FLACSO

FACULTAD LATINOAMERICANA DE CIENCIAS SOCIALES

MAESTRIA EN ECONOMÍA

Prestar como locos y obtener beneficios: ¿Es realmente posible?

(Un análisis logit multinomial para los determinantes del comportamiento de pago de una cartera de consumo)

Carlos Bambino Contreras

Quito, Agosto de 2005

FLACSO

FACULTAD LATINOAMERICANA DE CIENCIAS SOCIALES

MAESTRIA EN ECONOMÍA

Prestar como locos y obtener beneficios: ¿Es realmente posible?

(Un análisis logit multinomial para los determinantes del comportamiento de pago de una cartera de consumo)

Carlos Bambino Contreras

Director: Mat. Enrique Navarrete

Quito, Agosto de 2005

Indice

Marcecachirula:
Por la luna hurtada a tus noches para darle luz a este proyecto; por tus paseos vacía; por la cena fría; por guardar el amor del yermo de mi apatía

Francisco y Alejandra:
Por no comprender y no juzgar; por esperarme con la risa; por combatir contra el sueño; por trocar, en desigual intercambio, cada día mi cansancio con sus besos

A la tripulación, los amigos:
Por haber sido infinitos cuando el mundo mío se volvía angosto; por sostenerme al timón; por dejarse embestir por mis ojos torvos; por haber sido la fe que necesitaba cuando el cielo se quedaba sordo.

Oh captain! my captain! our fearful trip is done;
the ship has weather'd every rack, the prize we sought is won; the port is near, the bells I hear, the people all exulting.

Walt Whitman

Síntesis

Bajo el esquema de concesión de créditos de la banca tradicional el objetivo es evitar cualquier pérdida (dado el plazo y el monto del crédito una pérdida involucra costos significativos para la institución). Sin embargo, la óptica de los préstamos de consumo es diferente. El punto no es evitar cualquier pérdida, que en términos marginales son insignificantes, sino buscar maximizar los beneficios, permitiéndose un pequeño y controlado nivel de cartera mala merced de una expansión de los créditos otorgados.

Empero, la ampliación de la cartera de clientes requiere de mecanismos que agiliten el proceso y que logren trabajar con un gran volumen de aplicaciones. Esto implica, bajo el método tradicional de aprobación de crédito, un vasto y bien entrenado grupo de analistas que logre oportunamente entrevistar a los solicitantes. Con este esquema no es difícil concluir que la oferta de crédito termine contrayéndose, reduciendo las oportunidades de incrementar ganancias, o que se reduzca la calidad del control y por ende se cometan errores de aprobación que involucran efectos similares sobre los beneficios.

Por otro lado, existe la posibilidad de desarrollar métodos que reduzcan el subjetivo criterio humano "gut feel" implícito en la aprobación tradicional de un crédito, que sean capaces de evaluar un conjunto de características en muy corto tiempo, que puedan procesar grandes cantidades de solicitudes en el menor tiempo posible y que emitan una medida de diferenciación entre buenos y malos solicitantes. Esta es precisamente la idea subyacente en la metodología 'scoring'. Básicamente el ***'credit scoring' es un método que se emplea para identificar diferentes grupos dentro de una población cuando no se pueden observar las características que los definen sino únicamente las relaciones con estas.***

Prestar como locos y obtener beneficios ¿Es realmente posible?

(Un análisis logit multinomial para los determinantes del comportamiento de pago de una cartera de consumo)

"We should not forget that the basic economic function of these regulated entities (banks) is to take risk. If we minimize risk taking in order to reduce failure rates to zero, we will, by definition, have eliminated the purpose of the banking system"

Alan Greenspan, Presidente de la Reserva Federal de E.E.U.U., Mayo de 1996

I.- Introducción

La aplicación estadística conocida como 'credit scoring', si bien carece del respaldo teórico de los modelos financieros, ha sido de vital importancia en los últimos 40 años, en especial en países como Estados Unidos o el Reino Unido, para la expansión del crédito de consumo[1]. Área en la cual, dado el volumen de operaciones, pequeñas mejoras en el otorgamiento de préstamos pueden representar para el prestamista cuantiosos incrementos en sus beneficios.

Esta metodología está compuesta por un conjunto de modelos de decisión que ayudan al prestamista en la concesión de créditos de consumo. Las técnicas subyacentes a estos modelos permiten decidir no sólo quién recibirá el dinero, sino cuánto se le debe entregar y cuál debe ser la estrategia operacional que mejorará la rentabilidad de los consumidores. No obstante, la técnica no puede asignar a un consumidor la categoría de 'sujeto de crédito', pues éste no es un atributo o característica inherente al individuo, como el peso, la altura o incluso el nivel de ingresos. La consideración 'sujeto de crédito' es una valoración del que acredita respecto al acreditado y refleja las circunstancias en las cuales ambos se encuentran, así como la percepción del primero sobre los futuros escenarios económicos.

Por lo tanto, no es saludable considerar 'no sujeto de crédito' a un consumidor cuyo perfil de riesgo no se acopla al requerido por la institución financiera. Resulta

[1] En realidad, la importancia de esta metodología es más amplia ('scoring' en términos más generales) extendiéndose a campos como el marketing, a sectores de ventas al por menor o incluso para microcréditos.

menos agraviante y refleja mejor el estado de la realidad decir que la solicitud de crédito del prestatario representa un riesgo que no se está dispuesto a asumir.

En términos de los requerimientos por etapas o fases del crédito[2] (nuevas aplicaciones para crédito o aplicaciones existentes), el 'credit scoring' deja al prestamista tomar dos tipos de decisiones[3]:

1.- entregar el crédito a un nuevo solicitante; y

2.- cómo manejar los créditos existentes, incluyendo la opción de mejorar sus límites crediticios.

Las técnicas que ayudan a tomar la primera decisión se conocen particularmente como 'credit scoring' o 'scoring de aprobación'. Las que asisten a la segunda se denominan 'behavioral scoring' o scoring de comportamiento'. Pero, independiente del tipo de técnica empleada, ambas demandarán una extensa muestra de clientes con el detalle de sus aplicaciones y la subsecuente disponibilidad de su historia crediticia. Es sobre esta muestra que se identifican las conexiones entre las características de los consumidores y cuán 'bueno' o 'malo' es su historial.

---

[2] Ver en este documento en el capítulo III el diagrama 'ciclo del crédito'.

[3] Mark Schreiner (2000), describe seis posibles modelos en función del riesgo que se quiera estimar 1) modelo para la probabilidad de que un préstamo vigente tenga un atraso **x** días o más; 2) modelo para la probabilidad de que un préstamo que lleva **x** días atrasado alcance eventualmente **y** días de mora; 3) modelo que estima la probabilidad de que un prestatario vigente, sin problemas de pago, opte por renovar una vez que ha repagado el crédito actual; 4) modelo para estimar el plazo de vencimiento esperado del próximo crédito de un prestatario vigente; 5) modelo para estimar el monto desembolsado esperado del próximo préstamo y 6) modelo de calificación que combina la información de los primeros cinco modelos con información sobre el ingreso esperado de un préstamo que posee un plazo de vencimiento y un monto de desembolso dados y con información sobre los costos esperados por deserciones, de las pérdidas de incumplimiento, y del seguimiento de los prestatarios en mora, para estimar en lugar del riesgo del cliente su rentabilidad.

II.- ¿Puede funcionar la calificación estadística en la banca de consumo? (Alcance del scoring de crédito)

Bajo el esquema de concesión de créditos de la banca tradicional el objetivo es evitar cualquier pérdida (dado el plazo y el monto del crédito una pérdida involucra costos significativos para la institución). Sin embargo, la óptica de los préstamos de consumo es diferente. El punto no es evitar cualquier pérdida, que en términos marginales son insignificantes, sino buscar maximizar los beneficios, permitiéndose un pequeño y controlado nivel de cartera mala merced de una expansión de los créditos otorgados.

Empero, la ampliación de la cartera de clientes requiere de mecanismos que agiliten el proceso y que logren trabajar con un gran volumen de aplicaciones. Esto implica, bajo el método tradicional de aprobación de crédito, un vasto y bien entrenado grupo de analistas que logre oportunamente entrevistar a los solicitantes. Con este esquema no es difícil concluir que la oferta de crédito termine contrayéndose, reduciendo las oportunidades de incrementar ganancias, o que se reduzca la calidad del control y por ende se cometan errores de aprobación que involucran efectos similares sobre los beneficios.

Por otro lado, existe la posibilidad de desarrollar métodos que reduzcan el subjetivo criterio humano "gut feel" implícito en la aprobación tradicional de un crédito, que sean capaces de evaluar un conjunto de características en muy corto tiempo, que puedan procesar grandes cantidades de solicitudes en el menor tiempo posible y que emitan una medida de diferenciación entre buenos y malos solicitantes. Esta es precisamente la idea subyacente en la metodología 'scoring'.

Básicamente el ***'credit scoring' es un método que se emplea para identificar diferentes grupos dentro de una población cuando no se pueden observar las características que los definen sino únicamente las relaciones con estas***[4].

Esto hace que el sistema se base en el desempeño pasado de clientes que tienen características similares a aquellos a ser evaluados. El 'credit scoring' termina, por tanto, siendo un predictor del riesgo, cuya fortaleza no radica en la habilidad para

[4] En términos más formales el 'credit scoring' es un método de evaluación del riesgo de crédito que emplea información histórica y técnicas estadísticas, para tratar de aislar los efectos que tienen las características de varios aplicantes sobre la delincuencia y el incumplimiento. El método arroja un puntaje o score que la institución financiera puede emplear para ranquear sus aplicaciones de crédito en términos de riesgo.

explicar causalidades (por qué algunos clientes incumplen y otros no) sino en la objetividad de la metodología aplicada.

Con el 'scoring' el prestamista aplica una fórmula para ciertos elementos claves de la aplicación de crédito. Este procedimiento arroja una cuantificación numérica[5] del riesgo y dependiendo de su valor la solicitud es aceptada o rechazada. De esta forma, el rol subjetivo en la valoración por parte del empleado se reduce a los casos donde sea genuina la oportunidad de añadir valor al proceso.

