando los resultados obtenidos en la Tabla 2-32 contra los resultados del *análisis de componentes principales* para las variables numéricas, se tiene lo siguiente:

Para los clientes nuevos se llega a la misma decisión en la mayoría de las variables, excepto por la variable $N2$; de la cual, se podría decir que se rechaza por no tener buena estabilidad y afectaría a la vigencia del modelo. Además, $N2$ explica a la componente 5; de la cual, se sabe que la variabilidad que explica por si sola es muy pequeña.

Para los clientes no nuevos hay discrepancia en las variables $N2$, $N8$ y $N9$. De estas, se podría concluir rechazarlas por no tener buena estabilidad y para que no afecten la vigencia del modelo.

Entonces, el conjunto de variables seleccionadas (Tabla 2-32) se utiliza para integrar las tablas *ABT* (una para clientes *nuevos* y otra para *no nuevos*). Estas tablas, son el insumo para la generación de los *modelos de originación*.

En el siguiente capítulo, se presenta la teoría de diversas técnicas de *minería de datos* que se emplean para el desarrollo de los modelos.

Capítulo 3

Técnicas de minería de datos

En el Capítulo 2, se describe la metodología para construir la variable **Objetivo** y las tablas *ABT* que sirven de insumo para el desarrollo de los modelos de minería de datos.

En este capítulo, se da una breve explicación de algunas de las técnicas que existen en la minería de datos para realizar tareas de clasificación y se explica, *grosso modo*, la teoría de las técnicas que se seleccionaron para el desarrollo de los *modelos de originación*.

3.1. Minería de datos

Conocida también como *data mining*, es un proceso computacional que sirve para extraer conocimiento que se encuentra oculto en los datos. Esto lo hace identificando patrones y relaciones existentes entre los datos que permiten la creación de modelos para resolver tareas en nuevos conjuntos de datos como clasificar, predecir, agrupar, asociar, etc.

La gran efectividad que muestra la minería de datos se debe a que es un área interdisciplinaria que utiliza diversas técnicas pertenecientes a campos como *inteligencia artificial*, *estadística*, *aprendizaje de máquinas*, *reconocimiento de patrones*, entre otras (Figura 3-1).

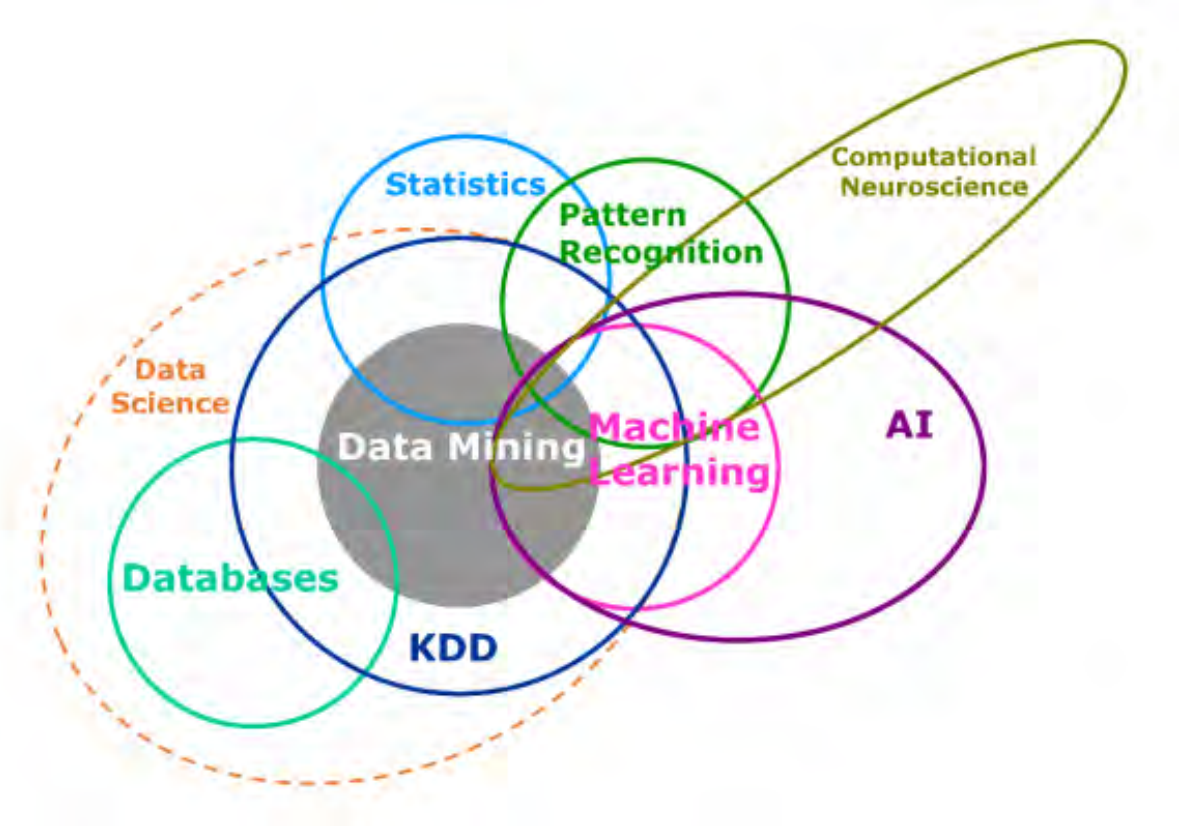

Figura 3-1: Hall, P., Dean, J., Kabul, I. K. & Silva, J. (2014). *An Overview of Machine Learning with SAS® Enterprise Miner™* [Multidisciplinary Nature of Machine Learning]. Recuperado de http://support.sas.com/resources/papers/proceedings14/SAS313-2014.pdf

3.1.1. Tipos de aprendizaje en los modelos

Dentro de la minería de datos, se pueden clasificar los modelos en base al tipo de *aprendizaje* que utilizan para resolver tareas determinadas; estos son: *supervisado* y *no supervisado*.

Los *modelos de aprendizaje supervisado* se caracterizan por utilizar un conjunto de observaciones; de las cuales, ya se conoce el resultado de la variable objetivo y a partir de estas observaciones se genera un modelo que será aplicado a nuevos casos para predecir dicha *variable objetivo*. De forma general, se puede decir que los *modelos de aprendizaje supervisado* aprenden de una base de conocimiento *a priori* y realizan tareas como clasificación, predicción, etc.

Por otra parte, los *modelos de aprendizaje no supervisado* se ajustan de forma automática a

los datos que se les presentan y dadas sus características son utilizados para resolver otras tareas distintas a las que se resuelven con *aprendizaje supervisado*, por ejemplo agrupación (*clusters*), reducción de dimensiones (componentes principales), etc.

3.1.2. Técnicas de minería para tareas de clasificación

La clasificación es una tarea que se ocupa constantemente en la vida diaria y tiene que ver con el distinguir objetos para asignarlos a categorías que son mutuamente excluyentes y que son conocidas como clases.

Existen varias técnicas (*modelos de aprendizaje supervisado*) que son útiles para llevar a cabo la tarea de clasificación; las que se utilizan en este trabajo son: *árboles de decisión*, *redes neuronales* y *regresión logística*; las cuales, se describen en la siguiente sección.

3.2. Árboles de decisión

3.2.1. Definición

Un *árbol de decisión* es un modelo que predice una variable objetivo con base en el aprendizaje de simples reglas de decisión, inferidas desde las características de los datos. Si la *variable objetivo* es categórica se le conoce como *árbol de clasificación* y si es numérica se le conoce como *árbol de regresión*.

En cuanto a los *árboles de clasificación*, si la variable objetivo tiene dos categorías, el modelo es un *árbol binario* y cuando la variable objetivo tiene más de dos niveles se le llama *árbol múltiple*. En particular, en este trabajo se utiliza el *árbol de clasificación binario*, debido a la naturaleza de la variable objetivo construida en el capítulo anterior.

Los árboles de decisión son llamados así por que están representados con una estructura similar a la de un árbol (tienen raíz, ramas y hojas). Para predecir el valor de la *variable obje-*

tivo, cada uno de los casos evaluados se pasan por el árbol y lo recorren de arriba hacia abajo comenzando por el primer nodo, llamado nodo raíz, avanzan a través de las ramas (son reglas o condiciones) para llegar a algún nodo interno (conocidos también como nodos hijo) y si la desigualdad que se les presenta en cada nodo interno es cierta el caso evaluado se mueve hacia la izquierda, en caso contrario avanza hacia la derecha.

De esta forma el caso evaluado se desplaza hasta llegar a un nodo terminal; el cual, es conocido como hoja y en este nodo terminal se encuentra el resultado de la predicción hecha por el modelo. En la Figura 3-2 se representa la estructura de un *árbol de decisión*.

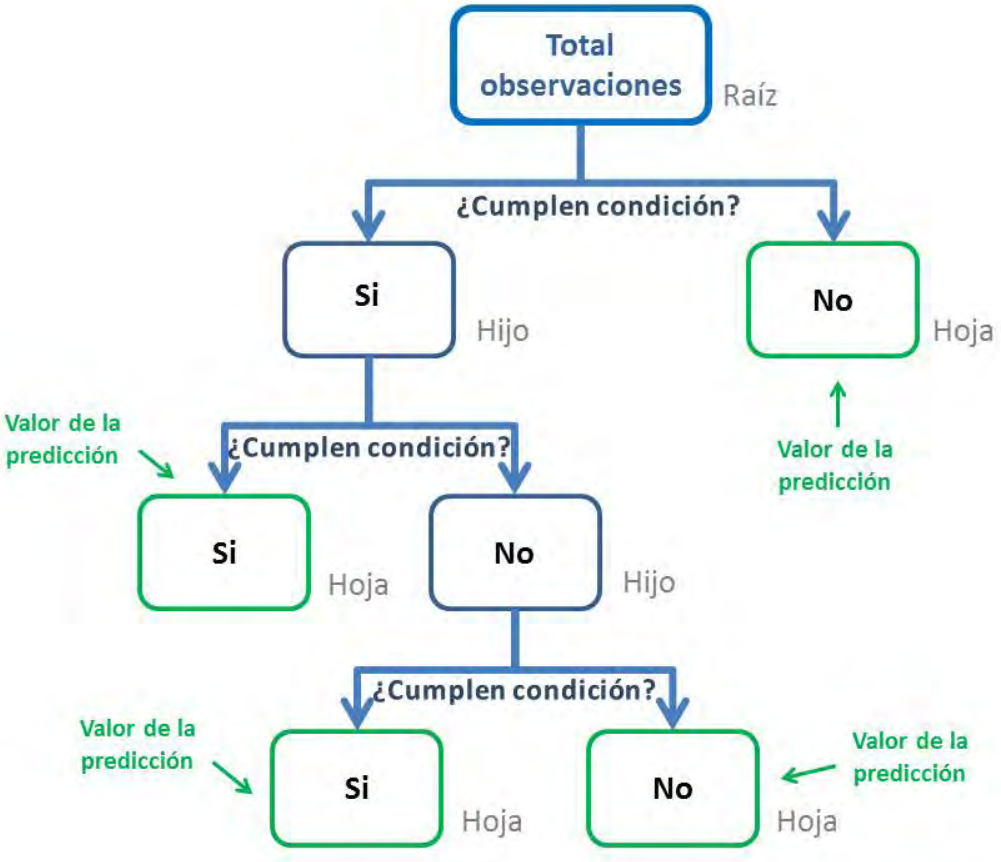

Figura 3-2: Estructura de un *árbol de decisión*.

3.2.2. Modelado

Para el desarrollo del *árbol de decisión* se utiliza el método *partición recursiva*; donde una partición se entiende como un punto de corte en alguna variable de entrada. Estos cortes son las reglas o condiciones que evalúa el árbol para poder separar las observaciones en las clases de la *variable objetivo*.

En este método se empieza partiendo el nodo raíz, el cual contiene todas las observaciones de entrenamiento. Para hacer la partición en este nodo, primero se busca la mejor partición que puede aportar cada una de las variables de entrada y cada una de ellas se pone a competir para de esta manera elegir a la mejor parición de todas que mejor discrimine las clases. Una vez que la partición ha sido elegida, se procede a dividir el nodo raíz.

Con la partición hecha, se generan nuevos nodos en el árbol; los cuales, son llamados nodos hijo. En cada uno de estos nodos se repite el procedimiento para encontrar la mejor partición y se itera el proceso hasta hacer crecer al árbol hasta su máximo tamaño; es decir, hasta que ya no se pueda partir ningún nodo. Los últimos nodos que se generan en el árbol son los que determinarán la clase a la que pertenece la observación evaluada por el árbol.

La meta que se busca al hacer las particiones, es que los nuevos nodos generados sean puros; es decir, que sean homogéneos en el tipo de observaciones que contienen. Entonces, cuando todas las observaciones que hay en un nodo son del mismo tipo, se puede decir que hay pureza en ese nodo. Si por el contrario las observaciones que hay en cada nodo son de diferentes tipos, se dice que este nodo es un nodo con impureza.

Por otra parte, para determinar cual es la mejor partición, se definen *métricas de pureza* que se utilizan para medir la *reducción de variabilidad (impureza)* (ver Apéndice E) en los nodos. Las métricas más comunes son las siguientes:

- **Índice de *Gini*** (ver Apéndice C)
- **Entropía** (ver Apéndice D)

- **Prueba χ^2 *Pearson*** (ver Apéndice F)

Una vez que se ha creado el árbol, se tiene que podar; es decir, se tiene que elegir un subconjunto de nodos llamado subárbol. Esto es debido a que el árbol generado se adapta por si mismo a las características de los datos de entrenamiento y generalmente no se ajusta bien cuando se le aplica un nuevo conjunto de datos.

Para elegir al subárbol, se deben decidir que nodos serán excluidos y para esto se debe utilizar alguno de los métodos conocidos, estos son los siguientes:

- **Evaluación**: selecciona el árbol que tenga el valor promedio de ajuste más grande y el valor promedio de pérdida más pequeño; es decir, selecciona el subárbol más pequeño con el mejor ajuste. Y el ajuste está dado por la *medida de evaluación* que sea elegida.

- **Largest**: este método selecciona el árbol completo.

- **N**: este método selecciona el subárbol con a lo más N niveles.

Finalmente, se debe evaluar el ajuste que tiene el subárbol elegido y para esto, se definen las *medidas de evaluación*. Las más comunes se muestra a continuación:

- **Decisión:** se puede definir una matriz de pérdida-beneficio y de esta manera se puede seleccionar al árbol que tenga que el beneficio promedio más alto y la perdida promedio más pequeña. Si no se define la matriz, entonces la medida se define como la *tasa de clasificación errónea*.

- **Error cuadrado promedio:** se selecciona el el error cuadrado promedio más pequeño.

- **Tasa de clasificación errónea:** se selecciona el árbol con el error de clasificación más pequeño.

- **Elevación:** la evaluación del árbol se basa en los calificaciones obtenidas en las primeras n observaciones. Y las calificaciones de las observaciones, están basadas en el valor de la variable objetivo predicho.

Las definiciones descritas para la técnica de *árboles de decisión*, fueron tomadas del módulo «Ayuda» que ofrece el Software SAS® Enterprise Miner™ 12.1.

3.2.3. Conclusiones

El modelo de *árboles de decisión*, presenta grandes ventajas, algunas de ellas son:

- Es simple de entender y de interpretar; además de que puede ser visualizado.
- Requiere poca preparación de los datos (no se tiene que normalizar datos, quitar valores ausentes, etc.).
- El costo de utilizar árboles en la predicción de datos es logarítmico en el número de datos usados para entrenar el árbol.
- Puede manejar datos numéricos y categóricos.

Por otra parte, las desventajas que presenta son:

- Si no se poda adecuadamente, se puede tener un árbol muy complejo que esté sobreajustado.
- Si se tienen clases dominantes en alguna variable que no se han identificado, se puede generar un árbol sesgado.
- Los árboles pueden ser inestables debido a pequeñas variaciones que se presenten en los datos de entrenamiento.

3.3. Redes neuronales

3.3.1. Definición

Este modelo es un paradigma de procesamiento de información, su arquitectura está inspirada en el sistema nervioso biológico y la forma como el cerebro procesa la información.

Su estructura se compone de un gran número de elementos de procesamiento de la información llamados *neuronas*. Estas trabajan en conjunto para resolver problemas específicos como el

reconocimiento de patrones o la clasificación de datos a través de un proceso de aprendizaje.

Las *redes neuronales* presentan diversas arquitecturas; en el caso de la arquitectura más simple, la *red neuronal* tiene una unidad de entrada (variable independiente), una *variable objetivo* (variable dependiente) y una sola unidad de salida (valores predichos).

