UNIVERSIDAD NACIONAL AUTÓNOMA DE MÉXICO

FACULTAD DE CIENCIAS

MODELO DE REGRESIÓN LOGÍSTICA EN CREDIT SCORING PARA EL SECTOR RURAL EN MÉXICO DE 2002 A 2011.

T E S I S

QUE PARA OBTENER EL TÍTULO DE:

ACTUARIA

PRESENTA:

Maria Leonor Zamora Maass

DIRECTOR DE TESIS:

Act. Francisco Sánchez Villareal

2013

Alumna
Zamora
Maass
Maria Leonor
Universidad Nacional Autónoma de México
Facultad de Ciencias
Actuaría
0305639747

Director de Tesis
Act.
Francisco
Sánchez
Villarreal

Sinodal 1
Dra.
María del Pilar
Alonso
Reyes

Sinodal 2
Dra.
Nora
Gavira
Durón

Sinodal 3
Dra.
María Araceli
Bernabe
Rocha

Sinodal 3
Act.
María del Rosario
Espinosa
Tufiño

Trabajo escrito
Modelo de Regresión Logística en Credit Scoring
para el sector rural en México de 2002 a 2011.
116 p
2013

AGRADECIMIENTOS

A la Facultad de Ciencias y la UNAM por el gran aprendizaje, sobre todo al Act. Francisco por su paciencia y disposición, a todos mis profesores (as) y amigos (as) que han hecho de estos años el reto más divertido.

A mi familia por su cariño y apoyo, en especial a mi abuelo (D.E.P) y mi abuela, a mis tíos y por supuesto, a Pepe. Gracias a Leonor por su fortaleza y sus interminables consejos, porque más que madre y padre, es mi mejor amiga.

A quienes han sido mis cómplices y compañeros (as) de vida, aquellos (as) que siempre van conmigo, incluso cuando no están presentes.

ÍNDICE

INTRODUCCIÓN

Objetivo – *Implementar un modelo de regresión logística en Credit Scoring para clasificar y describir el comportamiento y los factores de riesgo crediticio, así como calcular la probabilidad de incumplimiento del sector rural en México del año 2002 al 2011.*

Clasificar es una acción de la vida diaria, es darle a algo sus propias características de identificación y un hecho que no siempre resulta ser sencillo, ya que requiere de un amplio conocimiento del tema. La Regresión Logística es una herramienta estadística de interpretación útil para ello, clasifica y explica los factores de ocurrencia de un evento.

En este caso, a través de esta herramienta, se busca impulsar el Credit Scoring correspondiente al riesgo de crédito como apoyo indispensable para la economía y el sistema financiero. Lo que se pretende es divulgar la forma de creación de un modelo capaz de aportar de manera eficiente conocimientos para clasificar a los clientes. Todo esto mediante la aplicación particular al sector rural en México, por ser una población con grandes áreas de oportunidad para la mejora socio-económica del país.

Adicionalmente, como parte de la importancia de conocer metodologías estadísticas e incentivar el uso de esta ciencia, se busca describir otros modelos de análisis de datos adecuados para apoyar en la implementación de mejoras e investigaciones innovadoras al respecto.

La estructura del estudio será implementada en tres capítulos:

- Capítulo 1: Otorgamiento y riesgos de crédito del sector rural en México

  De manera inicial se dará una introducción de la importancia y el concepto del crédito, partiendo del análisis del Sistema Financiero Mexicano y la Banca de Desarrollo. Se hará énfasis en la población, el financiamiento al sector rural y a sus respectivas áreas de oportunidad. Serán descritos los principales conceptos de riesgo, particularmente del riesgo de crédito, con el fin de conocer sus métodos de estimación y la importancia de la probabilidad de incumplimiento.

- Capítulo 2: Credit Scoring y algunos métodos alternativos

  Esta parte tiene como herramienta principal el Credit Scoring, como una solución al cálculo de la probabilidad de incumplimiento, al análisis de las características de los clientes y la identificación de factores de riesgo crediticio. Serán conceptualizados los principios y procedimientos de su aplicación, aspectos de regulación y la importancia que la estadística tiene para ello. Se incluye además la descripción de algunos métodos alternativos para su construcción.

- Capítulo 3: Modelo de Regresión Logística en el Credit Scoring

  En este capítulo se describirá la función logística y la transformación logit, partiendo de sus orígenes, supuestos y aplicaciones. Se buscará dar a conocer la forma de estimar los parámetros desconocidos y ajustar la regresión con diversos estadísticos para conseguir su optimización. Este desarrollo será complementado con una muestra y la construcción del Credit Scoring para el sector rural, obteniendo su estimación, ajuste e interpretación.

Finalmente se encontrarán las conclusiones y la bibliografía consultada, lo cual permitirá dar a conocer las fuentes donde se pueden profundizar los temas de interés particular.

OTORGAMIENTO Y RIESGOS DE CRÉDITO DEL SECTOR RURAL EN MÉXICO

1.1. El crédito y la Banca de Desarrollo en México

1.1.1. Actividades económicas y el Sistema Financiero Mexicano

En la actualidad referirse a temas importantes en México corresponde en gran parte a mencionar actividades económicas, es decir, a todas las acciones que tienen por objeto la producción, distribución y consumo de bienes y servicios, donde el dinero resulta ser el eje principal y cuyo uso representa una parte fundamental de la vida diaria.

Se tratan de actividades dependientes del Sistema Financiero Mexicano (SFM), que consiste en la agrupación de instituciones relacionadas entre sí y encargadas de captar, administrar y canalizar la inversión y el ahorro de los oferentes a los demandantes, a través de los intermediarios financieros (asignar recursos de aquellas personas que cuentan con ellos hacia aquellas que los necesitan con el fin de generar ganancias). Está compuesto por mercados organizados con actividades y activos propios, manteniendo la siguiente estructura:

- Oferentes o ahorradores: Personas físicas, inversionistas nacionales y extranjeros; Personas morales, empresas privadas o entidades como Instituciones de Seguros y Fianzas, Sociedades de Inversión, Fondos Laborales e Instituciones de Seguridad Social.
- Intermediarios Financieros: Banca Comercial, Banca de Desarrollo, Grupos Financieros, Casas de Bolsa y entidades auxiliares de crédito.
- Demandantes: Personas físicas, personas morales y el Gobierno Federal.

Su regulación se da por diversos organismos como la Secretaría de Hacienda y Crédito Público (SHCP), una dependencia gubernamental con la misión de proponer, dirigir y controlar la política económica del Gobierno Federal en materia financiera, fiscal, de gasto, de ingreso y deuda pública, las estadísticas, geografía e información, con el propósito de consolidar un país con crecimiento económico de calidad, equitativo, incluyente y sostenido, que fortalezca el bienestar de sus habitantes.

En particular, en el Art. 31 de la Ley Orgánica de la Administración Pública Federal, se hace referencia a una de sus actividades más relevantes: “Fracc. VII.- Planear, coordinar, evaluar y vigilar el sistema bancario del país que comprende al Banco Central, a la Banca Nacional de Desarrollo y las demás instituciones encargadas de prestar el servicio de banca y crédito.”

De la SHCP se desprenden órganos desconcentrados, encargados de supervisar y regular diversos sectores como la Comisión Nacional Bancaria y de Valores (CNBV) encargada de las entidades financieras, la Comisión Nacional del Sistema de Ahorro para el Retiro (CONSAR) encargada de los participantes en los sistemas de pensiones y la Comisión Nacional de Seguros y Fianzas (CNSF) encargada de las afianzadoras, de Instituciones y Sociedades Mutualistas de Seguros. Así como por el organismo central, Banco de México (Banxico), que tiene como objetivos principales regular la masa monetaria, ser un banco de reserva y prestador de servicios de tesorería del Gobierno Federal, sus disposiciones influyen en todo el sistema financiero.

Por otro lado, todas las acciones del SFM están regidas por leyes como:

1. Constitución Política de los Estados Unidos Mexicanos
2. Ley Orgánica de la Administración Pública Federal
3. Ley Orgánica de la Secretaría de Hacienda y Crédito Público
4. Ley Orgánica del Banco de México
5. Ley Orgánica de la Comisión Nacional Bancaria y de Valores
6. Ley Orgánica de la Comisión de Seguros y Fianzas
7. Ley General de Títulos y Operaciones de Crédito

Sin embargo, referirse al SFM no sólo implica el conocimiento del funcionamiento, ya que como cualquier otro aspecto de la vida diaria es un tema que cuenta con grandes problemáticas y una gran cantidad de campos de investigación. En esta ocasión se hará énfasis en uno de los temas que más han destacado en los últimos años: El otorgamiento de créditos en la Banca de Desarrollo. Mismo que ha obligado a diversas entidades a instrumentar medidas para cambiar y/o mejorar las políticas, pues ante una situación de crisis la principal estrategia se localiza en las oportunidades que surgen para sobrevivir, es un momento ideal para reorganizar los procesos.

1.1.2. Origen y principales conceptos de crédito

La evolución del proceso de crédito es un tema fundamental para referirse al otorgamiento. La palabra crédito proviene del latín *Creditum y Credere* que significa creer, está relacionado con la confianza y se define como la entrega de un valor actual ya sea en dinero, mercancía o servicio a cambio de un valor equivalente esperado en un futuro, pudiendo existir un interés pactado. Desde el punto de vista jurídico, es una promesa de pago con un vínculo entre el deudor y el acreedor.

En realidad existen diferentes opiniones sobre su origen pero en la mayoría coinciden en que se trata de operaciones muy antiguas que se efectuaban como préstamos pagados en especie y no en moneda, reconocidos como una actividad económica al haber sido establecido el concepto de *trueque*, es decir, el intercambio de un bien por otro de la misma naturaleza o valor.

Algunos análisis dedicados a la gestión de crédito mencionan que este origen se pudo ver con los préstamos y escritos que dejaron algunas civilizaciones, donde ordenaban pagos posteriores similares a letras de cambio para evitar el transporte de material de recursos[1]. Los Fenicios en la zona Mediterránea, por ejemplo, cobraban hasta después de entregar su mercancía, al ser ellos quienes ofrecían productos a los aborígenes y se alejaban para volver con señales de fuego a recoger los bienes o valores en pago. Este proceso fue expandido al comercio marítimo, los prestamistas y comerciantes que se convertían en auténticos socios, situación que llevó a crear incluso una combinación de préstamos y una especie de seguros.

Sin embargo, la esencia del crédito fue en realidad conocida hasta la Edad Media, época en la que se registraron operaciones bancarias donde se determinaba la calidad de los clientes en función del conocimiento que se tenía acerca de ellos, decidiendo individualmente cuánto le prestarían y cuándo es que debían pagar. Poco a poco se comenzaron a pedir garantías que redituaran el valor del préstamo en caso de no querer prescindir de los bienes obtenidos mediante estos recursos.

En 1350 se inició el fenómeno de las casas de empeño[2], las cuales no solían cargar intereses durante el tiempo que se esperaba la recuperación del préstamo para devolver la prenda dejada a cambio pero con el tiempo lo encontraron como una buena oportunidad de aumentar sus ingresos.

[1] Santandreu, Eliseu. *Manual del Credit Manager.* Barcelona. Gestión 2000, S.A.: 2002.
[2] Thomas, Lyn C.; Edelman, David B.; Crook, Jonathan N. *Credit Scoring and Its Applications*. USA. SIAM: 2002.

Más tarde aparecieron formalmente las instituciones bancarias y en 1920 inició la verdadera revolución del consumo de crédito, con el deseo acelerado de comprar coches de motor y con la demanda de ropa y artículos para amas de casa, que al encontrarse localizadas lejos de los centros más poblados encargaban productos que pagaban después.

Con dichos cambios se pudo ver un crecimiento acelerado de la operación. Se desarrolló una cultura que se fue generalizando hasta convertirse en un estilo de vida, con el uso de recursos de aquellos que querían mantener su dinero seguro y obtener algún beneficio, por aquellos que lo necesitaban para subsistir o mejorar sus condiciones económicas, misma que ha requerido una atención especial por los riesgos que conlleva pero que finalmente son necesarios de tomar con el aumento de competencia en el mercado y en los sistemas financieros. El problema radica en que con un alto consumo crediticio existe una gran diversidad en las formas de pago y un posible aumento también en los niveles de morosidad que no sólo generan desconfianza, si no que obligan a realizar este tipo de estudios para dar el conocimiento necesario y disminuir el otorgamiento de préstamos a clientes incumplidos, consiguiendo con ello la atención adecuada a los aspectos en los que se involucra el crédito como son: El incremento de la producción, la creación de fuentes de trabajo, el desarrollo de mercados y el mantenimiento de las actividades económicas en niveles favorables.

Características del crédito

La operación crediticia está compuesta tanto por el *Acreedor*, aquella persona que cuenta con los recursos y otorga el crédito, como por el *Deudor,* aquella persona que recibe el crédito. Cuando existen *Intermediarios* de por medio, éstos se convierten en los acreedores y son ellos quienes cuentan con los recursos y fungen como inversionistas.

Cuentan con diversas *tasas de interés* que el deudor pagará al acreedor a cambio del préstamo o financiamiento de recursos, entre las cuales se encuentra la tasa de interés ordinaria, que se cobra por el simple hecho de la existencia del financiamiento y que puede ser fija o variable. La tasa de interés fija es aquella que se conoce previamente y se mantiene constante a lo largo de la vida del crédito, la tasa variable depende de otras tasas o valores del mercado que pueden subir o bajar en cualquier momento. También existen la tasa de interés por apertura, que es una comisión para los

gastos iniciales del crédito, así como la tasa de interés moratoria, que se cobra al caer en mora o impago cuando los deudores no devuelven los recursos en el tiempo y forma.

El *plazo o duración* del crédito es otra característica y representa el tiempo durante el cual se mantiene el préstamo, mismo en el que el deudor tiene la obligación de devolver los recursos, ya sea en plazos de pago o en un solo pago al final del crédito. Puede ser clasificado como corto o largo, siendo la diferencia marcada comúnmente por 12 meses; esto es que los créditos menores o iguales a un año suelen ser vistos como créditos a corto plazo y los que tienen duración mayor a un año suelen ser referidos como de largo plazo.

Adicionalmente se utilizan las *garantías*, que consisten en respaldos por parte del deudor para asegurar que el acreedor no perderá los recursos otorgados. Son garantías reales cuando implican bienes materiales; garantías personales por el historial crediticio, la solvencia o los avales (terceras personas encargadas de devolver recursos en caso de que el deudor no lo haga), y garantías naturales cuando el respaldo es el destino del crédito (aquél objeto o fin que representa la necesidad del deudor de recibir recursos).

Es no sólo una operación de deuda, ya que el saldo de los clientes representa un activo que supone una inversión, cuyo monto pretende recuperarse y recibir una rentabilidad. Sus etapas principales están divididas en cuatro: Promoción de los créditos, evaluación de los sujetos y condiciones bajo las cuales se hará el préstamo (otorgamiento de crédito), apertura y entrega de los recursos, cobranza y recuperación.

Tipos de crédito

Debido a la diversidad de operaciones de crédito y las múltiples actividades en las que se involucra tiene diferentes clasificaciones y sub-clasificaciones según el área y tema referidos.

La principal clasificación conforme a Banxico toma en cuenta el objeto del crédito: Los bancos ofrecen créditos para *consumo* y *actividades productivas*. Los créditos de consumo son ofrecidos a las personas físicas por medio de tarjetas de crédito, créditos para adquisición de bienes, crédito hipotecario y automotriz. Por otro lado, los créditos para actividades productivas están a disposición de personas físicas y morales, sub-clasificados como créditos PYME, Microcréditos,

de Habilitación y Avío, Prendarios y Refaccionarios, entre otros relacionados con socios comerciantes y empresarios.

Aunque existen otras clasificaciones de acuerdo al resto de las características y condiciones del crédito, como la actividad[3]. Créditos *por el uso*, que son los créditos de inversión, con el objeto de colocar capitales en manos de terceros; créditos bancarios, basados en la intermediación lograda a partir de los depósitos de clientes y que hace funcionar la producción, distribución y consumo; créditos entre comerciantes, que son aquellos en los cuales la operación crediticia se constituye por mercancías o servicios; y créditos al consumidor, en los que las empresas conceden plazos y tasas de interés para recibir el pago de los productos entregados. Además de existir los créditos *por el sujeto*, que representan los créditos públicos, para uso del Estado que comprende los créditos a Instituciones Gubernamentales, Estados y Municipios; y los créditos privados, que ejercen los particulares y están regidos por normas del mercado.

Factores en el otorgamiento de créditos

El otorgamiento es la principal etapa en la operación. Es el momento en el que se decide aceptar o rechazar el acceso a los recursos, basando la conclusión en la probabilidad de pago. Tiene como objetivo identificar sus riesgos inherentes, con el fin de poder obtener los ingresos y ganancias deseados, considerando una investigación previa del cliente y el crédito que se desea otorgar.

Al inicio de la operación se requiere de la entrega de una solicitud de crédito, adecuada al tipo de población a quien va dirigida, cuestionando datos para conocer al solicitante y proporcionar elementos de juicio para autorizar el préstamo, acompañados de documentación e informes que lo avalen. Las fuentes de información de los documentos se dividen en internas y externas con base en cómo se originan (fuera o dentro de la propia empresa y/o lugar de trabajo) y deben ser avaladas por entidades autorizadas, pudiendo ser complementadas con otras herramientas como recurrir a compañeros de trabajo, clientes y/o proveedores relacionados con el solicitante. Acción que no siempre es posible realizar debido a los costos de transporte, tiempo invertido y personal.

[3] Villa Señor Fuente, Emilio. *Elementos de Administración de Crédito y Cobranza.* México. Ed. Trillas: 2ª Ed. 1985

Los factores más relevantes durante este proceso parten de lo que se conoce tradicionalmente como las 4C's del crédito[4], actualmente llamado 5C's o método experto con la inclusión de un nuevo factor que mediante conocimientos empíricos estructuran la información para facilitar la decisión del otorgamiento:

1) Carácter: Es la responsabilidad, honradez e integridad del cliente, cualidades que lo harán pagar al momento del vencimiento. La mejor evaluación de este riesgo se consigue con el historial de pago de otros créditos, ya que aquellas personas con antecedentes de buen pago, difícilmente se desviarán. Los puestos laborales, relaciones comerciales, estado civil, régimen matrimonial y la naturaleza de su educación son también parte del carácter.

2) Capacidad: Es la aptitud, la experiencia y fortaleza, en otras palabras, la posibilidad para pagar la deuda. Es un riesgo difícil de evaluar por la cantidad de elementos que intervienen, no sólo los ingresos de los que dispone, sino por las ventas y deudas de las cuales depende ese ingreso. Las variables como ubicación comercial, organización, antigüedad, apoyos y competencia del mercado están sumamente relacionadas, el patrón de gastos del cliente, la cantidad de hijos que mantiene y el nivel de vida acorde con su profesión, edad o antigüedad del empleo.

3) Capital: Es el soporte financiero y patrimonio del cliente. Se determina a partir de estados financieros para personas morales y de bienes como automóvil, objetos personales, dinero en efectivo, títulos financieros y bienes inmuebles para personas físicas.

4) Condiciones y ciclos: Representa las circunstancias externas y elementos adicionales del crédito, son fluctuaciones que pueden ocurrir a corto o largo plazo. Para ello se debe realizar un estudio sobre las zonas geográficas, tipo de población, destino del crédito, tipo de crédito y plazos. Se pueden referir a temas sociales, económicos e incluso físicos.

5) Colateral: Es un elemento adicional que toma en cuenta el derecho que se tiene por la garantía dejada en referencia. Dicha garantía puede ser real, natural o personal, entre mayor sea su valor la posibilidad de tener pérdidas por incumplimiento será menor.

[4] Villa Señor Fuente, Emilio. *Elementos de Administración de Crédito y Cobranza.* México. Ed. Trillas: 2ª Ed.

1.1.3. El crédito y la Banca de Desarrollo en México

De acuerdo a los objetivos es necesario abordar el tema de la Banca de Desarrollo que, con base en las definiciones de la SHCP[5], es el conjunto de entidades de la Administración Pública Federal, con personalidad jurídica y patrimonio propios. Se trata de un conjunto de Instituciones constituidas con el carácter de Sociedades Nacionales de Crédito, un organismo público y diversos Fideicomisos. Su objetivo es facilitar el acceso al financiamiento a personas físicas y morales de diversos sectores productivos del país y conforme a los lineamientos y estrategias del Plan Nacional de Desarrollo, para apoyar la competitividad y la capitalización, proporcionarles asistencia técnica y capacitación, fomentando una mayor coordinación con otras dependencias.

Desde marzo de 1988 la SHCP fue la encargada de la Dirección General de la Banca de Desarrollo para formular políticas financieras y de planeación, encargándose de las autorizaciones y límites de financiamiento. En el marco regulatorio, la CNBV es quien emite las reglas para dichas instituciones, Banxico emite las disposiciones generales y es la Ley General de Instituciones de Crédito donde se encuentran las leyes orgánicas.

Las principales instituciones, entidades de fomento y sectores que conforman la Banca de Desarrollo son:

- Sector Empresarial: Busca impulsar el desarrollo de las empresas mexicanas, principalmente micro, pequeñas y medianas, fomentar empleos y el comercio exterior. Está representado por Nacional Financiera y el Banco Nacional de Comercio Exterior.
- Sector Rural: Tiene como mandato proporcionar recursos financieros, capacitación, asistencia técnica e información a las empresas y productores del campo. Está constituido por Fideicomisos Instituidos con Relación a la Agricultura (FIRA), por el Fondo de Capitalización e Inversión del Sector Rural y por un organismo público, Financiera Rural.
- Sector Infraestructura: Otorga recursos financieros para el desarrollo de proyectos como carreteras, plantas de tratamiento de agua, puertos y aeropuertos, entre otros. Está constituida por el Banco Nacional de Obras y Servicios Públicos.

[5] SHCP. *Apartado Hacienda para Todos: Banca de Desarrollo.* [Recurso en línea 2012]

- ❖ Sector Vivienda: Tiene como objetivo apoyar a familias de bajos ingresos con el otorgamiento de recursos financieros para construcción y adquisición de viviendas. Está representado por la Sociedad Hipotecaria Federal.

- ❖ Sector Servicios Financieros: Representado por el Banco de Ahorro Nacional y Servicios Financieros y el Banco Nacional del Ejército, Fuerza Aérea y Armada, que proporcionan servicios financieros a los miembros de dichas entidades.

La Banca de Desarrollo durante los últimos años

Una vez conceptualizada la Banca de Desarrolla, vale la pena tener un panorama completo de su situación actual, a través de algunos resultados presentados por la SHCP en el último sexenio[6]:

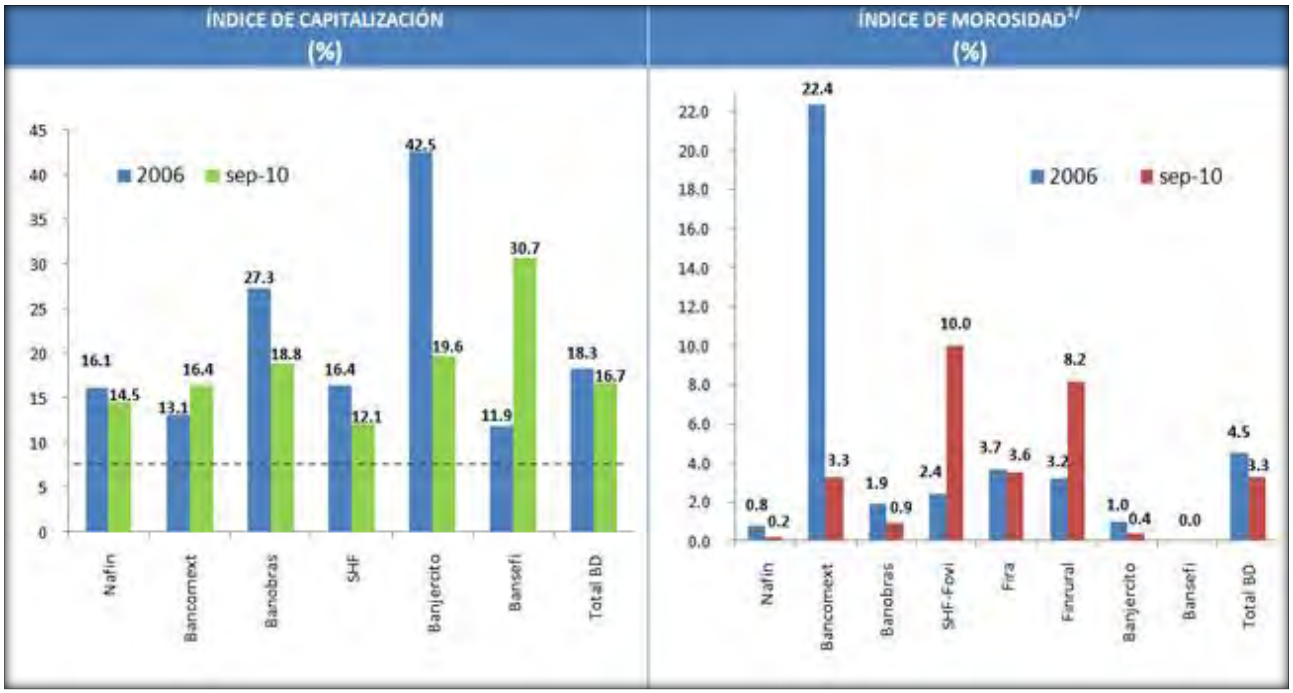
Cartera de la Banca: Es la participación de la Banca de Desarrollo como porcentaje de la Banca Comercial, que son aquellos casos cuyo mercado objetivo también atienden a esos sectores (sin que esto implique que tengan las mismas condiciones). Muestra una caída en el año 2008 (35.2%) con respecto al inicio del sexenio en 2006, a causa de prepagos de los intermediarios financieros por la recuperación de los financiamientos privados después de la crisis, y un incremento desde entonces del 15.8% hasta el tercer trimestre de 2010.

Fuente: Secretaría de Hacienda y Crédito Público

[6] SHCP. *Apartado Hacienda para Todos: Banca de Desarrollo.* [Recurso en línea 2012]

Estabilidad financiera: Se analiza el índice de capitalización (ICAP) y morosidad (IMOR). El primero es el crecimiento de la cartera de cada una de las entidades, con el fin de preservar su capital y productividad, es un indicador que representa la fortaleza financiera de una institución para soportar pérdidas no esperadas. Por otro lado, el Índice de Morosidad representa la cartera de crédito vencida (créditos en impago) como proporción de la cartera total. Es uno de los mejores indicadores para observar el buen o mal manejo de créditos.

El total de la Banca de Desarrollo para finales del 2010 muestra una mejora con respecto al año 2006 ya que a pesar de las caídas principales de Banjército y Banobras, el ICAP se mantiene con el doble del mínimo regulatorio (8%) y el IMOR disminuyó en 1.2%, aún cuando también destacan entidades con problemáticas como SHF y Finrural a causa de los conflictos con el uso de financiamientos para vivienda y actividades productivas durante y después de la crisis.



1/ IMOR de la Cartera de Crédito Directo al Sector Privado de la Banca de Desarrollo
Fuente: Secretaría de Hacienda y Crédito Público

Sectores: Conforme al objetivo de la Banca de apoyar sectores específicos de la población, se alcanzaron buenos resultados los últimos años, comparando el tercer trimestre de 2010 con respecto al de 2007. Dentro de dicha mejora es de interés el 43% de mayor atención a

productores rurales de bajos ingresos (por FIRA y Financiera Rural), así como 159 municipios de alta y muy alta marginación más en cartera (por Banobras).

Atención a Sectores Prioritarios		
Sector de atención	sep-07	sep-10
MIPYMES atendidas	618,887	928,681
Productores rurales con ingresos menores a 3,000 smd	790,755	1,128,019
Municipios con alto y muy alto grado de marginación	218	377
Proporción de créditos individuales de vivienda a población con ingresos inferiores a 6 vsm	47%	90%

Fuente: Secretaría de Hacienda y Crédito Público

De acuerdo a esto se puede hablar de un mantenimiento de estabilidad en general en el último sexenio pero sin representar progresos importantes en comparación con otras fuentes de financiamiento ya que la Banca de Desarrollo está concebida como un fomento nacional para el otorgamiento de créditos a empresas públicas y personas físicas pero que actualmente realiza también actividades de intermediación y apoyo por medio de garantías o fideicomisos, provocando que se dedique a privilegiar el crédito privado de consumo antes que a fomentar ocupaciones productivas y a incentivar la Banca Comercial (en lugar de complementarla), la cual se dedica a financiar plazos menores y con fines lucrativos.

Este hecho complementa la importancia del presente estudio al incentivar la atención que se le debe dar a los créditos directos para personas físicas y en particular al sector rural, dando a conocer los temas involucrados para conseguir la mejora de los procesos crediticios.

1.2. Población y financiamiento rural

1.2.1. Elementos fundamentales de la población rural

En apego al Instituto Nacional de Estadística y Geografía (INEGI), en el último reporte de Población Rural y Rural ampliada (2000 – 2006), existen distintos criterios adoptados en las oficinas nacionales de estadística del mundo para definir la cuantificación rural y urbana:

- Demográfico. Define lo rural a partir del tamaño de la localidad. Esto en algunos países se da por la densidad de población o por el agrupamiento de manzanas.

- Político-administrativo. Que la localidad sea o no cabecera municipal o que por decreto cuente con determinada categoría política (villa o ciudad), otorga el nivel de urbana o rural, independientemente de su tamaño poblacional.

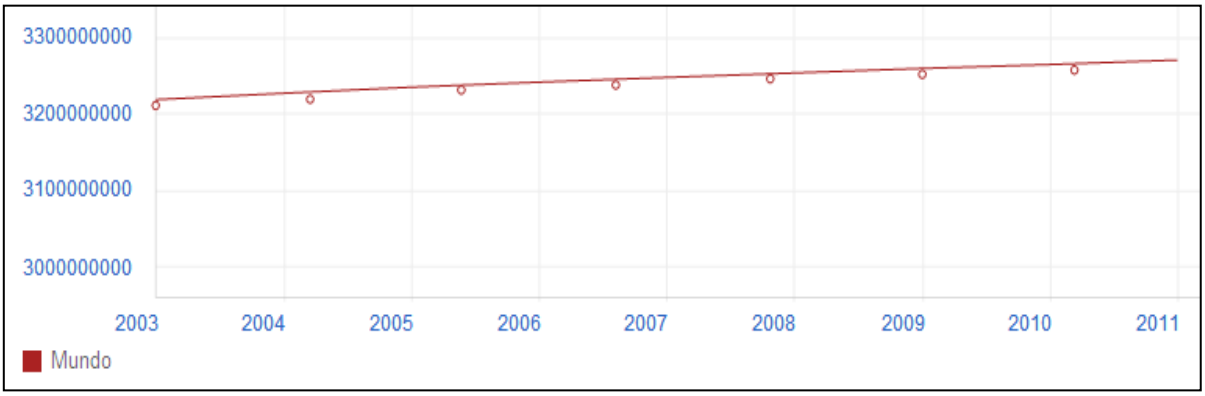
- Económico. Define a la población por el perfil económico (industrial, comercial, etc). Por ejemplo el hecho de que en una pequeña localidad se asiente una ciudad industrial o una empresa grande, la clasifica como urbana aún cuando sea un pequeño poblado.

- De infraestructura y equipamiento urbano. La disposición de carreteras, el nivel de los servicios educativos, de salud y gubernamentales, la cobertura de agua potable, la electricidad y telefonía son determinantes.

- Geográfico. En un sentido físico la dispersión geográfica y la distancia de una localidad a carreteras y centros urbanos categorizan lo rural. En el ámbito funcional los vínculos de localidades con los lugares centrales establecen la ruralidad.

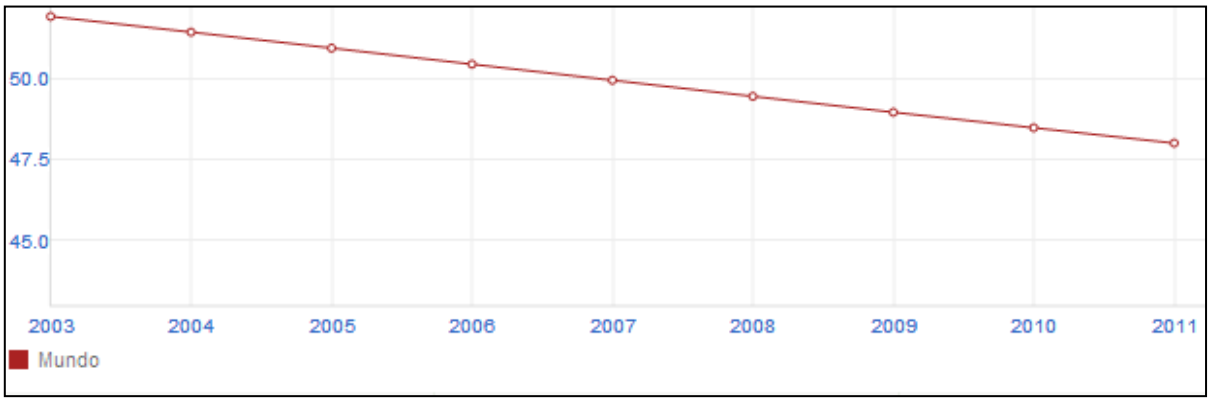
A grandes rasgos son poblaciones con diferencias destacables en comparación con las zonas urbanas y es que aunque la cantidad de sus habitantes sigue aumentando en la mayoría de los países, las poblaciones urbanas asentadas en territorios con actividades económicas industriales y de servicios han sido predominantes con los desarrollos tecnológicos, desde representar un 10% de la población total hasta quedar en un 50%. Dichos resultados pueden ser confirmados con los indicadores dados por el Banco Mundial:

Población rural: Son aquellos habitantes de zonas rurales, según las definiciones anteriores. Se calcula como la diferencia entre población total y población urbana. Se comprueba el aumento de la población rural del 2003 al 2011 de 3,283,792,714 a 3,335,890,194 personas.



Fuente: Banco Mundial

Porcentaje de la población rural (% de la población total): Se calcula como proporción de la diferencia entre población total y población urbana con respecto a la población total. Se comprueba la disminución de la población rural del 2003 al 2011 de 51.9% a 48%.



Fuente: Banco Mundial

Pero entender el sector rural no sólo es tener la capacidad de definirlo, implica tambien entender su estilo de vida. Son regiones donde las actividades de los habitantes están enfocadas a la cría, cultivo, comercialización y servicios derivados del campo. Sus medios de subsistencia se diversifican en cada región. Los hogares más pobres dependen en su mayoría de la agricultura. El nivel de desarrollo en esta población es bajo, tienen falta de servicios públicos e incluso de sanidad, condiciones de vida con carencias suficientes para buscar continuamente convertirse en

poblaciones urbanas, propiciadas por factores como oportunidades de empleo y educación. Por ello tratan de evitar conflictos como el de la degradación de tierra y agotamiento de los recursos naturales. Buscan centros urbanizados con desarrollos tecnológicos e industrialización para procesar sus productos.

