TESINA:

CONSTRUCCIÓN DE UN MODELO PREDICTIVO DE RIESGO-RENTABILIDAD PARA UNA ENTIDAD BANCARIA

MBA – UNIVERSIDAD DEL CEMA

PROFESORES-TUTORES:

PABLO RAIES

DIEGO SODOR

AUTOR:

MARIO MARTÍN PALLARES

SEGUNDO TRIMESTRE DE 2009

Índice

INTRODUCCIÓN **3**

TERMINOLOGÍA BÁSICA PARA COMPRENDER LA TESINA **5**

CARACTERÍSTICA 5
MODELO 5
RIESGO 5

OBJETIVO DE LA TESINA **7**

CLASIFICACIÓN DE LOS MODELOS DE CALIFICACIÓN **8**

EN FUNCIÓN DEL OBJETO DE LA CALIFICACIÓN 8
Modelos de rating: califican clientes *8*
Modelos de scoring: califican cliente-operación *8*
EN FUNCIÓN DE SI SE CONOCE LA VARIABLE INDEPENDIENTE O NO 9
Modelos no supervisados: *9*
Modelos supervisados: *9*
MODELOS DICOTÓMICOS O ELECCIÓN BINARIA: 10
EN FUNCIÓN A LA CANTIDAD DE INFORMACIÓN QUE UNO CUENTA DEL CLIENTE O CLIENTE-OPERACIÓN. 10
Modelo de admisión *10*
Modelo de seguimiento *10*

METODOLOGÍA DE CONSTRUCCIÓN **12**

FASES DE LA CONSTRUCCIÓN 14
ANÁLISIS DE LA CARTERA E IDENTIFICACIÓN DE LA POBLACIÓN OBJETIVO *14*
ANÁLISIS DE LA CALIDAD DE LA INFORMACIÓN *17*
TRATAMIENTO DE LA POBLACIÓN *20*
SELECCIÓN Y CONSTRUCCIÓN DE VARIABLES *24*
ANÁLISIS UNIVARIANTE DE LAS VARIABLES *28*
ANÁLISIS MULTIVARIANTE DE LAS VARIABLES *30*
IDENTIFICACIÓN Y SELECCIÓN DEL ALGORITMO DE CALIFICACIÓN *31*
ANÁLISIS DE BONDAD DEL MODELO *35*
ESTIMACIONES DE LOS PARÁMETROS DE LA REGRESIÓN *37*
ANÁLISIS DEL PODER PREDICTIVO *40*
GRANULADO DEL MODELO *48*

COMO AFECTARÁ A LA ESTRATEGIA DE LA ENTIDAD LA UTILIZACIÓN DEL MODELO DE RIESGO-RENTABILIDAD **50**

DEFINICIÓN DE LAS FUENTES DE DATOS **51**

BIBLIOGRAFÍA **51**

Introducción

Los bancos tienen un gran dilema: confiar en cada uno de sus clientes, ya que dan dinero a través de préstamos y créditos, sin solicitar garantía, a cambio de un costo financiero compuesto por interés, seguro y gastos de envío/renovación.

El negocio de los créditos al consumo es un negocio que ejemplifica de manera perfecta la relación riesgo-rentabilidad en cada operación comercial.

La entidad presta dinero de manera rápida y solicitando, información declarada y escasa documentación al cliente.

La situación óptima para la entidad sería:

- Contar con una estimación de si el cliente le podrá devolver el dinero solicitado
- Si utilizará el producto de manera que la entidad tenga ingresos
- Si la inversión realizada por la entidad para captar el cliente será repagada en la vida útil del mismo.

Por todo lo expuesto, es fundamental que la política de riesgo, marketing y venta esté dirigida a clientes con alta probabilidad que devuelvan el crédito otorgado, pagando los intereses y que luego sean clientes fieles que permitan a la entidad repagar el costo de adquisición.

Entonces las preguntas que surgen serían:

¿Cuales son las características de los clientes de bajo riesgo, rentables y fieles?

¿Como debemos ponderar cada una de esas características, variables, atributos para encontrar la combinación adecuada?

¿Cual es la probabilidad que un cliente que me solicita un producto sea de bajo riesgo, rentable y genere valor para la empresa?

Estas preguntas se contestarán en el documento, desarrollando un modelo predictivo, que el banco integrará a su estrategia de adquisición de clientes. Es decir calculando la calificación del modelo al momento que un cliente solicita un producto crediticio para la toma de diferentes decisiones.

Terminología básica para Comprender la Tesina

Característica

Cuando hablemos de característica, nos estamos refiriendo a toda variable o campo de información disponible del cliente. Esta información puede provenir de distintas fuentes: la solicitud de crédito, documentación del cliente, comportamiento del cliente en el mercado financiero, etc.

Ejemplos de variables de la solicitud son: la edad, nivel de ingresos, estudios, modelo del automóvil, profesión, domicilio, etc. Las variables pueden ser tanto campos originales como otras variables creadas a partir de las originales.

Modelo

Un modelo propone una ponderación para cada una de las variables que lo componen, creando en el fondo una nueva variable que corresponde a los diferentes niveles de puntaje que el modelo asigna[1].

Un modelo se puede obtener por métodos estadísticos, métodos de inteligencia artificial, métodos de minería de datos, métodos de redes neuronales, métodos de algoritmos genéticos, encuestas entre expertos, o por simple intuición e inspiración. Lo importante es que tanto para el modelo, como cualquier variable original o intermedia, puede medirse su capacidad para discriminar entre buenos y malos clientes.

Mas adelante veremos como se pueden clasificar los modelos en función de si se conoce la variable independiente o no, en que momento se calculan, etc.

Riesgo

[1] Fundamentos de Evaluación de Capacidad de Discriminación de Variables y Modelos en Análisis Crediticio (visto en www.automind.cl visitado el 09/07/2009)

Cuando nos referimos al riesgo en sentido amplio, podemos resumirlo en la siguiente frase "posibilidad de que se produzca un acontecimiento que puede ocasionar una pérdida económica para una entidad"

Los riesgos que se diferencian en una entidad bancaria son:

- Riesgo de crédito
- Riesgo de mercado
- Riesgo operacional
- Otros riesgos

La gestión de distintas clases de riesgos, es una actividad fundamental del negocio bancario.

Todos los productos y servicios bancarios están sujetos a algún tipo de riesgo, este es asumido en contraprestación de un interés económico.

El objetivo es maximizar el ingreso minimizando el riesgo asumido, para ello las entidades tienen que ser capaces de medir el riesgo y de gestionarlo intentando cubrirse frente a el o traspasándolo fuera de la entidad.

En la tesina nos centraremos en el riesgo de crédito y la definición que utilizaremos será:

Riesgo de pérdida que se puede producir por el incumplimiento de los pagos debidos a la entidad.[2]

El riesgo de crédito se produce como consecuencia del posible no cumplimiento por parte de la contrapartida del contrato firmado entre ambas partes.

[2] Management Solutions, Junio 2004: "Curso de riesgo de crédito" jornadas de capacitación para empleados de la consultora.

Objetivo de la Tesina

El objetivo es desarrollar un modelo predictivo que estime la probabilidad que un cliente tenga un perfil deseado.

¿Cómo definimos un buen cliente o un perfil deseado para la entidad?

Los buenos clientes de un banco deberían reunir las siguientes características:

- Altos ingresos netos – margen porcentual[3] positivo superior a la tasa de morosidad
- Bajo riesgo de incobrabilidad – baja tasa de morosidad
- Permanente - fiel

Si ahondamos en los conceptos anteriores para sentar las bases de futuras conclusiones, debemos tener en cuenta que:

- Ingresos netos, se referirá a que el cliente periódicamente genere un saldo de dinero positivo para el banco.
- Bajo riesgo de incobrabilidad, todos los clientes tienen asociada una probabilidad de no pago, o default.
- Permanente, ser fiel y utilizar los servicios financieros del banco todo el tiempo posible, permitiendo repagar el costo de adquisición.

Una vez definido el perfil deseado y las variables disponibles para explicar este fenómeno, con el modelo calcularemos la probabilidad de que un cliente cumpla con dicho perfil para tomar decisiones al momento de la aprobación crediticia.

[3] Margen porcentual: tasa libre de costos de fondeo y de costos atribuibles al producto.

Clasificación de los Modelos de Calificación [4] [5]

En este capítulo se explicará la clasificación y diferentes modelos que se pueden realizar, para situarnos luego en la metodología de la construcción de un modelo real, en la cual se comenzará con una explicación teórica cerrando con la aplicación práctica del modelo real.

Los modelos de calificación tienen por objeto, a partir de la información definida por las variables seleccionadas, clasificar la inversión crediticia de una entidad en grupos homogéneos a efectos del criterio seleccionado.

En función del objeto de la calificación

Los modelos se clasifican en:

Modelos de rating: califican clientes

Con estos modelos la inversión crediticia de una entidad queda clasificada en grupos homogéneos de riesgo a partir de la calificación otorgada por el modelo al titular de la exposición.

Estos modelos se utilizan para calificar grandes empresas y PYMEs, dado que miden la capacidad de la empresa para atender a sus compromisos financieros en su conjunto.

Modelos de scoring: califican cliente-operación

Con estos modelos la inversión crediticia queda clasificada en grupos homogéneos de riesgo a partir de la calificación que el modelo otorga al binomio cliente-operación, en

[4] Métodos estadísticos básicos de discriminación con árboles de decisión

[5] Management Solutions, Junio 2004: "Curso de riesgo de crédito"

función de información sobre las características de dicha operación y características de su titular.

Estos modelos se utilizan en Banca de Particulares y Pequeños Negocios, donde está probado que un titular puede comportarse de manera diferente según el tipo de producto de que se trate.

En función de si se conoce la variable independiente o no

Según este criterio los modelos se dividen en modelos supervisados y no supervisados.

Modelos no supervisados:

Se utilizan estos modelos cuando no se dispone de una muestra con una situación histórica o calificación dada.

Modelos supervisados:

Se emplean cuando se dispone de una muestra con una histórica del cliente o calificación dada con la que entrenar el modelo.

El entrenamiento del modelo permite que reconozca los atributos, características y patrones de la cartera estudiada y lo vincule a la definición objetivo o variable independiente.

El conjunto de atributos, características y patrones afectados en una proporción establecida determina el modelo.

Posteriormente se introducen en el modelo nuevos prospectos de los que no se conoce su clasificación para obtener una calificación.

Como ejemplos se destacan el análisis discriminante, los árboles de decisión, la regresión logística y las redes neuronales, y, más recientemente, los algoritmos genéticos y las máquinas de vector soporte.

Modelos dicotómicos o elección binaria:

Los modelos de calificación clasifican la inversión crediticia de una entidad en grupos homogéneos a efectos del objetivo buscado.

En función a la cantidad de información que uno cuenta del cliente o cliente-operación.

Los modelos se clasifican en:

Modelo de admisión

Asigna una probabilidad a las operaciones en el momento de formalización de la operación.
Característica general: no se posee información acerca del comportamiento del solicitante si éste no es cliente de la entidad.

Modelo de seguimiento

Asigna una probabilidad a la operación en un momento de la vida de la misma distinta de su fecha de formalización.
Característica general: se posee información acerca del comportamiento del cliente tanto con esta operación como con otras.