Los beneficios no sólo se leen en una reducción de la subjetividad del analista inmersa en la concesión del crédito, al estar los métodos tradicionales apoyados fundamentalmente en información cualitativa mantenida en la mente del evaluador, sino que al basarse en información cuantitativa mantenida en los sistemas de cómputo de la institución financiera y por ende cuantificables a bajo costo, se pueden lograr reducciones de costos de morosidad y de evaluaciones de préstamos de tal forma que se mejore la eficiencia (mejores colocaciones) y por ende la rentabilidad.

En Bolivia y Colombia, en el campo de las microfinanzas, experimentos de este tipo llevados a cabo por Schreiner[6], han confirmado las mejoras en la valoración del riesgo y por ende la disminución de costos. En Colombia se ha estimado un ahorro al prestamista de alrededor de $75.000 al año.

No obstante, si bien el scoring de crédito, aplicado a créditos de consumo o microcréditos, parece ser el siguiente salto a la eficiencia[7] no está exento de debilidades.

A la par coexisten tanto ventajas como desventajas en esta metodología, dejando siempre al final la opción al prestamista de optar por un sistema de otorgamiento de créditos subjetivo (scoring implícito) en donde el analista valora el riesgo del prestatario comparando mentalmente las características de un aplicante con su experiencia acumulada de otras aplicaciones similares o por un scoring estadístico que haga uso del conocimiento cuantitativo del desempeño de clientes pasados almacenados en una base

[5] Estrictamente no todos los métodos conducen a una 'puntuación' o 'scorecard'. Algunos indican directamente la posibilidad de que un cliente sea bueno y si la aprobación del crédito vale la pena.

[6] Mark Schreiner es Director de Microfinance Risk Managment e investigador del Center for Social Development de la Universidad de Washington en St.Louis

[7] La evolución de este tipo de crédito ha pasado de la no concesión, a la concesión por parte de bancas de desarrollo, en el caso de microcréditos, a la conformación de grupos solidarios o a evaluaciones minuciosas de las solicitudes de crédito, para actualmente situarse en modelos estadísticos que pueden ir desde sencillas herramientas como las funciones discriminantes a sistemas complejos como las redes neuronales o los algoritmos genéticos.

de datos con el fin de pronosticar comportamientos futuros. Esto conduce, en virtud de la seriedad académica, a exponer los pros y contras de hacer uso de esta técnica.

Ventajas	Desventajas
Cuantifica el riesgo como probabilidad	Requiere la existencia de una sólida y extensa base de datos
Consistencia	Puede necesitar un consultor externo y por ende generar riesgo operativo
Transparencia	Su funcionamiento depende de la implementación en el sistema transaccional
Multivariable	No trabaja sobre solicitudes rechazadas (ignora toda la información no existente en la base de datos de la institución)
Puede ser validado	Sólo destaca los casos de alto riesgo
Permite evaluar opciones de políticas de aprobación	Asume que el futuro será como el pasado
Establece relaciones entre riesgo y características del prestatario, el prestamista y el préstamo	Es un modelo estocástico no determinístico
Reduce tiempos de concesión de créditos	
Reduce tiempo de cobranza	
Permite estimar efectos sobre la rentabilidad	
Mayor efectividad que una calificación basada en cualquier "sistema experto[8]".	
Orienta los esfuerzos de venta hacia mercados más rentables	
Asigna recursos internos de una manera más eficaz	

[8] Método de asignación de ponderaciones de características basadas en la experiencia y en supuestos.

La primera ventaja y tal vez la más importante es la **cuantificación del riesgo como una probabilidad,**

> *"mientras el scoring subjetivo expresa que un préstamo tiene un riesgo por debajo del promedio, un juicio basado mayormente en un sentimiento cualitativo" (Schreiner, M. 2004)*[9].

La **consistencia** surge de la homogeneidad en la calificación otorgada a cada grupo de clientes que comparta las mismas características, todas las solicitudes idénticas tendrán por tanto el mismo valor de riesgo predicho; el hecho de contar con una fórmula matemática que arroje un valor π_i hace que el proceso de obtención de esta probabilidad sea explícito y como consecuencia de fácil comunicación tanto a nivel de alta gerencia como de analistas.

Uno de los riesgos de realizar **análisis univariantes** es el conocido como "Paradoja de Simpson"[10] según la cual la asociación entre dos variables (numéricas o cuantitativas) puede cambiar de sentido cuando se controla el efecto de una tercera variable, esto fortalece el uso de modelos estadísticos multivariantes, en donde se enmarca la metodología scoring, pudiendo realizar evaluaciones sobre riesgo de crédito con base en 30, 50, 100 o más características simultáneamente[11], permitiendo análisis mucho más precisos.

El contar con modelos estadísticos da la **posibilidad de contrastación** de los mismos, validando los modelos sea con información previa al periodo muestral empleado, permitiendo observar como habría funcionado scoring si hubiera estado implementado al momento de estos desembolsos o verificando, previo a la fase de implementación, la consistencia del modelo con los nuevos clientes. Esta fase suele denominarse **backtesting**.

La obtención de una calificación y de una distribución de clientes asociada a este puntaje permite **establecer límites de aceptación y determinar el porcentaje de potenciales clientes que estamos excluyendo** (la tasa de aprobación respecto al total de aplicaciones), por otro lado la opción de emplear información histórica permite calcular

---

[9] "Benefits and Pitfalls of Statistical Credit Scoring for Microfinance", Savings and Development, Vol. 28 No1, pp. 63-68.

[10] Ver anexo 1 "De estadística, paradojas y resoluciones financieras". Comentario realizado a una resolución emitida por el organismo regulador ecuatoriano respecto al riesgo de emplear análisis univariantes.

[11] Claro que siempre debe premiarse el principio de parsimonia.

el riesgo asumido con los clientes vigentes en la institución y el efecto que podría tener el modificar la política vigente sobre la relación buenos:malos y por ende sobre la reducción en el número de créditos desembolsados, estas posibilidades de evaluación dejan al tomador de decisiones conocer las consecuencias de implementar diferentes opciones de política.

A diferencia de las valoraciones subjetivas el método scoring deja conocer no sólo que característica tiene un comportamiento habitual de riesgo, por ejemplo tanto en créditos de consumo como en microfinanzas se encuentra que el género es una variable determinante del riesgo, premiando a las mujeres en relación a los hombres como mejores sujetos de crédito al cumplir estas históricamente mejor con sus obligaciones financieras, sino que también **permite saber en cuánto esta característica es más o menos riesgosa**. De esta forma y dependiendo de la base estadística disponible en la institución financiera se podrá no sólo confirmar la orientación respecto al riesgo dada a la variable con base en el juicio subjetivo del analista, sino cuantitativamente ver cuán fuerte o débil es esta relación.

En mercados competitivos donde los precios de los productos financieros no son mandatorios en las preferencias de los clientes es importante para el postulante a crédito el **tiempo de aprobación** del mismo (se supone que los clientes siempre preferirán la opción que les reporte el menor costo de oportunidad), por lo tanto el tiempo de respuesta de una institución financiera ante la demanda de crédito puede resultar un factor determinante al momento de querer colocar el dinero que intermedia. La solución no radica en la sola reducción de los días o las horas de aprobación, pues esto puede degenerar el riesgo como consecuencia de la reducción de controles, sino en la reducción del plazo de entrega del crédito con un correcto control del riesgo, este beneficio se incorpora al proceso de crédito con el uso de modelos estadísticos.

De la misma forma el **tiempo que puede gastar un analista de crédito en gestiones de cobranzas se reduce**, primero porque el otorgarle un puntaje a cada cliente reduce el número, monto, y plazo de los créditos entregados a clientes de alto riesgo, reduciéndose el número de veces que los créditos sufrirán atrasos y consecuentemente el tiempo en gestiones de cobranza; segundo porque una vez desembolsado el préstamo, todos los clientes que tienen una calificación que indica mayor probabilidad de incumplimiento podrían ser sujetos de estrategias preventivas de

cobro, como llamadas recordatorias, que permitan reforzar la presencia de la obligación financiera en la mente del prestatario[12]; tercero al dar calificaciones a los clientes se cuenta con un criterio de ordenación que permite a los analistas priorizar esfuerzos de cobranza, por ejemplo contactarse con aquellos de mayor riesgo de incumplimiento durante largo tiempo.

Con el conocimiento adicional por parte de la institución financiera del costo de un crédito malo y el beneficio de uno bueno **se puede determinar el impacto sobre la utilidad que tiene el fijar un umbral de créditos extremamente malos**. El cambio en la rentabilidad no es más que el resultado de la diferencia entre el número de créditos malos que se niegan multiplicados por el beneficio de evitar estos préstamos menos el número de préstamos buenos multiplicados por el costo por préstamo bueno perdido.

Muchas instituciones financieras asignan notas subjetivas a los préstamos en función de algunas variables, por ejemplo atraso promedio y atraso máximo es una pareja de variables comúnmente empleada para calificar al cliente como bueno o malo, el problema con este tipo de calificaciones es que suponen una relación entre los atrasos pasados y el riesgo futuro, mientras que **el modelo scoring deriva la relación histórica**. En el caso de los nuevos solicitantes que no disponen de un historial de pago las calificaciones basadas en sistemas expertos simplemente no aplican, mientras que los modelos scoring justamente permiten pronosticar el riesgo para este tipo de modelos.

Al **identificar y segmentar mejor el universo de clientes potenciales** permite: a).- focalizar las estrategias de venta hacia los clientes más rentables, b).- realizar ofertas diferenciadas con el fin de obtener incrementos en el nivel de respuesta y c).- desarrollar estrategias para mercados de aparentemente alto riesgo.

La asignación **interna de recursos de una manera más eficaz** pasa por la posibilidad de evaluación objetiva y consistente de la relación riesgo beneficio en la adquisición de nuevos clientes. El empleo del scoring permite optar entre dos decisiones: 1).- reducir el riesgo manteniendo un volumen fijo de aprobaciones, 2).- aumentar el número de clientes aprobados pero manteniendo el mismo nivel de riesgo.

---

[12] Esta gestión preventiva de cobranza debe reducir las caídas de clientes a etapas de cobranza posteriores que involucren mayores costos para la institución. En términos de análisis costo-beneficio debe esperarse que esta opción permita reducir costos de cobranza, de lo contrario optar por una estrategia de este estilo podría generar resultados no esperados.