Con base en esto se puede comprender que es un sector al cual se le debe prestar una atención especial, tal y como lo menciona el Fondo Internacional de Desarrollo Agrícola (FIDA) en el "Informe Sobre la Pobreza Rural del año 2011", donde se define la pobreza de la población rural como una consecuencia de falta de activos, escasez de oportunidades económicas, educación, capacidades deficientes y una serie de desventajas derivadas de desigualdades sociales y políticas. En este informe se describe que para el año 2050 se espera que existan 9,000 millones de personas en este sector y que la producción agrícola en los países en desarrollo se vea duplicada, con lo cual los agricultores desempeñarán un papel aún más importante, razón suficiente para este estudio que busca promover al sector rural con el fin de fomentar que se aprovechen al máximo las oportunidades de crecimiento y atender las preocupaciones que se refieren principalmente a los pequeños productores.

1.2.2. Población Rural en México

En México son zonas rurales las localidades de menos de 2,500 habitantes y zonas rurales ampliadas las que cuentan con menos de 5,000 habitantes. Es un sector cuya base demográfica y estadística analiza aspectos sociales y económicos, resultando de suma importancia indagar en las características de la población con base en los "Principales Resultados del Censo de Población y Vivienda 2010" aportados por el INEGI para localidades menores a 5,000 habitantes.

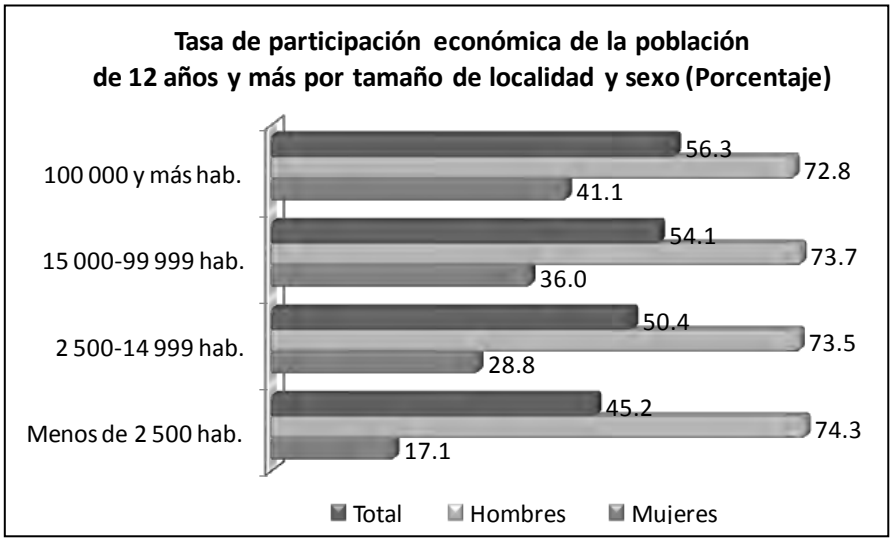
Es una población que tuvo un crecimiento constante en el siglo XX, de 1930 a 2010 ascendió de 12.3 a 32.4 millones de personas, sin embargo, la migración hacia zonas urbanas en búsqueda de mejores condiciones de vida ha sido un factor determinante para el tamaño de la población rural que en los últimos años ha representado aproximadamente el 23.2% de la población total y que mantiene la siguiente distribución:

DISTRIBUCIÓN DE LA POBLACIÓN RURAL POR TAMAÑO DE LOCALIDAD

Tamaño de localidad	Localidades	Porcentaje	Población	Porcentaje
Total	**190 432**	**100.0**	**32 410 077**	**100.0**
1-249 habitantes	**159 820**	**83.9**	**5 743 745**	**17.7**
250-999 habitantes	**22 852**	**12.0**	**11 328 495**	**35.0**
1 000-2 499 habitantes	**5 921**	**3.1**	**8 976 888**	**27.7**
2 500-4 999 habitantes	**1 839**	**1.0**	**6 360 949**	**19.6**

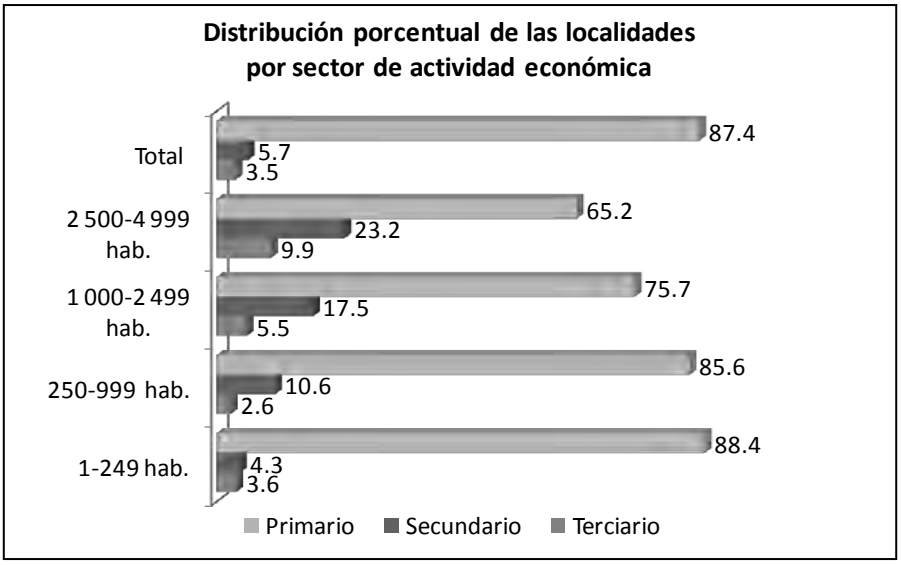
Fuente: INEGI Censo de Población y Vivienda 2010

La participación económica de los habitantes mayores a 12 años, a la cual se le conoce como Población Económicamente Activa (PEA), es del 45.2% en estas zonas. A diferencia de la PEA en localidades más grandes, por arriba del 50%. En este sector el género masculino tiene mayor participación y está representado por el 74.3%.



Fuente: INEGI Censo de Población y Vivienda 2010

Las ramas o sectores de producción son también parte vital del análisis. El sector principal es el primario y las actividades destacables son las agropecuarias, seguidas por las de artesanía y de obreros, comerciantes y dependientes, ayudantes y trabajadores domésticos. La agricultura resulta ser la actividad preponderante también durante los últimos años pero ha resultado ser un medio de subsistencia más que una actividad de producción, lo cual ha hecho vulnerable el sistema rural y ha dificultado la existencia de pequeños y medianos propietarios.



Fuente: INEGI Censo de Población y Vivienda 2010

Su grado de alfabetismo es menor al del sector urbano, la asistencia escolar a Secundarias o Telesecundarias es del 56.1% y a Preparatorias del 13.8%. Tiene una gran cantidad de viviendas construidas sobre tierra (una tercera parte), falta de servicios públicos y drenaje, poca afiliación a servicios de salud y disponibilidad de bienes en niveles bajos. En medios de comunicación el mayor porcentaje se ubica en la pertenencia de televisión y radio con 67% y 80% respectivamente, mientras que el menor está presente en computadoras (6.8% a diferencia de las zonas urbanas con 42.7%) e internet. Los medios de transporte por su parte están presenten en 29.8% de las viviendas, muy bajo en comparación con 52.3% del sector urbano.

1.2.3. Otorgamiento de créditos en el sector

Ya que ha sido posible dar a conocer los principales aspectos del crédito y del sector rural de manera independiente, es necesario involucrar ambos temas, iniciando por la descripción del tipo de créditos que son otorgados a los habitantes de estas zonas. Como fue mencionado anteriormente, la clasificación de los tipos de crédito se puede hacer bajo diversos puntos de vista, ya sea por su naturaleza, sujetos a quienes se otorga, actividades o destinos.

En este caso la clasificación está impuesta por las actividades productivas que lo involucran y las garantías otorgadas a cambio. Conforme a las definiciones de Banxico se tiene lo siguiente[7]:

[7] Banco de México. *Divulgación. Sistema Financiero: 4.2 Servicios de Crédito.* [Recurso en línea 2012]

a) Habilitación o avío: Créditos otorgados para la industria y la adquisición de materias primas, para el pago de salarios o los gastos directos de las empresas, garantizados con las mismas materias primas y materiales adquiridos o con los productos finales.

b) Refaccionarios: Créditos destinados a la producción agrícola. Sirven para la adquisición de instrumentos para labrar la tierra, compra de abono y adquisición de ganado, animales de cría, o bien para la plantación de cultivos, para la apertura de tierras o en la compra de maquinaria. Estos créditos quedan garantizados con la maquinaria o instrumentos adquiridos y con los productos agrícolas generados.

c) Prendarios: Créditos con el objetivo de proporcionar recursos para que los sujetos de crédito realicen sus productos primarios o para terminarlos en mejores condiciones de precio, ante situaciones temporales de desequilibrio del mercado.

d) Quirografarios: Créditos sin una garantía específica. Son aquellos en los que el deudor no ha dejado algún bien mueble o inmueble ni tampoco existe una tercera persona (fiador o aval) que se comprometa a pagar el crédito en caso de que el deudor no lo haga. En caso de incumplimiento, la entidad acreedora puede reclamar una parte o la totalidad del patrimonio del deudor por la vía judicial, buscando embargar los bienes para poderlos vender y recuperar el monto colocado. Estos créditos suelen ser con pagarés a plazos cortos, usados para que el deudor pueda cubrir sus necesidades de liquidez.

Características del Crédito Rural

Como detalles específicos del crédito, el otorgamiento a este sector conserva sus características pero queda condicionado a determinados factores:

Largo Plazo: El plazo de estos créditos suele ser mayor en comparación con los créditos de consumo por la naturaleza del sector, ya que los agricultores solicitan recursos para hacer frente a las exigencias del cultivo siempre sujeto a las leyes de la naturaleza. El crédito debe coincidir con el tiempo suficiente para obtener la utilidad de la obra o cultivo.

Interés bajo: Como consecuencia del plazo se suele imponer un interés bajo porque no se podría sostener el interés como el de un crédito de consumo (con plazos de 30 a 90 días).

Garantías: A su vez, con un plazo largo y un interés bajo, es lógica la búsqueda de una compensación con la solidez de una garantía, algunas veces dada por apoyo y fideicomisos. Este sector tiene un uso de garantías especial, desde luego la garantía real, que corresponde al producto o la tierra misma en la que se cultivará, teniendo en cuenta que muchos agricultores no son propietarios o el valor de sus propiedades es menor. Otro tipo de garantías son, por ejemplo, las de un agricultor que posee maquinaria que puede ofrecerse en prenda, siempre y cuando no se requiera para sus actividades ya que el sistema legal exige que se dé la posesión real.

Ubicación: Un factor determinante es la zona de colocación, es decir, la cantidad de créditos que se logran otorgar en áreas geográficas específicas, que debe contemplar incluso los lugares donde habiten campesinos que en virtud de sus propias ocupaciones están alejados de los centros o mercados de capital, teniendo accesos limitados o falta de conocimiento sobre la posibilidad de obtener financiamientos. Un buen manejo de la ubicación al colocar créditos permitirá cumplir con la cobertura del mercado objetivo y presenciar a fondo la situación real del cliente para analizar correctamente sus características y ayudar a conocer el riesgo de la operación.

Actividades involucradas en el sector

Como parte del financiamiento es preciso conocer las actividades del sector que pueden estar sujetas a crédito. Son actividades relacionadas directa o indirectamente con los servicios de la zona y la explotación de recursos naturales o sus derivados. Son procesos en contacto con la naturaleza y donde se utilizan instrumentos que reducen y facilitan el trabajo en el campo.

Están divididas en sectores y sub-sectores conceptualizados por el INEGI:

- Sector primario: Incluye actividades relacionadas con la extracción de bienes y recursos naturales conocidos como materia prima.
    - Agricultura. Actividad a la cual corresponden las técnicas dedicadas al cultivo de diferentes plantas, semillas y frutos, para proveer de alimentos al ser humano o al ganado y de materias primas a la industria.

- Ganadería. Es una actividad del sector primario que se refiere al cuidado y alimentación de animales como cerdos, vacas, pollos, borregos y abejas, entre otros, para aprovechar su carne, leche, huevos, lana, miel y derivados.

- Aprovechamiento forestal. Es la intención de uso de un territorio o espacio declarado jurídicamente como ambiente natural no-acuático en la Ley Agraria 2010, fuera de zonas urbanas o parcelas de cultivo. Incluye bosques, selvas y matorrales.

- Pesca. Es la captura y extracción de peces y otros organismos en aguas salada, salobre o dulce. Es una recolección realizada con el fin de proveer alimentos a la sociedad. La mayor producción proviene del mar.

- Minería. Se refiere a la exploración, explotación y aprovechamiento de minerales, ya sea sólidos (oro y níquel), líquidos (mercurio o el petróleo), quebradizos (yeso o cal) y gaseosos (gas natural). Se encuentran localizados en yacimientos al aire libre o en el subsuelo, a diferentes niveles de profundidad

- Sector secundario: Son actividades que suponen la transformación de los recursos extraídos, muchas de ellas se realizan también en zonas urbanas. Las transformaciones se realizan a través de diversos procesos productivos.

    - Industria manufacturera. Es la actividad económica que transforma una gran diversidad de materias primas en diferentes artículos para el consumo. Está constituida por empresas de todo tipo y tamaño (empacadoras, embotelladoras, fábricas, tortillerías, panaderías, etc.)

    - Construcción. Es una actividad encargada del desarrollo de un país con elementos de bienestar básicos en una sociedad. Sus insumos provienen de otras industrias como el acero, hierro, cemento, arena, cal, madera y aluminio, entre otros, para la construcción de infraestructura.

    - Electricidad. Actividades realizadas por unidades económicas dedicadas a la generación de energía eléctrica y transmisión de la planta generadora a la planta receptora de energía eléctrica para su venta.

- Sector terciario: Implica actividades sin productos tangibles, están conformados en su mayoría por los servicios a la sociedad y el trabajo humano.

    - Comercio. Es la actividad en la cual se intercambian, venden o compran productos. Cuenta con el mayor número de establecimientos y personal ocupado en el país. Están dedicados a alimentos, bebidas y tabaco, seguido por artículos de papelería y uso personal, producto textil y calzado.

    - Servicios. Agrupan una serie de actividades que proporcionan comodidad o bienestar a las personas. Se refieren a servicios públicos como de salud, educativos, financieros, de asistencia social, turismo y entretenimiento.

    - Transporte. Actividades encargadas del mantenimiento y mejora de los medios de transporte para el traslado de personas, mercancías y materias primas de un lugar a otro, ya sea dentro de México o hacia otros países. Se realiza por medios terrestres (carreteras y ferrocarril), aéreos y marítimos.

1.2.4. Financiamiento Rural en la Banca de Desarrollo

De manera específica, la Banca de Desarrollo promueve préstamos a este sector desempeñando un papel determinante y especializado para la disponibilidad y acceso al crédito, busca mejorar las condiciones de vida y un crecimiento económico al impulsar el crédito de largo plazo. Es un sector de financiamiento con montos y riesgos importantes, que se encarga de mantener estables los factores macroeconómicos y sobre todo incrementar mejoras en el mercado.

Tiene la posibilidad de ofrecer créditos tanto a personas físicas como a personas morales, para lo cual otorga créditos directos o indirectos. Los créditos directos o de primer piso, son los créditos tradicionales que existen entre la institución (acreedor) y los clientes del sector rural (deudores), en los cuales se especializará este estudio. Los créditos indirectos o de segundo piso son los que se realizan a través de empresas intermediarias, que son aquellas dedicadas a solicitar altos montos de financiamiento para atender grandes cantidades de demanda de créditos de clientes finales (deudores), pequeños productores rurales con los cuales tienen mayores facilidades de acceso que las instituciones.

Antecedentes del financiamiento al sector rural

Para hablar del financiamiento rural en la actualidad vale la pena conocer su origen, que en la mayoría de los análisis históricos lo refieren a partir de que el país tuvo una mejor organización, en una época de reconstrucción y al haber sido creada la Constitución de 1917 que sirvió como base para reformas sociales, económicas y políticas, buscando restablecer el orden interno[8].

En estas fechas se inició contemplando la redistribución de la propiedad territorial que requería de apoyo financiero, lo cual generó que el Gobierno publicara la Ley de Crédito Agrícola en 1926 junto con la cual se creó el Banco Nacional de Crédito Agrícola, establecido para proporcionar recursos para que los pequeños propietarios cultivaran. Más tarde se autorizó a la Secretaría de Agricultura y Fomento fundar bancos agrícolas ejidales y facilitar el crédito a poseedores de parcelas ejidales organizadas como cooperativas, hasta que se creó el Banco Nacional de Crédito Ejidal para atender a sociedades, uniones y cooperativas, construyendo el eje central del sistema en conjunto con el Banco Nacional de Crédito Agrícola .

A partir de 1954 se crearon Fideicomisos Instituidos en Relación con la Agricultura (FIRA), por parte del Gobierno Federal y con Banco de México, como institución fiduciaria. Buscaban impulsar las labores agrícolas de ejidatarios y pequeños productores con el acceso a descuentos y garantías, desvinculado con intereses políticos e integrado a la Banca Privada. Para 1955 se dio origen a dos leyes, la Ley General de Crédito Agrícola, preocupándose por sistematizar la estructura de manera descentralizada y ajustándose a disposiciones como la Ley General de Títulos y Operaciones de Crédito, y la Ley General de Instituciones de Crédito y Organizaciones Auxiliares. Diez años después se creó el Banco Nacional Agropecuario, S.A. con el fin de descentralizar el crédito en menores tiempos gracias a instituciones regionales.

En 1975 se fusionaron los bancos Agrícola, Ejidal y Agropecuario en lo que se conoció como Banrural. Su objetivo consistió en financiar la producción primaria agropecuaria y forestal, cumpliendo con la tarea de otorgar créditos a productores de bajos ingresos. Se incorporaron posibilidades de participación de asociaciones de interés colectivo, ejidos y comunidades, se hizo una distinción entre los recursos para la producción y la inversión. En los años ochenta fue

[8] Anaya Mora, Miguel Luis. *La Banca de Desarrollo en México.* Chile. CEPAL: 2007.

autorizada la realización de operaciones de Banca Múltiple, motivo que permitió al Estado adecuar la estructura de las instituciones y generar su ley orgánica.

Años después continuaron surgiendo diversos programas[9], el primero fue llamado Apoyos y Servicios a la Comercialización Agropecuaria (ASERCA), creado en 1991 para fortalecer la comercialización en las regiones que producen excedentes ante la eliminación de los precios de garantía. Más tarde PROCAMPO se presentó como un programa compensatorio y transitorio, con vigencia hasta el 2008 y destinado a la producción de cultivos básicos. En 1995 se creó Alianza para el Campo (Alianza Contigo), para aumentar las exportaciones y el ingreso de la producción agropecuaria a una tasa superior a la del crecimiento demográfico.

Sin embargo, el desequilibrio financiero ocasionado por el alto gasto operativo y la contracción crediticia explicada por una arraigada cultura de no pago, consecuencia de décadas de protección gubernamental al sector, generó una reestructuración integral de la Banca de Desarrollo y llevó a la liquidación del sistema Banrural. En 2002 se creó un organismo de la Administración Pública Federal: Financiera Rural, en sustitución a Banrural y sectorizada en la SHCP, con responsabilidad jurídica y patrimonio propio. Su objeto es coadyuvar con el Estado en el impulso al desarrollo de las actividades agropecuarias, forestales, pesqueras y todas aquellas vinculadas al medio rural, con la finalidad de elevar la productividad y mejorar el nivel de vida de la población.

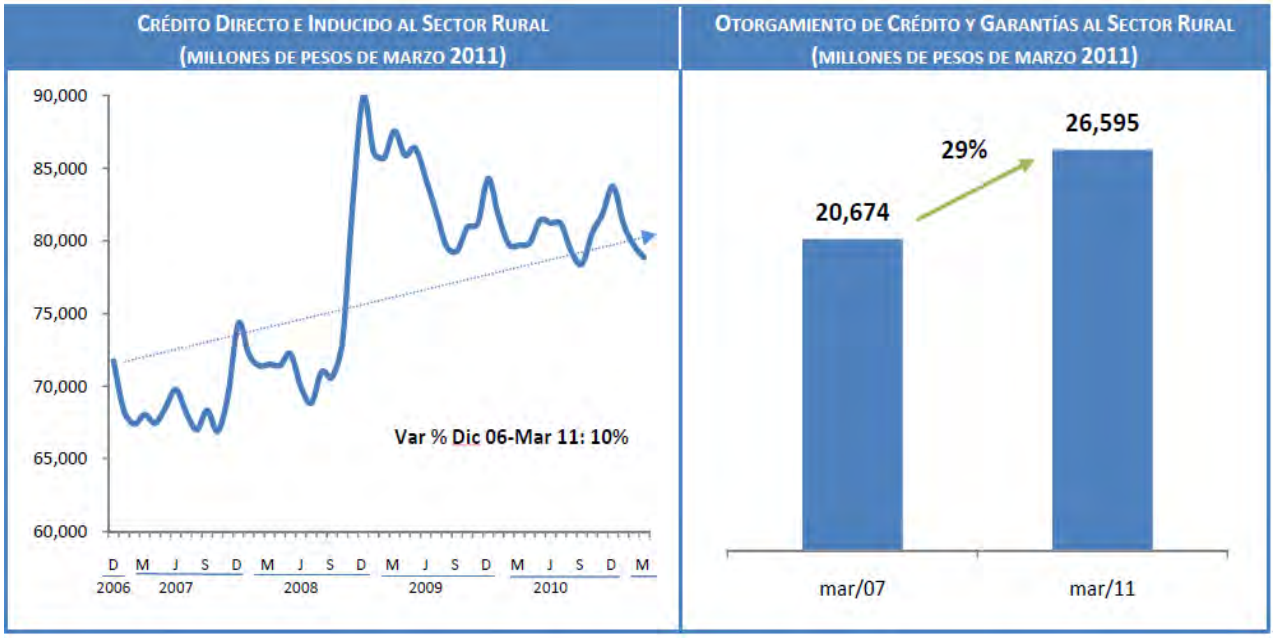
Instituciones y financiamiento rural hoy en día

Como las políticas públicas en el sector rural han cambiado durante los últimos años, debiendo tener mejoras estructurales en las condiciones de vida, los esfuerzos han sido enfocados en hacer más eficiente la cantidad de los gastos públicos para generar bienes requeridos por el Gobierno, más que para mejorar los precios, la oferta y la producción como se pretendía.

Para describir este panorama se mencionan los resultados de la SHCP en el "Informe sobre la Evolución y Resultados de la Banca de Desarrollo" del tercer trimestre de 2010 donde se asegura un incremento en el saldo de crédito al sector rural a lo largo del sexenio con caídas justificadas por fuentes de financiamiento privadas.

[9] Ardila, Sergio. *El Sector Rural en México: Desafíos y Oportunidades.* Nota de política del Banco Interamericano de Desarrollo. Departamento Regional de Operaciones II. División de Recursos Naturales y Medio Ambiente. 2006

Se reporta un aumento en el otorgamiento de créditos y garantías de 29% a productores rurales con ingresos netos anuales menores a 3,000 salarios mínimos.



Fuente: Secretaría de Hacienda y Crédito Público

Dicho reporte corresponde a resultados de las principales instituciones que conforman hoy en día el sector rural en la Banca de Desarrollo, como FIRA y Financiera Rural, regidas por la Ley de Instituciones de Crédito y las leyes orgánicas de sus instituciones, aprobadas con el H. Congreso de la Unión, encargadas de regular y establecer sus objetivos. Sin olvidarse, de acuerdo al Artículo 30º, de preservar y mantener su capital, garantizando la sustentabilidad de su operación, mediante la canalización eficiente, prudente y transparente de sus recursos.

1.3. Riesgos de crédito y probabilidad de incumplimiento

1.3.1. Introducción de temas relacionados con el riesgo

En relación con el otorgamiento de crédito al sector rural están los riesgos involucrados y las técnicas que permiten controlarlos. La palabra riesgo proviene del latín *Riscare* que significa atreverse o arriesgarse, se refiere a la posibilidad de sufrir un daño que resulte en cualquier tipo de pérdida y que por ende requiere una atención especial, la cual ha sido razón suficiente de debates, investigaciones y mejoras continuas.

Para realizar la identificación de los riesgos en cada evento es imprescindible saber que estos se clasifican de acuerdo al tipo de pérdidas y las actividades en las que se presentan. Se relacionan con ciencias naturales y sociales junto con sus diferentes ramas y se dividen en cuantificables (aquellos que cuentan con cifras capaces de medir pérdidas y generar estadísticas) o no cuantificables (eventos imprevistos que no pueden ser estimados cuantitativamente).

En este caso los riesgos a estudiar son los riesgos financieros cuantificables, relacionados con aspectos económico–administrativos, cuyas pérdidas involucran a los participantes del sistema financiero y se derivan de los factores relacionados con la volatilidad del mercado. Son producto de la incertidumbre que existe sobre el valor de los activos financieros y en las operaciones.

En México, los tipos de riesgo financiero están regidos por la CNBV en la Circular Única de Bancos (CUB) y son:

- Riesgos discrecionales: Resultantes de la toma de una posición de riesgo.
    - Riesgo de crédito o crediticio. Pérdida potencial por la falta de pago de un acreditado o contraparte, incluyendo las garantías reales o personales que les otorguen y cualquier otro mecanismo de mitigación de riesgos.
    - Riesgo de liquidez. Pérdida potencial por la imposibilidad o dificultad de renovar pasivos o de contratar otros en condiciones normales para la institución, por la venta anticipada o forzosa de activos a descuentos excesivos para hacer frente a sus obligaciones, o bien, por el hecho de que una posición no pueda ser adquirida oportunamente.
    - Riesgo de mercado. Pérdida potencial por cambios en los factores de riesgo que inciden sobre la valuación o resultados esperados de las operaciones activas, pasivas o causantes de pasivo contingente, tales como tasas de interés, tipos de cambio e índices de precios.
- Riesgos no discrecionales: Resultantes de la operación del negocio.
    - Riesgo operativo. Representa la pérdida potencial por fallas o deficiencias en los controles internos, por errores en el procesamiento y almacenamiento de las operaciones o en la transmisión de información y por resoluciones administrativas o judiciales:

- Riesgo tecnológico. Pérdida potencial por daños, interrupción, alteración o fallas derivadas del uso en el hardware, software, sistemas, redes y cualquier otro canal de distribución de información en la prestación de servicios bancarios.

- Riesgo legal. Pérdida potencial por incumplimiento de las disposiciones legales, de la emisión de resoluciones administrativas y aplicación de sanciones.

Al ser indispensables en la mayoría de actividades de cualquier organización o entidad es de esperarse que cuenten con un área de gestión específica. El área de administración o gestión de riesgos es el área encargada de realizar continuamente los procesos de análisis de la incertidumbre financiera y es requerida para la toma de decisiones en actividades como:

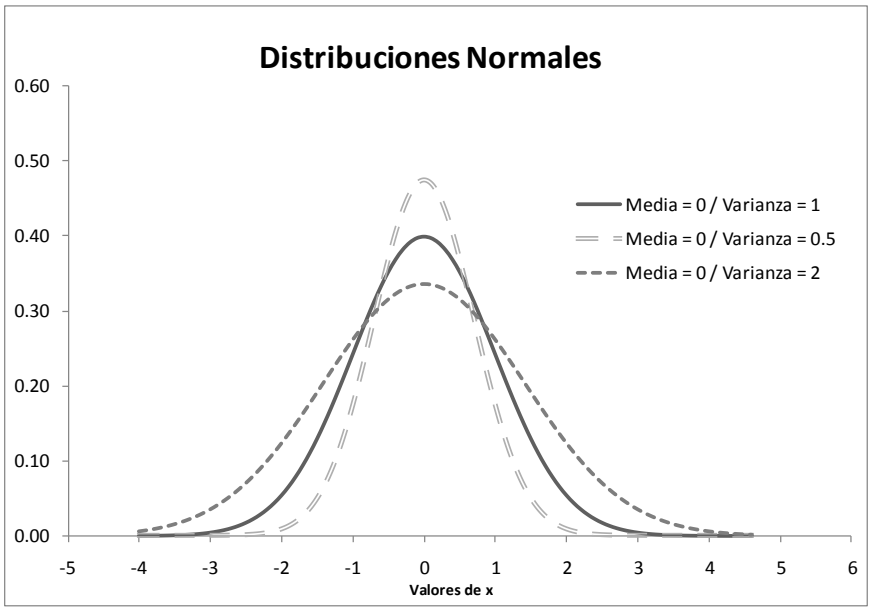
I. Identificar y ponderar los riesgos inherentes a las actividades desempeñadas
II. Determinar niveles apropiados de tolerancia
III. Implementar gestiones estratégicas de riesgo acorde con las políticas
IV. Medir, reportar, controlar y mejorar la operación

El conocimiento de esta gestión inicia con la primera etapa que corresponde al momento de asumir riesgos, sabiendo que tal acción tiene ventajas y desventajas importantes ya que por ejemplo, al asumir mayor riesgo se logra obtener un mayor *rendimiento*[10], es decir, mayores ganancias reflejadas al comparar al tiempo *t* el valor de un activo con respecto a su valor inicial.

Estos rendimientos suelen presentar una distribución de probabilidad de tipo *Normal,* que es también una herramienta fundamental en los riesgos y la estadística, lo cual será corroborado en los capítulos posteriores. Se caracteriza a través de la media *(μ)* y la desviación estándar *(σ),* quedando representada por una curva simétrica y la siguiente ecuación de densidad:

$$\boldsymbol{f(x) = \frac{e^{\frac{-(x-\mu)^2}{\sigma^2}}}{\sigma\sqrt{2\pi}}} \quad \boldsymbol{\sigma > 0, -\infty < \mu < \infty, -\infty < x < \infty}$$

[10] Horche, Karen A. *Essentials of Financial Risk Management*. USA. John Wiley & Sons: Aug 2011.



Dicha distribución es el origen de gran cantidad de modelos, incluso en la forma de distribución de las posibles pérdidas y las posturas frente al riesgo que permiten tomar acciones de manera preventiva y se resumen en: Asumir el riesgo haciendo frente a las consecuencias; disminuirlo al minimizar los factores de riesgo y sus pérdidas; distribuirlo o transferirlo, es decir, repartirlo o traspasarlo a otras entidades; y en caso de ser de alto riesgo, eliminar las variables posibles que pueden hacer que se presente.

En conclusión, los riesgos y los principales aspectos que lo involucran implican una toma de decisiones importante que en un mercado financiero con tal rapidez de avance como el que existe requiere de estudios específicos con el fin de lograr un mejor desempeño en las operaciones y disminuir la cantidad de posibles pérdidas, es por ello que continuando con los objetivos de este estudio, el riesgo de crédito será analizado de manera particular.

1.3.2. Riesgos de Crédito

El riesgo de crédito puede ser referido bajos distintos contextos. Aquél *Riesgo de Crédito* en el que interviene únicamente la falta de pago o el *Riesgo de Contraparte* en el que se incluye la probabilidad de que empeore la capacidad financiera de una entidad:

- "El Riesgo de Crédito se define como la pérdida potencial que se registra con motivo del incumplimiento de una contraparte en una transacción financiera (o en alguno de los

términos y condiciones de la transacción). También se concibe como un deterioro en la calidad crediticia de la contraparte o en la garantía o colateral pactada originalmente" (De Lara Haro, 2005, p. 163).

- "El Riesgo de Contraparte existe cuando se da la posibilidad de que una de las partes de un contrato financiero sea incapaz de cumplir con las obligaciones financieras contraídas, haciendo que la otra parte del contrato incurra en una pérdida. El riesgo de crédito es el caso particular cuando el contrato es uno de crédito, y el deudor no puede pagar su deuda. Recientemente, además del caso de incumplimiento, se han incorporado eventos que afectan el valor de un crédito, sin que necesariamente signifique incumplimiento del deudor. Esto ocurre típicamente por cambios en la calidad de un crédito, cuando una calificadora lo degrada. Cuando esto ocurre, significa que la calificadora considera que ha aumentado la probabilidad de incumplimiento del emisor de la deuda, y por lo tanto el crédito vale menos ya que se descuenta a una tasa mayor" (Recurso en línea Banco de México, 2005, p. 7).

Existen otros aspectos de riesgo en la operación de crédito como la concentración de cartera, que representa la acumulación de un conjunto de créditos con montos altos en ciertos sectores de la población y que en caso de catástrofe, afectarían drásticamente la cartera. Su solución se da con la diversificación, evitando concentraciones y cubriendo la mayor cantidad posible de mercados.

Para los fines de este estudio se usará el riesgo de crédito, existente a causa de un incumplimiento por el atraso en el pago de las obligaciones. Los incumplimientos suelen estar causados de manera evolutiva con la alteración de aspectos de todo tipo, ya sean retrocesos en la solvencia de los clientes, problemas de liquidez, quiebra o disminución de ingresos, conflictos con la actividad que se desempeña, nivel de vida, cambios ambientales, condiciones sociales e incluso por falta de voluntad. Son aspectos que deben ser evaluados en cantidad y calidad lo más confiable posible, por cada cartera de crédito y cada institución, haciendo congruencia con el sector de la población al que se refiere. Debido a su existencia se da el cálculo de la *probabilidad de incumplimiento (PI),* que mide la posibilidad de que un cliente se encuentre en esta situación.

La *Exposición (EAD Exposure at Default),* comúnmente conocida como *Monto Expuesto (ME*), es otro factor del riesgo de crédito que representa la cantidad máxima que se encuentra en

posibilidades de sufrir pérdidas. De igual modo lo es la *Severidad (LGD Loss Given Default),* que se trata de la proporción real y efectiva que se pierde al momento del incumplimiento e incluye todos los gastos, incluso los generados por mora; puede ser vista en términos de la tasa de recuperación *(T).* Ambos se definen formalmente conforme a los términos de los Acuerdos de Capital de Basilea I y Basilea II que serán vistos a continuación.

"Exposición al Riesgo de Crédito (EAD) es el monto total comprometido con el deudor al momento de Incumplimiento; en consecuencia, su estimación comprende la exposición potencial por operaciones contingentes que puedan convertirse en cartera en el futuro" (SBEF 2005, P.87)

"Pérdida en caso de incumplimiento (LGD) es la pérdida que se produce después de que el Prestatario ha Incumplido. Se expresa como un porcentaje de la exposición al riesgo de crédito (EAD), también puede definirse como uno menos la tasa de recuperación (1 – R). La pérdida en caso de incumplimiento (LGD) se puede minimizar cuando el crédito está respaldado por un colateral o una garantía" (SBEF, 2005, P.154-155)

Estos factores son de apoyo para calcular la *Pérdida Esperada (PE):* Pérdida promedio asociada al deterioro normal de los activos. Relacionada con las reservas preventivas como un costo:

$$PE = ME * (1 - R) * PI$$

Junto a este valor existe la *Pérdida No Esperada*, que es la pérdida no anticipada y está por encima de la pérdida esperada. Es donde realmente se presenta el riesgo por no saber el tamaño real que tendrá. Ambas son parte de modelos y medidas para las posturas de acción que permiten disminuir las pérdidas y contingencias.

Métodos para calcular el Riesgo de Crédito

En la administración de riesgos de crédito de cualquier entidad, incluyendo aquellas dedicadas al sector rural, la cuantificación es el campo más explorado, no obstante representa día a día una mayor dificultad. Existen gran cantidad de modelos para su cálculo, autorizados por comités internacionales como el *Comité de Basilea*, que consiste en una organización internacional financiera y de investigación para la regulación y los estándares de prácticas sanas en el mercado.