Modelo de anti-fuga

Caso particular de un modelo de seguimiento. Asigna una probabilidad a la operación en un momento de la vida de la misma distinta.

Característica general: se posee información acerca del comportamiento del cliente tanto con esta operación como con otras.

Intenta predecir si un cliente dará de baja su producto o lo dejará de usar definitivamente.

Con la introducción teórica realizada, nos centraremos para desarrollar un modelo de scoring, supervisado, dicotómico o de elección binaria y de admisión.

Metodología de Construcción[6]

En el siguiente apartado se expondrá una manera de estructurar el desarrollo de un modelo de decisión. Luego de cada explicación o introducción teórica se acompaña con el desarrollo de las fases prácticas del modelo.

Como veremos en el documento, generalmente el primer paso de todo análisis es entender donde estamos parados, por lo tanto, debe realizarse un estudio previo de la cartera que permita captar las peculiaridades de la estructura y composición de ésta. Condición necesaria para este paso es contar con la información suficiente.

Luego de este entendimiento general del problema, hay que definir el conjunto de clientes buscado para utilizar en los análisis.

El desarrollo del modelo se basa en la asociación entre datos particulares, financieros y de negocio con una situación crediticia, para lograr el reconocimiento de patrones de comportamiento en función de la información disponible. El algoritmo del modelo contiene las variables y la ponderación de cada uno de ellas para el cálculo del score.
Con dicha ecuación se produce una calificación de los prospectos en función de la puntuación ofrecida por el modelo.

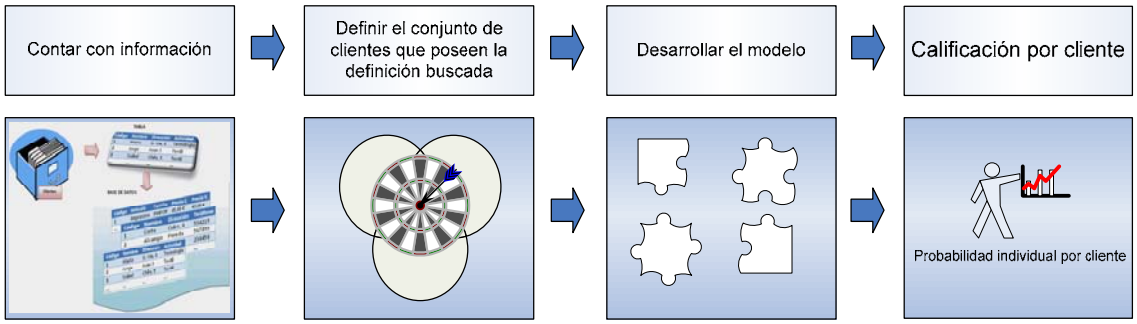
El score obtenido se valida y se contrasta a cada uno de los niveles crediticios internos que estén calculadas en un período histórico amplio y con una muestra de clientes representativa[7].

Estos modelos, tienen la condición de contar con una muestra suficiente de resultados buscados, consistente en términos estadísticos, que permita el entrenamiento del modelo.

[6] La base teórica de la siguiente sección se deriva de Management Solutions, Junio 2004: "Curso de riesgo de crédito"

[7] Es muy relativo al caso analizado.

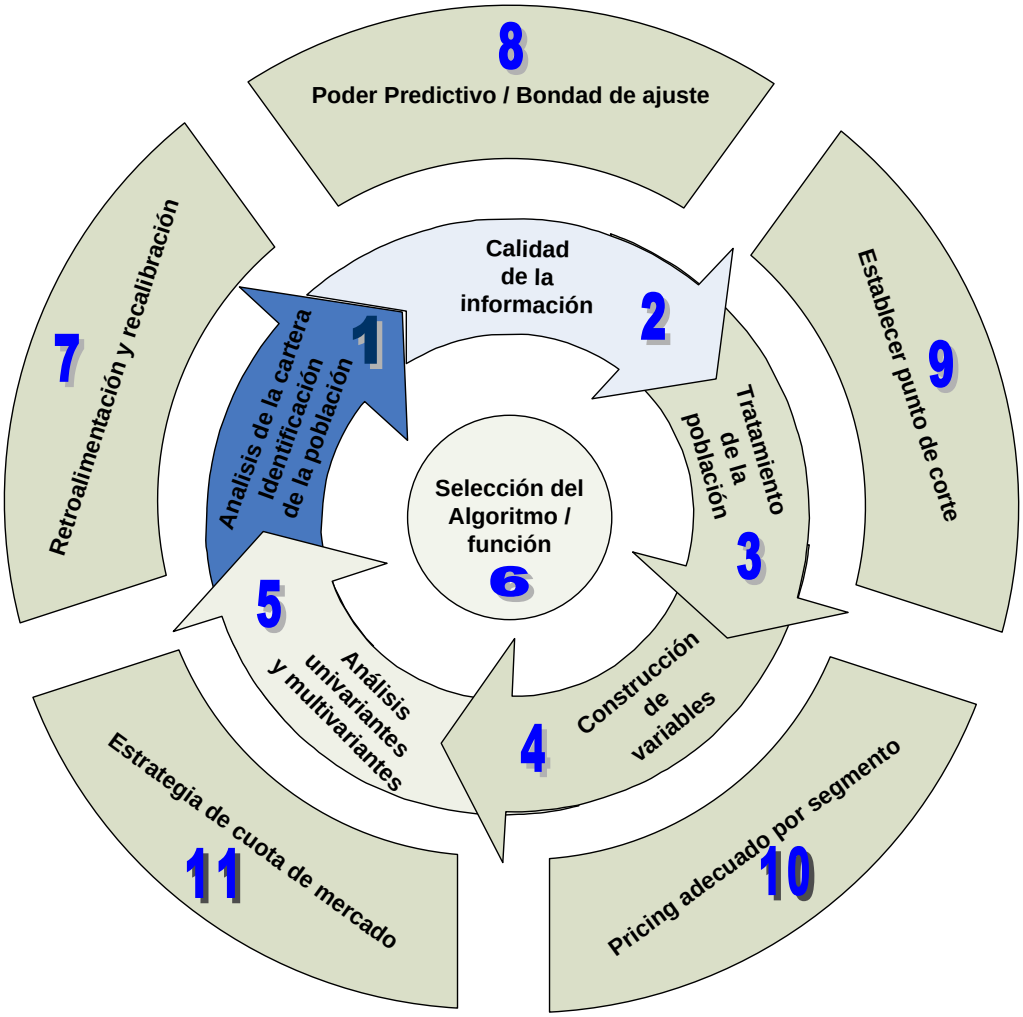
El siguiente diagrama busca generalizar la metodología de construcción[8].



[8] El diagrama contiene cuatro pasos conocidos, pero su formato es propio para esta tesina.

Fases de la construcción

El siguiente diagrama muestra un paneo general de los pasos que realizaremos y como interactúa con la realidad una vez construido el modelo[9].



ANÁLISIS DE LA CARTERA E IDENTIFICACIÓN DE LA POBLACIÓN OBJETIVO

[9] El diagrama contiene las fases típicas de una metodología de construcción, pero su formato es propio para esta tesina.

Como se comentó anteriormente, el análisis de la cartera de la entidad permite conocer cuáles son las características financieras y, eventualmente, las particularidades de los prestatarios de la entidad.

En el caso de que se observasen aspectos muy específicos, se tendrían en cuenta a la hora de construir los modelos. De esta forma, se tienen en cuenta las estrategias y políticas de crédito y financiación de la entidad, así como sus posibles nichos de mercado.

Identificar la población para la que se va a desarrollar el modelo de calificación crediticia. Por ejemplo, se segmenta principalmente por tipos de productos (préstamos de consumo, créditos hipotecarios, tarjetas de crédito, cuentas corrientes, etc.). Asimismo, se estudian otros tipos de segmentación (clientes, no clientes, etc.)

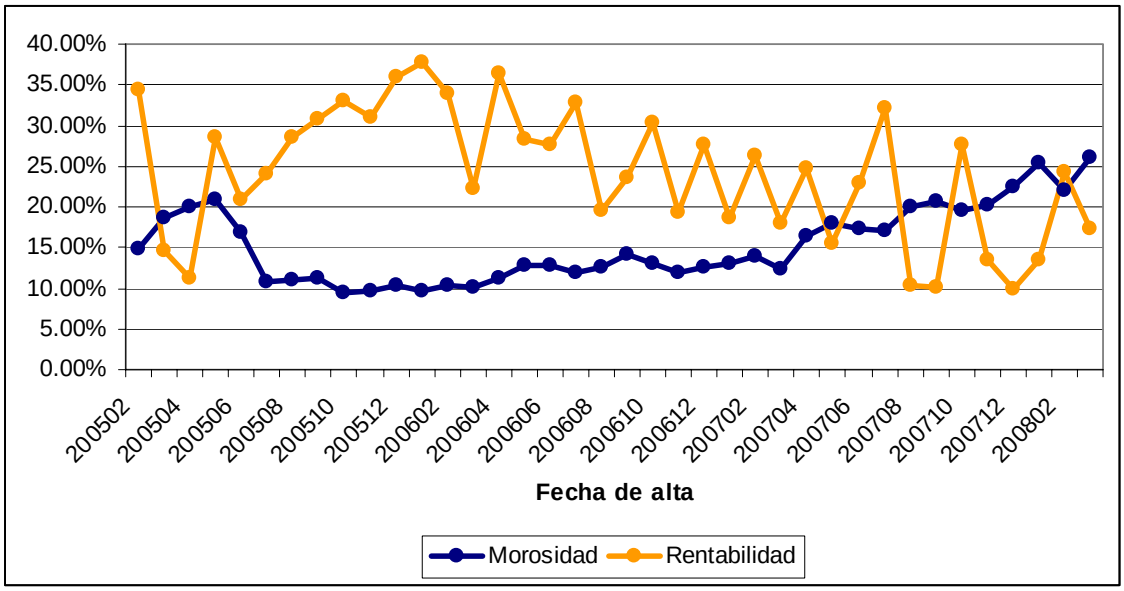
Para el modelo que construiremos se toma una cartera constituida con nuevos clientes que solicitaron una tarjeta de crédito, que tienen fechas de alta desde febrero de 2005 hasta marzo de 2008.

La población se compone por 128.257 tarjetas de crédito.

La cantidad tarjetas de crédito dadas de altas por mes se distribuye de la siguiente manera.

La política crediticia para la venta de tarjetas de crédito se basaba únicamente en antecedentes negativos, ejemplo: juicios, quiebras o situación en el banco central de la republica argentina (BCRA) y Bureaux[10].

La tasa de morosidad[11] y la rentabilidad financiera[12] de la cartera por mes de alta la podemos observar con el siguiente gráfico.



En la curva de morosidad se denota que desde abril de 2006 la morosidad comenzó en alza sin mostrar mejoría en el último mes analizado. Las principales causas se debieron a la forma de instrumentar el producto, todas las tarjetas salían con un préstamo personal preaprobado, no se utilizaba un modelo predictivo para la calificación de las cuentas y la asignación de límite no era en función al riesgo.

La rentabilidad o margen financiero, en determinados meses no era suficiente para cubrir los gastos directos e indirectos que ocasionaba ofrecer dicho producto.

