Universidad Andina Simón Bolívar

Sede Ecuador

Área de Gestión

Programa de Maestría en Finanzas y Gestión de Riesgos

Análisis y preparación estadística de variables para el diseño de un modelo credit score de gestión de riesgo de crédito

Freddy Hernán Carranza Vergara

2008

Al presentar esta tesis como uno de los requisitos previos para la obtención del grado de magíster de la Universidad Andina Simón Bolívar, autorizo al centro de información o a la biblioteca de la universidad para que haga de esta tesis un documento disponible para su lectura según las normas de la universidad.

Estoy de acuerdo en que se realice cualquier copia de esta tesis dentro de las regulaciones de la universidad, siempre y cuando esta reproducción no suponga una ganancia económica potencial.

Sin perjuicio de ejercer mi derecho de autor, autorizo a la Universidad Andina Simón Bolívar la publicación de esta tesis, o de parte de ella, por una sola vez dentro de los treinta meses después de su aprobación.

……

Freddy H. Carranza V.

Marzo, 2008

Universidad Andina Simón Bolívar

Sede Ecuador

Área de Gestión

Programa de Maestría en Finanzas y Gestión de Riesgos

Análisis y preparación estadística de variables para el diseño de un modelo credit score de gestión de riesgo de crédito

Freddy Hernán Carranza Vergara

Tutor: Paúl Noboa

Quito - Ecuador

2008

ABSTRACT

El presente estudio busca enfatizar la labor previa a la construcción de cualquier modelo de gestión de riesgo crediticio basado en un sistema tipo *scoring*. Esta importante fase inicial involucra la identificación de las fuentes de información, la verificación de la cantidad y calidad de los datos, la determinación y tipificación de variables cualitativas, cuantitativas, demográficas, así como la verificación y cuantificación del poder discriminante de dichas variables respecto el objetivo planteado. Con este fin se profundiza en el análisis estadístico a nivel descriptivo, en forma individual y conjunta de los datos, además de bosquejar los pasos esenciales en la arquitectura de un modelo *credit score* de gestión de riesgo crediticio.

DEDICATORIA

A mi esposa Ibeth, por su apoyo, comprensión, compañía, por hacer suyas mis preocupaciones y desvelos desde el inicio hasta el fin de cada jornada.

A mis padres Ernesto y Argentina, que siempre han estado pendientes de mi superación, con su ejemplo se han constituido en la luz y el motor para alcanzar nuevas metas.

TABLA DE CONTENIDO

INTRODUCCION

En el mundo de la modelación es común utilizar la expresión: “si basura entra, basura sale” para referirse despectivamente a la poca o nula importancia que antiguamente se proporcionaba a las fuentes de información, la calidad de las bases de datos, el origen de las mismas o la interrelación que dichos datos pudiesen tener entre sí, y más aún, la concordancia existente entre las variables de entrada y el objetivo central de la modelación. He aquí el punto de partida de nuestro trabajo, orientado a la búsqueda diaria de herramientas adecuadas, que conduzcan a una eficiente gestión de riesgos, acompañadas de fuertes requerimientos de información e ingentes procesos de análisis y preparación de dichos datos.

En la actualidad se reconoce la importancia de modelos de credit score para una eficaz administración y gestión del riesgo crediticio, en la mayoría de los casos se hace hincapié en los resultados del modelo obtenido, las puntuaciones alcanzadas por las variables participantes del score, las estrategias de colocación o la trascendencia de fijar un punto de corte y su interacción con la rentabilidad del negocio, etc., pero difícilmente se puede encontrar una reseña de todo el trabajo existente detrás de dichos resultados, de los pasos a seguir en función de la disponibilidad de información.

Bajo esta perspectiva nuestro estudio procura plasmar toda esa gran labor inicial que se desarrolla en el proceso de modelación, como resultado de los conocimientos y experiencias adquiridas en el desarrollo de *scoring* de crédito que acompañados de un sólido sustento técnico, guarden estrecha relación entre la teoría y la realidad para la gestión del riesgo crediticio.

¿Qué es un *Credit Score o Scoring*?

El *scoring* se refiere al uso del conocimiento sobre el desempeño y las características de préstamos en el pasado, para pronosticar los desempeños de préstamos en el futuro. Así, cuando un analista de crédito valora el riesgo de una nueva solicitud de crédito, comparando mentalmente el presente, con la experiencia que este mismo analista ha acumulado con otros clientes al analizar solicitudes parecidas está aplicando un *scoring,* aunque sea un credit score implícito y subjetivo. De igual manera, cuando una institución financiera adopta una política de no renovar préstamos a clientes que han tenido atrasos mayores a 30 días, en su préstamo anterior, está aplicando un *scoring*, aunque sea un score simple y unidimensional.

El *credit score o scoring* es una metodología estadística que asigna en rangos la probabilidad de un resultado desconocido al otorgar puntajes a variables conocidas. Ha sido utilizado aproximadamente por 50 años para tomar *decisiones* crediticias y su aplicación es cada vez más común.

Muchas instituciones ajenas al sector financiero están utilizando técnicas de *scoring*. Por ejemplo, la telefonía celular puede utilizar el *scoring* para decidir si otorga un teléfono en prepago o post-pago; las compañías de servicios públicos pueden recurrir al *scoring* para decidir si el medidor debe ser instalado; en los almacenes de cadenas comerciales de electrodomésticos y ropa pueden utilizar el *scoring* para decidir si un cliente puede comprar productos con un crédito instantáneo; adicionalmente, en términos de crédito, puede aplicarse una metodología credit score para la aprobación, calificación y cobranza de un producto crediticio.

La gestión de riesgos basada en *credit score* tiene fundamentalmente cuatro fortalezas:

Primera fortaleza: Puede acortar el tiempo requerido para evaluar solicitudes crediticias, de esta manera, hay más tiempo y recursos disponibles para los casos que no son sencillos;

Segunda fortaleza: Hacer el riesgo más directamente susceptible a la política de la institución. Por ejemplo, una entidad financiera que quiere expandir su cartera podría pedir a los analistas de crédito que acepten clientes con más riesgo, pero cada analista actuará sobre esta instrucción a su propia manera. En contraste, un una entidad financiera con un modelo de calificación estadística podría reducir el umbral de corte de un determinado número de puntos porcentuales y saber el incremento esperado en las solicitudes aceptadas así como el aumento esperado de riesgo;

Tercera fortaleza: Radica en la homogenización y consistencia de los criterios de evaluación de riesgo, tanto en una sola oficina, como en todas las agencias u oficinas donde se evalúen solicitudes; y,

Cuarta fortaleza: Es la más importante, implica la efectividad de los modelos de calificación estadística que pueden ser comprobados antes de su uso. Esto alivia los temores de los gerentes y analistas de crédito, quienes dudan que una computadora pueda ayudarles a hacer su trabajo.

Partiendo entonces de las fortalezas de un modelo *credit score*, como de sus muy diversas formas de utilización que coadyuven en la gestión del riesgo crediticio, el trabajo que desarrollaremos a continuación procurará determinar ¿Cuáles son los

elementos que configuran el instrumental metodológico de análisis a priori en la elaboración de un modelo *credit score* de gestión de riesgo de crédito?

Para ello a lo largo de este trabajo nos concentraremos apenas en el primer peldaño, pero sin duda el más importante en la construcción de un *score* crediticio, detallando el arduo análisis estadístico que está detrás de los resultados finales, de cada grupo y tipo de variables, sin el cual hasta el más sofisticado de los modelos perdería sentido y aplicación práctica. Para alcanzar nuestro objetivo primero abordaremos los aspectos teóricos que han de sustentar el trabajo con variables cualitativas o cuantitativas, su identificación, su forma, las herramientas, procedimientos de medición y representación de dichas variables acompañados siempre de ejemplos ilustrativos.

Luego y una vez cumplida esta trascendental etapa repasaremos las fases fundamentales en la construcción de un modelo tipo *scoring*, empezando por una breve exploración sobre la integridad de la información, los elementos que ayuden en la definición del tipo de cliente, el período de exposición, etc. Posteriormente se desarrolla una escueta aplicación práctica con una base de trabajo ficticia, se trabaja estadísticamente un pequeño grupo de variables sobre las que se aplica un modelo logístico de discriminación y finalmente se exhiben los parámetros que permiten cuantificar la robustez del modelo obtenido.

I. BASE CONCEPTUAL

En esta primera fase, exploraremos el complejo mundo de la información, consolidado a través de diferentes tipos de variables, su forma de medición, cuantificación y representación gráfica, así como el análisis estadístico implícito univariante y bivariante según corresponda.

I.1 TIPOS DE VARIABLES

La fase inicial en la construcción de un modelo credit score de gestión de riesgo de crédito, parte con la preparación de la base de datos; la que, a la postre se convertirá en información útil para alcanzar el objetivo planteado en este caso, del modelo la clasificación y calificación de clientes. Para alcanzar este propósito es necesario que los datos con los cuales se ha de trabajar mantengan consistencia y coherencia para lo cual se requiere de una codificación o representación numérica de las características tanto cualitativas como cuantitativas primordiales, para la aplicación de técnicas de modelación.

I.1.1 Variables cualitativas

Cuando se hace referencia a este tipo de variables, intuitivamente se relacionan con aquellas que brindan cierta “cualidad” del elemento a medir. Desde el punto de vista de la gestión de riesgo crediticio pueden asociarse a este grupo variables como: género, estado civil, nivel de educación, el comportamiento de los agentes económicos en decisiones de consumo, de oferta de trabajo, profesión de una persona, etc. Como estas variables no aparecen en forma numérica, sino como

categorías o atributos, se pueden valorar utilizando dos tipos de escalas de medición, la escala nominal y la ordinal[1]

Escala nominal: Se utiliza para medir características que no son susceptibles de jerarquía o rango alguno; los objetos así medidos se asocian a conjuntos o categorías mutuamente excluyentes, a cada conjunto se le asigna un número. Por ejemplo, la variable género (sexo) puede tomar valores como: Mujer = 1, Hombre = 2, estos números constituyen únicamente un identificador de la variable, pues no existe siquiera una relación de orden entre ellas, es decir el número asignado solo indica la pertenencia a una categoría de la variable.

Escala ordinal: Esta escala es utilizada para medir características cualitativas en las cuales es posible establecer un escalafón u orden. Al igual que el caso anterior, se puede asignar números a los casos en estudio para indicar el grado relativo que posee de la cualidad medida. Por ejemplo la variable educación puede tomar los siguientes valores: primaria = 1, secundaria = 2, superior = 3, en este caso los valores asignados indican un mayor nivel de educación, pero no se establece una cuantía al respecto; es decir, una escala ordinal indica una posición relativa, pero no la magnitud de la diferencia de la variable medida.

Los datos correspondientes a variables cualitativas se agrupan, de manera natural en diferentes categorías, clases o familias y se cuenta el número de datos

---

1 Carmen Stella Verón, *La utilización de variables cualitativas y el análisis de datos categóricos en la investigación*, U. Nacional de Rosario, 2006, p. 4.

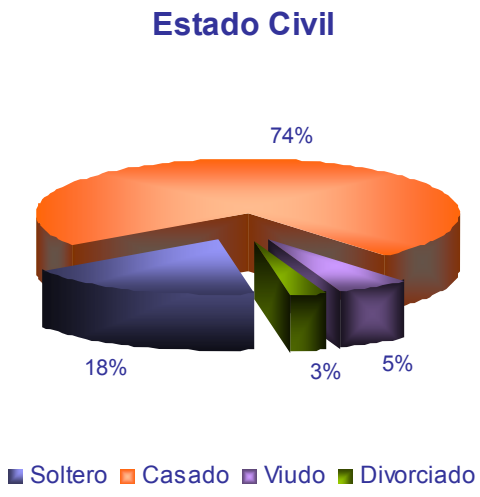
que aparecen en cada una de ellas. Suelen representarse mediante diagramas de barras, sectores o líneas.

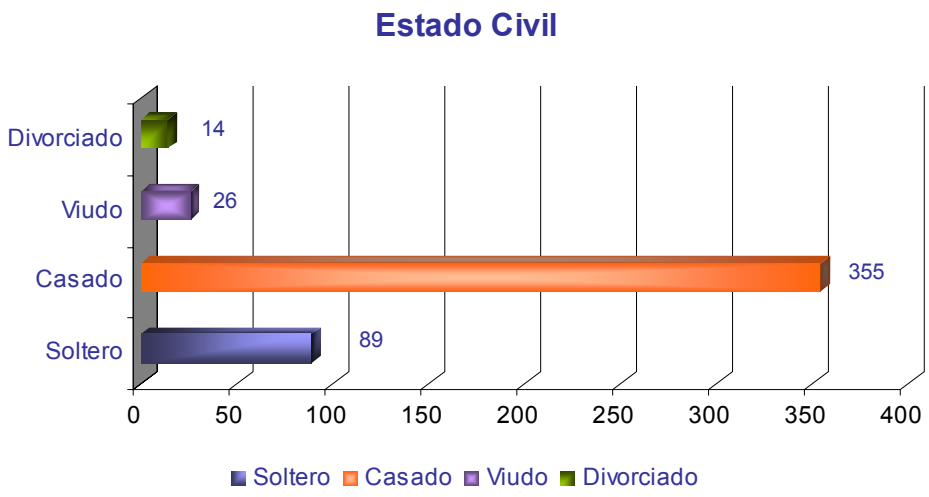
Por ejemplo, si sobre un grupo de clientes de una institución financiera tenemos medida la variable estado civil con la siguiente información:

Estado Civil

	No. Clientes	Porcentaje
Soltero	89	18%
Casado	355	73%
Viudo	26	5%
Divorciado	14	3%
Total	**484**	**100%**

Con los datos mostrados podemos representar la variable estado civil en forma gráfica de dos formas distintas:





En cualquiera de las dos gráficas utilizadas, vemos de manera inmediata que la presencia de la característica casado es predominante entre el grupo seleccionado, así como la poca presencia de personas con estado civil divorciado.

I.1.2 Variables cuantitativas

Con las variables cuantitativas se produce un fenómeno contrario al expuesto hasta este momento, ya que se las puede expresar numéricamente. Una primera clasificación, basada en el tipo de valores que puede tomar, permite distinguir entre variables cuantitativas discretas y continuas. El primer grupo es el resultado de conteos y, por tanto, toman sólo valores enteros, mientras que el segundo grupo resultan de mediciones y pueden contener cifras decimales.

Variables discretas: Pueden ser por ejemplo el número de lavadoras producidas por una empresa en un año (100, 20, 3.476), el número de hijos de una pareja (0, 1, 3), el número de pagos a realizar por un préstamo o en una compra a plazo de un artículo o un bien (12, 24, 120), etc.

Variables continuas: Son aquellas cuyo valor puede ser cualquier cantidad en un intervalo, así por ejemplo la temperatura (38,53 ºC, 121,3 ºF), el peso (112,8 Kg., 245,66 lb.), la altura de una persona (1,70 m., 182,45 cm.) o la superficie de las viviendas (124,5 m^2).

Al igual que las variables cualitativas, las cuantitativas son susceptibles de ser medidas mediante una escala específica. Si se encuentran variables cuantitativas discretas con un número pequeño de valores se tratarían de manera similar a las variables cualitativas antes descritas.

Escala numérica: Establece intervalos entre cada número observado de la característica cuantitativa y éste es asociado a un significado que permite establecer comparaciones. Así por ejemplo, si el sueldo de dos personas son tres mil y mil dólares respectivamente, se puede concluir que el primero gana tres veces más que el segundo, por tanto, tiene un mayor nivel de ingresos.

Las variables cuantitativas generalmente van acompañadas por medidas descriptivas, que permiten tener información complementaria de la variable y su distribución en sí, expresado usualmente por diagrama de barras.

Cargas Familiares

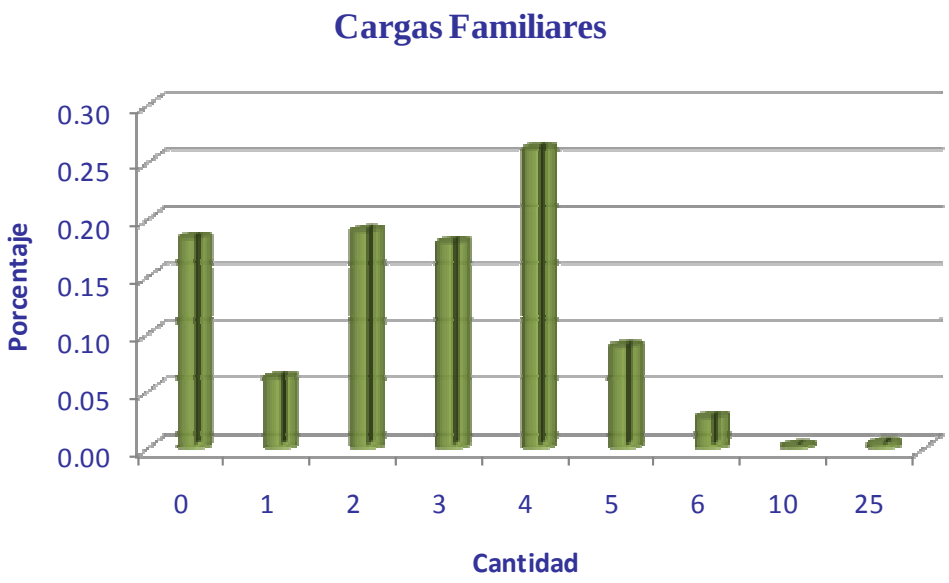
Tabla de Frecuencias

Cantidad	No. Clientes	Porcentaje	Porc. Acumulado
0	89	18,4%	18,4%
1	30	6,2%	24,6%
2	92	19,0%	43,6%
3	87	18,0%	61,6%
4	127	26,2%	87,8%
5	43	8,9%	96,7%
6	13	2,7%	99,4%
10	1	0,2%	99,6%
25	2	0,4%	100,0%
Total	**484**	**100%**	

No. de Cargas Familiares

Estadísticas Descriptivas

Media	2,8
Mediana	3,0
Moda	4,0
Desv. Tip	2,2
Mínimo	0,0
Máximo	25,0
Curtosis	38,8
Pct. 25	2,0
Pct. 50	3,0
Pct. 75	4,0



En este ejemplo (diagrama anterior) podemos ver que de la muestra analizada de 484 clientes, el valor promedio de cargas familiares es de aproximadamente tres personas, pero el valor de la moda nos indica que cuatro cargas familiares es el valor más frecuente en el grupo analizado, Además se puede apreciar que hay un valor que sale del contexto del resto de datos, pues hay dos clientes que afirman tener 25 cargas familiares, que podría considerarse como un error atribuible a la captación o al ingreso de la información.

I.2 ANÁLISIS ESTADÍSTICO

Una vez que se logre obtener consistencia en la información contenida en la base de trabajo depurando datos aberrantes, perdidos y las cotas de los mismos acorde a las políticas institucionales, se determinan las variables que eventualmente traduzcan la esencia del modelo a construir; a pesar que no todos los datos se analizan por igual, ya que será necesario identificar si se tratan de variables nominales, ordinales, o numéricas, en cuyo caso el estudio cambia en gran medida.

Si tratamos de medir características cualitativas de la población donde no se han de establecer jerarquías, la forma de trabajar con este tipo de variables diferirá, por completo respecto a variables cuya característica cuantitativa establezca, por sí sola, un orden comparativo. Específicamente, no es lo mismo trabajar con la variable género de la población (hombres y mujeres), que con el nivel de ingresos mensuales de la misma ($100, $1.000, $5.000), por tanto, el tratamiento y análisis previos para estas dos variables será totalmente distinto.

I.2.1 Análisis univariante

Permite analizar el comportamiento de las variables por separado, sin cruzar información de otras variables. Busca seleccionar aquellos indicadores más discriminantes del caso para que vayan de acuerdo a la realidad.

Variables cualitativas: Una de las herramientas más empleadas para representar este tipo de variables es la distribución de frecuencias, que consiste en una tabla que presenta las categorías de una variable y sus repeticiones. Si tenemos una variable cualitativa al asar que toma *N* valores u observaciones que se agrupan

en *k* clases o categorías, se representan con letras minúsculas los datos n_1, n_2, n_3,…, n_k que aparecen en cada categoría *k*, bajo estos elementos se define:

- Frecuencia absoluta de la clase i-ésima (n_i): número de observaciones en la clase i.

- Frecuencia relativa de la clase i-ésima (f_i): es la proporción de datos en la clase i-ésima, es decir:

$$f_i = \frac{n_i}{N}.$$

- La suma de las *k* frecuencias relativas es igual a la unidad: $f_1 + f_2 + \ldots + f_k = 1$

La distribución de frecuencias permite comparar las frecuencias de las categorías en conjuntos de datos con distinto número de observaciones. Para entender mejor este concepto lo podemos explicar de la siguiente manera

En un conjunto de datos de 20 clientes, tomamos la variable nivel de estudios, la misma que pueden ser las siguientes categorías:

Código	**Categoría**
1	Sin estudios
2	Estudios primarios
3	Estudios medios
4	Estudios superiores

Los valores encontrados (observaciones) se encuentran codificados de la siguiente manera:

1 1 4 3 3 3 2 2 4 2 2 1 4 2 3 2 3 4 2 3

Frecuencias absolutas:

n_1= 3; n_2= 7; n_3= 6; n_4= 4

$$N = n_1 + n_2 + n_3 + n_4 = 3 + 7 + 6 + 4 = 20$$

Frecuencias relativas:

$$f_1 = \frac{3}{20} = 0{,}15; \quad f_2 = \frac{7}{20} = 0{,}35; \quad f_3 = \frac{6}{20} = 0{,}3; \quad f_4 = \frac{4}{20} = 0{,}2$$

$$f_1 + f_2 + f_3 + f_4 = 0{,}15 + 0{,}35 + 0{,}3 + 0{,}2 = 1$$

Distribución de frecuencias:

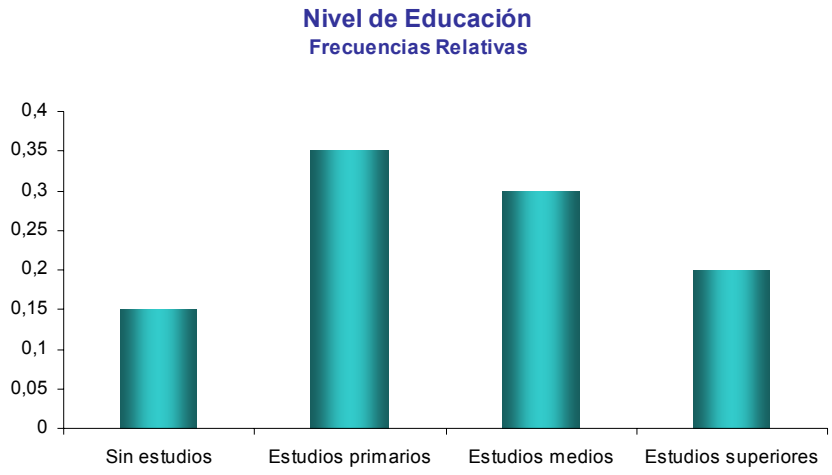
Variable:	Nivel de Estudios
No. de observaciones (N):	20
No. de categorías (k):	4

Categoría	Cod. Categoría	n_i	f_i
Sin estudios	1	3	15%
Estudios primarios	2	7	35%
Estudios medios	3	6	30%
Estudios superiores	4	4	20%
		20	**100%**

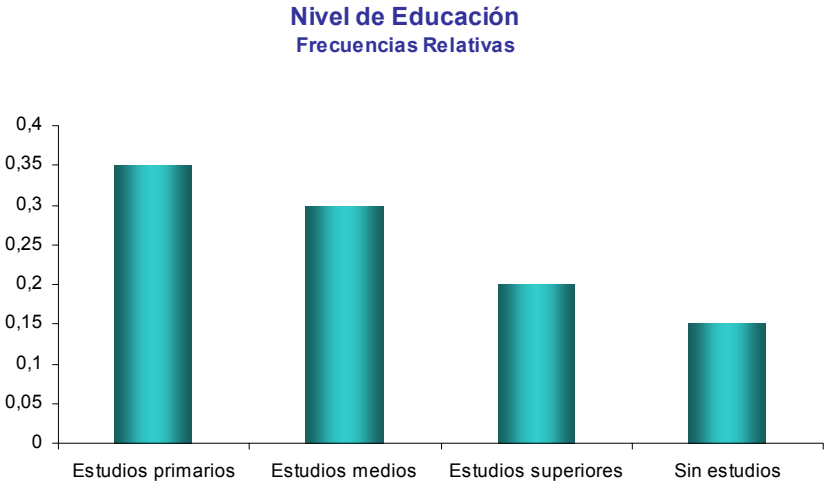
Según las frecuencias mostradas tenemos que en el grupo analizado la frecuencia más alta está asociada a la categoría estudios primarios; mientras que, la menos frecuente es la categoría sin estudios. Estas tablas de frecuencias tienen

diversas maneras de ser representadas gráficamente. A continuación, mencionaremos las más utilizadas

- Diagrama de barras: Permite visualizar de forma sencilla la distribución de una variable cualitativa. Se dibuja sobre cada categoría una barra (o rectángulo) cuya altura coincida con la frecuencia absoluta o relativa de dicha categoría. Por ejemplo, para la variable analizada tendríamos:



- Diagrama de Pareto: Es como un diagrama de barras descrito en el párrafo anterior, pero en este caso se ordenan las categoría de mayor a menor frecuencia (absoluta o relativa).