En 1988 emitió el primer Acuerdo de Capital *(Basilea I)* y el método estándar para calcular los requerimientos de capital por riesgos de crédito basándose en sus activos donde por cada tipo de activo o préstamo, divididos en cuatro categorías, se debían aplicar ponderaciones de los riesgos inherentes a él sin importar elementos particulares de instituciones, obteniendo un factor mínimo del 8% como resultado de dividir los recursos de capital entre la suma de ponderaciones.

Para 2003 después de diversos debates e inconsistencias, el comité realizó cambios en el Acuerdo de Capital *(Basilea II),* indicando para el método estándar nuevas categorías para clasificar a los clientes en portafolios y poder realizar la ponderación más adecuada al perfil de cada activo, además de utilizar el riesgo de contraparte en métodos como *Credit Risk*. Fueron cambios que establecieron metodologías internas a través de un sistema conocido como *IRB (Internal Rating Based Approach),* que da libertad de hacer estimaciones en apego a las características, estadística, productos y activos de cada entidad, optimizando con esto los requerimientos de capital y haciéndolos más precisos. Últimamente se han tenido mejoras en los estándares y guías para las prácticas del mercado, con el acuerdo que estará siendo el nuevo requisito de los próximos años *(Basilea III)* y en el cual se mantienen los sistemas internos de calificación IRB.

El modelo de *Valor en Riesgo (VAR*) es uno de los principales involucrados en los Acuerdos de Basilea, es un método que asocia probabilidades para aproximar el nivel de pérdidas (esperadas y no esperadas) posibles de sufrir en condiciones normales del mercado. Es una medida de posición en una proporción de la distribución de pérdidas, mejor conocida como cuantil. Está evaluada en un intervalo de tiempo y con un nivel de confianza determinado que hace referencia al porcentaje de casos que se pretende no rebasen los niveles de pérdida establecidos. Para Basilea II se usa un nivel de confianza 99.9%.

Todos estos métodos son benéficos para optimizar las operaciones de crédito, para analizar la cartera de forma desagregada por área geográfica, línea de negocio y/o sector económico, entre otros, además de aportar respuestas sobre variaciones en cualquier variable del entorno, permitiendo hacer más eficiente el establecimiento de precios y políticas. Por ello es imprescindible tener la información adecuada para construirlos y poder tener la capacidad de medir otros elementos análogos.

En esta ocasión el análisis será enfocado en la probabilidad de incumplimiento, uno de los requerimientos primordiales de estos modelos. Su estimación, en apego al Acuerdo de Basilea II, depende del riesgo (crédito o contraparte) y de la fuente que califique (interna o externa). Para el riesgo de contraparte están las fuentes externas (Calificaciones o Ratings), que en respuesta a la necesidad de información sobre la solvencia de las empresas cuentan con un tratamiento en el cual se etiquetan con indicadores que hacen referencia a su calidad financiera. Dichos valores se dan por parte de agencias como Fitch, HR ratings, Moody's y Standard & Poors, conocidas como calificadoras. Por otro lado, para el riesgo de crédito donde se hará énfasis, se requieren los sistemas internos *(IRB),* sobre todo cuando se trata de personas físicas como en este caso.

1.3.3. Cálculo de la Probabilidad de Incumplimiento

La probabilidad de incumplimiento también llamada *Default*, es un indicador que da como resultado un valor máximo de 1 para alertar el posible retraso y/o falta de pago de un crédito y un valor mínimo de 0 que indica la posibilidad de que el crédito a otorgar sea recuperado. Es un valor que debe ser evaluado continuamente y ajustado a cada tiempo, cliente y entidad.

En México su cálculo es un requerimiento legal en la circular única de bancos expedida por la CNBV para la Banca de Desarrollo y el Sector Bancario:

"Artículo 80.- Las Instituciones en la administración del riesgo de crédito o crediticio, como mínimo deberán:

II. Por lo que hace al riesgo de la cartera crediticia en específico:

   a) Medir, evaluar y dar seguimiento a su concentración por tipo de financiamiento, calificación, sector económico, zona geográfica y acreditado.

   b) Dar seguimiento periódico a su evolución y posible deterioro, con el propósito de anticipar pérdidas potenciales

   c) Calcular la probabilidad de incumplimiento, así como la exposición al riesgo por parte de los deudores.

d) Desarrollar sistemas de medición que permitan cuantificar las pérdidas esperadas de toda la cartera.

e) Estimar las pérdidas no esperadas de toda la cartera.

f) Comparar sus exposiciones estimadas de riesgo de crédito o crediticio, con los resultados efectivamente observados. En caso de que los resultados proyectados y los observados difieran significativamente, se deberán realizar las correcciones necesarias.

g) Calcular las pérdidas potenciales bajo distintos escenarios, incluyendo escenarios extremos."

Uno de los métodos más utilizados para calcular esta probabilidad de un modo cuantitativo, eficiente y objetivo, es el *Credit Scoring o Score de Crédito*, que engloba un conjunto de técnicas donde se maneja información de las condiciones de vida de las personas o el estatus de las empresas, posterior a la recopilación de datos durante la solicitud de un crédito. Algunas de estas dan el valor real de la probabilidad y otros simplemente clasifican a los clientes.

Es un mecanismo de evaluación de los clientes que otorga una puntuación o calificación final de manera inmediata, comparable con valores fijos para aceptar o rechazar el otorgamiento de créditos a cada sujeto. Es útil para cuando existe un alto volumen de clientes y su correcta aplicación es de gran apoyo para la colocación de recursos en el mercado.

CREDIT SCORING Y ALGUNOS MÉTODOS ALTERNATIVOS

2.1. Aspectos generales del Credit Scoring

2.1.1. Introducción al método

El Scoring es un método que ha sido ampliamente desarrollado a través de los años y cuya utilidad radica en calificar a los individuos de cualquier población con información propia de cada entidad, hecho que permite dar una aplicación a cualquier mercado pero que puede llegar a ser desatendido por la importancia de cubrir otras necesidades. Es una técnica de minería de datos (*Data Mining*) donde se busca descubrir patrones y relaciones con el fin de clasificar. En este caso la evaluación que se realiza es a nivel crediticio para diferenciar entre clientes "buenos" o "malos", en otras palabras cumplidos o incumplidos en cuanto a obligaciones de pago se refiere.

Como tal el *Credit Scoring* es un término que se define como una herramienta para medir la probabilidad de que un crédito sea recuperado satisfactoriamente[11]. Se basa en el hecho de predecir el comportamiento de futuros solicitantes de crédito con características similares a solicitantes previos. Calcula el nivel de riesgo y hace objetivas las decisiones en el proceso de otorgamiento con el uso de fórmulas que asignan información (atributos o características) a las variables para dar un resultado (probabilidad de incumplimiento y/o clasificación). Por otro lado, el término *Score* técnicamente se define como la suma total de los puntos asignados a cada variable de los clientes. Es el puntaje final de la evaluación y se compara con el límite de aceptación o rechazo (*Cut-off*) del préstamo. Su construcción se realiza con diversas metodologías cuyos supuestos y procedimientos deben analizarse con base en los objetivos y la información disponible. En el presente estudio se hará hincapié en algunas técnicas estadísticas, buscando fomentar la aplicación de esta ciencia.

En este ámbito existe también la posibilidad de otorgar una calificación a los clientes durante la vida del crédito, esto es lo que se conoce como *Scoring de Comportamiento* (*Behavioural Scoring).* Controla las cuentas de créditos existentes y determina el perfil de riesgo de los

[11] Finance & Leasing Association. *Guide to Credit Scoring.* United Kingdom 2000

deudores de manera continua, sirve para administrar límites, identificar cuentas rentables o para el ofrecimiento de productos. No representa parte del interés pero su cálculo puede realizarse a través de la misma metodología, es sólo la información de los clientes la que cambia al incluir variables que hagan referencia al historial que se va generando con el transcurso del crédito y las fechas de pago establecidas, información que no se tiene al principio de una solicitud como en el credit scoring de origen.

Es preciso saber que aún con la importancia de este método, es inevitable que llegue a ser sustituido debido a que no es el único método existente, sobre todo en aquellos lugares con historial de clientes no disponible y/o información poco confiable o incongruente. Para esto se buscan soluciones como datos de otras entidades o productos con características semejantes. En las operaciones de crédito las evaluaciones tradicionales dadas por el juicio humano son una opción alternativa pero que implican la revisión por parte de un analista financiero de cada una de las variables, lo cual resulta ser tardado y costoso.

2.1.2. Origen del Credit Scoring

Se tienen registros iniciales de la idea de scoring después de la Segunda Guerra Mundial, con la aparición de las calculadoras, cuando existió una formación científica propuesta por R. A. Fisher (1936) quien describió una técnica llamada Análisis Discriminante pero fue en 1941 que D. Durand se dedicó a distinguir entre buenos y malos préstamos, proyecto para Estados Unidos que al final nunca fue usado para predecir. No tuvo un gran auge desde su inicio porque no aceptaban sustituir los procesos y automatizarlos, lo consideraban injusto por la idea de discriminar casos que de ser evaluados más a fondo o bajo otros criterios de los analistas pudieran ser correctos.

Con el crecimiento en el mercado crediticio el uso del scoring se convirtió en una herramienta capaz de reducir el tiempo que ocupaba el proceso de decisión. Su aceptación se dio con la aparición de tarjetas de crédito, altos consumos y requerimientos legales. En 1967 apareció un estudio en el que se analizaron indicadores de empresas en bancarrota, de manera individual e independiente, que podía clasificar y predecir la quiebra. Un año después se realizó un análisis para la evaluación simultánea de los indicadores *(multivariado),* planteado por Edward Altman.

En este modelo llamado *Z-score,* se seleccionaron 66 corporaciones del sector manufacturero que cotizaban en la bolsa, 33 en bancarrota y 33 en no-bancarrota, calculó sus indicadores y determinó la solvencia de las empresas, logrando conseguir una función que distinguía entre una y otra, mostrando que el comportamiento de esta función era distribuido normalmente. Para este análisis las razones financieras se combinaban linealmente para predecir un resultado (*Z*), partiendo de la definición básica que considera la suma de elementos $X = X_1, X_2, \dots, X_s$ de un conjunto ponderado a través de coeficientes constantes $a = a_1, a_2, \dots, a_s$ expresados de forma vectorial en cualquier espacio:

$$Z = a \cdot X = \begin{bmatrix} a_1 \\ a_2 \\ \vdots \\ a_s \end{bmatrix} [X_1 \quad X_2 \quad \dots \quad X_s] = a_1X_1 + a_2X_2 + \cdots + a_sX_s$$

En el vector *X* se colocaron las características a evaluar y para conocer los coeficientes (*pesos*) Altman usó una regresión lineal múltiple que analiza la relación entre las variables independientes para poder explicar una variable dependiente de ellas (*Z)*:

$$Z = 1.2X_1 + 1.4X_2 + 3.3X_3 + +0.6X_4 + 0.99X_5$$

Donde: X_1 = *Capital de trabajo / activos totales*

X_2 = *Utilidades retenidas / activos totales*

X_3 = *Utilidades antes de impuestos/ activos totales*

X_4 = *Valor de mercado de la acción / valor de mercado de la deuda*

X_5 = *Ventas / activos totales*

Era una función con variables estadísticamente significativas y cuyos coeficientes eran fijos. Su uso se daba al incorporar las condiciones *X* de una nueva empresa, que si conseguía un valor de $Z > 2.99$ era tomada como una empresa solvente que no caería en bancarrota y si tenía un valor de $Z \leq 1.81$ era muy probable que en un futuro fuera insolvente. Altman hizo modelos para otros sectores y fue a partir de estos estudios que nuevos modelos fueron surgiendo, con los recursos estadísticos y tecnológicos implementados año tras año, buscando la creación de nuevas funciones cada vez más precisas y adecuadas a la realidad, con modelos cada vez más eficaces pero complejos, todos con el objetivo de poder tomar la mejor decisión con respecto al otorgamiento de créditos en cualquier sector del mercado.

2.1.3. Principios y procedimientos

El credit scoring cuenta con principios y conceptos indispensables para lograr su objetivo de clasificación. Uno de ellos incluye en la mayoría de los modelos la creación de una *Scorecard,* que son el conjunto de atributos y puntos de evaluación para poder clasificar a cada individuo solicitante de algún producto, se encuentran como tablas para colocar las variables y los posibles valores de respuesta con sus respectivos puntos que formarán el puntaje final (*score*).

Estás tablas son llenadas con variables que cuentan con información de los solicitantes de crédito, en forma de preguntas y respuestas, con características que varían dependiendo del mercado y la persona. Un credit scoring de empresas utiliza activos, inventarios, costos de mercaderías, pasivos, activos, ganancias retenidas, apalancamiento, tamaño de la empresa y liquidez, mientras que uno para grandes volúmenes de venta, como es el caso de las tarjetas de crédito, utilizan características como edad, estado civil, cantidad de dependientes económicos, tiempo de permanencia en el domicilio y empleo actual, nivel educativo, bienes y propiedades, ingresos e historial de créditos. En general son variables apegadas a los productos por ofrecer, al sector y la población que se busca atender, con una estructura adecuada como para que una vez construida la scorecard con las preguntas y posibles respuestas, las características de solicitantes de crédito puedan apegarse a la realidad y ser contestadas en su totalidad.

A nivel internacional, conforme a la *Guía para Credit Scoring*, existen principios que involucran a la scorecard y se dividen en cada momento de la operación crediticia: Diseño, implementación, operación, toma de decisiones, comunicación con los solicitantes y apelaciones. Son útiles para no perder de vista los objetivos y poder hacer congruente el modelo, pudiendo incluirlos como políticas de crédito. Los principios más relevantes son:

- Los datos deben ser capturados y codificados correctamente.
- El resto de los departamentos deben estar informados del sistema.
- El sistema tendrá un control de resguardo suficiente para conservar su integridad.
- Los procedimientos estarán documentados y serán revisados regularmente.
- Las Scorecards serán monitoreadas periódicamente para asegurar que aún son apropiadas para la población, mediante revalidaciones periódicas de la predicción.

El *Cut-off* que determina el umbral para el límite de aceptación o rechazo, está también involucrado en el método. Es un balance en los datos que pondera la estructura y prioridades del negocio, con respecto a los principios básicos de la institución para asumir el riesgo. Es elegido con base en los resultados del modelo, metas de la entidad y en el criterio de los responsables del otorgamiento, es sin duda un punto de equilibrio y optimización.

Hay distintas formas de establecer el cut-off conforme a las propuestas de algunos autores. Una posibilidad es tener un área con intervalos de ± 5 y 10% donde el puntaje de las solicitudes que caen en ella será evaluado aparte. Otra solución es usar límites de aceptación, que lo vuelven más complejo y asociando límites de crédito con una función. Existen otros casos que toman en cuenta los activos y las cuentas, se acostumbra a definir el cut-off ya sea para reducir costos, al evitar el mayor número de pérdidas y asumir el menor riesgo posible al aceptar con cierto rigor los créditos, o de aumentar ganancias, buscando colocar en préstamo la mayor cantidad de recursos posibles pero asumiendo mayores riesgos. Otra buena opción es trabajar con diferenciales de precios, ajustándolos al riesgo y teniendo un mayor número de rangos de cut-off. Puede así existir un primer rango alto, que asume mucho riesgo, para que los mejores candidatos tengan el producto más actualizado; otro para un producto estándar con una tasa baja de interés para colocar más préstamos; un tercero muy bajo, en el que se asuma poco riesgo y el acceso sea menor (para productos con problemáticas); y alguno intermedio adecuado a un producto estándar con tasas y precios promedio[12].

La muestra es otro tema prioritario. Para el método de scoring será aleatoria, representativa y suficiente para garantizar una estadística significativa lo más reciente posible, aunque vale la pena no olvidar que como cualquier modelo tiene un sesgo inherente por contar con información histórica sólo de los clientes que fueron aceptados porque se creía que pagarían, quedando excluidos aquellos rechazados por los analistas y los cuales no se sabe si realmente habrían incumplido. La determinación del tamaño es lo más relevante y es un tema que ha tenido mucho debate. Se cree que una proporción de 50% de cada tipo de cliente (buenos y malos) es un buen punto de inicio, aunque la realidad es que en la práctica difícilmente se encontrará la posibilidad de contar con las proporciones precisas, la solución suele ser conseguir la mayor cantidad de clientes malos posibles y un número igual o mayor de clientes buenos.

[12] Thomas, Lyn C.; Edelman, David B.; Crook, Jonathan N. *Credit Scoring and Its Applications*. USA. SIAM: 2002.

Construcción del modelo

Se pueden enlistar diversos pasos para la construcción del credit scoring, como en el siguiente caso que se incluye información aportada por parte del CONAC[13]:

1) Conformar la base de datos: Es la obtención de la muestra en un periodo de tiempo, con información de la solicitud de crédito y el resto de las fuentes disponibles.

2) Depurar los datos: Es el momento para eliminar o corregir los datos erróneos, inconsistentes y nulos. Se evalúa cada variable y elemento de la muestra, tomando una postura crítica en la cual si una gran cantidad de datos cuenta con errores será mejor eliminar la variable o el elemento de la muestra por completo. Una vez depurada, se separa la muestra tomando un 75%, para el modelo y un 25% para la validación.

3) Seleccionar y estudiar las características: Se analizan las variables, con herramientas de estadística descriptiva, buscando saber cuáles son aquellas que presentan mayores diferencias en las proporciones de clientes buenos y malos, es el primer punto decisivo para saber que variables incorporar o no al modelo, cuáles pueden ser categóricas o numéricas.

4) Determinar la clasificación y el modelo: La clasificación se puede estimar con cualquiera de los modelos que serán mencionados a continuación. Se busca una función y/o herramienta que de como resultado la probabilidad de incumplimiento.

5) Asignar el Cut-off y validar: Se establece con las políticas de crédito y el riesgo que se pretende asumir, pudiendo hacer segmentaciones para generar distintos niveles de riesgo. Se valida tomando la parte restante de la muestra (25%) y corroborando que los resultados se apegan al modelo original (backtesting).

6) Elaborar la Scorecard: En caso de que el modelo lo permita se construirá la tabla con cada variable (preguntas) y sus respectivos valores (respuestas), a los cuales se les pueden hacer traslaciones o cambios de escala para hacer más cómodas las respuestas. Dicha tabla tendrá los pesos de cada variable para otorgar la calificación.

[13] Nieto Murillo, Soraida; Pérez Salvador, Blanca Rosa; Soriano Flores, José Fernando. *Crédito al Consumo: La estadística aplicada a un problema de riesgo crediticio.* México. Colegio Nacional de Actuarios: 2011.

2.2. Métodos estadísticos de implementación

2.2.1. Modelo de Regresión Lineal

El modelo de regresión lineal es el primero de los métodos estadísticos alternativos por ser una herramienta indispensable y base de otros métodos desarrollados y mejorados para el credit scoring a partir de modelos como el de Altman. Es una regresión multivariada para cada elemento *i* de la muestra que explica y predice su comportamiento a través de la relación que existe entre las variables independientes y una variable dependiente:

$$y_i = c_0 + c_1 x_{1i} + c_2 x_{2i} + \cdots + c_k x_{ki} + \varepsilon_i$$

Donde: y_i = Variable dependiente con $i = 0, 1, 2, \ldots, n$.

x_{ji} = Variables independientes con $j = 0, 1, 2, \ldots, k$.

c_j = Coeficientes o parámetros.

ε_i = Errores aleatorios independientes.

Visto de manera matricial son dos vectores de orden $n \times 1$ (vectores ε y *Y),* una matriz *X* de rango k y orden $n \times k$ con $k < n$ para garantizar la dependencia lineal entre los vectores, además del vector *C* de orden $k \times 1$ con los coeficientes que dan valor a la relación:

$$Y = \begin{bmatrix} Y_1 \\ Y_2 \\ \vdots \\ Y_n \end{bmatrix} \quad X = \begin{bmatrix} 1 & X_{11} & X_{12} & \cdots & X_{1k} \\ 1 & X_{21} & X_{22} & & X_{2k} \\ & \vdots & & \ddots & \vdots \\ 1 & X_{n1} & X_{n2} & \cdots & X_{nk} \end{bmatrix} \quad C = \begin{bmatrix} c_0 \\ c_1 \\ \vdots \\ c_k \end{bmatrix} \quad \varepsilon = \begin{bmatrix} \varepsilon_0 \\ \varepsilon_1 \\ \vdots \\ \varepsilon_n \end{bmatrix}$$

Dicha estructura junto con los supuestos básicos forman los principios del modelo:

- Linealidad – La relación entre las variables debe adoptar la forma $Y = XC + \varepsilon$.
- Normalidad – Los errores al ser quienes evitan que la relación de las variables sea exacta, usualmente se suponen distribuidos normalmente con media 0 y matriz de varianza y covarianza $E(\varepsilon\varepsilon') = \sigma^2 I_n$ es decir $\varepsilon \rightarrow N(0, \sigma^2 I_n)$. Sus variaciones generan los valores aleatorios de la variable *Y*, la cual se distribuye normalmente.

- Homocedasticidad – Este supuesto, que va unido al supuesto de normalidad, indica que la varianza de los errores es constante. Dado el vector transpuesto ε' de orden $1 \times n$, el producto $\varepsilon\varepsilon'$ es una matriz cuadrada simétrica de orden $n \times n$ y su valor esperado es una matriz con valor de ceros fuera de la diagonal I_n por la falta de correlación y valor de σ^2 en la diagonal, tal que $E(\varepsilon\varepsilon') = \sigma^2 I_n$.

- Variables independientes – Los valores de la matriz *X* corresponden a *n* observaciones de las *k* variables de la muestra, las cuales son carentes de propiedades estocásticas.

Desarrollo y solución

A continuación surge la resolución del problema: Estimar los parámetros o coeficientes desconocidos para indicar la relación de cada variable. Para la estimación del vector *C* existe el procedimiento de mínimos cuadrados, que parte de la estructura inicial, utilizando un vector de $n \times 1$ con valores residuales de la muestra con la forma $e = Y - X\hat{C}$. Dichos residuales se espera que sean independientes: Cualquier par de elementos $(e_{\mathrm{m}}, e_{\mathrm{n}})$ del vector de residuales no debe afectar entre sí sus valores y/o probabilidades $\mathrm{E}(e_{\mathrm{m}}, e_{\mathrm{n}}) = 0 \quad \forall\, \mathrm{m} \neq \mathrm{n}$. Se busca tomar la suma de cuadrados, multiplicando al vector *e* por su transpuesto *e'*:

$$\sum_{i=1}^{n} \mathrm{e}^2 = ee' = \left[Y - X\hat{C}\right]'\left[Y - X\hat{C}\right]$$

$$= \left[Y' - \hat{C}'X'\right]\left[Y - X\hat{C}\right]$$

$$= Y'Y - 2\hat{C}'X'Y + \hat{C}'X'X\hat{C}$$

Observando que por ser escalares cada uno de los sumandos son equivalentes a su transpuesto. Más tarde se busca la minimización de la suma de cuadrados al derivando con respecto a $\hat{C}$:

$$\frac{\delta ee'}{\delta \hat{C}} = -2X'Y + 2X'X\hat{C}$$

Y se iguala a cero la ecuación, para poder despejar $\hat{C}$:

$$-2X'Y + 2X'X\hat{C} = 0 \quad \Rightarrow \quad X'X\hat{C} = X'Y$$

El vector resultante que estima el valor de los coeficientes se encuentra al tomar la matriz inversa, tal que $\hat{C} = (X'X)^{-1} X'Y$. Logrado esto, se pretende confirmar que este estimador tiene la propiedad de ser insesgado, comprobándolo al sustituir el vector *Y*:

$$\begin{aligned} \hat{C} &= (X'X)^{-1} X'(XC + \varepsilon) \\ &= (X'X)^{-1} X'XC + (X'X)^{-1} X'\varepsilon \\ &= C + (X'X)^{-1} X'\varepsilon \end{aligned}$$

Resultado que con el valor esperado da como resultado el insesgamiento, ya que $E[\varepsilon] = 0$.

$$E[\hat{C}] = E[C + (X'X)^{-1} X'\varepsilon] = E[C] + E[(X'X)^{-1} X'\varepsilon]$$
$$= E[C] + (X'X)^{-1} X'\, E[\varepsilon] = C$$

Al presentarse en combinación lineal con los errores que se distribuyen normalmente, este resultado hereda el comportamiento, conservando propiedades de insesgamiento y varianza con una distribución normal multivariada $\hat{C} \to N(C, \sigma^2(X'X)^{-1})$.

Pruebas de ajuste y medición

Resuelto el problema de estimación, vale la pena evaluar el ajuste del modelo a la realidad. Para ellos existen estadísticos que comparan las estimaciones con los valores reales.

Coeficiente de Correlación Simple: Es un coeficiente para medir la relación lineal entre dos variables (x_i, x_j) del modelo, tanto en magnitud como en dirección. Se calcula como un cociente de la covarianza de las dos variables entre sus desviaciones estándar, producto del momento de *Pearson.* Toma valores entre (-1,1) siendo los valores extremos -1 ó 1 donde hay una relación lineal directa o inversa perfecta y 0 el valor que indica ausencia de correlación.

$$r = \frac{cov(x_i, x_j)}{\hat{\sigma}(x_i)\hat{\sigma}(x_j)}$$

Su cálculo empírico utiliza *n* observaciones para estimar la relación de la variable dependiente con cada variable independiente y desviaciones que trasladan el origen al punto de medias:

$$\begin{matrix} y = Y - \bar{Y} \\ x = X - \bar{X} \end{matrix} \Rightarrow r = \frac{\sum xy}{\sqrt{\sum x^2}\sqrt{\sum y^2}}$$

Coeficiente de Correlación Parcial: Es una variación del coeficiente de correlación simple, que se usa para medir la relación entre dos variables manteniendo constante la presencia de otras variables. Se supone el coeficiente de modelos con la relación de la variable x_1 y la variable x_2, fijando x_3 con residuales para cada elemento *i* de la muestra tal que:

$$\begin{matrix} k_i = x_{1i} - \hat{c}_3^{(k)} x_{3i} \\ l_i = x_{2i} - \hat{c}_3^{(l)} x_{3i} \end{matrix} \Rightarrow r_{x1,x2-x3} = \frac{\sum k_i l_i}{\sqrt{\sum {k_i}^2}\sqrt{\sum {l_i}^2}}$$

Resultado que puede expresarse con los coeficientes de correlación simple:

$$r_{x1,x2-x3} = \frac{r_{x1,x2} - r_{x1,x3} r_{x2,x3}}{\sqrt{1 - {r_{x1,x3}}^2}\sqrt{1 - {r_{x2,x3}}^2}}$$

Coeficiente de Determinación: Medida para precisar la proporción explicada de la variable dependiente, se asocia con la variabilidad de ajuste de la recta y llamado coeficiente R^2. Se estima con los valores normales y promedio $\bar{Y} = c_0 + c_1\bar{x}_1 + c_2\bar{x}_2 + \cdots + c_k\bar{x}_k + \bar{\varepsilon}$, buscando desviaciones de cada elemento *i* de la muestra:

$$Y_i - \bar{Y} = (c_0 - c_0) + c_1(x_{1i} - \bar{x}_1) + \cdots + c_k(x_{ki} - \bar{x}_k) + (\varepsilon_i - \bar{\varepsilon})$$

De esta forma el modelo anula la ordenada al origen, pasando a través de él y quedando las variables independientes x_i y los residuales e_i como desviaciones que bien pueden calcular estimadores mediante el procedimiento de mínimos cuadrados $y_i = \hat{c}_1 x_{1i} + \cdots + \hat{c}_k x_{ki} + e_i$

Se hace la multiplicación de residuales con $\hat{C} = (X'X)^{-1} X'y$:

$$e'e \quad = \left[y - X\hat{C}\right]'\left[y - X\hat{C}\right] = \left[y' - \hat{C}'X'\right]\left[y - X\hat{C}\right] = y'y - \hat{C}'X'y$$

Así el producto escalar $y'y$ representa suma de cuadrados de las desviaciones de la variable dependiente $y'y = \sum y^2$ con $y'y = \hat{C}'X'y + e'e$. Si se divide entre $y'y$ la proporción de variabilidad es lo que se conoce como coeficiente de determinación de R^2.

$$1 = \frac{\hat{C}'X'y}{y'y} + \frac{e'e}{y'y}$$

$$R^2 = \frac{\hat{C}'X'y}{y'y} \quad \text{O análogamente} \quad R^2 = 1 - \frac{e'e}{y'y}$$

Si a la suma de cuadrados de los residuales y de la variable independiente se dividen sus grados de libertad, se tiene el coeficiente de determinación ajustado, para comparar modelos que provengan de los mismos datos pero con distinto número de variables

$$R^2{}_{ajustado} = 1 - \frac{e'e/n-k}{y'y/n-1}$$

Coeficiente de Correlación Múltiple: Medida capaz de medir la relación lineal entre la variable dependiente y la combinación lineal de variables independientes, denotada por *R*. Se calcula con la raíz cuadrada positiva del coeficiente de determinación.

Limitantes y aplicación en Credit Scoring

El uso de la regresión lineal ha sido de gran importancia a lo largo del tiempo, para el credit scoring el modelo analiza las características de *n* clientes de la muestra para calificar el cumplimiento o incumplimiento del pago de un crédito con las variables:

y_i = Variable dependiente: Resultado de clasificación del cliente con *i = 0, 1, 2,..., n* elementos.

x_{ji} = Variables independientes: Característica *j* del cliente *i* donde *j = 0, 1, 2,..., k.*

Sin embargo esta alternativa suele ser sustituida por otros modelos debido a las desventajas que tiene la regresión lineal, en principio a causa los supuestos básicos que requiere y sobre todo al considerar que la matriz *X* utiliza valores sin propiedades estocásticas y que la variable dependiente, como resultado del comportamiento de los errores y la combinación lineal, tiene un comportamiento distribuido normalmente en el que se pueden encontrar valores mayores a uno o menores a cero, lejos de ser probabilidades de incumplimiento.

Este modelo puede ser comprendido con mayor claridad mediante un ejemplo como el que se presenta a continuación, evaluando características de individuos con el fin de clasificar su estatus

para saber si se trata de una persona que fuma o no (estatus de Tabaquismo)[14]. Dicha información ha sido calculada con pruebas en el ámbito de la psicología y es de utilidad únicamente para dar a entender una correcta aplicación de la regresión lineal:

- Id del sujeto: Código de identificación
- Tabaquismo - Va. dependiente
  - No fumador = 0
  - Fumador = 1
- Responsabilidad: Nivel de minuciosidad
  - Números enteros
- Extroversión: Nivel de extroversión
  - Números enteros

La muestra es de tamaño $n = 1200$ y la evaluación fue hecha con un análisis de regresión lineal en el paquete estadístico SPSS, donde la primera tabla representa estadísticas descriptivas de las variables: Medias, desviación estándar y número de casos de la muestra con esta información.

Descriptive Statistics

	Mean	Std. Deviation	N
Tabaquismo	.36	.481	1200
Responsabilidad (Minuciosidad)	48.2433	4.71886	1200
Extroversión	53.7958	5.25298	1200

Más tarde se muestra el resumen de la recta ajustada con la regresión, con el coeficiente de correlación R para conocer la proporción de variabilidad de la variable dependiente Y explicada por las variables independientes x_j, en este caso Responsabilidad y Extroversión son las variables que otorgan en un 81.3% el pronóstico (explicación) del estatus de Tabaquismo.

También se muestra el valor de R^2 y R corregido por el número de casos y variables, ambos difieren de manera importante con el coeficiente R al tener un valor de $.660$, sin embargo, el error estándar de estimación es adecuado para hacer referencia a un buen ajuste. Dicho error se refiere

[14] Hugo. H. Salinas. Departamento de Matemáticas, Facultad de Ingeniería. Universidad de Atacama. Chile 2012. [Recurso en línea]

a la desviación de los residuos obtenidos de la variable *Y* y sus valores estimados $\hat{Y}$, es decir, es una medida de variabilidad de la variable dependiente.

Model Summary

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate
1	.813[a]	.660	.660	.281

a. Predictors: (Constant), Extroversión, Responsabilidad (Minuciosidad)

Adicionalmente se presenta el *ANOVA (Análisis de Varianza)* que la separa en suma de cuadrados global (diferencia entre *Y* y valor medio, con *n-1* grados), explicada (diferencia entre *Y* y valor medio, con *k* grados) y residual (diferencia entre *Y* y valor estimado, con *n-[k+1]* grados). Se evalúa con el estadístico *F* con *(k, n-[k+1])* grados de libertad, valora la significancia en la relación lineal y contrasta la hipótesis nula que supone al vector de coeficientes igual a cero $H_0: c_1 = \cdots = c_k = 0$ y la hipótesis alternativa de que algún elemento del vector de coeficientes sea distinto de cero. Como la prueba generó *sig=.000,* se rechaza la hipótesis, confirmando que las variables independientes influyen realmente en la variable dependiente.

ANOVA[b]

Model		Sum of Squares	df	Mean Square	F	Sig.
1	Regression	183.488	2	91.744	1163.684	.000[a]
	Residual	94.371	1197	.079		
	Total	277.859	1199			

a. Predictors: (Constant), Extroversión, Responsabilidad (Minuciosidad)

b. Dependent Variable: Tabaquismo

El siguiente resultado representa en primer lugar los *Coeficientes no Estandarizados* para definir las puntuaciones directas, con lo cual se genera la ecuación del modelo:

$$Y = -3.757 + .004\ responsabilidad + .073\ extroversión$$

Dicha estructura permite ver los pronósticos y mostrar la desventaja del scoring en términos de resultados de probabilidad, ya que los valores pronosticados salen del intervalo *(0,1)*.

Id	Responsabilidad	Extroversion	Tabaquismo (Y)	Predicción (Y estimada)
1	42.00	51.00	0	0.13 = 0
2	48.00	57.00	1	0.6 = 1
3	36.00	48.00	0	**-0.11 = 0**
4	47.00	53.00	0	0.3 = 0
5	47.00	50.00	0	0.08 = 0
6	54.00	44.00	0	**-0.33 = 0**
7	45.00	47.00	0	**-0.15 = 0**
8	52.00	55.00	0	0.47 = 0
9	45.00	52.00	0	0.22 = 0
10	54.00	56.00	1	0.55 = 1

Otro aspecto que se ve presente es el de los C*oeficientes Estandarizados* que dan los valores típicos de los coeficientes para valorar la importancia de cada variable en el modelo individualmente, sin tomar en cuenta las otras variables. En este ejemplo ambas tienen la misma importancia *(.002)* para conseguir un pronóstico del estatus de Tabaquismo.

Adicionalmente se observa el *Estadístico t* y sus significancias para contrastar la hipótesis nula de que los coeficientes de regresión valgan cero en toda la población, para obtener sus valores basta con dividir cada coeficiente entre su desviación estándar y compararlo con una distribución *t* de student con *n-2* grados de libertad. Se puede ver que las dos variables son significativas, recordando que el valor óptimo de significancia es menor a .05 y confirmando la importancia de su presencia en el modelo.