La forma como se relaciona la unidad de entrada con la *variable objetivo* es a través de una *función de activación*; la cual, puede ser lineal o puede ser la función identidad. En términos estadísticos, se podría decir que este modelo es una *simple regresión lineal* y si la *función de activación* fuera una función logística, se convertiría en un modelo de *regresión logística*. En la práctica, es muy común que se utilice la función logística como *función de activación*.

Por otra parte, existe una arquitectura más compleja, el *perceptrón*; el cual, es un modelo lineal discriminante. Este modelo utiliza una combinación lineal de varias entradas como *función de activación*. La arquitectura de esta red se muestra en la Figura 3-3.

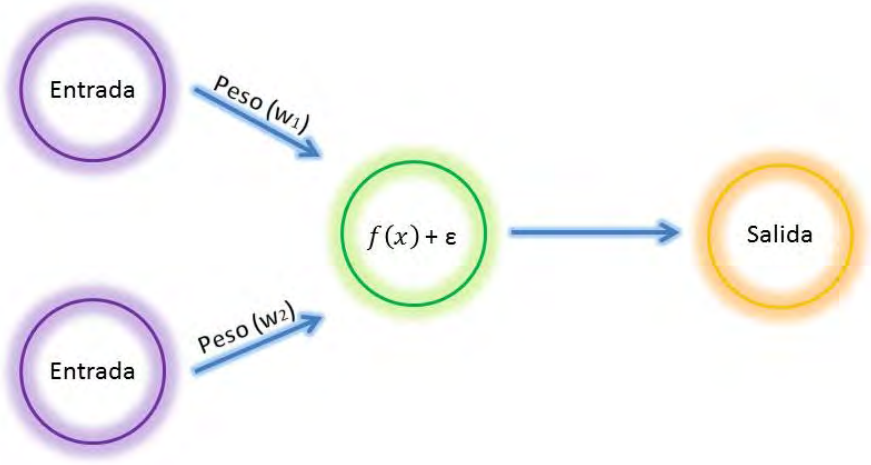

Figura 3-3: Arquitectura de un *perceptrón* con dos entradas.

En los modelos de *perceptrón*, se pueden aplicar transformaciones adicionales a través de *capas ocultas*. Normalmente, cada unidad de entrada es conectada a una unidad de la *capa oculta* y esta a su vez está conectada a una unidad de salida. Las *capas ocultas* combinan los valores de

entrada y aplican una *función de activación*; la cual, puede ser lineal o sigmoidea (aunque puede presentar otro tipo de *funciones de combinación* y *de activación*). Estos valores calculados que se generan, son nuevamente combinados con una *función de activación* en las unidades de salida para de esta manera generar el resultado.

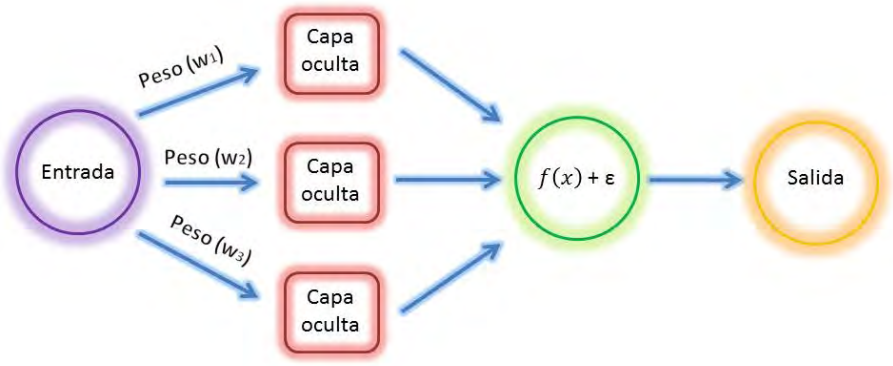

Figura 3-4: Forma de un *perceptrón* con una capa.

Si la *red neuronal* presenta más de una capa, entonces la red es conocida como *perceptrón multicapa*. Esta arquitectura, es la más utilizada y sus características son las siguientes:

- Puede tener cualquier cantidad de entradas.
- Capas ocultas con cualquier cantidad de unidades.
- Utilizar funciones de combinación lineales en las capas ocultas y de salida.
- Utilizar funciones de activación sigmoideas en las capas ocultas.
- Tener cualquier cantidad de salidas con cualquier función de activación.
- Tener conexiones entre la capa de entrada y la primer capa oculta, entre las capas ocultas y entre las últimas capas ocultas y la capa de salida.
- Dada suficiente información, capas ocultas y tiempo de entrenamiento, un *perceptrón multicapa* puede aprender a aproximar cualquier función a cualquier grado de precisión.

3.3.2. Modelado

Para el entrenamiento de una red se utiliza el método *back-propagation*. En este método, se define el valor de los pesos y sesgos en la *red neuronal*, después los valores de las variables de

entrada son ingresados a la *red neuronal* para calcular los valores de salida y compararlos con los valores de la variable objetivo a través de una *función de error* (también conocida como *criterio de estimación*).

Si el error es muy grande, se redefine el valor de los pesos y sesgos de la red, en caso contrario se mantienen los valores. Este proceso se repite varias veces buscando ajustar la *red neuronal* de la forma más precisa para que los resultados de salida sean lo más cercanos a los valores que tiene la variable objetivo. En cada actualización de los parámetros, se busca converger a una solución en un número de iteraciones finitas; siempre y cuando la solución exista.

Para iniciar el entrenamiento de una *red neuronal* se aplican diferentes criterios; ya que no hay un método conocido para calcular el valor inicial de los pesos. Uno de estos criterios es asignar valores de forma aleatoria a los pesos y sesgos de la red para después irlos ajustando. Por su parte, el software SAS® Enterprise Miner™ 12.1, emplea los siguientes criterios:

- **Redes con funciones de combinación lineal:** los pesos aleatorios de inicio se ajustan dividiendo entre la raíz cuadrada del número de conexiones de la unidad.
- **Redes con funciones de combinación no lineal:** se entrena la red para algunas pequeñas iteraciones, partiendo de números grandes (pude ser 10, 100, o 1000, etc.) para inicializaciones aleatorias.

Por otra parte, para calcular el valor de los pesos en cada iteración se utilizan diversos métodos. Algunos de ellos se mencionan a continuación:

- ***Back-propagation*:** Es el más popular de todos; aunque es lento, poco confiable y requiere que el usuario afine de forma manual la tasa de aprendizaje; lo cual, lo vuelve tedioso.
- **Región de confianza (*trust-region*):** es recomendado para problemas pequeños y medianos de optimización con más de 40 parámetros.
- ***Levenberg-Marquardt*:** es muy rápido y confiable para redes neuronales pequeñas pero requiere de grandes cantidades de memoria.
- ***Conjugate gradient*:** es bueno para redes neuronales grandes donde no hay suficiente memoria. Usualmente requiere más iteraciones que las técnicas anteriormente descritas pero requiere menos cálculo de

punto flotante.

En general, en todas las técnicas de optimización convencionales, la *función objetivo* decrece en cada iteración hasta que un mínimo local es alcanzado y es en ese punto donde el algoritmo termina. En el caso del *Back-propagation*, la *función objetivo* tiende a bajar y subir repetidamente durante el entrenamiento. Pero si se especifica una tasa de aprendizaje alta, los pesos divergerán y la *función objetivo* se incrementará sin limite.

En cuanto a las *funciones de error*, las más comunes para modelar una *variable objetivo* de tipo numérica son las siguientes:

- **Distribución Normal:** es adecuada para variables objetivos que no están acotadas y tienen una varianza condicional constante, no tiene datos atípicos y tienen distribución simétrica.
- **Distribución *Gamma*:** es adecuada cuando se presenta sesgo, la variable objetivo pertenece a intervalos positivos y su desviación estándar es proporcional a la media condicional.
- **Distribución *Poisson*:** es adecuada cuando se presenta sesgo, la variable objetivo pertenece a intervalos no negativos, especialmente cuando se tiene conteo de eventos raros y cuando la varianza condicional es proporcional a la media condicional.

Las *funciones de error* más comunes para modelar una *variable objetivo* de tipo categórica son las siguientes:

- **Distribución Bernoulli:** es adecuada para variables objetivo que toman los valores 0 y 1.
- **Bernoulli Múltiple:** es adecuada en general para varaibles objetivos nominales u ordinales.

Las definiciones descritas para la técnica de *red neuronal*, fueron tomadas del módulo «Ayuda» que ofrece el Software SAS® Enterprise Miner™ 12.1.

3.3.3. Conclusiones

El modelo de *red neuronal*, presenta grandes ventajas, algunas de ellas son:

- Tiene la habilidad de aprender como hacer ciertas tareas basándose en cierta información.
- Crea su propia organización o representación de la información que recibe durante el tiempo de entrenamiento.
- Todos los procesos computacionales se llevan a cabo en paralelo.
- La destrucción parcial de una red neuronal dirige a una degradación en el desempeño del modelo. Sin embargo, la red neuronal puede retener algunas de sus capacidad aunque se dañe la red principal.

Por otra parte, las desventajas que presenta son:

- La naturaleza de *caja negra* que tiene.
- La gran carga computacional que presenta.
- La propensión al sobreajuste en su modelado.
- La naturaleza empírica para el desarrollo del modelo.

3.4. Regresión logística

3.4.1. Definición

Este modelo también es conocido como *regresión logit* o *modelo logit* y es una transformación del *modelo lineal generalizado*[1]. Su objetivo es clasificar observaciones y lo hace con base en la probabilidad que presenta la observación de pertenecer a alguna de las categorías de la *variable objetivo*, dado que presenta determinados valores en las variables de entrada.

Cuando la *variable objetivo* a explicar tiene dos clases, se le conoce como *regresión logística binaria* y cuando tiene más de dos clases, se le conoce como *regresión logística multinomial*. En particular, en este trabajo se explica de forma general el caso binario, debido a que la *variable objetivo* construida en el Capítulo 2, para el caso de estudio presentado, es de tipo binaria.

[1] Los *modelos lineales generalizados* son una generalización de la *regresión lineal* que permiten que la distribución de error en las variables de respuesta sean diferentes a la distribución normal.

La forma como este modelo relaciona la *variable objetivo* con las variables de entrada es a través de la *función logística* (ecuación 3-1).

$$\sigma(x) = \frac{1}{1 + e^{-(\beta_0 + \beta_1 x_1 + \cdots + \beta_n x_n)}} \tag{3-1}$$

Donde:

β_0 es el interceptor de la ecuación de regresión.

$x_{i\ ,\ i \in 1,2,\cdots,n}$ representa las i-ésima variable independiente (*regresor*) del modelo.

$\beta_{i\ ,\ i \in 1,2,\cdots,n}$ representa el i-ésimo coeficiente de regresión del modelo.

Ya que los valores de la combinación lineal $(\beta_0 + \beta_1 x_1 + \cdots + \beta_n x_n)$ pueden tomar valores en el intervalo $(-\infty, \infty)$ y los valores que regresa la *función logística* están acotados en el intervalo [0,1] (Figura 3-5); se puede interpretar esta función como la probabilidad de que la variable dependiente sea igual a cierto valor dado los valores de x_i y β_i.

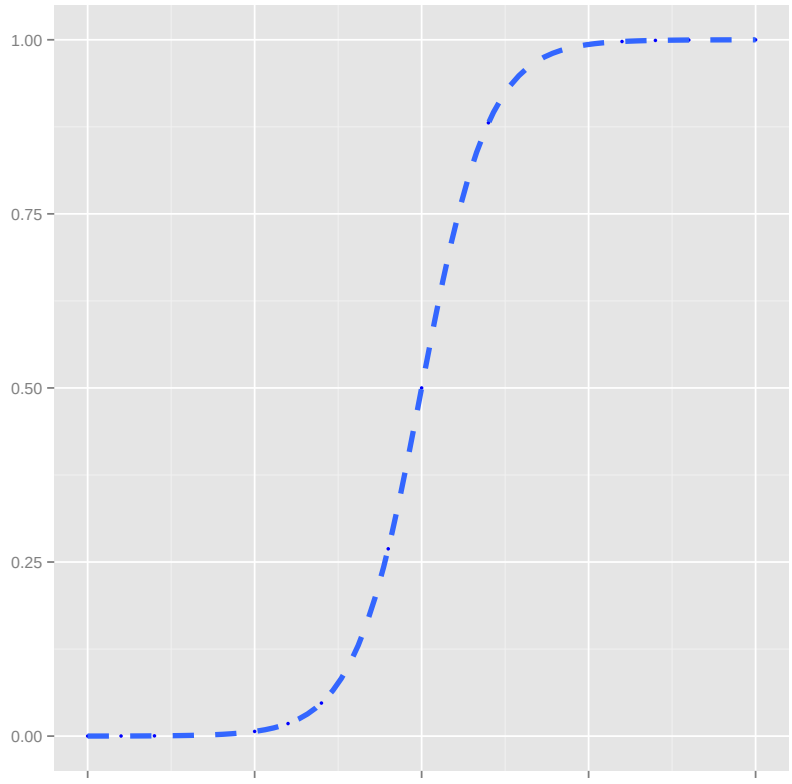

Figura 3-5: Gráfica de la función logística.

Para el caso de estudio de este trabajo, la función de probabilidad que describe el modelo de *regresión logística* es la probabilidad de que el cliente al que se evalúa esté en la categoría de $Malo$, dado las características que presenta en su solicitud.

3.4.2. Modelado

Para la estimación de los *coeficientes de regresión*, se utiliza la *función de máxima verosimilitud*; la cual, se calcula con el método *NewtonRaphson*[2]. Con este método, además de calcular los *coeficientes de regresión*, se obtienen los errores estándar y las covarianzas de las variables de entrada.

[2] *NewtonRaphson* es un método iterativo que permite aproximar la solución de una ecuación del tipo $f(x) = 0$.

En cuanto a la selección del modelo (conjunto de regresores), se prueban diferentes combinaciones de variables de entrada, para ver como se afecta el modelo. Este proceso solo consta de quitar o agregar regresores y se hace de forma iterativa utilizando los siguientes métodos:

- **Hacia delante (*Forward*):** este método comienza con un modelo sin variables de entrada, sistemáticamente se va agregando una de las variables candidatas y se evalúa el efecto utilizando el criterio definido. Si la variable agregada mejora el modelo se mantiene, de lo contrario se rechaza y se sigue con el proceso.
- **Hacia atrás (*Backward*):** este método comienza con un modelo formado con todas las variables de entrada, sistemáticamente se va quitando una de las variables candidatas y evalúa el efecto utilizando el criterio definido. Si al quitar la variable mejora el modelo se descarta la variable, de lo contrario se mantiene y se sigue con el proceso.
- **Paso por paso (*Stepwise*):** este método es una combinación de los dos anteriores, se evalúa el modelo cada vez que se quita o se incluye una variable.

Por otra parte, para medir el desempeño general del modelo, se tienen definidos los siguientes criterios:

- **Criterio de información de *Akaike* (CIA):** evalúa la sobreparametrización. El modelo que presente el valor de CIA más bajo es el elegido. Se define como: $n \ln\left(\frac{SSE}{n}\right) + 2p$, donde n es el número de casos, SSE es la suma del error cuadrado y p es el número de parámetros que tiene el modelo.
- **Criterio *bayesiano de Schwarz* (CBS):** evalúa la sobreparametrización. El modelo que presente el valor de CBS más bajo es el elegido. Se define como: $n \ln\left(\frac{SSE}{n}\right) + p \ln(n)$, donde n es el número de casos, SSE es la suma del error cuadrado y p es el número de parámetros que tiene el modelo.
- **Tasa de clasificación errónea:** evalúa la imperfección del modelo al hacer clasificaciones. El modelo que presente el error más pequeño es el elegido. Se define como: $\frac{1}{n}\sum_i \mathbb{I}(y_i \neq \hat{y}_i)$, donde n es el número de casos, y_i representa al valor de la variable objetivo en la observación i, $\hat{y}_i$ representa la predicción de la variable objetivo en la observación i y $\mathbb{I}(y_i \neq \hat{y}_i)$ es la función indicadora que vale 1 cuando y_i es diferente de $\hat{y}_i$ y 0 en otro caso.
- **Tasa de clasificación errónea con validación cruzada:** evalúa la imperfección del modelo al hacer clasificaciones y prueba que los resultados obtenidos sean independientes de los datos de entrenamiento utilizados, es decir evalúa el sobreajuste; por lo que, el modelo que presente el error más pequeño es el elegido. La validación cruzada consiste en dividir los datos en k subconjuntos, uno de los subconjuntos se utiliza como datos de prueba y los $k-1$ restantes se utilizan como datos de entrenamiento. El proceso de

validación se repite k veces y se obtiene la media aritmética de la tasa de clasificación errónea.