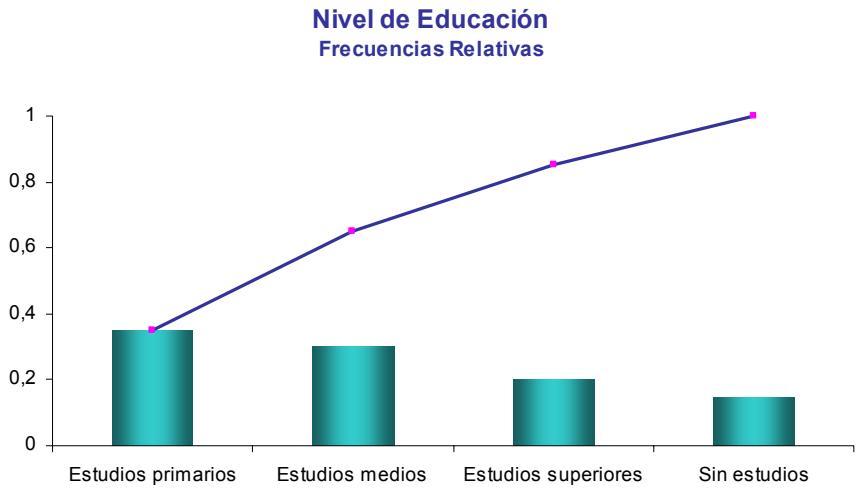
En la parte superior de la figura suele trazarse una línea que representa la suma de la frecuencia de cada categoría y las que la preceden:

f_2=0,35

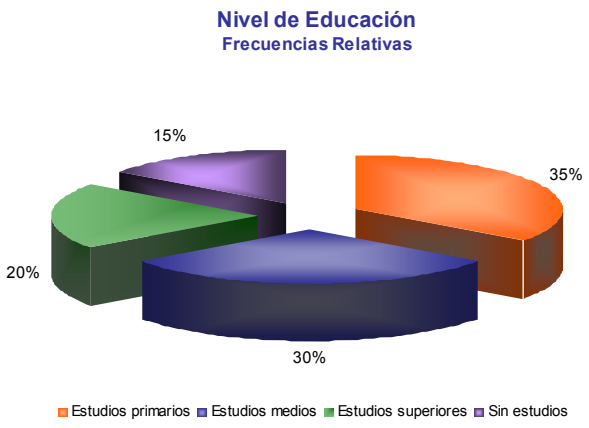
f_2+ f_3=0,35+0,3=0,65

f_2+ f_3+ f_4=0,35+0,3+0,2=0,85

f_2+ f_3+ f_4+f_1=0,35+0,3+0,2+0,15=1



- Pictograma: Consiste en un círculo en el que se representan sectores o porciones con áreas proporcionales a las frecuencias de cada una de las categorías.

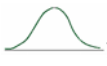
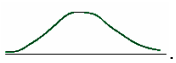


Otro de los elementos utilizados para analizar variables cualitativas es la denominada “moda” o clase modal que corresponde al dato o categoría con mayor frecuencia o número de repeticiones; A pesar que no sirve como una medida de descripción de una variable, ya que la moda puede no ser única. En el ejemplo que se ha venido desarrollando vemos que los estudios primarios también corresponden a la moda de los datos analizados.

Variables cuantitativas: Para este tipo de variables, dependiendo si se trata de variables discretas, el análisis es muy similar al aplicado para las variables cualitativas, es decir mediante el uso de los diagramas de frecuencia que fueron descritos con anterioridad. Para el caso de variables continuas se puede incorporar la utilización de histogramas y polígonos de frecuencia, diagramas de caja, tallo y hoja. Adicionalmente este tipo de variables deben ser analizadas a través de medidas de posición central, dispersión, asimetría y curtosis[2].

- Histogramas de frecuencias: Las clases o rangos de las variables continuas no están dados en forma implícita como en las variables cualitativas o en las discretas; por tanto, es necesario construirlas manualmente, para ello, se divide el conjunto de posibles valores de la variable en intervalos que no se intersequen o se solapen; aquí se puede

---

2 Se define curtosis como la medida o grado de apuntamiento de una curva de distribución respecto a un estándar, mide la mayor o menor concentración de datos alrededor de la media, pudiendo darse el caso de una curva muy puntiaguda o leptocúrtica  , medianamente puntiaguda o mesocúrtica  y platicúrtica o curva completamente aplastada  .

identificar un punto central de cada intervalo (marca de categoría $\mathbf{c_i}$), luego se puede proceder al igual que en las variables cualitativas.

Por ejemplo, los datos que se muestran a continuación, corresponden a los egresos familiares mensuales a un determinado grupo de 75 clientes:

No. Obs.	Egreso Familiar ($)	No. Obs.	Egreso Familiar ($)	No. Obs.	Egreso Familiar ($)	No. Obs.	Egreso Familiar ($)	No. Obs.	Egreso Familiar ($)
1	81,86	**16**	142,68	**31**	90,5	**46**	531,10	**61**	152,08
2	105,63	**17**	510,22	**32**	89,5	**47**	475,76	**62**	228,81
3	110,69	**18**	158,83	**33**	466,9	**48**	316,50	**63**	76,92
4	134,25	**19**	278,85	**34**	87,1	**49**	279,59	**64**	255,20
5	226,18	**20**	168,62	**35**	309,8	**50**	48,59	**65**	241,99
6	273,87	**21**	176,20	**36**	247,4	**51**	96,67	**66**	417,10
7	142,38	**22**	179,11	**37**	427,8	**52**	256,55	**67**	752,44
8	309,96	**23**	113,07	**38**	195,7	**53**	514,33	**68**	352,71
9	101,43	**24**	876,16	**39**	257,6	**54**	161,60	**69**	259,47
10	276,27	**25**	64,43	**40**	176,7	**55**	228,37	**70**	225,39
11	662,80	**26**	112,35	**41**	285,9	**56**	638,37	**71**	174,34
12	493,73	**27**	255,47	**42**	450,6	**57**	442,16	**72**	308,71
13	308,79	**28**	321,31	**43**	56,3	**58**	65,06	**73**	455,13
14	254,42	**29**	434,38	**44**	306,5	**59**	160,58	**74**	122,70
15	172,93	**30**	707,44	**45**	156,8	**60**	197,39	**75**	479,79

Elaboración: Propia
Fuente: Bdd APEVSC

Se puede estructurar rangos o intervalos de tamaño igual a $100 para construir la distribución de frecuencias, de este modo, el primer rango corresponderá a los egresos mensuales inferiores a cien ($100) dólares, cuya marca de categoría c_1 es igual a cincuenta dólares ($50), procediendo igual con el resto de datos se pueden encontrar nueve categorías diferentes (k = 9).

Rangos de Clase	*Frecuencias absolutas*	*Frecuencias absolutas acumuladas*	*Frecuencias relativas*	*Frecuencias relativas acumuladas*
(Gasto en $)	n_i	N_i	f_i	F_i
100	10	10	13,33%	13,33%
200	22	32	29,33%	42,67%
300	17	49	22,67%	65,33%
400	8	57	10,67%	76,00%
500	10	67	13,33%	89,33%
600	3	70	4,00%	93,33%
700	2	72	2,67%	96,00%
800	2	74	2,67%	98,67%
900	1	75	1,33%	100,00%

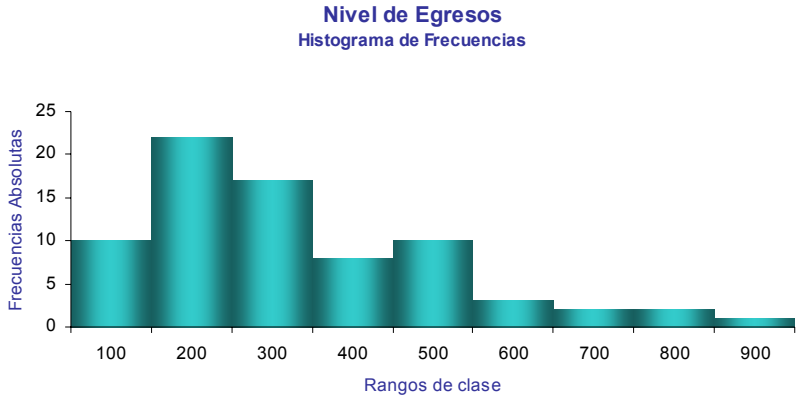
De la tabla de frecuencias generadas, se puede extraer aquellos clientes que tienen un egreso familiar inferior a $200 es igual al 42,67%, mientras que aquellos que gastan más de $600 mensuales representan apenas el 6,67% del total evaluado (1 – 93,33%). La proporción de clientes que han declarado gastar entre $100 y $300 corresponde al 52%. Con los datos obtenidos, podemos elaborar un histograma, que no es más que la representación de las frecuencias mediante áreas, para ello, sobre cada rango o clase se levanta un rectángulo, cuya área representa la frecuencia o número de observaciones de esa clase.

- Cuando las clases (o intervalos) en que dividimos los datos son de distinta longitud, el eje vertical no tiene sentido. Como la frecuencia es el área de cada rectángulo, si dibujamos rectángulos con distinta base su mayor o menor altura no nos da información.

- Cuando las clases son de la misma longitud, las frecuencias son proporcionales a las alturas de los rectángulos. La altura nos informa sobre la densidad o concentración de datos en ese intervalo.

- Si los rectángulos son más altos hay más datos de la variable.

- Si los rectángulos son más bajos los datos de la variable son más escasos.



Los rectángulos se dibujan en forma contigua (a diferencia del diagrama de barras o de Pareto) para transmitir la idea de variable continua. La forma del histograma es la misma si representamos frecuencias absolutas o relativas, sólo cambia la escala del eje vertical. La forma del histograma depende de

- El ancho de las clases o tamaño de los intervalos.
- Elección del punto donde empieza la primera clase

Para la selección del número de intervalos para un histograma, es preferible empezar con pocos grupos y a medida que se quieran aumentar éstos se debe verificar que exista un aumento de información. Si tenemos N observaciones se puede elegir el número de familias como el entero más próximo a $\sqrt{N}$. En nuestro ejemplo tenemos N=75 clientes, entonces $\sqrt{N} = \sqrt{75} = 8{,}6 \approx 9$.

La forma del histograma refleja las propiedades y características que tiene la variable, las que se pueden apreciar visualmente. Por ejemplo, se puede ver si la variable presenta algún tipo de simetría, si presenta un solo valor máximo (pico o

moda) o en su defecto, tiene varias modas; al respecto, podemos mencionar las formas más frecuentes de los histogramas

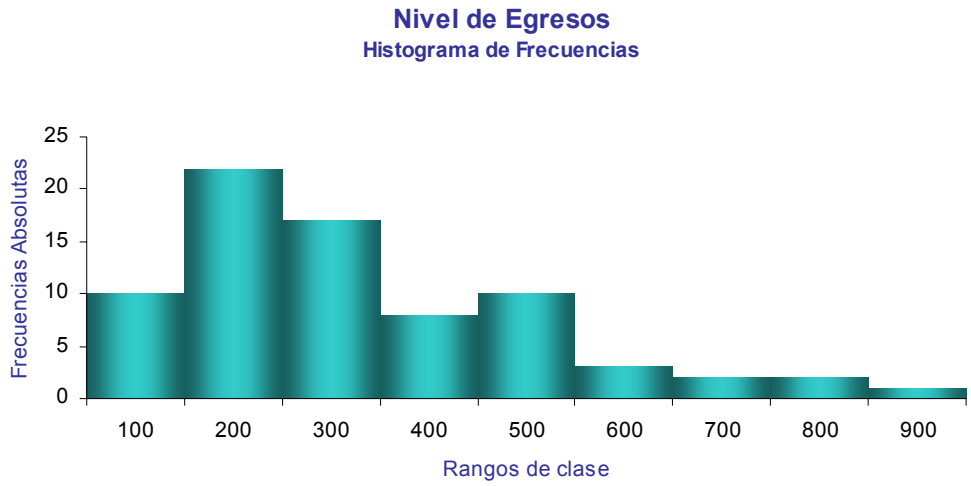
- Unimodal simétrico: se suele dar en variables en las que hay una gran cantidad de observaciones con valores intermedios y algunos valores en ambos extremos (notas, peso, altura…).

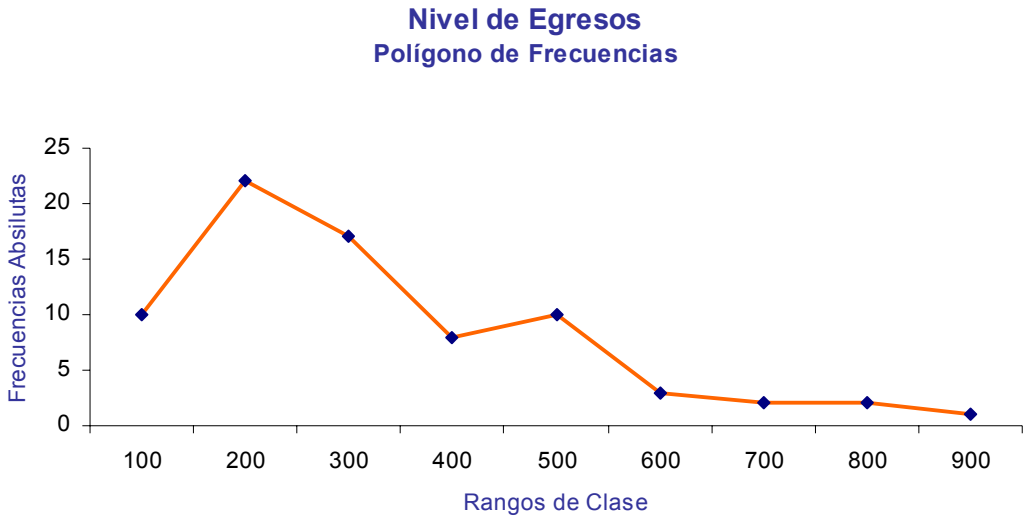
- Unimodal asimétrico a la derecha: se da en variables que tienen una gran cantidad de observaciones pequeñas o intermedias y algunos datos grandes (gasto, ingreso…).

- Unimodal asimétrico a la izquierda: variables con muchas observaciones de valor alto o intermedio (esperanza de vida en los distintos países).

- Bimodal simétrico: suele aparecer cuando los datos son de 2 grupos heterogéneos y conviene estudiarlos por separado (un objeto que se hiciera en dos tamaños distintos en cantidades iguales).

- Polígono de Frecuencia: consiste en una representación gráfica de las frecuencias de una variable, similar al histograma y se obtiene al unir los centros de la base superior de los rectángulos del histograma.





Tanto el histograma como el polígono de frecuencias pueden determinarse en forma acumulada, la diferencia fundamental de éstas dos representaciones es la forma más suavizada que proporciona el polígono.

Otra de las herramientas del análisis univariante que permiten conocer más sobre una variable cualitativa es el Diagrama de Tallo y Hojas, el cual permite obtener en forma simultánea la distribución de frecuencias de la variable y su representación gráfica. Para construirlo hay que separar en cada dato el último dígito de la derecha (la hoja) del resto de las cifras (el tallo). De este modo los tallos aparecen a la izquierda de una línea vertical y a la derecha de cada uno anotamos las cifras finales (hojas) de todos los datos de cada rango o clase.

Si al grupo de 75 clientes que se analizó los egresos familiares, tabulamos sus edades, tenemos la siguiente tabla:

No. Obs.	Edad	No. Obs.	Edad	No. Obs.	Edad	No. Obs.	Edad	No. Obs.	Edad
1	18	**16**	18	**31**	75	**46**	19	**61**	76
2	80	**17**	84	**32**	74	**47**	76	**62**	64
3	76	**18**	71	**33**	67	**48**	31	**63**	62
4	31	**19**	32	**34**	61	**49**	35	**64**	54
5	35	**20**	36	**35**	57	**50**	29	**65**	53
6	28	**21**	28	**36**	56	**51**	28	**66**	48
7	23	**22**	25	**37**	52	**52**	27	**67**	46
8	23	**23**	23	**38**	47	**53**	61	**68**	42
9	68	**24**	67	**39**	43	**54**	65	**69**	42
10	40	**25**	40	**40**	40	**55**	41	**70**	35
11	44	**26**	44	**41**	39	**56**	42	**71**	30
12	45	**27**	46	**42**	35	**57**	47	**72**	29
13	57	**28**	57	**43**	28	**58**	57	**73**	28
14	57	**29**	51	**44**	25	**59**	55	**74**	27
15	51	**30**	52	**45**	19	**60**	53	**75**	19

Elaboración: Propia
Fuente: Bdd APEVSC

Aplicando el procedimiento señalado para los datos mostrados tenemos el siguiente diagrama de tallo y hojas:

1	8 8 9 9 9
2	3 3 3 5 5 7 7 8 8 8 8 8 9 9
3	0 1 1 2 5 5 5 5 6 9
4	0 0 0 1 2 2 2 3 4 4 5 6 6 7 7 8
5	1 1 2 2 3 3 4 5 6 7 7 7 7 7
6	1 1 2 4 5 7 7 8
7	1 4 5 6 6 6
8	0 4

Al igual que el histograma o el polígono de frecuencias, este tipo de diagramas proporciona una impresión visual del número de observaciones de cada clase, con la ventaja de que al darnos un mayor detalle nos permite recuperar los datos, lo que no puede hacerse con el histograma o el polígono.

- Diagrama de caja: Los diagramas de caja proporcionan información visual completa referente a la distribución de los datos. Pueden ser de gran utilidad como técnica de análisis exploratorio de datos, ya que nos proporcionan información sobre la mediana (o media), sobre el 50% y 90%

de los datos, sobre la existencia de valores atípicos, así como, de la simetría de la distribución, su construcción se realiza de la siguiente manera:

- Se ordenan los datos de la muestra y se obtiene el valor mínimo, el máximo, y los tres cuartiles Q_1, Q_2 y Q_3.

- Dibujar un rectángulo cuyos extremos son Q_1 y Q_3 e indicar la posición de la mediana, Q_2, mediante una línea.

- Calcular unos límites admisibles inferior y superior, L_i y L_s, que identifiquen a los valores atípicos.

- Considerar como valores atípicos los situados fuera del intervalo (L_i, L_s).

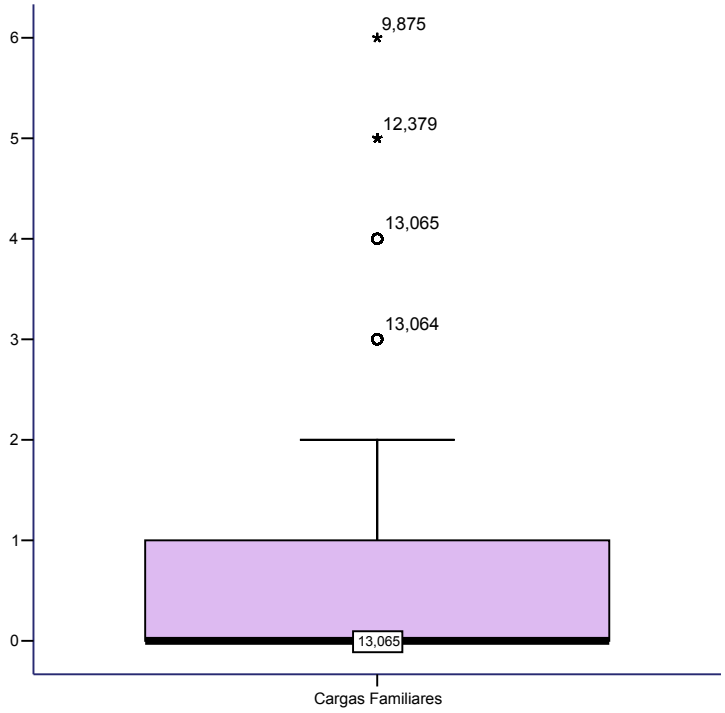
- Dibujar una línea que va desde cada extremo del rectángulo central hasta el valor más alejado no atípico, es decir, que está dentro del intervalo (L_i, L_s).

- Identificar todos los datos que están fuera del intervalo (L_i, L_s), marcándolos como atípicos.

Es posible introducir algunas variaciones en la construcción de estos diagramas, dependiendo del tipo de estudio y de la información disponible[3]. La caja o rectángulo contiene un porcentaje de la muestra y puede construirse con diferentes rangos de variación, como el 80% de los datos y ser cortada por la media; sin embargo, lo más usual es que sea cortada por la mediana, de este modo se tiene

3 Antonio Calvo-Flores y Antonio Arques Pérez, *Modelos Estadísticos Teóricos*, Facultad de Economía y Empresa, Universidad de Murcia, p. 12.

de antemano conocimiento del comportamiento del 50% de la población en estudio sobre una variable específica. Los diagramas de caja proporcionan una idea intuitiva de la simetría de la distribución de los datos; si la media no está en el centro del rectángulo, eso significa que la distribución no es simétrica, conociendo además a qué lado se escora o desvía.



Complementariamente al análisis gráfico desarrollado para las variables cuantitativas, es conveniente estudiar los estadísticos descriptivos de las diferentes variables agrupadas en medidas de posición y de variabilidad. Las medidas de posición (tendencia central) forman parte de las medidas descriptivas numéricas, cuya función es darnos la orientación del conjunto de datos. Por su parte las medidas de variabilidad se encargan de proporcionarnos información correspondiente a la dispersión de los datos, puesto que varios conjuntos de datos pueden presentar iguales valores promedios pero diferente variabilidad.