Coefficients[a]

Model		Unstandardized Coefficients		Standardized Coefficients			95.0% Confidence Interval for B	
		B	Std. Error	Beta	t	Sig.	Lower Bound	Upper Bound
1	(Constant)	-3.757	.099		-37.990	.000	-3.951	-3.563
	Responsabilidad (Minuciosidad)	.004	.002	.042	2.241	.025	.001	.008
	Extroversión	.073	.002	.794	42.852	.000	.069	.076

a. Dependent Variable: Tabaquismo

Los intervalos de confianza son parte final de esta tabla, son los límites mínimos y máximos cuya probabilidad (nivel de confianza) da la cobertura asociada a la localización de los coeficientes. Entre más cortos sean los rangos la estimación será más precisa.

2.2.2. Análisis discriminante

Es un modelo estadístico multivariado para la clasificación de *n* elementos de una muestra en *m* grupos. Su desarrollo está en función de diversos contextos como el de distancias, inferencia por máxima verosimilitud, enfoque bayesiano con probabilidad a priori, entre otros. En esta ocasión se hará énfasis en el enfoque clásico *(Discriminante de Fisher),* que desde el punto de vista geométrico representa los valores de variables dependientes *X* mediante nubes de puntos capaces de ser divididas en grupos designados a la variable independiente, lo anterior a través de la identificación de un área en común llamada regla de discriminación.

En otras palabras, es una partición del espacio en regiones disjuntas y se obtiene con las diferencias de los valores medios (*centroides*) por la proyección de cada grupo con distribución normal. Su planteamiento se conforma por combinaciones lineales no correlacionadas para evaluar a cada elemento *i*. Cada combinación representa una función donde los elementos pueden o no estar presentes, clasificándolos en un grupo con la variable dependiente $d^{(g)}$, lo anterior apoyado por las k variables independientes en una matriz X de orden $n \times k$ con $k < n$ (n corresponde al tamaño de la muestra), así como a vectores de coeficientes $C^{(g)}$ desconocidos.

$$d_i^{(g)} = {c_0}^{(g)} + {c_1}^{(g)} x_{1i} + {c_2}^{(g)} x_{2i} + \cdots + {c_k}^{(g)} x_{ki}$$

El número de combinaciones lineales es el mínimo entre el número k de variables a evaluar y el número m de grupos menos uno. La perspectiva matricial en cada grupo es la misma que en el modelo de regresión lineal y sus supuestos básicos consisten en:

- Linealidad, homocedasticidad e independencia – Mismos que en regresión lineal.
- Variable dependiente categórica: La variable $d_i^{(g)}$ toma valores de 0 ó 1 para la clasificación, esto surge de una regla creada con la combinación lineal, destacando con valor de 1 aquél caso en el que un elemento *i* es clasificado en el grupo *g* y dando valor de 0 al resto de las combinaciones.
- Variables independientes numéricas: Las variables independientes de la matriz *X* podrán ser aleatorias pero únicamente cuantitativas (numéricas).

Desarrollo y solución

El desarrollo de este modelo busca que cada una de las combinaciones lineales puedan contar con coeficientes para diferenciar los grupos, es por ello que se busca maximizar la varianza entre cada grupo de manera externa *(Entre-grupos)* y minimizarla dentro de ellos (*Intra-grupos*), para poder homogeneizar lo más posible. La maximización se inicia con la covarianza de la matriz *X* dada cada variable x_j *y* $x_{j'}$ como:

$$cov(x_j\ , x_{j'}) = \frac{1}{n}\sum_{i=1}^{n}(x_{i,j} - \bar{x}_j)(x_{i,j'} - \bar{x}_{j'}).$$

Si se toma el valor medio de cada variable para cada g grupo $\bar{x}_{g,j} = \frac{1}{n_g}\sum_{i\in g} x_{i,j}$ se tendrá que $n_g\bar{x}_{g,j} = \sum_{i\in g} x_{i,j}$ lo cual es útil para sustituir:

$$\bar{x}_j = \frac{1}{n}\sum_{i=1}^{n} x_{i,j} = \frac{1}{n}\sum_{g=1}^{m}\sum_{i\in g} x_{i,j} = \frac{1}{n}\sum_{g=1}^{m} n_g\bar{x}_{g,j} = \sum_{g=1}^{m}\frac{n_g}{n}\bar{x}_{g,j}$$

Quedando la covarianza como $cov(x_j, x_{j'}) = \frac{1}{n}\sum_{g=1}^{m}\sum_{i\in g}(x_{i,j} - \bar{x}_j)(x_{i,j'} - \bar{x}_{j'})$ para transformar con el valor positivo y negativo de $\bar{x}_{g,j}$ sin afectar su valor:

$$cov(x_j, x_{j'}) = \frac{1}{n}\sum_{g=1}^{m}\sum_{i\in g}(x_{i,j} - \bar{x}_{g,j})(x_{i,j'} - \bar{x}_{g,j'}) + \sum_{g=1}^{m}\frac{n_g}{n}(\bar{x}_{g,j} - \bar{x}_j)(\bar{x}_{g,j'} - \bar{x}_{j'})$$

Dichos valores son la suma de covarianza dentro *(Intra)* y fuera *(Entre)* de los grupos*:*

$$cov(x_j\ ,\ x_{j'}) =\ cov_{intra}(x_j\ ,\ x_{j'}) + cov_{entre}(x_j\ ,\ x_{j'}).$$

Es en este momento cuando se busca, análogo a la regresión, estimar los coeficientes $C^{(g)}$ de las combinaciones lineales para cada grupo $d^{(g)}$, que en el enfoque clásico de Fisher se realiza con análisis vectorial al maximizar la variabilidad entre-grupos. Se toma la covarianza como:

$$C'cov(x_j\ ,\ x_{j'})\ C =\ C'\ cov_{intra}(x_j\ ,\ x_{j'})\ C + C'\ cov_{entre}(x_j\ ,\ x_{j'})\ C.$$

Se usa la función $f(C) = \frac{C'\, cov_{entre}(x_j, x_{j\prime})\, C}{C' cov(x_j, x_{j\prime})\, C}$ para ser maximizada con un método de optimización sujeto a restricciones como el de multiplicadores de *Lagrange*. Bajo este esquema, suponemos que la covarianza total es igual a uno, quedando la función *L* para maximización:

$$L = C'\, cov_{entre}(x_j\ ,\ x_{j\prime})\, C - \lambda\, (\ C' cov(x_j\ ,\ x_{j\prime})\, C\ \ - 1).$$

Calculando su derivada e igualando a cero, la ecuación en términos de los vectores de *C* es:

$$\frac{\partial L}{\partial a} = -2\, cov_{entre}(x_j, x_{j\prime})\, C - 2\lambda\, cov(x_j, x_{j\prime})\, C = 0$$

$$\Rightarrow \quad cov_{entre}(x_j, x_{j\prime})\, C = \lambda\, cov(x_j, x_{j\prime})\, C$$

$$\Rightarrow \ \left(cov(x_j, x_{j\prime})^{-1}\, cov_{entre}(x_j, x_{j\prime}) \right) C = \lambda\ C$$

En otras palabras, al contar con la matriz cuadrada $A = cov(x_j, x_{j\prime})^{-1}\, cov_{entre}(x_j, x_{j\prime})$, se realiza el cálculo de los valores reales λ *(Valores Propios o Característicos)* que podrán ser asociados a vectores *(Eigenvectores o Vectores Propios) C* diferentes de cero que cumplan con la solución de la ecuación $AC = \lambda C$ o su equivalente $(A - \lambda I)C = 0$, donde *I* es la matriz identidad, dando como resultado que sus vectores sean los coeficientes de discriminación estimados para cada grupo ($\hat{C}^{(g)}$), mismos que serán vectores invariantes bajo la multiplicación de la matriz *A*, ordenados de mayor a menor que coincide con el poder de discriminación.

Pruebas de ajuste y medición

El ajuste se presenta al contrastar la hipótesis nula de que los centroides son iguales en ambos grupos, con el fin de identificar el poder de discriminación. Se realiza con el estadístico *Lambda de Wilks*: Una proporción de variabilidad por diferencia intra-grupos.

Entre más homogéneos sean los elementos dentro de los grupos, el valor del cociente será más cercano a cero, asumiendo que se tiene una buena discriminación entre-grupos.

$$\lambda = \frac{cov_{intra}\left(x_j{}^{(1)}, x_j{}^{(2)}\right)}{cov\left(x_j{}^{(1)}, x_j{}^{(2)}\right)}$$

Además del estadístico *F de Snedecor,* que es una medida de la variabilidad intra-grupos para evaluar los centroides. Su estimación se logra con un cociente de covarianzas comparado con la distribución F de *(m-1, n-m)* grados de libertad:

$$F = \frac{\left(cov_{intra}\left(x_j{}^{(1)}, x_j{}^{(2)}\right) - cov\left(x_j{}^{(1)}, x_j{}^{(2)}\right)\right)(n-m)}{cov_{intra}\left(x_j{}^{(1)}, x_j{}^{(2)}\right)(m-1)} \sim F_{m-1,n-m}$$

Con esto se comprueba la significancia de cada variable en el modelo, recordando que se espera un valor menor a .05 para no rechazar la hipótesis nula de igualdad en los centroides.

Limitantes y aplicación en Credit Scoring

El credit scoring logra mediante esta herramienta crear una función con un punto de corte para la clasificación. Para su mejor comprensión se presenta el ejemplo anterior del estatus de Tabaquismo, reiterando que su uso es sólo para ejemplificar la correcta aplicación del modelo:

- Id del sujeto: Código de identificación
- Tabaquismo: Estatus de tabaquismo (Va. Dependiente)
  - No fumador = 0
  - Fumador = 1
- Responsabilidad: Nivel de responsabilidad o minuciosidad
  - Números enteros
- Extroversión: Nivel de extroversión
  - Números enteros
- Control: Escala de control personal
  - Números reales
- Or_Espacial: Escala de orientación espacial
  - Números enteros
- Raz_Verbal: Nivel de razonamiento verbal
  - Números enteros

La muestra se mantiene de tamaño *n = 1200* y la evaluación se realiza con un análisis de clasificación discriminante en el paquete estadístico SPSS, el cual presenta en primer lugar la tabla con los valores de las medias, la desviación estándar y el número de casos en cada grupo:

Group Statistics

Tabaquismo		Mean	Std. Deviation	Valid N (listwise) Unweighted	Valid N (listwise) Weighted
No fumador	Responsabilidad (Minuciosidad)	46.9109	4.04079	763	763.000
	Extroversión	50.5701	3.26739	763	763.000
	Escala de Control	21.6173	3.96147	763	763.000
	Orientación Espacial	53.0156	35.52772	763	763.000
	Razonamiento Verbal	31.3224	4.55193	763	763.000
Fumador	Responsabilidad (Minuciosidad)	50.5698	4.91607	437	437.000
	Extroversión	59.4279	2.68746	437	437.000
	Escala de Control	22.9108	5.45993	437	437.000
	Orientación Espacial	45.8518	36.70941	437	437.000
	Razonamiento Verbal	31.9245	4.22870	437	437.000
Total	Responsabilidad (Minuciosidad)	48.2433	4.71886	1200	1200.000
	Extroversión	53.7958	5.25298	1200	1200.000
	Escala de Control	22.0883	4.60451	1200	1200.000
	Orientación Espacial	50.4068	36.11232	1200	1200.000
	Razonamiento Verbal	31.5417	4.44463	1200	1200.000

Aquí se puede ver en un principio la diferencia en las medias de cada variable para ambos grupos. La variable de Razonamiento verbal es la que cuenta con mayor semejanza:

Tests of Equality of Group Means

	Wilks' Lambda	F	df1	df2	Sig.
Responsabilidad (Minuciosidad)	.861	193.934	1	1198	.000
Extroversión	.341	2314.574	1	1198	.000
Escala de Control	.982	22.316	1	1198	.000
Orientación Espacial	.991	11.026	1	1198	.001
Razonamiento Verbal	.996	5.116	1	1198	.024

El estadístico *Lambda de Wilks*, que busca un valor cercano a cero y el *Estadístico F* buscan contrastar la igualdad de medias en los grupos. En este caso, aún cuando los valores de Lambda no son cercanos a cero, se revisan los valores y las pruebas de F para ver que la misma variable (Razonamiento verbal) es aquella con menor poder de discriminación, pudiendo confirmar la significancia de todas las variables con valores menores a .05.

También está presente la matríz de correlaciones dentro de los grupos, con valores muy débiles:

Pooled Within-Groups Matrices

		Responsabilidad (Minuciosidad)	Extroversión	Escala de Control	Orientación Espacial	Razonamiento Verbal
Correlation	Responsabilidad (Minuciosidad)	1.000	.211	.244	.083	.025
	Extroversión	.211	1.000	.091	-.063	.066
	Escala de Control	.244	.091	1.000	.149	-.100
	Orientación Espacial	.083	-.063	.149	1.000	-.229
	Razonamiento Verbal	.025	.066	-.100	-.229	1.000

Otra evaluación del método es el cálculo de la *Correlación Canónica*, una correlación entre la combinación lineal de las variables independientes (función discriminante) y la combinación de variables dependientes para la pertenencia de grupos (como son dos grupos, es una sola variable indicadora para agrupar en fumadores y no fumadores). Cuando la correlación canónica es cercana a 1 como esta, las variables tienen un buen poder de discriminación.

Eigenvalues

Function	Eigenvalue	% of Variance	Cumulative %	Canonical Correlation
1	1.946[a]	100.0	100.0	.813

a. First 1 canonical discriminant functions were used in the analysis.

La clasificación es el elemento final. Se asignan los elementos en cada grupo con ayuda de los coeficientes de cada combinación creada por el punto de corte de la función Fisher:

Classification Function Coefficients

	Tabaquismo	
	No fumador	Fumador
Responsabilidad (Minuciosidad)	1.530	1.588
Extroversión	4.765	5.690
Escala de Control	.436	.428
Orientación Espacial	.086	.085
Razonamiento Verbal	1.542	1.526
(Constant)	-187.992	-241.427

Fisher's linear discriminant functions

Los casos nuevos se colocan en el grupo que les de mayor valor como se ve en la muestra:

Id	Responsabilidad	Extroversion	Control	Or_Espacial	Raz_Verbal	Tabaquismo	*Función No Fumadores*	*Función Fumadores*	Predicción Status (Valor mayor)
1	42.00	51.00	23.00	30.51	28.00	0	*175*	*171*	0
2	48.00	57.00	24.00	37.83	32.00	1	*220*	*221*	1
3	36.00	48.00	15.00	46.38	19.00	0	*136*	*128*	0
4	47.00	53.00	22.00	106.03	29.00	0	*200*	*197*	0
5	47.00	50.00	21.00	61.17	33.00	0	*187*	*182*	0
6	54.00	44.00	24.00	57.23	31.00	0	*167*	*157*	0
7	45.00	47.00	20.00	57.48	20.00	0	*149*	*141*	0
8	52.00	55.00	27.00	112.84	32.00	0	*224*	*224*	0
9	45.00	52.00	19.00	8.17	37.00	0	*195*	*191*	0
10	54.00	56.00	15.00	26.46	37.00	1	*227*	*228*	1
11	47.00	52.00	24.00	44.79	33.00	0	*197*	*194*	0
12	49.00	54.00	27.00	60.86	26.00	0	*201*	*200*	0
13	45.00	54.00	19.00	18.53	35.00	0	*202*	*200*	0
14	47.00	50.00	22.00	27.58	36.00	0	*190*	*184*	0
15	42.00	47.00	23.00	111.97	21.00	0	*152*	*144*	0
16	54.00	58.00	29.00	8.18	30.00	1	*231*	*233*	1
17	49.00	59.00	22.00	4.18	37.00	1	*235*	*238*	1
18	52.00	57.00	22.00	78.06	27.00	1	*221*	*223*	1
19	47.00	48.00	21.00	109.30	26.00	0	*171*	*164*	0
20	38.00	50.00	18.00	41.19	29.00	0	*165*	*159*	0
21	46.00	57.00	21.00	7.66	36.00	1	*219*	*221*	1
22	45.00	53.00	24.00	19.75	28.00	0	*189*	*186*	0
23	47.00	52.00	19.00	67.42	34.00	0	*198*	*195*	0
24	53.00	59.00	33.00	46.76	35.00	1	*247*	*250*	1
25	42.00	53.00	23.00	12.20	34.00	0	*192*	*190*	0
26	49.00	51.00	28.00	93.09	27.00	0	*192*	*188*	0
27	47.00	49.00	25.00	36.12	35.00	0	*185*	*179*	0

Terminando con la validación de tablas cruzadas para comprobarlo. Aquí, el 100% de los casos de estatus de tabaquismo dado por no fumadores (Y = 0) fueron correctamente clasificados, igual que aquellos con el estatus de fumadores (Y = 1), aunque no en su totalidad (más del 95%).

Tabaquismo * Predicted Group for Analysis 1 Crosstabulation

Count

		Predicted Group for Analysis 1		
		No fumador	Fumador	Total
Tabaquismo	No fumador	763	0	763
	Fumador	8	429	437
Total		771	429	1200

Con esto se puede decir que es un método útil para la clasificación pero no es óptimo para el credit scoring que se busca en este estudio por no contar con un resultado inmediato de probabilidad de incumplimiento y por no ser adecuado al tipo de variables de la muestra, ya que el análisis discriminante requiere de supuestos de normalidad y no acepta variables cualitativas como las que suelen representar a las características de la población rural.

2.2.3. Árboles de decisión

Es un modelo jerárquico origen de la idea de creación de diversos modelos no estadísticos y atribuido a Breiman y Friedman en 1973 pero utilizado para credit scoring hasta 1985 por Makowski y Coffman. Se trata de una técnica en la que predominan las formas gráficas para representar variables asociadas a la ocurrencia de eventos y sus respectivos resultados, que pueden o no estar basadas en lógica de probabilidad, los cuales se dividen o segmentan en forma de combinaciones para ir generando dos o más grupos *m* de clasificación o resultados de posibles decisiones. Se construye con la categorización de las variables ya sea cuantitativas para árboles de regresión, o cualitativas para árboles de clasificación.

Sus principales aplicaciones existen en la investigación de mercados (para identificar perfiles de clientes), evaluación de créditos (describir grupos de alto riesgo) y medicina (para resultados de los tratamientos), todo esto debido a su forma intuitiva y fácil de interpretación.

El problema se define con variables $X = (x_1, x_2, \dots, x_k)$ asociadas a los eventos y la segmentación de cada uno de sus posibles resultados x_{jf}, con sus respectivas probabilidades de ocurrencia, donde $j = 1, 2 \dots k$ y f son los elementos del espacio muestral de cada variable. Dicha segmentación provoca el uso de la probabilidad condicional que calcula la relación entre dos variables *(A,B)* hallando la distribución condicional de una de ellas *(A)*, dado el valor de la otra *(B),* para ir segmentando los resultados de nuevas variables en función de las previas.

$$P(A|B) = \frac{P(A \cap B)}{P(B)}$$

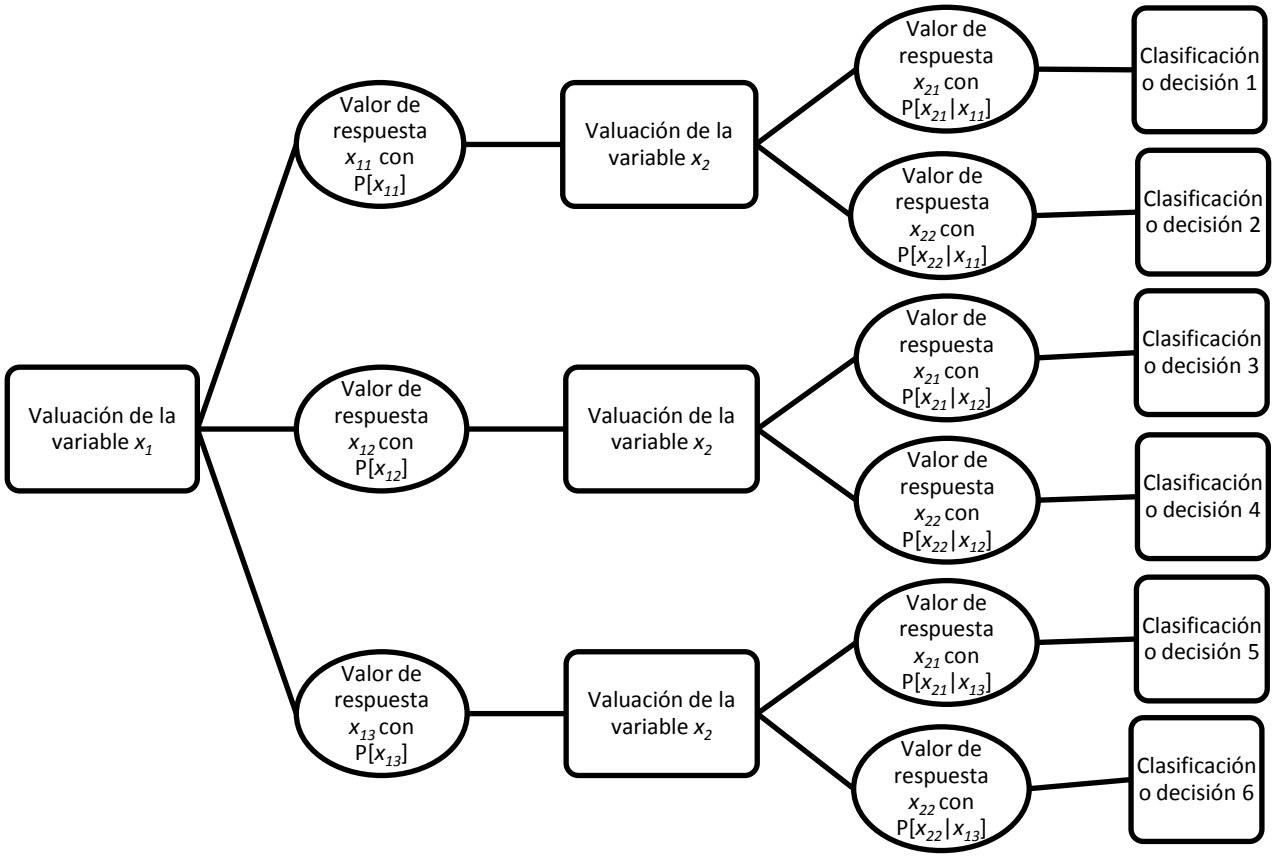
Sus aspectos básicos radican en:

- No requiere del supuesto de relación lineal.
- No es paramétrico, ni inferencial.
- Utiliza sucesos o eventos condicionados.

La representación se forma de izquierda a derecha o de arriba hacia abajo, llamando nodo a cada variable y sus resultados, existe un *nodo raíz* en el origen el árbol y *nodos intermedios* para el

resto de las evaluaciones, todo unido a través de lo que se conoce como *ramas*, que son los caminos para las combinaciones. Los nodos terminales son llamados *hojas* y contienen los resultados de decisión o clasificación final, permitiendo encontrar combinaciones óptimas. La forma de estos nodos es circular cuando se trata de eventos probabilísticos y cuadrada cuando no.

Gráficamente es visto del siguiente modo, pudiendo existir variaciones:



Desarrollo y evaluación del modelo

Para la construcción existen diversos algoritmos, que se diferencian por sus criterios de segmentación, su forma de agregar o eliminar variables (conocido como *poda* del árbol) y el tratamiento a los valores faltantes de la muestra. Algunos utilizan técnicas para la evaluación de la significancia durante la construcción y otros lo hacen después, algunos admiten sólo variables dicotómicas y otros variables politómicas.

Los más referidos en los últimos años han sido CART (Classification Regression Tree), CHAID (Chi-square automatic interaction delection) y QUEST (Quick, Unbiased, Efficient, Statistical Trees)[15], aunque comúnmente se usan también ID3, C4.5 y CIS. Dichos algoritmos realizan análisis para variables con ayuda de la media, la desviación estándar, los valores mínimos y máximos, estimando puntos para la categorización en los cuales la varianza sea mínima; más tarde se calculan las tablas de contingencia, como a las que se recurre directamente con las variables categóricas, con el fin de observar el número de casos clasificados con la segmentación.

Además utilizan medidas y criterios de comparación para evaluar los atributos que caracterizan los nodos o subdivisiones del árbol y son principalmente:

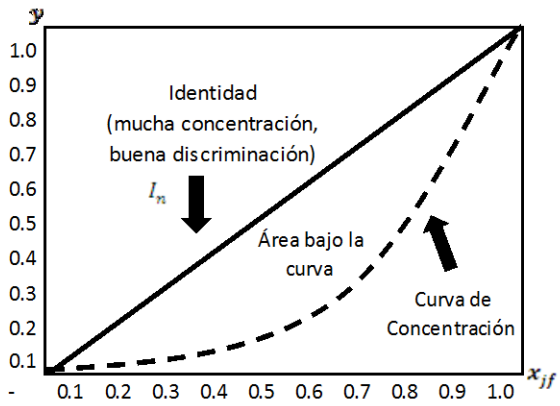
- Índice de Gini

Calculado para medir concentración en las distribuciones. Si todos los elementos de un grupo (como en credit scoring todos los clientes incumplidos) pertenecen a una de las x_{jf} segmentaciones de la clasificación *S* dada a la variable X_j (por ejemplo para la variable edad, una segmentación de mayores de 40 años en la que caen todos los casos incumplidos), entonces la variable y su respectivo nodo serán *puros*, es decir, la clasificación óptima.

De acuerdo al estudio, se busca conseguir aquella clasificación *S* con el índice de Gini $G(t) = 1 - \sum^{m} P\left(\boldsymbol{x_{jf}}\right)^2$ que tenga la mejor pureza, la cual corresponde al mayor valor de $\phi(\mathbf{S}) = G(t-1) - \sum P(t)G(t)$. Donde $G(t-1)$ es el índice del nodo anterior previamente segmentado, $G(t)$ son los índices según la clasificación a evaluar y $P(t)$ la proporción de casos incumplidos resultante.

Cabe mencionar que esta medida, usualmente para estudios de la desigualdad de ingresos en Economía, puede comparar la concentración de elementos de cada grupo con la curva dada por la concentración (proporción en cada segmentación x_{jf} de la variable x_j junto con los valores y de la clasificación) y con la recta de la función identidad I_n.

[15] Vieira Braga, Luis Paulo; Ortiz Valencia, Luis Iván; Ramirez Carvajal, Santiago Segundo. *Introducción a la Minería de Datos.* Río de Janeiro. E-papers: 2009.



- Coeficiente Ji-Cuadrado

Compara la frecuencia de la variable de clasificación bajo diversas segmentaciones estimadas n y observadas n' para cada i cliente de una variable de evaluación X_j, mediante una tabla cruzada y con el fin de conocer la asociación entre ellas, suponiendo independencia.

$$\mathcal{X}^2 = \sum \frac{\left(n_{ij} - n_{ij}{}'\right)^2}{n_{ij}{}'}$$

Entre menor sea la relación, la información que otorga esa segmentación de la variable X_j será menos significativa: Valores cercanos a cero suponen que hay poca asociación y la información de la variable no aportará mucho para clasificar correctamente.

- Índice de Impureza (Algoritmo ID3)

Se obtiene generando combinaciones y colocando las variables en distinto orden para identificar aquella X_j con la segmentación S y valores de respuesta x_{jf} que maximicen la información, es decir, realiza la segmentación más conveniente para clasificar en m grupos. En un árbol binario, como el requerido en del credit scoring, el índice $I(S, \boldsymbol{x_j})$ es:

$$I(S, \boldsymbol{x_j}) = H(S) - H(S|\boldsymbol{x_j})$$

Donde $H(S) = -\sum_{w=1}^{m} P(w) \log_2 P(w)$

$$H(S|\boldsymbol{x_j}) = -\sum_{w=1}^{m}\sum_{f} P(\boldsymbol{x_{jf}} \cap \boldsymbol{w}) \log_2 P(w|\boldsymbol{x_{jf}})$$

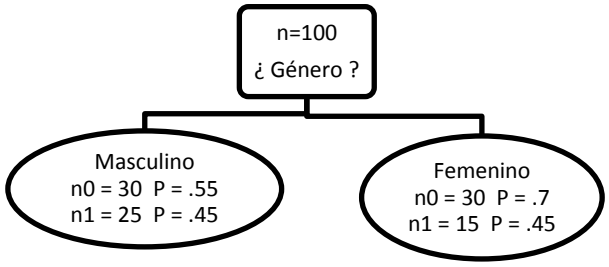
El valor $H(S)$ es conocido como *entropía* y es útil para ponderar subconjuntos de un grupo total. Comúnmente se usa como un promedio de información, para evaluar los símbolos y palabras que realmente aportan información en un texto. Es una medida de incertidumbre que aumenta su valor cuando la información aportada (proporción conseguida con la segmentación *S*) es igual de relevante en todas las variables o grupos.

Limitantes y aplicación en Credit Scoring

En credit scoring tomando los eventos de evaluación como las características de los clientes y sus posibles resultados como las respuestas, se puede identificar en cada nodo la frecuencia de cumplidos e incumplidos *(Y)* hasta lograr una predicción final.

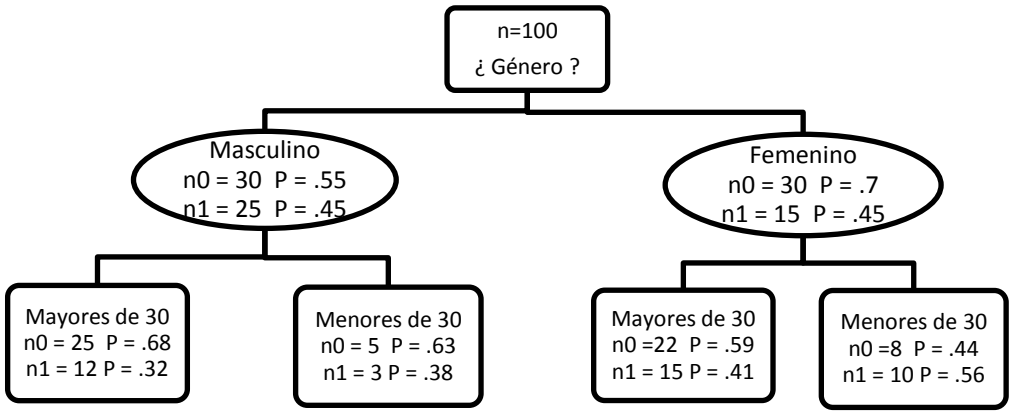
A continuación se tiene un ejemplo que supone una primera característica como género (X_1), una muestra de $n = 100$ individuos y sus x_{1f} respuestas dadas por x_{11} = *femenino* o x_{12} = *masculino.* Para cada respuesta se evalúa la cantidad de cumplidos *(*$n0$*)* e incumplidos *(*$n1$*)*, construyendo una tabla de contingencia:

TABLA DE CONTINGENCIA GÉNERO				X_{11} = FEMENINO			X_{12} = MASCULINO		
				CASOS	PROBABILIDAD		CASOS	PROBABILIDAD	
TOTAL	100	$P(Y)$	1.00	45	$P(X = X_{11})$	0.45	55	$P(X = X_{11})$	0.55
CUMPLIDOS	60	$P(Y = Y_{n0})$	0.60	30	$P(Y = Y_{n0} \mid X = X_{11})$	0.67	30	$P(Y = Y_{n0} \mid X = X_{12})$	0.55
INCUMPLIDOS	40	$P(Y = Y_{n1})$	0.40	15	$P(Y = Y_{n1} \mid X = X_{11})$	0.33	25	$P(Y = Y_{n1} \mid X = X_{12})$	0.45



De aquí se toma otra característica (X_2) y sus respuestas x_{2f} para ser clasificadas de igual forma pero partiendo de la condición dada por la característica anterior. Esto es, si la próxima característica a evaluar es edad, suponiendo que se categoriza en una partición de mayores de 30 años (x_{21}) y menores o iguales (x_{22}), se hará la clasificación de cumplidos e incumplidos para el caso de clientes de género masculino y otra para el de género femenino. Así sucesivamente con todas las características.

TABLA DE CONTINGENCIA GÉNERO ? EDAD				X_{11} = FEMENINO						X_{12} = MASCULINO					
				X_{21} = MAYORES DE 30			X_{22} = MENORES O IGUALES DE 30			X_{21} = MAYORES DE 30			X_{22} = MENORES O IGUALES DE 30		
				CASOS	PROBABILIDAD		CASOS	PROBABILIDAD		CASOS	PROBABILIDAD		CASOS	PROBABILIDAD	
TOTAL	100	$P(Y)$	1.00	37	$P(X_{11}, X_{21})$	0.37	8	$P(X_{11}, X_{22})$	0.08	37	$P(X_{12}, X_{21})$	0.37	18	$P(X_{12}, X_{22})$	0.18
CUMPLIDOS	60	$P(Y=Y_{n0})$	0.60	25	$P(Y_{n0}, X_{21} \mid X_{11})$	0.68	5	$P(Y_{n0}, X_{22} \mid X_{11})$	0.63	22	$P(Y_{n0}, X_{21} \mid X_{12})$	0.59	8	$P(Y_{n0}, X_{22} \mid X_{12})$	0.44
INCUMPLIDOS	40	$P(Y=Y_{n1})$	0.40	12	$P(Y_{n1}, X_{21} \mid X_{11})$	0.32	3	$P(Y_{n1}, X_{22} \mid X_{11})$	0.38	15	$P(Y_{n1}, X_{21} \mid X_{12})$	0.41	10	$P(Y_{n1}, X_{22} \mid X_{12})$	0.56



En conclusión, es un modelo que tiene ventajas por su fácil comprensión, por sus pocos supuestos y porque para este modelo la estadística descriptiva es suficiente. El problema surge con el aumento de variables y de posibles respuestas en cada característica, pues se genera un exceso de combinaciones a evaluar. Sin olvidar, que la clasificación de clientes nuevos y la identificación de probabilidad de incumplimiento es más intuitiva que de predicción, alejándose del propósito inicial.

2.2.4. Análisis Clúster

Son técnicas de elección cualitativa, con fundamentos matemáticos importantes. Es conocido como Análisis de Conglomerados, sobre todo en la Minería de Datos e incluso en otras ciencias como la Biología. Usan variables categóricas o variables numéricas. Para su desarrollo existen medidas de similitud, que comparan semejanzas de los clientes a través de variables categóricas, o medidas de distancia para clasificar variables numéricas.