Del otro lado, se tiene un conjunto de desventajas que se deben considerar antes de incurrir en un proceso de obtención e implementación de un modelo scoring para el otorgamiento de créditos. La **disponibilidad de una sólida y extensa información** es uno de los principales obstáculos al momento de desarrollar un modelo estadístico. Lewis (1992) sugiere que una base que cuente con 1500 buenos y 1500 malos puede considerarse suficiente para modelar el comportamiento de pago. No obstante, mientras mayor sea el número de clientes mejor pueden hablar los datos. Makuch (1999) en cambio, sugiere que 100 mil clientes buenos se pueden considerar información suficiente sobre estos, en este caso se podría formar una base de análisis con todos los clientes malos existentes y tomar 100 mil buenos. Además, la cantidad de información registrada no debe sólo ser buena sino vasta[13] para que los modelos desarrollados puedan compensar la ausencia de información financiera con mucha capacidad de pronóstico empleando un gran número de características menos significativas. La calidad de las bases de datos es fundamental, no sólo en cantidad de registros almacenados en cada variable (porcentaje bajo de clientes en blanco[14]) sino en la confiabilidad de la información archivada. Por ejemplo, puede ser que la variable sueldo no sea un buen predictor dentro de los modelos tal vez porque los clientes tienden a mentir sobre su salario, por lo que la correlación entre sueldo y riesgo es espúrea, o por que los vendedores de crédito adulteran la cifra ante la posibilidad de obtener una aprobación segura del crédito vendido por parte del analista[15].

Si bien el desarrollo de este tipo de modelos no es complejo, su elaboración, monitoreo y administración **puede necesitar alguien con experiencia** (consultor externo), al menos hasta el desarrollo de una cultura organizacional basada en una administración conciente del riesgo y de los beneficios de su medición, monitoreo y control; sin embargo, el uso de un consultor reduce la flexibilidad del sistema y aumenta la dependencia de la institución financiera al mismo, en caso de disolver el convenio y

[13] En especial para instituciones de microfinanzas en las que los prestatarios suelen no poseer historial de crédito en central de riesgos o son trabajadores cuenta propistas sin registro o con registros poco fiables de ingresos laborales. En el caso ecuatoriano esta situación se extiende al crédito de consumo.

[14] Se suele considerar que una variable es buena si tiene menos del 5% de los registros como blanco. Note que blanco no significa cero. Cero debe ser siempre en la base el registro del valor "0" mientras que blanco significa ausencia de información respecto a esa variable. Por ejemplo, un cliente casado puede tener cero cargas familiares, mientras que un casado que tenga esta variable en blanco no implica que también tiene cero cargas familiares, sino que no se ha almacenado esta información.

[15] Siempre que no exista la posibilidad de verificar el sueldo del cliente, como es el caso usual de los trabajadores por cuenta propia que no facturan (sector informal).

de no poseer la capacidad de autogestión en este tema se puede incurrir en riesgo operativo, sin mencionar la dificultad de adaptación al método por parte de la gente dentro de la organización que puede devenir del rechazo al cambio o de la puesta en duda del aporte de un consultor que desconoce los clientes de la institución y que mediante una fórmula "dice" ser capaz de indicar el riesgo alto o bajo de un cliente, juicio que era propiedad del analista de crédito. Para vencer este obstáculo es fundamental la capacitación continua no sólo de la alta gerencia sino de los analistas de crédito, educación que implica un conocimiento del modelo, que genere confianza en el método y no una caja negra que arroja resultados[16].

Para que el scoring tenga éxito en la práctica, además de su correcto modelado, es **necesaria su implementación dentro del sistema transaccional**, si no existe personal asignado a este trabajo o se cometen errores en el proceso, el proyecto scoring puede fracasar en esta etapa.

El modelo **compara las solicitudes actuales con aquellas almacenadas históricamente en la base de datos**, estas solicitudes fueron aprobadas pasando por los filtros subjetivos de los analistas y las medidas de política adoptadas hasta la fecha por la institución, por ende los únicos préstamos existentes en las bases son aquellos que han sido analizados previamente, por ejemplo si una medida de política es no prestar a los trabajadores informales y por ende no existe información respecto a ellos en las bases de datos, el scoring no solucionará este problema, en otras palabras el modelo ignora todos los factores de riesgo que no están cuantificados ni registrados en la base de datos.

Los modelos **no pueden pronosticar algo que no haya sucedido con suficiente frecuencia en periodos anteriores y que no esté almacenado en la base de datos**, pese a esta deficiencia el modelo no pierde capacidad de pronosticar el riesgo relativo, es decir el cliente con mayor probabilidad de incumplimiento lo seguirá siendo ante cambios en el contexto socio-económico, sin embargo, sí se pierde poder de pronóstico del valor de riesgo absoluto.

---

[16] Posiblemente la sencillez en la comprensión del funcionamiento de modelos como análisis discriminantes, árboles de decisión o regresiones ha favorecido su difusión por encima de métodos como las redes neuronales o los algoritmos genéticos.

Por último, es relevante recordar que el modelo pronostica probabilidades de éxito (buen cliente) o fracaso (mal cliente)[17], pero en la realidad el riesgo observado es siempre cero (malo) o uno (bueno), por esto el **scoring no logra nunca dar en el blanco para un préstamo dado**, la función del modelo es lograr en promedio determinar el riesgo asumido por el prestamista. Siguiendo el ejemplo de Schreiner, el scoring funciona si entre 1000 préstamos cada uno con un riesgo pronosticado de 10 por ciento, el riesgo promedio observado es de más o menos 10 por ciento. No se puede juzgar el scoring por su desempeño en casos particulares, pues éste siempre fallará en algunos casos. Los escépticos en el uso de estas técnicas pueden objetar que con un umbral de malos del 50% se están rechazando solicitudes que tienen alrededor del 50% de probabilidades de ser buenos clientes, pero este problema no es salvado tampoco por el scoring subjetivo, sino que el scoring estadístico lo hace explícito, mientras que el subjetivo no permite medir la magnitud de la pérdida.

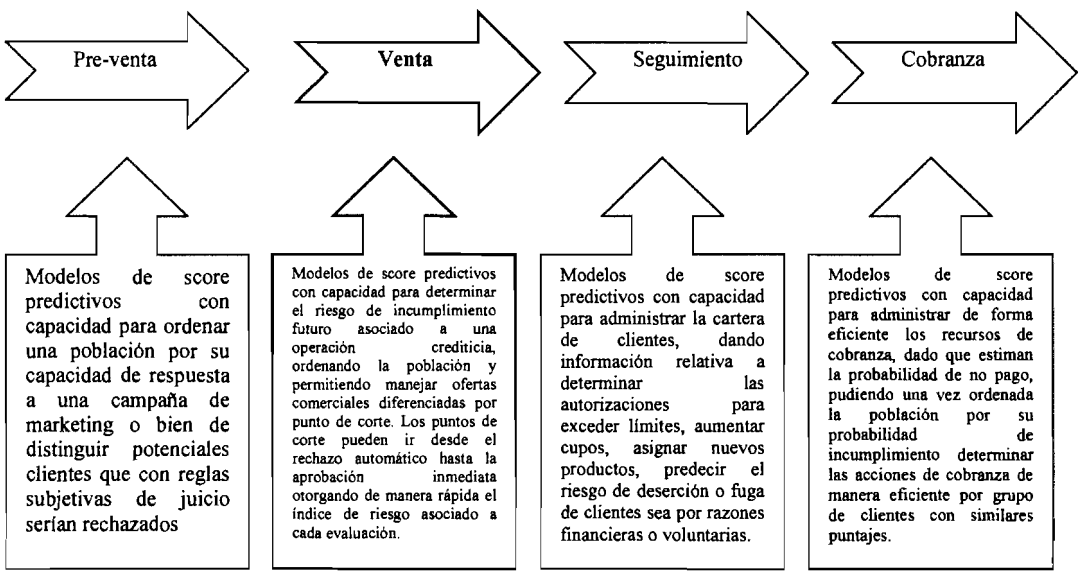
En conclusión y salvando las limitaciones que pueden presentar los modelos estadísticos empleados para la valoración del riesgo de crédito en banca de consumo, es inevitable admitir que la esencia de las finanzas es la predicción del riesgo de incumplimiento de la promesa de pago de un prestatario. Estas estimaciones de riesgo sean subjetivas u objetivas siempre se basan en información acumulada sea en la mente de los analistas o en los sistemas de información de las instituciones crediticias. La calificación estadística apuesta por la segunda, datos de carácter cuantitativo que permitan no sólo asignar un valor de mejor o peor cumplimiento sino revelar también cómo estos datos (características del crédito, acreditado y acreditador) afectan al riesgo, por ende siempre tendrá algún poder predictivo, que en fin último sino logra suplir en su totalidad la valoración del analista, será un complemento valioso en la predicción del riesgo y en la reducción de costos.

[17] En términos estadísticos se considera éxito a la ocurrencia del evento que queremos pronosticar, en este caso la probabilidad de que el cliente sea bueno.

III.- Esquema metodológico para la elaboración de un proyecto scoring

Si bien los modelos scoring pueden ser aplicados en cualquier momento del ciclo continuo del crédito este esquema hace referencia al momento de la venta o de la aprobación del crédito.

CICLO DEL CRÉDITO



Fuente: Equifax

Elaboración propia

Por tanto el modelo busca determinar el riesgo de no pago o de incumplimiento futuro asociado a una operación crediticia, ordenando la población y dando la posibilidad de manejar ofertas comerciales diferenciadas por punto de corte. Dando adicionalmente la facilidad de utilizar múltiples puntos de corte que van desde el rechazo (puntaje más bajo) hasta la aprobación inmediata (puntaje más alto) otorgando rápidamente el índice de riesgo asociado a esa evaluación.

La construcción de un modelo scoring de aprobación se puede resumir en un "procedimiento de modelado" que incluye seis fases:

1.- Selección de la muestra

2.- Definición de buenos y malos

3.- Definición y selección de datos

4.- Análisis preliminar de los datos

4.1.- Verificación de la integridad de la información contenida en la base de datos.

4.2.- Análisis preliminar, que incluye transformación de variables y tratamiento de datos nulos.