- Media Aritmética: o simplemente media, es la medida de posición más utilizada, representa el centro físico del conjunto de datos y se define como la suma de todos los posibles valores observados, ponderada por el total de observaciones registradas. Si x_1, x_2,…, x_n son n observaciones numéricas, entonces la media aritmética de dichas observaciones se define de la siguiente manera:

  Es decir, si la tabla de valores de una variable X es

$$\bar{x} = \frac{x_1 + x_2 + x_3 + \cdots + x_n}{n}$$

  O lo que es lo mismo:

$$\bar{x} = \frac{\sum_{i=1}^{n} x_i}{n}$$

- Desviación Típica: es una medida de variabilidad de mayor uso se determina mediante la raíz cuadrada de la varianza, definida como la media de las diferencias cuadráticas de n observaciones respecto a su media aritmética y se calcula como:

$$s^2 = \frac{\sum_{i=1}^{n} \left(x_i - \bar{x}\right)^2}{n}$$

En la varianza al tomar el cuadrado de las desviaciones se obtienen unidades al cuadrado, para evitar que se magnifique dicha diferencia real se establece la desviación estándar (o desvío típico) como sigue:

$$s = \sqrt{\frac{\sum_{i=1}^{n}\left(x_i - \bar{x}\right)^2}{n}}$$

Por ejemplo si tomamos los siguientes datos:

44, 59, 36, 55, 47, 61, 53, 32, 65, 51

Podemos obtener su media aritmética de la siguiente manera:

$$\bar{x} = \frac{44 + 59 + 36 + 55 + 47 + 61 + 53 + 32 + 65 + 51}{10}$$

$$\bar{x} = \frac{503}{10} = 50{,}3$$

La varianza de los datos presentados estará dada por:

$$s^2 = \frac{\sum_{i=1}^{10}\left(x_i - \bar{x}\right)^2}{10}$$

$$\left(x_1 - \bar{x}\right)^2 = \left(44 - 50{,}3\right)^2 = 39{,}69$$

$$\left(x_2 - \bar{x}\right)^2 = \left(59 - 50{,}3\right)^2 = 75{,}69$$

$$\left(x_3 - \bar{x}\right)^2 = \left(36 - 50{,}3\right)^2 = 204{,}49$$

$$\left(x_4 - \bar{x}\right)^2 = \left(55 - 50{,}3\right)^2 = 22{,}09$$

$$\left(x_5 - \bar{x}\right)^2 = \left(47 - 50{,}3\right)^2 = 10{,}89$$

$$\left(x_6 - \bar{x}\right)^2 = \left(61 - 50{,}3\right)^2 = 114{,}49$$

$$\left(x_7 - \bar{x}\right)^2 = \left(53 - 50{,}3\right)^2 = 7{,}29$$

$$\left(x_8 - \bar{x}\right)^2 = \left(32 - 50{,}3\right)^2 = 334{,}89$$

$$\left(x_9 - \bar{x}\right)^2 = \left(65 - 50{,}3\right)^2 = 216{,}09$$

$$\left(x_{10} - \bar{x}\right)^2 = \left(51 - 50{,}3\right)^2 = 0{,}49$$

$$s^2 = \frac{1026{,}10}{10} = 102{,}61$$

Y su desviación estándar se calcularía así:

$$s = \sqrt{102{,}61} = 10{,}13$$

Como se mencionó, tanto la media como la desviación estándar nos permiten conocer de mejor manera las características de un conjunto de datos, los cuales, en el caso de entidades financieras, podrían estar representados por los montos de los préstamos concedidos durante un tiempo específico; los plazos de concesión, las edades de los sujetos de crédito, la morosidad de dichos créditos, los ingresos y/o egresos de una persona dedicada al micro crédito que solicita un producto crediticio.

Cuando tratamos de datos es común hablar de las medidas descriptivas que los caracterizan, tales como la media, su desviación estándar y la moda. Pero cuando ya se empieza a hablar de variables aleatorias se procura homologar dichas medidas descriptivas hacia las distribuciones o el comportamiento que pueden seguir dichas variables.

- Valor Esperado: Sea X una variable aleatoria (v.a.) discreta que toma los valores x_1, x_2,..., x_n y cuya función de probabilidad es p_1, p_2,..., p_n respectivamente. Se define el valor esperado de X, como:

$$\mu_x = E[X] = x_1 p_1 + x_2 p_2 + x_3 p_3 + \cdots + x_n p_n$$

$$\mu_x = E[X] = \sum_{i=1}^{n} x_i P(x_i)$$

Si la v.a. X es de tipo continuo, con función de densidad f(x), definimos el valor esperado E(X), como:

$$\mu_x = E[X] = \int_{-\infty}^{+\infty} x f(x) dx$$

La expresión de E(X) en el caso que X sea una v.a. discreta, este valor es la media ponderada de los posibles valores que puede tomar la variable X, en donde los pesos o ponderaciones son las probabilidades, $P(x_i) = P(X = x_i)$, de ocurrencia de los posibles valores de X. Luego el valor esperado de X se interpreta como una media ponderada de los posibles valores de X, y no como el valor que se espera que tome X, pues puede suceder que E(X) no sea uno de los posibles valores de X. En el caso de v.a. continua, E(X) nos indica el centro de la función de densidad, es decir, nos indica el centro de gravedad de la distribución.

Propiedades

- La esperanza de una constante es la propia constante. Es decir si k es una constante entonces:

$$E(k) = k$$

- Si una v.a. X y k una constante, entonces:

$$E(k.X) = k.E(X)$$

- Si una v.a. X y k una constante, entonces:

$$E(X + k) = E(X) + k$$

- Si una v.a. X esta acotada, es decir existen dos valores a y b tales que $a \leq X \leq b$, entonces se verifica que:

$$a \leq E(X) \leq b$$

- Si X y Y son variables aleatorias, entonces:

$$E(X + Y) = E(X) + E(Y)$$

- Sea X y Y variables aleatorias; a, b constantes cualesquiera, entonces:

$$E[a \cdot X + b \cdot Y] = a \cdot E[X] + b \cdot E[Y]$$

- No se cumple que:

$$E[X^2] = (E[X])^2$$

- **Varianza:** Sea X una v.a. que toma los valores $x_1, x_2,..., x_n$ y cuya función de probabilidad es $p_1, p_2,..., p_n$ respectivamente. La varianza de una distribución se denota y define así:

$$\sigma_x^2 = Var[X] = E[(X - \mu_x)^2]$$

También puede escribirse como

$$\sigma_x^2 = V[X] = E\left([X - E(X)]^2\right)$$

Es una medida de dispersión de los valores de la variable respecto de su media, y nos permite conocer el grado de separación de los valores de la distribución, pudiendo realizar comparaciones con otras distribuciones. La

varianza se expresa en las mismas unidades que la variable X, pero al cuadrado. La desviación estándar o desviación típica, se expresa en las mismas unidades de medida que la variable X.

Propiedades:

- La varianza no puede ser negativa
- La varianza de una constante k es cero.

$$Var(k) = 0$$

Sea X una v.a. cuya varianza existe. Entonces:

$$Var(X) = E[X^2] - (E[X])^2$$

- Sea X una v.a. cuya varianza existe, y k una constante cualquiera. Entonces:

$$Var\ (k.X) = k^2\ .Var(X)$$

- Sea X una v.a. cuya varianza existe y a, b dos constantes cualesquiera. Entonces:

$$Var(\ aX + b) = a^2.Var(\ X)$$

- Sean X e Y dos v.a. independientes cuyas varianzas existen, entonces se verifica que la varianza de la suma o de la diferencia de ambas v.a. independientes es igual a la suma de las varianzas. Es decir:

$$Var\ (X \pm Y) = Var\ (X) + Var(Y)$$

Si las v.a. no son independientes entonces:

$$E[(X-E(X))(Y-E(Y))] = Cov(X,Y)$$

Y se verificará que:

$$Var(X \pm Y) = Var(X) + Var(Y) \pm 2Cov(X,Y)$$

En términos generales, el valor medio representa las concentraciones de los valores observados de una variable y permite a priori tener una idea de la composición de los datos. Además mediante la desviación estándar veremos el distanciamiento del resto de observaciones respecto al valor medio, es decir, si en la variable analizada existe distanciamiento o agrupamiento de datos. Estas dos primeras conclusiones se ven complementadas con los valores máximos y mínimos, pues a más de acotar el conjunto de valores de la variable nos dan un indicio si dichos valores (máximo y mínimo) deben o no ser considerados en el análisis a fin de evitar distorsiones por información mal recabada.

I.2.2 Análisis bivariante

Al igual que en el caso univariante, se puede distinguir el análisis a emplear, dependiendo si las variables son cualitativas o cuantitativas, es necesario mencionar que en esta parte del análisis deja de ser tan intuitivo como en el análisis preliminar.

Variables cualitativas: Una de las principales herramientas utilizadas para el análisis bivariante de variables cualitativas son las tablas de contingencia.

- Tablas de Contingencia: Son tablas de frecuencias conjuntas en donde cada entrada representa un criterio de clasificación, que da como resultado que las frecuencias aparecen organizadas en casillas que contienen información sobre la relación existente entre los criterios que conforman la tabla. Las tablas de contingencia según la cantidad de variables que clasifique se denominan bidimensionales o de 2 x 2, cuando la cantidad de

variables son dos, tridimensionales o de 3 x 3, cuando la cantidad es 3, hasta llegar a las multidimensionales o de n variables.

Si consideramos dos variables cualitativas, la notación estándar de las tablas de contingencias es la siguiente:

n_{ij} : frecuencia observada en la casilla formada por la combinación del nivel i de la variable *X*, el nivel j de la variable *Y*

i: 1,2,….I (I= número de categorías de la variable X)

j: 1,2,….J (J= número de categorías de la variable Y)

Así por ejemplo, la frecuencia n_{12} se refiere a la frecuencia de la casilla resultante de combinar la categoría 1 de la variable *X* y la categoría 2 de la variable *Y*. La notación aquí utilizada corresponde a una tabla de contingencia bidimensional pero también resulta aplicable para tablas de más de dos variables. Por ejemplo, para una tabla tridimensional la notación será n_{231} y se refiere a la frecuencia resultante de combinar la categoría 2 de la variable *X*, la categoría 3 de la variable *Y*, así como la categoría 1 de la variable *Z*.

La utilidad de las tablas de contingencia no es únicamente la obtención de las frecuencias conjuntas de las variables analizadas, sino que permiten la aplicación de estadísticos para estudiar las posibles pautas de asociación existentes entre dichas variables.

El grado de asociación existente entre dos variables categóricas no puede ser establecido simplemente observando las frecuencias de una tabla de contingencia. Para determinar si dos variables se encuentran relacionadas necesitamos utilizar

algún índice de asociación acompañado de su correspondiente prueba de significación. La prueba Chi-cuadrado (χ^2) de Pearson proporciona un estadístico que permite contrastar la hipótesis que dos criterios de clasificación utilizados sean independientes; es decir, podemos establecer si dos variables cualitativas son independientes entre sí. Para ello se comparan las frecuencias obtenidas (frecuencias observadas) con las frecuencias que teóricamente deberíamos haber encontrado en cada casilla si las dos variables fueran independientes.

Las frecuencias esperadas se estiman de la siguiente forma:

$$\widehat{m}_{ij} = \frac{n_{i+} n_{+j}}{n}$$

Donde i se refiere a una fila cualquiera, j a una columna cualquiera e ij a una casilla cualquiera.

Una vez obtenidas las frecuencias esperadas para cada casilla el estadístico χ^2 se calcula de la siguiente forma:

$$\chi^2 = \sum_i \sum_j \frac{\left(n_{ij} - \widehat{m}_{ij}\right)^2}{m_{ij}}$$

Donde n_{ij} se refiere a las frecuencias observadas y $\widehat{m}_{ij}$ se refiere a las esperadas. De la ecuación se desprende que el estadístico χ^2 valdrá cero cuando las variables sean completamente independientes, ya que las frecuencias esperadas coincidirán con las observadas, y que será tanto mayor cuanto mayor sea la discrepancia entre las frecuencias observadas y las esperadas, señalando la inexistencia de independencia entre las variables analizadas.

El estadístico χ^2 resulta de utilidad para establecer si existe asociación o no entre variables categóricas, pero no nos permite establecer el grado o fuerza de asociación entre dichas variables. Esta limitación es consecuencia directa de que su valor no solo depende del grado en que los datos se ajustan al modelo de independencia sino del número de casos que consta la muestra, ya que con tamaños muestrales muy grandes, diferencias relativamente pequeñas entre las frecuencias observadas y las esperadas pueden dar lugar a valores de Chi-cuadrado muy altos. Es por ello que para estudiar el grado de asociación entre dos variables categóricas se utilizan índices o medidas que intentan cuantificar ese grado de asociación, eliminando el efecto del tamaño muestral.

La selección de una medida de asociación concreta deberá tener en cuenta el tipo de variable analizada (ordinal o nominal) y la hipótesis que se intenta contrastar (independencia). Así, para analizar variables cualitativas nominales tenemos las siguientes medidas[4]:

- Coeficiente de contingencia.
- Lambda.
- Tau.
- Coeficiente de incertidumbre.

Con respecto a las variables cualitativas ordinales tenemos los siguientes estadísticos:

---

4 Ana María Aguilera del Pino, *Tablas de Contingencia Bidimensional*, Ed. La Muralla, S.A., Madrid, 2001, pp. 20-22.

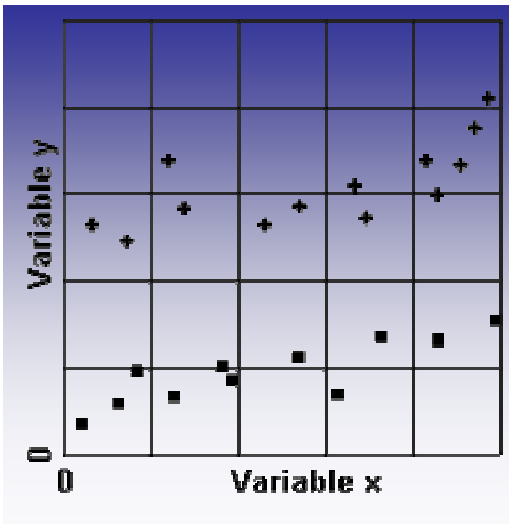
- Gamma.
- Tau-b de Kendall.
- Tau-c de Kendall.

Variables cuantitativas: Para el caso de variables cuantitativas fundamentalmente nos apoyamos en el análisis de correlación y de covarianza existente entre las variables a analizar. Por ejemplo en educación se ha comprobado la relación entre las notas de lenguaje y matemática, la fortaleza de personas altas respecto a las de menor estatura, la relación entre los precios de venta de los productos de primera necesidad, respecto a su disponibilidad.

- Correlación: mide la relación entre dos variables y su sentido (si es directo o inverso), cuando dicha relación es perfectamente lineal, dicho coeficiente vale 1 (ó -1), cuando el coeficiente tiene un valor próximo a cero, se puede afirmar que o bien, no existe relación entre las variables analizadas o bien, dicha relación no es lineal. La correlación está ligada directamente al concepto de covarianza, y podemos encontrar varias formas de escribir su ecuación entre las que tenemos:

$$\rho_{XY} = \frac{Cov(X,Y)}{\sigma_X \sigma_Y}$$

$$\rho(X,Y) = \frac{Cov(X,Y)}{\sqrt{Var(X)*Var(Y)}}$$



$$r = \frac{\sigma_{xy}^2}{\sigma_x \sigma_y}$$

De dichas ecuaciones vemos claramente que para la determinación de la correlación, primero hay que calcular la covarianza entre las variables, pese a que su significado intuitivo es más complejo que el del coeficiente de correlación.

- Covarianza: La fórmula que expresa la covarianza entre dos variables es la siguiente:

$$\sigma_{xy}^2 = \frac{1}{N}\sum_{i=1}^{n}\left(x_i - \bar{x}\right)\left(y_i - \bar{y}\right)$$

Donde:

x_i es la observación de la variable *X*,

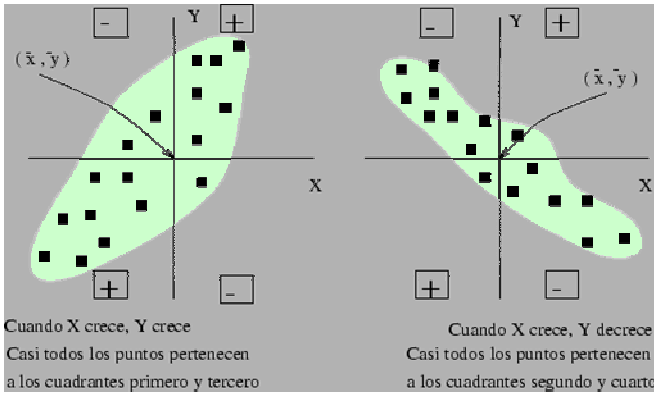
y_i corresponde a la observación de la variable *Y*,

Además de sus correspondientes valores medios $(\bar{x}, \bar{y})$.

Una covarianza positiva significa que existe una relación lineal positiva entre las dos variables; es decir, valores bajos de la primera variable *X* se asocian con valores bajos de la segunda variable *Y*, y viceversa. Una covarianza negativa significa que existe una relación lineal inversa perfecta entre las dos variables, lo que significa que valores bajos en *X* se asocian con los valores

altos en *Y*, mientras los valores altos en *X* se asocian con los valores bajos en *Y*.

Una covarianza igual a 0 se interpreta como la no existencia de una relación lineal entre las dos variables estudiadas. Una aproximación gráfica al significado de la covarianza es la que se muestra en la siguiente gráfica:



Desde el punto de vista técnico la relación entre dos variables está dada por el coeficiente de correlación de Pearson que cumple las siguientes características.

- El índice de correlación de Pearson no puede valer menos de -1 ni más de +1.
- El índice de correlación de Pearson (en valor absoluto) no varía cuando se transforman linealmente las variables.
- Es importante señalar que correlación no implica causación, es decir que si dos variables están altamente correlaciones no implica que *X* causa *Y* ni que *Y* causa *X*.
- Es importante indicar que el coeficiente de correlación de Pearson puede verse afectado por la influencia de terceras variables.
- Además el valor del coeficiente de Pearson depende en parte de la variabilidad de la muestra de trabajo o población en estudio.

II. CONSTRUCCIÓN DE UN MODELO CREDIT SCORE

Una vez que hemos realizado un recorrido por las variables y sus respectivas clasificaciones, se ha procurado detallar las herramientas o los análisis estadísticos necesarios para una adecuada interpretación de las variables a utilizar, ahora nos corresponde sintetizar los siguientes pasos que permitirán estructurar el modelo de credit score que coadyuve a la adecuada gestión de riesgo crediticio.

Integridad de la información

Uno de los más importantes elementos a tomar en consideración, a la hora de desarrollar un modelo de gestión de riesgo tipo *credit score*, es la calidad de la información con la que se trabaja y ello dependerá, en gran medida, de la manera como esta fue capturada y consolidada; si los datos con los que se va a trabajar fueron introducidos directamente por algún módulo de captura de información, diseñado por la institución, presentarán un mayor grado de confiabilidad y certeza, frente a aquella información proveniente de digitación de formularios.

La primera base de información que se puede solicitar debe contener campos correspondientes a fecha de concesión, identificadores de clientes, fechas de análisis con corte mensual, morosidad en días a la fecha de análisis, identificador del producto en estudio, montos otorgados; y, saldo a la fecha de análisis. Usualmente, se puede tomar toda la historia que posea la institución sobre el producto a modelar para estructurar esta base, esto dependerá fundamentalmente de la estabilidad del producto en términos de sus características intrínsecas, así como de las políticas que rigen su aprobación y seguimiento.

Un primer paso será la verificación de la consistencia de los campos solicitados, es deseable que todos los clientes incluidos en esta base contengan el historial de información requerida, es decir que si a un cliente X se le otorgó el producto el 15 de enero 2006 y la fecha de corte se determina el 30 de septiembre 2007, la base de datos debe contener información del cliente X al 31 de enero 2006, 28 de febrero 2006, 31 de marzo 2006 y así sucesivamente hasta llegar a septiembre 2007.

Definición de buenos y malos

Otro de los elementos en el que se debe tener mucho cuidado y debe ser discutido a profundidad, es aquel que tiene como objetivo clasificar al portafolio de clientes entre buenos y malos, que fundamentalmente depende del conocimiento de la cartera en mora dentro de la institución y del proceso de cobranzas entre otros. Por ejemplo, utilizar un mes de cartera en mora, no sería la mejor representación del estado del producto a modelar en cualquier institución financiera[5].

Evidentemente, la definición de "*Mal Cliente"* se ha de referir a aquellas cuentas o casos, que a base de su experiencia, la institución no quiere seleccionar para la generación de crédito debido a su expediente. Desde el punto de vista de la gestión de riesgo nos referimos a todos esos clientes a los que se les ha negado un producto crediticio, de haber sabido que no cumplirían la obligación contraída con la institución.

---

5 Lilian Simbaqueba, ¿*Qué es un Scoring*?, Instituto de Riesgo Financiero, 2004, p. 5.

Una de las definiciones que más comúnmente se utilizan para un cliente malo, está basada en la tasa de morosidad de 90 días, ya que hay una gran diferencia entre el desempeño de 60 y 90 días, mientras que los rangos entre 90 y 120 días son menos marcados, según el producto crediticio que se este analizando para la modelación. Otra técnica que se utiliza como alternativa consiste en establecer una matriz de contingencia para los períodos de morosidad promedio y morosidad máxima de la cartera de clientes, a fin de, obtener una primera clasificación de mal cliente al estructurar las intersecciones entre el tipo de cliente al que ya no deseo prestarle mis servicios.

Tabla de Contingencia Mora Promedio & Mora Máxima

Rangos Mora Promedio	Rangos Mora Máxima: Cero	De 1 a 15	De 15 a 30	De 30 a 45	De 45 a 60	De 60 a 75	De 75 a 90	De 90 a 180	De 180 a 360	Más de 360	Total general
Cero	29.291	250									**29.541**
De 1 a 15		1.342	2.625	760	578	272	68	27			**5.672**
De 15 a 30			24	25	94	85	271	342	15	1	**857**
De 30 a 45				10		6	18	352	31	34	**451**
De 45 a 690					11			237	43	29	**320**
De 60 a 75						9		75	102	14	**200**
De 75 a 90							2	10	116	1	**129**
De 90 a 180								23	230	32	**285**
De 180 a 360									30	37	**67**
Más de 360										28	**28**
Total general	**29.291**	**1.592**	**2.649**	**795**	**683**	**372**	**359**	**1.066**	**567**	**176**	**37.550**

La matriz diseñada, muestra los cruces entre morosidades máximas y promedios de los clientes, en función del apetito al riesgo de la institución, se puede establecer la siguiente definición de *mal cliente*:

- Cliente con una mora promedio superior o igual a 15 días y una mora máxima superior o igual a 30 días, ó

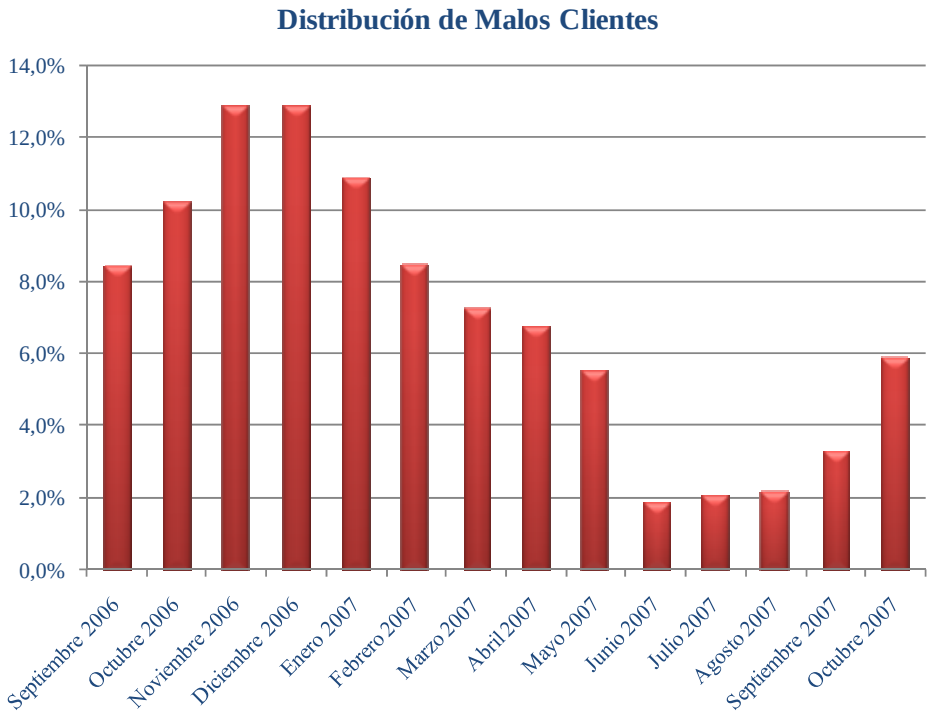
Período de exposición

Una vez establecidas las características asociadas a un mal cliente, se reclasifica la base de trabajo con esta definición, a fin de establecer la cantidad de clientes malos por fecha de concesión del producto a modelar (este proceso se

conoce como cosecha por morosidad). Un cuadro de cartera en mora respecto a las concesiones del producto de crédito mostrará cuándo la cartera en mora se ha estabilizado, punto en donde el período de exposición debe ser determinado este, es usualmente establecido entre 12 y 18 meses, dependiendo del producto. Por ejemplo, una tarjeta de crédito tenderá a madurar más rápidamente que un crédito hipotecario. El objetivo de este pequeño estudio probará si la muestra para el desarrollo del modelo es suficientemente madura para que el análisis sea válido.