Las medidas de distancia o funciones de distancia métrica *(d)* entre dos conjuntos *U* de elementos tal que $d: U \times U \rightarrow \mathbb{R}$ se caracterizan por cuatro propiedades para todo *u1, u2 ϵ U*:

- Ser valores positivos $d(u1, u2) > 0$
- Tener valor cero $d(u1, u2) = 0$, si y sólo si $u1 = u2$
- Ser simétricas $d(u1, u2) = d(u2, u1)$
- Cumplir con la desigualdad triangular $d(u1, u2) + d(u2, u3) \geq d(u1, u3) \quad \forall u3 \in U$

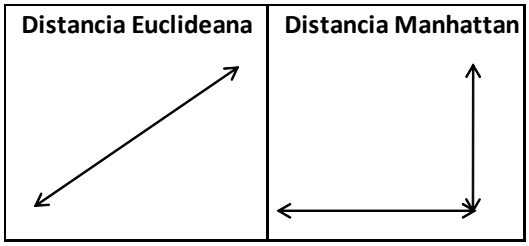
Estas distancias se usan para calcular diferencias y similitudes entre dos de las *k* variables a evaluar, con sus respectivos vectores *(X_a, X_b)* que cuentan con la información y coordenadas de los *n* clientes, identificando así a cada uno de sus *j-ésimos* elementos y pudiendo diferir de acuerdo a las escalas de cada característica. Son distancias que parten de una familia de la forma:

$$d_{a,b} = \left\{ \sum_{j=1}^{k} \left| x_{a,j} - x_{b,j} \right|^r \right\}^{\frac{1}{r}}$$

Esta medida, en su forma más sencilla tiene valor de *r=1,* llamada distancia *Manhattan* por tratarse de distancias recorridas por bloques, tal y como recorrería un automóvil por cuadras. Es en si la suma de diferencias absoluta de los valores:

$$\delta^2_{Manh}(X_a, X_b) = |x_a - x_b|$$

Lo más frecuente es la generalización del Teorema de Pitágoras, cuando *r=2* y se conoce como *Distancia Euclidiana*, puede o no usar raíz cuadrada (*Distancia Euclideana Cuadrada*) pero tiene diversos inconvenientes, por variaciones en las diferencias de escala de las variables:



$$\delta^2_{Euc}(X_a, X_b) = \sqrt{(x_a - x_b)'\,(x_a - x_b)}$$

$$\delta^2_{Euc.Cuad}(X_a, X_b) = (x_a - x_b)'\,(x_a - x_b)$$

Una solución es la corrección por normalización, con la matriz *S* diagonal que incluye las varianzas, es una distancia mejor conocida como *Distancia Euclídea Generalizada*:

$$\delta^2_{Euc.Gen.}(X_a, X_b) = (x_a - x_b)'S\,(x_a - x_b)$$

La distancia de *Mahalanobis* corrige este problema y es invariante a los cambios de escala, ya que pondera a través de la matriz de varianzas y covarianzas cada par de elementos con pesos proporcionales a la inversa de su correlación:

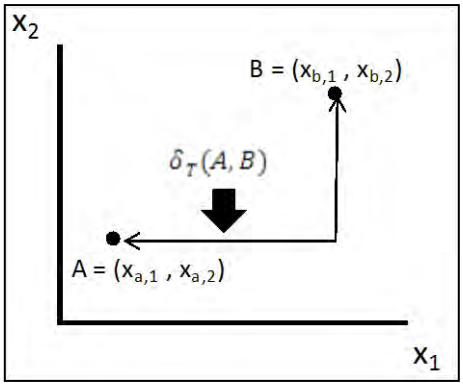
$$\delta^2_{Mahalanobis}(X_a, X_b) = (x_a - x_b)'\sum^{-1}(x_a - x_b) \quad \text{con } i = p, q.$$

También se presentan otros casos como cuando $r = \infty$ con la distancia de *Chebychev*, calculada con la diferencia absoluta máxima de los valores:

$$\delta^{\infty}_{T}(X_a, X_b) = \max\{j = 1, \dots, k\}\,\left|x_{a,j} - x_{b,j}\right|$$

Esto puede verse gráficamente para dos puntos, por ejemplo A y B, mediante la medición de los catetos del triángulo que los conecta, como en la siguiente imagen que calcula la distancia sobre el eje X_1 y X_2, notando que la mayor corresponde al eje X_1 con:

$$\delta_T(A, B) = \left|x_{a,1} - x_{b,1}\right|$$



Por otro lado, para la similitud están las medidas que otorgan el porcentaje de semejanza o coincidencia que existe entre los elementos de la muestra para confirmar a qué grupo pertenecen.

La selección del índice es subjetiva pero coincide en la posibilidad de generar tablas de frecuencia para evaluar los grupos clasificados. Son funciones $s: U \times U \to \mathbb{R}$ donde $\forall\, u1,\ u2\ \epsilon\ U$ y un valor S^* real finito arbitrario, se tiene:

- $s(u1, u2) \leq S^*$
- $s(u1, u1) = S^*$
- $s(u1, u2) = s(u2, u1)$

Para el enfoque de credit scoring, por ser datos dicotómicos, se tienen índices definidos con $N_{c1,\ c2}$ que identifica a *c1* como el subíndice para referir la presencia (1) o ausencia (0) de cierta característica de un cliente, en comparación con el subíndice *c2* que indica la misma información para la evaluación de un segundo cliente. Algunos de estos índices son:

Índice de Coincidencia simple $\frac{N_{00} + N_{11}}{N_{11} + N_{01} + N_{10} + N_{00}}$	Índice de Russell y Rao $\frac{N_{00}}{N_{11} + N_{01} + N_{10} + N_{00}}$
Índice de Jaccard $\frac{N_{11}}{N_{01} + N_{10} + N_{00}}$	Índice de Dice $\frac{2 * N_{11}}{2N_{11} + N_{01} + N_{10}}$

Desarrollo y construcción

Con los índices y/o distancias se forma la matriz de comparación para realizar el análisis que establece la formación de los grupos por medio de dos métodos: Jerárquicos y No Jerárquicos.

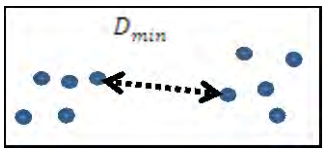
Los primeros métodos van realizando divisiones de las características de los individuos, por cada característica, según sus distancias y asociaciones, con tres criterios básicos a elegir libremente: ligación simple (distancia mínima), ligación completa (distancia máxima) y asociación media. Para esto es de ayuda la matriz, localizan las distancias o similitudes más próximas de todos los elementos y se van incorporando a grupos, estimando la nueva matriz por el criterio seleccionado y así sucesivamente hasta lograr un solo grupo que los identifique.

Sea G_{1i} el primer grupo cuyos elementos i desean ser comparados ($i = 1,2 \dots n_1$) con los del grupo G_{2j} ($j = 1,2 \dots n_2$) se tiene que:

- *Distancia Mínima*

  Es un método de contracción definido con la medición de cada una de las distancias entre los elementos i y los elementos j, seleccionando la menor de ellas (a los más cercanos). Es llamado *Single Linkage y* asegura que la distancia entre los elementos de un grupo más cercanos sea menor que si se incorporan con otros. Construye grupos grandes.

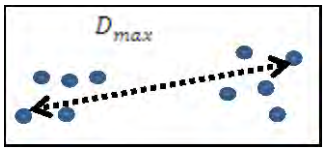
$$D_{min}(G_{1i}, G_{2j}) = min[d(G_{11}, G_{21}), \dots, d(G_{11}, G_{2n_1}), \dots, d(G_{1n_1}, G_{21}), \dots, d(G_{1n_1}, G_{2n_2})]$$



- *Distancia Máxima*

  Es un método de expansión con la medición de cada una de las distancias entre los elementos i y los elementos j, seleccionando la mayor (a los más lejanos). Es llamado *Complete Linkage* y asegura la distancia máxima entre los elementos. Construye grupos pequeños.

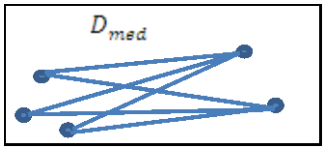
$$D_{max}(G_{1i}, G_{2j}) = max[d(G_{11}, G_{21}), \dots, d(G_{11}, G_{2n_1}), \dots, d(G_{1n_1}, G_{21}), \dots, d(G_{1n_1}, G_{2n_2})]$$



- *Distancia Media*

  Es un método que a diferencia de los demás, no tiene distorsión de expansión o contracción. Compara únicamente a los elementos de un grupo con otro por el valor de la media de todas las distancias entre cada par de elementos i y j. Es llamado *Group Average*. Construye grupos de tamaño mediano en comparación con los otros métodos.

$$D_{med}(G_{1i}, G_{2j}) = \frac{d(G_{11}, G_{21})}{2} + \dots + \frac{d(G_{11}, G_{2n_1})}{2} + \dots + \frac{d(G_{1n_1}, G_{21})}{2} + \dots + \frac{d(G_{1n_1}, G_{2n_2})}{2}$$



A diferencia de lo anterior, los métodos no jerárquicos suponen previamente la clasificación óptima y el número de grupos. Lo que se busca es poder dividir los elementos de la muestra y no a las variables, por lo tanto es útil para cuando las muestras son demasiado grandes.

El más conocido es el procedimiento de *k-Medias*: Se selecciona un número inicial de elementos, lo más lejanos posible, para dividir en grupos iniciales y calcular lo que se conoce como *Centroides* en conglomerados (medias en la acumulación de elementos de un conjunto). Más tarde se localizan los elementos o centroides de otros grupos que estén más próximos para poder clasificarlos o unirlos, tener un menor número de grupos y nuevos centroides, así sucesivamente hasta conseguir agrupar a los *n* elementos en los *m* grupos deseados.

La localización se hace estimando sus distancias, con la media $\bar{X}$ de una *j-ésima* variable en el *l-ésimo* conglomerado (*l = 1, 2,..., k*) y con el valor *f* de respuesta de una característica ($\boldsymbol{x}_{jf}$), asociada a su vez al *a* elemento, como será mostrado a continuación. Esta notación de valores $\boldsymbol{x}_{jf}$

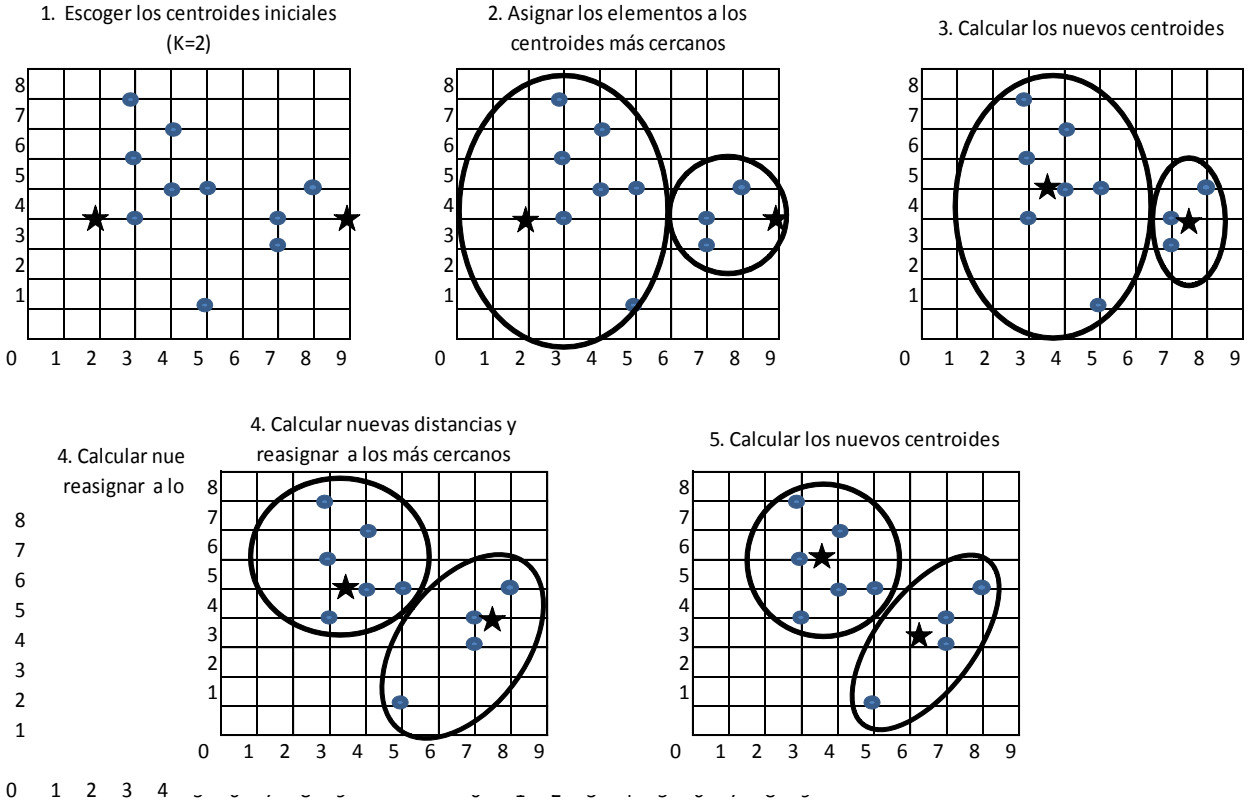
es la misma de los árboles de decisión pero en lugar de buscar una proporción de todos los casos posibles de la muestra que cuenten con determinada característica $\boldsymbol{x}_{jf}$ en cada *f* valor de respuesta, se toma la distancia que se genera de cada *a* elemento al conglomerado *l*.

- Distancia de las características del individuo *a=1, 2,...,n* al conglomerado *l*:

$$d(a,l) = \sqrt{\sum_{j=1}^{k}\left[X_{jf} - \bar{X}(l,j)\right]^2}$$

- Suma de las distancias, representa el error y busca ser minimizado en los conglomerados:

$$E[P(n,m)] = \sum_{j=1}^{m}\sum_{a=1}^{n} d(a,l_j)$$



Limitantes y aplicación en Credit Scoring

En el análisis clúster de credit scoring surge la incertidumbre sobre la elección de la distancia o similitud a criterio de quien crea el modelo. Es una técnica para cuando se busca agrupar pero no permite la identificación adecuada de los factores de riesgo. Además no permite mezclar variables numéricas y categóricas, se deben hacer modelos de agrupaciones por separado para cada tipo de variable y en presencia de grandes tamaños de muestra la cantidad de distancias y la matriz crean dificultades en la capacidad de cálculo.

A continuación se presenta un ejemplo de *K-Medias*, con base en la muestra de análisis discriminante y regresión lineal sobre el estatus de Tabaquismo. Las variables son: Extroversión, Control, Orientación y Razonamiento. El primer resultado de SPSS es un cuadro que representa la ubicación de los centroides (medias) en cada grupo al momento de la clasificación inicial, seleccionados por ser los más distantes.

Initial Cluster Centers

	Cluster	
	1	2
Extroversión	57.00	54.00
Escala de Control	17.00	23.00
Orientación Espacial	2.73	135.93
Razonamiento Verbal	31.00	34.00

Después se evalúa cada elemento de la muestra para asignarlo a los grupos e ir actualizando los centroides de manera iterativa hasta que los cambios sean mínimos.

Iteration History[a]

	Change in Cluster…	
Iteration	1	2
1	25.114	42.732
2	3.466	3.903
3	1.207	1.461
4	.343	.437
5	.195	.252
6	.050	.064
7	.050	.064
8	.000	.000

También se tiene la tabla de *ANOVA*, como en el modelo de regresión lineal. Se presenta la prueba F y se puede identificar que las variables son significativas, junto con la cantidad de elementos clasificados en cada grupo.

ANOVA

	Cluster		Error			
	Mean Square	df	Mean Square	df	F	Sig.
Extroversión	236.138	1	27.420	1198	8.612	.003
Escala de Control	324.955	1	20.948	1198	15.512	.000
Orientación Espacial	1258649.984	1	254.562	1198	4944.367	.000
Razonamiento Verbal	1142.895	1	18.817	1198	60.737	.000

Number of Cases in each Cluster

Cluster	1	676.000
	2	524.000
Valid		1200.000
Missing		.000

Cerrando con la tabla de los centroides de grupos al final de la clasificación:

Final Cluster Centers

	Cluster	
	1	2
Extroversión	54.19	53.29
Escala de Control	21.63	22.68
Orientación Espacial	21.89	87.19
Razonamiento Verbal	32.40	30.43

2.2.5. Modelos no estadísticos

Los modelos no estadísticos son también relevantes hoy en día, muchos de ellos aún en estudio. Surgieron de técnicas estadísticas pero su desarrollo se basa en sistemas tecnológicos, siendo evidente que cuentan con áreas de oportunidad. Aunque en esta ocasión no serán descritos a fondo por no ser parte de los objetivos, serán mencionados únicamente como parte de posibles soluciones para la realización de un score.

Redes neuronales

Iniciadas en 1953 por Nathaniel Rochester con los estudios en 1943 de McCulloch y Pitts, surgieron las redes neuronales que modelan el proceso de decisión con unidades de procesamiento conectadas, en lugar de una base de datos.

Son sistemas para imitar el comportamiento humano del sistema nervioso, con una red de neuronas interconectadas que hacen la función de las variables o características al momento de la solicitud. Funcionan a prueba y error, por medio de la recepción y emisión de señales o estímulos registrados como "entrenamiento", a partir de la experiencia y la capacidad humana para reconocer e identificar patrones.

Las principales redes son llamadas: perceptrón, función de base radial y mapas auto organizables. La red perceptrón es análoga al análisis univariado, es la existencia de una neurona con una sola capa para realizar todas las funciones (entrada, interna y salida). Cuando existen más neuronas es en dichas capas donde se realiza todo el proceso, permitiendo el flujo de elementos a través de *n* nodos de entrada y *m* nodos de salida. Cuentan con funciones de *activación* para las variables regresivas, de *propagación* para las capas intermedias, su variable *(Y)* predictiva y los grupos de clasificación.

Es un modelo para la aplicación en cualquier tipo de decisión pero al no tener claro su comportamiento en la industria crediticia tiene la desventaja de no poder emitir las justificaciones precisas del rechazo, sin embargo, es útil al momento de adaptarse a los cambios en el comportamiento y cuando no se tienen los suficientes datos para una muestra significativa.

Algoritmos genéticos

Construidos con la idea de un procedimiento sistemático de búsqueda que cuente con parámetros (atributos) de la población a optimizar. Propuestos por John Holland en 1975 de la Universidad de Michigan para credit scoring pero analizados por la Biología y sus procesos de selección natural desde años anteriores.

Son algoritmos de búsqueda heurística con un resultado óptimo dentro de un espacio de resultados, pensado en la ley de evolución referida como "la ley del más fuerte". En otras palabras, se genera el conjunto de posibles valores otorgados a los atributos de los clientes, incluso con combinaciones entre ellos, para buscar la solución con selección óptima de variables. A diferencia de los métodos estadísticos que parten de un punto dado hasta llegar a un criterio de solución, en este método hay un conjunto de soluciones para la optimización, aún cuando ésta no sea necesariamente la más adecuada a las situaciones de la vida diaria, siendo esta una de sus mayores desventajas.

No requiere de conocimientos sobre las distribuciones, ni de una función de clasificación pero si sistemas tecnológicos muy avanzados por cantidad de resultados y combinaciones.

Programación lineal

Con ayuda de Freed & Glover (1981) se buscó el planteamiento de las funciones de clasificación como un problema de Programación Lineal, técnica originada de la investigación de operaciones y apoyada por herramientas de programación.

El objetivo principal fue apoyar la toma de decisiones en la asignación de recursos, como en ocurrió en 1930 con los medios de transporte (Kantorovich) y la teoría de juego (Morgenstern y von Neumann), sin embargo la mejora ocurrió en 1947 con George Dantzig, quien desarrolló el método de puntos interiores (simplex) al buscar soluciones logísticas para la fuerza aérea.

Con los desarrollos tecnológicos, el uso de la programación lineal fue avanzando hasta consistir en la maximización o minimización de una función lineal desconocida sujeta a ciertas restricciones (igualdades o desigualdades).

Bajo este contexto, la idea de aplicar el método es minimizar los posibles errores de una regresión, aquellos elementos de la muestra que no pueden ser clasificados correctamente. La programación lineal pretende encontrar valores estimados ($\hat{C}$) para los coeficientes dadas ciertas restricciones:

Minimizar $\sum_{i=1}^{n} \varepsilon^2$ sujeto a $Y = c_0 + c_1X_1 + c_2X_2 + \cdots + c_kX_k + \varepsilon$

Con determinados coeficientes $\hat{C} > 0$

Es apto para cualquier tipo de variables y tiene flexibilidad para las distribuciones. Las desventajas radican, entre otras cosas, en que no arroja resultados en términos de probabilidad, tiene bases complejas de programación y su significancia estadística no puede ser medida para evaluar factores de riesgo, siendo un modelo capaz de aceptar la concesión parcial de créditos en torno a la continuidad de la función objetivo, lo cual en la práctica no es posible.

2.2.6. Benchmark y comparaciones

Al ser cada vez más necesario el poder tomar decisiones correctas sobre las solicitudes de crédito, la comparación de los modelos que buscan ser cada vez más eficientes y objetivos es un aspecto básico. Explicar los modelos resulta vital para el conocimiento de sus requerimientos, ya que no es posible usar cualquier modelo en cualquier caso, para ello sus ventajas y desventajas han sido aclaradas y serán confirmadas.

Los modelos de regresión lineal y análisis discriminante son modelos paramétricos y robustos, con una base matemática y estadística consistente, representan el origen del credits y de otros modelos pero cuentan con limitantes notables en este caso. La regresión lineal se ve limitada ya que cuenta con variables independientes carentes de propiedades estocásticas y requiere de un comportamiento de los errores determinado por la distribución normal con resultados que salen de los posibles valores de la probabilidad de incumplimiento. El análisis discriminante es mucho más útil, sobre todo cuando se tiene una gran cantidad de grupos, para la búsqueda de una combinación lineal que clasifique, sin embargo, tampoco otorga un resultado que pueda ser asociado a la probabilidad de incumplimiento, además de utilizar variables únicamente numéricas.

Adicionalmente, los modelos estadísticos no paramétricos, como los árboles de clasificación o análisis clúster, buscan particiones óptimas de la muestra, no requieren de supuestos de distribución y son visualmente más sencillos, siempre y cuando no se cuente con una gran cantidad de características a evaluar. Las desventajas de los árboles radican en la existencia condicionada de cada nodo y los resultados en función de estrategias. Las desventajas del análisis

clúster son las diferencias por la elección subjetiva de las medidas de distancia y similitud, así como por la gran cantidad de cálculos y el tamaño de matrices, las cuales pueden llegar a ser de un orden lo suficientemente grande como para no ser calculado por cualquier sistema, perdiendo eficiencia. En ambos casos la interpretación de los modelos no otorga una probabilidad de incumplimiento, ni información sobre los factores de riesgo, sólo clasifica los datos conocidos para encontrar el óptimo, convirtiendo la predicción de probabilidades para nuevos clientes en una simple intuición.

Finalmente se presentan los modelos no estadísticos, como redes neuronales, algoritmos genéticos y programación lineal, que son aún más complejos en cuanto a estimaciones que los modelos estadísticos. Son muy útiles cuando la relación de las características del cliente y el resultado no es lineal, sin embargo, sus desventaja son la difícil interpretación e implementación, son poco intuitivos y representan altos costos en tecnología y aprendizaje.

	REGRESIÓN LINEAL	***DISCRIMINANTE LINEAL***	***ÁRBOLES DE DECISIÓN***	***ANÁLISIS CLUSTER***	***MÉTODOS NO ESTADÍSTICOS***
TIPO DE MODELO	Regresión	Combinaciones lineales	Optimización y combinaciones	Agrupación y clasificación	Algoritmos y sistemas
VARIABLES INDEPENDIENTES	Cuantitativas y categóricas	Cuantitativas	Cuantitativas y categóricas	Cuantitativas y categóricas	Cuantitativas y categóricas
VARIABLE DEPENDIENTE	Cuantitativa	Regla para discriminación	Categórica	Categórica	Categórica
	Distribución normal	Valores {0,1}	Combinación óptima	Valores {0,1}	Valores {0,1}
DISTRIBUCIÓN DE ERRORES	Distribución Normal	Distribución Normal	Sin distribución particular	Sin distribución particular	Sin distribución particular
	Homocedasticidad	Homocedasticidad	Heterocedasticidad	Heterocedasticidad	Heterocedasticidad

MODELO DE REGRESIÓN LOGÍSTICA EN EL CREDIT SCORING

3.1. Análisis del modelo de Regresión Logística

3.1.1. Orígenes y aplicaciones del modelo

El modelo logístico ha sido empleado durante varios años en muchas áreas de la ciencia, sobretodo en la Biología, Medicina y en análisis econométricos, usando elementos de Matemáticas Aplicadas, Estadística experimental y Teoría Económica, como algunos autores lo sostienen[16]. Su fuente principal es la función logística, diseñada en 1845 para darle un uso de curva de crecimiento, misma que describía el comportamiento de tamaño de la población con el paso del tiempo. El desarrollo comenzó con la cantidad de elementos de una población en un cierto tiempo, definida como *N(t)* y una tasa o porcentaje de crecimiento con un $\propto$ constante, la cual se entendía en un principio como:

$$N(t)' = \frac{d\,N(t)}{d\,t} \quad \text{Donde} \quad N(t)' = \propto N(t).$$

Identificando la existencia de una ecuación diferencial, con solución en la forma exponencial y base de lo que se conoce como el crecimiento geométrico, es decir $N(t) = N(o)\,exp\,(\propto t)$. Lo anterior, ya que $\frac{d\,e^z}{d\,z} = e^z$ y por lo tanto $N(t)' = \propto N(o)\,exp(\propto t)$.

Sin embargo Pierre-François Verhulst junto con Quetelet, interesados en la estadística de vida, buscaron un factor que disminuía el problema de crecimiento infinito al recordar que la población va teniendo límites de tamaño *M(t),* donde *0 < N(t) < M(t),* ya que el crecimiento poblacional va disminuyendo su rapidez con el paso del tiempo a causa de factores del medio ambiente.

Este procedimiento generó un modelo en el cual se incluyó a la proporción $N(t)' = \propto N(t)$, un factor constante de razón $\frac{M(t)-N(t)}{M(t)}$ que junto con la constante $\propto$, se convierte en una nueva

[16] Hosmer, David W.; Lemeshow, Stanley. *Applied logistic regression.* USA. Wiley & Sons: 2000.

proporción, $\frac{d\,N(t)}{d\,t} = \beta N(t)[M(t) - N(t)]$, estableciendo que el crecimiento es una proporción del producto de la población y la diferencia con el límite.

Si hacemos con esto la separación de variables $\frac{d\,N(t)}{N(t)[M(t)-N(t)]} = \beta dt$ e incluimos la integral para resolver, entonces $\int \frac{1}{N(t)[M(t)-N(t)]} d\,N(t) = \int \beta\, dt$. Para resolver el lado izquierdo de esta ecuación se hace una integración por fracciones parciales, donde se buscan A, B constantes mediante un sistema de ecuaciones tal que:

$$\int \frac{1}{N(t)[M(t) - N(t)]} dN(t) = \frac{A}{N(t)} + \frac{B}{M(t) - N(t)}$$

$$\Rightarrow\ A\,[M(t) - N(t)] + B\,N(t) = 1$$

$$\Rightarrow -A + B = 0$$
$$M(t)\,A = 1$$

Cuya solución es $A = \frac{1}{M}$ y $B = \frac{1}{M}$:

$$\int \frac{1}{N(t)[M(t) - N(t)]} dN(t) = \frac{1}{M(t)} \int \frac{d\,N(t)}{N(t)} + \frac{1}{M(t)} \int \frac{d\,N(t)}{M(t) - N(t)}$$

$$= \frac{1}{M(t)} [ln|N(t)| + ln|M(t) - N(t)|]$$
$$= \frac{1}{M(t)} ln \left| \frac{N(t)}{M(t) - N(t)} \right|$$

Despejando así el valor de cantidad de la población:

$$\frac{1}{M(t)} ln \left| \frac{N(t)}{M(t) - N(t)} \right| = \beta t$$

$$ln \frac{N(t)}{M(t) - N(t)} = M(t)\beta t$$

$$\frac{N(t)}{M(t) - N(t)} = e^{M(t)\beta t}$$

$$N(t) = [M(t) - N(t)]\ e^{M(t)\beta t}$$

$$N(t)e^{M(t)\beta t} + N(t) = M(t)\ e^{M(t)\beta t}$$

$$N(t)\left[\ e^{M(t)\beta t} + 1\right] = M(t)\ e^{M(t)\beta t}$$

$$N(t) = \frac{M(t)\ e^{M(t)\beta t}}{e^{M(t)\beta t} + 1}$$

Lo cual fue designado como la función logística pero no recibió la atención suficiente hasta 1920 que fue analizada de manera experimental por los norteamericanos Raymond Pearl y Lowell Reed, incrementando poco a poco su presencia en otros estudios para conocer tendencias de comportamiento de población.

Por otro lado, la idea de incluir transformaciones para el modelo de regresión lineal surgió de forma bivariada en 1930, con la función de transformación *Probit* y variables continuas. Fue alrededor de 1950 cuando se dio la teoría para la selección de variables discretas o categóricas, considerando valores aleatorios para la utilidad económica, como en 1954 cuando se consideró para la clasificación de dos grupos: Posesión o falta de un automóvil en función del ingreso obtenido en las familias. Surgiendo más tarde la forma multivariada.

La función probit ha sido de gran utilidad por ser una alternativa semejante de la transformación logit, misma que proviene de la función logísitica. Es un tipo de transformación que tiene ventajas por admitir variables categóricas a diferencia de los métodos alternativos anteriores, además de tomar valores entre 0 y 1 para la variable dependiente, lo cual se puede asociar a una probabilidad de incumplimiento. Sin embargo, el supuesto de distribución de los errores asume que son de tipo normal, razón por la cual se ha optado por encontrar otras transformaciones. Ha sido utilizada en toxicología para conocer la tolerancia y proporción de individuos frente a ciertas dosis de sustancias químicas.

Fue en 1970 cuando se generalizó el procedimiento con ayuda de D. R. Cox, para introducir al modelo de regresión la transformación y función logística que han incrementado sus aplicaciones con el paso del tiempo entre las que se encuentran:

- Modelos para predecir el padecimiento de una enfermedad.
- Modelos para predecir la sobrevivencia de pacientes.
- Modelos para conocer la probabilidad de éxito de una empresa/negocio.

3.1.2. Descripción de la función Logística

Como es evidente, referirse al modelo logístico para credit scoring es partir de la regresión que, como un procedimiento para analizar la relación entre variables, cuenta con diversos tipos de estimaciones para explicar variables que representan eventos de la vida diaria.

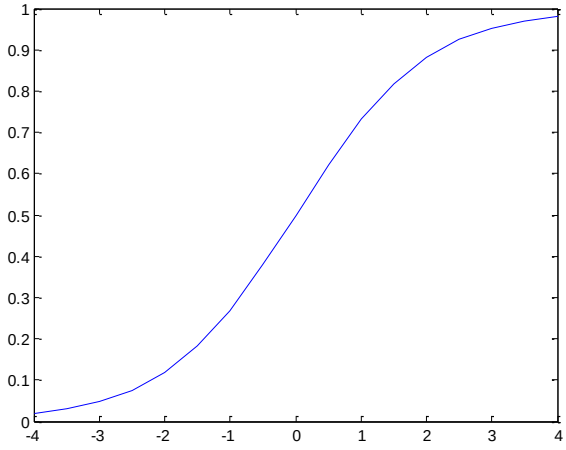
Recordando el tema del capítulo anterior, un tipo de estimación para representar los eventos es por ejemplo la regresión lineal, basadas en la distribución normal y con varianza constante. El análisis discriminante es otra alternativa con variables aleatorias para la predicción de nuevas observaciones y la clasificación de clientes, sin llegar a ser óptima por no limitar el rango de resultado a una interpretación en términos de probabilidad y por mantener el requerimiento de un comportamiento normal de las variables, sin aceptar variable categóricas.

La regresión logística es una solución más adecuada, tiene la posibilidad de incluir variables categóricas y brindar resultados directos de la probabilidad de incumplimiento y los factores de riesgo. Se logra al suponer un límite $M(t) = 1$ a la función. Es una regresión sin supuestos de distribución normal y sin la necesidad de grandes tamaños de muestra como la función probit. Para su descripción se iniciará con el análisis de su función.

Como función logística es una relación entre dos o más variables de forma que a cada elemento *x* del conjunto independiente *X*, le corresponde un único elemento $\pi(x)$ del conjunto imagen $\pi(X)$ y está representada por:

$$\pi(x) = \frac{1}{1+e^{-x}} = \frac{e^x}{e^x}\,\frac{1}{1+e^{-x}} = \frac{e^x}{e^x+1}$$

Su gráfica es una curva *S* o *Sigmoidea,* tiene un único punto de inflexión en el que cambia la concavidad y la rapidez del crecimiento:



Función Logística en Matlab.

Su dominio (valores *x*) se encuentra en el intervalo $(-\ ,\ )$ y su codominio (valores $\pi(x)$) en el intervalo $(0,1)$, de acuerdo al comportamiento de sus límites.

$$\lim_{x\to} \frac{1}{1+e^{-x}} = 1$$

Ya que $\lim_{x\to}\ e^{-x} = 0$, es decir $\forall\varepsilon > 0\ \exists\ \delta > 0 : x > \delta\ \Rightarrow\ |e^{-x}|\ < \varepsilon$

Donde $0 < \delta < x\ \Rightarrow\ 0 < \frac{1}{x} < \frac{1}{\delta}\ \Rightarrow\ 0 < e^{\frac{1}{x}} < e^{\frac{1}{\delta}}$, tomando $\delta = \frac{1}{\ln\varepsilon}$

$$\lim_{x\to -} \frac{1}{1+e^{-x}} = 0$$

Ya que $\lim_{x\to -}\ e^{-x} =$, es decir $\forall\varepsilon > 0\ \exists\ \delta > 0 : x < -\delta\ \Rightarrow\ |e^{-x}| > \varepsilon$

Donde $x < -\delta\ \Rightarrow\ \frac{1}{x} > -\frac{1}{\delta}\ \Rightarrow\ e^{\frac{1}{x}} >\ e^{-\frac{1}{\delta}}$, tomando $\delta = -\frac{1}{\ln\varepsilon}$

Además $\lim_{x\to}\ \frac{1}{x} = 0$, es decir $\forall\varepsilon > 0\ \exists\ \delta > 0 : x >\ \delta\ \Rightarrow\ \left|\frac{1}{x}\right|\ < \varepsilon$

Donde $\delta < x\ \Rightarrow \frac{1}{x} < \frac{1}{\delta}$ tomando $\delta = \frac{1}{\varepsilon}$

Es una función inyectiva, esto es, que a cada valor de *x* le corresponde uno del conjunto $\pi(x)$, siendo demostrado al tomar dos elementos del codominio, cada con su propio elemento $x_1,\ x_2$ tal que si $\pi(x1) = \pi(x2)$ se tiene que:

$$\pi(x1) = \pi(x2) \Rightarrow \frac{1}{1+e^{-x1}} = \frac{1}{1+e^{-x2}} \Rightarrow 1 + e^{-x1} = 1 + e^{-x2} \Rightarrow e^{-x1} = e^{-x2}$$

$$\Rightarrow \ln e^{-x1} = \ln e^{-x2} \Rightarrow -x1 = -x2 \Rightarrow x1 = x2$$

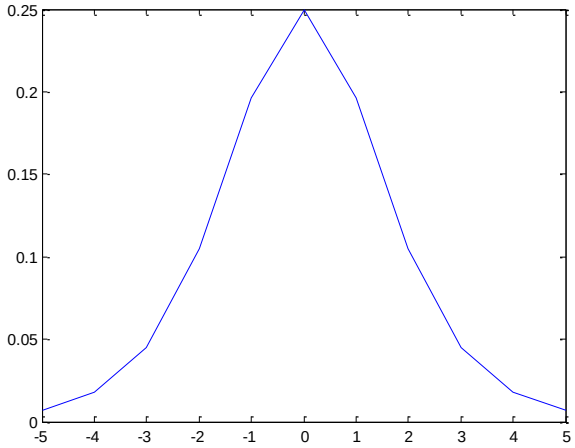
Es también creciente, ya que tomando dos elementos cualesquiera *x1, x2* ϵ *X* tal que *x1 < x2,* encontraremos que $\pi(x1) < \pi(x2)$ *:*

$$x1 < x2 \Rightarrow -x2 < -x1 \Rightarrow e^{-x2} < e^{-x1} \Rightarrow 1 + e^{-x2} < 1 + e^{-x1} \Rightarrow \frac{1}{1+e^{-x1}} < \frac{1}{1+e^{-x2}}$$

Vista como distribución logística en probabilidad es una distribución continua y simétrica, con función de distribución (*F*) y densidad (f) dadas por dos parámetros, μ en los reales y $s > 0$:

$$F(x;\mu,s) = \frac{1}{1+e^{\frac{-(x-\mu)}{s}}} \quad ; \quad f(x;\mu,s) = \frac{e^{\frac{-(x-\mu)}{s}}}{s\left(1+e^{\frac{-(x-\mu)}{s}}\right)^2}$$

Su gráfica es semejante a la una distribución normal con parámetro μ pero con mayor *Curtosis* (mayor concentración de frecuencias alrededor de la media). Tiene una varianza de $\frac{\pi^2}{3}s^2$.