En resumen, para elegir al mejor modelo se tiene que seleccionar el conjunto de variables que aporten más en la explicación de la variable dependiente y que no sean redundantes entre ellas. Es decir, se busca, que el modelo presente un buen ajuste y a su vez cumpla el *principio de parsimonia*[3].

3.4.3. Conclusiones

Este modelo presenta ciertas ventajas, la principal es que los datos no deben de cumplir ninguna hipótesis de las que piden los *modelos lineales generalizados*, como: *linealidad*, *normalidad* y *homocedasticidad*.

Por otra parte, uno de los problemas con los que se enfrenta la *regresión logística* es que requiere suficiente información para que los resultados sean estables y significativos.

3.5. Conclusiones

Estas técnicas de *minería de datos*, descritas de forma general, son las que se emplean para el desarrollo del *modelo de originación*. Las tres técnicas se desarrollan en cada uno los conjuntos de datos generados (tablas *ABT* de clientes nuevos y no nuevos), se comparan su desempeños y se selecciona al mejor modelo. Todo esto es lo que se describe en el siguiente capítulo.

[3]Principio metodológico atribuido al fraile Guillermo de Ockham, según el cual: «en igualdad de condiciones, la explicación más sencilla suele ser la correcta».

Capítulo 4

Generación de modelos

La construcción de los modelos se hace con las tablas ABT que se construyeron en el Capítulo 2 y utilizando las técnicas de minería de datos descritas en el Capítulo 3. En este capítulo, se describe el proceso de modelado que consiste en la partición de los datos, selección de las variables para el modelo, selección de la técnica de minería, el refinamiento de la técnica elegida y las conclusiones. Todo el desarrollo se hizo con SAS® Enterprise Miner™ 12.1.

4.1. Partición de los datos

Previo al modelado de datos, las solicitudes deben de ser clasificadas en grupos para que se puedan entrenar, validar y probar los modelos. Estos grupos, se forman para las dos ABT y son los siguientes:

- **Entrenamiento**: conjunto de datos que sirve de entrenamiento para los modelos, se conforma con el 60 % de las solicitudes de crédito.
- **Validación**: conjunto de datos que se asigna para la validación de los modelos, se conforma con el 20 % de las solicitudes de crédito.
- **Prueba**: conjunto de datos que se utiliza para probar la asertividad del modelo, se conforma con el 20 % de las solicitudes de crédito.

En la Tabla 4-1, se muestra como están conformados estos grupos.

ABT	Grupo	Objetivo	Solicitudes	Proporción (%)
Nuevos	Total	*Bueno*	271,644	78.25
		Malo	75,461	21.74
	Entrenamiento	*Bueno*	162,986	78,25
		Malo	15,093	21.74
	Validación	*Bueno*	54,328	78.25
		Malo	15,092	21.74
	Prueba	*Bueno*	54,330	78.25
		Malo	15,093	21.75
No nuevos	Total	*Bueno*	231,625	82.52
		Malo	49,049	17.47
	Entrenamiento	*Bueno*	138,974	82,51
		Malo	29,428	17.47
	Validación	*Bueno*	46,325	82.52
		Malo	9,810	17.47
	Prueba	*Bueno*	46,326	82.52
		Malo	9,811	17.47

Tabla 4-1: Partición de las *ABT*.

Estos grupos, se generan con una partición estratificada, que ya está predefinida en el *software*, y de esta manera se logra conservar las mismas proporciones de la variable **Objetivo** en cada uno de los grupos.

Después de que segmentan las solicitudes, se hace una selección de las variables que tienen un mayor aporte estadístico para explicar a la variable **Objetivo**, esto se describe en la sección 4.2.

4.2. Selección de variables

Cada una de las *ABT*, está integrada por variables (independientes) que pueden ser utilizadas para la predicción de la variable **Objetivo** (variable dependiente). Pero antes de ingresarlas a los modelos, se debe evaluar si estas variables independientes presentan alguna relación con la variable dependiente que permita predecirla. Para hacer esta selección se ejecuta el nodo

«Selección de variables».

4.2.1. Parametrización de la selección de variables

Los parámetros que se utilizaron en el nodo «Selección de variables», son los que están predeterminados por el *software* y se muestran en la Figura 4-1.

Entrenamiento	
Variables	...
Nivel de clase máx.	100
Porcentaje de ausentes máximo	50
Modelo objetivo	Predeterminado
Selector manual	...
Rechaza las entradas no utilizadas	Sí
⊟ Opciones de bypass	
┊ Variable	Ninguno
┊ Rol	Entrada
⊟ Opciones Chi-cuadrado	
┊ Número de clases	50
┊ Número máximo de pasadas	6
┊ Chi-cuadrado mínimo	3.84
⊟ Opciones R-cuadrado	
┊ Número máximo de variables	3000
┊ R-cuadrado mínimo	0.0050
┊ Detener R-cuadrado	5.0E-4
┊ Utilizar variables AOV16	No
┊ Utilizar variables de grupo	Sí
┊ Utilizar interacciones	No
Usar la librería SPD Engine	Sí
Opción Imprimir	Predeterminado

Figura 4-1: Parámetros del criterio de selección de variables.

La descripción de las opciones con las que cuenta el nodo «Selección de variables» se pueden ver, *grosso modo*, en el Apéndice G.

4.2.2. Resultados de la selección de variables

Dado que la variable dependiente no tiene más de 400 grados de libertad, condición para que el *software* utilice el criterio de la *Ji-cuadrada* (χ^2), entonces el criterio que se utiliza es el de la *R-cuadrada* (R^2). Este criterio, consiste en aplicar una regresión lineal con el método *hacia adelante* que maximice el valor de R^2. De esta manera, se descartan fácilmente las variables que no aportan nada para la explicación de la variable dependiente.

El valor de la R^2 que tiene asociado cada variable, es el que se utiliza para decidir si la variable se descarta o se acepta. Este valor es importante porque muestra el porcentaje de variabilidad que se puede explicar de la variable dependiente a través de la variable independiente.

Resultados de la selección de variables para el modelo de clientes nuevos

De los resultados que arroja el nodo «Selección de variables», los más relevantes son: *efectos elegidos para la variable objetivo* y *tabla ANOVA final*; que se muestran en la Figura 4-5 y 4-6.

The DMINE Procedure

Effects Chosen for Target: Objetivo

Effect	DF	R-Square	F Value	p-Value	Sum of Squares	Error Mean Square
Group: C4	1	0.363968	119176	<.0001	12896	0.108214
Var: N6	1	0.106886	42068	<.0001	3787.290237	0.090029
Var: N4	1	0.047096	20347	<.0001	1668.758422	0.082016
Class: C1	1	0.016292	7284.798290	<.0001	577.280086	0.079244
Var: N5	1	0.011576	5307.862774	<.0001	410.166864	0.077275

Figura 4-2: *Efectos elegidos para la variable objetivo.*

En los resultados obtenidos en *efectos elegidos para la variable objetivo* (Figura 4-5), se muestra que son 5 variables las que muestran mayor efecto sobre la *variable objetivo*; estas son $C4$, $N6$, $N4$, $C1$ y $N5$.

De estas variables, se puede observar que $C4$ presenta una R^2 de 0.3639; esto significa que con esta variable se logra explicar el 36.39 % de la variabilidad de la *variable objetivo*. Por otra parte, $C4$ presenta un $p - valor$ menor a 0.0001; con lo cual, se concluye que la media de esta variable en cada uno de los niveles de la *variable objetivo* es diferente. Además, el estadístico F *de Snedecor* ($F - valor$) que presenta la variable $C4$ es el más grande y esto indica que discrimina mejor los niveles de la *variable objetivo* en comparación con las otras variables.

Contextualizando los resultados, se concluye que la variable $C4$, que contiene el número de cuentas con atraso que tiene el cliente en BC, discrimina bien las categorías *Bueno* y *Malo* de la variable **Objetivo**. De forma análoga, se obtienen las conclusiones para las variables $N6$, $N4$, $C1$ y $N5$.

The DMINE Procedure

The Final ANOVA Table for Target: Objetivo

Effect	DF	R-Square	Sum of Squares
Model	5	0.545818	19340
Error	208256	.	16093
Total	208261	.	35433

Figura 4-3: *Tabla ANOVA final.*

En los resultados de la *tabla ANOVA final* (Figura 4-6), se puede ver que el modelo integrado por las variables $C4$, $N6$, $N4$, $C1$ y $N5$, presenta una R^2 de 0.5458; es decir, con esas cinco variables se logra explicar el 54.58 % de la variabilidad de la *variable objetivo*. Por otra parte, se tiene que el *total de la suma de cuadrados* es de 35,433; el cual, de forma aislada no dice nada, pero si se compara con el *total de la suma de cuadrados* de otros modelos se tiene un parámetro para medir el error del modelo. Y lo que se busca es que este número sea lo más pequeño posible; lo cual, es un indicador de que el modelo tiene un buen ajuste.

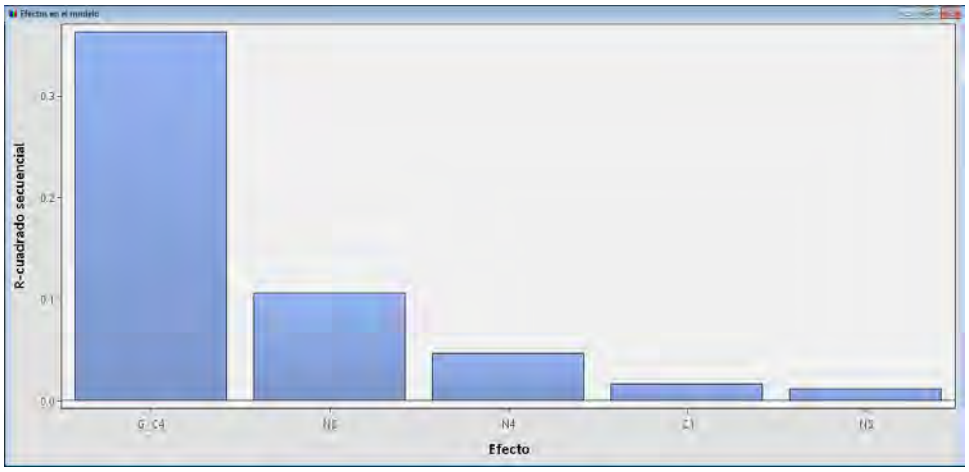

Figura 4-4: Variables con mayor efecto sobre la variable objetivo.

En resumen, las variables $C4$, $N6$, $N4$, $C1$ y $N5$ son elegidas para integrar el modelo, porque tienen un mayor efecto sobre la *variable objetivo* (Figura 4-7). Las demás variables no son relevantes estadísticamente y se excluyen del modelo.

Cabe mencionar que la variable $C4$ es una variable categórica ordinal y para el análisis de selección de variables fue agrupada con el fin de reducir sus niveles; es por eso que en la Figura 4-7 aparece con el sufijo «G_».

Resultados de la selección de variables para el modelo de clientes no nuevos

Los resultados *efectos elegidos para la variable objetivo* y *tabla ANOVA final*, del nodo «Selección de variables», se muestran en la Figura 4-5 y 4-6.

The DMINE Procedure

Effects Chosen for Target: Objetivo

Effect	DF	R-Square	F Value	p-Value	Sum of Squares	Error Mean Square
Group: C4	1	0.268983	61964	<.0001	6532.375345	0.105422
Var: N6	1	0.092605	24427	<.0001	2248.954539	0.092068
Var: N4	1	0.055832	16139	<.0001	1355.914904	0.084017
Class: C1	1	0.019685	5888.895410	<.0001	478.051299	0.081178
Var: N15	1	0.016388	5049.773533	<.0001	398.000110	0.078815
Var: N5	1	0.012707	4008.739498	<.0001	308.605893	0.076983
Var: N2	1	0.006148	1962.147653	<.0001	149.313628	0.076097
Var: N1	1	0.002891	927.570031	<.0001	70.199070	0.075681
Var: N14	1	0.001839	592.229311	<.0001	44.663465	0.075416
Group: C7	3	0.002261	243.697632	<.0001	54.898607	0.075091
Var: N12	1	0.001612	522.869261	<.0001	39.141547	0.074859

Figura 4-5: *Efectos elegidos para la variable objetivo.*

En los resultados *efectos elegidos para la variable objetivo* (Figura 4-5), se tiene que son 11 variables las que muestran mayor efecto sobre la *variable objetivo*; estas son $C4$, $N6$, $N4$, $C1$, $N15$, $N5$, $N2$, $N1$, $N14$, $C7$ y $N12$.

De estas variables, se puede observar que $C4$ presenta una R^2 de 0.2689; esto significa que con esta variable se logra explicar el 26.89 % de la variabilidad de la *variable objetivo*. Por otra parte, $C4$ presenta un $p - valor$ menor a 0.0001; con lo cual, se concluye que la media de esta variable en cada uno de los niveles de la *variable objetivo* es diferente. Además, el estadístico $F - valor$ que presenta la variable $C4$ es el más grande y esto indica que discrimina mejor los niveles de la *variable objetivo* en comparación con las otras variables.

Contextualizando los resultados, se concluye que la variable $C4$, que contiene el número de cuentas con atraso que tiene el cliente en BC, discrimina bien las categorías *Bueno* y *Malo* de la variable **Objetivo**. De forma análoga, se obtienen las conclusiones para las variables $N6$, $N4$, $C1$, $N15$, $N5$, $N2$, $N1$, $N14$, $C7$ y $N12$.

The DMINE Procedure

The Final ANOVA Table for Target: Objetivo

Effect	DF	R-Square	Sum of Squares
Model	13	0.480950	11680
Error	168388	.	12605
Total	168401	.	24286

Figura 4-6: *Tabla ANOVA final.*

En los resultados de la *tabla ANOVA final* (Figura 4-6), se puede ver que el modelo integrado por las variables $C4$, $N6$, $N4$, $C1$, $N15$, $N5$, $N2$, $N1$, $N14$, $C7$ y $N12$, presenta una R^2 de 0.4809; es decir, con esas variables se logra explicar el 48.09 % de la variabilidad de la *variable objetivo*.

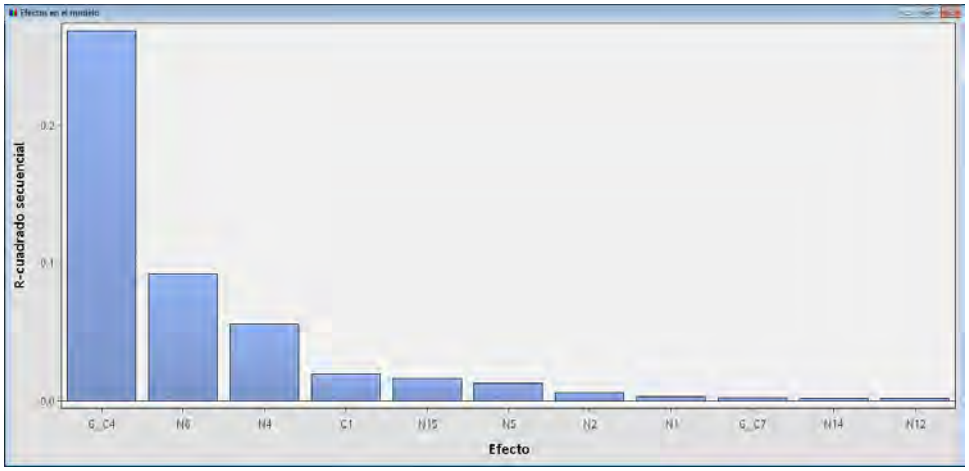

Figura 4-7: Variables con mayor efecto sobre la variable objetivo.

En resumen, las variables $C4$, $N6$, $N4$, $C1$, $N15$, $N5$, $N2$, $N1$, $N14$, $C7$ y $N12$ son elegidas para integrar el modelo, porque tienen un mayor efecto sobre la *variable objetivo* (Figura 4-7). Las demás variables no son relevantes estadísticamente y se excluyen del modelo.

Las variable $C4$ y $C7$ son variables categóricas ordinales y para el análisis de selección de

variables se agruparon con el fin de reducir sus niveles; es por eso que en la Figura 4-7 aparecen con el sufijo «G_».

Una vez que se ha seleccionado el conjunto de variables que mejor discriminan los niveles de la *variable objetivo*, el siguiente paso en el proceso del modelado es la selección de la mejor técnica de minería de datos. Los análisis y resultados correspondientes, se presenta en la sección 4.3.