Ventana de aplicación

Es importante escoger una ventana donde la tasa de morosidad o la cosecha se haya estabilizado y el conglomerado sea comparable con la actual población de clientes. Por ejemplo, podemos establecer el período de estabilidad de la base que corresponderá a las fechas de concesión, en las cuales no se evidencie una disminución abrupta del porcentaje de malos clientes, respecto al siguiente período.



En este gráfico mostrado vemos que un período de estabilidad de la base analizada, puede corresponder a las concesiones realizadas entre septiembre de 2006 y mayo de 2007, ya que, luego de dicha fecha se evidencia un cambio radical en la proporción de malos clientes, respecto a los meses mencionados. Acorde a lo referido en el párrafo anterior, se puede solicitar la generación de la base de modelación para las concesiones realizadas en los nueve meses que se han determinado como período de estabilidad del producto a modelar.

Modelo de clasificación

Uno de los modelos de mayor difusión y utilización a la hora de clasificar un grupo de clientes, es el modelo de regresión lineal[6] sea esta simple o múltiple. Para abordar el modelo que nos ocupa, es decir clasificar clientes entre buenos y malos (variable dependiente binaria), vemos que el modelo lineal presenta algunos problemas, lo cual nos llevará a usar modelos de regresión no lineales, que entre otras mantiene las siguientes características[7]:

- No atienden a suposiciones distribucionales en el mismo sentido que lo hace el análisis discriminante.
- La solución puede ser más estable si las variables explicativas o predictoras tienen una distribución normal multivariante.

---

6 Damodar N. Gujarati, *Essentials of Econometrics* , New York, McGraw Hill, 2004, p. 112.

7 Freddy H. Carranza V, *Clasificación de clientes mediante la aplicación de modelos econométricos a índices financieros*, Quito, UASB, 2006, p. 14.

- Adicionalmente, al igual que con otras formas de regresión, la multicolinealidad entre las variables puede conducir a estimaciones sesgadas y errores estándar inflados.
- Puede acondicionarse un modelo no lineal, para datos cuyas respuestas son ordinales o binarias, por medio del método de máxima verosimilitud.
- Como en el análisis discriminante los modelos no lineales ponderan las variables independientes y asigna una puntuación en una forma de probabilidad de incumplimiento para cada individuo.

Un modelo de regresión múltiple (no necesariamente lineal) nos permite explicar el comportamiento de una variable dependiente *Y* en función de una serie de variables independientes X_1, $X_{2,...}$, X_k y de un término de perturbación *u*, es decir:

$$Y = f(X_1, X_2, ..., X_k, u)$$

En el caso particular que el modelo de regresión sea lineal, tendremos una expresión con la siguiente estructura:

$$Y = \beta_1 + \beta_2 \cdot X_2 + ... + \beta_k \cdot X_k + u$$

El objetivo de la regresión será estimar los parámetros del modelo $\beta_1, \beta_2, \ldots, \beta_k$, de forma que el modelo resultante se ajuste lo mejor posible a las observaciones iniciales. Cuando la variable dependiente *Y* es continua, resulta frecuente utilizar un modelo de regresión lineal múltiple como el señalado. En tal caso la estimación de los parámetros $\beta_1, \beta_2, \ldots, \beta_k$ se lleva a cabo mediante los métodos de mínimos cuadrados ordinarios o generalizados según corresponda.

Tal como se mencionó anteriormente, a diferencia de la regresión lineal, en las regresiones para variables dicotómicas[8] como la regresión logística[9] se emplean los métodos de máxima verosimilitud para llevar a cabo la estimación de los parámetros del modelo. La idea fundamental al igual que la regresión lineal consiste en explicar la decisión *Y* que toma un individuo de entre un número limitado de posibles opciones a partir de un conjunto de variables explicativas X_1, X_2,..., X_k.

La variable dependiente en este caso puede denominarse como de decisión discreta y en general se trata de variables de tipo categórico, dentro de la que podemos distinguir varios tipos:

- Variables categóricas binarias: son aquellas que sólo pueden tomar dos valores como por ejemplo éxito ó fracaso, 0 ó 1, sí ó no etc.
- Variables categóricas ordinales: pueden tomar múltiples valores, entre los cuales es posible establecer una relación de orden así tenemos ninguno, alguno o muchos; primero, segundo, tercero o cuarto; pequeño, mediano, grande o muy grande, etc.
- Variables categóricas nominales: pueden tomar múltiples valores que no son posibles de ordenarlos, por ejemplo azul, rojo, verde, blanco; Quito, Tena, Guayaquil, Puerto Baquerizo Moreno, etc.

---

8 Variable dicotómica (dummy, indicadora, binaria, cualitativa) : Tenemos el caso de sexo, raza, religión, nacionalidad, tipo de cliente, etc. Indican la presencia o ausencia de una cualidad o atributo, para ello se construyen variables artificiales de 1 ó 0 (1 ó 2), donde 0 indica la ausencia de un atributo y 1 indica la presencia (o posesión) de ese atributo. El 1 puede indicar, por ejemplo, que una persona es hombre y 0 designar a una mujer.

9 La regresión logística parte de los coeficientes estimados para el conjunto de variables independientes, permite calcular la probabilidad de ser considerado como futuro cliente de riesgo (incumplir las obligaciones crediticias).

Lamentablemente el modelo de probabilidad lineal no está libre de pequeños inconvenientes a la hora de utilizarse como un método de clasificación. Consideremos el caso de una variable dependiente binaria *Y*, la cual viene explicada por un conjunto de predictores X_1, X_2,…, X_k. Podemos observar que por la característica de *Y* (al ser binaria puede tomar solo valores de 0 y 1), entonces siempre se cumplirá que:

$$E[Y] = 0 \cdot P(Y=0) + 1 \cdot P(Y=1) = P(Y=1)$$

Es decir: $E[Y] = P(Y=1)$

Por otra parte, podemos pensar en utilizar un modelo de regresión lineal múltiple para explicar el comportamiento de la variable *Y*, es decir:

$$Y = \beta_1 + \beta_2 X_2 + \ldots + \beta_k X_k + u$$

Bajo el supuesto habitual de que E[*u*] = 0[10], y suponiendo conocidos los valores que toman las variables explicativas, tendremos que:

$$E[Y] = \beta_1 + \beta_2 X_2 + \ldots + \beta_k X_k$$

Igualando las dos expresiones obtenidas para E [*Y*] llegamos al resultado que le da nombre al modelo de probabilidad lineal:

$$P(Y=1) = \beta_1 + \beta_2 X_2 + \ldots + \beta_k X_k = Y - u$$

10 El error aleatorio u se asume sigue una distribución Normal $(0, \sigma^2)$

Observamos que según esta última ecuación se puede expresar la variable dependiente binaria *Y* como la probabilidad de "éxito" más un término de perturbación.

$$Y = P(Y = 1) + u = E[Y] + u$$

Sin embargo, este modelo inicial no será válido para explicar el comportamiento de variables dependientes binarias, pues no restringe los valores que toman las variables independientes que pueden ser continuas, ordinales o nominales, por tanto la variable dependiente será continua y tomará valores desde menos infinito a más infinito, a más de ello podemos mencionar otros inconvenientes como los siguientes:

1. El término de perturbación

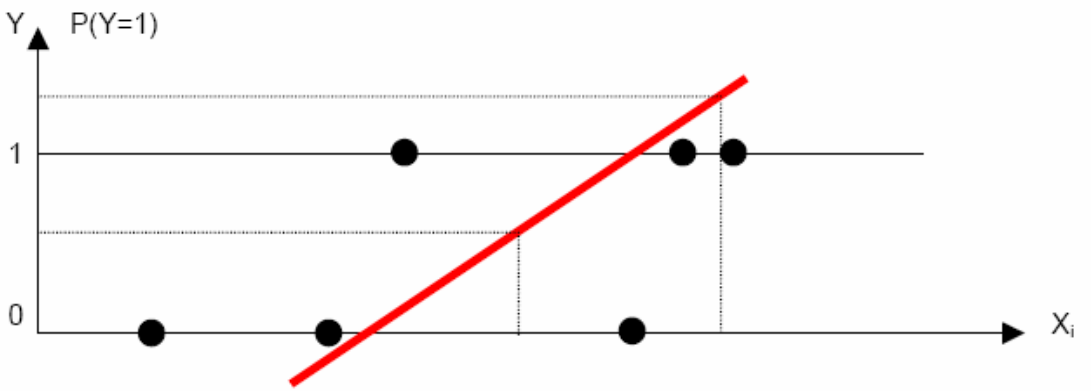
$$u = Y - (\beta_1 + \beta_2 X_2 + \ldots + \beta_k X_k)$$

No será una variable aleatoria continua como en la regresión lineal múltiple, sino que será una variable aleatoria discreta, puesto que, conocidos los valores de las variables explicativas, *u* sólo puede tomar dos valores determinados; por tanto, *u* ya no se distribuirá de forma normal que es uno de los supuestos básicos del modelo de regresión. Aunque este supuesto no resulta estrictamente necesario para aplicar mínimos cuadrados, sí es fundamental a la hora de realizar cualquier tipo de inferencia posterior sobre el modelo (intervalos de confianza para los parámetros estimados, contrastes de hipótesis, etc.).

2. El término de perturbación *u* no cumple la hipótesis de homoscedasticidad. Debido a este problema, los estimadores de mínimos cuadrados

no serán eficientes, por lo que resultará necesario recurrir a la estimación por mínimos cuadrados generalizados.

3. Como la variable dependiente *Y* sólo puede tomar los valores 0 y 1, si representamos gráficamente la nube de puntos formada por los pares de observaciones de *Y* con una de las variables explicativas X_i, obtendremos puntos situados sobre las rectas *Y* = 1 e *Y* = 0



Modelo de Regresión Lineal

Al estimar los parámetros del modelo de probabilidad lineal, estaremos ajustando una recta a la nube de puntos anterior (recta en rojo). El uso de dicha recta para predecir nuevos valores de *Y*, es decir valores de P (*Y* = 1) = *Y*-*u*, a partir de valores dados de X_i puede proporcionar valores mayores que 1 o menores que 0, lo que estaría en contradicción con la definición de probabilidad utilizada.

4. Finalmente, la expresión:

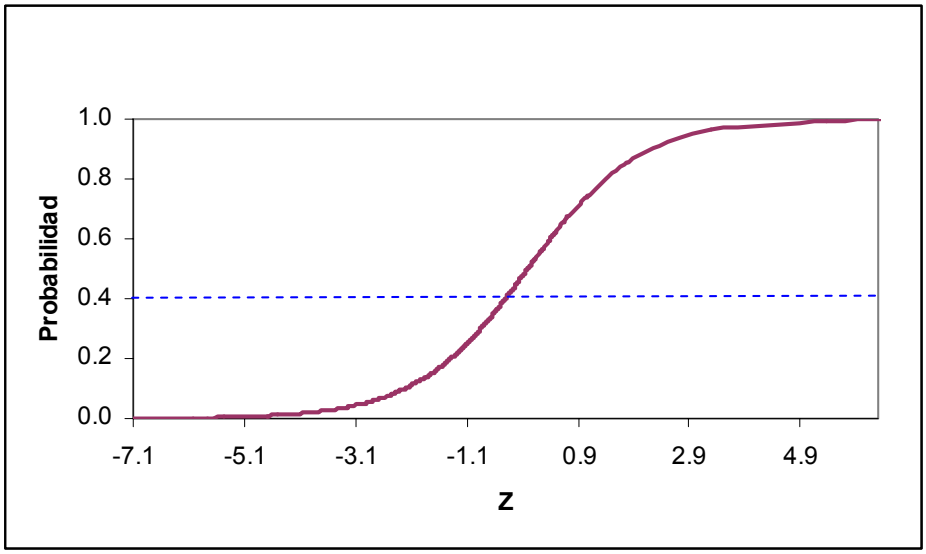
$$P(Y=1) = \beta_1 + \beta_2 X_2 + \ldots + \beta_k X_k$$

Indica que la probabilidad del suceso “éxito” viene determinada por una combinación lineal de variables explicativas. De ello se deduce que

$$\frac{\partial P(Y=1)}{\partial X_i} = \beta_i \quad \forall i = 2,3,\ldots,k\,.$$

En otras palabras, la variación en P (*Y* = 1) causada por cambios en alguna de las variables explicativas es constante, lo cual es una hipótesis muy poco realista.

Como hemos visto uno de los principales problemas al procurar usar un modelo de regresión lineal para estimar la probabilidad de éxito medida por la variable independiente, por lo tanto no podemos asegurar que los valores resultantes de utilizar este modelo estén necesariamente en el intervalo $[0,1]$ (que coincidan con el valor de probabilidad), por ello es más conveniente utilizar una ecuación cuya función de distribución tenga la siguiente estructura:



Modelo de Regresión No Lineal

Modelos con características similares a la gráfica mostrada son el Logit y Probit, que procuran solucionar las inconsistencias ya mencionadas de un modelo de regresión lineal sobre una variable dependiente de tipo binario, sobre estos modelos y fundamentalmente respecto al modelo Logit, hablaremos al final de este

trabajo al abordar brevemente el tema de la modelación que tiene como paso fundamental lo discutido a lo largo del presente desarrollo.

Punto de corte

Una vez que se haya determinado el modelo de clasificación y este ha arrojado resultados que han sido validados mediante pruebas estadísticas, definimos como punto de corte a aquel que nos permita maximizar las características que diferencian los grupos de buenos y malos clientes. A partir de este punto se tomaran una serie de decisiones tanto hacia arriba como hacia abajo del mismo, por ejemplo podemos decidir que aquellos clientes que se encuentran sobre este punto tengan facilidades o menos restricciones a la hora de conceder un crédito, y que en su defecto aquellos clientes que se encuentran por debajo del mismo, deban cumplir mayores exigencias o requerimientos a la hora de acceder al mismo crédito.

III. APLICACIÓN PRÁCTICA

Para un entendimiento práctico de todo lo que se ha expuesto a lo largo de este trabajo, vamos a ejemplificar el tratamiento aplicado a un grupo de variables, así como su interacción en el proceso de modelación. Para ello primero determinaremos una base ficticia de trabajo compuesta por 15.000 clientes que tienen un producto crediticio, se determinaron algunas variables de las usualmente recopiladas por las instituciones financieras en la gestión crediticia, dichas variables se muestran en la siguiente tabla:

Nombre de la Variable	Descripción	Tipo de variable
Fecha de Corte	Fecha de análisis para la base de datos	Informativa
ID	Identificación del cliente	Informativa
Monto Aprobado	Valor concedido para el producto crediticio	Cuantitativa
Sueldo Depurado	Ingresos menos egresos	Cuantitativa
Ciudad Domicilio	Ciudad de residencia del cliente	Cualitativa
Celular	Tenencia del cliente de teléfono móvil	Cualitativa
Mail	Tenencia del cliente de dirección de correo electrónico	Cualitativa
Provincia Domicilio	Provincia de residencia del cliente	Cualitativa
Telefono Domicilio	Tenencia del cliente de teléfono en domicilio	Cualitativa
Actividad Económica	Area en la que el cliente se desempeña laboralmente	Cualitativa
Cargo	Estatus laboral del cliente	Cualitativa
Antiguedad Laboral	Tiempo en el trabajo actual	Cuantitativa
Deuda SIF	Tenencia de deudas en el Sistema Financiero	Cualitativa
Profesión	Area de especialización del cliente	Cualitativa
Tipo de Vivienda	Tenencia de vivienda	Cualitativa
Buró	Calificación del cliente en central de riesgo	Cualitativa
Estado Civil		Cualitativa
Edad		Cuantitativa
Nivel Educación	Estudios alcanzados	Cualitativa
Cargas Familiares	No. de personas que dependen del cliente	Cuantitativa
Sexo	Género del cliente	Cualitativa
Saldo por vencer	Monto pendiente de pago al corte	Cuantitativa
Saldo vencido	Monto no cancelado al corte	Cuantitativa
Mora Promedio	Atraso promedio en los últimos 12 meses	Cuantitativa
Mora Máxima	Atraso máximo en los últimos 12 meses	Cuantitativa
B ó M	Definición de mal cliente	Cuantitativa
Antiguedad Vivienda	Tiempo en la vivienda actual	Cuantitativa

Elaboración: Propia
Fuente: Bdd APEVSC

Como se ha señalado a lo largo de este trabajo, uno de los primeros pasos a dar consiste en la validación de los datos, esto implica verificar las fuentes de información de dónde proceden, más aún si se tratan de bases internas o externas. Por ejemplo se pueden tener diferentes fuentes de información para una misma variable, la contenida en la solicitud crediticia, la recabada de los procesos de verificación telefónica y física, las existentes en el Registro Civil e inclusive bases del Tribunal Supremo Electoral, estas fuentes alternativas permitirán constatar y depurar algunas variables, como edad, sexo, estado civil, ciudad de nacimiento, dirección domiciliaria e inclusive nivel de educación de los clientes.

Para variables como antigüedad laboral, actividad económica, lugar de trabajo, se pueden evitar distorsiones en los datos, realizando el cruce de información entre los datos entregados por el cliente, los certificados expedidos por la empresa o institución dónde él labora y la información consolidada en los registros del IESS. Otro par de variables que se pueden analizar con mayor énfasis son los ingresos y gastos que tiene el cliente, a fin de obtener la variable sueldo depurado.

Para el caso de los ingresos, se debe verificar cuán actualizada está la variable, ya que si se está trabajando con bases de datos históricas seguramente existirá mucha volatilidad e inconsistencias entre las cifras obtenidas, para evitar esto lo lógico es procurar una antigüedad no mayor a dos años, y paralelamente se puede intentar cruzar la información proporcionada del cliente con documentos de respaldo como roles de pago, tablas de homologación salarial de la SENRES, Seguro Social, etc.

Tratándose de los egresos, aparte de los valores declarados por el cliente como gastos de luz, agua, teléfono, pensiones, se puede intentar determinar si el

cliente tiene o maneja tarjetas de crédito de casas comerciales (Casa Tosi, De Prati, Fybeca, etc.), la utilización de teléfonos celulares y el tipo de plan al que está afiliado el cliente (pos pago o pre pago), la afiliación a clubes, revistas o servicios específicos como internet, televisión por cable, etc. que demanden desembolsos mensuales. De este modo pese a no contar con la información del buró de crédito se puede calcular de mejor manera la capacidad de endeudamiento del cliente y junto con los ingresos, obtener efectivamente una mejor estimación actualizada del sueldo depurado.

En este sentido mientras a mayores y diversas fuentes de información se tenga acceso, garantizará de alguna maneja una base de datos depurados y confiables que facilitará el proceso de análisis de las variables en aras de la construcción de un modelo tipo score para una adecuada gestión de riesgo crediticio.

Una vez que se han identificado el grupo de variables de la base de trabajo vamos a analizar cada una de ellas según su característica cualitativa o cuantitativa tal como se detallo anteriormente en el Capitulo No. 2. Para el caso de las variables cuantitativas como monto, sueldo, saldo, etc., se determinan estadísticas descriptivas que permitan tener un mejor conocimiento de dichas particulares de la población en análisis. De la población en estudio, encontramos variables cuya información no es completa para la base de 15.000 clientes tal es el caso de antigüedad en la vivienda actual, las cargas familiares, edad, etc., esto puede obedecer a errores en la consolidación de los datos y/o sus fuentes de origen.

Si se tratara de una base real, hay diversas alternativas que nos permitirán solucionar este aparente inconveniente de falta de información, uno de ellos sería acudir nuevamente al origen de los datos o a las carpetas de los clientes para tratar de validar la información de la base de trabajo. Para el caso de variables numéricas y dependiendo de la cantidad de casos faltantes se suele aplicar técnicas sencillas de compleción, como la utilización de las medias, medianas, el cálculo de promedios entre valores contiguos. Aunque puede ser más recomendable utilizar métodos de pronóstico de datos como regresiones lineales, si se desea dar un tratamiento más técnico se ha de optar por técnicas de estadística bayesiana, extrapolaciones alambradas o basadas en cálculo numérico como *splines*[11] cúbicos, todos estos métodos procuran evitar sesgos adicionales al pre existente en la información.

Adicionalmente es necesario identificar qué corresponde a un valor perdido (existe pero no está registrado) y qué a un vacío (no existe valor), sobre esto último se tomará la decisión de continuar trabajando con éstos datos, borrarlos o completarlos con lo señalado en el párrafo anterior. En este sentido se han desarrollado diversas técnicas de investigación y tratamiento de datos faltantes recopilados en textos de *minería de datos*, lamentablemente ese tipo de análisis está fuera del objetivo trazado en nuestro trabajo[12].

---

11 Un spline es una curva definida a trozos mediante polinomios. En los problemas de interpolación, se utiliza a menudo la interpolación mediante splines porque da lugar a resultados similares requiriendo solamente el uso de polinomios de bajo grado a la vez que se evitan las oscilaciones, que en la mayoría de las aplicaciones resultan indeseables, que aparecen al interpolar mediante polinomios de grado elevado.

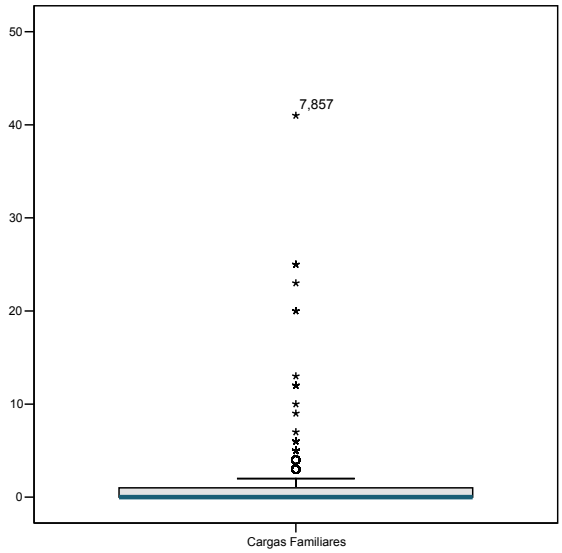
12 Juan Gómez García, Javier Palarea Albaladejo y Joseph Antoni Martín Fernádez, *Métodos de inferencia estadística con datos faltantes*. *Estudio de simulación sobre los efectos en las estimaciones*, Estadística Española, Vol. 48, 2006.