Distribución Logística en Matlab.

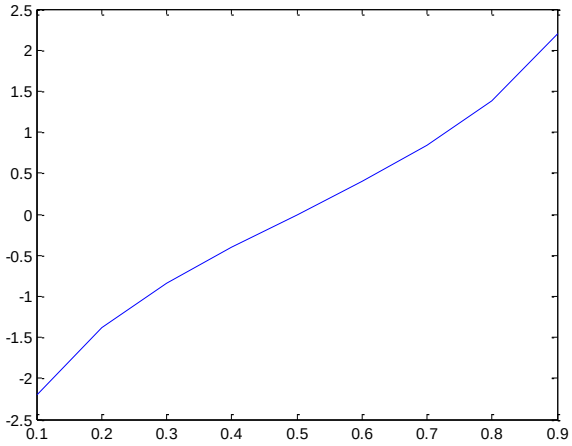
3.1.3. Transformación Logit y construcción del modelo

La función logística cuenta con una función inversa llamada transformación logit, indispensable para la construcción de la regresión. El desarrollo de esta función inversa (π^{-1}), se basa en un despeje de variables:

$$\pi(x) = \frac{1}{1+e^{-x}} \Rightarrow 1+e^{-x} = \frac{1}{\pi(x)} \Rightarrow e^{-x} = \frac{1}{\pi(x)} - 1 \Rightarrow e^{x} = \frac{1}{\frac{1-\pi(x)}{\pi(x)}}$$

$$e^{x} = \frac{\pi(x)}{1-\pi(x)} \quad \Rightarrow \quad x = \ln\frac{\pi(x)}{1-\pi(x)}$$

Su gráfica es representada por una curva de la función logaritmo:



Función Logit en Matlab.

Dicha función tiene codominio (valores *x*) en el intervalo $(-\infty, \infty)$ y dominio (valores $\pi(x)$) en el intervalo $(0,1)$, de acuerdo a la función logaritmo natural, que es una función positiva con base en el número racional *e=2.718…*, teniendo como único caso posible para que $\frac{\pi(x)}{1-\pi(x)} > 0$, que el numerador y el denominador sean positivos.

a. $\pi(x) > 0$

b. $1-\pi(x) > 0$ Donde $1-\pi(x) > 0 \Rightarrow 1 > \pi(x)$

Como en ambos casos, para la función logit y la función logística, el valor de cualquier $\pi(x)$ se encuentra en el intervalo $(0,1)$, se trata de una función suprayectiva.

Tomando dos elementos del codominio con sus respectivos valores $\pi(x1)$ y $\pi(x2)$, tal que si $x1 = x2$, es una función inyectiva, ya que cumple $\pi(x1) = \pi(x2)$. Con esto se confirma que son funciones biyectivas.

$$x1 = x2 \;\Rightarrow\; \ln\frac{\pi(x1)}{1-\pi(x1)} = \ln\frac{\pi(x2)}{1-\pi(x2)} \;\Rightarrow\; \frac{\pi(x1)}{1-\pi(x1)} = \frac{\pi(x2)}{1-\pi(x2)}$$

$$\Rightarrow \frac{1}{\frac{1}{\pi(x1)}-1} = \frac{1}{\frac{1}{\pi(x2)}-1} \Rightarrow \frac{1}{\pi(x2)} - 1 = \frac{1}{\pi(x1)} - 1 \;\Rightarrow\; \frac{1}{\pi(x2)} = \frac{1}{\pi(x1)}$$

$$\therefore \pi(x1) = \pi(x2)$$

Otra característica es que es creciente, donde con dos elementos cualesquiera $\pi(x1)$ y $\pi(x2)$, tal que $\pi(x1) < \pi(x2)$, se tiene que $x1 < x2$ *:*

$$\pi(x1) < \pi(x2) \;\Rightarrow\; \frac{1}{\pi(x2)} < \frac{1}{\pi(x1)} \;\Rightarrow\; \frac{1}{\pi(x2)} - 1 < \frac{1}{\pi(x1)} - 1 \Rightarrow \frac{1}{\frac{1}{\pi(x1)}-1} < \frac{1}{\frac{1}{\pi(x2)}-1}$$

$$\Rightarrow \frac{\pi(x1)}{1-\pi(x1)} < \frac{\pi(x2)}{1-\pi(x2)} \Rightarrow \ln\frac{\pi(x1)}{1-\pi(x1)} < \ln\frac{\pi(x2)}{1-\pi(x2)} \;\Rightarrow\; x < x2$$

Con base en esto surge la regresión que cuenta con un error ε y en la cual se incorpora la variable *Y,* en este caso dicotómica (o indicadora) de valores cero y uno. Dicha variable *Y* da a $\pi(x)$ una interpretación de probabilidad, que dentro de sus características está en el intervalo $(0,1)$. Dicha probabilidad está relacionada con la pertenencia al éxito (*Y=1)* o al fracaso (*Y=0)* de determinado evento, usando una distribución Bernoulli, a diferencia de la regresión lineal (una desviación de la esperanza condicional distribuida normalmente). Esta regresión tiene una ecuación para todo *x* dada por lo siguiente:

$$y = \pi(x) + \varepsilon = P(y|x) + \varepsilon = \frac{1}{1+e^{-x}} + \varepsilon$$

Como se busca sostener que $E[\varepsilon] = 0$, con lo cual el valor de los errores depende de la variable *Y,* se tiene $\varepsilon = 1 - \pi(x)$ cuando se considera un éxito con probabilidad $\pi(x)$ y $\varepsilon = -\pi(x)$ cuando se considera un fracaso con probabilidad $1 - \pi(x)$.

Para calcular la varianza del error se tiene que:

$$V[\varepsilon] = E[\varepsilon^2] - E[\varepsilon]^2 = E[\varepsilon^2]$$

$$\Rightarrow \quad E[\varepsilon^2] = [1 - \pi(x)][-\pi(x)]^2 + \pi(x)[1 - \pi(x)]^2$$

$$= [1 - \pi(x)][\pi(x)]^2 + \pi(x)[1 - \pi(x)]^2$$

$$= \pi(x)[1 - \pi(x)]$$

$$\therefore \; V[\varepsilon] = \pi(x)[1 - \pi(x)]$$

Dadas una o más características a evaluar $x = x_1, x_2, \ldots, x_k$ para cada *i* individuo se forma la regresión, que como un modelo matemático supone un vector en el lugar de la variable *X*, mediante una combinación lineal con *k* variables y *k+1* coeficientes. La cantidad de variables en *X* hará del modelo una regresión bivariada o multivariada.

Enfocando el modelo a credit scoring ambas funciones son fundamentales. Se toma el valor de $\pi(x) = P(Y = i|\, x)$ asociado a la existencia de una probabilidad que clasifica el cumplimiento (con $y_i = 0$) o incumplimiento (con $y_i = 1$) de las obligaciones crediticias de todos los clientes con probabilidad $1 - \pi(x)$ y $\pi(x)$, respectivamente. En particular, en términos se utiliza la regresión multivariada dicotómica para cada cliente:

$$\pi(X) = \frac{1}{1 + e^{-(c_0 + c_1 x_1 + \cdots + c_k x_k)}} = \frac{1}{1 + e^{-(X \cdot C)}}$$

Visto desde la transformación logit como:

$$\ln \frac{\pi(X)}{1 - \pi(X)} = c_0 + c_1 x_1 + c_2 x_2 + \cdots + c_k x_k = X \cdot C$$

3.1.4. Tipo de estudio y uso de variables

Una vez construido el modelo de regresión logística llega el momento de conocer el tipo de estudio que se puede realizar con ayuda del modelo y del manejo de las variables independientes.

Dado que las variables del modelo hacen referencia a las características de los clientes, este modelo es adecuado por aceptar diferentes tipos de variables que las representan correctamente. Cómo ya se mencionó, existen variables *numéricas y categóricas*, sin embargo, cada una de ellas cuenta con sub-clasificaciones. Las variables categóricas (o cualitativas) se dividen en ordinales y nominales, las ordinales son aquellas que guardan un orden (como bueno, regular y malo) y las nominales no mantienen ningún orden (como las que se refieren a la ausencia o presencia de algún factor). Las variables numéricas por el contrario se pueden cuantificar numéricamente y se pueden realizar operaciones como el cálculo de intervalos, su división se da con las variables discretas (números enteros sin valores intermedios) o continúas (con valores intermedios). De acuerdo a esto la regresión logística para credit scoring es dicotómica, ya que sólo existe la posibilidad de evaluar por ser una respuesta nominal (de clientes cumplidos o incumplidos).

Existen así las variables continuas, que en el estudio se tratan como tal y sus diferencias en intervalos son posibles de calcular bajo operaciones algebraicas tradicionales, sin embargo, las variables categóricas o discretas requieren de cambios específicos cuando no son binarias.

Si una j variable categórica toma r valores asociados a distintas respuestas, x_j con valores x_{jf} $\forall\, f = 1, \dots, r$ donde $r > 2$, serán creadas *r-1* variables binarias asociadas a la variable x_j, tal que cada $x_{j(f)}$ tendrá valor de uno cuando su valor de respuesta sea el correspondiente a la característica del cliente o cero si no. Se conocen como variables *Dummy* (también llamadas internas o de diseño) y se tratan de forma asociada, esto es que cuando se trabaje con cualquier $x_{j(f)}$ se tomarán en cuenta los efectos de las restantes, ya sea al evaluarlas, eliminarlas o agregarlas al modelo. Su codificación (de 0 y 1) es a veces modificada por -1 u otras combinaciones.

$$x_j \begin{cases} x_{j1} \\ x_{j2} \\ \vdots \\ x_{jr} \end{cases} = \begin{matrix} x_{j(1)} \begin{cases} x_{j(1)} = 1 \\ x_{j(1)} = 0 \end{cases} \\ x_{j(2)} \begin{cases} x_{j(2)} = 1 \\ x_{j(2)} = 0 \end{cases} \\ \vdots \\ x_{j(r-1)} \begin{cases} x_{j(r-1)} = 1 \\ x_{j(r-1)} = 0 \end{cases} \end{matrix}$$

Cuando un individuo tiene como característica la variable *r,* el resto de las variables construidas deberán valer cero, asumiendo que ninguno de los *r-1* valores son respuesta y en su lugar lo es *r.*

Otro aspecto relevante es la *interacción*. Sea la variable x_i y la variable x_j, es posible suponer que entre ellas existe alguna interacción o una posible mejora en el ajuste. Está construcción forma parte de las posibles combinaciones de variables, se confirma su importancia en el modelo al ver que modifican drásticamente los valores de coeficientes o estadísticos. Bajo este esquema cualquier combinación de variables puede formar una nueva variable de interacción x_{int} a través de su multiplicación $x_{int} = x_i x_j$.

Al comprender el manejo de variables, la combinación lineal usada quedará especificada de aquí en adelante con la existencia de *k* variables en total, pudiendo ser *d* variables dummies, *w* variables de interacción y el resto de las *k-d-w* variables iniciales, todas con sus respectivos coeficientes. Procurando siempre que el tamaño de muestra *n* sea sustancialmente mayor que el total de las *k* variables independientes.

$$c_0 + c_1x_1 + c_2x_2 + \cdots + \overbrace{c_d x_{j(d)}}^{va.\ dummy} + \cdots + \overbrace{c_w x_{int}}^{va.\ interacción} = X \cdot C$$

Con esta construcción es preciso conocer las formas de selección de las variables:

1. Inclusión de variables *(Forward)* – Selección en la que se incluyen las variables una a una, iniciando de ceros y tomando las más significativas. El modelo se evalúa cada vez que se agrega una variable.

2. Eliminación de variables (*Backward)* – Selección en la que se inicia con todas las variables, eliminando una a una y evaluando el ajuste cada vez que esto ocurre.

3. Combinación (*Stepwise*) – Es una mezcla de las selecciones anteriores, se puede empezar por cualquiera y meter o sacar variables en cualquier momento, incluso aunque ya hayan sido evaluadas, debido a que se trata de nuevas combinaciones.

El tipo de estudio es fundamental. Su clasificación es básica para las evaluaciones de demografía, estadística y muestreo:

I. Descriptivos:

- Caso: Estudios que realizan una descripción de las variables y características asociadas al evento para un solo caso en el que ocurre.
- Series de caso: Describen el evento para más de un caso en el que ocurre (éxito).
- Transversal: Estudio que observa los factores de riesgo que pueden desencadenar la aparición o permanencia de un evento.

II. Analíticos

- Cohorte: Estudios en los que se observa a elementos de la población previos a la ocurrencia del evento, divididos en aquellos expuestos y no expuestos a ciertos factores, con el fin de encontrar aquellos en los que ocurre el éxito del evento.
- Caso – Control: Se seleccionan elementos de la población con el evento de éxito y de fracaso, buscando el estudio de los factores que lo desencadenaron. Es un estudio retrospectivo y es el más común para scoring.

3.1.5. Estimación de parámetros del modelo

La estimación de los parámetros C que dan la relación entre las variables y los diversos pesos a cada característica de los clientes implica tomar la función logística y la matriz X de la muestra, con n elementos de cada variable X_j y sus resultados conocidos de la variable Y.

Mientras que en la regresión lineal se usa el procedimiento de mínimos cuadrados para minimizar la suma de cuadrados de los errores, en la regresión logística, por contar con la existencia de variables de respuesta categóricas, se recurre a la estimación por Máxima Verosimilitud, en la

que se calculan valores cuya probabilidad de obtener el conjunto verdadero de la muestra sea máxima.

El *EMV (Estimador Máximo Verosimil) o MLE (Maximum Likelihood Estimate),* tiene la posibilidad de inferir uno o más parámetros $C = c_1, c_2, \dots, c_k$ de cualquier función de probabilidad, basándose en las *n* observaciones hechas $\hat{X}$. Brinda el valor estimado que maximiza la *Función de Verosimilitud*, una función conjunta de los parámetros desconocidos y las observaciones $\hat{X}$ de la muestra: $L(\hat{X}; C) = p(\hat{\boldsymbol{x}}_1; C) \cdots p(\hat{\boldsymbol{x}}_n; C) = \prod_{i=1}^{n} p(\hat{\boldsymbol{x}}_i; C)$. La maximización para encontrar la estimación de los coeficientes $\hat{C}$ deberá cumplir con la desigualdad $L(\hat{X}; \hat{C}) \geq L(\hat{X}; C)$.

En esta estimación es común el uso de la función Logaritmo, para formar el *Logaritmo de la Verosimilitud* definido como $l(\hat{X}; C) = ln\, p(\hat{X}; C)$, ya que la función *ln(x)* es creciente y unívoca para *x>0*, fácilmente derivable y no afecta a $L(\hat{X}; C)$, tal que $\ln L(\hat{X}; \hat{C}) \geq \ln L(\hat{X}; C) \Rightarrow l(\hat{X}; \hat{C}) \geq l(\hat{X}; C)$ permanece y la maximización basada en la propiedad de suma de logaritmos queda como $l(\hat{X}; C) = \sum_{i=1}^{n} \ln p(\hat{\boldsymbol{x}}_i; C)$.

Es aquí cuando el valor de la estimación $\hat{C}$ se encuentra como en el procedimiento de maximización de cualquier función, diferenciando la función logaritmo de la verosimilitud e igualando la derivada (ya sea total o parcial, por el número de coeficientes con el que se cuenta) con respecto a cero, para así despejar el conjunto de parámetros, tomando en cuenta que la derivada en segundo orden deberá ser negativa.

$$\left.\frac{\partial L(\hat{X}; C)}{\partial C}\right|_{C=\hat{C}} = 0 \qquad \left.\frac{\partial^2 L(\hat{X}; C)}{\partial C^2}\right|_{C=\hat{C}} < 0$$

Una vez comprendida la idea se usa para emplearla a la función logística $\pi(X)$, que vista desde el punto de vista de la variable *Y* representa un evento de distribución Bernoulli para entender que cada elemento observado tiene la posibilidad de ser un éxito o un fracaso con su respectiva probabilidad. La matriz $\hat{X}$, que cuenta con las observaciones hechas de la muestra, será de $(n \times k + 1)$ elementos con la información de *n* registros y sus *k* características con los *k+1* coeficientes incluidos.

$$\begin{bmatrix} 1 & \hat{x}_{11} & \hat{x}_{12} & \cdots & \hat{x}_{1k} \\ 1 & \hat{x}_{21} & \hat{x}_{22} & & \hat{x}_{2k} \\ & \vdots & & \ddots & \vdots \\ 1 & \hat{x}_{n1} & \hat{x}_{n2} & \cdots & \hat{x}_{nk} \end{bmatrix}$$

Teniendo la posibilidad de calcular la función de verosimilitud a través de cada elemento $\hat{x}_i$ de la muestra, con su respectivo vector de características $\hat{x}_{ij} = \hat{x}_{i1}, \hat{x}_{i2}, \dots, \hat{x}_{ik}$:

$$L(\hat{X}; C) = \pi(\hat{x}_1; C) \cdots (1 - \pi(\hat{x}_h; C)) \cdots \pi(\hat{x}_n; C) = \prod_{i=1}^{n} \pi(\hat{x}_i; C)^{y_i} [1 - \pi(\hat{x}_i; C)]^{1-y_i}$$

Para así aplicar la función: $l(\hat{X}; C) = \sum_{i=1}^{n} y_i \ln \pi(\hat{x}_i; C) + (1 - y_i) \ln[1 - \pi(\hat{x}_i; C)]$

Se deriva el logaritmo de la verosimilitud con respecto al vector de parámetros, teniendo así un vector de *k+1* derivadas parciales. Primero se realiza con respecto a $\pi(\hat{\boldsymbol{x}}_{\boldsymbol{i}}; C)$, que contiene los parámetros:

$$\frac{\partial\, l(\hat{X}; C)}{\partial\, \pi(\hat{\boldsymbol{x}}_{\boldsymbol{i}}; C)} = \sum_{i=1}^{n} \left[\frac{\boldsymbol{y_i}}{\pi(\hat{\boldsymbol{x}}_{\boldsymbol{i}}; C)} + \frac{\boldsymbol{1 - y_i}}{1 - \pi(\hat{\boldsymbol{x}}_{\boldsymbol{i}}; C)} \right] \frac{\partial\, \pi(\hat{\boldsymbol{x}}_{\boldsymbol{i}}; C)}{\partial\, C}$$

Más tarde se deriva con respecto a la combinación lineal $X \cdot C$ y finamente, con respecto a cualquier coeficiente c_i:

$$\frac{\partial\, \pi(\hat{\boldsymbol{x}}_{\boldsymbol{i}}; C)}{\partial\, X \cdot C} \frac{\partial\, X \cdot C}{\partial c_i} = \frac{\partial\, \pi(\hat{\boldsymbol{x}}_{\boldsymbol{i}}; C)}{\partial\, C}$$

Lo cual se realiza del siguiente modo:

$$\frac{\partial\, \pi(\hat{\boldsymbol{x}}_{\boldsymbol{i}}; C)}{\partial\, C \cdot X} = \frac{-e^{-(X \cdot C)}(-1)}{(1 + e^{-(X \cdot C)})^2} = \frac{e^{-(X \cdot C)}}{(1 + e^{-(X \cdot C)})^2} = \frac{e^{-(X \cdot C)}}{1 + e^{-(X \cdot C)}} \frac{1}{1 + e^{-(X \cdot C)}}$$

$$= \left[1 - \frac{1}{1 + e^{-(X \cdot C)}} \right] \frac{1}{1 + e^{-(X \cdot C)}} = [1 - \pi(\hat{\boldsymbol{x}}_{\boldsymbol{i}}; C)]\, \pi(\hat{\boldsymbol{x}}_{\boldsymbol{i}}; C)$$

Así como para cada $c_i \neq c_0$ dado $i = 0,1, \dots k$ con $\frac{\partial\, X \cdot C}{\partial\, c_i} = X_i$ y por otro lado $\frac{\partial\, X \cdot C}{\partial\, c_0} = 1$

Si se incluye la derivada en la derivada del logaritmo de la verosimilitud, se tienen los dos casos, la derivada parcial para c_0 y la de los k coeficientes restantes c_i:

$$\frac{\partial\, l(\hat{X};C)}{\partial\, c_0} = \sum_{i=1}^{n}\left[\frac{\boldsymbol{y_i}}{\pi(\hat{\boldsymbol{x}}_{\boldsymbol{i}};C)} + \frac{\boldsymbol{1-y_i}}{1-\pi(\hat{\boldsymbol{x}}_{\boldsymbol{i}};C)}\right][1-\pi(\hat{\boldsymbol{x}}_{\boldsymbol{i}};C)]\,\pi(\hat{\boldsymbol{x}}_{\boldsymbol{i}};C)$$

$$\frac{\partial\, l(\hat{X};C)}{\partial\, c_i} = \sum_{i=1}^{n}\left[\frac{\boldsymbol{y_i}}{\pi(\hat{\boldsymbol{x}}_{\boldsymbol{i}};C)} + \frac{\boldsymbol{1-y_i}}{1-\pi(\hat{\boldsymbol{x}}_{\boldsymbol{i}};C)}\right][1-\pi(\hat{\boldsymbol{x}}_{\boldsymbol{i}};C)]\,\pi(\hat{\boldsymbol{x}}_{\boldsymbol{i}};C)\,\hat{\boldsymbol{x}}_{\boldsymbol{i}}$$

Logrando simplificar la notación:

$$\frac{\partial\, l(\hat{X};C)}{\partial\, C_i} = \sum_{i=1}^{n}[y_i\hat{x}_i - y_i\pi(\hat{x}_i;C)\hat{x}_i - \pi(\hat{x}_i;C)\hat{x}_i + y_i\pi(\hat{x}_i;C)\hat{x}_i] = \sum_{i=1}^{n}[y_i\hat{x}_i - \pi(\hat{x}_i;C)\hat{x}_i]$$

$$\Rightarrow \frac{\partial\, l(\hat{X};C)}{\partial\, c_i} = \sum_{i=1}^{n}[y_i - \pi(\hat{x}_i;C)]\hat{x}_i$$

$$\frac{\partial\, l(\hat{X};C)}{\partial\, c_0} = \sum_{i=1}^{n}[y_i - \pi(\hat{x}_i;C)]$$

Estas ecuaciones se igualan a cero con el fin de obtener los estimadores, lo cual es un procedimiento iterativo por tratarse de ecuaciones no lineales, donde para $c_i \neq c_0$ se tiene $\sum_{i=1}^{n}[y_i - \pi(\hat{x}_i;C)]\hat{x}_i = 0$ y para c_0 que $\sum_{i=1}^{n}[y_i - \pi(\hat{x}_i;C)] = 0$.

Uno de los procedimientos más usados es el de *Newton Raphson*, eficiente por encontrar las raíces r de ecuaciones o funciones no lineales en una convergencia rápida con ayuda de las tangentes a la curva, interceptándola con valores t hasta cruzar con los ejes en un punto en el que la función $f = y_i - \pi(\hat{x}_i;C)$ sea cero. Este procedimiento se desarrolla en serie de Taylor:

$$f(r) = f(\hat{C}_t) + (r-\hat{C}_t)f'(\hat{C}_t) + \cdots$$

$$\Rightarrow\ 0 = f(\hat{x}_t) + (r-\hat{C}_t)f'(\hat{C}_t) + \cdots$$

$$\Rightarrow\ r \approx \hat{C}_{t+1} = \hat{C}_t - \frac{f(\hat{C}_t)}{f'(\hat{C}_t)}$$

Con esta función se dan valores arbitrarios de partida al vector de coeficientes $\hat{C}$, obteniendo la segunda derivada para calcular las tangentes a este vector estimado con la esperanza negativa de la matriz Hessiana (segundas derivadas parciales), lo que se conoce como *Matriz de Información* $I(\hat{C})$. Dichas derivadas son calculadas de manera análoga a las primeras y comprueban la existencia del valor máximo por ser menores a cero:

$$\frac{\partial\, l(\hat{X}; C)}{\partial\, {c_i}^2} = -\sum_{i=1}^{n} \pi(\hat{x}_i; C)[1 - \pi(\hat{x}_i; C)]{\hat{x}_i}^2$$

$$\frac{\partial\, l(\hat{X}; C)}{\partial\, c_i c_h} = -\sum_{i=1}^{n} [y_i - \pi(\hat{x}_i; C)]\hat{x}_i\hat{x}_h \quad con\ i \neq h$$

$$\frac{\partial\, l(\hat{X}; C)}{\partial\, {c_0}^2} = -\sum_{i=1}^{n} \pi(\hat{x}_i; C)[1 - \pi(\hat{x}_i; C)]$$

Tomando estas ecuaciones negativas para cumplir con la ecuación, la matriz inversa de tamaño $(k+1 \times k+1)$ forma las varianzas y covarianzas tal que $cov(\hat{C}) = I(\hat{C})^{-1} = (\hat{X}'V\hat{X})^{-1}$ donde la matriz $\hat{X}'$ es la traspuesta de $\hat{X}$ y la matriz V es la matriz diagonal de tamaño $(n \times n)$ con las varianzas de cada individuo $\hat{x}_i$. Lo anterior proviene de lo que se conoce como la *Cota de Cramer Rao* que proporciona la mínima varianza que puede alcanzar cualquier estimador insesgado de un vector de parámetros.

$$V = \begin{bmatrix} \pi(\hat{\boldsymbol{x}}_{\boldsymbol{1}}; C)[\mathbf{1} - \pi(\hat{\boldsymbol{x}}_{\boldsymbol{1}}; C)] & 0 & 0 \\ 0 & \ddots & 0 \\ 0 & 0 & \pi(\hat{\boldsymbol{x}}_{\boldsymbol{n}}; C)[\mathbf{1} - \pi(\hat{\boldsymbol{x}}_{\boldsymbol{n}}; C)] \end{bmatrix}$$

$$cov(\hat{C}) = \begin{bmatrix} \hat{\sigma}^2(\hat{c}_0) & \hat{\sigma}^2(\hat{c}_0, \hat{c}_1) & \dots & \hat{\sigma}^2(\hat{c}_0, \hat{c}_k) \\ \hat{\sigma}^2(\hat{c}_1, \hat{c}_0) & \ddots & & \hat{\sigma}^2(\hat{c}_1, \hat{c}_k) \\ \vdots & & & \vdots \\ \hat{\sigma}^2(\hat{c}_k, \hat{c}_0) & \cdots & & \hat{\sigma}^2(\hat{c}_k) \end{bmatrix}$$

Esta matriz, junto con el vector de primeras derivadas que se anula en el punto máximo, son vitales para la iteración en la que se buscará iniciar con $\hat{C}_t$ y estimar un siguiente valor posible $\hat{C}_{t+1}$ para el vector de coeficientes, el cual deberá ser probado para revisar si se trata de la solución final que resuelve el sistema. En caso de que no sea así, se tomará el valor de $\hat{C}_{t+1}$ como inicial, repitiendo el procedimiento hasta hallar el máximo estimador $\hat{C}$, donde la diferencia entre $\hat{C}_{t+1}$ y $\hat{C}_t$ sea prácticamente cero (generalmente se usa una diferencia de 10^{-6}).

Lo anterior cumpliendo el esquema $\hat{C}_{t+1} = \hat{C}_t + \left(\hat{X}'V\hat{X}\right)^{-1}X'\left[\boldsymbol{y_i} - \pi\left(\hat{X};\hat{C}_t\right)\right]$.

Como se puede ver, se trata de un proceso robusto, complejo a mayor cantidad de elementos de la muestra y características a evaluar, implica calcular una gran cantidad de operaciones, para lo cual hoy en día existen sistemas tecnológicos que facilitan el proceso como SPSS, LOGIT, RELODI, STATA y SAS, entre otros.

3.1.6. Ajuste y validación

Para la verificación del modelo se buscará que las variables influyan significativamente con diversos estadísticos. El primero es el *cociente de verosimilitud*, un estadístico que mide los efectos de ajuste y cambios en las variables del modelo.

Para el efecto de cambio se calcula la *Devianza (D)* y la construcción de varios modelos a los cuales se les disminuya una cantidad *h* de variables con respecto al modelo saturado (que incluye las *k* variables de información disponibles en la muestra). También se pueden ir aumentando variables con respecto a un modelo inicial que contenga la más significativa.

La evaluación se realiza mediante ambas hipótesis y una expresión:

$$H_0: Las\ variables\ no\ influyen\ en\ el\ modelo\ (modA) \Rightarrow\ c_i = 0\ \ \forall\, i = 1, \dots, h$$

$$H_1: Las\ variables\ influyen\ en\ el\ modelo\ (modB) \Rightarrow\ c_i \neq 0\ \ \forall\, i = 1, \dots, h$$

$$D = -2\ln\left[L\left(\hat{X};\hat{C}_{modA}\right) - L\left(\hat{X};\hat{C}_{modB}\right)\right] = -2\ln\left(\frac{L\left(\hat{X};\hat{C}_{modA}\right)}{L\left(\hat{X};\hat{C}_{modB}\right)}\right)$$

Si $D > \mathcal{X}^2_{a,\ n-h}$ entonces se rechaza H_0, concluyendo que las h variables evaluadas son significativas y el modelo en el cual se eliminan, no ajusta correctamente. Cuando existen variables dummies, la eliminación o inclusión de las mismas será asociada.

Si lo que se busca es medir el efecto de ajuste del modelo, se busca que la expresión de devianza tenga valor de cero:

$$D = -2\ln\left[L\left(\hat{X};\hat{C}\right)\right] = -2\sum_{i=1}^{n}\ln\pi\left(\hat{\boldsymbol{x}}_{\boldsymbol{i}};\hat{C}\right) + (\mathbf{1}-\boldsymbol{y}_{\boldsymbol{i}})\ln\left[1-\pi\left(\hat{\boldsymbol{x}}_{\boldsymbol{i}};\hat{C}\right)\right]$$

Un estadístico más es el de *Wald,* para evaluar la significancia del coeficiente de cada variable en el modelo. Su prueba es más sencilla que en el caso anterior, ya que sólo ocupa el cociente del coeficiente $\hat{c}_i$ entre su desviación estándar estimada $\hat{\sigma}(c_i)$. Esta prueba se realiza con la hipótesis $H_0: c_i = 0$ y la alternativa $H_1: c_i \neq 0$.

$$Wald = \frac{\hat{c}_i}{\hat{\sigma}(c_i)}$$

Para su evaluación se calcula el valor de una Ji-Cuadrada con *1* grado de libertad $\left(\mathcal{X}^2_{a,1}\right)$, que representa la variable a evaluar y se rechaza H_0 cuando el valor del cociente (Wald) es mayor.

Otro elemento para valorar la capacidad de clasificación es el de *Hosmer Lemeshow*, prueba en la que se construyen tablas para comparar los resultados de estimación del modelo contra los resultados reales de la muestra, hecha la clasificación de éxitos y fracasos para ambos casos. Para su uso se toman las probabilidades estimadas de la muestra, ordenadas de mayor a menor y se dividen en Q percentiles de los individuos que hayan tenido riesgos semejantes. Se cuentan los casos observados $(O_{y,q})$ y estimados $(E_{y,q})$ para cada cuantil y se calcula *HL.*

$$HL = \sum_{q=1}^{Q}\frac{\left(O_{y,q} - E_{y,q}\right)^2}{E_{y,q}} \quad con\ y = 1,0$$

Este estadístico se compara con una distribución Ji-cuadrada, a un nivel de confianza α y puede modificarse cuando no se tiene la posibilidad de calcular en la práctica la probabilidad de los *n* individuos, en su lugar se calculan probabilidades promedio por cuantil $\bar{\pi}_q$ y con una tabla de

frecuencias la cantidad de elementos en cada uno de ellos para así obtener *HL*. En esta modificación los elementos tomados serán sólo los éxitos (*y=1),* dados por n_q.

$$HL = \sum_{q=1}^{Q} \frac{\left(O_q - n_q \bar{\pi}_q\right)^2}{n_q \bar{\pi}_q \bar{\pi}_q}$$

Si $HL \geq \chi^2_{\alpha,Q-2}$ se rechaza el modelo debido a que no clasifica los datos adecuadamente. Es una de las pruebas más utilizadas.

De manera gráfica, incluso previo al ajuste del modelo, es recomendable hacer una *Gráfica de Dispersión*, un diagrama simple para el análisis de relaciones entre 2 o más variables. Identifica tendencias, relaciones y agrupaciones. Se crean con ayuda de la representación de los valores de la muestra ajustados a escalas semejantes y otorga información sobre la correlación: positiva, negativa o nula, lineal o no-lineal.

Finalmente existe la *Matriz de Correlación*, con los coeficientes que indican la relación entre las variables, mismas que se eliminarán si dependen totalmente una de otra. Calculando la matriz se considerará eliminar las variables con coeficientes de correlación muy elevados, probando su eliminación con los modelos de ajuste previos.

La desventaja es que sólo se consigue conocer colineanilidad entre dos variables y no multicolinealidad (entre más de dos). Esta matriz es simétrica, con valor de uno en la diagonal:

$$\begin{bmatrix} 1 & \rho(\hat{X}_1, \hat{X}_2) & \cdots & \rho(\hat{X}_1, \hat{X}_k) \\ \rho(\hat{X}_2, \hat{X}_1) & 1 & & \rho(\hat{X}_2, \hat{X}_k) \\ \vdots & & & \vdots \\ \rho(\hat{X}_k, \hat{X}_1) & \cdots & & 1 \end{bmatrix}$$

Donde cada $\rho(\hat{X}_i, \hat{X}_j)$ para cada par de variables $\hat{X}_i \neq \hat{X}_j$ es el coeficiente de correlación, pudiendo ser evaluado con el valor *r* del coeficiente de correlación lineal de *Pearson,* calculado con la covarianza y desviaciones, como fue analizado en el capítulo anterior:

$$\rho(\hat{X}_i, \hat{X}_j) = r = \frac{\hat{\sigma}^2(X_i, X_j)}{\hat{\sigma}(X_i)\hat{\sigma}(X_j)}$$

Un cálculo alternativo no paramétrico que toma en cuenta los elementos de la muestra más aislados es el coeficiente de correlación de *Spearman.* Para este coeficiente no se usan supuestos de linealidad, es una correlación por orden de rango.