4.3. Selección de la técnica de minería de datos

Para generar el *score de originación*, se eligieron las siguientes técnicas:

- *Árboles de clasificación*
- *Redes neuronales*
- *Regresión logística*

Estos algoritmos se aplican al conjunto de variables anteriormente seleccionadas. Posteriormente para la comparación de modelos, se aplican diferentes medidas estadísticas para comprobar la eficacia de cada técnica y elegir la más conveniente. En la Figura 4-8 se muestra el flujo que se lleva a cabo para la selección de la técnica más adecuada.

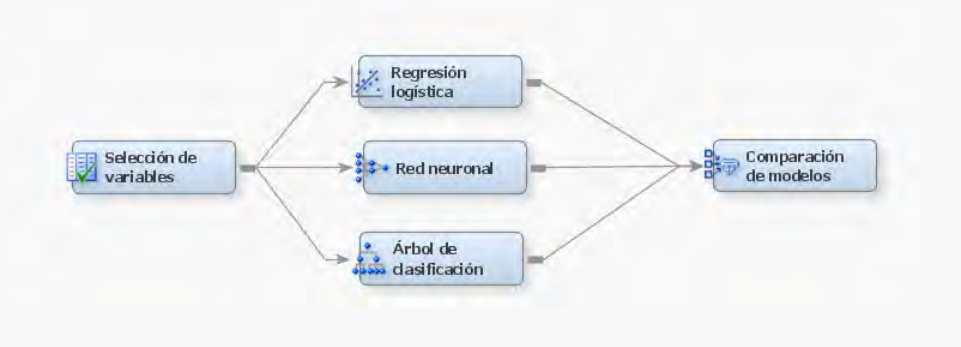

Figura 4-8: Flujo de la selección de la técnica.

4.3.1. Parametrización de las técnicas propuestas

Los nodos para modelar *árboles de clasificación, redes neuronales* y *regresión logística*, se ejecutan para cada uno de los grupos *entrenamiento* con los parámetros que están predefinidos (Figura 4-9, 4-10 y 4-11).

Entrenamiento	
Variables	
Interactivo	
Usar árbol fijo	No
Usar múltiples variables objetivo	No
Precisión	4
Regla de división	
Criterio de intervalo	ProbF
Criterio nominal	ProbChisq
Criterio ordinal	Entropía
Nivel de significación	0.05
Valores ausentes	Utilizar en Búsqueda
Utilizar entrada una vez	No
Ramas máximas	2
Profundidad máxima	6
Tamaño categórico mínimo	5
Precisión de división	4
Nodo	
Tamaño de hoja	5
Número de reglas	5
Número de reglas de sustitución	0
Tamaño de división	.
Búsqueda de división	
Utilizar decisiones	No
Utilizar probabilidades a priori	No
Exhaustivo	5000
Muestra nodo	20000

Subárbol	
Método	Evaluación
Número de hojas	1
Medida de evaluación	Decisión
Fracción de evaluación	0.25
Validación cruzada	
Realizar validación cruzada	No
Número de subconjuntos	10
Número de repeticiones	1
Semilla	12345
Importancia según las observaciones	
Importancia según las observaciones	No
Número de importancia de una sola variable	5
Ajuste de P-valor	
Ajuste Bonferroni	Sí
Tiempo de ajuste Kass	Antes
Entradas	No
Número de entradas	1
Ajuste de división	Sí
Variables de salida	
Variable de hoja	Sí
Rendimiento	Disco

Figura 4-9: Parámetros de la técnica *árboles de clasificación*.

Entrenamiento	
Variables	
Continuar entrenamiento	No
Red	
Optimización	
Semilla de inicialización	12345
Criterio de selección del modelo	Beneficio/Pérdida
Suprimir salida	No

Architecture	Multilayer Perceptron
Direct Connection	No
Number of Hidden Units	3
Randomization Distribution	Normal
Randomization Center	0.0
Randomization Scale	0.1
Input Standardization	Standard Deviation
Hidden Layer Combination Function	Default
Hidden Layer Activation Function	Default
Hidden Bias	Yes
Target Layer Combination Function	Default
Target Layer Activation Function	Default
Target Layer Error Function	Default
Target Bias	Yes
Weight Decay	0.0

Figura 4-10: Parámetros de la técnica *redes neuronales*.

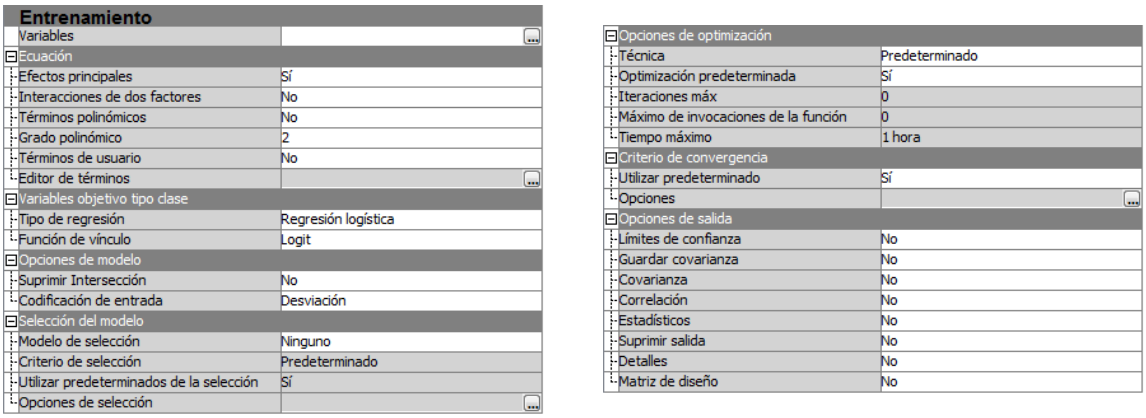

Figura 4-11: Parámetros de la técnica *regresión logística*.

La descripción de las opciones de parametrización que se tienen en cada una de las técnicas de modelado se describen, *grosso modo*, en el Apéndice H, Apéndice I y Apéndice J.

4.3.2. Resultados de la selección de la técnica de modelado

Para evaluar el desempeño de cada una de las técnicas de minería aplicadas, se recurre al estadístico *ROC* (acrónimo de *Receiver Operating Characteristic*) (ver Apéndice L); con este índice, se puede medir la asertividad total del algoritmo aplicado. En general, es una métrica de clasificación muy eficiente para la determinación del modelo óptimo.

Otra métrica que se analiza es la *tasa de clasificación errónea* (ver Apéndice K) que tiene cada modelo aplicado. Con esta tasa se puede apreciar, de forma general, el grado de imperfección que muestra el modelo al clasificar las solicitudes.

También, se analiza el *error tipo 1*; el cual, para este contexto de negocio muestra la deficiencia que tiene el modelo al predecir las solicitudes que están clasificadas con la categoría *Malo*. Cabe mencionar que, el *error tipo 1* es el error que más caro le cuesta a la empresa; por lo que lo ideal sería tener un modelo que minimice este error.

Resultados de la selección de la técnica de minería para el modelo de clientes nuevos

De los resultados que se obtienen en la comparación de modelos, se muestran los más relevantes en la Figura 4-12.

Estadísticos de ajuste
Selección de modelo basada en Entrenar: índice Roc (_AUR_)

Descripción del modelo	Entrenar: índice Roc	Train: Average Squared Error	Train: Misclassification Rate	Valid: Average Squared Error	Valid: Misclassification Rate
Red Neuronal	0.982	0.039368	0.051469	0.038278	0.050375
Árbol clasificación	0.971	0.041013	0.050681	0.040096	0.049424
Regresión logística	0.949	0.067619	0.092028	0.066992	0.091285

Figura 4-12: Estadísticos de ajuste de los modelos evaluados.

En cuanto a los resultados obtenidos en la *curva ROC*, el modelo mejor calificado es la *red neuronal* con una asertividad de 0.982, le sigue el *árbol de clasificación* con un valor de 0.971 y finalmente la *regresión logística* con un valor de 0.94. Estos resultados, también se pueden apreciar en la Figura 4-13, donde se observa que el área bajo la curva más grande, pertenece al modelo de *red neuronal*.

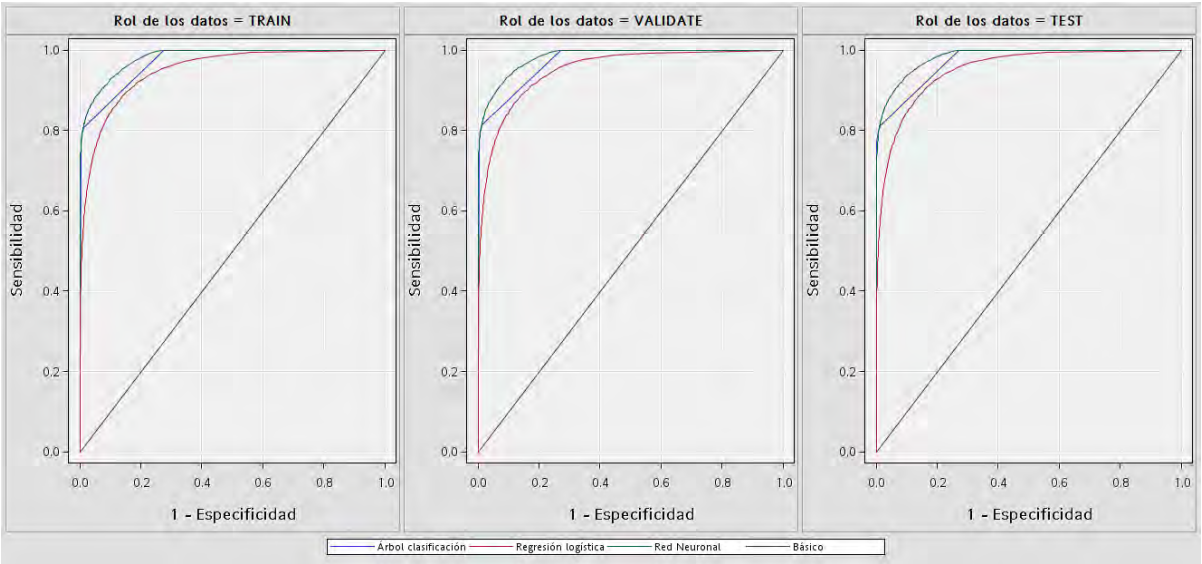

Figura 4-13: Curva ROC de los modelos evaluados.

En cuanto a los resultados obtenidos en la *tasa de clasificación errónea*, se puede apreciar que el modelo que tiene el peor desempeño al clasificar las solicitudes correctamente, es la *regresión logística* con una tasa de error de 9.20 %, le sigue la *red neuronal* con una tasa de 5.14 % y finalmente el *árbol de clasificación* con una tasa de 5.06 %.

A partir de estos resultados, se puede concluir que la técnica de *regresión logística* es la que muestra el peor desempeño en la clasificación de solicitudes; por lo cual, se descarta. Por otra parte, con la métrica *ROC* el modelo mejor calificado es la *red neuronal* pero con la *tasa de clasificación errónea* el mejor calificado es el *árbol de clasificación*.

Entonces, para tomar la decisión de que técnica es la más adecuada se analiza el *error tipo 1* de cada modelo, estos resultados se muestran en la Figura 4-14 y 4-15.

Tabla de clasificación

Rol de los datos=TRAIN Variable objetivo=malo Etiqueta objetivo=' '

Objetivo	Resultado	Porcentaje objetivo	Porcentaje resultado	Número de ocurrencias	Porcentaje total
0	0	95.2726	98.3011	160217	76.9305
1	0	4.7274	17.5590	7950	3.8173
0	1	6.9061	1.6989	2769	1.3296
1	1	93.0939	82.4410	37326	17.9226

Rol de los datos=VALIDATE Variable objetivo=malo Etiqueta objetivo=' '

Objetivo	Resultado	Porcentaje objetivo	Porcentaje resultado	Número de ocurrencias	Porcentaje total
0	0	95.4067	98.2955	53402	76.9260
1	0	4.5933	17.0355	2571	3.7035
0	1	6.8863	1.7045	926	1.3339
1	1	93.1137	82.9645	12521	18.0366

Figura 4-14: Tabla de clasificación de la red neuronal.

En los resultados mostrados en la tabla de clasificación de la *red neuronal* (Figura 4-14), se resalta con rojo el caso que es de mayor interés, el *error tipo 1*. Para el caso del grupo *entrenamiento*, se puede observar que hay 7,950 solicitudes; en las cuales, el modelo predijo incorrectamente la categoría *Malo* y esas solicitudes representan el 4.74 % de 168,167 solicitudes que fueron clasificadas en la categoría *Bueno*. Y representan el 17.55 % de las 45,276 solicitudes que realmente son categoría *Malo*. Los resultados, se interpretan de igual manera para el grupo *validación* y se puede observar que son consistentes. Por otra parte, se puede observar que la tasa de error en la predicción de la categoría *Malo* para el grupo *entrenamiento* es de 3.81 % y en el grupo *validación* es de 3.70 %.

Tabla de clasificación

Rol de los datos=TRAIN Variable objetivo=malo Etiqueta objetivo=' '

Objetivo	Resultado	Porcentaje objetivo	Porcentaje resultado	Número de ocurrencias	Porcentaje total
0	0	94.5166	99.2840	161819	77.6997
1	0	5.4834	20.7350	9388	4.5078
0	1	3.1494	0.7160	1167	0.5604
1	1	96.8506	79.2650	35888	17.2321

Rol de los datos=VALIDATE Variable objetivo=malo Etiqueta objetivo=' '

Objetivo	Resultado	Porcentaje objetivo	Porcentaje resultado	Número de ocurrencias	Porcentaje total
0	0	94.6676	99.2766	53935	77.6937
1	0	5.3324	20.1299	3038	4.3763
0	1	3.1574	0.7234	393	0.5661
1	1	96.8426	79.8701	12054	17.3639

Figura 4-15: Tabla de clasificación del árbol de clasificación.

Interpretando de igual manera los resultados del *error tipo 1* para el *árbol de clasificación* en los grupos *entrenamiento* y *validación* se tiene que, la tasa de error en la predicción de la categoría *Malo* para el grupo *entrenamiento* es de 4.50 % y en el grupo *validación* es de 4.37 % (Figura 4-15).

Comparando estas tasas de error con la del modelo de *red neuronal*, se tiene que la *red neuronal* clasifica mejor la categoría que es de interés para el negocio.

En conclusión, con las métricas analizadas: *curva ROC*, *tasa de clasificación errónea* y *error tipo 1*, se tiene que la técnica de *redes neuronales* es la más adecuada para desarrollar el modelo de *score* para clientes nuevos, debido al desempeño que mostró en la clasificación de solicitudes.

Resultados de la selección de la técnica de minería para el modelo de clientes no nuevos

De los resultados que se obtienen en la comparación de modelos, se muestran los más relevantes en la Figura 4-16.

Estadísticos de ajuste
Selección de modelo basada en Entrenar: índice Roc (_AUR_)

Descripción del modelo	Entrenar: índice Roc	Train: Average Squared Error	Train: Misclassification Rate	Valid: Average Squared Error	Valid: Misclassification Rate
Red Neuronal	0.978	0.039470	0.051674	0.039685	0.052677
Árbol clasificación	0.964	0.038408	0.046116	0.039112	0.047172
Regresión logística	0.950	0.060135	0.081139	0.059966	0.081571

Figura 4-16: Estadísticos de ajuste de los modelos evaluados.

En cuanto a los resultados obtenidos en la *curva ROC*, el modelo mejor calificado es la *red neuronal* con una asertividad de 0.978, le sigue el *árbol de clasificación* con un valor de 0.964 y finalmente la *regresión logística* con un valor de 0.950. Estos resultados, también se pueden apreciar en la Figura 4-17, donde se observa que el área bajo la curva más grande, pertenece al modelo de *red neuronal*.