Estadísticos Descriptivos

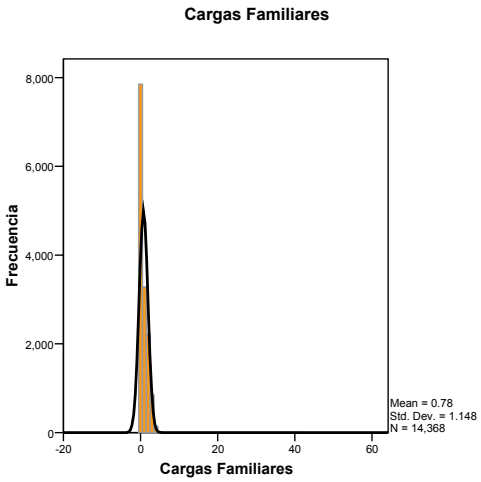
	No. Obs.	Mínimo	Máximo	Media	Desv. Típ.	Curtosis
Monto Aprobado	15.000	150,0	1.420,0	433,2	223,5	0,040
Sueldo Depurado	14.203	-2.813,4	3.885,8	305,4	185,5	0,041
Anti. Laboral	15.000	-	54,0	6,1	6,3	0,040
Edad	14.364	21,7	66,7	32,7	9,0	0,041
Cargas Familiares	14.368	-	41,0	0,8	1,1	0,041
Antigüedad Vivienda	14.367	-	61,0	11,1	9,8	0,041
Saldo por vencer	15.000	-	2.073,0	239,5	212,2	0,040
Saldo Vencido	15.000	-	1.171,0	6,8	46,7	0,040
Mora Promedio	15.000	-	163,0	5,3	16,3	0,040
Mora Máxima	15.000	-	375,0	19,7	50,3	0,040

Como se puede apreciar en la tabla anterior, al revisar las columnas de valores mínimos, máximos y promedios vemos que existe la necesidad de depurar la base de información, pues mientras que unas variables presentan valores bastante lógicos y acordes a la realidad del producto en estudio, como por ejemplo los datos señalan que la concesión crediticia tiene un monto promedio de $433, con valores que van entre los 150 y 1.420 dólares. Mientras que por otro lado vemos errores en los valores de la variable sueldo puesto que existe un dato negativo en su valor mínimo.

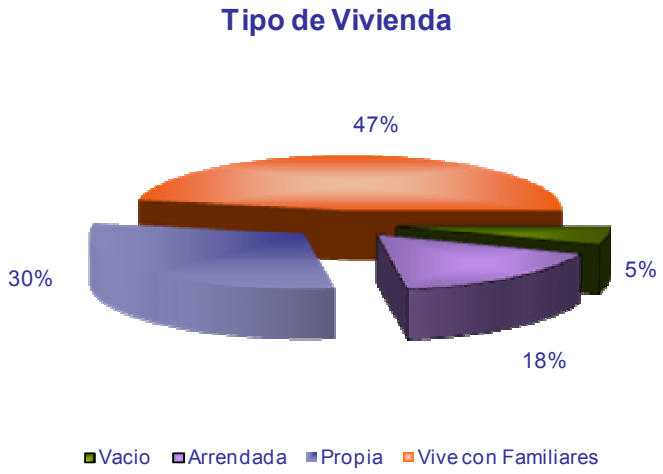
Otro dato curioso se encuentra en la variable cargas familiares, tal como se observa en el diagrama de caja de esta variable, existen datos que pueden ser considerados como errores, tal es el caso de aquellos registros con valores superiores a diez cargas familiares, mientras que valores entre tres y diez podrían ser tratados como atípicos.



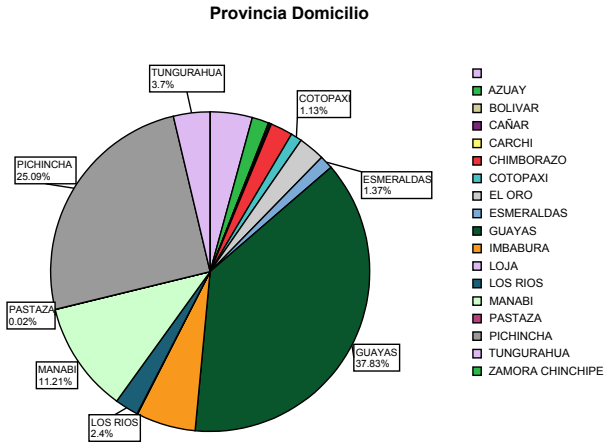
Si analizamos el diagrama de distribución de frecuencias de esta variable, podemos confirmar su valor medio mientras que en los extremos de la distribución se encuentran los valores que hemos anotado anteriormente.



Por otro lado se analizan las variables cualitativas, por ejemplo en el caso de la variable tipo de vivienda, vemos que el 30% señala vivir en casa propia y así mismo existe un 5% que carece de información.



Situación similar se evidencia en el caso de la variable provincia de domicilio, donde entre Guayas y Pichincha albergan el 62% de los clientes y también existe un porcentaje de datos nulos.



Adicionalmente al análisis descriptivo ilustrado en los párrafos anteriores, es necesario verificar la consistencia de la información mediante el cruces de variables, por ejemplo contrastar la edad de los clientes respecto a la antigüedad en la vivienda, la antigüedad laboral, o las cargas familiares ya que resulta extraño que la base de datos tenga clientes de 25 o 30 años de edad que vivan 40 o 51 años en la misma casa, que tengan una antigüedad laboral igual o mayor a su edad, que el

cliente de 25 años tenga más de 11 años de antigüedad laboral, o que pese a ser soltero tenga 6 cargas familiares, etc.

Del resultado del análisis descriptivo, más los cruces de variables se realiza una depuración de la base de datos original, de este modo nuestra base de trabajo se reduce de 15.000 clientes a 13.065, tal como se puede apreciar en la siguiente tabla:

Estadísticos Descriptivos Base Depurada

	Edad	Cargas Familiares	Antigüedad Vivienda	Antigüedad Laboral	Sueldo Depurado	Saldo por Vencer	Saldo Vencido
No. Obs.	13065	13065	13065	13065	13065	13065	13065
Media	33,3	0,8	11,1	6,1	309,5	248,1	7,4
Mediana	30,8	0,0	8,0	3,9	270,2	213,9	0,0
Moda	32,0	0,0	10,0	3,0	272,0	0,0	0,0
Desv. Tip	8,9	1,0	9,8	5,9	188,9	217,2	49,1
Mínimo	22,0	0,0	0,0	1,0	33,9	0,0	0,0
Máximo	67,0	6,0	60,0	44,0	3885,8	2073,0	1171,0
Pct. 25	26,4	0,0	3,0	2,4	192,7	60,1	0,0
Pct. 50	30,8	0,0	8,0	3,9	270,2	213,9	0,0
Pct. 75	38,0	1,0	20,0	7,1	369,9	372,7	0,0

Ahora bien, el siguiente paso será analizar la relación que mantiene cada una de las variables independientes de la base de datos respecto a la variable dependiente a fin de determinar la discriminación de la primeras, este proceso se realiza tanto para las variables cualitativas como cuantitativas, esto puede hacerse tanto mediante la utilización de tablas de contingencia como en forma gráfica según corresponda.

Partiendo de la definición de cliente malo formulada anteriormente[13] podemos tomar la variable cualitativa tenencia de teléfono móvil, vemos que el 14,9% de la

13 Cliente con una mora promedio superior o igual a 15 días y una mora máxima superior o igual a 30 días

población en estudio no dispone de celular del cual el 63,9% de identifica como buen cliente y su restante 36,1% como mal cliente. Proporciones similares se evidencian en el caso de los clientes que si tienen teléfono celular tal como se muestra en la tabla siguiente.

Tabla de Contingencia Celular & B ó M

			B ó M		Total
			Bueno	Malo	
Celular	no	% con celular	63,9%	36,1%	100,0%
		% total	9,5%	5,4%	14,9%
	si	% con celular	63,7%	36,3%	100,0%
		% total	54,2%	30,8%	85,1%
Total		% con celular	63,8%	36,2%	100,0%
		% total	63,8%	36,2%	100,0%

En el caso de la variable tenencia de correo electrónico es totalmente marcada la cantidad de clientes que no disponen de este medio con un 99,3% de la población en estudio.

Tabla de Contingencia Mail & B ó M

			B ó M		Total
			Bueno	Malo	
Mail	no	% con Mail	63,8%	36,2%	100,0%
		% total	63,3%	35,9%	99,3%
	si	% con Mail	58,8%	41,2%	100,0%
		% total	0,4%	0,3%	0,7%
Total		% con Mail	63,8%	36,2%	100,0%
		% total	63,8%	36,2%	100,0%

En el caso de teléfono en domicilio el 35,7% de quienes si disponen de este dispositivo se presentan como malos clientes.

Tabla de Contingencia Teléfono Domicilio & B ó M

			B ó M		Total
			Bueno	Malo	
Teléfono Domicilio	no	% con teléf. dom	62,3%	37,7%	100,0%
		% total	16,4%	9,9%	26,3%
	si	% con teléf. dom	64,3%	35,7%	100,0%
		% total	47,4%	26,3%	73,7%
Total		% con teléf. dom	63,8%	36,2%	100,0%
		% total	63,8%	36,2%	100,0%

Para actividad económica, vemos que actividades financieras e industriales apenas alcanzan un 13,3% de la población en estudio con un 32% en promedio de malos clientes.

Tabla de Contingencia Actividad Económica & B ó M

			B ó M		Total
			Bueno	Malo	
Actividad Económica	Comercial	% con actividad económica	63,2%	36,8%	0,0%
		% total	15,5%	9,0%	24,4%
	Financiera	% con actividad económica	70,3%	29,7%	0,0%
		% total	0,8%	0,3%	1,2%
	Industria	% con actividad económica	65,7%	34,3%	0,0%
		% total	8,0%	4,2%	12,1%
	Servicios	% con actividad económica	63,5%	36,5%	0,0%
		% total	39,5%	22,8%	62,3%
Total		% con actividad económica	63,8%	36,2%	0,0%
		% total	63,8%	36,2%	0,0%

El 64% de los clientes categorizados en cargos de administración y operaciones se catalogan como buenos clientes, mientras que este grupo alberga el 59,3% de la variable.

Tabla de Contingencia Cargo & B ó M

			B ó M		Total
			Bueno	Malo	
Cargo	Administración y Operaciones	% con cargo	64,1%	35,9%	100,0%
		% total	38,0%	21,3%	59,3%
	Ejecutivo	% con cargo	62,0%	38,0%	100,0%
		% total	2,2%	1,4%	3,6%
	Mando Medio	% con cargo	62,7%	37,3%	100,0%
		% total	13,5%	8,0%	21,6%
	Obrero	% con cargo	64,7%	35,3%	100,0%
		% total	6,9%	3,8%	10,7%
	Propietario	% con cargo	65,5%	34,5%	100,0%
		% total	0,1%	0,1%	0,2%
	Vendedor	% con cargo	64,1%	35,9%	100,0%
		% total	3,0%	1,7%	4,7%
Total		% con cargo	63,8%	36,2%	100,0%
		% total	63,8%	36,2%	100,0%

Los clientes que tienen alguna deuda en el sistema financiero y han sido clasificados como buenos clientes representan el 51,8%, mientras que el 11,9% de

clientes buenos restantes corresponde a clientes sin deudas en el sistema financiero.

Tabla de Contingencia Deuda SIF & B ó M

			B ó M		
			Bueno	Malo	Total
Deuda SIF	no	% con deuda SIF	63,8%	36,2%	100,0%
		% total	11,9%	6,8%	18,7%
	si	% con deuda SIF	63,7%	36,3%	100,0%
		% total	51,8%	29,5%	81,3%
Total		% con deuda SIF	63,8%	36,2%	100,0%
		% total	63,8%	36,2%	100,0%

Resultado fundamental de este análisis es identificar atributos o categorías de una variable que por carecer de una cantidad significativa de clientes que justifiquen ser tratados en forma independiente, lo que conlleva a la redistribución de sus elementos en otros atributos. Por ejemplo si nos fijamos en la variable profesión, encontramos que profesiones como abogado, arquitecto, auditor, economista, enfermero, etc., tienen menos del 1% de la población en análisis, esto sugiere que dichos atributos deban ser reagrupados sea en una sola gran categoría o en función de otras profesiones afines como unir contadores y economistas, ingenieros con arquitectos y tecnólogos o en su defecto aplicar técnicas más elaboradas de análisis estadístico como componentes principales, clúster, árboles de decisión, etc., esto último según prefiera la persona a cargo de la elaboración del score.

Tabla de Contingencia Profesión & B ó M

			B ó M		
			Bueno	Malo	Total
profesión	Abogado	% con Profesión	50,0%	50,0%	100,0%
		% total	0,0%	0,0%	0,1%
	Administrador	% con Profesión	62,8%	37,2%	100,0%
		% total	1,4%	0,8%	2,2%
	Arquitecto	% con Profesión	73,3%	26,7%	100,0%
		% total	0,1%	0,0%	0,1%
	Auditor	% con Profesión	57,1%	42,9%	100,0%
		% total	0,1%	0,1%	0,2%
	Contador	% con Profesión	63,4%	36,6%	100,0%
		% total	1,0%	0,6%	1,6%
	Economista	% con Profesión	56,3%	43,8%	100,0%
		% total	0,1%	0,1%	0,2%
	Enfermero	% con Profesión	70,6%	29,4%	100,0%
		% total	0,5%	0,2%	0,6%
	Ingeniero	% con Profesión	67,2%	32,8%	100,0%
		% total	1,9%	0,9%	2,8%
	Licenciado	% con Profesión	64,8%	35,2%	100,0%
		% total	3,1%	1,7%	4,7%
	Médico u Odontólogo	% con Profesión	79,2%	20,8%	100,0%
		% total	0,5%	0,1%	0,6%
	Militar	% con Profesión	61,3%	38,7%	100,0%
		% total	1,8%	101,0%	2,9%
	Ninguna	% con Profesión	64,1%	35,9%	100,0%
		% total	38,0%	21,3%	59,3%
	Periodista	% con Profesión	52,0%	48,0%	100,0%
		% total	0,1%	0,1%	0,2%
	Policia	% con Profesión	63,8%	36,2%	100,0%
		% total	5,6%	3,2%	8,8%
	Profesor	% con Profesión	60,0%	40,0%	100,0%
		% total	2,4%	1,6%	4,0%
	Secretario	% con Profesión	64,8%	35,2%	100,0%
		% total	0,5%	0,3%	0,8%
	Sociólogo	% con Profesión	46,7%	53,3%	100,0%
		% total	0,1%	0,1%	0,1%
	Técnico	% con Profesión	61,9%	38,1%	100,0%
		% total	5,2%	3,2%	8,4%
	Tecnólogo	% con Profesión	63,9%	36,1%	100,0%
		% total	1,3%	0,8%	2,1%
Total		% con Profesión	63,8%	36,2%	100,0%
		% total	63,8%	36,2%	100,0%

Análisis similar al mencionado para la variable profesión se ha de aplicar a las variables que en términos generales cuenten con muchas categorías como el caso de provincia y ciudad de domicilio cuyas tablas no son presentadas en este documento.

Tabla de Contingencia Tipo de Vivienda & B ó M

			B ó M		Total
			Bueno	Malo	
Tipo de Vivienda	Arrendada	% con tipo de vivienda	63,8%	36,2%	100,0%
		% total	11,8%	6,7%	18,5%
	Propia	% con tipo de vivienda	63,7%	36,2%	100,0%
		% total	21,1%	12,0%	33,1%
	Vive con Familiar	% con tipo de vivienda	63,7%	36,3%	100,0%
		% total	30,9%	17,6%	48,4%
Total		% con tipo de vivienda	63,8%	36,2%	100,0%
		% total	63,8%	36,2%	100,0%

Como muestra la tabla anterior, a priori no se puede establecer una diferencia entre el comportamiento de pago de aquellos clientes con casa propia, arrendada o que vivan con familiares.

Tabla de Contingencia Buró & B ó M

			B ó M		Total
			Bueno	Malo	
Buró	A	% con Buró	62,8%	37,2%	100,0%
		% total	18,2%	10,8%	29,0%
	AA	% con Buró	64,7%	35,3%	100,0%
		% total	31,6%	17,3%	48,9%
	AAA	% con Buró	63,0%	37,0%	100,0%
		% total	13,9%	8,2%	22,1%
Total		% con Buró	63,8%	36,2%	100,0%
		% total	63,8%	36,2%	100,0%

Situación similar a la del tipo de vivienda se presenta para la variable buró de crédito frente a tipo de cliente.

Tabla de Contingencia Sexo & B ó M

			B ó M		Total
			Bueno	Malo	
SEXO	Femenino	% con sexo	63,4%	36,2%	100,0%
		% total	23,6%	13,6%	37,1%
	Masculino	% con sexo	64,0%	36,0%	100,0%
		% total	40,2%	22,6%	62,9%
Total		% con sexo	63,8%	36,2%	100,0%
		% total	63,8%	36,2%	100,0%

De la base de trabajo, vemos que el 37% corresponden a mujeres de las cuales un el 63,4% son catalogadas como buenas clientas.

Tabla de Contingencia Estado Civil & B ó M

			B ó M		Total
			Bueno	Malo	
Estado Civil	Casado	% con Estado	63,2%	36,8%	100,0%
		% total	21,7%	12,6%	34,3%
	Divorciado	% con Estado	60,5%	39,5%	100,0%
		% total	1,7%	1,1%	2,8%
	Soltero	% con Estado	64,2%	35,8%	100,0%
		% total	39,2%	21,9%	61,1%
	Unión Libre mayor a 2 años	% con Estado	67,2%	32,8%	100,0%
		% total	1,0%	0,5%	1,4%
	Viudo	% con Estado	56,7%	43,3%	100,0%
		% total	0,3%	0,2%	0,5%
Total		% con Estado	63,8%	36,2%	100,0%
		% total	63,8%	36,2%	100,0%

Para la variable estado civil, aparentemente la categoría unión libre evidencia un mejor comportamiento que el resto de atributos, pero en este caso nuevamente se torna necesario prestar atención a su representatividad en la población, pues apenas constituye el 1,4% de la base en análisis.

Tabla de Contingencia Nivel de Educación & B ó M

			B ó M		Total
			Bueno	Malo	
Estado Civil	Ninguno	% con Nivel	62,5%	37,5%	100,0%
		% total	0,2%	0,1%	0,2%
	Primaria	% con Nivel	64,4%	35,6%	100,0%
		% total	12,2%	6,8%	19,0%
	Secundaria	% con Nivel	63,5%	36,5%	100,0%
		% total	42,2%	24,2%	66,5%
	Superior	% con Nivel	64,1%	35,9%	100,0%
		% total	9,1%	5,1%	14,1%
	Técnico	% con Nivel	47,4%	52,6%	100,0%
		% total	0,1%	0,1%	0,1%
Total		% con Nivel	63,8%	36,2%	100,0%
		% total	63,8%	36,2%	100,0%

En el caso de nivel de educación se podría determinar una tendencia del atributo técnico a ser mal cliente, pero nuevamente entra en juego la significancia de la categoría dentro de la variable.

En el caso de variables cuantitativas, podemos realizar el análisis tanto de las estadísticas descriptivas como de las distribuciones de las variables en función de la

variable tipo de cliente, donde se podrá verificar si existe o no diferencias en su estructura.

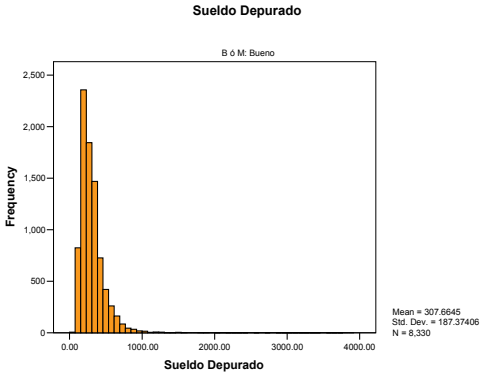
Estadísticos Descriptivos para Buen Cliente

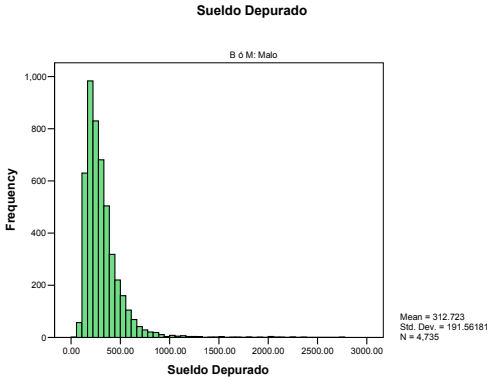
	No. Obs.	Mínimos	Máximos	Media	Desviación Std.
Sueldo Depurado	8.330	33,9	3.885,8	307,7	187,4
Antig. Laboral	8.330	1,0	44,0	6,1	5,9
Edad	8.330	22,0	67,0	33,3	8,9
Cargas Familiares	8.330	-	6,0	0,8	1,0
Saldo por vencer	8.330	-	2.073,0	249,5	215,1
Saldo vencido	8.330	-	1.168,0	5,5	42,6
Antigüedad Vivienda	8.330	-	55,0	11,1	9,8
Valid N (listwise)	8.330	-			

Estadísticos Descriptivos para Mal Cliente

	No. Obs.	Mínimos	Máximos	Media	Desviación Std.
Sueldo Depurado	4.735	47,7	2.750,0	312,7	191,6
Antig. Laboral	4.735	1,0	42,0	6,1	5,9
Edad	4.735	22,0	67,0	33,2	9,0
Cargas Familiares	4.735	-	6,0	0,8	1,0
Saldo por vencer	4.735	-	1.327,0	245,6	221,1
Saldo vencido	4.735	-	1.171,0	10,8	58,6
Antigüedad Vivienda	4.735	-	60,0	11,0	9,8
Valid N (listwise)	4.735	-			

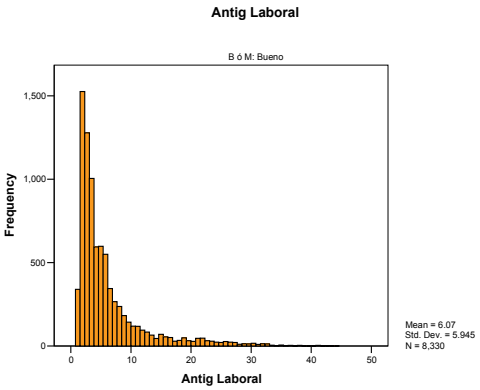
Al revisar brevemente los estadísticos descriptivos de los cuadros anteriores, no se evidencian diferencias en las características de los dos grupos poblacionales, esto se confirma una vez que se analiza la distribución de frecuencias en forma gráfica.

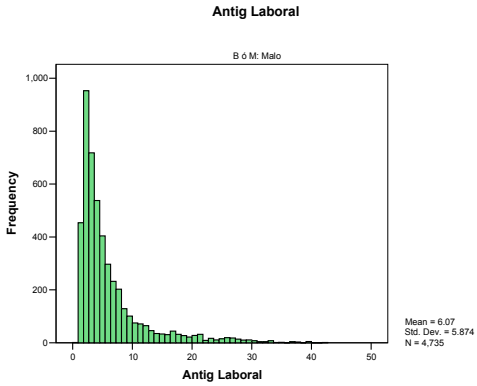


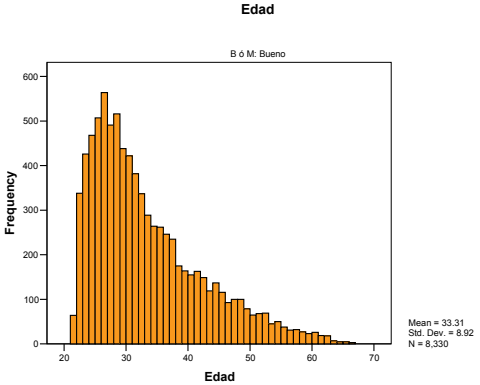


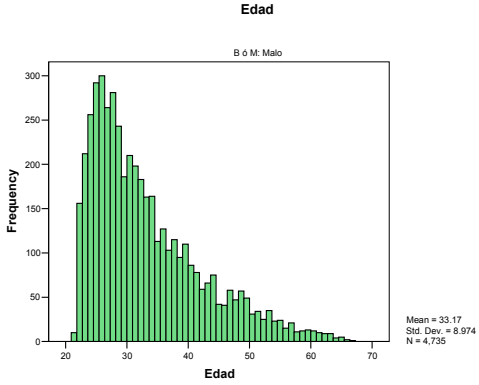
En el caso de sueldo depurado, hay un ligero incremento en la desviación de la distribución de los malos clientes.