4.3.- Decidir sobre los esquemas de segmentación

4.4.- Selección preliminar de variables

5.- Análisis multivariado

5.1.- Verificación de las dimensiones de los datos

5.2.- Selección fina de variables

5.3.- Selección gruesa de variables

5.4.- Escoger el algoritmo de scoring

6.- Validación del modelo

6.1.- Análisis de percentiles

6.2.- Indice de Gini

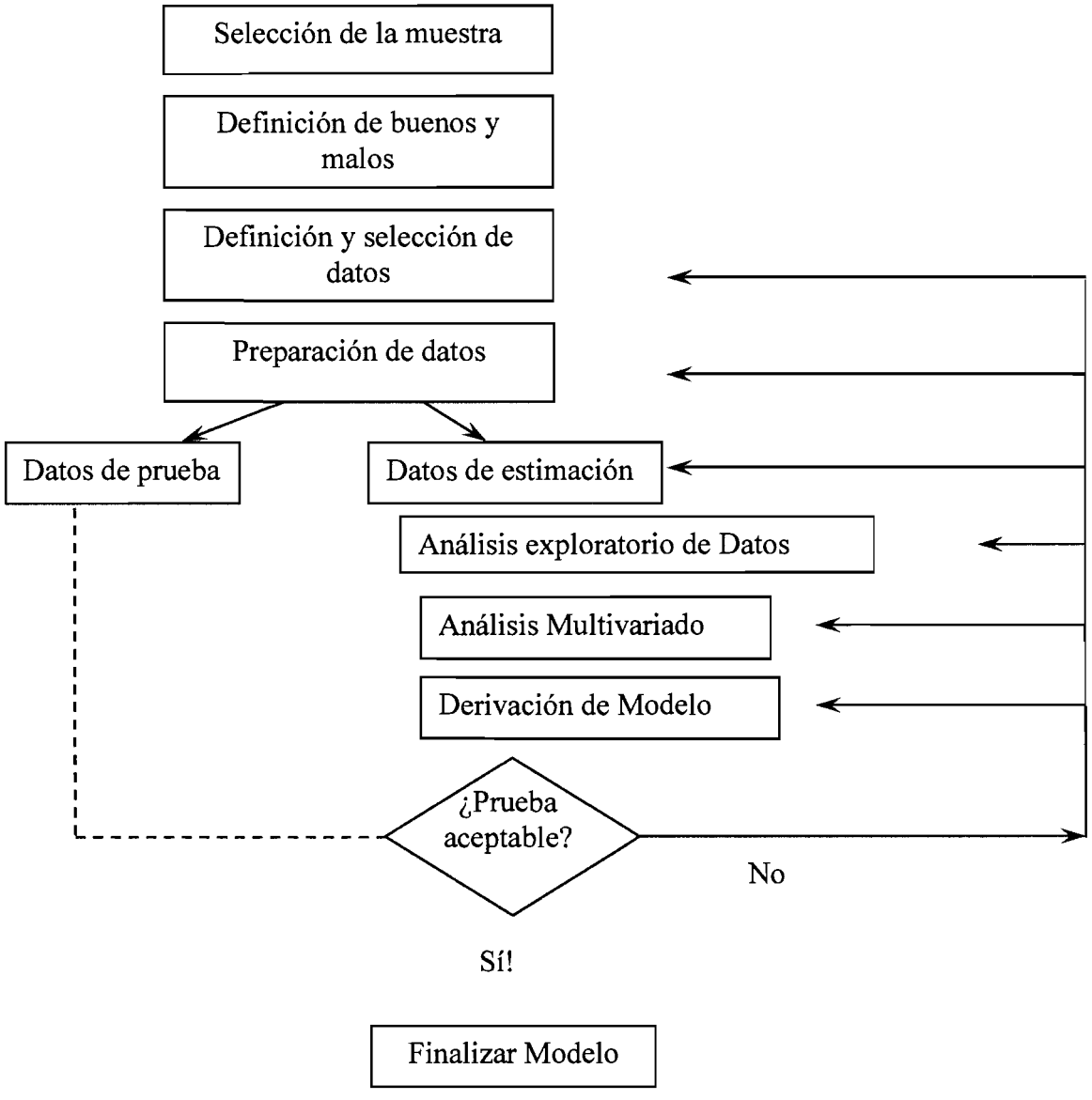
6.3.- Rendimiento

7.- Diseño de la scorecard

8.- Determinación de los puntos de corte (cutoff)

9.- Implementación en el sistema transaccional

ESQUEMA DE MODELADO DE UN SCORING CREDITICIO



Elaboración propia

1.- Selección de la muestra

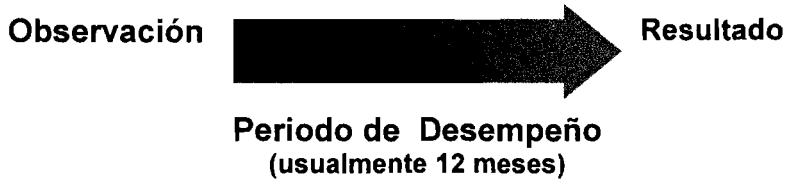
Con la finalidad de desarrollar un sistema scoring es necesario contar con una muestra de clientes que contenga tanto información histórica como sociodemográfica de los mismos. Este requerimiento genera una dicotomía entre dos elementos de decisión: a) la muestra debe ser representativa de aquellas personas (clientes potenciales) que deseen aplicar a una línea de crédito en el futuro (through-the-door-population), b) la muestra debe incorporar información suficiente acerca de las diferentes conductas de

pago de los clientes (buena y mala conducta en los extremos) a fin de identificar las mejores características que lograrán recoger estos comportamientos en los clientes potenciales.

Por ende, al momento de querer seleccionar la muestra a emplear en un modelo de aprobación tenderemos a buscar aquella que esté más cerca de los clientes potenciales. Es aquí justamente donde se genera el conflicto, dado que con esta muestra se necesita definir el criterio de bueno y malo a emplear para encontrar las características relevantes en el modelo. Para esto, en cambio, se necesitará información histórica suficiente y por lo tanto un razonable horizonte temporal.

Usualmente se fija un periodo de 12 meses para un scoring de aprobación. La interrogante es determinar cuáles 12 meses se deben seleccionar para contar con información actualizada que a la vez recoja una madurez adecuada (comportamiento estable), es decir minimizar el costo del dilema anterior.

Lo arriba expuesto se sintetiza en un diagrama que permite establecer los dos puntos que se deben determinar al momento de seleccionar la muestra:



El periodo de observación y el momento en que queremos ver los resultados de pago de los clientes a seleccionar como muestra. El periodo de observación es el tiempo t en el que el investigador decide situarse y observar el desempeño del cliente. Es este periodo de desempeño o performance el que va a ser empleado para predecir el comportamiento futuro de los potenciales clientes. En el punto de resultado se asigna una calificación (bueno o malo) al cliente con base en un resumen del comportamiento en el periodo de desempeño. De ahí la importancia de madurez de la cartera para no calificar como bueno a un cliente que es malo, pero que no logra denotar un comportamiento porque inicia a pagar su obligación.

Periodo de performance

Como se mencionó anteriormente es importante escoger un periodo que reflejando la actual población de clientes muestre un comportamiento estable de cartera. Un indicador útil es la tasa de morosidad.

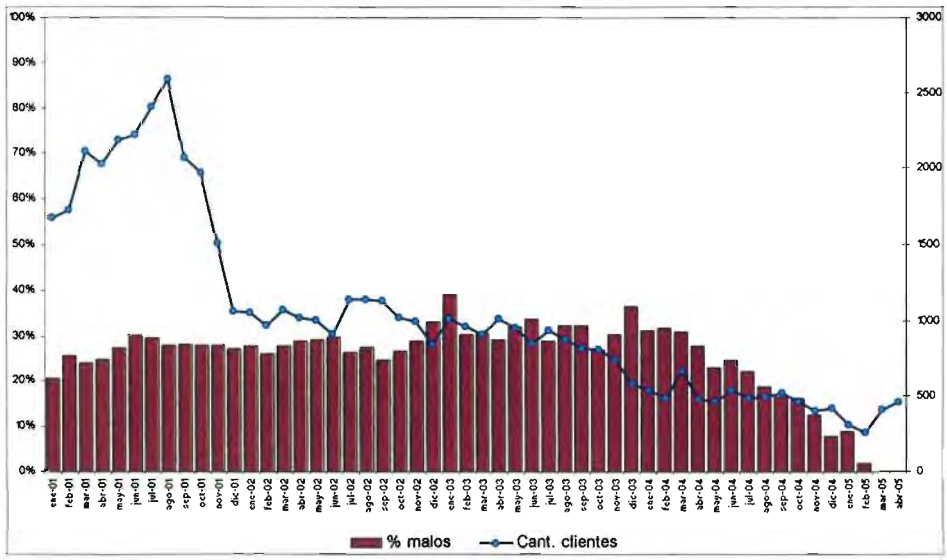
$$\text{Tasa de morosidad}_t = \frac{\#\text{ de malos clientes}_t}{\#\text{ total de clientes}_t}$$

La tasa de morosidad se construye por periodo de cosecha (fecha de venta de las operaciones de crédito) y tiene por objeto mostrar en forma gráfica y por mes de colocación la relación clientes malos sobre total de clientes, con el fin de señalar los periodos en que esta razón se estabiliza como equivalentes a un comportamiento estable de la cartera.

Un periodo se considera como estable en la medida que la razón de la cosecha t presente pequeñas variaciones en relación a la cosecha t-1 y t+1. En términos estadísticos se puede argumentar que se necesita, durante la ventana temporal elegida, que la razón tasa de morosidad siga una distribución uniforme. La elección de periodos con tasas de morosidad decrecientes no implica necesariamente una mejora en el comportamiento de la cartera, puede ser que estas cosechas por ser cercanas a la fecha actual estén reflejando carteras poco maduras y por ende no comparables con otros periodos de análisis. A este tipo de estudios se los conoce como **análisis de cosechas**.