Su procedimiento usa valores en escalas ordinales, tomando los elementos de cada variable por separado y las diferencias *D* de orden que existen para cada de pareja $(x_i, \hat{x}_j)$, calculando:

$$\rho_{sp}(\hat{X}_i, \hat{X}_j) = 1 - \frac{6\sum_{i=1}^{n} D^2}{n(n^2-1)}$$

La *Multicolinealidad* es otra opción de análisis que puede realizarse a través de diversos modelos que incluyan o eliminen la presencia de algunas variables independientes, debiendo calcular para cada uno de ellos la medida R^2 conforme a una regresión lineal y su raíz cuadrada positiva para el coeficiente de correlación múltiple.

3.1.7. Interpretación de los coeficientes

Al contar con un modelo que tenga un buen ajusta, la interpretación de los coeficientes es lo que permite la comprensión de la presencia de cada una de las variables.

Las estimaciones existentes más comunes y útiles hasta hoy para este tipo de regresión son los cocientes llamados *Momios,* que son valores mayores a cero y sin cotas superiores, vistos como medidas de disparidad, sin importar el tipo de estudio que sea o el método de selección de variable. Tienen una interpretación directa sobre la contribución de cada variable al modelo, lo cual no ocurre en todos los métodos para identificación de los niveles de riesgo.

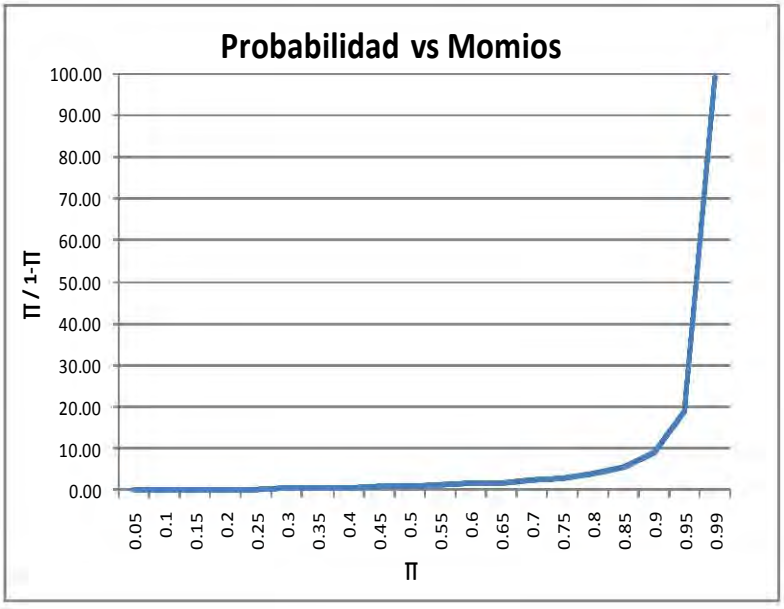
Es una medida que intuitivamente ha sido ubicada para referir expectativas de ganancia económica. En los eventos deportivos se habla de apuestas M a N, por ejemplo 8:2, donde se espera que en caso de éxito se tengan ganancias de $8 por cada apuesta, asumiendo una probabilidad de $\frac{8}{8+2} = \frac{8}{10} = .8$ de ganar.

Dichos cocientes, en pocas palabras, cuentan el número de veces que será más probable que ocurra un éxito del evento correspondiente con cada variable *i* (para credit scoring el incumplimiento de un cliente):

$$Momio = \frac{\pi(\hat{X};\hat{C})}{1-\pi(\hat{X};\hat{C})}$$

Su gráfica se puede ver a continuación con probabilidad de éxito y fracaso:

$\pi(\hat{X};\hat{C})$	$1-\pi(\hat{X};\hat{C})$	$\frac{\pi(\hat{X};\hat{C})}{1-\pi(\hat{X};\hat{C})}$
0.05	0.95	0.05
0.1	0.9	0.11
0.15	0.85	0.18
0.2	0.8	0.25
0.25	0.75	0.33
0.3	0.7	0.43
0.35	0.65	0.54
0.4	0.6	0.67
0.45	0.55	0.82
0.5	0.5	1.00
0.55	0.45	1.22
0.6	0.4	1.50
0.65	0.35	1.86
0.7	0.3	2.33
0.75	0.25	3.00
0.8	0.2	4.00
0.85	0.15	5.67
0.9	0.1	9.00
0.95	0.05	19.00
0.99	0.01	99.00



Los *parámetros* o *coeficientes* dan a conocer en qué dirección se mueve la probabilidad con el incremento o decremento del cada una de las variables, sin embargo, cuando se trata de conocer el valor específico de dicha probabilidad es necesario realizar una evaluación mayor del modelo.

Para ello existen los *Odds Ratios (OR) o factores de cambio*, que son razones para evaluar cuantitativamente el impacto al cambiar el valor de una variable, ya sea mediante un incremento o por el cambio entre la ausencia y presencia ($\hat{c}_i = 0$ ó $\hat{c}_i \neq 0$). Se diferencian del valor de un Riesgo Relativo (Risk Ratio RR) debido a que este último, solo tiene la capacidad de evaluar a una población estudiada por medio de un cohorte, no un estudio de tipo caso-control en los que ya se cuenta con una muestra de factores y resultados (éxito o fracaso) conocidos.

Estos factores de cambio requieren de dos modelos para el análisis de cada variable *i*, uno que contenga la totalidad de variables (*mod A*) y otro que la elimine (*mod B*). La razón de momios de ambos modelos con logaritmo natural es:

$$OR = \frac{\dfrac{\pi\left(\hat{X};\hat{C}_{modA}\right)}{1-\pi\left(\hat{X};\hat{C}_{modA}\right)}}{\dfrac{\pi\left(\hat{X};\hat{C}_{modB}\right)}{1-\pi\left(\hat{X};\hat{C}_{modB}\right)}}$$

$$\Rightarrow ln(OR) = \ln\left[\frac{\dfrac{\pi\left(\hat{X};\hat{C}_{modA}\right)}{1-\pi\left(\hat{X};\hat{C}_{modA}\right)}}{\dfrac{\pi\left(\hat{X};\hat{C}_{modB}\right)}{1-\pi\left(\hat{X};\hat{C}_{modB}\right)}}\right] = \ln\left[\frac{\pi\left(\hat{X};\hat{C}_{modA}\right)}{1-\pi\left(\hat{X};\hat{C}_{modA}\right)}\right] - \ln\left[\frac{\pi\left(\hat{X};\hat{C}_{modB}\right)}{1-\pi\left(\hat{X};\hat{C}_{modB}\right)}\right]$$

$$\Rightarrow ln(OR) = \sum_{j=1}^{k} \hat{c}_j \cdot \hat{x}_j - \sum_{j=1,\ j\neq i}^{k} \hat{c}_j \cdot \hat{x}_j = \hat{c}_i \cdot \hat{x}_i$$

Para eliminar el logaritmo, se aplica la función exponencial en ambos lados:

$$\ln\left[\frac{\dfrac{\pi\left(\hat{X};\hat{C}_{modA}\right)}{1-\pi\left(\hat{X};\hat{C}_{modA}\right)}}{\dfrac{\pi\left(\hat{X};\hat{C}_{modB}\right)}{1-\pi\left(\hat{X};\hat{C}_{modB}\right)}}\right] = \hat{c}_i \cdot \hat{x}_i \ \Rightarrow \ \frac{\dfrac{\pi\left(\hat{X};\hat{C}_{modA}\right)}{1-\pi\left(\hat{X};\hat{C}_{modA}\right)}}{\dfrac{\pi\left(\hat{X};\hat{C}_{modB}\right)}{1-\pi\left(\hat{X};\hat{C}_{modB}\right)}} = e^{\hat{c}_i \cdot \hat{x}_i}$$

Quedando $e^{\hat{c}_i \cdot \hat{x}_i}$ como el factor de cambio, que para ser utilizado serán tomados intervalos, tal que si buscamos saber el efecto que tendrá la probabilidad de éxito al modificar en un intervalo de u unidades (comprendidas en el dominio de la variable), se obtendrá $e^{\hat{c}_i \cdot u}$.

Al tratarse por ejemplo de una variable independiente dicotómica, evaluar la ausencia o presencia dará un valor de $u = x_{i\,(presencia)} - x_{i\,(ausencia)} = 1 - 0 = 1$, siendo suficiente $e^{\hat{c}_i}$ para saber cuánto cambia la probabilidad con este movimiento. Análogamente, con $u = 1$, si no se aplica la función exponencial la simple razón de los momios dará el valor del coeficiente $\hat{c}_i$:

$$\ln\left[\frac{\frac{\pi(\hat{X};\hat{C}_{modA})}{1-\pi(\hat{X};\hat{C}_{modA})}}{\frac{\pi(\hat{X};\hat{C}_{modB})}{1-\pi(\hat{X};\hat{C}_{modB})}}\right] = \hat{c}_i \cdot \hat{u}_i \quad \Rightarrow \quad \ln\left[\frac{\frac{\pi(\hat{X};\hat{C}_{modA})}{1-\pi(\hat{X};\hat{C}_{modA})}}{\frac{\pi(\hat{X};\hat{C}_{modB})}{1-\pi(\hat{X};\hat{C}_{modB})}}\right] = \hat{c}_i$$

Para esta estimación la valoración se hace con todos los coeficientes involucrados, ya sea *d* variables dummies o *w* variables de interacción, con lo cual para una variable *i* un factor de cambio queda definido como $OR = e^{\sum_{j=1}^{d} \hat{c}_j \cdot \hat{x}_{i(j)} + \sum_{j=1}^{w} \hat{c}_j \hat{x}_i \cdot \hat{x}_j}$ y en términos de unidades como $OR = e^{\sum_{j=1}^{d} \hat{c}_j \cdot u_{(j)} + \sum_{j=1}^{w} \hat{c}_j u \cdot \hat{x}_j}$.

3.1.8. Clasificación de la población

Una ventaja fundamental del método aplicable al credit scoring es la posibilidad de clasificar. La parte medular de este aspecto es conocer el punto en el cual a un individuo se le considera como incumplido, dicho en otras palabras definir el punto de corte óptimo (Cut-off). Para esto se analizará el riesgo que se quiere tomar, ya que entre mayor sea el punto de corte, mayor será la probabilidad de incumplimiento pero podría aumentar la utilidad al colocar un mayor número de créditos, caso en el cual la administración de riesgos tendrá que tomar una decisión.

Una acción como clasificar o discriminar es sencilla de hacer con este modelo a partir del punto medio en la que si un cliente obtiene una probabilidad arriba de 0.5 será incumplido, sin embargo esto no es lo óptimo, a veces es posible aceptar un mayor riesgo o incluso necesario disminuirlo, lo cual se puede ver con ayuda de elementos como la especificidad y la sensibilidad, relacionados directamente con los errores tipo I y II de estadística.

Ya hecha la clasificación de éxitos y fracasos (incumplidos y cumplidos, respectivamente), existe la diferencia entre observaciones del modelo y casos reales de la muestra, dados por los elementos verdaderos y falsos:

COMPARACIÓN DEL MODELO	*MUESTRA (ÉXITOS) Y = 1*	*MUESTRA (FRACASOS) Y = 0*
MODELO (ÉXITOS) Y=1	*A (VERDADERO)*	*B (FALSO)*
MODELO (FRACASOS) Y = 0	*C (FALSO)*	*D (VERDADERO)*

Sensibilidad: Es la proporción de elementos de la población clasificados en el grupo de éxito de ocurrencias correctamente (al incluirlos en el modelo su clasificación fue de clientes incumplidos cuando el estatus real en la muestra lo era). Es llamado el grupo de verdaderos positivos y son aquellos elementos del sub-conjunto A.

$$Se = \frac{A}{A + C}$$

Especificidad: Es la proporción de elementos de la población clasificados en el grupo de fracaso de ocurrencias correctamente (al incluirlos en el modelo su clasificación fue de clientes en cumplimiento cuando el estatus real en la muestra lo era). Es llamado el grupo de verdaderos negativos y son aquellos elementos del sub-conjunto D. Su complemento es la cantidad de éxitos falsos, valor que implica a clientes incumplidos que pudieron ser clasificados como cumplidos.

$$Sp = \frac{D}{B + D}$$

La gráfica ROC *(Receiver Operating Characteristic)* es un buen apoyo al momento de estas clasificaciones binarias, va siempre en función de los éxitos y representa la sensibilidad en un eje frente al porcentaje equivalente a uno menos la especificidad $(1 - Sp)$. El punto elegido de corte óptimo (Cut-off) se da con al realizar diversos escenarios y seleccionar el punto con mayor sensibilidad y menor complemento de especificidad, lo cual se da con exactitud en *(1,0)*. El área bajo la curva que forma esta gráfica con la recta de la función identidad es conocida como AUC *(Area Under Curve),* calculada con el valor *U de Mann-Whitney,* compara dos muestras apoyado por el tamaño de cada población (*n1* y *n0*) y la suma de valores de rango para tomar el mínimo:

$$AUC = 1 - \frac{U}{n1 * n0},$$

Esta área deberá ser lo más cercano a uno posible, para que esté lo más cercano al valor óptimo y lejos de la recta de identidad, en la cual existe una gran cantidad de fracasos y éxitos falsos.

3.2. Modelo de aplicación al sector rural en México

3.2.1. Introducción al modelo de aplicación

Es en este punto se vuelve preciso realizar la aplicación que permitirá aterrizar la forma en que se usa el modelo de regresión logística en credit scoring, misma que como fue mencionado se ha decidido realizar para el sector rural en México.

El modelo ha sido construido con ayuda de la información de los créditos otorgados por Financiera Rural (FR), el organismo que canaliza recursos financieros, da asistencia técnica, capacitación y asesoría al sector rural. Propicia condiciones para la recuperación del nivel de vida de los productores rurales, sobre todo a través de entidades intermediarias que aseguren impacto en ellos por sus posibilidades de acceso. Se rige por la Ley Orgánica de Financiera Rural y cuenta con premisas que se desprenden del Programa Sectorial de Desarrollo Agropecuario y Pesquero, el Plan Nacional de Desarrollo y los Lineamientos de Política de la Banca de Desarrollo de la cual es parte, junto con la mejora continua del país, razón suficiente para este estudio.

Al año 2011 Financiera Rural fue un organismo independiente y descentralizado de la SHCP, parte de un sector de la banca de desarrollo del país para apoyar actividades vinculadas con el medio rural. Su estructura está compuesta por un Consejo Directivo y un Director General del que dependen: Un Comité de Administración de riesgos, un Comité de Crédito, un Comité de Operación y un Comité de recursos Humanos. Sus operaciones están reguladas por la CNBV y desarrolladas en 95 agencias ubicadas a lo largo del territorio nacional, una agencia corporativa en el Distrito Federal y 5 Coordinaciones Regionales:

- Coordinación Norte, sede en Monterrey, N.L., y cobertura en los estados de Chihuahua, Coahuila, Durango, Nuevo León, San Luís Potosí, Tamaulipas y Zacatecas.
- Coordinación Noroeste, sede en Hermosillo, Son., y cobertura en los estados de Baja California, Baja California Sur, Sinaloa y Sonora.
- Coordinación Centro - Occidente, sede en Guadalajara, Jal., y cobertura en los estados de Aguascalientes, Colima, Guanajuato, Jalisco, Michoacán, Nayarit y Querétaro.

- Coordinación Sur, sede en Puebla, Pue., y cobertura en Estado de México, Guerrero, Hidalgo, Morelos, Oaxaca, Puebla, Tlaxcala y Veracruz.

- Coordinación Sureste, sede en Mérida, Yuc., y cobertura en los estados de Campeche, Chiapas, Quintana Roo, Tabasco y Yucatán.

Esta institución busca elevar la capacidad de producción y calidad de los productos que maneja en cada sector, ya sea mediante recursos monetarios y/o con programas de capacitación, que ayudan a mejorar la calidad de vida de las familias. Se atienden aspectos tanto de producción, como de comercio y consumo para el sector agrícola, ganadero, pecuario, pesquero y forestal. Es sin duda una de las financieras del Gobierno Mexicano para los programas de Desarrollo Social que obtiene utilidades más importantes, a diferencia de su antecesor Banrural. Es un organismo que ha incrementado su patrimonio de 15,000 millones de pesos con el que inicio en el 2003 a más de 25,000 millones y que ha colocado más de 130,000 millones en todo el país[17].

Los productos de crédito que maneja son: Habilitación o Avío, Simple, Prendario, Cuenta Corriente, Refaccionario y Reporto. Se otorgan tanto a personas físicas como a personas morales y se acompañan por programas de crédito auxiliares que son:

~ Programa para Financiamiento Mezcla de Recursos: Combina potencialmente los productos de crédito con recursos de apoyo de los gobiernos y FR para inversión en equipamiento e infraestructura, tecnificación de riego, jóvenes emprendedores, etc.
~ Programa para Financiamiento a Empresas de Intermediación Financiera: Programas especiales con apoyos para los créditos otorgados a este tipo de empresas.
~ Programa para Financiamiento Pre-Autorizado: Programa de atención masiva para productos específicos con esquema de paquete tecnológico (aspectos recurrentes y homologados) y sin paquete (evaluaciones financieras previamente definidas).
~ Programa para Financiamiento de compra de Cobertura de Precios: Se trata de contratos a futuro de precios a la alta o a la baja, que protejan las inversiones
~ Programa para Financiamiento del Sector Cañero: Apoyo a organizaciones cañeras, legalmente constituidas, para que otorguen créditos individuales a los productores.

[17] SHCP. *Apartado Hacienda para Todos: Banca de Desarrollo.* [Recurso en línea 2012]

Son financiamientos que sin duda han apoyado al país. Sus características en común son:

- Tasas de interés: Tasas fijas o variables de acuerdo al historial crediticio del cliente y a las garantías puestas a disposición.

- Fondo FIRA: Se otorga la posibilidad de obtener tasas fijas o variables preferenciales mediante un fondo por parte de FIRA, cuando se cumplan requisitos especiales.

- Monto: El mínimo es de 7,000 UDIS y el máximo según las necesidades de cada proyecto y de las garantías otorgadas.

Aunque también cuentan con particularidades especiales como en el cuadro a continuación:

Producto de Crédito	Garantías	Plazo
Habilitación o Avío	Garantías con materias primas, materiales o artefactos adquiridos con el Crédito	De dos años hasta diez
Crédito Simple	Garantías reales, líquidas, hipotecarias y avales, entre otras.	Hasta dos años
Crédito Prendario	Garantías como bienes o productos, granos, fertilizantes y bonos de prenda.	Hasta diez años
Refaccionario	Garantizado con Fincas, edificios, construcciones, maquinarias, muebles y otros instrumentos.	Hasta diez años
Cuenta Corriente	Garantías reales, líquidas, hipotecarias y avales, entre otras.	Hasta diez años
Reporto	Garantías como bienes o productos, granos, fertilizantes y bonos de prenda.	Máximo 45 días naturales

Como parte del SFM la Financiera Rural tiene un compromiso notable, por lo que su proceso de otorgamiento de créditos inicia con la búsqueda de mejoras en los tiempos y lugares de colocación, tratando de abarcar por completo las necesidades del medio y evitar, a su vez, el incremento de la cartera vencida. De acuerdo a esto es que surge el interés de predecir y clasificar el cumplimiento o incumplimiento de los clientes, para dar a conocer la forma en que se debe evaluar la información y poder llevar a cabo mejores operaciones crediticias.

3.2.2. Descripción de la muestra

Se obtuvo una muestra simple aleatoria de clientes de tamaño *n=600,* con clientes cumplidos (*n_0=300)* y clientes incumplidos *(n_1=300),* en la cual se tienen datos de personas físicas que han recibido créditos de tipo Avió, Simple y Refaccionario por parte de esta institución desde el año 2002 al año 2011 con requerimientos como:

- Solicitud de crédito e identificación oficial
- Referencias bancarias, personales y comerciales
- Proyecto de inversión e información financiera
- Garantías de bienes muebles e inmuebles.
- Permisos y licencias de propiedades y unidades de producción.

Toda la información fue extraída con el programa SAS de las bases de datos solicitudes de crédito, características de créditos y comportamiento de los clientes. Fueron manejadas en de Visual Fox Pro con un análisis de las variables y del proceso de crédito.

3.2.3. Variables asignadas al modelo

La asignación se inició con una variable de identificación para cada crédito asignado (*Num_credito)*, continuando con las que se suelen considerar relevantes como son:

- *Historial Crediticio*: Evaluación de cumplimientos o incumplimientos de créditos anteriores, cuando existen. Se espera una menor posibilidad de pérdida con un mejor historial de crédito.

- *Activos Fijos:* Variable que evalúa la cantidad o monto de los activos fijos del cliente, con la cual se espera que a mayor cantidad se tenga menor posibilidad de pérdida.

- *Edad:* Años de vida de una persona, influye comúnmente ya que a menor edad, menor será la capacidad de pago y mayor la posibilidad de pérdida.

- *Plazo:* Es un factor que suele ser de interés porque entre más largo sea el tiempo de uso del crédito, hay mayor posibilidad de falta de pago.

Variable Dependiente

La variable dependiente del modelo es llamada *Marca_Score* y es una variable dicotómica, establecida para hacer referencia a los clientes cumplidos con valor de 0 y a los clientes incumplidos con valor de 1. Sus atrasos se consideran definidos como incumplimientos al pertenecer al rango de 30 días y por al menos una cuota de amortización de acuerdo a la fecha correspondiente de pago de cada uno de los créditos, esto porque conllevan costos adicionales para la institución, al menos de tipo administrativo por el seguimiento cobranza.

Variables Independientes

Las variables independientes seleccionadas conforme a las características de la población rural y para identificar los casos de riesgo de incumplimiento se presentan a continuación, de acuerdo a los aspectos frecuentemente evaluados e iniciando por las variables categóricas.

- *Trimvenc*

  - Nombre: Cuatrimestres de vencimiento
  - Definición: Variable construida a partir de las fechas de vencimiento de los créditos. Tomando el mes, se asigna al cuatrimestre correspondiente.
  - Tipo: Categórica (Politómica).

    Tr1 = Cuatrimestre 1 (Enero a Abril)
    Tr2 = Cuatrimestre 2 (Mayo a Agosto)
    Tr3 = Cuatrimestre 3 (Septiembre a Diciembre)

Trim venc / Marca Score	0	1	Total general
tr1	7%	30%	**19%**
tr2	33%	62%	**48%**
tr3	60%	7%	**34%**
Total general	**100%**	**100%**	**100%**

El segundo cuatrimestre del año es el más común en los vencimientos de créditos y contiene más incumplimientos, seguido por el tercer cuatrimestre con el menor porcentaje de incumplimientos.

- *Actividad*

  - Nombre: Actividad de desarrollo del cliente
  - Definición: Variable construida por la actividad de destino del crédito.
  - Tipo: Categórica (Dicotómica).

    Agrícola = Créditos del sector Agrícola.

    Otros = Créditos de los sectores Pecuario, Pesquero y Forestal.

Actividad / Marca Score	0	1	Total general
AGRICOLA	97%	82%	**90%**
OTROS	3%	18%	**11%**
Total general	**100%**	**100%**	**100%**

La mayor parte de los créditos son de tipo Agrícola, que son incluso aquellos con más clientes cumplidos, lo cual indica que la presencia de un crédito agrícola debería disminuir sin duda el riesgo de incumplimiento.

- *Tipo_Plazo*

  - Nombre: Tipo de plazo
  - Definición: Variable construida con base en el plazo de cada crédito.
  - Tipo: Categórica (Dicotómica).

    CP - Corto plazo para aquellos créditos menores a 6 meses.

    LP – Largo plazo para aquellos créditos mayores a 6 meses.

Tipo_Plazo / Marca Score	0	1	Total general
CP	69%	98%	**84%**
LP	31%	2%	**16%**
Total general	**100%**	**100%**	**100%**

La muestra tiene en su mayoría créditos de corto plazo, con la mayor parte de los incumplimientos.

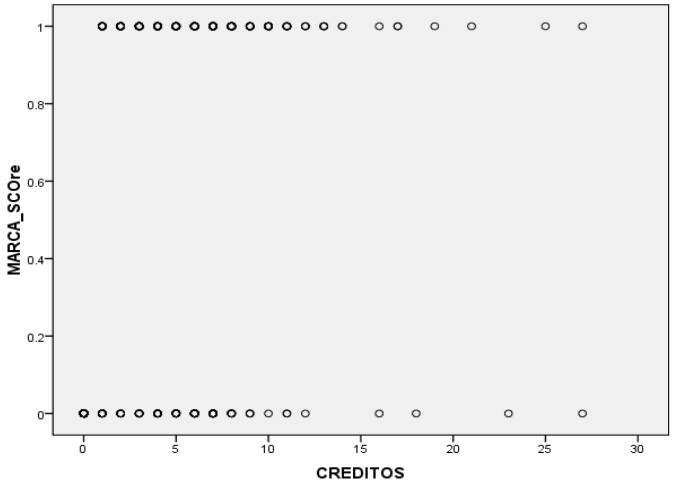
Por otro lado las variables numéricas con gráficas de dispersión y estadísticas descriptivas:

	N	Minimum	Maximum	Mean	Std. Deviation
CREDITOS	600	0	27	4.10	4.126
EDAD	600	22	85	59.78	14.178
ANIOS_ARRA	600	0	79	15.76	16.431
ACTIVO_VEH	600	0	926	70.73	162.032
ACTIVO_BIE	600	0	980	133.15	248.627
MAXMORA_HI	600	0	1819	38.24	156.704
Valid N (listwise)	600				

- *Créditos*

  - Nombre: Número de Créditos
  - Definición: Cantidad de créditos con los que ha contado el cliente anteriormente.
  - Tipo: Numérica (Cantidad de créditos)

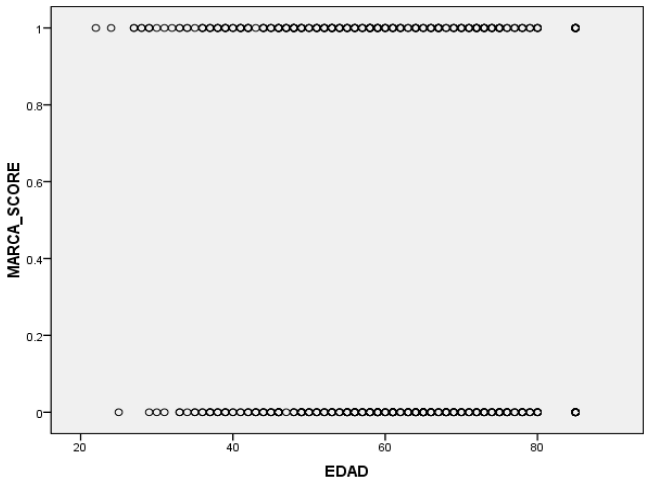
  El valor promedio y desviación estándar de la cantidad de créditos anteriores que han tenido los clientes es de 4. La cantidad mínima es 0 por aquellos que no han tenido créditos previos y el valor máximo en la muestra es de 27.



- *Edad*

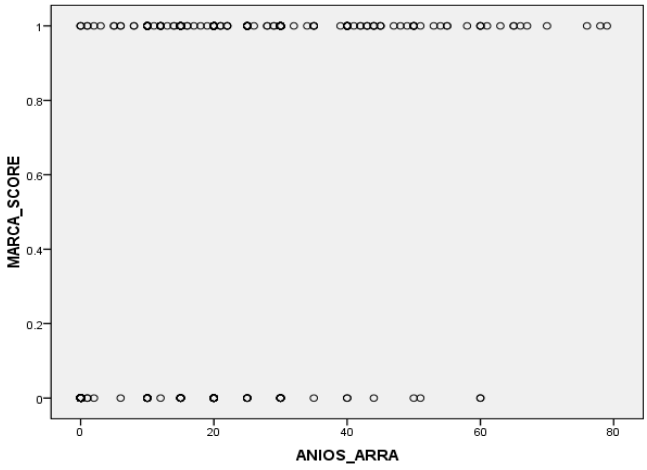
  - Nombre: Edad del cliente
  - Definición: Variable construida con la fecha de nacimiento por la cantidad de años que tiene el cliente.
  - Tipo: Numérica (Años)

El valor promedio de la edad es de 60, con una desviación de 14. El valor mínimo es de 22 y máximo de 85. A mayor edad el incumplimiento es menor por la mejoría en la capacidad de pago, gracias a una mayor estabilidad.



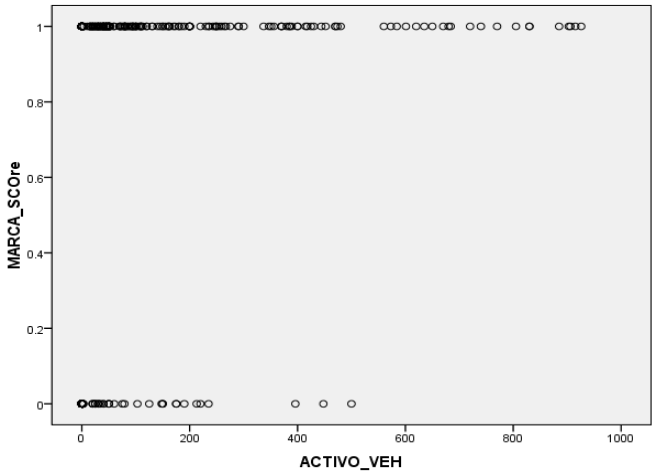
- *Anios_Arra*
  - Nombre: Años de arraigo
  - Definición: Cantidad de años que tiene un cliente de radicar en la zona actual.
  - Tipo: Numérica (Años)

El valor promedio de los años de arraigo es de 16, con una desviación de 16. Existe el valor mínimo de 0, por clientes que llegan a nuevas regiones y un máximo de 79. A mayor cantidad de años se puede ver que se tiene un mayor número de incumplimientos.



- *Activo_Veh*

  - Nombre: Valor de los activos de vehículos
  - Definición: Valor de los activos vehiculares de los clientes, en caso de existir.
  - Tipo: Numérica (Miles de pesos)

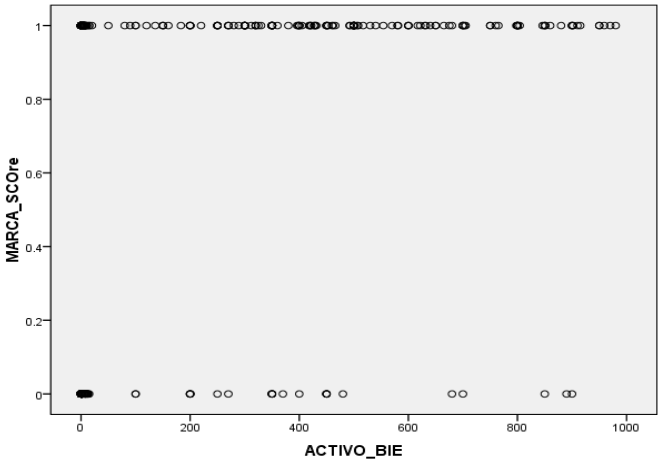


El valor promedio de los activos vehiculares de los clientes es de 71 mil pesos, con una desviación estándar alta de 162 mil. Existen aquellos que no cuentan con activos y cuyo valor es 0, así como el valor máximo de la muestra que es 926 mil pesos. Se sabe que los valores de activos vehiculares en esta población no suelen ser muy altos y que a menor valor de activos, existe una mayor concentración de incumplimientos.

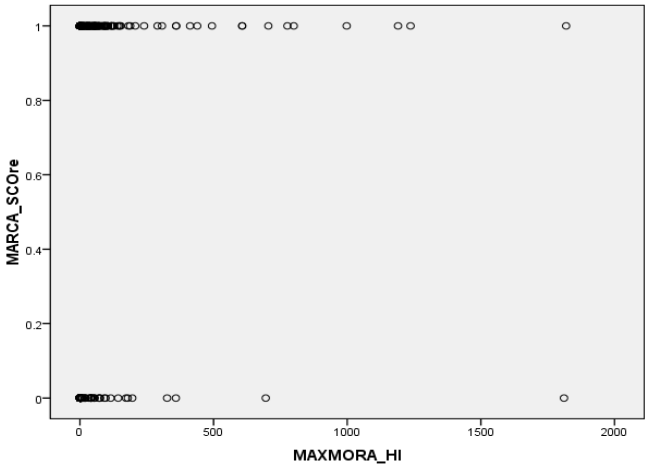
- *Activo_Bie*

  - Nombre: Valor de los activos de bienes
  - Definición: Variable de los activos asociados a bienes inmuebles, en caso de existir.
  - Tipo: Numérica (Miles de pesos)

El valor promedio de los activos asociados a bienes es de 133 mil pesos, con una desviación de 249 mil. El valor mínimo es 0 como en el caso anterior y el valor máximo es 980 mil pesos, más altos que los de los vehículos. La concentración con respecto a la cantidad de activos y sus incumplimientos en la población es similar a la variable anterior debido al nivel de pobreza.



- *MaxMora_Hi*
    - Nombre: Máximo de días históricos en mora
    - Definición: Variable del número máximo de días en mora que ha llegado a tener un cliente históricamente.
    - Tipo: Numérica



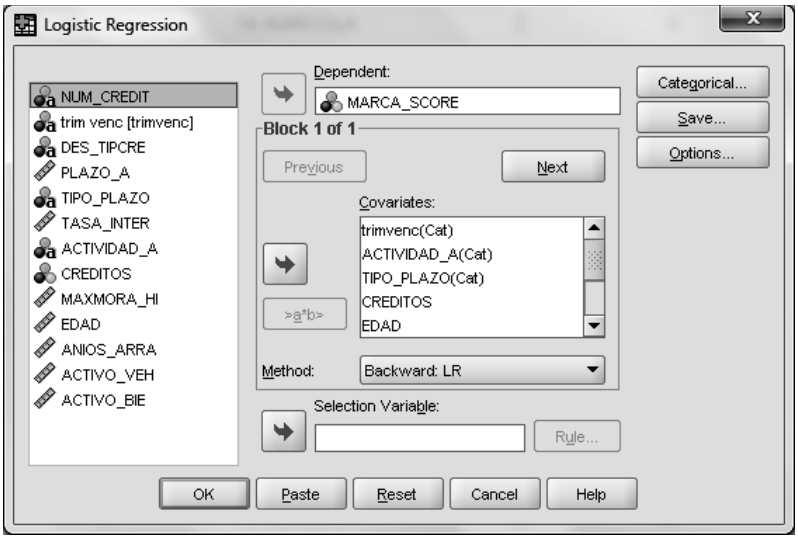
El valor promedio de los días máximos de mora es 38, con una desviación estándar alta debido a la diversidad en los clientes. El valor mínimo es 0, por aquellos que nunca han incumplido y el valor máximo es 1819, por aquellos clientes que han incumplido hasta 5 años. En esta variable a mayor número de días máximos aumentan los incumplimientos.

3.2.4. Estimación del modelo en SPSS

Ya que se cuenta con las variables analizadas y la base de datos depurada, se realiza la evaluación de la regresión logística en el programa SPSS, que en sus herramientas cuenta con este método de tipo binario para variables dependientes dicotómicas como es el caso.