Para el caso de antigüedad laboral y edad el comportamiento de la distribución para buenos y malos clientes mantiene similares características, con concentraciones alrededor de los seis y treinta y tres años de antigüedad y edad respectivamente.

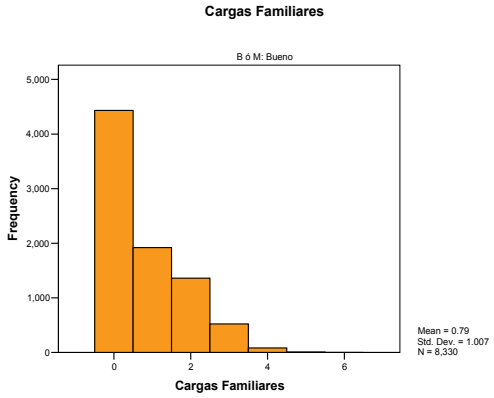


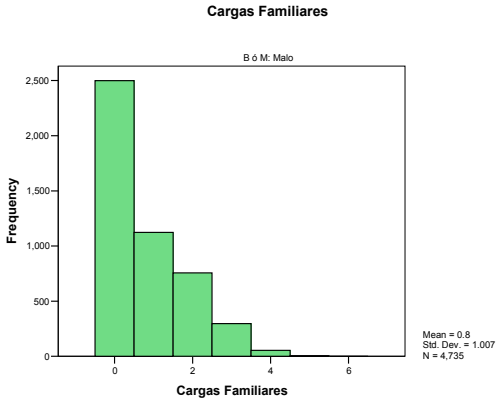




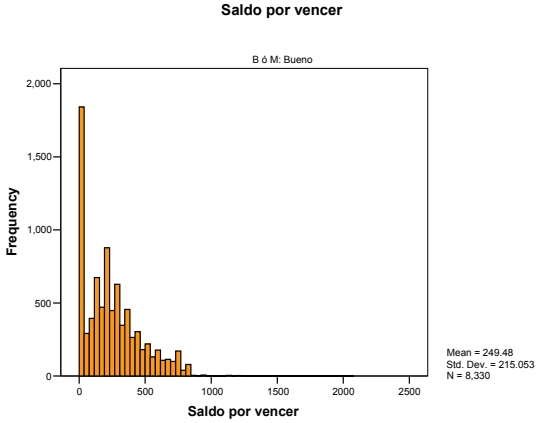


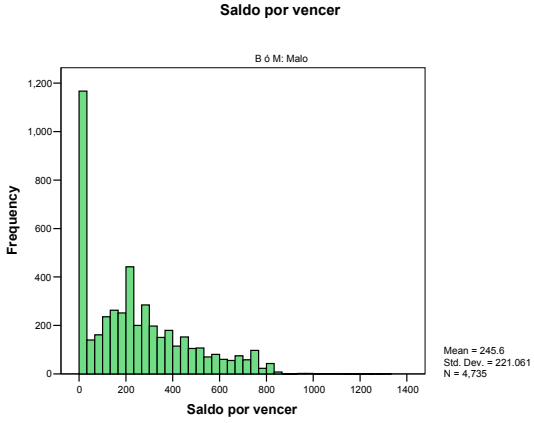
En cuanto a cargas familiares, en ambos grupos poblacionales se evidencia la mayor concentración en torno a cero cargas.



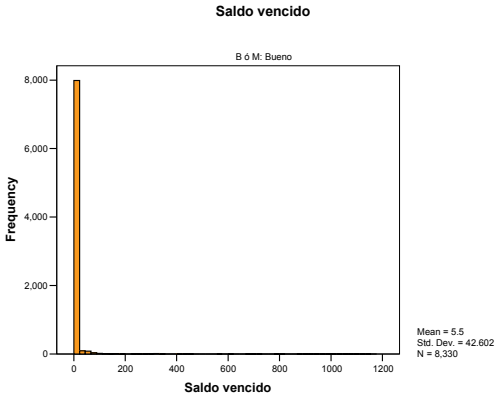


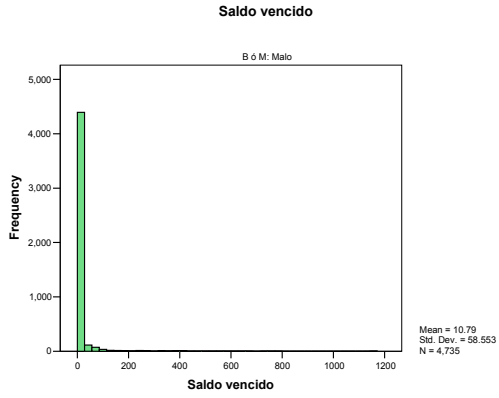
En el caso de saldo por vencer la dispersión de los malos clientes es mayor, aunque las medias poblacionales mantienen valores similares.



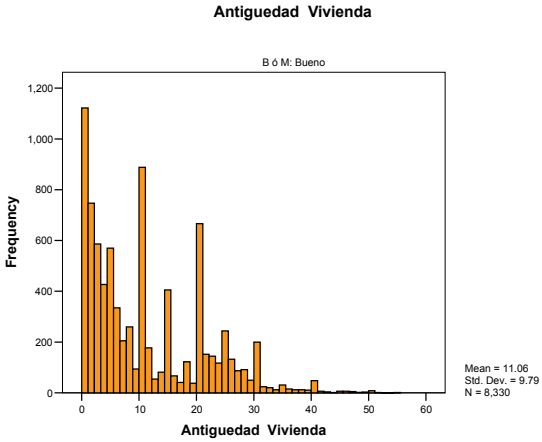


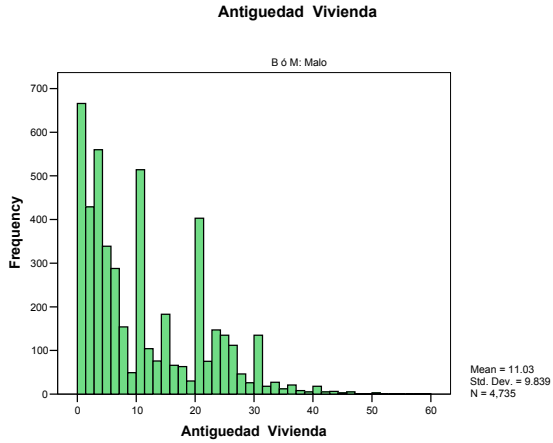
En cuanto al saldo vencido en cambio, tanto los valores promedio, como de la desviación estándar aumentan en el grupo que representa a los malos clientes.





Finalmente en cuanto a la variable antigüedad de vivienda, en los dos grupos poblacionales, se puede apreciar una alta concentración de personas que tienen menos de un año en su residencia actual.





Del diagnóstico realizado tanto a variables cuantitativas como cualitativas y la interacción de las mismas con la variable de decisión, nos permiten a priori establecer una idea respecto de la capacidad de discriminación o clasificación de cada una de ellas, además el profesional a cargo de la modelación del *score* podrá evaluar si es necesario realizar algún trabajo adicional con las variables según sus características, la cantidad de clientes por categoría, agrupaciones por conjuntos similares para edad, antigüedad laboral, cargas familiares, construcción de índices, etc.

En este sentido, un procedimiento muy empleado para la elección de los puntos de corte se basa en escoger los valores de los cuartiles o de percentiles específicos de la distribución de los datos en función de la variable a analizar. Este método se suele utilizar para fijar intervalos de referencia de pruebas analíticas a partir de una muestra representativa de la población, eligiéndose dos percentiles centrados en torno a la mediana de la distribución, concretamente los valores 2,5 y 97,5, que definen entre sí un intervalo de referencia del 95 %.

Otra alternativa a emplear para los puntos de corte, podría ser considerar como valores elevados los que están por encima del tercer cuartil, cualquiera sea la alternativa escogida, hay que estar conciente que la categorización de una variable cuantitativa supone siempre una pérdida importante de información, y si además los puntos de corte se eligen en base a la información proporcionada por los propios datos del estudio puede dar lugar a que las conclusiones sean menos extrapolables a otras situaciones[14]; por ello en los modelos de regresión siempre es preferible utilizar las variables cuantitativas como tales, y no convertidas a cualitativas ya que, además de no perder eficiencia, nos permite calcular la modificación en el riesgo para cambios diferentes en la magnitud del factor de pronóstico.

Una vez obtenidas las variables definitivas de las base de datos y seleccionado el método de modelación (análisis discriminante, regresiones, árboles de decisión, etc.),[15] se han de tomar en cuenta aquellas medidas estadísticas

---

14 Luis Miguel Molinero*, Elección de los puntos de corte para convertir una variable cuantitativa en cualitativa*, Asociación de la Sociedad Española de Hipertensión, 2003, p. 1.

15 Roberto Araya, *Métodos estadísticos básicos de discriminación con árboles de decisión*, AutoMind, Santiago de Chile, 2005, p. 3.

usualmente utilizadas para evaluar la contribución de una variable en el modelo a desarrollar, podemos mencionar algunas según sea el caso:

Coeficiente de determinación (R^2)

A veces también llamado coeficiente correlación múltiple al cuadrado, es una medida descriptiva que sirve para evaluar la bondad de ajuste del modelo a lo datos, ya que mide la capacidad predictiva del modelo ajustado. Si el valor de R^2 es alto, se asocia como una buena capacidad de pronóstico de la variable dependiente a través de las independientes. Se define como el cociente entre la variabilidad explicada por la regresión y la variabilidad total, esto es:

$$R^2 = \frac{\text{Varianza explicada}}{\text{Varianza total}} = 1 - \frac{\text{Varianza no explicada}}{\text{Varianza total}}$$

Se utiliza para medir la reducción en la variabilidad total de Y debido a la inclusión de las variables regresoras X_1, X_2,… X_k. Un valor grande de R^2 no necesariamente implica que el modelo es bueno. Adicionar variables al modelo siempre incrementa el valor de R^2, ya sea que las variables contribuyan o no al modelo. Es posible que modelos con valor de R^2 grande sean malos en la predicción o estimación.

1. R^2 mide la correlación entre Y y $\hat{Y}$, dado que $0 \leq R^2 \leq 1$.

2. Si existe error puro, es imposible que R^2 alcance el valor de 1. La única manera en que podría dar R^2 =1, sería que se tuviera un perfecto ajuste de los datos en el cual $\hat{Y}_i = Y_i$, lo cual es un improbable evento en la práctica,

3. Si $\hat{Y}_i = Y_i$, esto es si $b_1 = b_2 = \cdots = b_{p-1} = 0$(suponiendo que el modelo $Y = \beta_0 + \varepsilon$ha sido ajustado), entonces R^2 =1.

4. R^2 es una medida de la utilidad de los términos en el modelo diferentes de β_0

El coeficiente de correlación de Spearman es una variante del coeficiente de correlación de Pearson en la que, en lugar de medir el grado de asociación lineal a partir de los propios valores de las variables, se mide a partir de la asignación de rangos a los valores ordenados (no paramétrico).

Estadístico Chi Cuadrado (χ^2)

El objetivo es conservar atributos que permitan diferenciar entre buenos y malos a la vez que se mantiene una relación buenos/malos que refleje la población. Lo que se buscará, en términos estadísticos, es determinar si existe una relación de dependencia[16] entre los atributos establecidos y las clasificaciones.

Para analizar si existe asociación entre variables cualitativas se puede usar el estadístico χ^2 de Pearson. Este contraste de homogeneidad, (o de independencia, que asume igualdad en todas las clases o categorías) mediante la prueba Chi-cuadrado entre dos variables cualitativas, se basa en la comparación de las frecuencias observadas con las frecuencias esperadas, estas últimas construidas bajo la hipótesis de independencia[17].

16 Se puede decir que existe una relación de dependencia si las variables no son independientes.

17 Ernesto Gonzalo Hernández, *Relación entre Variables Cualitativas*, 2000, p. 2.

La prueba por ende testea la siguiente hipótesis:

H_o: Las variables son independientes

H_a: Las variables no son independientes

Si las dos variables son independientes se puede expresar este supuesto, en términos de probabilidades, como:

$$p_{ij} = p_{i.}p_{.j}, \qquad i = 1,2,\ldots.a;\ j = 1,2,\ldots.b.$$

Para calcular el estadístico Chi-cuadrado que va a permitir contrastar esta hipótesis se debe primero construirse la tabla de contingencia ($a_x b$) con las frecuencias absolutas observadas n_{ij} resultado de contar el número de individuos para cada par de posibilidades de los distintos niveles i de la primera variable y j de la segunda variable.

Estadístico radio del logaritmo de verosimilitud

Un mejor criterio para determinar que variable debe ser excluida de un modelo es test de *Radio Verosimilitud* (**LR**). Para estimar este estadístico, se calcula el valor del logaritmo de la verosimilitud con cada variable que es borrada y se la compara con el valor de verosimilitud del modelo antes de que se borre la variable. El estadístico **LR** se obtiene de la diferencia del valor de la verosimilitud del modelo completo y la verosimilitud del modelo reducido a través de la siguiente relación[18]:

---

18 Diego Calvache y Freddy Carranza*, Diseño y Elaboración Estadística de un Sistema de Evaluación para la Otorgación de Crédito de Consumo en una Institución Financiera*, Quito, EPN, 2000, p. 207.

$$LR = 2 * \left[l(\pi_{\max}; y) - l(\pi; y) \right]$$

Donde:

$\pi_{\max}$, es el valor de la verosimilitud para un modelo con todas las variables.

π , es el valor de verosimilitud estimado para el modelo reducido

Si la hipótesis nula es verdadera en una muestra de tamaño suficientemente grande, este estadístico sigue una distribución Chi-cuadrado con *r* grados de libertad, donde *r* es la diferencia entre el número de términos del modelo completo y el modelo reducido.

La corrección por continuidad de Yates

Consiste en restar 0,5 puntos al valor absoluto de las diferencias entre las frecuencias observadas y las frecuencias esperadas (antes de elevarlas al cuadrado). La corrección de Yates, al tratarse simplemente de una corrección por continuidad, se interpreta exactamente igual que el estadístico Chi-cuadrado de Pearson. Dado que las diferencias entre lo observado en la muestra y lo esperado bajo la hipótesis nula son estadísticamente significativas[19].

Tal como lo mencionamos en el capítulo anterior y tratando de ejemplificar el uso y la interpretación de los estadísticos que ayudan a definir tanto las variables discriminantes como la calidad del modelo de gestión de riesgo de crédito que se ha

19 Pablo Andrés Salgado, *El uso de la estadística en electrofisiología*, 2007, p. 87.

de utilizar, en las siguientes líneas abordaremos brevemente las características del modelo Logit y la forma como se trabaja con las variables según su tipo.

El modelo Logit *(Modelo Logístico)*

Se estructura el modelo de la siguiente forma:

$$Y = f(\beta_1 + \beta_2 X_2 + ... + \beta_k X_k) + u$$

$$f(z) = \frac{\exp(z)}{1 + \exp(z)}$$

Por tanto, tendremos que:

$$E[Y] = P(Y = 1) = \frac{\exp(\beta_1 + \beta_2 X_2 + ... + \beta_k X_k)}{1 + \exp(\beta_1 + \beta_2 X_2 + ... + \beta_k X_k)}$$

Siendo Y la variable dependiente (para nuestro caso Y será igual a tipo de cliente), de la cual pretendemos encontrar el valor de probabilidad de ser mal cliente (es decir que Y tome el valor de 1). Como ya se ha mencionado anteriormente en modelos no lineales como el *Logit* y *Probit*, la estimación de los parámetros se realiza mediante el método de máxima verosimilitud. Además, en este tipo de modelos no resulta posible interpretar directamente las estimaciones de los parámetros β, ya que son modelos no lineales. Lo que haremos, en la práctica, es fijarnos en el signo de los estimadores, si el estimador es positivo, significará que incrementos en la variable asociada causan incrementos en P (Y= 1). Por el contrario, si el estimador muestra un signo negativo, ello supondrá que incrementos en la variable asociada causarán disminuciones en P (*Y* = 1).

En el modelo *Logit* se suelen usar otros dos conceptos para profundizar más en la interpretación de los estimadores:

Se llama *odds* al siguiente cociente de probabilidades:

$$Odds = \frac{P(Y=1)}{1-P(Y=1)} = \exp(\beta_1 + \beta_2 X_2 + ... + \beta_k X_k)$$

Tomando logaritmos neperianos en la expresión anterior, obtenemos una expresión lineal para el modelo:

$$Logit[P(Y=1)] \equiv \ln\left(\frac{P(Y=1)}{1-P(Y=1)}\right) = \beta_1 + \beta_2 X_2 + ... + \beta_k X_k$$

Aquí se aprecia claramente que el estimador del parámetro β_2 se podrá interpretar como la variación en el término Logit (el logaritmo neperiano del cociente de probabilidades) causada por una variación unitaria de la variable X_2 suponiendo constantes el resto de variables explicativas.

Cuando se hace referencia al incremento unitario en una de las variables explicativas del modelo, aparece el concepto de *odds-ratio* como el cociente entre los dos *odds* asociados (el obtenido tras realizar el incremento y el anterior al mismo). Así, si suponemos que ha habido un incremento unitario en la variable X_i, tendremos:

$$Odds - ratio = \frac{Odds2}{Odds1} = \exp(\beta_i)$$

De la expresión anterior se deduce que un coeficiente β_i cercano a cero, significará que cambios en la variable explicativa X_i asociada no tendrán efecto alguno sobre la variable dependiente *Y*.

El Modelo Logístico frente a otros modelos de clasificación

Más allá de los resultados de predictibilidad, ajuste y consistencia más discretos que tienen tanto el análisis discriminante como la regresión lineal, la regresión logística es utilizada en parte porque le permite al analista superar muchas de las suposiciones restrictivas anteriormente mencionadas:

- La regresión logística no asume una relación lineal entre las variables dependiente e independiente. Por lo tanto, puede manejar efectos no lineales incluso con términos exponenciales o polinomiales que no son explícitos o evidentes. Este tipo de variables son incluidas debido a que la función discriminante logit del lado izquierdo de la ecuación de la regresión logística es no lineal. Adicionalmente, es posible si fuese necesario incluir explícitamente dichas interacciones y poner como términos exponenciales o polinómicos dichas variables en el lado derecho de la ecuación logística.
- La variable dependiente no necesita ser normalmente distribuida, aunque se asume que su distribución forma parte de la familia de distribuciones exponenciales.
- La variable dependiente no necesita ser homocedástica para cada nivel de las independientes; es decir, no existe la suposición de la homogeneidad de la varianza.
- No se asume que los términos del error estén normalmente distribuidos.
- La regresión logística no requiere que las variables independientes sean continuas.
- La regresión logística no requiere que las variables independientes no sean acotadas.

- El modelo logit ya me proporciona la probabilidad de que un cliente sea buen o mal cliente, a diferencia del modelo lineal o el análisis discriminante.
- En términos de la gestión de riesgo crediticio el modelo logístico es uno de los más utilizados por diversas empresas de consultoría dedicadas a la generación y creación de moldeos de score crediticio.

Por las razones expuestas en los párrafos anteriores, desarrollaremos un pequeño ejemplo con una regresión logística, no con la finalidad de encontrar el mejor modelo de clasificación ya que esa no es la finalidad de nuestro estudio, sino más bien a fin de ejemplificar la manera cómo se ha de tratar con variables con diferentes propiedades, sean estas categóricas, continuas, discretas, etc. Procederemos entonces a trabajar con la base de datos depurada, es decir con los 13065 clientes que nos quedaron una vez que se eliminaros los datos que presentaban inconsistencias, para ello vamos a recordar la definición de tipo de cliente presentada anteriormente

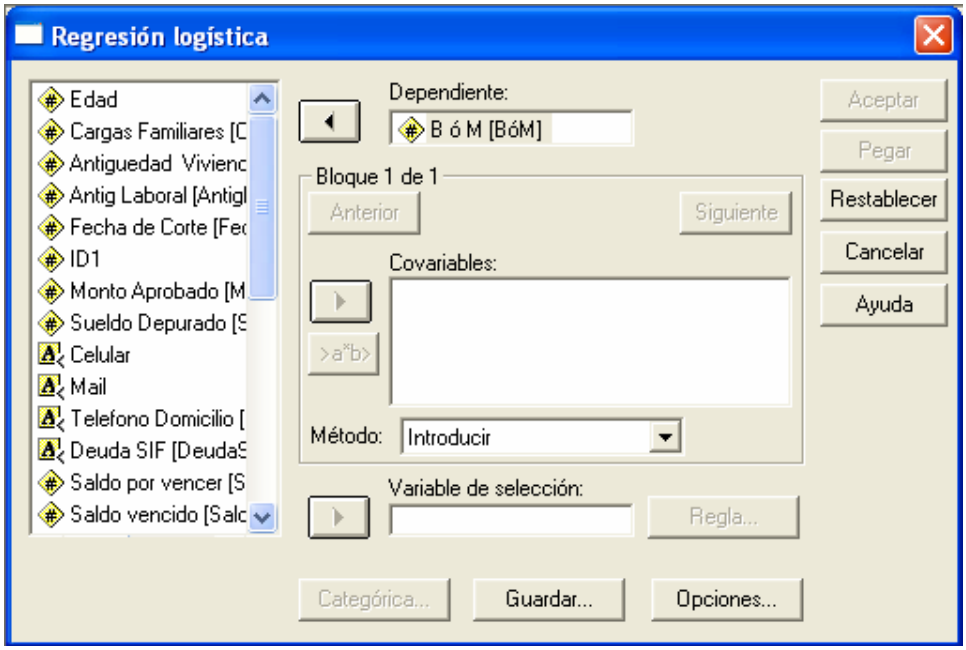
Mal cliente: es aquel con una mora promedio superior o igual a 15 días y una mora máxima superior o igual a 30 días.

Mediante la utilización del paquete estadístico SPSS, trabajaremos con el siguiente grupo de variables:

Nombre de la Variable	Descripción	Tipo de variable
Profesión	Area de especialización del cliente	Cualitativa
Edad		Cuantitativa
Sexo	Género del cliente	Cualitativa
Cargas Familiares	No. de personas que dependen del cliente	Cuantitativa
Telefono Domicilio	Tenencia del cliente de teléfono en domicilio	Cualitativa

Vamos a generar un modelo logístico con las variables presentadas tal cual se encuentran en la base de datos de trabajo y adicionalmente si la variable así lo

requiere realizaremos las transformaciones que se consideren necesarias según corresponda. En primer lugar haremos uso de la variable profesión, al tratar de utilizarla en el modelo planteado nos encontramos con la dificultad que una variable cualitativa como esta no puede ser incluida directamente en el modelo de regresión, es más ni siquiera aparece en el listado de variables posible, según se puede apreciar en la siguiente pantalla.