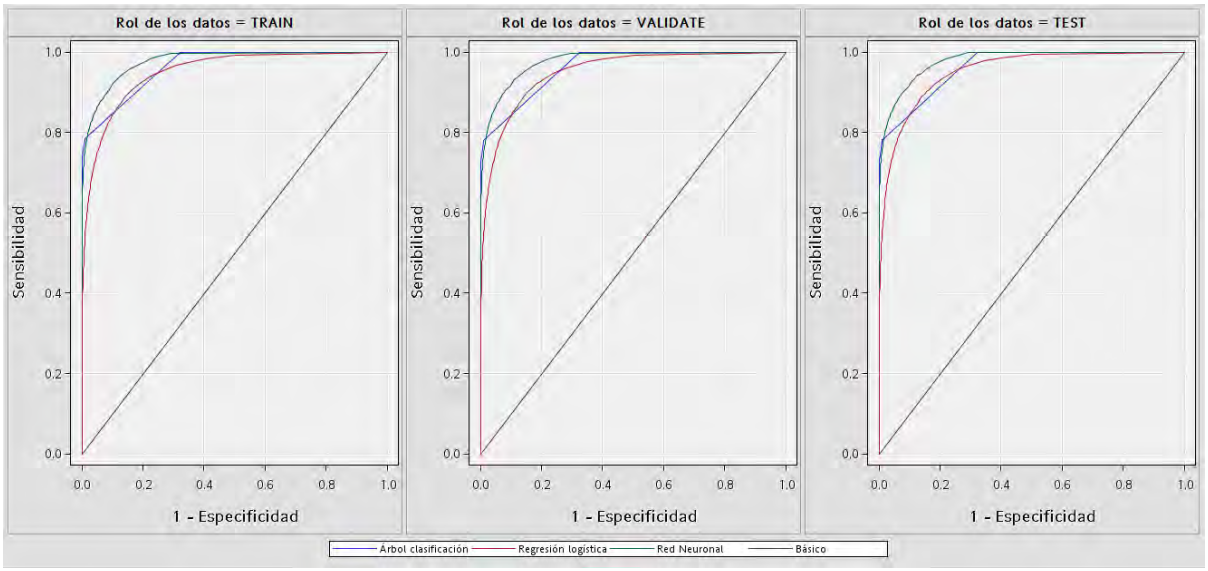

Figura 4-17: Curva ROC de los modelos evaluados.

En cuanto a los resultados obtenidos en la *tasa de clasificación errónea*, se puede apreciar que el modelo que tiene el peor desempeño al clasificar las solicitudes correctamente, es la *regresión logística* con una tasa de error de 8.11 %, le sigue la *red neuronal* con una tasa de 5.67 % y finalmente el *árbol de clasificación* con una tasa de 4.61 %.

Entonces, se concluye que la técnica de *regresión logística* tiene el peor desempeño en la clasificación de solicitudes; por lo cual se descarta. Por otra parte, se observa que la métrica *ROC* califica mejor a la *red neuronal* pero con la *tasa de clasificación errónea* el mejor calificado es el *árbol de clasificación*.

Para decidir que técnica es la más adecuada se analiza el *error tipo 1* de cada modelo, estos resultados se muestran en la Figura 4-18 y 4-19.

Tabla de clasificación

Rol de los datos=TRAIN Variable objetivo=Objetivo Etiqueta objetivo=' '

Objetivo	Resultado	Porcentaje objetivo	Porcentaje resultado	Número de ocurrencias	Porcentaje total
0	0	95.5051	98.3680	136706	81.1784
1	0	4.4949	21.8635	6434	3.8206
0	1	8.9779	1.6320	2268	1.3468
1	1	91.0221	78.1365	22994	13.6542

Rol de los datos=VALIDATE Variable objetivo=Objetivo Etiqueta objetivo=' '

Objetivo	Resultado	Porcentaje objetivo	Porcentaje resultado	Número de ocurrencias	Porcentaje total
0	0	95.3811	98.3810	45575	81.1882
1	0	4.6189	22.4975	2207	3.9316
0	1	8.9788	1.6190	750	1.3361
1	1	91.0212	77.5025	7603	13.5441

Figura 4-18: Tabla de clasificación de la red neuronal.

En los resultados mostrados en la tabla de clasificación de la *red neuronal* (Figura 4-18), se resalta con rojo el caso que es de mayor interés, el *error tipo 1*. Para el caso del grupo *entrenamiento*, se puede observar que hay 6,434 solicitudes; en las cuales, el modelo predijo incorrectamente la categoría *Malo* y esas solicitudes representan el 4.49 % de 143,140 solicitudes que fueron clasificadas en la categoría *Bueno*. Y representan el 21.86 % de las 29,428 solicitudes que realmente son categoría *Malo*.

Los resultados, se interpretan de igual manera para el grupo *validación* y se puede observar que son consistentes. Por otra parte, se puede observar que la tasa de error en la predicción de la categoría *Malo* para el grupo *entrenamiento* es de 3.82 % y en el grupo *validación* es de 3.93 %.

Tabla de clasificación

Rol de los datos=TRAIN Variable objetivo=Objetivo Etiqueta objetivo=' '

Objetivo	Resultado	Porcentaje objetivo	Porcentaje resultado	Número de ocurrencias	Porcentaje total
0	0	95.5615	99.0106	137599	81.7086
1	0	4.4385	21.7174	6391	3.7951
0	1	5.6325	0.9894	1375	0.8165
1	1	94.3675	78.2826	23037	13.6798

Rol de los datos=VALIDATE Variable objetivo=Objetivo Etiqueta objetivo=' '

Objetivo	Resultado	Porcentaje objetivo	Porcentaje resultado	Número de ocurrencias	Porcentaje total
0	0	95.4618	98.9897	45857	81.6906
1	0	4.5382	22.2222	2180	3.8835
0	1	5.7792	1.0103	468	0.8337
1	1	94.2208	77.7778	7630	13.5922

Figura 4-19: Tabla de clasificación del árbol de clasificación.

Interpretando de igual manera los resultados del *error tipo 1* para el *árbol de clasificación* en los grupos *entrenamiento* y *validación* se tiene que, la tasa de error en la predicción de la categoría *Malo* para el grupo *entrenamiento* es de 3.79 % y en el grupo *validación* es de 3.88 % (Figura 4-19).

Comparando estas tasas de error con la del modelo de *red neuronal*, se tiene que el *árbol de clasificación* discrimina mejor la categoría que es de interés para el negocio.

En conclusión, con las métricas analizadas: *curva ROC*, *tasa de clasificación errónea* y *error tipo 1*, se tiene que la técnica *árbol de clasificación* es la más adecuada para desarrollar el modelo de *score* para clientes no nuevos, debido al desempeño que mostró en la clasificación de solicitudes.

4.4. Refinamiento de las técnicas seleccionadas

En esta etapa del modelado, se busca el conjunto de parámetros que optimicen el desempeño de los modelos. Esto se hace cambiando las parametrizaciones de las técnicas seleccionadas y comparando los resultados con los obtenidos en los modelos predefinidos en el *software* que fueron utilizados para la selección de la técnica.

Para seleccionar el conjunto óptimo de parámetros, se cuida que no se altere tanto el *índice de Gini*; el cual, mide la pureza de la clasificación realizada por cada modelo. Y al igual que en la selección de la técnica, se utiliza el *error tipo 1*.

4.4.1. Resultados del modelo final en la población de clientes nuevos

En el caso del modelo para los clientes nuevos, se juega con la composición de la *red neuronal* hasta encontrar el conjunto de parámetros que mejoran la tarea de clasificación. Los parámetros con los que se mejora el desempeño del modelo, se muestran en la Figura 4-20.

Entrenamiento	
Variables	
Continuar entrenamiento	No
Red	
Optimización	
Semilla de inicialización	12345
Criterio de selección del modelo	Beneficio/Pérdida
Suprimir salida	No

Architecture	Multilayer Perceptron
Direct Connection	Yes
Number of Hidden Units	3
Randomization Distribution	Normal
Randomization Center	0.0
Randomization Scale	0.1
Input Standardization	Standard Deviation
Hidden Layer Combination Function	Default
Hidden Layer Activation Function	Default
Hidden Bias	Yes
Target Layer Combination Function	Default
Target Layer Activation Function	Default
Target Layer Error Function	Default
Target Bias	Yes
Weight Decay	0.0

Figura 4-20: Parámetros óptimos de redes neuronal.

Los resultados obtenidos en el modelo con los parámetros predeterminados por el *software* (*red neuronal*) y los resultados obtenidos con el conjunto de parámetros óptimos (*red neuronal afinada*), se muestran a continuación.

Fit Statistics
Model Selection based on Train: Gini Coefficient (_GINI_)

Model Description	Train: Gini Coefficient
Red neuronal afinada	0.970
Red neuronal	0.969

Figura 4-21: Índice de Gini.

En los resultados de la Figura 4-21, se puede observar que con el modelo *red neuronal afinada* se reduce la pureza de la clasificación ya que el el *índice de Gini* es de 0.97 y en el modelo *red neuronal* es 0.969.

Classification Table

Data Role=TRAIN Target Variable=Objetivo Target Label=' '

Target	Outcome	Target Percentage	Outcome Percentage	Frequency Count	Total Percentage
0	0	95.4889	98.6128	160725	77.1744
1	0	4.5111	16.7705	7593	3.6459
0	1	5.6604	1.3872	2261	1.0857
1	1	94.3396	83.2295	37683	18.0940

Data Role=VALIDATE Target Variable=Objetivo Target Label=' '

Target	Outcome	Target Percentage	Outcome Percentage	Frequency Count	Total Percentage
0	0	95.6443	98.6195	53578	77.1795
1	0	4.3557	16.1675	2440	3.5148
0	1	5.5962	1.3805	750	1.0804
1	1	94.4038	83.8325	12652	18.2253

Figura 4-22: Tabla de clasificación del modelo refinado.

Por otra parte, en los resultados de la Figura 4-22, se puede ver que el *error tipo 1* es de 3.64 para el grupo *entrenamiento* y 3.51 para el grupo *validación*. Cabe recordar, que los primeros resultados obtenidos en el *error tipo 1* eran de 3.81 y 3.70 para los grupos *entrenamiento* y *validación* respectivamente (Figura 4-14).

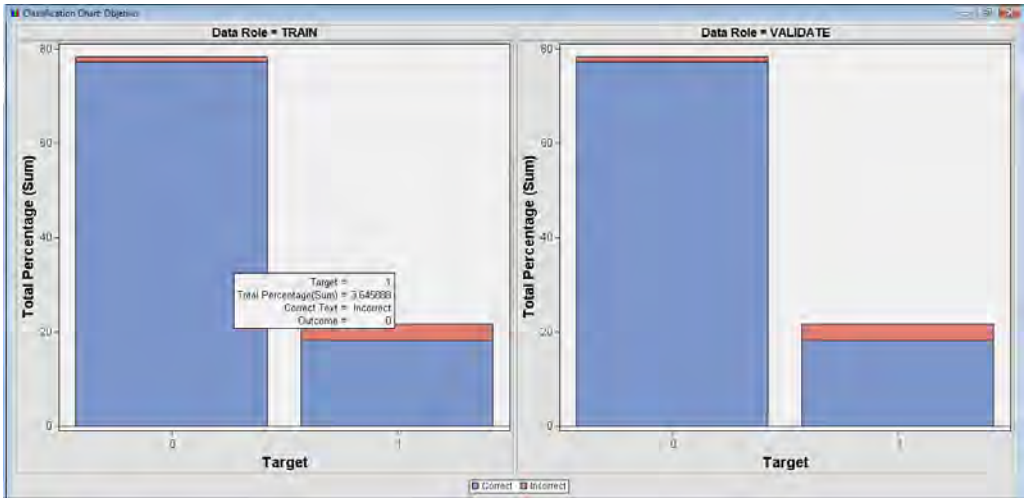

Figura 4-23: Gráficas de clasificación.

En la Figura 4-23, se muestran las gráficas de clasificación en los grupos *entrenamiento* y *validación* obtenidas con el modelo *red neuronal afinada*. En estas gráficas, se puede apreciar que el *error de clasificación* es pequeño, en general. Y dado los resultados anteriores (Figura 4-22 y 4-14), se tiene que se mejoró la precisión de la clasificación en la categoría *Malo*. Por lo tanto, el modelo *red neuronal afinada* se considera que tiene mejor desempeño.

Optimization Start
Parameter Estimates

N	Parameter	Estimate	Gradient Objective Function	N	Parameter	Estimate	Gradient Objective Function
1	N4_H11	0.930788	-0.000059381	16	BIAS_H11	0.351479	-0.000220
2	N5_H11	-0.619029	0.000112	17	BIAS_H12	-2.649087	0.000300
3	N6_H11	0.062995	-0.000010359	18	BIAS_H13	-2.541301	-0.000062595
4	N4_H12	0.179275	0.000213	19	H11_Objetivol	0.773310	-0.000201
5	N5_H12	-2.138045	-0.000167	20	H12_Objetivol	4.151493	-0.000257
6	N6_H12	-0.002465	0.000073669	21	H13_Objetivol	4.283719	-0.000270
7	N4_H13	0.072318	0.000037350	22	N4_Objetivol	-1.586062	0.000114
8	N5_H13	-0.081872	-0.000041897	23	N5_Objetivol	0.752996	-0.000004914
9	N6_H13	-2.652673	0.000107	24	N6_Objetivol	0.098327	0.000110
10	C1H_H11	-1.798775	0.000276	25	C1H_Objetivol	1.250374	-0.000005624
11	G_C40_H11	-0.965654	0.000226	26	G_C40_Objetivol	-5.374574	-0.000012638
12	C1H_H12	-0.057991	-0.000067048	27	BIAS_Objetivol	0.613435	0.000044368
13	G_C40_H12	0.121413	-0.000240				
14	C1H_H13	-0.000504	0.000052656	Value of Objective Function = 0.1148284036			
15	G_C40_H13	0.228461	0.000145				

Figura 4-24: Pesos de los vínculos en el modelo.

En la Figura 4-24, se muestra un listado con los pesos asignados a cada uno de los vínculos que forman la *red neuronal*. Cabe recordar, que los vínculos son relaciones formadas a partir de las variables de entrada, las capas ocultas de la red y la variable **Objetivo**. Con estos pesos, se puede determinar la importancia que tiene cada vínculo en la red, siendo que entre más alejado del cero este el peso del vínculo, mayor será la importancia de este.

Esta importancia que se obtiene con los pesos, nos indica que pareja de elementos son los que tienen mayor efecto sobre la predicción de la *variable objetivo*. Por ejemplo, el vínculo que se forma con la variable de entrada *C4* y la variable **Objetivo** tiene un peso de -5.37 y es el peso más grande en valor absoluto, lo que indica que esta pareja de elementos es la que tiene mayor efecto en los resultados de la clasificación. Por otra parte, la pareja de elementos *C1* y $H13$ es la que tiene el menor efecto en el modelo ya que su peso es de -0.000504.

Figura 4-25: Importancia de los vínculos en el modelo.

En la Figura 4-25, se muestra de forma gráfica la importancia de los vínculos que forman la red. También, se muestra una escala de color que se degrada de azul a rojo. Esta escala indica que los vínculos con pesos más pequeños (color azul) están más relacionados con la categoría *Malo* de la variable **Objetivo**. Y por el contrario, los vínculos con pesos más grandes (color rojo) están más relacionados con la categoría *Bueno*.

4.4.2. Resultados del modelo final en la población de clientes no nuevos

En el caso del modelo para los clientes no nuevos, se juega con la profundidad y el número de ramas que forman el *árbol de clasificación* hasta encontrar el conjunto de parámetros que mejoran la tarea de clasificación. Los parámetros con los que se mejora el desempeño del modelo, se muestran en la Figura 4-26.

Train	
Variables	
Interactive	
Import Tree Model	No
Tree Model Data Set	
Use Frozen Tree	No
Use Multiple Targets	No
Splitting Rule	
Interval Target Criterion	ProbF
Nominal Target Criterion	ProbChisq
Ordinal Target Criterion	Gini
Significance Level	0.2
Missing Values	Most correlated branch
Use Input Once	No
Maximum Branch	2
Maximum Depth	10
Minimum Categorical Size	10
Node	
Leaf Size	10
Number of Rules	5
Number of Surrogate Rules	1
Split Size	.
Split Search	
Use Decisions	No
Use Priors	No
Exhaustive	5000
Node Sample	20000

Subtree	
Method	Assessment
Number of Leaves	1
Assessment Measure	Decision
Assessment Fraction	0.25
Cross Validation	
Perform Cross Validation	No
Number of Subsets	10
Number of Repeats	1
Seed	12345
Observation Based Importanc	
Observation Based Importand	No
Number Single Var Importance	5
P-Value Adjustment	
Bonferroni Adjustment	Yes
Time of Bonferroni Adjustmen	Before
Inputs	No
Number of Inputs	1
Depth Adjustment	Yes
Output Variables	
Leaf Variable	Yes
Performance	Disk

Figura 4-26: Parámetros óptimos del árbol de clasificación.

Los resultados obtenidos en el modelo con los parámetros predeterminados por el *software* (*árbol de clasificación*) y los resultados obtenidos con el conjunto de parámetros óptimos (*árbol de clasificación afinado*), se muestran a continuación.

Fit Statistics
Model Selection based on Train: Gini Coefficient (_GINI_)

Model Description	Train: Gini Coefficient
Árbol de clasificación afinado	0.958
Árbol de clasificación	0.927

Figura 4-27: Índice de Gini.

En los resultados de la Figura 4-27, se puede observar que con el modelo *árbol de clasificación afinado* se reduce la pureza de la clasificación ya que el el *índice de Gini* es de 0.958 y en el modelo *árbol de clasificación* es 0.927.