A modo de ejemplo supongamos que en mayo de 2003 se desembolsaron alrededor de 1000 créditos (línea azul), de los cuales 300 han presentado morosidad máxima (barras rojas). En este caso se tiene un porcentaje de vencimiento del 30%, así se puede leer para cada fecha de venta la tasa de morosidad del total de créditos colocados, no obstante al acercarnos a fechas más recientes se aprecia una mejora en el comportamiento de los clientes, pues la tasa de morosidad decrece a niveles del 10%, el error consiste en creer que esto se debe a mejoras en las cosechas, cuando lo que realmente está ocurriendo es que falta maduración o tiempo de observación y por ende

GRÁFICA DE COSECHAS PARA DETERMINAR EL PERIODO DE DESEMPEÑO



Fuente: Institución financiera ecuatoriana

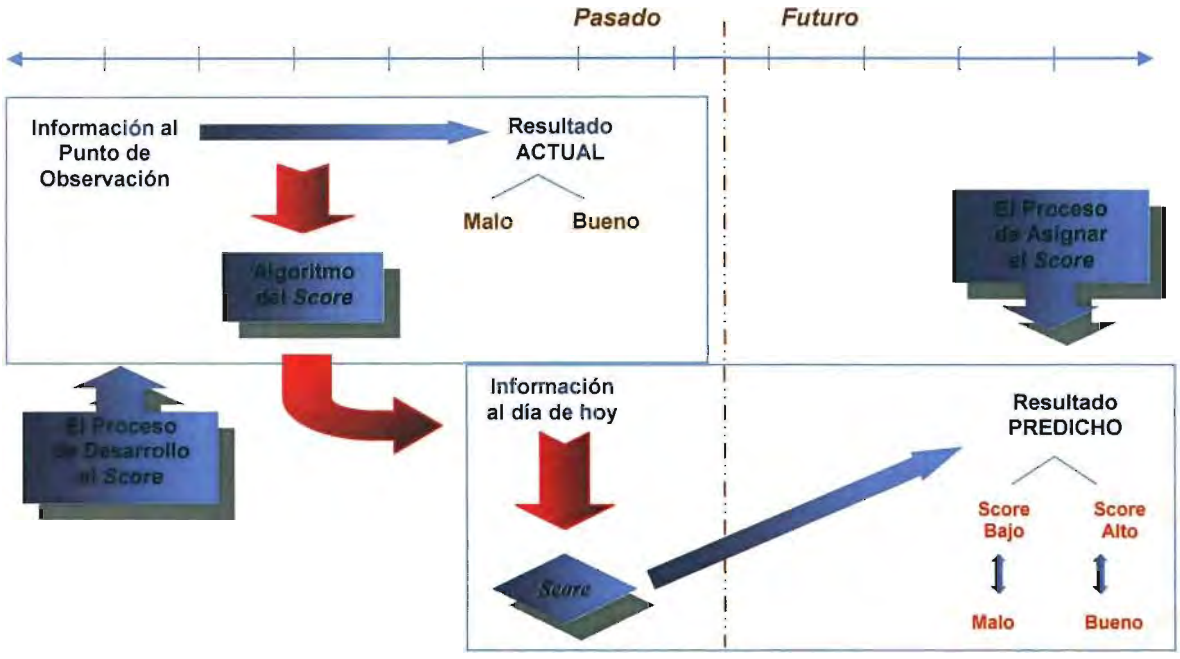
Elaboración propia

de desempeño que permita comparar con otros periodos de análisis. Si, por ejemplo se define como cliente malo a aquel que haya caído en mora alguna vez más de 90 días, no hay forma de que existan clientes malos en los últimos dos meses de desembolsos, porque no llegan a cumplir aún el requisito, mas no por que son buenos.

Una buena cosecha será por tanto aquella que mostrando una adecuada maduración tiene un bajo porcentaje de malos clientes y una cosecha será mejor que otra (en el mismo producto o productos similares) siempre que su porcentaje de malos sea menor. En la gráfica se puede considerar una cartera ya madura hasta mayo de 2004.

ESQUEMA DE SELECCIÓN DE LA MUESTRA

PARA PREDECIR COMPORTAMIENTOS FUTUROS



Fuente: Equifax

El esquema arriba expuesto sintetiza la importancia del proceso de selección de la muestra, un acertado periodo de performance, asegura que la definición de bueno y malo permita modelar un adecuado comportamiento de pago en el futuro.

2.- Definición de buenos y malos

La implementación de un modelo credit scorecard requiere la definición de buen y mal cliente. Establecer que un cliente es malo no implica necesariamente que todos los restantes sean buenos. En el camino se pueden encontrar al menos dos definiciones adicionales. Los 'indeterminados' o aquellos casos que no se pueden definir como buenos o como malos, y los 'experiencia insuficiente' o aquellos casos en que la cuenta, producto de su poco o corto uso, no se puede definir como buena sin caer en un juicio prematuro.

Por ejemplo, en el desarrollo de un scorecard para un portafolio de tarjetas de crédito, es común definir como malo al cliente que en algún momento se encuentra con

tres pagos de atraso. Esta definición se conoce como 'ever 3+ down' o 'worst 3+ down'. En este caso un cliente indeterminado sería aquel que tiene sólo dos cuotas impagas[18].

Resulta evidente que la definición escogida no va a afectar la metodología empleada para calificar al cliente. (esta asume que la definición crea una partición). Sin embargo, la forma como definamos buenos y malos sí va a tener efecto sobre los resultados del scorecard. Por lo tanto su definición requiere el conocimiento del sistema de cartera en mora dentro de la compañía, del proceso de cobranzas, entre otros.

Cuando se habla de malos se busca describir al conjunto de clientes de la institución (o las cuentas), que dada la experiencia no quiere seleccionar para su actividad intermediadora. Para el caso de los modelos de riesgo, usualmente esta definición hace alusión a esa cartera que de conocer su comportamiento no se hubiese aceptado. De aquí se sigue que la definición de bueno y malo se basará prácticamente en el comportamiento de pago de los clientes: mora máxima histórica, mora promedio, contadores de mora (número de veces que ha caído en mora o reincidencia)[19]. Es evidente que estas definiciones pasan por un grado de subjetividad siendo tan conservadoras como lo quiera la entidad o viceversa. En el caso de contar con varios productos se puede optar por una definición de bueno y malo para cada uno pues se sabe que los productos financieros no son homogéneos y sus características pueden influir en el comportamiento de pago de los clientes. Por ejemplo, la definición de bueno para el caso de un crédito hipotecario puede ser más estricta que la requerida para una tarjeta de crédito. Es habitual toparnos con moras de 30 días en tarjeta habientes, por lo que la definición de bueno en este caso no requiere ser tan ácida (muy bajo atraso medio y atraso máximo). Aunque finalmente todo dependerá de los objetivos en reducción del riesgo que se quiera plantear la institución.

Una técnica sencilla para definir los buenos y malos es la matriz atraso promedio y atraso máximo que consiste en listar en filas los rangos de atraso máximo y en

[18] Los indeterminados pueden causar problemas adicionales cuando este comportamiento se repite frecuentemente.

[19] Se pueden incluir definiciones en función de la rentabilidad que reporte el grupo de clientes por rango de atraso. Este criterio de selección es mucho más deseable pues asocia el criterio de bondad a una tasa de rendimiento, decidiendo la elección de bueno o malo con base en la máxima pérdida que se está dispuesto a asumir.

columnas los rangos de atraso medio, en las celdas de la matriz se van situando los clientes distribuidos de acuerdo al par de valores (atraso medio; atraso máximo)[20]

MATRIZ ATRASO PROMEDIO ATRASO MÁXIMO

Rango Atraso Max	Rango Atraso Prom							Total general
	0	1 - 15	16 - 30	31 - 60	61 - 90	91 - 120	> 120	
0	188							188
1 - 15	6.594	7.145						13.739
16 - 30	801	4.696	10					5.507
31 - 60	82	5.207	732	40				6.061
61 - 90	1	1.178	1.820	530	7			3.536
91 - 120		130	855	975	58	4		2.022
> 120		23	254	1.351	1.235	1.045	2.241	6.149
Total general	7.666	18.379	3.671	2.896	1.300	1.049	2.241	37.202

Fuente: Institución financiera ecuatoriana

Elaboración propia

En la matriz se observa que el 37,4% de la población empleada para la construcción de esta tabla se encuentra entre los 15 días de atraso promedio y 15 de atraso máximo, la elección de este par de valores para los buenos, bastante ácido por la definición, puede afectar la tasa de aceptación de clientes o incrementar la tasa de rechazos. Estos criterios también deben sumarse al momento de realizar una definición, al menos como una aproximación a los efectos que podrían tener sobre la rentabilidad. Si estiramos un poco los días de atraso máximo, por ejemplo al rango 16-30 vemos que ahora tenemos una definición de buenos que alcanza al 52,2% de la población estudiada (20 puntos porcentuales por encima de la anterior) sin necesidad de ser demasiado laxos. Claro que siempre se debe tener en cuenta el producto que se está estudiando, el nivel de riesgo que se quiere asumir (definiciones más blandas implican mayor riesgo), los costos de estos clientes y su impacto sobre la rentabilidad del negocio, pues el objetivo no es discriminar para perder, sino para mejorar los beneficios.

Una vez establecida la definición de buenos, se entiende como malos al complemento, es decir al resto de la población distribuido en las restantes edades de mora.

[20] Recuerde que la construcción de esta matriz exige la previa elección de una cartera madura.

INDICADOR DE BUENOS Y MALOS

Rango Atraso Max	Rango Atraso Prom							Total general
	0	1 - 15	16 - 30	31 - 60	61 - 90	91 - 120	> 120	
0	188							188
1 - 15	6.594	7.145						13.739
16 - 30	801	4.696	10					5.507
31 - 60	82	5.207	732	40				6.061
61 - 90	1	1.178	1.820	530	7			3.536
91 - 120		130	855	975	58	4		2.022
> 120		23	254	1.351	1.235	1.045	2.241	6.149
Total general	7.666	18.379	3.671	2.896	1.300	1.049	2.241	37.202

Buenos

Malos

Fuente: Institución financiera ecuatoriana

Elaboración propia

3.- Definición y selección de datos

Como se mencionó inicialmente una de las dificultades que puede afrontar un modelo scoring es la escasez de una buena base de datos o incluso la ausencia de esta. Es importante, por ende, contar no sólo con una base de datos sino con un sistema de información adecuadamente construido con una lógica de almacenamiento de datos que permita contar con una codificación o representación numérica de las características cualitativas y cuantitativas que servirán para la aplicación de técnicas estadísticas. La definición y selección de los datos a incluir en el modelo requiere identificar las escalas de medida que pueden presentarse.

Existen dos grupos de variables a estudiar: las cualitativas y las cuantitativas. Las primeras son aquellas que no aparecen en forma numérica, sino como categorías o atributos (género, actividad económica, tipo de vivienda) y sólo pueden ser nominales u ordinales.