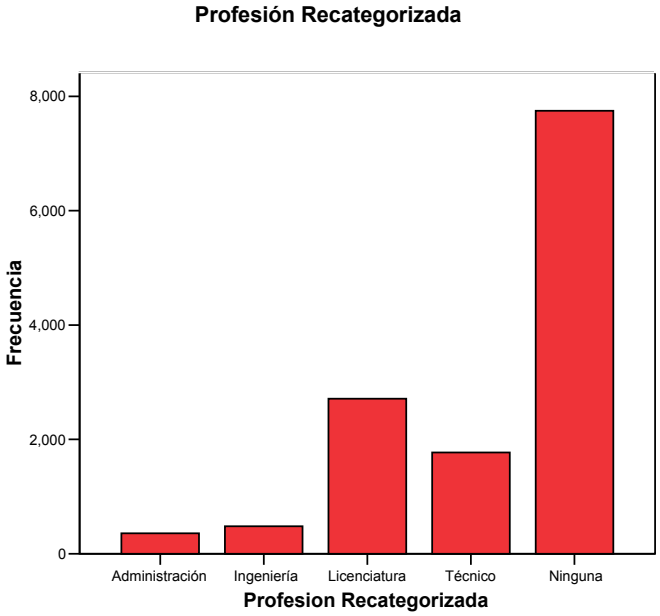


Por tal razón y al tratarse de una variable categórica, en la que anteriormente habíamos determinado la existencia de alrededor de 20 profesiones diferentes, muchas de las cuales no alcanzan el cinco por ciento (5%) de participación en la base de trabajo, para casos similares al de esta variable que contienen muchas categorías es preferible realizar una reagrupación de las mismas, de modo que no sobrepasen cuatro o máximo cinco nuevos conjuntos. Bajo este concepto y procurando agrupar carreras afines se obtuvo una nueva variable denominada Profesión Recategorizada notada como ProfCod con cinco grupos, donde todas las profesiones como administración, auditoria y economía fueron congregadas en un

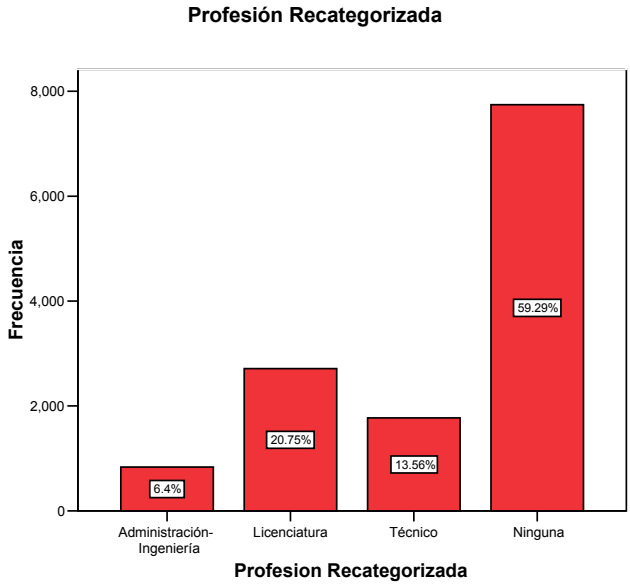
solo grupo denominado Administración; profesiones que se relacionan con medicina, ingeniería y arquitectura entraron a un segundo conjunto, en el grupo notado como Licenciatura están los clientes con profesiones tal como policía, militar, profesor, etc. quedando en un cuarto grupo todas las carreras técnicas de la base de datos y un grupo final con aquellos clientes que señalaron no tener profesión alguna, según se puede observar en la siguiente tabla.

Tabla de contigencia Profesión Recategorizada & B ó M

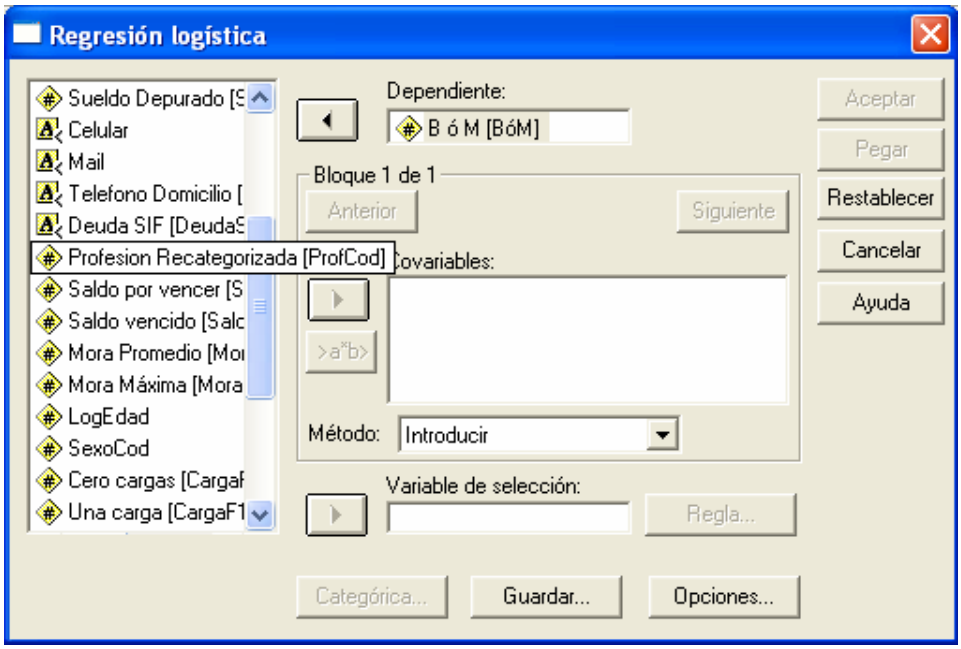
			B ó M		Total
			Bueno	Malo	
Profesión Recategorizada	Administración	Recuento	220,0	136,0	356,0
		% del total	1,7%	1,0%	2,7%
	Ingeniería	Recuento	331,0	149,0	480,0
		% del total	2,5%	1,1%	3,6%
	Licenciatura	Recuento	1.700,0	1.011,0	2.711,0
		% del total	13,0%	7,7%	20,7%
	Técnico	Recuento	1.118,0	654,0	1.772,0
		% del total	8,6%	5,0%	13,6%
	Ninguno	Recuento	4.961,0	2.785,0	7.746,0
		% del total	38,0%	21,3%	59,3%
Total		Recuento	8.330,0	4.735,0	13.065,0
		% del total	63,8%	36,2%	100,0%



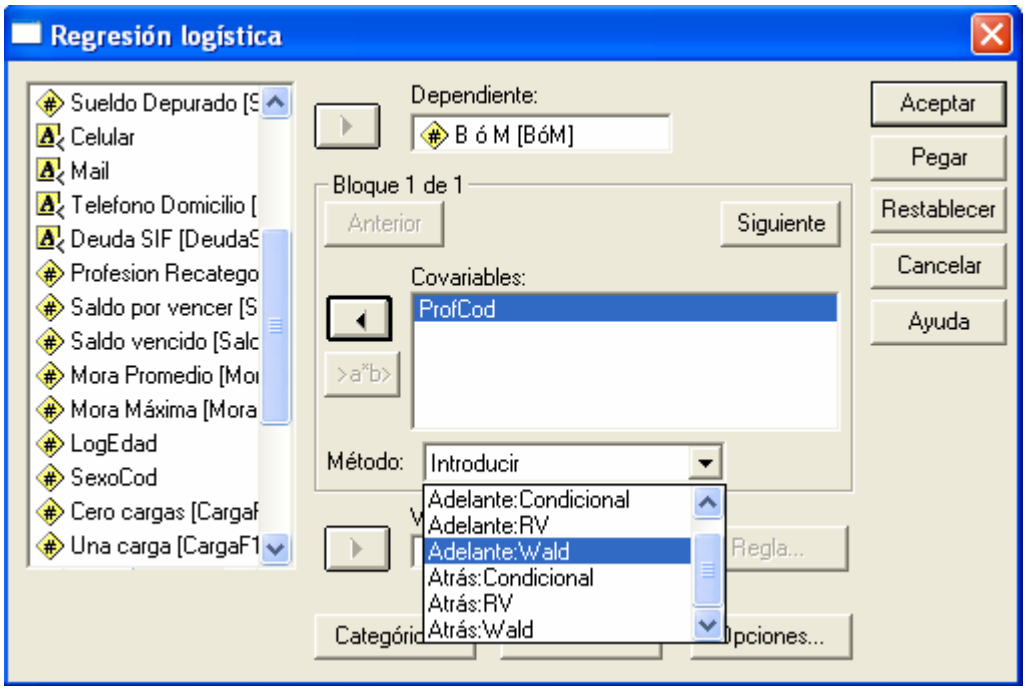
Pese al trabajo realizado, vemos que aún existen dos categorías que tienen menos del cinco por ciento de la población, por tanto es recomendable agruparlas en una sola clase que podríamos llamarla Administración – Ingeniería.



Una vez realizada la transformación sugerida a la variable veamos los resultados que se puedan alcanzar en el caso del modelo logístico.

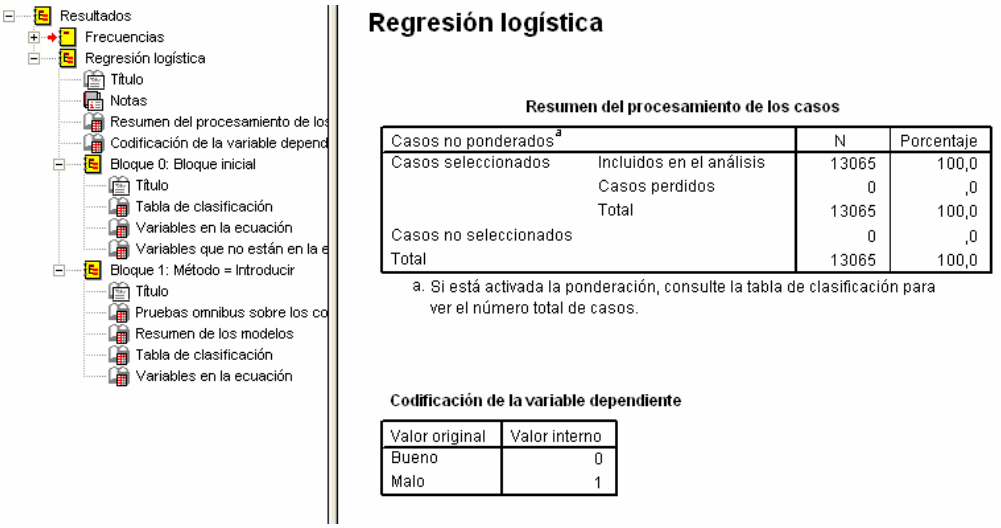


Verificamos ahora sí la presencia de la variable transformada, la cual ya puede ser incluida en el modelo en construcción.



En la gráfica anterior se puede apreciar cómo debería ser la carga de las variables en un paquete estadístico que disponga de la opción de regresión logística, en este caso se muestra una pantalla del programa SPSS el cual permite escoger el método para incorporar (o extraer) variables al modelo, para nuestro ejemplo se ha optado por la opción Introducir (Enter) la cual permite paso a paso verificar la interacción de la variable seleccionada al modelo a aplicar.

Empezando por la nueva variable Profesión Recategorizada (ProfCod) vamos a analizar las pantallas de salida que se puede obtener al modelizar un score de crédito basado en regresión logística, las pantallas que veremos a continuación se repiten cada vez que se incorpora o extrae una variable del modelo, por tal razón procuraremos mencionarlas con detalle solo en una ocasión.



Esta pantalla inicial nos da un resumen de los casos en estudio, que como habíamos señalado anteriormente la base de trabajo se redujo a 13.065 clientes, adicionalmente se hace referencia a la codificación utilizada para la variable dependiente que en nuestro caso es tipo de cliente, donde se asigna con el número uno (1) a los malos clientes y cero (0) a los buenos clientes.

Bloque 0 : Bloque Inicial

Tabla de clasificación a, b

			Pronosticado		
			B ó M		Porcentaje
	Observado		Bueno	Malo	correcto
Paso 0	B ó M	Bueno	8330	0	100,0
		Malo	4735	0	0,0
	Porcentaje global				63,8

a. En el modelo se incluye una constante
b. El valor de corte es 0,500

Variables en la ecuación

		B	E.T	Wald	gl	Sig.	Exp (B)
Paso	Constante	-0,565	0,018	963,32	1	0,00	0,568

Variables que no están en la ecuación

			Puntuación	gl	Sig.
Paso 0	Variables	Prof Cod	0,002	1	0,960
	Estadísticos globales		0,002	1	0,960

La segunda pantalla por su lado nos muestra el punto de partida sobre el cual girará la evolución del modelo en cuanto a su capacidad de discriminación, se nos muestra la llamada matriz de confusión (Tabla de clasificación), que reasume la cantidad de clientes buenos y malos observados vs. la predicción del modelo, por tratarse del punto de partida se puede observar 8.330 clientes buenos catalogados como tal y 4.735 clientes malos también clasificados como buenos, el poder de predicción de este modelo sin haber hecho absolutamente nada es del 63,8% con un 100% de predicción de los clientes buenos pero 0% para los clientes malos.

Luego se puede observar una pequeña tabla con valores para el modelo hasta aquí resultante aplicado sólo sobre la constante, tenemos los valores de los coeficientes de la variable (columna B), para el error de estimación (columna, E.T.), el estadístico de Wald, los grados de libertad de la variable (columna gl) el nivel de significación de la variable (columna Sig.) y el exponencial del coeficiente de la variable Exp(B) (recordemos que estamos trabajando en un modelo Logit).

De estos valores el que mayor importancia tiene en la modelación es el correspondiente a la significación, pues este valor determinará si se acepta o no la incorporación de la variable seleccionada al modelo en construcción según la especificidad del modelo al 5 o al 10% de nivel de confianza. Finalmente en esta pantalla del bloque inicial (Bloque 0) aparecerán las variables que no están en la ecuación, en este caso ProfCod.

Pruebas sobre los coeficientes del modelo

		chi-cuadrado	gl	Sig.
Paso 1	Paso	0,002	1	0,960
	Bloque	0,002	1	0,960
	Modelo	0,002	1	0,960

Resumen de los modelos

Paso	_2 log de la verosimilitud	R cuadrado de Cox y Snell	R cuadrado de Nagelkerke
1	17109,847[a]	0,000	0,000

a. La estimación ha finalizado en el número de iteración 3 porque las estimaciones de los parámetros han cambiado en menos de 0,001

En la tercera pantalla presentada, tenemos la primera parte de la información correspondiente al Bloque 1, aquí se muestran fundamentalmente los valores del estadístico Chi-cuadrado (Ch2) y de Log de Verosimilitud (LogV) que son los que determinarán mejoras en el modelo a causa de la inclusión de una nueva variable al modelo, que en este caso señalan un valor de 0,02 con un grado de libertad y 17.109,87 respectivamente.

Tabla de clasificación [a]

			Pronosticado		
			B ó M		Porcentaje
	Observado		Bueno	Malo	correcto
Paso 0	B ó M	Bueno	8330	0	100,0
		Malo	4735	0	0,0
	Porcentaje global				63,8

a. valor de corte es 0,500

Variables en la ecuación

		B	E.T	Wald	gl	Sig.	Exp (B)
Paso 1 [a]	ProfCod	0,001	0,016	0,002	1	0,960	1,001
	Constante	-0,568	0,068	68,921	1	0,000	0,567

a. variable (s) introducida (s) en el paso 1: ProfCod.

Al igual que en el Bloque 0 tenemos la tabla de clasificación que no presenta variaciones en cuanto a la información del modelo, la tabla que si presenta variantes es la de los estadísticos (Variables en la ecuación) donde podemos constatar que la variable introducida al modelo no pasa las pruebas de significación al 5% pues tiene un valor de 0,96 muy superior al límite fijado. Como adicionalmente el modelo tiene una constante se puede intentar probar nuevamente la variable ProfCod pero sin la constante para saber si entra o no a la ecuación.

Pruebas sobre los coeficientes del modelo

		chi-cuadrado	gl	Sig.
Paso 1	Paso	931,557	1	0,000
	Bloqueo	931,557	1	0,000
	Modelo	931,557	1	0,000

Resumen de los modelos

Paso	_2 log de la verosimilitud	R cuadrado de Cox y Snell	R cuadrado de Nagelkerke
1	17180,379[a]	0,069	0,092

a. La estimación ha finalizado en el número de literación 3 porque las estimaciones de los parámetros han cambiado en menos de 0,001

Según se puede constatar en el cuadro anterior, efectivamente los indicadores del modelo cambian al quietar la constante como elemento inicial del modelo, se puede observar que el estadístico Chi-cuadrado ahora se ha elevado a 931,55, mientras que el logaritmo de verosimilitud pasa de 17.109,87 a 17.180,37. Adicionalmente si revisamos los estadísticos de la variable en este nuevo paso, verificamos que el nivel de significación de ProfCod satisface la condición al 5%, según se puede observar en la tabla de Variables de la ecuación, es decir la variable se constituye (al menos momentáneamente) como elemento de discriminación para la variable tipo de cliente.

Variables en la ecuación

		B	E.T	Wald	gl	Sig.	Exp (B)
Paso 1[a]	ProfCod	-0,125	0,004	892,208	1	0,000	0,882

a. variable (s) introducida (s) en el paso 1: ProfCod.

Nombre de la Variable	Descripción	Tipo de variable
Profesión	Area de especialización del cliente	Cualitativa ✓
Edad		Cuantitativa
Sexo	Género del cliente	Cualitativa
Cargas Familiares	No. de personas que dependen del cliente	Cuantitativa
Telefono Domicilio	Tenencia del cliente de teléfono en domicilio	Cualitativa

Ahora el siguiente paso consiste en incorporar nuevas variables al modelo contribuyan a la identificación de los buenos y malos clientes, para ello del listado planteado tomamos la variable edad y la ingresamos al y verificamos las pantallas finales.

Pruebas sobre los coeficientes del modelo

		chi-cuadrado	gl	Sig.
Paso 1	Paso	976,572	2	0,000
	Bloqueo	976,572	2	0,000
	Modelo	976,572	2	0,000

Resumen de los modelos

Paso	_2 log de la verosimilitud	R cuadrado de Cox y Snell	R cuadrado de Nagelkerke
1	17135,364[a]	0,072	0,096

a. La estimación ha finalizado en el número de literación 2 porque las estimaciones de los parámetros han cambiado en menos de 0,001

En este caso nos encontramos con un incremento en el valor de Ch2 y una ligera disminución de LogV, adicionalmente verificamos el poder de predicción de la variable, así como la significación de las variables.

Tabla de clasificación [a]

Observado			Pronosticado: B ó M		Porcentaje correcto
			Bueno	Malo	
Paso 0	B ó M	Bueno	8330	0	100,0
		Malo	4735	0	0,0
Porcentaje global					63,8

a. valor de corte es 0,500

Variables en la ecuación

		B	E.T	Wald	gl	Sig.	Exp (B)
Paso 1 [a]	ProfCod	-0,056	0,011	25,733	1	0,000	0,945
	Edad	-0,009	0,011	44,380	1	0,000	0,991

a. variable (s) introducida (s) en el paso 1: ProfCod.

Según los cuadros presentados, vemos que la predicción del modelo no mejora aunque las dos variables introducidas mantienen un nivel de significación que pasan las pruebas. En este sentido y analizando con un poco más de detenimiento la variable edad mediante sus estadísticas descriptivas encontramos que el producto crediticio en estudio presenta clientes con edades entre 22 y 67 años.

Estadísticos descriptivos

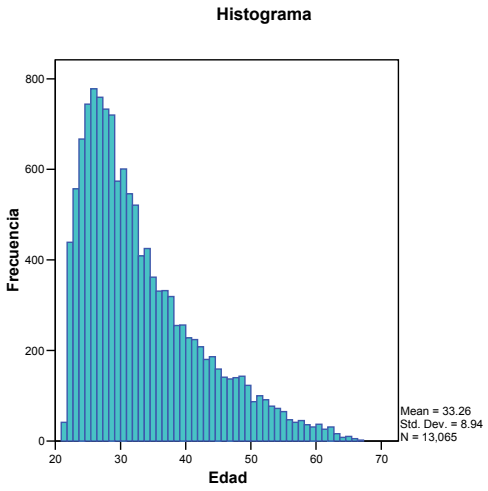
	N	Mínimos	Máximos	Media	Desviación Std.
Edad	13065	22	67	33,26	8,940

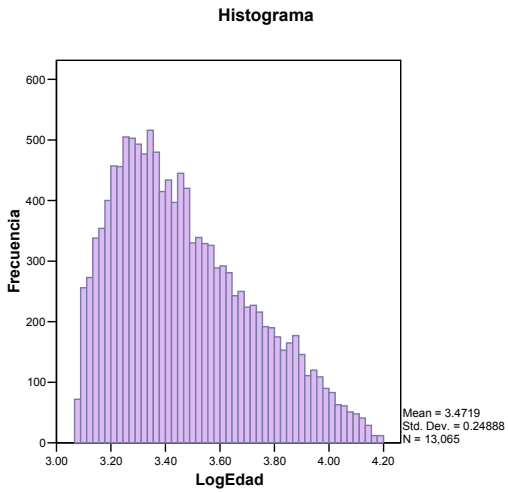
Edad

No. Obs.	Válidos	13065
Percentiles	25	26,41
	50	30,81
	75	37,97

Por otro lado se puede constatar que un 25% de la población tiene una edad inferior a 27 años, la media poblacional supera a la mediana la cual señala que la mitad de los clientes de la base de trabajo tiene menos de 31 años de edad, mientras que apenas el 25% de clientes restantes están en el grupo entre 38 años y la edad tope considerada para este producto crediticio. Por las características mencionadas de esta variable uno de los métodos que se sugiere para disminuir la

volatilidad de la misma es la utilización de logaritmos naturales, otra alternativa podría ser la de normalizar la variable respecto a su media o respecto a una medida específica, pero esto nuevamente dependerá del criterio y experiencia de la persona a cargo de la elaboración del score de crédito.





Como se esperaba según la gráfica anterior la dispersión de la variable se reduce considerablemente al aplicar el logaritmo natural a la variable edad y obtener la variable denominada LogEdad sus valores mínimos y máximos están entre 3 y 4,20 lo que permite que se pueda manejar con mayor facilidad y ver si esta transformación permite que la variable en sí contribuya a la capacidad de predicción del modelo en construcción.

Pruebas sobre los coeficientes del modelo

		Chi-cuadrado	gl	Sig.
Paso 1	Paso	1001,835	2	0,000
	Bloqueo	1001,835	2	0,000
	Modelo	1001,835	2	0,000

Resumen de los modelos

Paso	_2 log de la verosimilitud	R cuadrado de Cox y Snell	R cuadrado de Nagelkerke
1	17110,101[a]	0,074	0,098

a. La estimación ha finalizado en el número de literación 3 porque las estimaciones de los parámetros han cambiado en

Variables en la ecuación

		B	E.T	Wald	gl	Sig.	Exp (B)
Paso 1 [a]	ProfCod	-0,056	0,015	0,136	1	0,712	0,994
	LogEdad	-0,156	0,019	68,799	1	0,000	0,856

a. variable (s) introducida (s) en el paso 1: ProfEdad.

Nuevamente ocurre un resultado similar al obtenido con la variable edad, pero en este caso adicionalmente la primera variable introducida ProfCod deja de satisfacer la prueba de significancia por lo que será necesario retirarla del modelo y probar con la siguiente variable de nuestra lista.

Nombre de la Variable	Descripción	Tipo de variable	
Profesión	Area de especialización del cliente	Cualitativa	✓
Edad		Cuantitativa	✓
Sexo	Género del cliente	Cualitativa	
Cargas Familiares	No. de personas que dependen del cliente	Cuantitativa	
Telefono Domicilio	Tenencia del cliente de teléfono en domicilio	Cualitativa	

Para el caso de la variable sexo, esta no puede ser ingresada al modelo tal como se encuentra en la base de datos de trabajo, por lo que será necesario realizar una pequeña transformación de la variable para pasarla de cualitativa a categórica binaria, por ejemplo asignando el género femenino como unos (1) y el masculino como dos (2), para ello creamos una nueva variable denominada SexoCod, esta transformación es necesaria ya que el modelo logístico al igual que el modelo lineal no reconoce variables cualitativas a menos que se traten de variables dicotómicas o se las trabaje como variables dummy. Hecho este pequeño cambio procedemos a probarlo en el modelo en construcción.

Variables en la ecuación

		B	E.T	Wald	gl	Sig.	Exp (B)
Paso 1 [a]	LogEdad	-0,147	0,018	69,332	1	0,000	0,864
	SexoCod	-0,034	0,036	0,873	1	0,350	0,967

a. variable (s) introducida (s) en el paso 1: LogEdad, SexoCod.

El resultado es similar al encontrado anteriormente, ya que no satisface las pruebas de significación y es necesario retirarla del modelo para probar con la variable cargas familiares que es la siguiente en nuestro listado.