Classification Table

Data Role=TRAIN Target Variable=Objetivo Target Label=' '

Target	Outcome	Target Percentage	Outcome Percentage	Frequency Count	Total Percentage
0	0	95.8582	99.2718	137962	81.9242
1	0	4.1418	20.2562	5961	3.5397
0	1	4.1342	0.7282	1012	0.6009
1	1	95.8658	79.7438	23467	13.9351

Data Role=VALIDATE Target Variable=Objetivo Target Label=' '

Target	Outcome	Target Percentage	Outcome Percentage	Frequency Count	Total Percentage
0	0	95.7549	99.1365	45925	81.8117
1	0	4.2451	20.7543	2036	3.6270
0	1	4.8936	0.8635	400	0.7126
1	1	95.1064	79.2457	7774	13.8488

Figura 4-28: Tabla de clasificación del modelo refinado.

Por otra parte, en los resultados de la Figura 4-28, se puede ver que el *error tipo 1* es de 3.53 para el grupo *entrenamiento* y 3.62 para el grupo *validación*. Cabe recordar, que los primeros resultados obtenidos en el *error tipo 1* eran de 3.79 y 3.88 para los grupos *entrenamiento* y *validación* respectivamente (Figura 4-15). Por lo tanto, el modelo *árbol de clasificación afinado* se considera que tiene mejor desempeño.

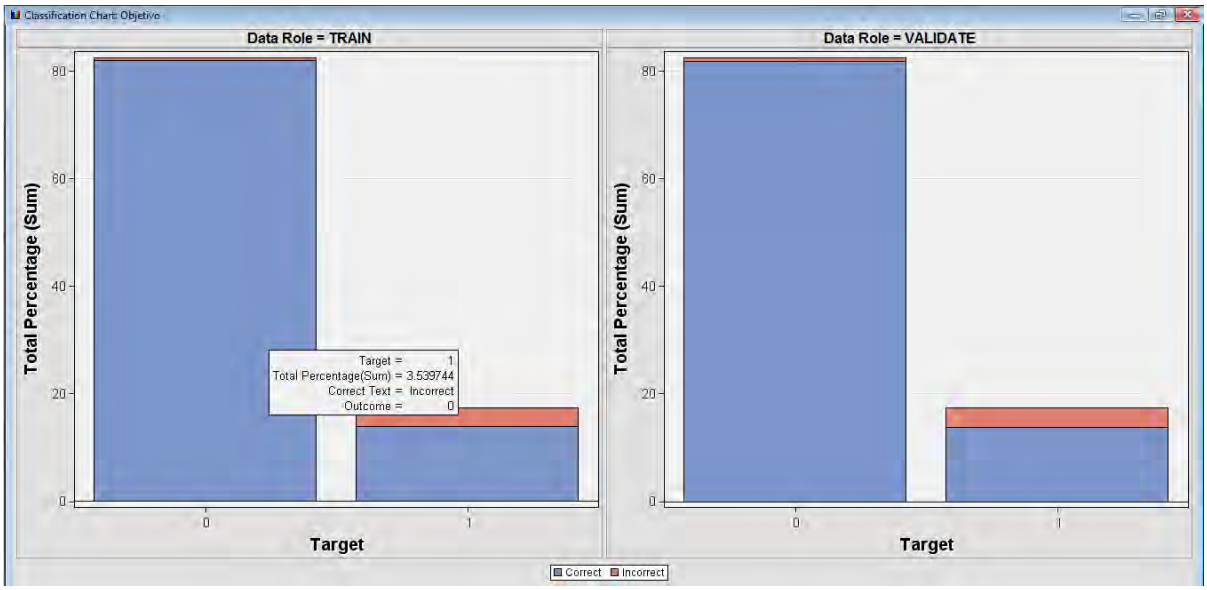

Figura 4-29: Gráficas de clasificación.

En la Figura 4-29, se muestran las gráficas de clasificación en los grupos *entrenamiento* y *validación* obtenidas con el modelo *árbol de clasificación afinado*. En estas gráficas, se puede apreciar que el error de clasificación es pequeño, en general. Y dado los resultados anteriores (Figura 4-28 y 4-15), se tiene que se mejoró la precisión de la clasificación en la categoría *Malo*. Por lo tanto, el modelo *árbol de clasificación afinado* se considera que tiene mejor desempeño.

Variable Importance

Variable Name	Label	Number of Splitting Rules	Number of Surrogate Rules	Importance	Validation Importance	Ratio of Validation to Training Importance
N15	v9 Transformation	2	2	1.0000	1.0000	1.0000
N6	time_home Transformation	3	0	0.9270	0.9236	0.9964
N5	time_job2 Transformation	6	3	0.6061	0.6109	1.0079
N1	loan_amount Transformation	0	4	0.5513	0.5667	1.0280
N4	edad2 Transformation	9	4	0.4845	0.4791	0.9887
N12	v7 Transformation	2	5	0.4815	0.4870	1.0114
G_C4	Grouped Levels for C4	3	0	0.4076	0.3961	0.9718
N2	ingresos Transformation	1	4	0.2199	0.2174	0.9886
N14	v8 Transformation	0	2	0.1913	0.1703	0.8899
C1		3	0	0.1674	0.1624	0.9700

Figura 4-30: Importancia de las variables en el modelo.

En la Figura 4-30, se muestra un listado de las variables que forman al árbol y la importancia

que estas representan en el modelo. Esta importancia sirve para determinar que variables son las que más ayudan a reducir la impureza de la clasificación, siendo que las variables con la importancia más alta reducen el *índice de Gini* y son las que se utilizan primero para comenzar la clasificación.

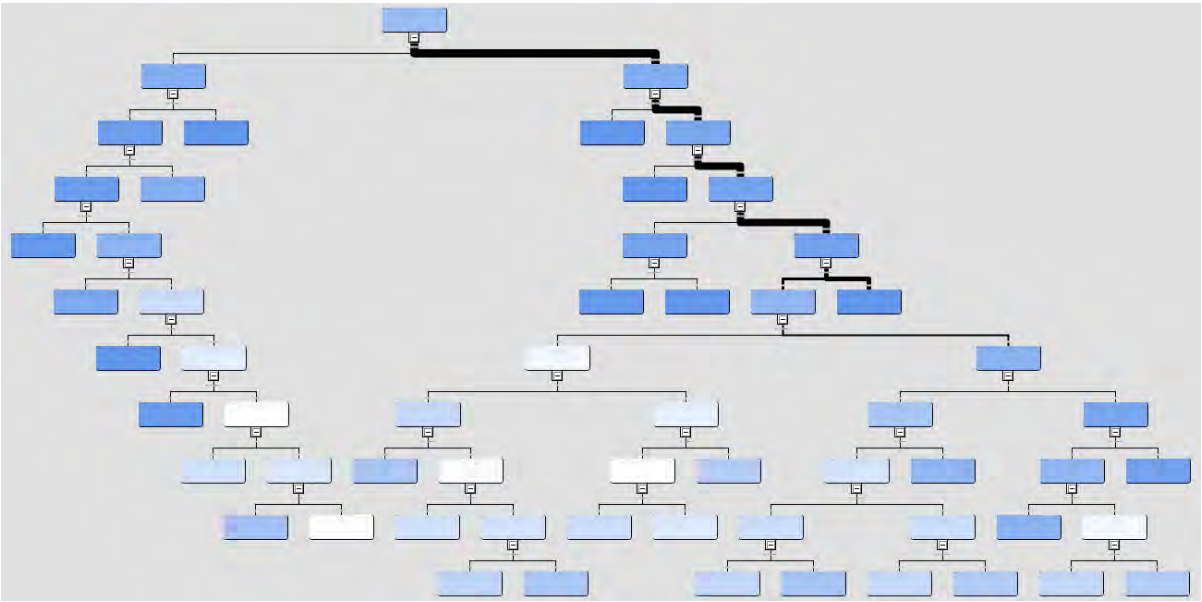

Figura 4-31: Gráfica del árbol final.

En la Figura 4-31, se muestra de forma gráfica el árbol de clasificación. En este diagrama se puede ver que sobresale una linea más gruesa que marca el conjunto de particiones que son más efectivas para clasificar clientes. Estas particiones se hacen con las variables que mostraron una mayor importancia en el modelo.

Por otra parte, en los nodos del árbol se muestra una escala de color azul; la cual, significa que si el color en ese nodo es intenso, se tiene una mayor cantidad de observaciones, mientras que si el azul es tenue se indica que las observaciones que quedaron en ese nodo han sido pocas. También, se puede apreciar que la profundidad del árbol no es muy grande y de esta manera se evita un sobreajuste.

4.5. Conclusiones

En los resultados conseguidos, se puede observar que para los clientes nuevos y no nuevos las técnicas de *árboles de clasificación* y de *redes neuronales* demostraron tener un mejor desempeño en la tarea de clasificar a clientes *Buenos* y *Malos*, esto es debido a que ambas son técnicas más robustas.

En cuanto a las métricas utilizadas, como el estadístico *ROC* y la *tasa de clasificación errónea*, se tiene que son métricas muy eficientes para medir la asertividad de un modelo. Pero es importante tener en cuenta que a los resultados obtenidos hay que darles sentido de negocio para que se forme un mejor criterio al elegir o descartar un modelo. Es por lo anterior, que el *error tipo 1* tiene mucha importancia; anteriormente ya se había dicho que es el error que más caro le cuesta a la empresa.

También, cabe mencionar que para los clientes nuevos las variables que tienen mayor efecto sobre el modelo son:

- **C4**: número de cuentas en las que el cliente tiene un atraso reportado en Buró de Crédito.
- **N5**: los meses de antigüedad en el trabajo que tiene el cliente.
- **N6**: los meses de antigüedad en el hogar que tiene el cliente.

Mientras que para los clientes no nuevos las variables con mayor efecto sobre el modelo son:

- **N5**: los meses de antigüedad en el trabajo que tiene el cliente.
- **N6**: los meses de antigüedad en el hogar que tiene el cliente.
- **N15**: esta variable es conocida en el negocio como *new money* e indica el monto que se le agrega al saldo del cliente al momento de renovar su crédito.

Con lo cual, se puede identificar que tipo de información es la que realmente ayuda a discriminar clientes *Buenos* de clientes *Malos* y sienta las bases para que en la operación se dejen de almacenar algunas otras variables que simplemente no tienen relevancia.

Por otra parte, se tiene que el *score de originación* que se genera sirve para ayudar en la operación del negocio; puesto que es una herramienta que ayuda a la toma de decisiones en el momento en que se presenta un cliente con una solicitud de crédito.

Pero se puede profundizar más en la metodología para la selección de clientes, esto es complementando el *score de crédito* con un estudio en los perfiles de los clientes a los cuales si se les otorga el préstamo. Esto es con la finalidad de saber que monto es el más adecuado para cada cliente y de esta forma mitigar más el riesgo en el negocio.

Apéndice A

Componentes principales

Es una técnica que se emplea para reducir la dimensionalidad de un conjunto de datos en donde hay un gran número de variables interrelacionadas mientras se trata de retener tanto como sea posible la variación del conjunto de variables originales.

Matemáticamente se define como una transformación lineal ortogonal que transforma los datos a un nuevo sistema de coordenadas, llamado *componentes principales*. En esta transformación se tiene que la varianza más grande bajo alguna proyección de los datos recae sobre la primer coordenada conocida como primer componente, la segunda con la varianza más grande recae sobre la segunda componente y así sucesivamente.

Ejemplo: Se tiene un conjunto de datos que contiene información sobre las incidencias en 4 tipos de arrestos que hay en cada estado de los Estados Unidos (Figura A-1).

	Murder	Assault	UrbanPop	Rape
Alabama	13.2	236	58	21.2
Alaska	10.0	263	48	44.5
Arizona	8.1	294	80	31.0
Arkansas	8.8	190	50	19.5
California	9.0	276	91	40.6
Colorado	7.9	204	78	38.7

Figura A-1: Muestra del conjunto de datos de arresto en Estados Unidos.

En este conjunto de datos, se aplica el *análisis de componentes principales*. Del conjunto de nuevas variables, se tiene que la componente 1 y la componente 2 explican el 86.75 % de variabilidad del conjunto de datos. Por lo cual, se puede concluir que estas componentes son suficientes para explicar los datos (Figura A-3).

```
Importance of components:
                          Comp.1    Comp.2    Comp.3     Comp.4
Standard deviation     1.5748783 0.9948694 0.5971291 0.41644938
Proportion of Variance 0.6200604 0.2474413 0.0891408 0.04335752
Cumulative Proportion  0.6200604 0.8675017 0.9566425 1.00000000
```

Figura A-2: *Componentes principales* obtenidas.

Por otra parte, en la Figura A-3, se aprecia la transformación lineal ortogonal que se le aplicó a los datos, visto desde la componente 1 y la componente 2. Además, se aprecia la correlación que guarda cada componente con las variables originales.

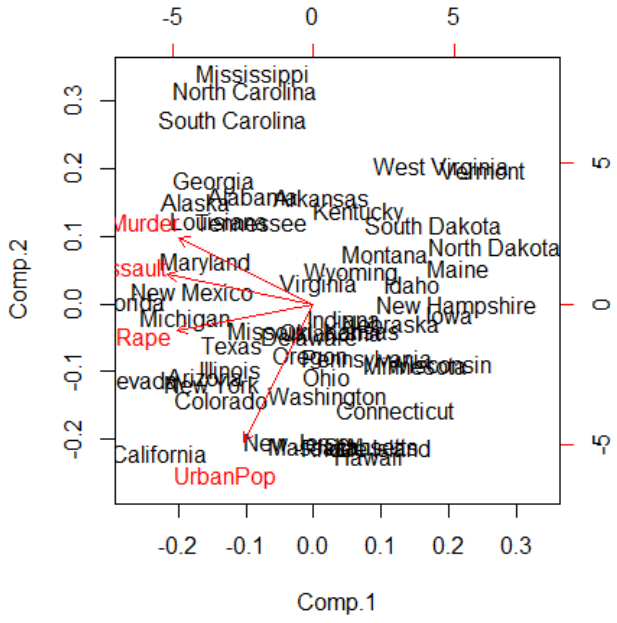

Figura A-3: Rotación de los datos vista desde la componente 1 y 2.

Apéndice B

Transformaciones *Box-Cox*

Son una familia de transformaciones, utilizadas para arreglar problemas de normalidad y heterocedasticidad en las variables.

Esta familia de transformaciones fueron propuestas por los estadísticos George E. P. Box y David Cox. Está definida de la siguiente manera (Ecuación B-1):

$$Y(\lambda) = \begin{cases} \dfrac{y^{\lambda} - 1}{\lambda} & si \quad \lambda \neq 0 \\ \\ log(y) & si \quad \lambda = 0 \end{cases} \tag{B-1}$$

Cabe mencionar que, para transformar una variable Y, sus valores muestrales deben de ser positivos. En caso contrario, se debe sumar una cantidad fija m tal que $Y + m > 0$.

Por otra parte, λ es un parámetro que se tiene que determinar. El método que se utiliza para estimarlo es el de *máxima verosimilitud*; con el cual, se elige $\hat{\lambda}$ tal que maximize la *función de verosimilitud* L(λ).

En la Figura B-1 y FiguraB-2, se muestra un ejemplo de su aplicación.

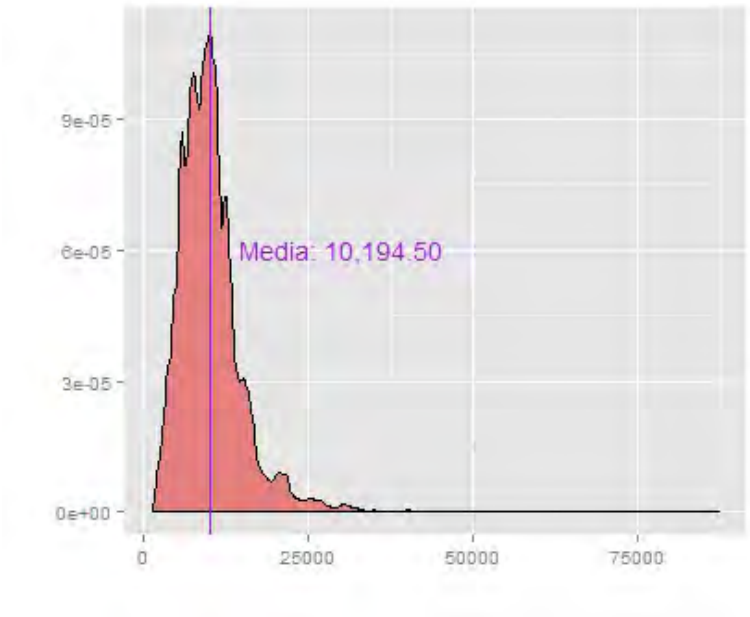

Figura B-1: Variable sin transformación.