[10] Centrales de información comercial y crediticia privados.
[11] Tasa de morosidad: mide la cantidad de personas por fecha de alta de la tarjeta que alcanzaron 90 días de atraso en el año posterior al alta de la tarjeta de crédito.
[12] Rentabilidad financiera: mide la diferencia entre la tasa cobrada al cliente, menos el costo de fondeo y menos la tasa de incobrabilidad.

ANÁLISIS DE LA CALIDAD DE LA INFORMACIÓN

Es necesario realizar un análisis de la calidad de la información obtenida de las bases de la entidad:

- Análisis de la bondad de la información suministrada. En general los datos utilizados provienen de las bases de datos internas de la Entidad.
- Análisis del porcentaje de información disponible para cada variable. Aquellas variables con un alto porcentaje de datos no informados son eliminadas para los estudios posteriores.
- Filtrado de registros. Con el fin de evitar la construcción del modelo con información desvirtuada se determina una serie de filtros lógicos sobre los componentes de la muestra inicial. Dichos filtros se establecen respecto de los siguientes criterios:
    - Se evalúa la coherencia lógica de la información. Por ejemplo, no se admitiría una operación en la que el titular figurase con una edad de doscientos años.
    - Se identifican los valores atípicos que pudieran distorsionar el modelo. Es el caso, por ejemplo, de un titular con 20 hijos.

Datos para el desarrollo

A nivel global de la población, podemos ver el nombre de la variable, la descripción de la variable para entender el significado, la completitud de cada variable y la cantidad de registros válidos.

Variable	Descripción de la variable	Completitud	Cantidad de registros válidos
NROCLI	Identificación del cliente	100.00%	128257
FALTA	Fecha de alta de la tarjeta de crédito	100.00%	128257
SEXO	Sexo	100.00%	128257
EST_CIV	Estado Civil	100.00%	128257
FECNAC	Fecha de nacimiento	100.00%	128257
LUGNAC	Lugar de nacimiento	99.84%	128058
PAISNAC	País de nacimiento	100.00%	128257
DOMICILIO	Domicilio	100.00%	128257
LOCALIDAD	Localidad del domicilio	100.00%	128256
PROFESION	Profesión	100.00%	128257
ACT_PRI	Actividad principal	99.99%	128248
EMPRESA	Empresa donde trabaja	92.73%	118934
ANTIGÜEDAD	Antigüedad laboral	100.00%	128257
AUTOMOVIL	Tiene automóvil	24.59%	31533
CANT_HIJOS	Cantidad de hijos	99.87%	128088
FAM_A_CARGO	Familiares a cargo	100.00%	128257
MODELO_VEHI	Modelo del vehículo	21.60%	27709
CAJA_JUB	Régimen de jubilación	100.00%	128256
COND_VIVIENDA	Condición de la vivienda	100.00%	128257
TBC_INGRESO_NETO	Ingreso neto	99.99%	128249
EDAD	Edad	100.00%	128257
PP	Tiene prestamos personal incluido en la tarjeta de crédito	100.00%	128257
TASAMORA_SUCURSAL_12	Tasa de mora de la sucursal de los últimos 12 meses	100.00%	128257

Variable	Descripción de la variable	Completitud	Cantidad de registros válidos
TASAMORA_SUCURSAL_TRI	Tasa de mora de la sucursal de los últimos 3 meses	100.00%	128257
PFRAUDE	Puntuación de prevención de fraude[13]	99.99%	128249
LIMITE	Máximo limite posible de la operación, calculado en función al ingreso y al nivel de endeudamiento	99.99%	128249

En el siguiente cuadro podemos observar las variables que disponemos para modelar, el tipo de dato, el valor mínimo, el valor máximo, la media, el desvío estándar, la asimetría[14], cantidad de valores posibles y la cantidad de valores válidos.

	Campo	Tipo	Mín	Máx	Media	Desv. Estandar	Asimetría	Únicos	Válidos
1	NROCLI	Range							128257
2	FALTA	Range	200502	200803	200601.847	83.385	0.515		128257
3	SEXO	Set	--	--	--	--	--	2	128257
4	EST_CIV	Set	--	--	--	--	--	4	128257
5	FECNAC	Set	--	--	--	--	--	250	128257
6	LUGNAC	Set	--	--	--	--	--	250	128257
7	PAISNAC	Set	--	--	--	--	--	48	128257
8	DOMICILIO	Set	--	--	--	--	--	250	128257
9	LOCALIDAD	Set	--	--	--	--	--	250	128257
10	RESDIENCIA	Set	--	--	--	--	--	25	128257
11	COND_IVA	Set	--	--	--	--	--	5	128257
12	PROFESION	Set	--	--	--	--	--	33	128257
13	ACTPRI	Set	--	--	--	--	--	6	128257

[13] La puntuación de prevención de fraude, es una ecuación o algoritmo, calculado para cada persona en función a variables y relaciones de los datos de la persona con una base de casos con irregularidades.

[14] Asimetría: Coeficiente de asimetría de Fisher $= \gamma_1 = \frac{\mu_3}{\sigma^3}$ donde μ3 es el tercer momento en torno a la media y σ es la desviación estándar.

Si γ1 = 0, la distribución es simétrica
Si γ1 > 0, la distribución es asimétrica positiva o a la derecha.
Si γ1 < 0, la distribución es asimétrica negativa o a la izquierda.

	Campo	Tipo	Mín	Máx	Media	Desv. Estandar	Asimetría	Únicos	Válidos
14	EMPRESA	Set	--	--	--	--	--	250	128257
15	ANTIGUEDAD	Range	1	32	6.977	7.843	1.485		128257
16	AUTOMOVIL	Set	--	--	--	--	--	250	128257
17	CANT_HIJOS	Range	0	5	1.265	1.453	0.942		128088
18	FAM_A_CARGO	Range	0	4	0.122	0.522	5.043		128257
19	MODELO_VEHI	Set	--	--	--	--	--	250	128257
20	CAJA_JUB	Set	--	--	--	--	--	25	128257
21	COND_VIVIENDA	Set	--	--	--	--	--	3	128257
22	TBC_INGRESONETO	Range	0	15000	1142.406	799.938	6.152		128249
23	EDAD	Range	20	72	43.449	13.645	0.586		128257
24	PP	Range	0	1	0.611	0.488	-0.455		128257
25	TASAMORA_SUC_12	Range	0.008	0.1	0.033	0.026	1.106		128257
26	TASAMORA_SUC_TRI	Range	0	0.5	0.034	0.041	3.281		128257
27	PFRAUDE	Range	-7.626	-0.929	-4.575	1.469	-0.405		128249

TRATAMIENTO DE LA POBLACIÓN

La construcción de los modelos requiere de la definición de una muestra de construcción específica para cada uno de ellos donde se recojan los distintos patrones crediticios.

Por lo tanto es necesario definir para cada componente de la muestra inicial, compuesta por un conjunto de operaciones, qué se entiende por registro moroso o no moroso, rentable o no rentable, que aporto valor al banco o no, en resumidas palabras se utiliza la clasificación dicotómica (malo o bueno), e identificar las variables que permiten definir el perfil del mismo.

La definición de registros buenos o malos se corresponde con el siguiente criterio, general, que se bajará a una definición concreta en el caso real desarrollado mas adelante:

Operaciones buenas: son aquellas operaciones en las que el cliente ha hecho frente a todos los compromisos de pago establecidos en el momento de la contratación, ha

generado una ganancia para el banco y permanece con una relación de cliente con la entidad.

Operaciones malas: son aquellas operaciones en las que no se han atendido los pagos establecidos en el momento de la contratación, habiendo estado la operación morosa en algún momento de su vida, pese a que en un momento posterior se hubiese regularizado la situación, que no haya generado una ganancia para el banco o se abandone la relación con la entidad.

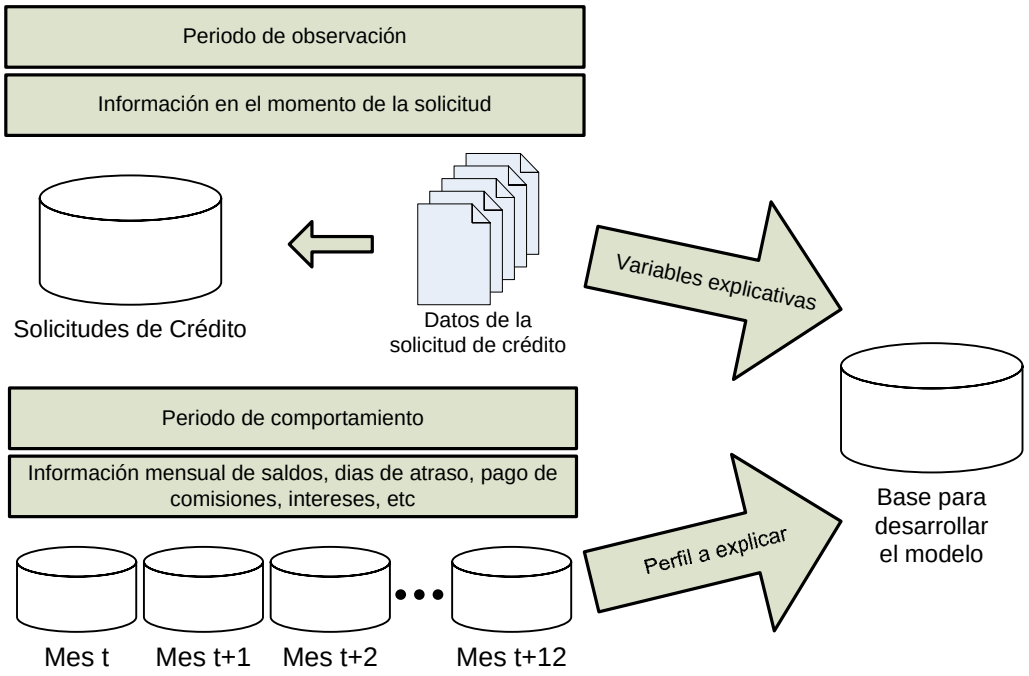
Comentario:

Para los modelos de scoring se establece en la bibliografía y también lo remarcan acuerdos internacionales que el conjunto de datos debe contemplar todas las operaciones solicitadas (aprobadas y rechazadas), puesto que un modelo estadístico únicamente basado en datos de operaciones concedidas proporcionará un resultado sesgado. Es necesario incluir en la construcción del modelo información acerca de las operaciones denegadas, para incluir los perfiles de éstas en los algoritmos de clasificación.
Es decir, en todos los modelos de admisión se deben incluir datos sobre operaciones rechazadas. Para ello la entidad debe capturar y guardar dicha información en la base de datos.
Si se cuenta con información acerca de las operaciones denegadas, se hace necesario estimar si dichas operaciones son buenas o malas a efectos de la construcción del modelo.

Proceso de creación de la base necesaria para el desarrollo del modelo

Según como se explica anteriormente, hay que trabajar con atributos y una definición real del perfil buscado, para poder realizar dicho proceso debemos llevar a una base única las variables estáticas de la solicitud y las variables de comportamiento para poder identificar y asignar el comportamiento buscado.

En el siguiente gráfico[15] se puede visualizar con un ejemplo este proceso. En el período de observación tenemos la información de las solicitudes de crédito y en los meses posteriores al alta de la tarjeta de crédito, las bases de las variables que muestran el comportamiento del cliente. Asociando a cada cliente su comportamiento tenemos la base para comenzar con el desarrollo del modelo.