Las variables nominales sólo permiten establecer frecuencias en cada atributo y la igualdad o desigualdad entre los diferentes casos, no es posible jerarquizar sus modalidades. Estas variables se denotan asignando a cada atributo dentro de la variable un número. La medida estadística que permite ver el grupo con mayor frecuencia es la moda. Por ejemplo, la variable sexo se define como Hombre: 1; Mujer: 0, donde los números son indicadores de pertenencia a una clase y no reflejan ninguna relación.

Las variables ordinales recogen la idea de orden, pero no tiene sentido realizar operaciones aritméticas con ellas ya que no puede medirse distancia entre una categoría y otra. En este tipo de variables se puede establecer igualdad y desigualdad y relaciones de mayor y menor que. Puede establecerse orden, pero no medirse distancia dentro de ese orden. La medida estadística de tendencia central más apropiada para estas escalas es la mediana. Un ejemplo de estas variables es el nivel de educación, en donde primaria es menor que secundaria y este a la vez menor que universitaria, reflejando una jerarquía pero no una medida de cuán lejos está un atributo del otro. La codificación de estas variables es similar a las variables nominales, por ejemplo primaria: 1; secundaria: 2 y universitaria: 3 pero en este caso los números sí indican un mayor nivel de educación pero no indican cuanto.

Las variables cuantitativas, en cambio son aquellas cuyas categorías pueden expresarse numéricamente. Su naturaleza numérica permite un tratamiento estadístico más elaborado debido a las operaciones matemáticas que permiten. Estas variables pueden ser discretas o continuas. Las primeras son aquellas cuyas categorías sólo pueden tomar valores enteros. Por ejemplo la variable número de cargas familiares, no existe un cliente que pueda tener 2,3 cargas familiares. Las segundas son aquellas cuyas categorías pueden fraccionarse según cualquier entero, por ejemplo la variable salario. Una vez definido los tipos de datos con los que se puede trabajar se procede a seleccionar la información en función de las características disponibles.

Características disponibles

Existen tres tipos de características disponibles para discriminar entre buenos y malos clientes:

- Las derivadas del formulario de aplicación
- Las que provienen de los buró de crédito (bases externas), y
- Aquellas que describen la historia transaccional del prestatario (sólo scoring de comportamiento)

En cuanto al formulario de aplicación surgen varias dificultades, como qué hacer cuando se tiene no respuesta. Un tratamiento habitual es tener una categoría 'no respuesta' para cada característica, pero pueden existir casos en que la no respuesta

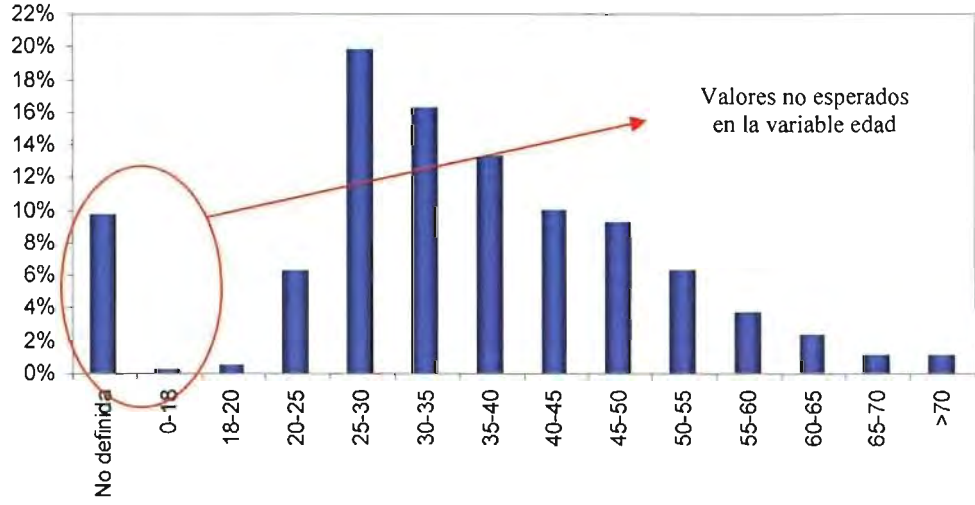
realmente implique un 'no' o 'ninguno' (p.ej. en respuesta a la pregunta ingreso del cónyuge).

Otro inconveniente, dado que existen cientos de ocupaciones, es la actividad laboral. Una solución a este problema, no del todo satisfactoria, es especificar atributos como ejecutivo, obrero, gerente, etc. Sin embargo, como se mencionó, esto no deja de implicar problemas, pues una persona puede gerenciar una empresa con cientos de empleados, mientras que otra puede hacerlo con su carrera de medio tiempo de vendedor de cosméticos. Por otro lado, se puede optar por simplificar el dilema y preguntar únicamente si se es empleado, desempleado o empleado por cuenta propia. Opción válida siempre y cuando se considere que esta alternativa puede recibir más de una respuesta.

El ingreso es otra característica que presenta dificultades. Es importante definir con claridad qué tipo de renta estamos solicitando: mensual, anual, sueldo básico, total o el total del hogar en que habita.

En general, es importante analizar cada característica a incluir en el formulario y una vez recopilada la información para el desarrollo del scoring, validarla. Para cada característica se puede revisar la distribución de las respuestas a fin de chequear que estas tengan sentido. (p.ej. no podemos esperar en una distribución de edades que existan individuos con 0 o con 180 años).

DISTRIBUCIÓN POR RANGO DE EDADES

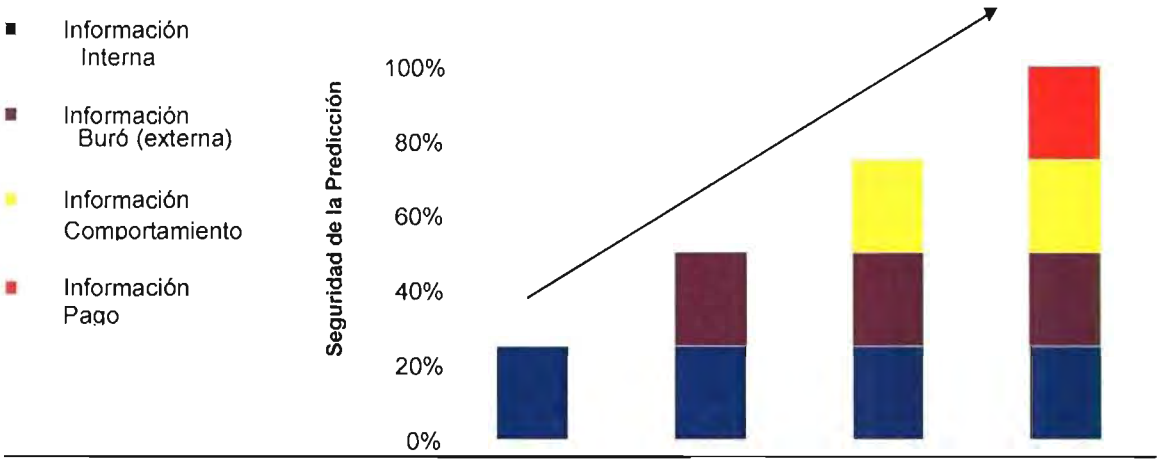


Fuente: Institución financiera ecuatoriana

Elaboración propia

Es importante el acceso a mayores y mejores fuentes de información y contar con excelentes bases internas pues este conjunto de datos permite alcanzar mejores niveles de captura de buenos y malos clientes y aumentar la eficiencia de separación de ambas poblaciones, tanto para el scoring de aprobación como para los de comportamiento o cobranzas.

RELACIÓN ENTRE LA SEGURIDAD DE LA PREDICCIÓN Y LAS FUENTES DE INFORMACIÓN



Fuente: Equifax

La inclusión de características que permitan discriminar clientes puede ser tan extensa como se quiera y es muy probable que estén fuertemente correlacionadas unas con otras. La decisión de cuáles mantener y cuáles ignorar es parte del arte de elaborar modelos scoring. Sin embargo, existen varios criterios que permiten eliminar las variables que tengan poco poder de predicción.

4.- Análisis preliminar de los datos

4.1.- Verificación de la integridad de la información contenida en la base de datos.

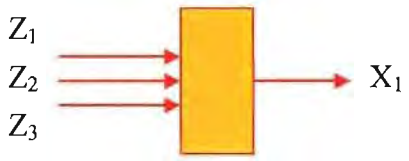
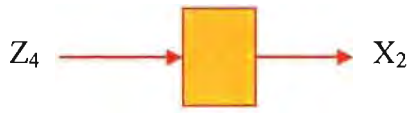
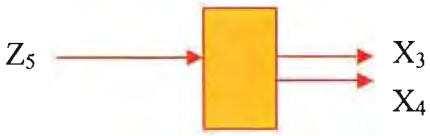
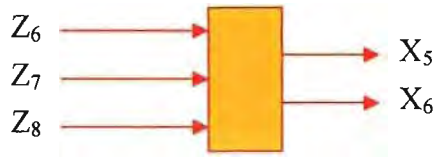
4.2.- Análisis preliminar, que incluye transformación de variables y tratamiento de datos nulos.

4.3.- Decidir sobre los esquemas de segmentación

4.4.- Selección preliminar de variables

En un principio todas las variables existentes en la base de datos deben ser seleccionadas e incluso se debe analizar la posibilidad de construir nuevas variables a partir de las ya existentes. Esto permitirá emplear no sólo variables simples sino incorporar variables cruzadas o de combinaciones de características como por ejemplo el tipo de vivienda cruzado con el rango de edad o el estado civil con el número de cargas familiares.

CONSTRUCCIÓN DE VARIABLES CASOS ESPECIALES

	
Ejemplo: X_1 = casado, educación universitaria y tiene menos de dos cargas familiares.	Ejemplo: Transformar la variable continua edad en rango de edades
	
Ejemplo: Z_5 = tipo de vivienda con tres atributos puede transformarse en X_3 = casa propia y X_4 = casa arrendada.	Ejemplo: X_5 = afiliado a la seguridad social, casa propia y es empleado. X_6 = afiliado a la seguridad social, casa arrendada y es empleado.