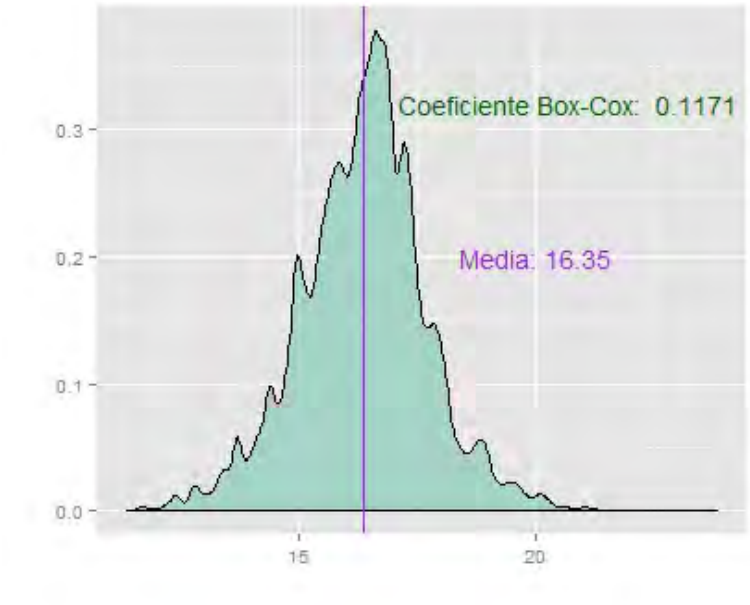

Figura B-2: Variable con transformación *Box-Cox*.

Apéndice C

Índice de *Gini*

Desarrollado por el estadístico *Corrado Gini* en 1912, es una medida de variabilidad para datos categóricos y es interpretado como la probabilidad de que cualesquiera dos elementos de un conjunto, elegidos aleatoriamente (con reemplazo) sean diferentes. La forma de obtenerlo es con la siguiente fórmula:

$$Gini = 1 - \sum_{j=1}^{r} \rho_j^2 \tag{C-1}$$

Donde:

ρ_j es la frecuencia relativa asociada a cada clase del conjunto.

r es el número de elementos que hay en el conjunto.

Para mostrar su uso se muestra el siguiente ejemplo:

Se tiene un conjunto de datos con 698 elementos; los cuales, están clasificados en tres categorías (*c1, c2, c3*) como lo muestra la Tabla C-1.

Clase	N	%
c1	293	41.98
c2	363	52.01
c3	42	6.02
Total	698	100

Tabla C-1: Resumen del conjunto de datos.

Para calcular el índice de *Gini* se sustituye la fórmula C-1, obteniendo lo siguiente:

$$Gini = 1 - \sum_{j=1}^{3} \rho_j^2 = 1 - (0.4198)^2 - (0.5201)^2 - (0.0602)^2 = 0.5497$$

Entonces, del resultado anterior se tiene que es 0.5497 la probabilidad de tomar dos elementos cualesquiera de mi conjunto de datos y que estos sean diferentes.

Apéndice D

Entropía

Es una medida de variabilidad para datos categóricos. Considera r clases mutuamente excluyentes con probabilidades ρ_1, ρ_2, $\cdots$, ρ_r. La singularidad de un resultado en particular puede ser medido como $-log_2(\rho_i)$. La *entropía* es el promedio de estas singularidades calculadas y esta mide la incertidumbre de un resultado. La forma como se calcula es de la siguiente manera:

$$Entropia = -\sum_{j=1}^{r} \rho_j log_2(\rho_j) \tag{D-1}$$

Donde:

ρ_j es la frecuencia relativa asociada a cada clase del conjunto.

r es el número de elementos que hay en el conjunto.

Para mostrar su uso, se utilizan los datos del ejemplo mostrado en el Apéndice C:

Se tiene un conjunto de datos con 698 elementos; los cuales, están clasificados en tres categorías (*c1, c2, c3*) como lo muestra la Tabla D-1.

Clase	N	%
c1	293	41.98
c2	363	52.01
c3	42	6.02
Total	698	100

Tabla D-1: Resumen del conjunto de datos.

Para calcular la *entropía* se sustituye la fórmula D-1, obteniendo lo siguiente:

$$Entropía = -\sum_{j=1}^{3} \rho_j log_2(\rho_j)$$
$$= -0.4198 log_2(0.4198) - 0.5201 log_2(0.5201) - 0.0602 log_2(0.0602)$$
$$= 2.2602$$

Entonces, del resultado anterior se tiene que la *entropía* es 2.2602; la cual, es la medida de incertidumbre de que se obtenga algún resultado en particular.

Esta medida de forma aislada no dice mucho, tendría que ser comparada con otros resultados para poderla interpretar.

Apéndice E

Reducción de variabilidad (impureza)

En las secciones anteriores (Apéndice C y Apéndice D) se definen métricas de impureza para medir la variabilidad que hay dentro de un conjunto de datos. En el contexto de la técnica de *árboles de decisión*, se utilizan para comparar las particiones generadas y elegir la mejor de ellas; entendiendo que la mejor partición es la que reduzca la *impureza* de un *nodo padre* al ser segmentado en m *nodos hijos*.

Esta medida de *reducción de impureza* (Δi), está dada por la fórmula siguiente:

$$\Delta i = i(0) - \left(\frac{n_1}{n_0} i(1) + \frac{n_2}{n_0} i(2) + \cdots + \frac{n_m}{n_0} i(m) \right) \tag{E-1}$$

Donde:

$n(0)$ es la cantidad de elementos en el nodo padre.

$i(j)$ es la impureza en el nodo padre.

$n(j)$ es la cantidad de elementos en el j-ésimo nodo hijo, con $j \in 1, 2, ..., m$.

$i(j)$ es la impureza en el j-ésimo nodo hijo, con $j \in 1, 2, ..., m$. La impureza se calcula con el *índice de Gini* o la *Entropía*

Para su entendimiento se propone el siguiente ejemplo:

Se tiene un nodo padre con 1,064 elementos; los cuales, están clasificados en tres categorías (*c1, c2, c3*). Dado un criterio de partición, se pretende reducir la impureza en el nodo padre segmentándolo en dos nodos hijos, como lo muestra la Tabla E-1.

Clase	**N**$nodo_0$	**N**$nodo_1$	**N**$nodo_2$
c1	364	293	71
c2	364	363	1
c3	336	42	294
Total	1,064	698	366

Tabla E-1: Resumen de la partición.

Utilizando el *índice de Gini* para medir la impureza se tiene que el $nodo_0$ tiene una impureza de 0.6662, mientras que el $nodo_1$ y el $nodo_2$ tiene 0.5497 y 0.3171 respectivamente.

Sustituyendo en la fórmula de *reducción de impureza*, se tiene:

$$\Delta i = 0.6662 - \left(\frac{698}{1,064} 0.5497 + \frac{698}{1,064} 0.3171 \right) = 0.1965$$

Por otro lado, utilizando el *entropía* para medir la impureza se tiene:

$$\Delta i = 2.5839 - \left(\frac{698}{1,064} 2.2602 + \frac{698}{1,064} 1.7360 \right) = 0.5040$$

Cabe mencionar que para los *árboles de clasificación* con particiones binarias, *Breiman* (1996) mostró que la *reducción de impureza* utilizando el criterio de *Gini* tiende a favorecer el aislamiento de la clase más grande en una rama; mientras que la *reducción de impureza* utilizando el criterio de *Entropía* tiende a favorecer las divisiones balanceadas.

Apéndice F

Prueba χ^2 *Pearson*

Otra forma de juzgar el valor de las particiones en los *árboles de clasificación* es utilizando la prueba χ^2 de *Pearson*; la cual, se utiliza para hacer pruebas de *bondad de ajuste*.

En el caso de los *árboles de clasificación*, el estadístico de la prueba se usa para medir la diferencia entre las proporciones de las clases que hay en cada nodo, tomando las ramas y las clases objetivo como si fueran independientes. En la mayoría de las pruebas el *p-valor* es muy pequeño; por lo cual, se utiliza la transformación $logworth = -log_{10}(p - valor)$. Y lo que se busca es que el valor *logworth* sea el más pequeño posible.

De esta forma se puede ver las particiones en el árbol como si fueran *tablas de contingencia*, donde las columnas representan los nodos hijo y los renglones las clases. De esta manera cada celda contiene la frecuencia de elementos que pertenezcan a cierto nodo y cierta clase.

Utilizando el ejemplo del Apéndice E, se obtienen las siguientes tablas F-1 y F-2:

Clase	N$nodo_1$	N$nodo_2$
c1	293	71
c2	363	1
c3	42	294

Tabla F-1: Valores observados (O).

Clase	N$nodo_1$	N$nodo_2$
c1	293	71
c2	363	1
c3	42	294

Tabla F-2: Valores esperados (E).

Y a partir de los valores *observados* y *esperados* se calcula la tabla de la prueba estadística (Tabla F-3):

Clase	N$nodo_1$	N$nodo_2$
c1	12	23
c2	65	123
c3	144	275

Tabla F-3: Tabla de la prueba estadística $\frac{(O-E)^2}{E}$.

Sumando todas las celdas de la Tabla F-3 se obtiene el valor 643; el cual, pertenece a una distribución χ^2 con 2 grados de libertad. El $p-valor$ obtenido es muy cercano a cero, entonces se le aplica la transformación $logworth = -log_{10}(p-valor)$. Con esta transformación se obtiene un valor de 139.725

El valor $logworth = 139.725$ por si solo no dice nada, se tiene que comparar con el valor *logworth* de otra partición generada para que pueda ser útil. La mejor partición será la que tenga el valor *logworth* más pequeño.

Apéndice G

Nodo: «Selección de variables»

En este apéndice se describen, *grosso modo*, los parámetros con los que cuenta el nodo «Selección de variables». Esta información fue consultada en la referencia de SAS® Enterprise Miner™ 12.1.

Parámetros generales del entrenamiento:

- **Variables:** permite al usuario seleccionar las variables a evaluar.
- **Nivel de clase máxima:** define el número máximo de clases que se permiten en una variable, si la variable excede este parámetro se rechaza.
- **Porcentaje de ausentes máximo:** define el porcentaje máximo de valores ausentes que puede tener una variable, si la variable excede este parámetro se rechaza.
- **Modelo objetivo:** tipo de criterio utilizado para la selección de variables Las opciones son $Ji-cuadrada$, $R-cuadrada$ y el predeterminado.
- **Selector manual:** permite la selección y rechazo de las variables de forma manual.

Opciones de bypass:

- **Variable:** permite seleccionar el tipo de variables (numéricas o categóricas) para asignarles un rol.
- **Rol:** rol que se le asigna a algún tipo de variable, puede ser entrada o rechazado.

Opciones de Ji-cuadrada:

- **Número de clases:** define el número de clases en las que puede ser cortada una variable numérica.

- **Número máximo de pasadas:** define la cota máxima de iteraciones hechas para encontrar el número óptimo de cortes en la variable.
- **Chi-cuadrado mínimo:** define la cota mínima para el valor de la $Ji-cuadrada$; con el cual, se permite que una variable sea elegible.

Opciones de R-cuadrada:

- **Número máximo de variables:** define la cota máxima del número de variables que pueden ser seleccionadas para el modelo.
- **R-cuadrado mínimo:** define la cota mínima del valor de la $R-cuadrada$; con el cual, se permite que una variable sea elegible.
- **Detener R-cuadrado:** define el valor incremental de la $R-cuadrada$; si se alcanza este valor, el proceso de selección se detiene.
- **Utilizar variables AOV16:** permite indicar si se quiere cortar las variables numéricas en 16 intervalos igualmente espaciados; con la finalidad de ayudar a identificar relaciones no lineales con la variable dependiente.
- **Utilizar variables de grupo:** permite indicar si se quiere reducir los niveles de las variables categóricas en menos grupos.
- **Utilizar interacciones:** permite indicar si se quiere la medición del efecto de clasificación que tiene una variable a través de los niveles de otra.
- **Usar la librería SPD Engine:** permite habilitar la opción de multiproceso.
- **Opción imprimir:** permite seleccionar o suprimir los detalles que se quieren ver en los resultados de la selección.

Apéndice H

Nodo: «Árboles de clasificación»

En este apéndice se describen, *grosso modo*, los parámetros con los que cuenta el nodo «Árboles de clasificación». Esta información fue consultada en la referencia de SAS® Enterprise Miner™ 12.1.

Parámetros generales del entrenamiento en árboles de clasificación:

- **Variables:** permite al usuario visualizar y definir el rol de las variables (*entrada*, *rechazo* y *objetivo*).
- **Interactivo:** permite al usuario desplegar una ventana interactiva de entrenamiento.
- **Usar árbol fijo:** especifica si una definición de árbol fijo se debería usar o si se debería crear un nuevo árbol durante el entrenamiento.
- **Usar múltiples variables objetivo:** especifica si se deberían estar disponibles múltiples variables objetivos durante el entrenamiento.
- **Precisión:** especifica el número de decimales.

Regla de división

- **Criterio de intervalo:** especifica el método para evaluar y buscar reglas candidatas de corte en presencia de una variable objetivo numérica.
- **Criterio nominal:** especifica el método para evaluar y buscar reglas candidatas de corte en presencia de una variable objetivo categórica nominal.
- **Criterio ordinal:** especifica el método para evaluar y buscar reglas candidatas de corte en presencia de una variable objetivo ordinal.

- **Nivel de significación:** especifica el $p - valor$ aceptable máximo para el valor de una regla de corte candidata cuando se elige el criterio $ProbChisq$ o $ProbF$.
- **Valores ausentes:** especifica cómo una regla de división trata una observación con valores ausentes. Los valores ausentes se utilizan como un valor durante la búsqueda de división o bien se asignan a un nodo basado en esta selección.
- **Utilizar entrada una vez:** especifica que ninguna regla de división se va a basar en una variable de entrada utilizada en una regla de división de un nodo antecesor.
- **Ramas máximas:** restringe el número de subconjuntos que una regla de división puede crear en el número especificado o menos.
- **Profundidad máxima:** especifica el número máximo de generaciones de nodos. Al nodo original, generación 0, se le llama nodo raíz. Los hijos del nodo raíz son de primera generación.
- **Tamaño categórico mínimo:** número mínimo de observaciones para un valor categórico. Una categoría debe aparecer en al menos el número de observaciones especificado para utilizarlo en la búsqueda de división.
- **Precisión de división:** especifica el número de decimales que se van a visualizar en los valores de división y en valores de promedio mostrados en los nodos del árbol.

Nodo

- **Tamaño de hoja:** especifica en número más pequeño de observaciones de entrenamiento que puede tener una hoja.
- **Número de reglas:** especifica cuántas reglas de división se guardaren cada nodo, El árbol solo utiliza una regla. El resto se guarda para compararlas.
- **Número de reglas de sustitución:** especifica el número máximo de reglas sustitutas buscadas en cada nodo no hoja. Una regla sustituta es un respaldo de la regla de división principal.
- **Tamaño de división:** especifica el número más pequeño de observaciones de entrenamiento que un nodo puede tener para que el procedimiento considere dividirlo.

Búsqueda de división

- **Utilizar decisiones:** especifica si se va a usar información de decisión durante la búsqueda de división.
- **Utilizar probabilidades a priori:** especifica si se van a usar probabilidades a priori durante la búsqueda de división.
- **Exhaustivo:** especifica el mayor número de divisiones candidatas que se van a encontrar en una búsqueda exhaustiva.

- **Muestra nodo:** especifica el tamaño muestral del nodo para buscar divisiones.

Subárbol

- **Método:** especifica cómo se va a construir el subárbol en términos de métodos de selección. Son posibles los siguiente métodos: *Evaluación* (se elige el subárbol más pequeño con el mejor valor de evaluación), $Mayor$ (selecciona el árbol completo) y N (selecciona el subárbol mayor con a lo más n hojas).
- **Número de hojas:** especifica el número de hojas que se utilizan al crear el subárbol cuando la opción subárbol es N.
- **Medida de evaluación:** especifica el método que se desea utilizar para seleccionar el mejor árbol, basándose en los datos de validación. Si no hay datos de validación disponibles, se utilizarán los datos de entrenamiento. Las medidas de evaluación disponibles son *Decisión*, *Error cuadrático de la media*, *Error de clasificación* y $Mejora$.
- **Fracción de evaluación:** especifica la fracción de observaciones que se van a utilizar cuando la propiedad $Medida$ sea $Mejora$.
- **Realizar validación cruzada:** especifica si desea realizar validación cruzada para cada subárbol de la secuencia.
- **Número de subconjuntos:** especifica el número de subconjuntos de validación cruzada.
- **Número de repeticiones:** especifica el número de veces que se repite la validación cruzada.
- **Semilla:** especifica el número aleatorio de semillas para generar los subconjuntos de validación.