Definiciones de Performance

Para construir dicha base debemos saber cuantos meses incluiremos en el análisis, pero esa respuesta la encontramos en la definición de performance, que se explica en el párrafo siguiente.

Muchos factores entran en juego al elegir la mejor definición de comportamiento. Los dos factores principales son la **clasificación empírica** (cual es el perfil buscado) y **longitud del período de resultado** (cuánto tiempo se estudia a los clientes antes de que se los clasifique).

[15] Desarrollo propio para la tesina

El período de análisis para definir el objetivo, se definió desde febrero de 2005 hasta marzo de 2008 y para cada uno de esos meses de observación se sigue el comportamiento o performance de 12 meses. El siguiente gráfico muestra una representación gráfica.

Período de observación	Período de Performance											
feb-05	mar-05	abr-05	may-05	jun-05	jul-05	ago-05	sep-05	oct-05	nov-05	dic-05	ene-06	feb-06
mar-05	abr-05	may-05	jun-05	jul-05	ago-05	sep-05	oct-05	nov-05	dic-05	ene-06	feb-06	mar-06
abr-05	may-05	jun-05	jul-05	ago-05	sep-05	oct-05	nov-05	dic-05	ene-06	feb-06	mar-06	abr-06
.	.	.	.	.	.	.	.	.	.	.	.	.
.	.	.	.	.	.	.	.	.	.	.	.	.
.	.	.	.	.	.	.	.	.	.	.	.	.
nov-07	dic-07	ene-08	feb-08	mar-08	abr-08	may-08	jun-08	jul-08	ago-08	sep-08	oct-08	nov-08
dic-07	ene-08	feb-08	mar-08	abr-08	may-08	jun-08	jul-08	ago-08	sep-08	oct-08	nov-08	dic-08
ene-08	feb-08	mar-08	abr-08	may-08	jun-08	jul-08	ago-08	sep-08	oct-08	nov-08	dic-08	ene-09
feb-08	mar-08	abr-08	may-08	jun-08	jul-08	ago-08	sep-08	oct-08	nov-08	dic-08	ene-09	feb-09
mar-08	abr-08	may-08	jun-08	jul-08	ago-08	sep-08	oct-08	nov-08	dic-08	ene-09	feb-09	mar-09

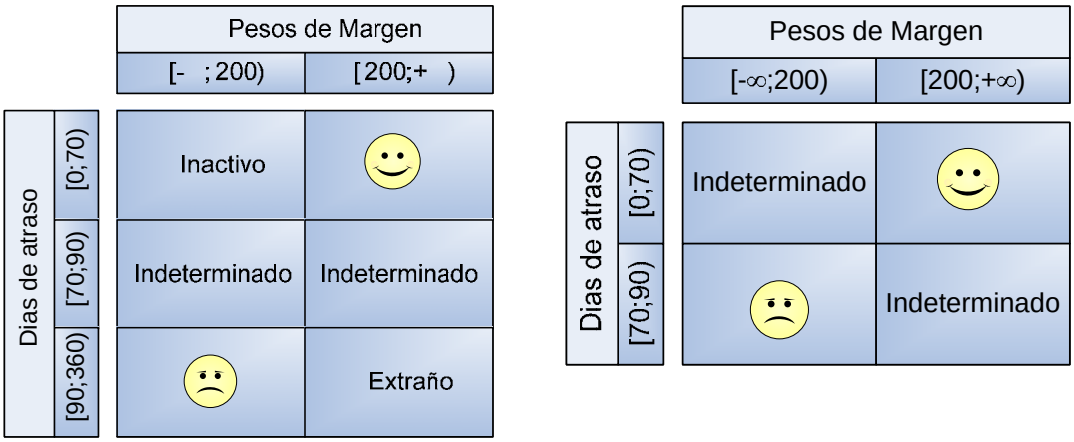
La definición del criterio o perfil buscado.

Dentro de la tesina se analizará la siguiente alternativa de análisis.

Perfil no deseado, malo: clientes que hayan alcanzado una morosidad mayor a 90 días de morosidad y que generen un margen negativo para la entidad en un año calendario desde el alta de su tarjeta de crédito.

Perfil deseado, bueno: clientes que no cumplan con la condición de malo o perfil no deseado.

Se plantean diversas alternativas, para definir el perfil buscado, donde se establece un grupo denominado indeterminado; por una cuestión de extensión para desarrollar otro modelo, no se tratarán en la tesina, pero se plantea la posibilidad de otros casos de estudio. En los siguientes diagramas se muestran estos ejemplos.



La diferencia de cada cuadro es la definición de los conjuntos posibles de clientes indeterminados que pueden provocar que el modelo no se entrene correctamente.

SELECCIÓN Y CONSTRUCCIÓN DE VARIABLES

Una vez analizada la información y realizados los filtros iniciales de la misma, se construyen las variables susceptibles de formar parte del modelo. Posteriormente se seleccionarán solamente aquellas que aporten un alto poder discriminante al modelo.

En ciertos casos las variables incrementan su capacidad discriminante si sus valores son agrupados en un menor número de valores o categorías, no depende del método utilizado para construir el modelo, sino de la oportunidad de darle sentido lógico a la variable. Por ello se procede a su categorización, la cual proporciona: transformación de las variables continuas en variables discretas o transformación de las variables discretas en otras variables discretas con menor número de categorías.

La categorización se lleva a cabo teniendo en cuenta, para cada variable, uno o más de los siguientes criterios:

- Homogeneidad de cada nueva categoría creada. Se realizan grupos en los que el porcentaje de morosidad dentro del mismo sea parecido, es decir, que la variable categorizada tenga una desviación típica reducida respecto a la morosidad dentro de cada categoría.
- Mantenimiento del sentido socioeconómico de la variable, es decir, formación de categorías con características socioeconómicas homogéneas.
- Homogeneidad de número de muestra. Si bien todas las categorías de cada variable no van a tener el mismo porcentaje de muestra, se evita la formación de categorías en las que se concentre la mayor parte de la muestra o en las que el porcentaje de la misma sea insuficiente para captar el patrón de comportamiento.

Una vez realizada la categorización de la variable, ésta se incluirá en el modelo, pudiéndose hacer de dos maneras:

- Incluir la variable categórica
- Incluir la variable continua

En el siguiente cuadro se muestra como se agrupo la variable lugar de residencia en 7 categorías reemplazando la variables alfanumérica por la tasa de malos de la agrupación realizada.

Residencia	Perfil no deseado	Cantidad total	Porcentaje por filas	Categoria
ENTRE RIOS	23	80	28.75%	22.8%
SAN LUIS	675	2,979	22.66%	
RIO NEGRO	397	2,166	18.33%	17.6%
JUJUY	409	2,243	18.23%	
SANTA FE	2,271	13,057	17.39%	
SANTA CRUZ	10	60	16.67%	
NEUQUEN	534	3,518	15.18%	15.0%
MENDOZA	1,705	11,344	15.03%	
STGO DEL ESTERO	890	5,931	15.01%	
CHUBUT	335	2,240	14.96%	
CAPITAL FEDERAL	535	3,589	14.91%	
BUENOS AIRES	3,113	21,345	14.58%	14.3%
MISIONES	701	4,872	14.39%	
TIERRA DEL FUEG	3	22	13.64%	
SALTA	1,195	8,889	13.44%	
LA PAMPA	1	9	11.11%	10.1%
SAN JUAN	871	7,961	10.94%	
TUCUMAN	1,292	12,786	10.10%	
CORDOBA	1,854	18,957	9.78%	
CORRIENTES	119	1,291	9.22%	9.2%
FORMOSA	1	11	9.09%	
CHACO	400	4,774	8.38%	8.3%
CATAMARCA	7	112	6.25%	
LA RIOJA	1	20	5.00%	
EN EL PAIS	0	1	0.00%	
	17,342	**128,257**	**13.52%**	

Otro ejemplo de creación de variables es el siguiente, la variable empresa agrupada tenía 61.543 valores posibles, se agrupó en las siguientes 15 categorías.

Empresa categorizada	Perfil no deseado	Cantidad total	Porcentaje por filas
MUNICIPAL	986	5,874	16.79%
ARMADA	21	129	16.28%
POLICIA	486	3,215	15.12%
PRIVADO	12,487	87,871	14.21%
ESTATAL	1,318	10,770	12.24%
EJERCITO	57	500	11.40%
GENDARMERIA	38	337	11.28%
VACIO	995	9,323	10.67%
AEREA	33	319	10.34%
ANSES	374	3,814	9.81%
DOCENTE	316	3,299	9.58%
MONOTRIBUTO	176	2,028	8.68%
AUTONOMO	30	354	8.47%
RELACIONADO	24	401	5.99%
DESCONOCIDO	1	23	4.35%
	17,342	**128,257**	**13.52%**

Composición de variables

Con la construcción de variables compuestas se trata de reflejar información adicional que cada variable por separado no recoge y sí la relación de varias.

Para reflejar este hecho, se estudia la construcción de nuevas variables estableciendo relaciones lógicas entre ellas. De este modo, se crean nuevas variables, que son composición de las originales, con el fin de optimizar la capacidad discriminante del modelo resultante.

En esta fase se construyen variables discretas, cuyas categorías formadas tienen comportamiento similar, es decir, existe homogeneidad en el resultado dentro de cada nueva categoría y presentan comportamientos lógicos.

Posteriormente, se estudia la importancia de cada nueva variable creada en la discriminación entre operaciones con resultado no deseado y operaciones con resultado deseado comparando los resultados obtenidos con las variables consideradas de manera individual.

De entre las variables creadas en este paso, se seleccionan para futuros análisis sólo aquellas en las que la capacidad discriminante se vea optimizada.

ANÁLISIS UNIVARIANTE DE LAS VARIABLES

El análisis univariante tiene como fin identificar aquellas variables con alta capacidad discriminante de manera individual. Se distinguen las siguientes fases:

- Análisis descriptivos de las variables
- Estudio de las variables frente al resultado obtenido.

Análisis descriptivos de las variables

Con ellos se obtiene una visión detallada de cada variable, con el fin de utilizar en los análisis posteriores aquellos contrastes más adecuados para el tipo de variable que se está analizando. Los análisis descriptivos son distintos dependiendo de si la variable es continua (numérica no dividida en categorías), discreta ordinal (dividida en categorías con un orden implícito dentro de las categorías) o discreta nominal (dividida en categorías sin orden entre las mismas):

- Variables continuas. Se estudian las medidas de tendencia central y la distribución de la variable (análisis de normalidad,...)
- Variables discretas ordinales y nominales. Se estudia la concentración de la variable (distribución de frecuencias) en cada categoría de la misma.

Las medidas de dispersión y tendencia central utilizadas son: la media, la mediana, la varianza.