Elaboración propia

En esta fase se deben tratar tanto los valores nulos (missing values) como una selección inicial de variables. En el caso más simple se deben considerar irrelevantes los valores nulos (no buscar ningún mecanismo de asignación de valores a estos registros) cuando estos, en porcentaje, no representan más del 5% de la información contenida en la variable, de lo contrario o se debe buscar una regla de asignación o eliminar la

variable por no contener información suficiente para incluirla en el proceso de modelado.

Esto conduce a pensar en la necesidad de buscar medidas descriptivas que indiquen los valores o atributos de las variables así como los casos asignados a cada uno. Adicional a este análisis el modelizador de riesgos debe conocer el grado de confiabilidad de las variables, por ejemplo en un proceso de entrega de créditos a través de terceros (local comercial de electrodomésticos) en donde el proceso de venta es controlado por la casa comercial y no por la institución financiera y en el cual la rapidez en el otorgamiento del crédito es característica diferenciadora al momento de elegir socios comerciales, es probable que ciertas variables de la solicitud no puedan pasar por procesos de verificación y puedan ser manipuladas por el vendedor, como es el caso del sueldo del cliente, por lo tanto esta no puede ser considerada dentro de la base de datos para modelar el comportamiento del cliente, aún cuando su información se encuentre disponible para todos los registros.

AUDITORÍA SIMPLE DE VARIABLES, EJEMPLO: TIPO DE VIVIENDA

Variable	Tipo de vivienda	
Categorías	**Cantidad**	**%**
	361	0,1%
No definido	68.942	22,7%
Arrendada	73.339	24,1%
Familiar	77.274	25,4%
Herencia	1.119	0,4%
Hipotecada	256	0,1%
Propia	83.016	27,3%
Total	**304.307**	**100,0%**

Fuente: Institución financiera ecuatoriana

Elaboración propia

La tabla adjunta muestra la variable tipo de vivienda obtenida de un formulario de aplicación levantado en una casa comercial. Como se observa, la categoría 'No definido' tiene 68942 clientes, o un equivalente al 22,7% de la base, esta distribución pone en duda la fiabilidad en el proceso de almacenamiento de la variable, pues indica que del 100% de los clientes que aplican a un crédito de este tipo declara no poder definir el tipo de vivienda en el que habita; intentar inferir la vivienda a través de algún otro conjunto de características no es recomendable dado el alto porcentaje de datos no

definidos y la eliminación de estos registros, en cambio produce una pérdida del 27% de la información recogida en esta muestra de datos.

AUDITORÍA SIMPLE DE VARIABLES, EJEMPLO: ESTADO CIVIL

Variable	Estado civil	
Categorías	Cantidad	%
	607	0,2%
CASADO	75.350	24,7%
CASADO SIN INF.CONYUGL	63.391	20,8%
DIVORCIADO	17.196	5,6%
SEPARADO	5.905	1,9%
SOLTERO	115.051	37,7%
UNION LIB. SIN INF.CONYU	15.117	5,0%
UNION LIBRE	6.793	2,2%
VIUDO	5.403	1,8%
Total	**304.813**	**100,0%**

Fuente: Institución financiera ecuatoriana

Elaboración propia

En el caso de la segunda variable, 'estado civil' se observa que el 0.2% de los datos no reportan información sobre el estado civil, eliminar este registro implica en este caso descartar 607 clientes de una muestra de 304813 por lo que con toda certeza no se está perdiendo representatividad de ningún atributo dentro de la muestra, ni su inclusión (en el caso de poder conocer el estado civil) afectaría la distribución actual de las categorías.

5.- Análisis multivariado

5.1.- Verificación de las dimensiones de los datos

5.2.- Selección fina de variables

5.3.- Selección gruesa de variables

5.4.- Escoger el algoritmo de scoring

Una vez realizada las auditorías simples de las variables contenidas en la base de datos se procede a elaborar una clasificación fina de variables, que consiste en analizar cada uno de sus atributos o categorías. El objeto es identificar el grado con el que

pueden contribuir en el modelo para discriminar entre buenos y malos. Si una característica es particularmente débil puede descartarse en esta etapa.

Un indicador que se emplea para la identificación del peso y signo que se debe esperar tenga un atributo dentro de un modelo es el porcentaje de referencia[21].

$$\%\text{ de referencia} = \frac{\%\text{ de malos en la población} - \%\text{ de malos en el atributo de la variable}}{\%\text{ de malos en la población}}$$

Este consiste en encontrar una medida porcentual que recoja la proximidad o lejanía de los malos contenidos en un atributo en relación al total de malos de la población. Si el porcentaje es negativo el atributo se espera ingrese con signo negativo en el modelo pues contribuye a desmejorar el comportamiento de pago del cliente, si el porcentaje es positivo ocurre el efecto contrario. Un porcentaje de referencia en términos absolutos por encima del 25% puede considerarse decidor al momento de seleccionar la variable que ingresará a la fase de modelado, aunque esta regla es empírica[22]. Un punto adicional a considerar al momento de seleccionar la variable es que su representación respecto al total muestreado supere el 5%. Este análisis se conoce como **de correlaciones,** aunque estrictamente en términos estadísticos el uso de un modelo de regresión estaría supliendo estas etapa. Sin embargo su uso puede ser de utilidad como proceso de conocimiento de las variables y de los posibles signos esperados en los coeficientes del modelo.

ANÁLISIS DE CORRELACIONES VARIABLE: AFILIACIÓN A LA SEGURIDAD SOCIAL

Afiliación al IESS	Bueno		Malo		%Ref	Total	% Total
	Cant	%	Cant	%			
NO	62.149	73,02%	22.960	26,98%	-2,69%	85.109	91,29%
SI	6.589	81,14%	1.532	18,86%	28,19%	8.121	8,71%
Total general	68.738	73,73%	24.492	26,27%	0,00%	93.230	100,00%

Fuente: Institución financiera ecuatoriana

Elaboración propia

[21] Término acuñado por la empresa consultora internacional LISIM, experta en la elaboración de modelos scoring.

[22] LISIM propone un porcentaje del 25%, aunque esta misma empresa en la práctica emplea valores de referencia por encima del 10% como indicadores de buena discriminación

La variable seguro social usada para ejemplificar el análisis empleando porcentajes de referencia indica que los clientes afiliados muestran un mejor comportamiento de pago que aquellos que no lo están (porcentaje de referencia del 28.19%), note que el porcentaje de clientes que tienen afiliación equivale al 8,71% (superior al 5% mínimo requerido) por lo que la variable puede considerarse "significativa".

ANÁLISIS DE CORRELACIONES VARIABLE: CIUDAD DEL CRÉDITO

Ciudad	Bueno		Malo		%Ref	Total	% Total
	Cant	%	Cant	%			
CUENCA	4.964	77,81%	1.416	22,19%	15,52%	6.380	6,84%
GUAYAQUIL	31.986	71,19%	12.947	28,81%	-9,68%	44.933	48,20%
IBARRA	2.175	85,73%	362	14,27%	45,69%	2.537	2,72%
LATACUNGA	648	85,38%	111	14,62%	44,33%	759	0,81%
LOJA	579	92,79%	45	7,21%	72,55%	624	0,67%
QUITO	27.457	74,37%	9.460	25,63%	2,46%	36.917	39,60%
RIOBAMBA	929	86,02%	151	13,98%	46,78%	1.080	1,16%
Total general	68.738	73,73%	24.492	26,27%	0,00%	93.230	100,00%

Fuente: Institución financiera ecuatoriana

Elaboración propia

En la segunda variable tomada como ejemplo se observa que los créditos entregados en Cuenca muestran un mejor comportamiento de pago que el resto de la población. Si bien ciudades como Ibarra, Latacunga, Loja o Riobamba muestran porcentajes de referencia por encima del 40% su participación en el total de la muestra no es significativo. La ciudad de Guayaquil muestra una correlación negativa con el hecho de ser buen cliente, pues su porcentaje de referencia es negativo y aunque este no es superior al 10% su relevancia dentro de la muestra (48% del total) sugiere incluirla en la fase de modelado.

Una segunda fase en el análisis de las variables es la denominada **clasificación gruesa o dura (coarse clasification)** de los atributos. Este tipo de agrupamiento consiste en reducir el número de categorías a unas más manejables.

La necesidad de clasificar (agrupar) los atributos de una característica se explica por dos razones, cada una asociada al tipo de variable (categórica o continua):

1. En el caso de las variables categóricas, contar con demasiados atributos puede agotar la muestra para cada respuesta y quitarle robustez al análisis.
2. Dado que el objetivo del credit scoring es predecir el riesgo en lugar de explicarlo, para las variables continuas es preferible contar con un sistema en el cual el riesgo sea no lineal en estas variables, siempre que esto permita una mejor predicción del mismo.

Ejemplo[23]**:** La tabla muestra la distribución de frecuencias de la pregunta ¿cuál es su estatus residencial?. Se sabe que la relación poblacional bueno malo es 9:1, pero las razones en los atributos varían en un rango de 20:1 a 1:1. La cuestión es si es correcto permitir estas relaciones y mantener todos los atributos.

DISTRIBUCIÓN POR ATRIBUTOS DE LOS CLIENTES BUENOS Y MALOS DE LA CARACTERÍSTICA ESTATUS RESIDENCIAL

Atributo	Propietario	Arriendo amoblado	Arriendo desamoblado	Dependencia	Otro	No responde	Total
Buenos	6000	1600	350	950	90	10	9000
Malos	300	400	140	100	50	10	1000
Razón bueno:malo	20	4	2.5	9.5	1.8	1	9
% atributo/total	63.0%	20.0%	4.9%	10.5%	1.4%	0.2%	

Fuente: Thomas, L., Edelman, D., Crook, J. Credit Scoring and Its Applications.

Observando los datos de la tabla notamos que sólo 20 individuos de la población pertenecen al atributo 'no responde' (0.2%), 140 están en 'otro' (1.4%) y apenas un 5% de la población se ubica en 'arriendo desamoblado'. No es difícil notar que posiblemente estas categorías tienen muy pocas respuestas para asegurar con confianza que sus razones buenos:malos se reproducen en la población total.

Agrupando estas tres categorías en una nueva, llamada 'otras respuestas' se mejora el número de individuos en este atributo y se incrementan las posibilidades de reproducir la razón poblacional.

[23] Tomado del texto de Thomas, L., Edelman, D., Crook, J, "Credit Scoring and Its Applications", pág 132.