Para la evaluación este programa asigna etiquetas a la variable dependiente, por lo cual los clientes incumplidos fueron etiquetados como "malos" con el valor de 1 y a los clientes cumplidos como "buenos" con el valor de 0. También se incluye la lista de variables independientes, que son todas las que han sido definidas, indicando cuáles son categóricas y cuáles son numéricas. Las variables categóricas suelen ser etiquetadas con un indicador de 1 y 0, manteniendo como referencia el último valor para las variables politómicas. Las variables independientes pueden incluirse en distintos *bloques* para comparar modelos. En esta ocasión se incluyeron sólo en un bloque, después de diversos intentos a prueba y error.

Se selecciona un método para la introducción de variables, que puede ser hacia adelante (forward), hacia atrás (backward) o de inclusión total (enter) donde se coloca la totalidad de variables independientes. Estos métodos pueden ser elegidos bajo dos criteros: Estadístico de Wald o Devianza *(LR).* Para el modelo de FR se seleccionó el método Backward: LR.



Se pueden solicitar resultados como la Bondad de ajuste de Hosmer-Lemeshhow, para evaluar los valores esperados y observados, tener un listado de los residuos y un historial de las iteraciones para observar el procedimiento seguido paso a paso, tanto de eliminación como de inclusión de variables, identificando la función de verosimilitud.

Finalmente se solicitan resultados de la matriz de correlaciones en cada paso y para la clasificación se elige el cut-off, que por default toma un valor de .5 pero puede ser modificado.

Resultados de la Regresión Logística en SPSS

- Resumen de casos

  Tabla para saber cuántos elementos de la muestra fueron utilizados y cuantos se perdieron por falta de datos, se presenta por número de casos y porcentaje. Para el modelo de aplicación de FR no se tuvieron casos perdidos.

Case Processing Summary

Unweighted Cases[a]		N	Percent
Selected Cases	Included in	600	100.0
	Missing Cases	0	.0
	Total	600	100.0
Unselected Cases		0	.0
Total		600	100.0

a. If weight is in effect, see classification table for the total number of cases.

- Codificación de variables

  Tabla que incluye los códigos de identificación de las variables categóricas independientes y sus frecuencias. Con esto se puede identificar en que valores de respuesta de las variables se basan los resultados y conocer las variables *dummies* generadas.

  En el ejemplo de aplicación se presentan las variables categóricas *Trimvenc, Actividad* y *Tipo_Plazo*. La variable *Trimvenc* por contar con 3 posibles valores de respuesta requirió de variables *dummies* y el resto de las variables fueron simplemente codificadas:

Categorical Variables Codings

		Frequency	Parameter coding (1)	Parameter coding (2)
Trimvenc	Tr1	113	1.000	.000
	Tr2	285	.000	1.000
	Tr3	202	.000	.000
TIPO_PLAZO	CP	503	1.000	
	LP	97	.000	
ACTIVIDAD_A	AGRICOLA	537	1.000	
	OTROS	63	.000	

- Prueba ómnibus

Es la prueba sobre los coeficientes, se basa en la hipótesis nula que supone que el vector de coeficientes (sin considerar la constante) es cero. El procedimiento se realiza mediante las tres diferencias en los cambios del logaritmo de la verosimilitud (Devianza) y una estadística comparada con la distribución Ji-Cuadrado:

1. Cambio en el último paso (Step): Significancia dada por la última inclusión o eliminación de variables. Al ser este un método backward se refiere a la eliminación.

2. Cambio en el último bloque (Block): Significancia dada por los distintos modelos generados en cada bloque. En este caso el bloque fue único.

3. Cambio del modelo (Model): Significancia dada por la comparación del modelo contra un modelo que contenga únicamente el valor constante.

Omnibus Tests of Model Coefficients

		Chi-square	df	Sig.
Step 1	Step	400.771	10	.000
	Block	400.771	10	.000
	Model	400.771	10	.000
Step 2[a]	Step	-.135	1	.714
	Block	400.636	9	.000
	Model	400.636	9	.000
Step 3[a]	Step	-.181	1	.670
	Block	400.455	8	.000
	Model	400.455	8	.000

a. A negative Chi-squares value indicates that the Chi-squares value has decreased from the previous step.

- Resumen del modelo

En esta tabla se pueden ver las iteraciones del modelo que van eliminando o agregando variables para crear diversos modelos y sus respectivos valores de logaritmo de la verosimilitud. Busca un valor pequeño de diferencia entre ellos (Devianza) para poder finalizar con el proceso e incluye los valores análogos a la R^2 del modelo de regresión lineal, que refieren la proporción de la variable dependiente explicada por las variables independientes:

- Cox & Snell R Square: Compara el logaritmo de verosimilitud para el modelo respecto al logaritmo de verosimilitud para el modelo base.
- Nagelkerke R Square: Versión corregida de Cox & Snell con valor máximo inferior a 1.

En el modelo se tuvieron 3 cambios y finalizó con 7 iteraciones. El último valor de determinación fue de .649, que para este tipo de modelo suele ser un valor ideal de explicación de la variable dependiente.

Model Summary

Step	-2 Log likelihood	Cox & Snell R Square	Nagelkerke R Square
1	431.006[a]	.487	.650
2	431.140[a]	.487	.650
3	431.322[a]	.487	.649

a. Estimation terminated at iteration number 7 because parameter estimates changed by less than .001.

- Bondad de ajuste Hosmer-Lemeshow

Se realiza una prueba Ji-cuadrada para evaluar la hipótesis nula de semejanza entre la probabilidad de los valores observados contra la probabilidad de los valores estimados en cada paso de cambio del modelo. Es una evaluación global del ajuste que especialmente en SPSS se busca que no sea un valor menor a .05.

En esta aplicación el resultado es significativo, sin embargo, no concluye el ajuste correcto o incorrecto debido a que es una prueba condicionada a los tamaños de muestra.

Hosmer and Lemeshow Test

Step	Chi-square	df	Sig.
1	21.635	8	.006
2	20.412	8	.009
3	17.948	8	.022

- Clasificación

El cuadro de clasificación indica paso a paso la clasificación de clientes cumplidos (buenos) e incumplidos (malos). Se puede ver el total de proporciones correctamente clasificadas en cada uno de los grupos.

Aquí se obtuvo un total de 85% de clasificaciones correctas con un punto de corte óptimo de .45 el cual pudo ser modificado, sin embargo, resultó óptimo por mantener una clasificación de clientes incumplidos cercana al 90%.

Classification Table[a]

Observed			Predicted		
			MARCA_SCORE		
			Buenos	Malos	Percentage Correct
Step 1	MARCA_SCORE	Buenos	240	60	80.0
		Malos	29	271	90.3
	Overall Percentage				85.2
Step 2	MARCA_SCORE	Buenos	240	60	80.0
		Malos	30	270	90.0
	Overall Percentage				85.0
Step 3	MARCA_SCORE	Buenos	241	59	80.3
		Malos	31	269	89.7
	Overall Percentage				85.0

a. The cut value is .450

Para comprobar que el punto de corte es óptimo se obtuvieron las clasificaciones en caso de que este punto tuviera valores alternativos. La especificidad y la sensibilidad son útiles al momento de este cálculo, ya que muestra las proporciones de clasificación.

Cut - Off	Pasos	Sensibilidad	Especificidad	1 - Especificidad
0.5	Paso 1	82.7%	86.3%	13.7%
	Paso 2	82.7%	86.3%	13.7%
	Paso 3	82.7%	86.7%	13.3%
Cut - Off	**Pasos**	**Sensibilidad**	**Especificidad**	**1 - Especificidad**
0.45	Paso 1	80.0%	90.3%	9.7%
	Paso 2	80.0%	90.0%	10.0%
	Paso 3	80.3%	89.7%	10.3%
Cut - Off	**Pasos**	**Sensibilidad**	**Especificidad**	**1 - Especificidad**
0.4	Paso 1	76.3%	91.0%	9.0%
	Paso 2	76.7%	92.0%	8.0%
	Paso 3	77.0%	91.7%	8.3%

Es claro que si se buscara aumentar la proporción de clasificación correcta de los clientes incumplidos (Especificidad por arriba del 90%) para evitar la mayoría de los incumplimientos, sin importar las posibles ganancias que no serían tomadas por la proporción de clasificación correcta de clientes cumplidos (sensibilidad por debajo del 80%), el cut-off de 0.4 sería el valor óptimo.

Si por otro lado se buscara arriesgar un poco la cantidad de incumplimientos (especificidad menor al 90%) para tomar la mayor cantidad de oportunidades posibles y una sensibilidad mayor al 80% para la correcta clasificación de clientes cumplidos, el valor de 0.5 sería adecuado. Sin embargo, lo ideal es conseguir lo más cercano posible a un punto medio, para minimizar la proporción de pérdidas en ambos casos y dando prioridad a mantener la menor cantidad de clientes incumplidos clasificados incorrectamente.

- Variables en la ecuación

Paso a paso se muestran las variables presentes en la ecuación, con sus respectivos parámetros estimados (*B*), error (*S.E.*), estadístico de Wald y su significancia, los valores de interpretación llamados OR (*Exp(b)*) y el intervalo de confianza al 95% para ubicar el rango de los coeficientes.

En este caso después de los 3 pasos se muestra la presencia de las variables independientes significativas: Trimvenc, Actividad, Tipo_Plazo, Edad, Anios_Arra, Activo_Veh, y

Activo_Bie. Se comprobó que la variable Tipo Plazo, a pesar de rebasar el valor de .05 óptimo, es útil para la explicación de la variable dependiente.

Variables in the Equation

		B	S.E.	Wald	df	Sig.	Exp(B)
Step 1[a]	trimvenc			24.326	2	0.00001	
	trimvenc(1)	2.124	0.474	20.124	1	0.00001	8.36618
	trimvenc(2)	1.984	0.426	21.737	1	0.00000	7.27138
	ACTIVIDAD_A(1)	-1.475	0.522	7.981	1	0.00473	0.22880
	TIPO_PLAZO(1)	1.549	0.827	3.505	1	0.06119	4.70720
	CREDITOS	-0.014	0.037	0.136	1	0.71276	0.98658
	EDAD	-0.020	0.009	4.566	1	0.03260	0.98062
	ANIOS_ARRA	0.055	0.010	29.697	1	0.00000	1.05703
	ACTIVO_VEH	0.005	0.001	16.365	1	0.00005	1.00497
	ACTIVO_BIE	0.003	0.001	22.997	1	0.00000	1.00286
	MAXMORA_HI	0.000	0.001	0.183	1	0.66923	1.00032
	Constant	-1.914	0.970	3.894	1	0.04845	0.14755
Step 2[a]	trimvenc			24.740	2	0.00000	
	trimvenc(1)	2.085	0.462	20.383	1	0.00001	8.04591
	trimvenc(2)	1.951	0.417	21.891	1	0.00000	7.03916
	ACTIVIDAD_A(1)	-1.446	0.515	7.884	1	0.00499	0.23558
	TIPO_PLAZO(1)	1.515	0.817	3.442	1	0.06356	4.54990
	EDAD	-0.020	0.009	4.559	1	0.03275	0.98063
	ANIOS_ARRA	0.055	0.010	30.666	1	0.00000	1.05612
	ACTIVO_VEH	0.005	0.001	16.285	1	0.00005	1.00492
	ACTIVO_BIE	0.003	0.001	22.813	1	0.00000	1.00285
	MAXMORA_HI	0.000	0.001	0.176	1	0.67498	1.00031
	Constant	-1.927	0.964	3.995	1	0.04565	0.14557
Step 3[a]	trimvenc			24.934	2	0.00000	
	trimvenc(1)	2.094	0.461	20.630	1	0.00001	8.11694
	trimvenc(2)	1.956	0.417	21.992	1	0.00000	7.06955
	ACTIVIDAD_A(1)	-1.472	0.511	8.280	1	0.00401	0.22952
	TIPO_PLAZO(1)	1.525	0.818	3.475	1	0.06232	4.59513
	EDAD	-0.020	0.009	4.962	1	0.02592	0.98005
	ANIOS_ARRA	0.055	0.010	30.759	1	0.00000	1.05621
	ACTIVO_VEH	0.005	0.001	16.291	1	0.00005	1.00492
	ACTIVO_BIE	0.003	0.001	22.867	1	0.00000	1.00285
	Constant	-1.870	0.956	3.824	1	0.05051	0.15405

a. Variable(s) entered on step 1: trimvenc, ACTIVIDAD_A, TIPO_PLAZO, CREDITOS, EDAD, ANIOS_ARRA, ACTIVO_VEH, ACTIVO_BIE, MAXMORA_HI.

- Variables fuera de la ecuación

Por el contrario al cuadro anterior, éste muestra el valor score de cada variable no incluida en el modelo y su significancia, con el fin de corroborar las acciones del programa de eliminación de variables. Para el modelo de la población rural las variables eliminadas no eran variables significativas al rebasar de manera importante el valor .05 óptimo.

Variables not in the Equation

			Score	df	Sig.
Step 2[a]	Variables	CREDITOS	.136	1	.713
	Overall Statistics		.136	1	.713
Step 3[b]	Variables	CREDITOS	.128	1	.720
		MAXMORA_HI	.178	1	.673
	Overall Statistics		.312	2	.855

a. Variable(s) removed on step 2: CREDITOS.

b. Variable(s) removed on step 3: MAXMORA_HI.

- Correlación

En el último cuadro se muestra la matriz de correlación formada que concluye para este estudio que las correlaciones finales son positivas y negativas pero lo suficientemente débiles como para mantener el modelo dentro de un buen ajuste.

Correlation Matrix

		Constant	trimvenc (1)	trimvenc (2)	ACTIVIDAD_A(1)	TIPO PLAZO(1)	CREDITOS	EDAD	ANIOS ARRA	ACTIVO VEH	ACTIVO BIE	MAXMORA HI
Step 1	Constant	1.000	-.074	.016	-.285	-.718	-.042	-.396	.018	-.128	-.027	-.133
	trimvenc(1)	-.074	1.000	.724	-.210	-.067	-.223	-.006	-.133	.034	-.055	-.034
	trimvenc(2)	.016	.724	1.000	-.333	-.144	-.208	-.051	-.186	.185	-.062	-.016
	ACTIVIDAD_A(1)	-.285	-.210	-.333	1.000	-.042	.160	-.190	.077	.020	.010	.109
	TIPO_PLAZO(1)	-.718	-.067	-.144	-.042	1.000	-.112	-.042	-.019	-.014	.070	-.024
	CREDITOS	-.042	-.223	-.208	.160	-.112	1.000	.005	-.240	-.094	-.068	-.025
	EDAD	-.396	-.006	-.051	-.190	-.042	.005	1.000	-.162	.099	-.096	.146
	ANIOS_ARRA	.018	-.133	-.186	.077	-.019	-.240	-.162	1.000	-.133	-.087	-.007
	ACTIVO_VEH	-.128	.034	.185	.020	-.014	-.094	.099	-.133	1.000	.020	.008
	ACTIVO_BIE	-.027	-.055	-.062	.010	.070	-.068	-.096	-.087	.020	1.000	.003
	MAXMORA_HI	-.133	-.034	-.016	.109	-.024	-.025	.146	-.007	.008	.003	1.000
Step 2	Constant	1.000	-.088	.007	-.284	-.724		-.401	.010	-.134	-.029	-.136
	trimvenc(1)	-.088	1.000	.711	-.176	-.097		-.006	-.195	.013	-.074	-.041
	trimvenc(2)	.007	.711	1.000	-.308	-.175		-.053	-.247	.169	-.079	-.022
	ACTIVIDAD_A(1)	-.284	-.176	-.308	1.000	-.027		-.191	.119	.036	.020	.115
	TIPO_PLAZO(1)	-.724	-.097	-.175	-.027	1.000		-.041	-.048	-.025	.062	-.028
	EDAD	-.401	-.006	-.053	-.191	-.041		1.000	-.167	.100	-.095	.148
	ANIOS_ARRA	.010	-.195	-.247	.119	-.048		-.167	1.000	-.159	-.106	-.013
	ACTIVO_VEH	-.134	.013	.169	.036	-.025		.100	-.159	1.000	.015	.007
	ACTIVO_BIE	-.029	-.074	-.079	.020	.062		-.095	-.106	.015	1.000	.002
	MAXMORA_HI	-.136	-.041	-.022	.115	-.028		.148	-.013	.007	.002	1.000
Step 3	Constant	1.000	-.095	.004	-.273	-.736		-.386	.007	-.133	-.028	
	trimvenc(1)	-.095	1.000	.712	-.171	-.098		.001	-.196	.013	-.074	
	trimvenc(2)	.004	.712	1.000	-.308	-.175		-.050	-.248	.169	-.078	
	ACTIVIDAD_A(1)	-.273	-.171	-.308	1.000	-.023		-.212	.122	.033	.022	
	TIPO_PLAZO(1)	-.736	-.098	-.175	-.023	1.000		-.038	-.048	-.024	.061	
	EDAD	-.386	.001	-.050	-.212	-.038		1.000	-.166	.100	-.099	
	ANIOS_ARRA	.007	-.196	-.248	.122	-.048		-.166	1.000	-.159	-.106	
	ACTIVO_VEH	-.133	.013	.169	.033	-.024		.100	-.159	1.000	.015	
	ACTIVO_BIE	-.028	-.074	-.078	.022	.061		-.099	-.106	.015	1.000	

3.2.5. Interpretación y uso del Credit Scoring

Finalmente es posible realizar el cálculo de la probabilidad de incumplimiento mediante la ecuación de regresión logística y los valores estimados de sus coeficientes, los cuales quedan definidos conforme a códigos específicos, con el fin de simplificar la notación:

VARIABLE	CÓDIGO	COEFICIENTES (c)
Trimvenc1	**Tr1**	
Trimvenc2	**Tr2**	2.0940
Trimvenc3	**Tr3**	1.9558
ACTIVIDAD_	**Act**	-1.4718
TIPO_PLAZO	**Pla**	1.5250
EDAD	**Ed**	-0.0202
ANIOS_ARRA	**An**	0.0547
ACTIVO_VEH	**Ave**	0.0049
ACTIVO_BIE	**Abi**	0.0028
CONSTANT	**Const**	-1.8704

$$\boldsymbol{\pi(X)} = \frac{1}{1 + e^{-1.87 + 2.094\,Tr_2 + 1.956\,Tr_3 - 1.472\,Act + 1.525\,Pla - .02\,Ed + .055\,An + .005\,Ave + .003\,Abi}}$$

Este es el elemento primordial para clasificar a los clientes, como se presenta en la próxima tabla, donde se toma una parte de la muestra, con 24 clientes para corroborar sus datos (características), su probabilidad y clasificación, tanto estimada como real:

NUM_CREDIT	Trimvenc	Tipo Plazo	Actividad	Créditos	MaxMora Hi	Edad	Años Arraigo	Activo Veh	Activo Bie	Marca Score	Marca Score (PROBABILIDAD ESTIMADA)	Marca Score (CLASIFICACIÓN ESTIMADA)
052010009860000014	tr3	CP	AGRICOLA	1	0	63	0	0	0	0	0.244	0
101400009710000001	tr3	LP	AGRICOLA	0	0	61	0	0	0	0	0.068	0
101500001660000002	tr1	CP	OTROS	11	30	59	25	40	1	1	0.508	1
102020004770000010	tr1	CP	AGRICOLA	13	52	66	15	255	3	1	0.256	0
102500002590000003	tr1	CP	OTROS	12	57	59	25	830	1	1	0.980	1
102600003770000001	tr2	CP	OTROS	10	0	39	15	250	2	1	0.953	1
102600003790000002	tr1	CP	OTROS	10	35	47	20	50	1	1	0.512	1
102600004290000001	tr3	CP	OTROS	4	1237	27	22	109	915	1	0.996	1
102600004530000001	tr1	CP	AGRICOLA	5	360	44	6	684	1	1	0.728	1
102600007520000001	tr1	CP	AGRICOLA	9	146	49	20	480	1	1	0.656	1
103010003740000018	tr3	CP	OTROS	2	172	45	0	0	0	0	0.669	1
103030002620000025	tr3	CP	AGRICOLA	0	0	44	0	0	0	0	0.321	0

NUM_CREDIT	Trimvenc	Tipo Plazo	Actividad	Créditos	MaxMora Hi	Edad	Años Arraigo	Activo Veh	Activo Bie	Marca Score	Marca Score (PROBABILIDAD ESTIMADA)	Marca Score (CLASIFICACIÓN ESTIMADA)
103300000090000001	tr3	CP	AGRICOLA	0	0	56	0	0	0	0	0.271	0
10350000327000003	tr3	CP	OTROS	9	1	49	30	50	1	1	0.925	1
103600001830000014	tr2	CP	OTROS	5	30	53	30	885	1	1	0.999	1
106400000800000002	tr3	LP	AGRICOLA	6	0	70	0	0	0	0	0.057	0
110400004130000002	tr3	LP	AGRICOLA	0	0	37	0	0	0	0	0.106	0
110400004960000003	tr3	LP	AGRICOLA	0	0	50	0	0	0	0	0.084	0
111400000620000014	tr3	LP	AGRICOLA	0	0	59	0	0	0	0	0.071	0
112030001060000114	tr3	CP	AGRICOLA	0	0	60	0	0	0	0	0.255	0
112400043310000001	tr3	LP	AGRICOLA	0	0	37	0	0	0	0	0.106	0
112500006570000003	tr1	CP	AGRICOLA	5	45	65	40	635	1	1	0.898	1
112700000130000002	tr2	CP	OTROS	2	4	61	20	235	1	1	0.941	1
114400006980000001	tr3	LP	AGRICOLA	0	0	36	0	0	0	0	0.108	0

Este procedimiento es a su vez con el que se clasifican a otros clientes para realizar la validación del modelo (*Backtesting*), lo cual debe hacerse con cierta frecuencia con el fin de ver que las variables siguen siendo significativas para esta predicción y que se eviten conflictos a futuro por los cambios que pueden irse dando en la población.

Para esta aplicación ha sido posible tomar una nueva muestra de 240 clientes (externos a la primera muestra) con el fin de hacer el Backtesting para probar y validar la utilidad del credit scoring, como en los casos de ejemplo que se encuentran en el próximo cuadro:

NUM_CREDIT	trim venc	TIPO PLAZO	ACTIVIDAD_	CREDITOS	MAXMORA HI	Edad	ANIOS ARRA	ACTIVO VEH	ACTIVO BIE	Marca Score	Marca Score (PROBABILIDAD ESTIMADA)	Marca Score (CLASIFICACIÓN ESTIMADA)
101400008690000001	tr3	LP	AGRICOLA	0	0	63	0	0	0	0	0.066	0
101400009020000001	tr3	LP	AGRICOLA	0	0	32	0	0	0	0	0.116	0
101400010610000001	tr3	LP	AGRICOLA	0	0	77	0	0	0	0	0.050	0
101700000330000004	tr1	CP	OTRO	10	6	57	20	80	3	1	0.500	1
102020001560000016	tr1	CP	AGRICOLA	13	97	28	5	40	225	1	0.219	0
102020002220000054	tr3	CP	AGRICOLA	0	0	51	0	0	0	0	0.291	0
102020004750000014	tr1	CP	AGRICOLA	10	32	53	25	55	370	1	0.451	1
102500001550000003	tr1	CP	OTRO	22	40	60	20	95	2	1	0.503	1
102500002750000001	tr1	CP	OTRO	2	27	49	25	185	2	1	0.721	1
102500003170000002	tr3	CP	OTRO	16	1182	46	35	321	1	1	0.985	1
102500007330000003	tr2	CP	OTRO	10	0	52	40	100	7	1	0.968	1
103300000560000001	tr3	CP	AGRICOLA	0	0	62	0	0	0	0	0.248	0
103500000070000013	tr1	CP	AGRICOLA	8	28	49	0	80	700	1	0.396	0
103500004890000002	tr2	CP	OTRO	8	1	63	40	714	47	1	0.998	1
103500006450000005	tr3	CP	OTRO	12	7	47	25	38	900	1	0.992	1
103600004530000005	tr3	CP	AGRICOLA	8	20	30	25	310	1	1	0.919	1
103700001000000009	tr1	CP	OTRO	11	3	52	40	260	2	1	0.889	1
104500015970000002	tr2	CP	AGRICOLA	33	23	41	10	0	0	1	0.499	1
106500000790000015	tr1	CP	AGRICOLA	4	5	85	77	0	1	1	0.665	1
106500002450000010	tr1	CP	AGRICOLA	10	0	82	74	123	2	1	0.766	1
106500007470000001	tr2	CP	AGRICOLA	11	0	72	30	85	3	1	0.709	1
106500017750000002	tr1	CP	AGRICOLA	18	4	67	61	20	915	1	0.946	1
106500025670000001	tr1	CP	AGRICOLA	7	0	64	50	165	4	1	0.610	1
106500027010000001	tr2	CP	AGRICOLA	12	133	55	54	24	2	1	0.904	1

La clasificación de esta nueva muestra (con el mismo punto de corte) comprueba el ajuste del modelo con la realidad, a través de los porcentajes de sensibilidad y especificidad que sólo varían en un 5% y 2% respectivamente con la muestra anterior, así como con el porcentaje de clasificación correcta con un nivel de 84%, que varía sólo 1% con respecto al modelo original.

Backtesting	Estimaciones			Clasificación Correcta	
Observaciones	**Buenos**	**Malos**	**Total**		
Buenos	47	16	**63**	Sensibilidad	**75%**
Malos	17	120	**137**	Especificidad	**88%**
Total general	**64**	**136**	**200**		**84%**

Con ello se puede también generar la *scorecard* para los clientes nuevos, con el fin de determinar si serán cumplidos o incumplidos, como se esperaba conseguir en un principio:

SCORECARD			
Variable	Coeficiente	Valor de Respuesta (Cliente)	
CONSTANTE	-1.870		
Trimvenc1	0.000	0	(Cuatrimestre 1 = 1, Otro = 0)
Trimvenc2	2.094	1	(Cuatrimestre 2 = 1, Otro = 0)
Trimvenc3	1.956	0	(Cuatrimestre 3 = 1, Otro = 0)
ACTIVIDAD_	-1.472	1	(Agrícola = 1, Otro = 0)
TIPO_PLAZO	1.525	0	(Corto Plazo = 1, Largo Plazo = 0)
EDAD	-0.020	45	(Años)
ANIOS_ARRA	0.055	20	(Años)
ACTIVO_VEH	0.005	30	(Miles de pesos)
ACTIVO_BIE	0.003	60	(Miles de pesos)
TOTAL Y	-0.891		
TOTAL P(X)	0.291		
CLASIFICACIÓN	0 = Cumplido		

SCORECARD			
Variable	Coeficiente	Valor de Respuesta (Cliente)	
CONSTANTE	-1.870		
Trimvenc1	0.000	0	(Cuatrimestre 1 = 1, Otro = 0)
Trimvenc2	2.094	0	(Cuatrimestre 2 = 1, Otro = 0)
Trimvenc3	1.956	1	(Cuatrimestre 3 = 1, Otro = 0)
ACTIVIDAD_	-1.472	0	(Agrícola = 1, Otro = 0)
TIPO_PLAZO	1.525	1	(Corto Plazo = 1, Largo Plazo = 0)
EDAD	-0.020	28	(Años)
ANIOS_ARRA	0.055	5	(Años)
ACTIVO_VEH	0.005	10	(Miles de pesos)
ACTIVO_BIE	0.003	0	(Miles de pesos)
TOTAL Y	1.319		
TOTAL P(X)	0.789		
CLASIFICACIÓN	1 = Incumplido		

Ahora establecido el credit scoring, la capacidad de clasificación y la estimación de probabilidad, se busca identificar los factores de riesgo y el cambio en la probabilidad con el cambio de una unidad en cada una de las variables (valores de OR), los cuales se localizan en la próxima tabla, recordando que dichas razones en caso de aplicar el logaritmo natural coinciden con el valor de los coeficientes de la regresión.

VARIABLE	COEFICIENTES (c)	OR = Exp (c)	Coef = ln (OR)
Trimvenc1			
Trimvenc2	2.0940	8.1169	2.0940
Trimvenc3	1.9558	7.0696	1.9558
ACTIVIDAD_	-1.4718	0.2295	-1.4718
TIPO_PLAZO	1.5250	4.5951	1.5250
EDAD	-0.0202	0.9800	-0.0202
ANIOS_ARRA	0.0547	1.0562	0.0547
ACTIVO_VEH	0.0049	1.0049	0.0049
ACTIVO_BIE	0.0028	1.0028	0.0028
CONSTANT	-1.8704	0.1541	-1.8704

Con esto sabemos que la variable *Trimvenc* (Variable que se evalúa con variables dummies para indicar el cuatrimestre de vencimiento de los créditos) es sin duda aquella que más poder tiene al momento de la evaluación, sobre todo al tratarse de clientes cuya fecha de vencimiento cae en el segundo cuatrimestre, el cual tiene un aumento de 8.1169 veces en la probabilidad de incumplimiento. A esta variable le sigue en términos de relevancia *Tipo_ plazo*, que aumenta la probabilidad en 4.5951 veces para los créditos de corto plazo, mientras que la variable *Actividad* es la que menos cambios aporta en todo el modelo, ya que con la presencia de un crédito agrícola el riesgo será .2295 veces menor.

Es entonces cuando se ha conseguido llegar al final del presente estudio, contando con el conocimiento y la identificación de algunos de los factores de riesgo y los respectivos rangos que manejan entre los clientes de esta entidad, permitiendo entender aquellos aspectos donde se podrían otorgar apoyos y mejoras para disminuir los incumplimientos, dando una referencia consistente de la decisión que debe tomarse al otorgar o rechazar el crédito de cualquier solicitante del sector y calculando una probabilidad de incumplimiento que es necesaria para el cálculo de la pérdida esperada, el siguiente paso en los riesgos de crédito.

CONCLUSIONES

Ya que ha sido posible entender que en la actualidad es indispensable contar con un amplio conocimiento de los riesgos y las técnicas de medición para la mejora en la operación crediticia, conforme al análisis realizado se logró implementar el modelo de regresión logística en credit scoring, dando a conocer el método, la forma en que se construye y logrando identificar sus ventajas sobre el modelo de regresión lineal, el análisis discriminante, el análisis clúster y los árboles de decisión. Dichas ventajas están dadas por el uso de variables categóricas y el cálculo directo de la probabilidad de incumplimiento, sin requerir supuestos de normalidad.

Como parte del estudio se hizo hincapié en que es inevitable la necesidad de evaluar continuamente el credit scoring con el fin de confirmar su correcto ajuste con los valores reales mediante un backtesting y la evaluación del tipo de variables que pueden utilizarse, mismas que se encuentran englobadas en el método experto que considera el carácter, la capacidad y el capital del cliente, las condiciones del crédito y los aspectos colaterales.

A través de esta técnica se logro construir una aplicación al sector rural en México con el uso de las variables: Tipo de plazo, cuatrimestre de vencimiento y actividad de destino del crédito; Edad, años de arraigo, valor de activos, días máximos de mora y cantidad de créditos anteriores del cliente. Donde tipo de plazo, cuatrimestre de vencimiento y actividad de destino del crédito son variables categóricas que el programa SPSS consideró como variables dummies.

Se utilizó una muestra de 600 registros de personas físicas con créditos del año 2002 al año 2011, encontrando mediante el criterio de eliminación (Backward) que de las 8 variables únicamente 6 resultaron significativas, descartando los días máximos de mora y la cantidad de créditos anteriores del cliente como factores de riesgo para la evaluación de la probabilidad de incumplimiento. El punto elegido de corte óptimo fue analizado con base en la sensibilidad y especificidad, validado mediante un backtesting de 200 observaciones.

Finalmente se logró generar la scorecard y se incluyó un ejemplo para cada tipo de cliente, analizando los resultados para comprender en qué casos se tiene mayor riesgo para esta aplicación, dando por concluida la construcción del scoring y del cálculo de la probabilidad de incumplimiento.

BIBLIOGRAFÍA

I. Anaya Mora, Miguel Luis. *La Banca de Desarrollo en México*. Chile. CEPAL: 2007.

II. Anda Gutiérrez, Cuauhtémoc. *La Nueva Banca Mexicana*. México. S. Ed.: 1992.

III. Ardila, Sergio. *El Sector Rural en México: Desafíos y Oportunidades*. Nota de política del Banco Interamericano de Desarrollo. Departamento Regional de Operaciones II. División de Recursos Naturales y Medio Ambiente. 2006

IV. Banco de México. *Definiciones Básicas de Riesgos*. México 2005 [Recurso en línea]

V. Banco de México. *Divulgación. Sistema Financiero: 4.2 Servicios de Crédito*. [Recurso en línea]

VI. Cramer, J.S. *The Logit Model: An introduction for economists*. London. E.Arnold: 1991.

VII. De Lara Haro, Alfonso. *Medición y Control de Riesgos Financieros*. México. Editorial Limusa: 2005.

VIII. Finance & Leasing Association. *Guide to Credit Scoring*. United Kingdom 2000

IX. Fuentes oficiales para datos estadísticos y demográficos: INEGI y SHCP.

X. Horche, Karen A. *Essentials of Financial Risk Management*. USA. John Wiley & Sons: Aug 2011.

XI. Hosmer, David W.; Lemeshow, Stanley. *Applied logistic regression*. USA. Wiley & Sons: 2000.

XII. Hugo. H. Salinas. Departamento de Matemáticas, Facultad de Ingeniería. Universidad de Atacama. Chile 2012. [Recurso en línea]

XIII. Johnson, Richard A.; Bhattacharyya, Gouri K. *Statistics: Principles and Method.* USA. John Wiley & Sons: 2009.

XIV. Kleinbaum, David G. *Logistic regression: A self-learning text*. USA. Springer V: 2002.

XV. Márquez, Javier. *An Introduction to Credit Scoring for Small and Medium Size Enterprises*. USA, Febrero 2008. [Recurso en línea].

XVI. McLachlan, Geoffrey J. *Discriminant analysis and statistical pattern recognition*. Hoboken, New Jersey : Wiley-Interscience, 2004.

XVII. Mirkin, Boris. *Clústering: A Data Recovery Approach*. USA. CRC Press: 2012.

XVIII. Nieto Murillo, Soraida; Pérez Salvador, Blanca Rosa; Soriano Flores, José Fernando. *Crédito al Consumo: La estadística aplicada a un problema de riesgo crediticio.* México. Colegio Nacional de Actuarios: 2011.

XIX. Notas de clase Act. Francisco Sánchez Villareal (Análisis Econométrico).

XX. Santandreu, Eliseu. *Manual del Credit Manager*. Barcelona. Gestión 2000: año 2002.

XXI. SHCP. *Apartado Hacienda para Todos: Banca de Desarrollo*. [Recurso en línea 2012]

XXII. Thomas, Lyn C.; Edelman, David B.; Crook, Jonathan N. *Credit Scoring and Its Applications*. USA. SIAM: 2002.

XXIII. Vieira Braga, Luis Paulo; Ortiz Valencia, Luis Iván; Ramirez Carvajal, Santiago Segundo. *Introducción a la Minería de Datos.* Río de Janeiro. E-papers: 2009.

XXIV. Villa Señor Fuente, Emilio. *Elementos de Administración de Crédito y Cobranza*. México. Ed. Trillas: 2ª Ed. 1985