Estudio de las variables frente al resultado

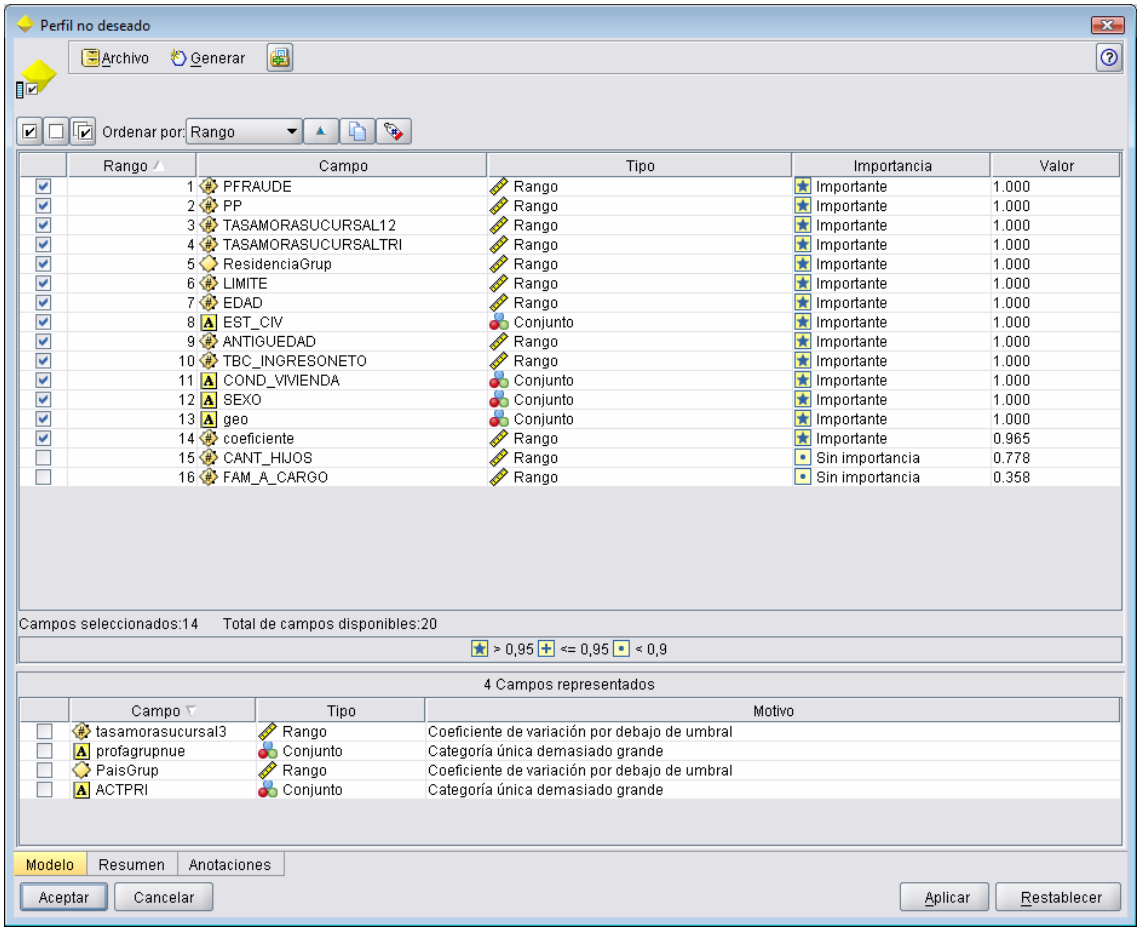
Este estudio tiene por objeto medir la capacidad discriminante de cada variable de manera individual, es decir, la relación existente entre ella y la variable a explicar.

En la práctica se busca realizar un análisis gráfico de la distribución de operaciones con resultado negativo y con resultado positivo y análisis de las tendencias de las variables frente al resultado de la operación.

El siguiente análisis es una variante para seleccionar las variables del futuro modelo, los parámetros para definir si una variable es importante, se lo denomina "Configuración de creación[16]" y los podemos resumir:

- Coeficiente mínimo de variación: 0,1
- Porcentaje máximo de registros en una única categoría: 90,0
- Porcentaje máximo de valores perdidos: 70,0
- Número máximo de categorías como un porcentaje de los registros: 95,0
- Desviación típica mínima: 0,0
- Basar el valor p (importancia) de los predictores categóricos en: Pearson
- Importancia baja desde 0,0 y hasta: 0,9
- Importancia media por encima y hasta: 0,95
- Etiqueta de importancia baja: Sin importancia
- Etiqueta de importancia media: Secundario
- Etiqueta de importancia alta: Importante
- Seleccionar en el modelo: Todos los campos con rango

[16] Para obtener el análisis univariante o las variables importantes para modelar se utilizó Clementine 10.1



Adicionalmente a las variables recomendadas en esta parte, al momento de modelar se pueden probar todas las variables simultáneamente utilizando distintos algoritmos que permiten entrar y salir variables del modelo.

ANÁLISIS MULTIVARIANTE DE LAS VARIABLES

Por medio del análisis multivariante se analizan las variables que han resultado discriminantes de manera individual para observar su capacidad discriminante teniendo en cuenta las posibles relaciones que existan con el resto de las variables. Este análisis se ha dividido en dos partes:

- Correlaciones: se estudia la relación existente entre las diferentes variables, para introducir en el modelo el menor número posible de variables con el máximo poder discriminante (principio de parsimonia).

- Selección multivariante de las variables: una vez identificadas las relaciones entre las variables, se realizan contrastes para analizar qué conjunto de variables maximiza el poder discriminante del modelo.

IDENTIFICACIÓN Y SELECCIÓN DEL ALGORITMO DE CALIFICACIÓN

Para la realización del modelo se selecciona la muestra de entrenamiento y la muestra de test.

- **Muestra de entrenamiento** será aquella con la que se desarrolla el modelo, es decir, a partir de la cual se asignan los pesos a las variables. Selección aleatoria, del 50%.

- **Muestra de test** será aquella con la que se comprueban los resultados del modelo construido. . Selección aleatoria, del 50%.

Una vez identificada la muestra de construcción a emplear, con los registros calificados como buenos o malos, y las distintas variables que han mostrado mayor capacidad de discriminación, se construye el modelo utilizando la metodología de regresión logística, que es un caso particular de red neuronal.

Los resultados resumidos del modelo[17], son lo siguientes:

[17] Para derivar el modelo se utilizó el software Clementine 10.1

Profundizando el análisis del modelo construido, en el siguiente cuadro podremos ver las variables que proceso el modelo para la muestra de entrenamiento, sus valores posibles, la cantidad de registros y su representación en dicha variable.

Resumen del procesamiento de los casos			
Variable y valores posibles		N	Porcentaje marginal
Perfil no deseado	0	55,232	86.40%
	1	8,696	13.60%
ACTPRI		5	0.00%
	COMERCIANTE	2,409	3.80%
	EMPLEADO	58,564	91.60%
	EMPRESARIO	231	0.40%
	JUBILADO	1,700	2.70%
	PROFESIONAL	1,019	1.60%
COND_VIVIENDA	NO PROPIETARIO	32,141	50.30%
	PROPIETARIO	31,787	49.70%
EST_CIV	Casado	26,688	41.70%
	Divorciado	2,967	4.60%
	Soltero	26,650	41.70%
	Viudo	7,623	11.92%
SEXO	Femenino	25,948	40.60%
	Masculino	37,980	59.40%
Geo	Centro	12,361	19.30%
	Norte	43,610	68.20%
	Sur	7,957	12.40%
profagrupnue	Empleado	61,209	95.70%
	Jubilado	1,700	2.70%
	Profesional	1,019	1.60%
Válidos		63,928	100.00%
Perdidos		0	
Total		63,928	
Subpoblación		63896(a)	

a. La variable dependiente sólo tiene un valor observado en 63886 (100.0%) subpoblaciones.

En el siguiente cuadro, podemos observar cuales fueron las variables que se probaron en el modelo, su test de significación o grado de aporte al modelo y su reemplazo en el caso de que una variable sea mas representativa que una variable que ya había ingresado.

Resumen de los pasos							
Modelo	N	Acción	Efecto(s)	Criterio de ajuste del modelo	Contrastes de selección de efectos		
				-2 log verosimilitud	Chi-cuadrado(a,b)	gl	Sig.
Paso 0	0	Introducido	Intersección	50830.630	.		
Paso 1	1	Introducido	PFRAUDE	47325.049	3085.307	1	.000
Paso 2	2	Introducido	LIMITE	47095.072	214.134	1	.000
Paso 3	3	Introducido	EDAD	46922.247	168.859	1	.000
Paso 4	4	Introducido	ResidenciaGrup	46763.274	160.149	1	.000
Paso 5	5	Introducido	TBC_INGRESONETO	46694.537	58.219	1	.000
Paso 6	6	Introducido	EST_CIV	46632.090	62.210	4	.000
Paso 7	7	Introducido	CANT_HIJOS	46545.794	87.787	1	.000
Paso 8	8	Introducido	SEXO	46484.740	60.512	1	.000
Paso 9	9	Introducido	TASAMORASUCURSALTRI	46432.551	53.285	1	.000
Paso 10	10	Introducido	PP	46375.423	54.651	1	.000
Paso 11	11	Introducido	ANTIGÜEDAD	46333.400	41.422	1	.000
	12	Eliminado	TBC_INGRESONETO	46334.070	.670	1	.413
Paso 12	13	Introducido	TASAMORASUCURSAL12	46290.016	44.499	1	.000
Paso 13	14	Introducido	Profagrupnue	46245.969	38.079	2	.000
Paso 14	15	Introducido	Geo	46230.545	15.466	2	.000
Paso 15	16	Introducido	PaisGrup	46222.024	7.655	1	.006
Método por pasos: Por pasos hacia delante							
a. El valor de chi-cuadrado para su inclusión se basa en la prueba de puntuación.							
b. El valor de chi-cuadrado para su eliminación se basa en la prueba de la razón de verosimilitudes.							

ANÁLISIS DE BONDAD DEL MODELO

La probabilidad de que los resultados observados, teniendo en cuenta los parámetros estimados, se conoce como verosimilitud. Se acostumbra a utilizar -2 veces el logaritmo natural de la verosimilitud (-2LL) como una medida de ajuste del modelo, ya que tiene vínculos con la distribución chi-cuadrado. Un buen modelo que tiene una alta verosimilitud se traduce en un escaso valor-2LL. Para un ajuste perfecto,-2LL sería igual a cero.

El contraste chi-cuadrado es una prueba estadística donde su hipótesis nula iguala todos los coeficientes del modelo a cero. Es equivalente a la prueba F global en la regresión.

Su valor, es simplemente la diferencia entre la inicial (basado en el modelo que contiene sólo la constante) y final-2LL. Tomando los grados de libertad como la diferencia entre el número de parámetros en los dos modelos, uno solamente con la constante y otro con cada parámetro beta asignado más la constante.

Si la significación o importancia es menor de .05, rechazamos la hipótesis nula y concluimos que el conjunto de variables mejora la predicción de las posibilidades de registro.

Información del ajuste del modelo				
	Criterio de ajuste del modelo	Contrastes de la razón de verosimilitud		
Modelo	-2 log verosimilitud	Chi-cuadrado	gl	Sig.
Sólo la intersección	50830.630			
Final	46222.024	4608.606	19	.000

La regresión logística también proporciona dos medidas que son análogos al R^2 de la regresión de mínimo cuadrados[18]. Debido a la relación entre la media y la desviación estándar de una variable dicotómica, la cantidad de varianza explicada por el modelo debe ser definido de manera diferente. Normalmente, el pseudo R2 Nagelkerke es preferible porque puede, a diferencia de la de Cox y Snell R2, alcanzar un valor máximo de uno.

Pseudo R-cuadrado	
Cox y Snell	.070
Nagelkerke	.127
McFadden	.091

[18] En un modelo de regresión lineal el coeficiente de determinación se interpreta como el porcentaje de variación de la variable dependiente explicado por el modelo.