Importancia según las observaciones

- **Importancia según las observaciones:** especifica si desea calcular las estadísticos de importancia basados en las observaciones.
- **Número de importancia de una sola variable:** especifica el número de variables que los estadísticos de importancia de una dirección debería generar.

Importancia según las observaciones

- **Ajuste Bonferroni:** solicita que se aplique un ajuste $Bonferroni$ a los $p - valores$.
- **Tiempo de ajuste Kass:** indica si el ajuste $Bonferroni$ se va a realizar antes o después de elegir la división.
- **Entradas:** solicita que se aplique un ajuste para el número de entradas a los $p - valores$.
- **Número de entradas:** especifica el número de entradas que se van a considerar no correlacionadas cuando la opción $Entradas$ es igual a Si.

- **Ajuste de división:** solicita que se aplique un ajuste para el número de divisiones de antecesores a los $p - valores$.

Variables de salida

- **Variable de hoja:** indica si se crea una variable tipo *id numérica* a la hoja; a la cual, se le asigna la observación.
- **Rendimiento:** rendimiento: especifica dónde se va a colocar la copia de los datos de entrenamiento (RAM o $Disk$).

Apéndice I

Nodo: «Redes neuronales»

En este apéndice se describen, *grosso modo*, los parámetros con los que cuenta el nodo «Redes neuronales». Esta información fue consultada en la referencia de SAS® Enterprise Miner™ 12.1.

Parámetros generales del entrenamiento en redes neuronales:

- **Variables:** permite al usuario visualizar y definir el rol de las variables (*entrada*, *rechazo* y *objetivo*).
- **Continuar entrenamiento:** especifica si las estimaciones actuales deberían utilizarse como valores de inicio para el entrenamiento.
- **Red:** abre un diálogo para personalizar las opciones de red.
- **Optimización:** abre un diálogo para personalizar las opciones de optimización.
- **Semilla de inicialización:** especifica el valor de la semilla aleatoria utilizada en la inicialización de ponderación de la red.
- **Criterio de selección del modelo:** especifica el criterio de selección del modelo (*Error de promedio*, *Error de clasificación* y *Beneficio/Pérdida*).
- **Suprimir salida:** suprime todas las salidas impresas del proceso.

Parámetros de la arquitectura de redes neuronales:

- **Arquitectura:** especifica la arquitectura de red que se utiliza en la construcción de la red. Las opciones son: modelo generalizado lineal, perceptrón multicapa, función de base radial ordinaria con anchos iguales, la función de base radial ordinaria con anchuras desiguales, normaliza la función de base radial con la misma

altura, normaliza la función de base radial con volúmenes iguales, normaliza la función de base radial con anchuras iguales, normaliza la función de base radial con igualdad de anchuras y alturas, normaliza la función de base radial con anchuras y alturas desiguales y una red de usuario especificado.

- **Conexión directa:** indica si debería haber una conexión directa de unidades de entrada con unidades de salida.
- **Número de unidades ocultas:** especifica el número de unidades ocultas en la capa oculta. Se permite seleccionar desde 1 hasta 64 unidades. El valor por defecto es 3.
- **Aleatoriedad de la distribución:** especifica el tipo de distribución que se le aplicará a los pesos.
- **Centro de aleatoriedad:** especifica el centro de la distribución de los pesos. Los valores permitidos son números reales. Por defecto es 0.
- **Escala de aleatoriedad:** especifica la escala de la distribución de los pesos. Los valores permitidos son números reales. El valor por defecto es 0.1.
- **Estandarización de entradas:** especifica el método que es utilizado para estandarizar las variables de entrada. Por defecto se elige la desviación estándar.
- **Función de combinación en capas ocultas:** especifica la función de combinación que se utiliza en la capa oculta.
- **Función de activación en capas ocultas:** especifica la función de activación que se utiliza en la capa oculta.
- **Sesgo oculto:** especifica si se deben utilizar capas ocultas con sesgo o sin sesgo.
- **Funciones de combinación en la capa objetivo:** especifica la función de combinación de la capa objetivo.
- **Función de activación en la capa objetivo:** especifica la función de activación de la capa objetivo.
- **Función error en la capa objetivo:** especifica la función de error de la capa objetivo.
- **Sesgo en la capa objetivo:** especifica si se deben utilizar sesgo en las copas objetivo.
- **Decadencia del peso:** especifica la decadencia del peso.

Apéndice J

Nodo: «Regresión logística»

En este apéndice se describen, *grosso modo*, los parámetros con los que cuenta el nodo «Regresión logística». Esta información fue consultada en la referencia de SAS® Enterprise Miner™ 12.1.

Parámetros generales del entrenamiento en regresión logística:

- **Variables:** permite al usuario visualizar y definir el rol de las variables (*entrada*, *rechazo* y *objetivo*).

Ecuación

- **Efectos principales:** incluye en el modelo todas las variables seleccionadas para utilizar.
- **Interacciones de los factores:** incluye en el modelo todas las interacciones de dos factores para las variables de clase seleccionadas para utilizar.
- **Términos polinómicos:** incluye en el modelo términos polinómicos hasta el nivel especificado para todas las variables de intervalo que se van a utilizar.
- **Grado polinómico:** especifica el grado polinómico cuando los términos polinómicos se incluyen en el modelo.
- **Términos de usuario:** incluye términos del usuario especificados en la opción *Términos*.
- **Editor de términos:** el editor de términos permite personalizar el modelo, especificando los términos de interacción y ordenando los términos de modelo.

Variables objetivo tipo clase

- **Tipo de regresión:** especifica el tipo de regresión que tiene que ejecutarse (*lineal* o *logística*). El tipo predeterminado para variables objetivo binarias es regresión es logística.
- **Función de vínculo:** para la regresión logística, se pueden seleccionar las siguientes funciones de vínculo: *Logit*, *Cloglog (log-log complementario)* y *Probit*. Para la regresión lineal, se utiliza la función de vínculo *Identidad*.

Opciones de modelo

- **Suprimir intersección:** indica si debería suprimirse la intersección. Se ignora el atributo para variables objetivo ordinales.
- **Codificación de entrada:** especifica la codificación que se utiliza para una variable de clase.

Selección del modelo

- **Modelo de selección:** especifica un método de selección de modelo. *Atrás:* el entrenamiento comienza con todos los efectos candidatos en el modelo y elimina los efectos hasta que se cumple el nivel de significación *permanecer* o *el criterio de parada*. *Hacia delante:* el entrenamiento comienza sin efectos candidato en el modelo y añade efectos hasta que se cumple el nivel de significación *entrada* o *el criterio de parada*. *Paso a paso:* el entrenamiento comienza como en el modelo *hacia delante*, pero puede eliminar efectos que ya estén en el modelo. Continúa así hasta que se cumple el nivel de significación *permanecer* o *el criterio de parada*. *Ninguno*: todas las entradas se utilizan para ajustar el modelo.
- **Criterio de selección:** especifica el criterio de selección: (*criterio de información de Akaike*, *criterio bayesiano de Schwarz*, *error de validación*, etc.)
- **Utilizar predeterminados de la selección:** especifica si se tiene que utilizar los valores predeterminados para la técnica de selección de modelos.
- **Opciones de selección:** abre un diálogo para personalizar las opciones de selección.

Opciones de optimización

- **Técnica:** especifica el método de optimización que se va a utilizar al ajustar el modelo logístico.
- **Optimización predeterminada:** especifica el uso de la optimización predeterminada del modelo.
- **Interacciones máximas:** número máximo de iteraciones.
- **Máximo de invocaciones de la función:** número máximo de invocaciones de función.
- **Tiempo máximo:** especifica el límite superior de tiempo *CPU* para el proceso de optimización.

Criterio de convergencia

- **Utilizar predeterminado:** utiliza opciones de criterios de convergencia predeterminado.

- **Opciones:** abre un diálogo para personalizar las opciones de convergencia.

Límites de confianza

- **Límites de confianza:** genera límites de confianza para las estimaciones de parámetro.
- **Guardar covarianza:** guarda la matriz de covarianza de las estimaciones de parámetro.
- **Covarianza:** imprime la matriz de covarianza de las estimaciones de parámetro.
- **Correlación:** imprime la matriz de correlación de las estimaciones de parámetro.
- **Estadísticos:** imprime los estadísticos descriptivos simples de las variables de entrada.
- **Suprimir salida:** suprime todas las salidas impresas.
- **Detalles:** imprime los detalles de cada paso de la selección del modelo.
- **Matriz de diseño:** imprime la matriz de diseño: entrada codificada para las variables de clase.

Apéndice K

Tasa de clasificación errónea

Se utiliza para conocer el grado de error que presenta alguna métrica al predecir el valor de la variable objetivo y se calcula a partir de la *matriz de confusión o tabla de contingencia*. En la Figura K-1, se muestra una *matriz de confusión* con una variable objetivo binaria.

Población total		Condición real	
		Condición positiva	Condición negativa
Predicción	Predicción positiva	Verdaderos positivos	Falsos positivos (Error tipo 1)
	Predicción negativa	Falsos negativos (Error tipo 2)	Verdaderos negativos

Figura K-1: Matriz de confusión.

En la *matriz de confusión*, se toman todas las observaciones que se tienen y se compara su valor real de la variable objetivo contra el valor predicho por la métrica evaluada. La finalidad es conocer que proporción de las observaciones fueron predichas correctamente (*verdaderos positivos* y *verdaderos negativos*) y en que proporción de observaciones se erró la predicción (*error tipo 1* y *error tipo 2*).

El cálculo de las proporciones de *verdaderos positivos*, *verdaderos negativos*, *error tipo 1* y *error tipo 2* se muestran en la Figura K-2. También, se muestra la precisión total de de la métrica evaluada; la cual, es conocida como *tasa de clasificación errónea*.

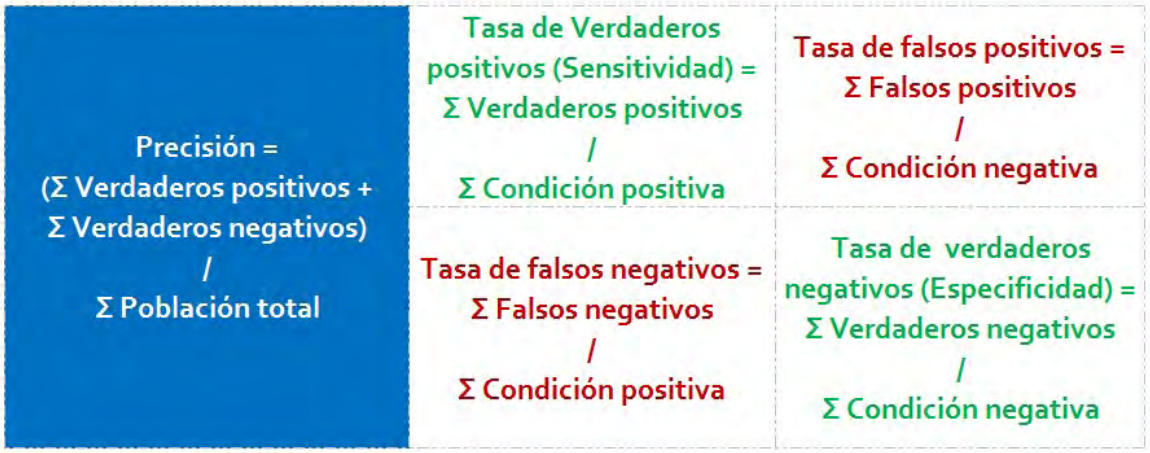

Figura K-2: Tasa de clasificación errónea.

Apéndice L

Curva ROC

La curva *ROC* o *Receiver Operating Characteristic*, es una gráfica que muestra el desempeño de un sistema de clasificación binario cuando su umbral es variado.

Para la construcción de esta gráfica, primero se define un *umbral* en la métrica evaluada (figura L-1) y se calculan los valores $1 - Especificidad$ y *Sensitividad* (definidas en el apéndice **??**, Figura K-2).

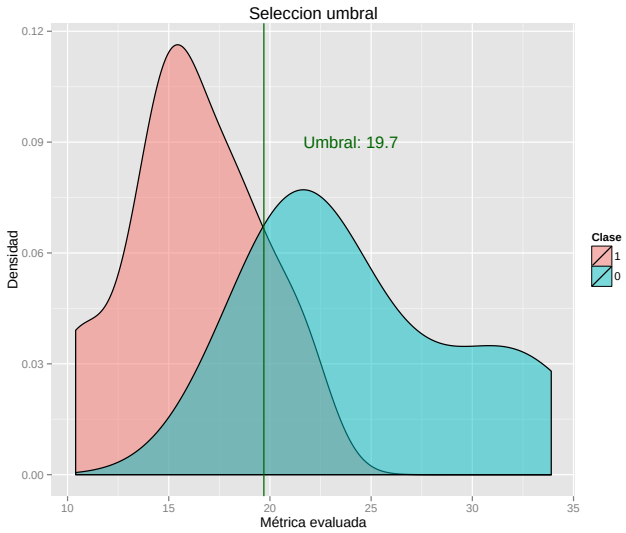

Figura L-1: Selección de umbral de un sistema de clasificación.

Este proceso se repite cambiando el valor del umbral para obtener un conjunto de coordenadas (Tabla L-1) como si se estuviera definiendo $Sensitividad$ en función de $1 - Especificidad$.

$Umbral$	$1 - Especificidad$	$Sensitividad$
u_1	x_1	y_1
u_2	x_2	y_2
$\ldots$	$\ldots$	$\ldots$
u_{n-1}	x_{n-1}	y_{n-1}
u_n	x_n	y_n

Tabla L-1: Conjunto de coordenadas de la curva ROC.

Con estas coordenadas se genera el gráfico ROC (Figura L-2).

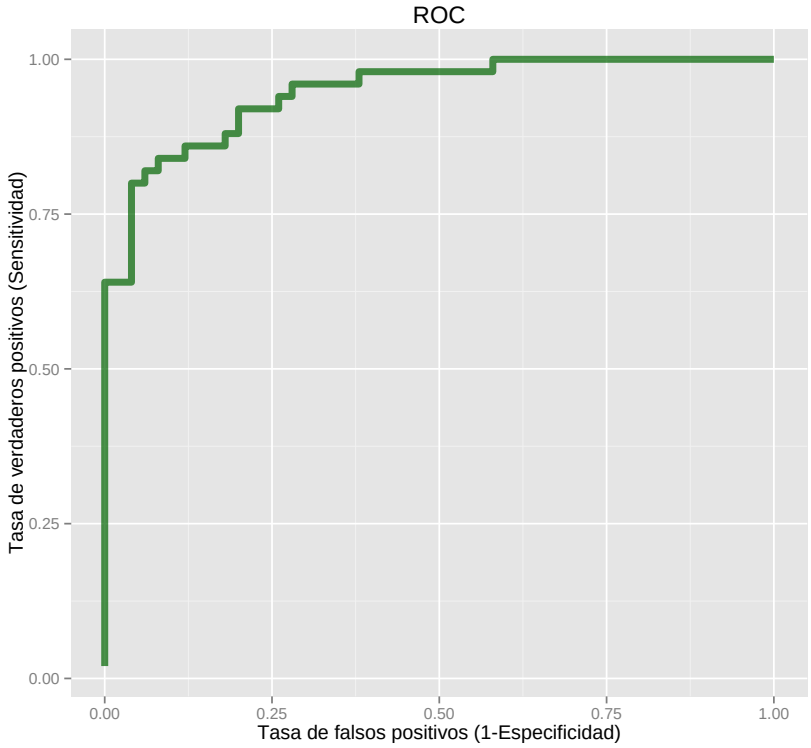

Figura L-2: Curva ROC.

Finalmente, para darle interpretación a la curva ROC, se calcula el área bajo la curva. Esta área calculada se interpreta como la precisión que tiene el clasificador evaluado, siendo que un área de 1 representa una precisión perfecta y un área de 0.5 representa una precisión deficiente.

Para entender de forma más fácil el valor del área bajo la curva, se puede utilizar la Tabla L-2; la cual, define una interpretación en función al intervalo donde se encuentre el valor.

Intervalo	Interpretación
$[0.5, 0.6)$	Precisión mala
$[0.6, 0.75)$	Precisión regular
$[0.75, 0.9)$	Precisión buena
$[0.9, 0.97)$	Precisión muy buena
$[0.97, 1)$	Precisión excelente

Tabla L-2: Interpretacióndel área bajo la curva *ROC*.