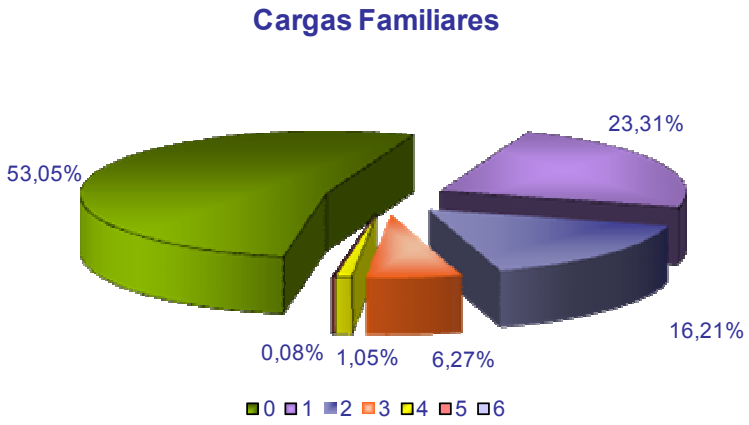
Nombre de la Variable	Descripción	Tipo de variable	
Profesión	Area de especialización del cliente	Cualitativa	✓
Edad		Cuantitativa	✓
Sexo	Género del cliente	Cualitativa	✓
Cargas Familiares	No. de personas que dependen del cliente	Cuantitativa	
Telefono Domicilio	Tenencia del cliente de teléfono en domicilio	Cualitativa	

Antes de probar esta variable en la regresión logística en construcción analizaremos el tratamiento que se pudiera dar a la variable en caso de necesitarlo, por tratarse de una variable discreta, básicamente puede optarse por dos alternativas, una de ellas sería proceder a reagrupar las cargas en dos o tres conjuntos, la segunda alternativa sería transformar la variable en nuevas variables dicotómicas representadas por ceros y unos según el número de cargas familiares, por ejemplo la variable CargaF0 representará a los clientes que tienen cero cargas familiares con el número uno (1) y con cero (0) a quienes tengan cargas familiares, la variable CargaF1 en cambio representará a los clientes que tengan una carga familiar con el número uno (1) y con cero (0) para quienes no tengan cargas familiares o tengan más de una carga familiar. Al observar la tabla de frecuencias de la variable y seguir con esta alternativa de análisis, podemos plantear la creación cinco variables nuevas, una que represente a cero cargas, otra a uno, dos, tres y finalmente una variable que reúna a todos aquellos clientes que tienen más de tres

cargas familiares, otro análisis pudiese haber resuelto crear menos grupos por ejemplo agrupando a los clientes con tres y más cargas en una sola variable.

Cargas Familiares

Validos	Frecuencia	Porcentaje	Porcentaje Valido	Porcentaje Acumulado
0	6931	53,1%	53,1%	53,1%
1	3046	23,3%	23,3%	76,4%
2	2118	16,2%	16,2%	92,6%
3	819	6,3%	6,3%	98,8%
4	137	1,0%	1,0%	99,9%
5	11	0,1%	0,1%	100,0%
6	3	0,0%	0,0%	100,0%
Total	13065	100,0%	100,0%	



Siguiendo el procedimiento escogido, generamos las variables CargaF0, CargaF1, CargaF2, CargaF3, CargaF4 cuya característica común es ser una variable discreta dicotómica que toma solo valores de ceros y unos según la construcción realizada. Ya fijada la alternativa de transformación de la variable cargas familiares, procedemos a incorporarla al modelo y analizamos sus resultados.

Pruebas sobre los coeficientes del modelo

		chi-cuadrado	gl	Sig.
Paso 1	Paso	1002,463	2	0,000
	Bloqueo	1002,463	2	0,000
	Modelo	1002,463	2	0,000

Resumen de los modelos

Paso	_2 log de la verosimilitud	R cuadrado de Cox y Snell	R cuadrado de Nagelkerke
1	17109,473[a]	0,074	0,098

a. La estimación ha finalizado en el número de literación 3 porque las estimaciones de los parámetros han cambiado en menos de 0,001

Variables en la ecuación

		B	E.T	Wald	gl	Sig.	Exp (B)
Paso 1 [a]	LogEdad	-0,166	0,007	600,376	1	0,000	0,847
	Cargas Familiares	0,016	0,018	0,765	1	0,382	1,016

a. variable (s) introducida (s) en el paso 1: LogEdad, Cargas Familiares

El efecto alcanzado no varía respecto a las ultimas corridas, ya que la variable incorporada no pasa las pruebas de significancia, excepto por un detalle adicional los valores de los estadísticos Ch2 y LogV no han variado su valor pese a tener en el modelo una variable codificada y una variable pura, por lo que podemos optar por probar nuevamente solo la variable pura y analizar su efecto.

Pruebas sobre los coeficientes del modelo

		chi-cuadrado	gl	Sig.
Paso 1	Paso	380,334	1	0,000
	Bloqueo	380,334	1	0,000
	Modelo	380,334	1	0,000

Resumen de los modelos

Paso	_2 log de la verosimilitud	R cuadrado de Cox y Snell	R cuadrado de Nagelkerke
1	17731,602[a]	0,029	0,038

a. La estimación ha finalizado en el número de literación 3 porque las estimaciones de los parámetros han cambiado en menos de 0,001

Tabla de clasificación [a]

Observado			Pronosticado B ó M		Porcentaje correcto
			Bueno	Malo	
Paso 1	B ó M	Bueno	3898	4432	46,8
		Malo	2236	2499	52,8
Porcentaje global					49,0

a. valor de corte es 0,500

Variables en la ecuación

		B	E.T	Wald	gl	Sig.	Exp (B)
Paso 1[a]	Cargas Familiare	-0,275	0,015	357902	1	0,000	0,760

a. variable (s) introducida (s) en el paso 1: Cargas Familiares

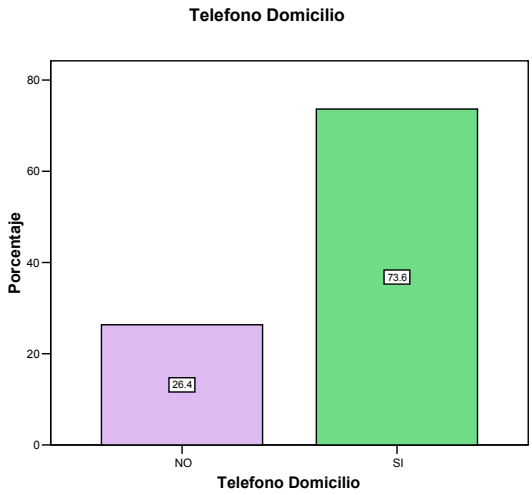
La variable log edad estaba distorsionando los resultados, se debe empezar de nuevo a probar con el resto de variables y ver los resultados, pero como se había manifestado anteriormente, aquí se busca ejemplificar brevemente el trabajo a seguir como parte del arduo proceso de modelación. Para terminar con las variables planteadas para nuestro ejemplo tomaremos la siguiente en nuestra lista.

Nombre de la Variable	Descripción	Tipo de variable	
Profesión	Area de especialización del cliente	Cualitativa	✓
Edad		Cuantitativa	✓
Sexo	Género del cliente	Cualitativa	✓
Cargas Familiares	No. de personas que dependen del cliente	Cuantitativa	✓
Telefono Domicilio	Tenencia del cliente de teléfono en domicilio	Cualitativa	

Trabajar con la variable teléfono en el domicilio es mucho más directo y no necesita de ningún procedimiento adicional, pues se trata de una variable categórica que puede tomar dos valores 1 si el cliente tiene teléfono y 0 en caso contrario, esta variable entrará tal cual al modelo de regresión.

Teléfono Domicilio

Validos	Frecuencia	Porcentaje	Porcentaje válido	Porcentaje acumulado
NO	3443	26,35%	26,35%	26,35%
SI	9622	73,65%	73,65%	100,00%
Total	13065	100,00%	100,00%	



Incorporada la variable a la regresión logística tenemos los siguientes resultados:

Tabla de clasificación [a]

Observado			Pronosticado B ó M		Porcentaje correcto
			Bueno	Malo	
Paso 1	B ó M	Bueno	7232	1098	86,8
		Malo	4073	662	14,0
	Porcentaje global				60,4

a. valor de corte es 0,500

Variables en la ecuación

		B	E.T	Wald	gl	Sig.	Exp (B)
Paso 1 [a]	Cargas Familiare	-0,095	0,017	32,597	1	0,000	0,909
	Tef. Domicilio	-0,515	0,025	437,204	1	0,000	0,597

a. variable (s) introducida (s) en el paso 1: Cargas Familiares, Teléfono Domicilio

En términos generales los valores de Ch2 y LogV aumentan pues pasan de 380,33 y 17.731,60 a 828,82 y 17.283,11 respectivamente, incluso se evidencia un incremento en el valor de predicción global del modelo al pasar de 49,0% a 60,4%, lo que indica una aparente mejoría del modelo pero a costa de sacrificar el poder de discriminación para los malos clientes, adicionalmente la variable incorporada pasa la prueba de significación al 5% esa decir si puede ser considerada dentro del

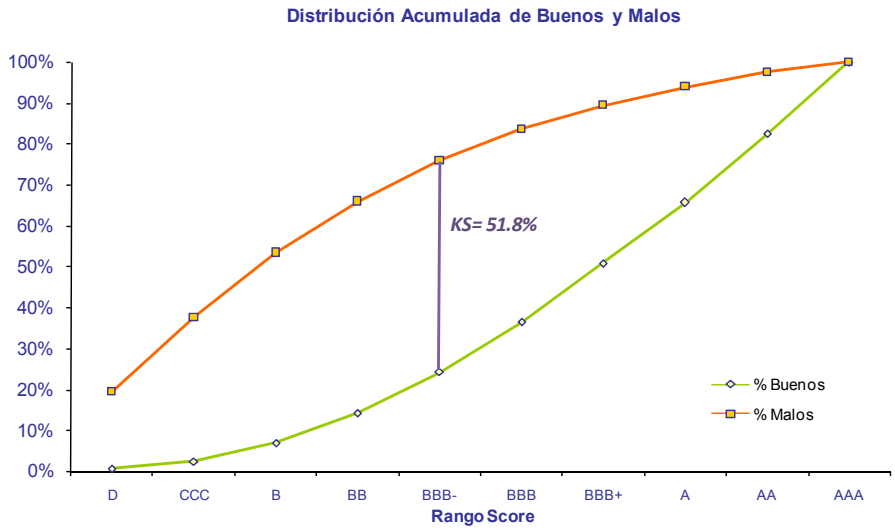
modelo final. El siguiente paso en este mini proceso de modelación sería volver a probar aquellas variables que se descartaron al no pasar las pruebas en los pasos anteriores.

Nombre de la Variable	Descripción	Tipo de variable	
Profesión	Area de especialización del cliente	Cualitativa	✓
Edad		Cuantitativa	✓
Sexo	Género del cliente	Cualitativa	✓
Cargas Familiares	No. de personas que dependen del cliente	Cuantitativa	✓
Telefono Domicilio	Tenencia del cliente de teléfono en domicilio	Cualitativa	✓

Una vez que hemos analizado todas las variables propuestas para este corto ejemplo y si bien es cierto hemos logrado encontrar un modelo de clasificación de clientes, no se ha profundizado en los detalles complementarios que permitan obtener un modelo que sea robusto desde el punto de vista estadístico, así como coherente y aplicable desde una perspectiva económica y crediticia, ya que solo el proceso de modelación implica un tratamiento particularizado que sobrepasa el objetivo del presente estudio. En forma complementaria al proceso de modelación mencionaremos algunos de los métodos más utilizados para determinar el poder de clasificación de un score de crédito y que son aplicables a modelos logísticos, lineales o de arboles de decisión, tales como la pruebas de Kolmogorov, el coeficiente de Gini, la prueba de Hosmer y la curva ROC.

Prueba de Kolmogorov Smirnov (*KS*)

Esta prueba permite determinar la consistencia del modelo, para ello mide la distancia entre las distribuciones acumuladas del modelo aplicado a las muestras de la estimación, por rangos de probabilidad. Se encuentra la distancia máxima entre ambas distribuciones, el cual, si es mayor al comparado con el valor crítico calculado al 99% de nivel de confianza, se demuestra que en cada rango se evidencian comportamientos diferentes y por tanto la discriminación es adecuada[20].

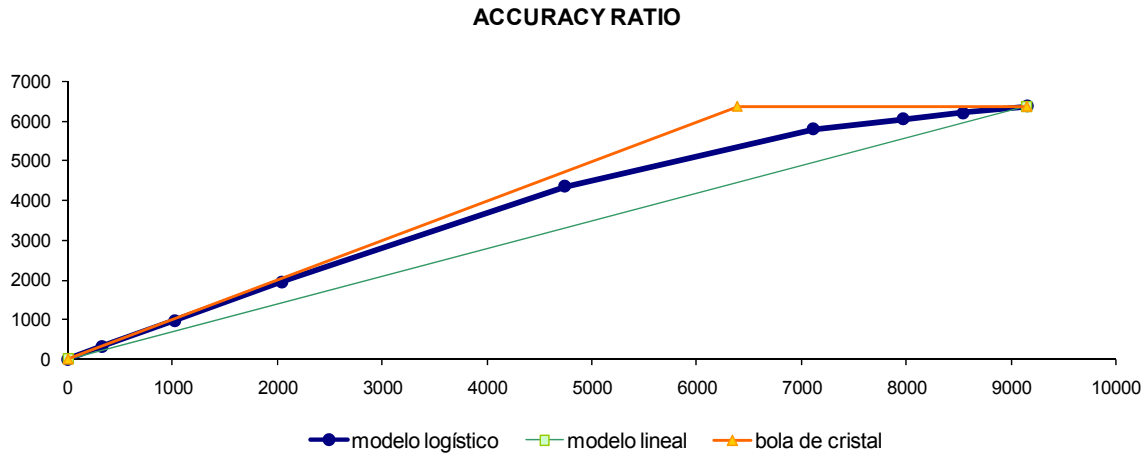


El Coeficiente de GINI *(Accuracy Ratio)*

Mide el grado de aciertos del modelo, en función de las probabilidades asignadas por el mismo y el comportamiento real (default o no). Mientras más alto mejor, lo que significa que el modelo propuesto se acerca al modelo perfecto.

---

20 Luis Miguel Molinero, *¿Y si los datos no siguen una distribución normal? Bondad de ajuste a una normal, transformaciones, pruebas no paramétricas*, Asociación de la Sociedad Española de Hipertensión, 2003, p. 2.



Prueba de Hosmer y Lemeshow

Evalúa la asignación de observaciones entre segmentos de probabilidad estimada, relacionándola con aquella que presenta la muestra o población. Si el nivel de significación de la prueba es menor que el 1% (0,01), se sugiere un buen desempeño del modelo[21].

Tabla de contingencia para la prueba de Hosmer y Lemeshow

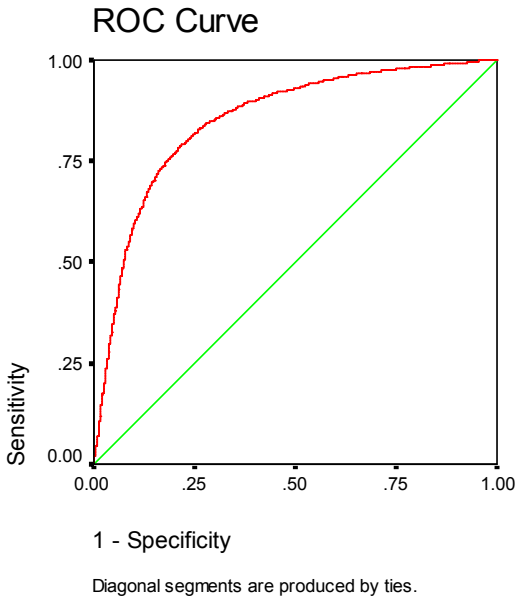
Paso 1	B ó M = Bueno		B ó M = Malo		Total
	Observado	Esperado	Observado	Esperado	
1	437	468,132	238	206,868	675
2	1004	1027,404	531	507,596	1535
3	1410	1452,144	831	788,856	2241
4	0	0,639	1	0,361	1
5	3334	3237,321	1837	1933,679	5171
6	1047	909,458	635	772,542	1682
7	1098	880,000	662	880,000	1760

21 Eva Medina Moral, *Modelos de Elección Discreta*, 2003, p. 20.

La curva ROC

Son curvas en las que se presenta la sensibilidad en función de los falsos positivos (complementario de la especificidad) para distintos puntos de corte[22]. Si la prueba fuera perfecta, es decir, sin solapamiento, hay una región en la que cualquier punto de corte tiene sensibilidad y especifidad iguales a 1: la curva sólo tiene el punto (0,1).

Si la prueba fuera inútil: ambas distribuciones de probabilidad coinciden y la sensibilidad (verdaderos positivos) es igual a la proporción de falsos positivos, la curva sería la diagonal de (0,0) a (1,1).



22 James A. Hanley, Barbara J. McNeil, *The meaning and use of the area under a receiver operating characteristic (ROC) curve*. Radiology. 1982 143: 29-36

IV. CONCLUSIONES

El desarrollo de cualquier modelo de crédito por más simple que éste sea, implica una serie de pasos o tareas que no se pueden evadir, tales como: la preparación de la base datos, la elección de la metodología a aplicar según el objetivo planteado, la selección de la muestra de trabajo; y, las pruebas de control, entre otras. Pero sin lugar a duda como hemos comprobado a lo largo de este trabajo, una de las etapas más importantes constituye el análisis y preparación de los datos a utilizar de modo que los resultados logrados por el modelo respondan eficaz y eficientemente al objetivo planteado.

Esta fase inicial en la construcción de un modelo credit score de gestión de riesgo de crédito parte con la preparación de la base de datos, misma que a la postre se convertirá en información útil para alcanzar el objetivo planteado del modelo sea este la clasificación o calificación de clientes. Para ello fue necesario constatar que los datos con los cuales se estaba trabajando mantengan consistencia y coherencia entre sí y respecto al producto crediticio a modelar, es necesario partir de una codificación o representación numérica de las características tanto cualitativas como cuantitativas primordiales de las variables implícitas en la técnica de modelación escogida.

De lo anterior se desprende el objetivo fundamental del presente trabajo, al plasmar las bases metodológicas que establecen las diferencias en el tratamiento y análisis de variables cualitativas y cuantitativas, su representación gráfica, las estadísticas que las diferencian, así como la interpretación de su información y el aporte que dichos datos pueden proporcionar al modelo de gestión de riesgo de crédito elaborado. El análisis de estabilidad de los datos de la base de trabajo debe

estar acompañado de una depuración de datos aberrantes, valores perdidos y delimitación de cotas máximas y mínimas de cada variable, que sobre todo respondan a las políticas institucionales establecidas para el producto sobre el que se construye el score crediticio.

Hemos evidenciado la importancia del análisis minucioso y exhaustivo a los datos de crédito de una institución de este modo podremos obtener información valiosa y eficiente para una adecuada gestión de riesgos mediante la aplicación de un credit score, para lo cual es necesario dar los siguientes pasos:

- Analizar la integridad de la información.
- Definir la variable de decisión (clientes buenos y malos, aceptados y rechazados, etc.).
- Analizar el período de exposición.
- Definir la ventana de aplicación.
- Determinar el modelo de clasificación a utilizar.
- Escoger el punto de corte.

Finalmente es necesario señalar que todos los pasos detallados en este trabajo constituyen apenas el primer pilar de cualquier modelo tipo score que se desee desarrollar, independientemente si el profesional a cargo opta por sistemas lineales, logísticos, árboles de clasificación, análisis discriminante o incluso sistemas basados en inteligencia artificial (redes neuronales), pues la experiencia dicta que la información es una poderosa herramienta de decisión y mientras más profundo y pormenorizado sea su análisis más amplio el abanico de oportunidades de aplicación.

V. BIBLIOGRAFIA

- Abraira Víctor, Zamora Javier y Muriel Alfonso, Unidad de Bioestadística, Hospital Universitario Ramón y Cajal, Madrid, 2005.
- Aguilera del Pino, Ana María, *Tablas de Contingencia Bidimensional*, Ed. La Muralla, S.A., Madrid, 2001.
- Aragon, Aker, *Discriminant Analysys of Default Risk*, MPRA, 2004.
- Araya Roberto*, Induction of decision trees when examples are described with noisy measurements and with fuzzy class membership*, INRIA, Paris, 1994.
- Araya Roberto, *Métodos estadísticos básicos de discriminación con árboles de decisión*, AutoMind, Santiago Chile, 2005.
- Botero Londoño Liliana, *Scoring en la Industria de Microfinanzas*, Accion Internacional, 2007.
- Calvache Diego y Carranza Freddy, *Diseño y Elaboración Estadística de un Sistema de Evaluación para la Otorgación de Crédito de Consumo en una Institución Financiera*, Quito, EPN, 2000.
- Calvo-Flores Segura Antonio y Arques Pérez Antonio, *Modelos Estadísticos Teóricos, Facultad de Economía y Empresa*, Universidad de Murcia.
- Carranza Vergara Freddy H., *Clasificación de clientes mediante la aplicación de modelos Econométricos a Índices Financieros,* Quito, UASB, 2006.
- COLAC, *Perspectivas de Desarrollo del Cooperativismo*, 2002.
- Consumer Federation of America, *Credit Score Accuracy and Implications for Consumers*, National Credit Reporting Association, 2002.
- Génesis Empresarial, *Experiencia con los Sistemas de Información de Créditos*, Guatemala, 2007.
- Gómez García Juan, Palarea Albaladejo Javier, y Martín Fernádez Josep

Antoni, *Métodos de inferencia estadística con datos faltantes*. Estudio de simulación sobre los efectos en las estimaciones, Estadística Española, Vol. 48, 2006.

- Gujarati, Damodar, *Essentials of Econometrics*, New York, McGraw-Hill, 2004.
- Hanley J.A., McNeil B.J. *The meaning and use of the area under a receiver operating characteristic (ROC) curve*. Radiology. Vol.143, 1982.
- Hernández, Ernesto Gonzalo, *Relación entre Variables Cualitativas*, 2006.
- Kosorok, Michael R., *Introduction to Empirical Processes and Smiparametric Inference*, Springer Science+Busines Media, Inc., 2006.
- Medina Moral, Eva, *Modelos de Elección Discreta*, 2003.
- Molinero, Luis Miguel, *¿Y si los datos no siguen una distribución normal? Bondad de ajuste a una normal, transformaciones, pruebas no paramétricas*, Asociación de la Sociedad Española de Hipertensión, 2003.
- Molinero, Luis Miguel., *Elección de los puntos de corte para convertir una variable cuantitativa en cualitativa*, Asociación de la Sociedad Española de Hipertensión, 2003.
- Salgado, Pablo Andrés, *El uso de la estadística en electrofisiología*, 2007.
- Schreiner, Mark, *El riesgo de deserción de prestatarios de un prestamista de microcrédito en Bolivia*, St. Louis, Center for Social Development Washington University, 2000.
- Schreiner, Mark, *Ventajas y Desventajas del Scoring Estadístico para las Microfinanzas*, St. Louis, Center for Social Development Washington University, 2002.
- Schreiner Mark, y Dellien Hans, *El scoring estadístico, los bancos y las microfinanzas: cómo lograr un balance entre el uso de tecnología y atención personalizada*, Banco Interamericano de Desarrollo, 2005.

- Scoring LiSim, *Seminario Estrategias para la Administración de Riesgo de Crédito, por medio de Scoring*, Quito, 2006.
- Scott Frame, Srinivasan Aruna, y Woosley Lynn, *The effect of Credit Scoring on small-business Lending*, Journal of Money, 2001.
- Simbaqueba, Lilian, *¿Qué es un Scoring? Una visión práctica de la gestión del riesgo de crédito*, Instituto de Riesgo Financiero, 2004.
- Stella Verón, Carmen, *La utilización de variables cualitativas y el análisis de datos categóricos en la investigación*, U. Nacional de Rosario, 2006.