Se calcula como: $R^2 = \frac{Varianza\ \text{explicada}\ por\ el\ \text{modelo}}{Varianza\ Total}$

ESTIMACIONES DE LOS PARÁMETROS DE LA REGRESIÓN

En el siguiente cuadro se expresan los estimadores de la regresión logística, seguido de un test de hipótesis para evaluar su significación estadística y luego un intervalo de confianza para el estimador puntual del parámetro $\exp(\beta_i)$.

Estimaciones de los parámetros de la regresión									
Perfil no deseado(a)	Variable	B	Error típ.	Wald	gl	Sig.	Exp(B)	Intervalo de confianza al 95% para Exp(B)	
								Límite inferior	Límite superior
0	Intersección	10.95	.823	176.909	1	.000			
	ANTIGÜEDAD	.017	.002	70.154	1	.000	1.018	1.013	1.022
	CANT_HIJOS	-.100	.010	107.998	1	.000	.905	.888	.922
	EDAD	.037	.002	286.390	1	.000	1.037	1.033	1.042
	[EST_CIV=C]	.665	.063	111.135	1	.000	1.944	1.718	2.200
	[EST_CIV=D]	-.044	.075	.345	1	.557	.957	.826	1.109
	[EST_CIV=S]	.206	.062	10.853	1	.001	1.228	1.087	1.388
	[EST_CIV=V]	.516	.079	42.121	1	.000	1.675	1.433	1.957
	LIMITE	.000	.000	51.663	1	.000	1.000	1.000	1.000
	PFRAUDE	1.147	.115	98.826	1	.000	3.149	2.511	3.947
	PP	-4.791	.339	200.020	1	.000	8.31E-003	4.28E-003	1.61E-002
	PaisGrup	-.008	.003	7.531	1	.006	.992	.986	.998
	ResidenciaGrup	-.004	.000	104.177	1	.000	.996	.995	.997
	[SEXO=F]	.361	.029	156.305	1	.000	1.434	1.355	1.518
	[SEXO=M]	0(b)	.	.	0	.	.	.	.
	TASAMORASUCURSAL12	-4.651	.619	56.427	1	.000	9.55E-003	2.84E-003	3.21E-002
	TASAMORASUCURSALTRI	-6.833	.507	181.462	1	.000	1.08E-003	3.99E-004	2.91E-003
	[geo=C]	-.160	.043	13.629	1	.000	.852	.782	.928
	[geo=N]	-.070	.037	3.579	1	.059	.932	.867	1.003
	[geo=S]	0(b)	.	.	0	.	.	.	.
	[profagrupnue=E]	-.845	.142	35.374	1	.000	.430	.325	.568
	[profagrupnue=J]	-.816	.160	25.925	1	.000	.442	.323	.605
	[profagrupnue=P]	0(b)	.	.	0	.	.	.	.
	a. La categoría de referencia es: 1.000000.								

En función a los parámetros del modelo construiremos la ecuación para el cálculo de las probabilidades que un cliente que se presenta a la entidad sea del perfil deseado.

$$P[\text{Perfil deseado}=1]= \frac{e^{\beta_0+\sum_{i=1}^{n}\beta_i * X_i}}{1+e^{\beta_0+\sum_{i=1}^{n}\beta_i * X_i}}$$

Siendo:

β_0=10,95

$\sum_{i=1}^{n}\beta_i * X_i$ =[0,01746 * ANTIGUEDAD + -0,1 * CANT_HIJOS + 0,03668 * EDAD + 0,6647 * [EST_CIV=C] + -0,04421 * [EST_CIV=D] + 0,2057 * [EST_CIV=S] + 0,5155 * [EST_CIV=V] + 0,000115 * LIMITE + 1,147 * PFRAUDE + -4,791 * PP + -0,008095 * PaisGrup + -0,003994 * ResidenciaGrup + 0,3606 * [SEXO=F] + -4,651 * TASAMORASUCURSAL12 + -6,833 * TASAMORASUCURSALTRI + -0,1603 * [geo=C] + -0,07029 * [geo=N] + -0,8447 * [profagrupnue=E] + -0,8159 * [profagrupnue=J]

Explicación de los términos de la ecuación

Lo que se puede observar del algoritmo obtenido en función a su signo (matemáticamente hablando a la derivada primera en función de la variable analizada) como afecta dicha variable al perfil buscado, entonces tenemos:

0,01746 * ANTIGÜEDAD

La antigüedad en el empleo afectando positivamente, es decir a mayor antigüedad en el empleo mejor probabilidad de perfil deseado, es una relación con coherencia económica.

-0,1 * CANT_HIJOS

La cantidad de hijos nos esta describiendo un impacto negativo en el perfil deseado.

0,03668 * EDAD

La edad representa un atributo que favorece en la ecuación.

0,6647 * [EST_CIV=C] + 0,5155 * [EST_CIV=V]+ 0,2057 * [EST_CIV=S] + -0,04421 * [EST_CIV=D]

La variable estado civil genera lo que se denomina variable dummy, es decir, genera distintas variables y ponderaciones que se activa cuando la persona tiene el estado civil establecido, asignando el valor 1 a la variable, por lo tanto podemos ver que los casados son los mas favorecidos, luego están los viudos y por últimos los solteros, y la característica divorciado resta en el perfil.

0,000115 * LIMITE

A mayor límite máximo posible el perfil mejora, porque la persona supone mayor ingreso y/o menor endeudamiento.

1,147 * PFRAUDE

La puntuación de prevención de fraude, es negativa, por lo tanto mientas mas cercano a cero el valor, el perfil mejora. Como se comentó al momento de describir las variables, la puntuación de prevención de fraude es un modelo en si mismo, representado por una ecuación del tipo $\beta_0 + \sum_{i=1}^{n} \beta_i * X_i$

-4,791 * PP

Si la persona solicita un préstamo personal en el mismo momento en el que solicita la tarjeta de crédito, el modelo lo castiga en casi 5 puntos en contra.

-0,008095 * PaisGrup

A mayor tasa de morosidad de la variable país agrupado el modelo lo castiga.

-0,003994 * ResidenciaGrup

A mayor tasa de morosidad de la variable residencia agrupada el modelo lo castiga.

+ 0,3606 * [SEXO=F]

Si el cliente es femenino, el modelo mejora el perfil.

-4,651 * TASAMORASUCURSAL12 + -6,833 * TASAMORASUCURSALTRI +

Si el punto de venta donde se esta vendiendo el producto tuvo malas tasas de morosidad en los últimos 12 y 3 meses el modelo empeora la calificación.

-0,1603 * [geo=C] + -0,07029 * [geo=N] +

Si la solicitud se completa en sucursales de centro o norte se realiza un pequeño castigo

-0,8447 * [profagrupnue=Empleado] + -0,8159 * [profagrupnue=Jubilado]

Si la persona declara empleado o jubilado es castigado.

El modelo tiene un valor constante valuado en: + 10,95

ANÁLISIS DEL PODER PREDICTIVO

Se evalúa la capacidad de distinción del modelo entre la ocurrencia o no del evento buscado. Los análisis empleados para la estimación de la bondad de ajuste del modelo son los siguientes:

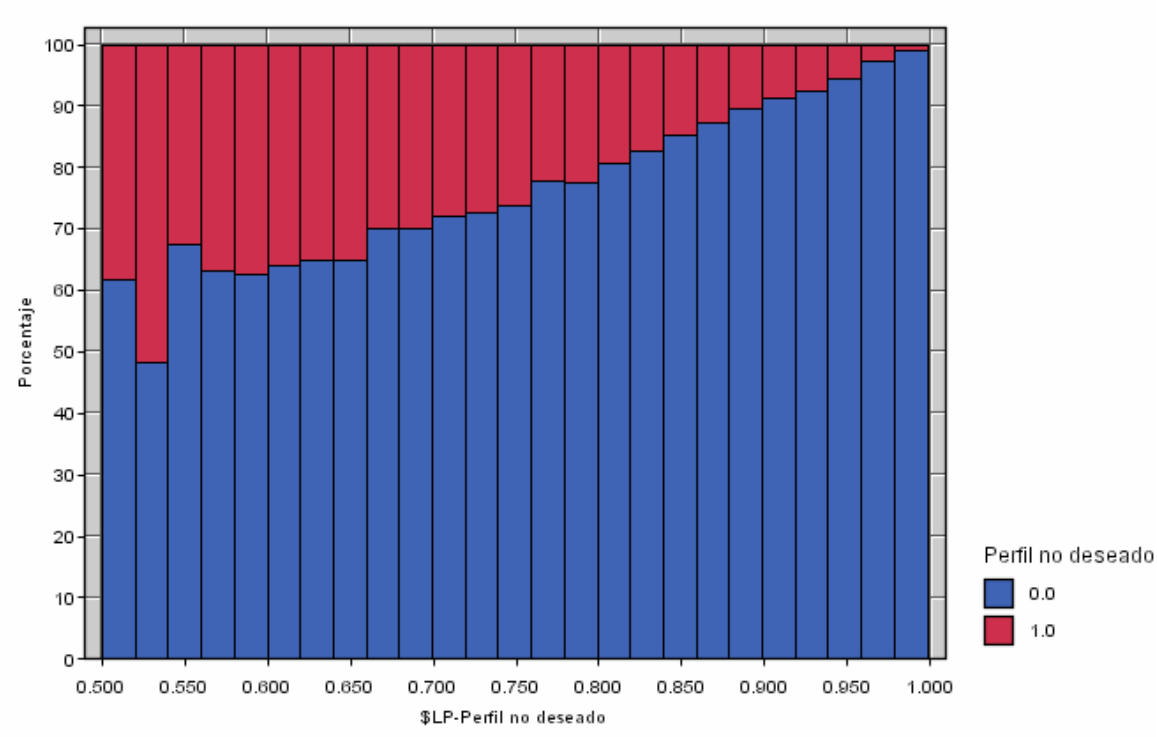
- Gráfico de las distribuciones de las calificaciones que asigna el modelo al grupo de las operaciones con resultado negativo y con resultado positivo.

- Índice de poder: medida que indica el grado de predicción del modelo, es decir, un índice de poder elevado indica que las operaciones con resultado negativo son calificadas en su mayor parte por el modelo en las peores calificaciones, mientras que las operaciones con resultado positivo son calificadas en los niveles altos.
- Curva ROC (Receiving Operating Characteristic), que compara distintos escenarios de aceptación o denegación de operaciones con el desempeño conocido de las mismas. De dicha comparación se obtienen distintos puntos que componen la curva ROC, cuya área comprendida indicará el nivel de predicción del modelo.
- Porcentaje de acierto de las operaciones con resultado negativo y de las operaciones con resultado positivo.

Todas estas medidas se evalúan en la muestra reservada para test, es decir, operaciones que no interviene en la construcción del modelo. Generalmente, el porcentaje de muestra reservado para este análisis esta entre el 25% y 50% de la muestra total de construcción.

Histograma 1

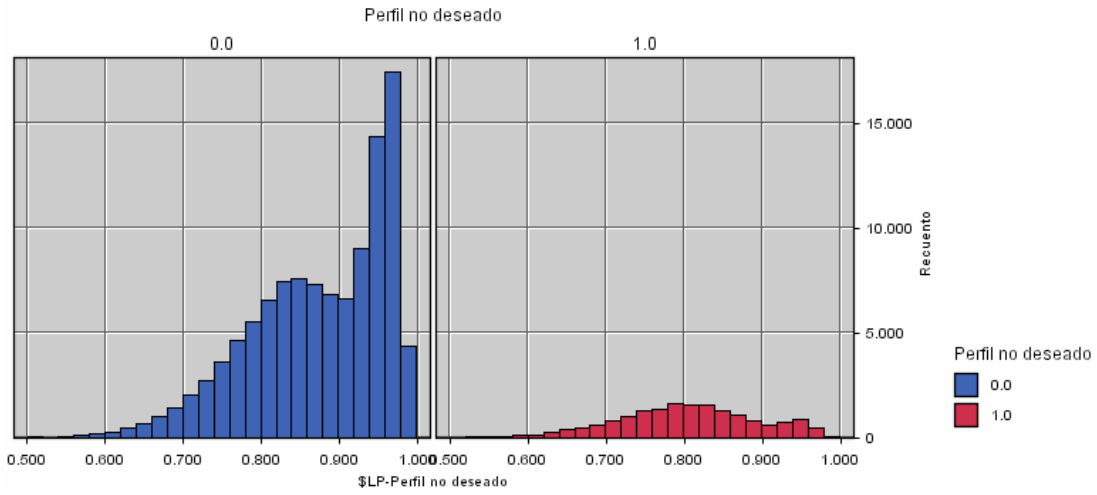
El siguiente histograma nos muestra para cada rango de la calificación del modelo el porcentaje de clientes "malos" o "buenos" tomando el total de dicho rango como 100%.
La muestra fue ordenada de peor a mejor, por probabilidad obtenida por el modelo, donde se denota que el modelo encuentra mayor porcentaje de clientes malos en las probabilidades mas bajas y mayor porcentaje de clientes buenos en las probabilidades mas altas.
Esto se representa en el gráfico por dos colores, el azul para los perfiles deseados y el bordo para los perfiles no deseados.
El modelo muestra que es mucho mejor su utilización que elegir un criterio de selección al azar, ya que muestra el orden en función a la probabilidad del modelo.



Ejemplo: El rango entre 0,5 y 0,55 de probabilidad calculada por el modelo, tiene un porcentaje de registros con perfil no deseado desque asciende al 40% y 60% de perfil deseado, mientras que el rango de 0,95 a 1 de probabilidad calculada por el modelo, tiene un porcentaje de registros con perfil no deseado del 3% y 97% de perfil deseado

Histograma 2

El siguiente histograma, es una variante del anterior, muestra la distribución de la muestra por rango de score, ordenado por probabilidad obtenida por el modelo, separando a la población de perfil deseado en el primer histograma de barras azules y el perfil no deseado en el segundo histograma de barras bordo.

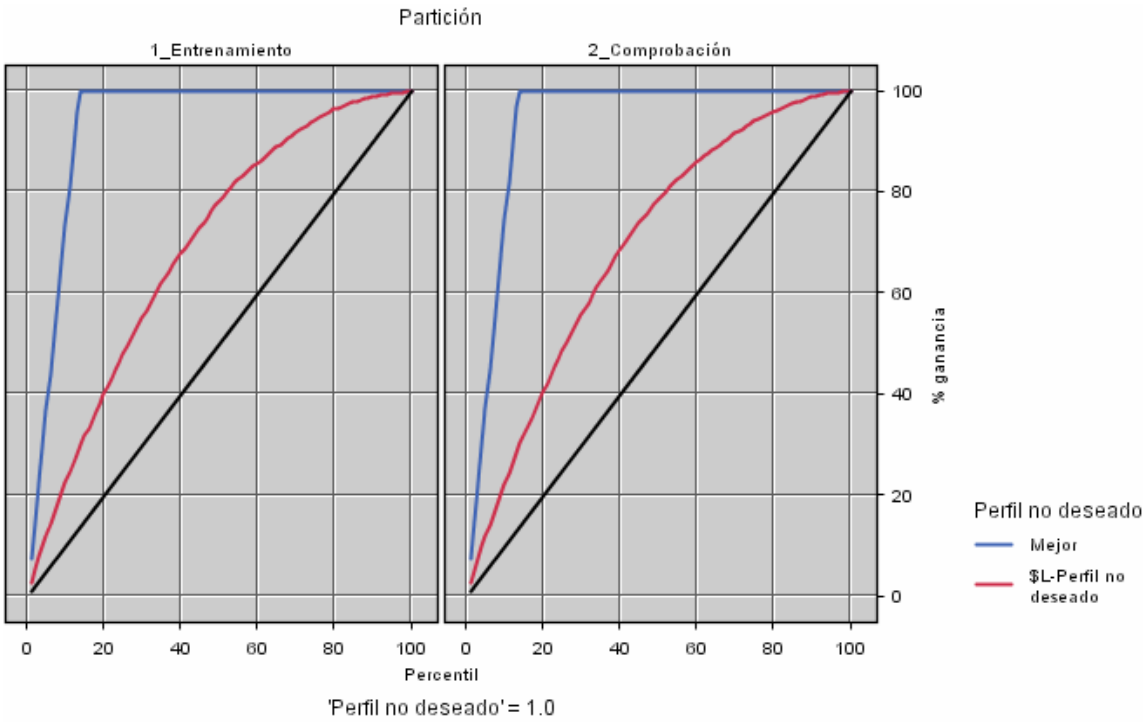


Power stat o curva de ganancia

Esta curva se basa en observar dónde se acumulan las calificaciones otorgadas por el modelo y comparar lo observado con la mejor de las situaciones posibles, es decir, con un modelo perfecto en el cual todos los clientes del perfil no deseado tengan asignadas las peores clasificaciones.

Para ello, se dibujan tres curvas:

- Curva de frecuencia acumulada de las calificaciones del modelo a los registros no deseados (Curva del modelo, línea roja).
- Curva de frecuencia acumulada de las calificaciones dadas a los registros "no deseados" por el "mejor modelo", es decir, aquel que asigna a los "clientes no deseados" las peores calificaciones (Curva mejor modelo, línea azul).
- Curva de frecuencia acumulada de las calificaciones dadas a los registros "de perfil no deseado" por "el peor modelo posible", es decir el que carece de discriminación alguna entre clientes "malos" y "buenos" (Curva Aleatoria, línea negra).



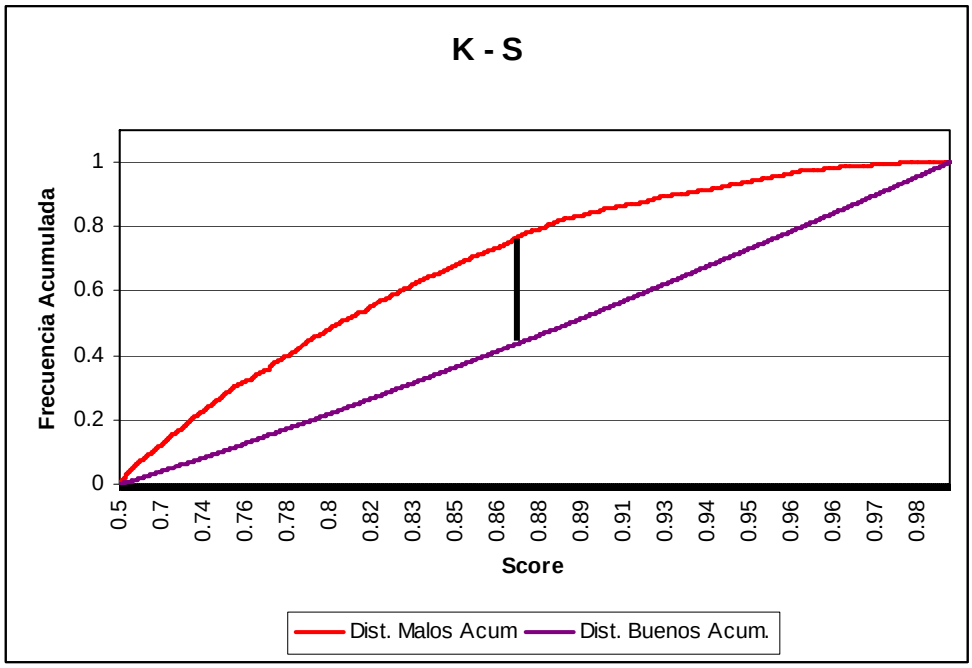
Podemos observar para la muestra de entrenamiento y la muestra de comprobación el grado de ganancia o poder de usar el modelo.

Ejemplo: Si la entidad utilizaría como punto de corte el percentil 20, es decir, no aceptó ninguna solicitud del 20% peor calificada por el modelo, estaría obteniendo una ganancia porque en ese grupo estaría dejando fuera el 40% de las solicitudes que tienen un perfil no conveniente para la entidad.

En función al indicador de Potencia del Modelo de 43.9% podemos concluir que el modelo tiene un poder explicativo importante para un modelo de admisión.

Distancia de Kolmogorov-Smirnov

Distancia de Kolmogorov-Smirnov entre las distribuciones de las calificaciones que asigna el modelo al grupo de las operaciones con resultado negativo y de las con resultado positivo. Una distancia de Kolmogorov elevada indica la existencia de diferencias entre las calificaciones otorgadas a la muestra de operaciones que destruyen valor y a la de operaciones que aportan valor, es decir, un alto poder predictivo.



El índice K-S es de 33,9% en la probabilidad 0,870, calculado caso a caso. Otro indicio que la calificación que realiza el modelo nos permite diferenciar los perfiles establecidos.

Tabla de performance con intervalos equiprobables

Además de los test para medir el poder predictivo o los histogramas, se utilizan las tablas de performance para entender la distribución de los malos por rango de score, la distribución de los buenos por rango de score y de una manera intuitiva explicar el punto de corte en función a que porcentaje de perfil no deseado estoy dispuesto a aprobar.

La siguiente tabla esta desarrollada con intervalos de la misma cantidad de casos por rango de score, donde se pueden interpretar las distribuciones de malos/buenos por score, relativa y acumulada.

Score		# perfil no deseado	# del rango	Dist. de malos por rango	Dist. de buenos por rango	Dist. de malos relativa	Dist. de buenos relativa	Dist. de malos acum.	Dist. de buenos acum.
500	698	2087	6404	32.59%	67.41%	12.05%	3.90%	12.05%	3.90%
698	738	1790	6404	27.95%	72.05%	10.33%	4.17%	22.38%	8.06%
738	763	1628	6404	25.42%	74.58%	9.40%	4.31%	31.78%	12.38%
763	784	1450	6404	22.64%	77.36%	8.37%	4.47%	40.15%	16.85%
784	801	1404	6404	21.92%	78.08%	8.11%	4.51%	48.26%	21.36%
801	817	1237	6404	19.32%	80.68%	7.14%	4.66%	55.40%	26.03%
817	831	1166	6405	18.20%	81.80%	6.73%	4.73%	62.13%	30.76%
831	845	1041	6404	16.26%	83.74%	6.01%	4.84%	68.14%	35.60%
845	859	897	6404	14.01%	85.99%	5.18%	4.97%	73.32%	40.57%
859	874	842	6404	13.15%	86.85%	4.86%	5.02%	78.18%	45.59%
874	891	737	6404	11.51%	88.49%	4.25%	5.12%	82.44%	50.71%
891	908	583	6404	9.10%	90.90%	3.37%	5.26%	85.80%	55.96%
908	924	511	6404	7.98%	92.02%	2.95%	5.32%	88.75%	61.28%
924	937	492	6405	7.68%	92.32%	2.84%	5.34%	91.59%	66.62%
937	947	428	6404	6.68%	93.32%	2.47%	5.40%	94.07%	72.02%
947	955	342	6404	5.34%	94.66%	1.97%	5.47%	96.04%	77.49%
955	962	277	6404	4.33%	95.67%	1.60%	5.53%	97.64%	83.02%
962	968	190	6404	2.97%	97.03%	1.10%	5.61%	98.74%	88.63%
968	976	147	6404	2.30%	97.70%	0.85%	5.65%	99.58%	94.28%
976	998	72	6405	1.12%	98.88%	0.42%	5.72%	100.00%	100.00%

Una manera práctica de ver la capacidad discriminante del modelo, es ver como ordena de mayor a menor la distribución de malos por rango y la distribución de malos relativa.

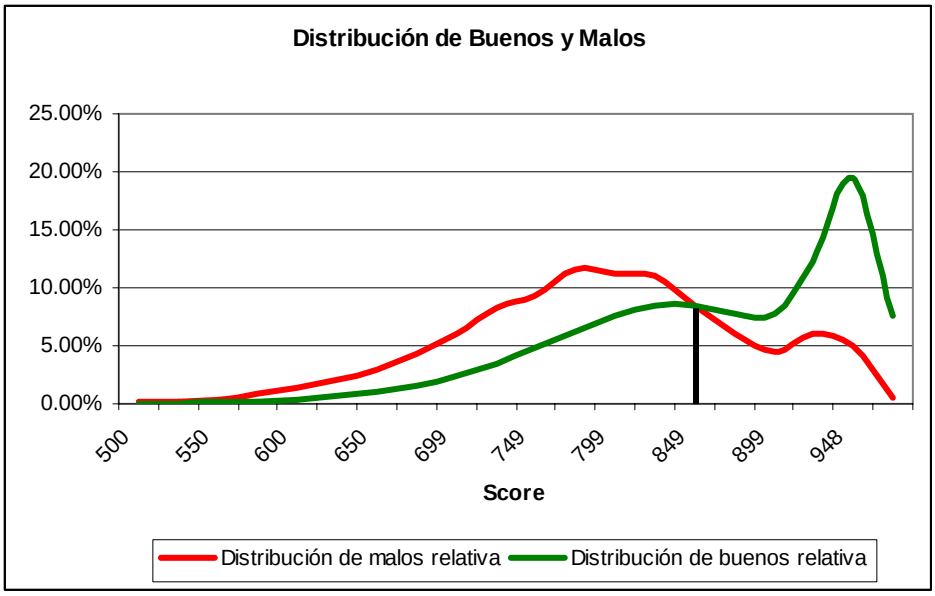
Tabla de performance con intervalos de ancho fijo

La siguiente tabla esta desarrollada con 20 intervalos de igual amplitud de score, 25 puntos de rango de score, donde se pueden interpretar las distribuciones de malos/buenos por score, relativa y acumulada.

Score		# perfil no deseado	# del rango	Dist. de malos por rango	Dist. de buenos por rango	Dist. de malos relativa	Dist. de buenos relativa	Dist. de malos acum.	Dist. de buenos acum.
500	524	32	75	42.67%	57.33%	0.18%	0.04%	0.18%	0.04%
525	550	41	95	43.16%	56.84%	0.24%	0.05%	0.42%	0.09%
550	575	68	194	35.05%	64.95%	0.39%	0.11%	0.81%	0.20%
575	600	139	379	36.68%	63.32%	0.80%	0.22%	1.62%	0.42%
600	625	231	621	37.20%	62.80%	1.33%	0.35%	2.95%	0.77%
625	650	365	1060	34.43%	65.57%	2.11%	0.63%	5.06%	1.40%
650	675	512	1643	31.16%	68.84%	2.96%	1.02%	8.01%	2.42%
675	699	747	2463	30.33%	69.67%	4.31%	1.55%	12.33%	3.97%
699	724	1034	3691	28.01%	71.99%	5.97%	2.40%	18.30%	6.37%
724	749	1434	5289	27.11%	72.89%	8.28%	3.48%	26.57%	9.85%
749	774	1599	6924	23.09%	76.91%	9.23%	4.81%	35.81%	14.65%
774	799	2001	8808	22.72%	77.28%	11.55%	6.15%	47.36%	20.80%
799	824	1955	10312	18.96%	81.04%	11.29%	7.55%	58.65%	28.35%
824	849	1902	11333	16.78%	83.22%	10.98%	8.51%	69.63%	36.86%
849	874	1456	10877	13.39%	86.61%	8.41%	8.51%	78.03%	45.37%
874	899	1043	9676	10.78%	89.22%	6.02%	7.79%	84.05%	53.16%
899	923	775	9367	8.27%	91.73%	4.47%	7.76%	88.53%	60.92%
923	948	1032	14510	7.11%	92.89%	5.96%	12.17%	94.49%	73.09%
948	973	852	22342	3.81%	96.19%	4.92%	19.40%	99.41%	92.49%
973	998	103	8424	1.22%	98.78%	0.59%	7.51%	100.00%	100.00%

En la tabla anterior no se puede ver el orden prolijo de la tabla con intervalos equiprobables, porque al tener intervalos de ancho fijo puede desordenar algún intervalo, pero podemos realizar gráficos de frecuencias que con la tabla de ancho fijo no tendrían sentido por la uniformidad de valores.

El siguiente gráfico nos muestra la distribución de la población de malos y buenos clientes, mientras menos área debajo de la curva compartan las distribuciones mejor es el modelo.



El score, donde se cruzan las distribuciones relativas de buenos y malos es el mismo score donde se establece la mayor diferencia de la distribuciones acumuladas de buenos malos, igual al estadístico K-S.

Una explicación práctica para este punto, donde la distribución de malos relativa y la distribución de buenos relativa tiene una sola intercepción, se presenta porque hasta ese punto de score la distribución acumulada de malos logro su crecimiento máximo relativo a la distribución acumulada de buenos que logro su crecimiento relativo mínimo, después de ese punto, la distribución acumulada de malos crecerá a un ritmo decreciente.

GRANULADO DEL MODELO

Con el fin de poder comparar el riesgo-rentabilidad inherente en distintas operaciones / contrapartes con distintas características en términos de riesgos-rentabilidad, conforme a una medida estandarizada basada en criterios de análisis objetivos, es necesario establecer distintos niveles, siendo los criterios seguidos para su establecimiento los siguientes:

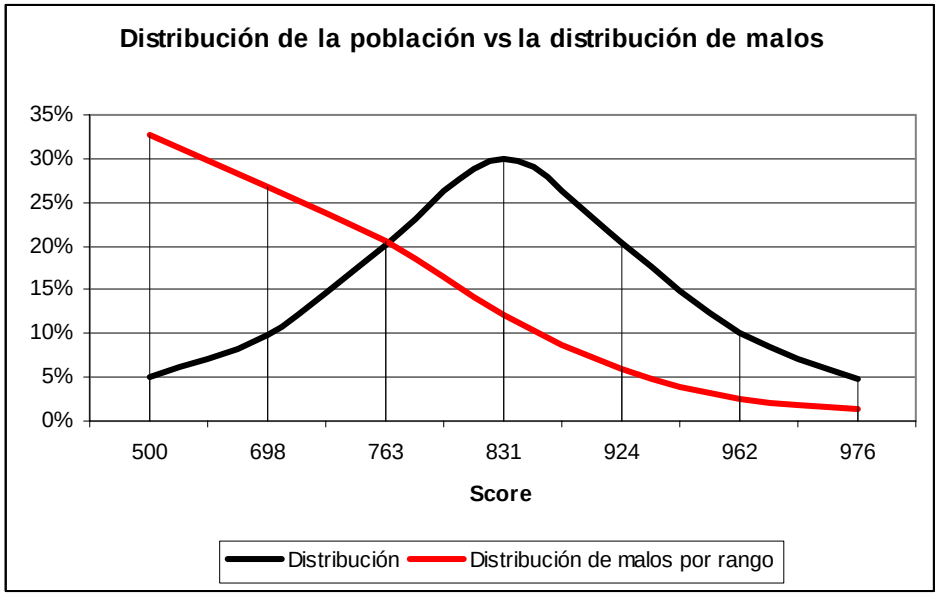
- Distancia entre los límites de los intervalos en términos de puntuación suministrada por el modelo.
- Concentración de operaciones / contrapartes por nivel.
- Concentración de importes de riesgo por nivel.

Por otro lado, hay que tener presente que los modelos deben cumplir con ciertos requisitos comúnmente aceptados:

- El modelo debe tener, como mínimo, 6 niveles crediticios
- Ningún grado o rango deberá corresponder a más de 30% de las exposiciones brutas, antes de compensación en el balance.

Para determinar los diferentes niveles en los modelos se divide la muestra en un número de categorías por percentiles de la puntuación que el modelo le otorga (es decir, la salida real del modelo ordenada de forma creciente), de manera que la distribución resultante sea simétrica y la mayor parte de las operaciones/contrapartes se concentre en los niveles centrales, mientras que en los niveles extremos (muy buenas o muy malas) se incluyan pocas operaciones. Asimismo, la tendencia de la tasa riesgo-rentabilidad ha de ser exponencialmente decreciente conforme mejoran los niveles de scoring.

Granulado	Score		# perfil no deseado	# perfil deseado	# del rango	Distribución	Distribución de malos por rango
1	500	698	2074	4282	6356	5%	33%
2	698	763	3405	9324	12729	10%	27%
3	763	831	5271	20412	25683	20%	21%
4	831	924	4606	33693	38299	30%	12%
5	924	962	1564	24508	26072	20%	6%
6	962	976	331	12602	12933	10%	3%
7	976	998	91	6094	6185	5%	1%



Como Afectará a la Estrategia de la Entidad la Utilización del Modelo de Riesgo-Rentabilidad

El modelo será utilizado en el proceso de decisión o política crediticia, para aceptar o declinar una tarjeta de crédito:

Para clientes con muy bajo score, la política declina el producto solicitado.
Superando dicho umbral mínimo, el score permitirá tomar diferentes decisiones, a nivel de control del analista de crédito, realizar o no una verificación telefónica, domiciliaria, laboral o de referencias.

Otra función importante del score es para asignar el límite de la tarjeta de crédito y la posibilidad de obtener un préstamo personal preaprobado en el mismo proceso.

Definición de las Fuentes de Datos

Las fuentes de datos serán internas del banco con 3 años de historia para el producto tarjeta de crédito y préstamos en efectivo administrado en la tarjeta.

Bibliografía

- Salvador Figueras, M (2000):"Modelos de regresión con respuesta cualitativa: regresión logística", 17-11-2004

- Management Solutions, Junio 2004: “Curso de riesgo de crédito”

- Roberto Araya, Fundamentos de Evaluación de Capacidad de Discriminación de Variables y Modelos en Análisis Crediticio.

- Group Model Development, Validation and Monitoring Standards, (For Retail and Scored SME Portfolios), Version 3.0, Junio 2005.