Atributo	Propietario	Arriendo amoblado	Dependencia	Otras respuestas	Total
Buenos	6000	1600	950	450	9000
Malos	300	400	100	200	1000
Razón bueno:malo	20	4	9.5	2.25	9
% atributo/total	63.0%	20.0%	10.5%	6.5%	

Fuente: Thomas, L., Edelman, D., Crook, J. Credit Scoring and Its Applications.

El argumento es similar para la agrupación de los atributos 'arriendo amoblado' y 'arriendo desamoblado' dado que sus razones buenos:malos no están tan distantes y ambos pertenecen a la categoría arriendo.

Si se prefieren sólo tres categorías se pueden agrupar en 'propietario' (6000; 300), 'arriendo' (1950;540) y 'otros' (1050;160); o en 'propietario' (6000; 300), 'dependencia' (950;100) y 'otros' (2050;600).

En fin, estas combinaciones de atributos y la forma como se las hace son las que terminan convirtiendo en un arte la agrupación de categorías. Sin embargo, existen estadísticos que pueden servir como guías al momento de agrupar.

Uno de los estadísticos que tienen mayor uso al momento de describir cuan bien las características, con una específica agrupación de atributos, están diferenciando buenos de malos es el estadístico χ^2 de Pearson

Estadístico χ^2

El objetivo es conservar atributos que permitan diferenciar entre buenos y malos a la vez que se mantiene una relación buenos:malos que refleje la poblacional. Lo que se buscará, en términos estadísticos, es determinar si existe una relación de dependencia[24] entre los atributos establecidos y las clasificaciones (análisis de tablas de contingencia).

Para analizar si existe asociación entre variables cualitativas se puede usar el estadístico χ^2 de Pearson.

Este contraste de homogeneidad, (o de independencia, que asume igualdad en todas las clases o categorías) mediante la prueba Chi cuadrado entre dos variables cualitativas, se basa en la comparación de las frecuencias observadas con las frecuencias esperadas, estas últimas construidas bajo la hipótesis de independencia.

La prueba por ende testea la siguiente hipótesis:

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

Ho: Las variables son independientes

Ha: Las variables no son independientes

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

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

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

Clasificación	**Atributo**				
	1	**2**	**.....**	**b**	**Total fila**
1	n_{11}	n_{12}	.....	n_{1b}	$n_{1.}$
2	n_{21}	n_{22}	.....	n_{2b}	$n_{2.}$
.	.	.	.....	.	$n_{3.}$
.	.	.	.....	.	$n_{4.}$
.	.	.	.....	.	$n_{5.}$
A	n_{a1}	n_{a2}	.....	n_{ab}	$n_{a.}$
Total columna	$n_{.1}$	$n_{.2}$	.....	$n_{.b}$	N

Elaboración propia

Si la Ho es cierta las mejores estimaciones de las probabilidades de cada atributo son:

$$\bar{p}_j = P(A_j) = \frac{n_{.j}}{n}$$

y por tanto la frecuencia esperada en cada columna, siempre que no haya diferencias, será el resultado de multiplicar la probabilidad estimada de la columna por el número de elementos en cada clase.

$$e_{ij} = \frac{n_{.j}}{n} x n_{i.}$$

Note que lo que se está expresando es que la frecuencia esperada (conjunta) es el producto de las frecuencias marginales, sólo que dividida para n. (Se está asumiendo para la construcción de las frecuencias esperadas la Ho: $\mathbf{p_{ij} = p_{i.}p_{.j}}$)

Con esta información podemos construir el estadístico Chi cuadrado definido como:

$$\chi^2 = \sum_{j=1}^{b}\sum_{i=1}^{a}\frac{(n_{ij}-e_{ij})^2}{e_{ij}} = \sum_{j=1}^{b}\sum_{i=1}^{a}\frac{(nf_{ij}-e_{ij})^2}{nn_{i.}n_{.j}} \; con\,(\text{a-1})(\text{b-1})\,\text{g.l.}$$

La forma de construcción del estadístico deja ver que en caso de tener diferencias cercanas a cero (frecuencias observadas y esperadas muy parecidas) no se podrá rechazar la hipótesis nula y por ende la independencia. Lo que implicará que la clasificación definida no influye sobre los atributos o que el atributo no permite diferenciar entre buenos y malos clientes.

Por lo tanto la regla de decisión establece que se rechaza la Ho a un nivel α de significancia si el valor del estadístico excede al valor crítico de la distribución Chi cuadrado. *En el caso de selección de agrupamientos de atributos escogeremos la agrupación que muestre, entre los valores que exceden el valor crítico, el mayor valor del estadístico de Pearson, lo que implica que este agrupamiento muestra la mayor dependencia (menor independencia).*

Ejemplo: Característica Estatus Residencial

¿Cuál agrupamiento deberíamos escoger de tal forma que podamos diferenciar entre buenos y malos a la vez que se mantiene una buena relación buenos:malos poblacional?

Siguiendo el esquema presentado para el empleo del criterio de Pearson tendremos:

Si agrupamos en tres categorías definidas como: 'propietario'; 'arriendo' y 'otros'.

	Propietario	Arriendo	Otros	Total
Buenos	6000	1950	1050	9000
Malos	300	540	160	1000
Total	6300	2490	1210	10000
Esperadas buenos	5670	2241	1089	
Esperadas malos	630	249	121	
(observadas - esperadas)^2/esperadas	19,206	37,787	1,397	
	172,857	340,084	12,570	
S^2	583,902			
$X^2_{(2)\ \alpha=95\%}$	5,99			

Fuente: Thomas, L., Edelman, D., Crook, J. Credit Scoring and Its Applications.
Elaboración propia

Si agrupamos en tres categorías definidas como 'propietario'; 'dependencia' y 'otros'.

	Propietario	Dependencia	Otros	Total
Buenos	6000	950	2050	9000
Malos	300	100	600	1000
Total	6300	1050	2650	10000
Esperadas buenos	5670	945	2385	
Esperadas malos	630	105	265	
(observadas - esperadas)^2/esperadas	19,206	0,026	47,055	
	172,857	0,238	423,491	
S^2	662,873			
$X^2_{(2)\ \alpha=95\%}$	5,99			

Fuente: Thomas, L., Edelman, D., Crook, J. Credit Scoring and Its Applications.
Elaboración propia

En ambos casos se rechaza la Ho, es decir no se puede inferir que exista independencia entre los atributos y las clasificaciones. Para la selección de la mejor agrupación, basados en la regla de decisión, escogeremos aquella que presente el mayor valor del estadístico. Para este ejemplo seleccionaríamos 'propietario', 'dependencia', 'otros'.

Una segunda opción, más recomendable, es evitar partir de agrupaciones a criterio del investigador analizando el aporte de cada atributo al rechazo de la Ho. De esta forma se podrán agrupar aquellas características que muestren poca dependencia o que no permitan diferenciar con claridad entre las clasificaciones.

Para el ejemplo anterior se tiene la siguiente tabla de contingencia:

	Propietario	Arriendo amoblado	Arriendo desamoblado	Dependencia	Otro	No responde	Total
Buenos	6000	1600	350	950	90	10	9000
Malos	300	400	140	100	50	10	1000
Total	6300	2000	490	1050	140	20	10000
Esperadas buenos	5670	1800	441	945	126	18	
Esperadas malos	630	200	49	105	14	2	
(observadas - esperadas)^2/esperadas	19,206	22,222	18,778	0,026	10,286	3,556	
	172,857	200,000	169,000	0,238	92,571	32,000	
S^2	740,7						
$X^2_{(5)\ \alpha=95\%}$	11,07						

Fuente: Thomas, L., Edelman, D., Crook, J. Credit Scoring and Its Applications.

Elaboración propia

Note que los atributos 'dependencia'; 'otro'; y 'no responde' son los que menor contribución tienen al rechazo de la Ho. (en coincidencia con el resultado obtenido de aplicar el método anterior) En consecuencia serían los atributos a agrupar.

Ejemplo: Institución financiera. Característica: Estado civil. # atributos 8

	Casado	Casado sin inf cóny.	Divorciado	Separado	Soltero	Unión libre sin inf cóny.	Unión libre	Viudo	Total
Malo	7307	10915	904	683	13651	2104	2245	504	38313
Bueno	495	1042	108	105	1953	369	324	47	4443
Total	7802	11957	1012	788	15604	2473	2569	551	42756
Esperadas malos	6991,3	10714,5	906,8	706,1	13982,5	2216,0	2302,0	493,7	
Esperadas buenos	810,7	1242,5	105,2	81,9	1621,5	257,0	267,0	57,3	
(observadas - esperadas)^2/ esperadas	14,26	3,75	0,01	0,76	7,86	5,66	1,41	0,21	
	122,97	32,36	0,08	6,52	67,77	48,83	12,19	1,84	
B:M	6,8%	9,5%	11,9%	15,4%	14,3%	17,5%	14,4%	9,3%	11,6%
S^2	326,48								
$X^2_{(7)\ \alpha=95\%}$	14,07								

Elaboración propia

Si bien la razón buenos:malos distribuida en los distintos atributos parece reflejar la relación poblacional, puede requerirse tener un menor número de atributos a efectos de simplificar la aplicación del scorecard, sin perder atributos de variables que sean buenos discriminadores. De lo contrario, por ejemplo, podríamos encontrarnos durante el desarrollo inicial del scoring con 30 variables (cada una con 10 atributos) y por ende con 300 categorías. Entonces, si el objetivo es finalizar con un modelo que sea comprensible y por ende aceptado por la administración puede ser preferible disminuir tanto el número de atributos como de características.

	Casado	Casado sin inf cóny.	Separado	Soltero	Unión libre sin inf cóny.	Unión libre	Viudo	Total
Malo	7307	10915	1587	13651	2104	2245	504	38313
Bueno	495	1042	213	1953	369	324	47	4443
Total	7802	11957	1800	15604	2473	2569	551	42756
Esperadas malos	6991,3	10714,5	1613,0	13982,5	2216,0	2302,0	493,7	
Esperadas buenos	810,7	1242,5	187,0	1621,5	257,0	267,0	57,3	
(observadas - esperadas)^2/esperadas	14,26	3,75	0,42	7,86	5,66	1,41	0,21	
	122,97	32,36	3,60	67,77	48,83	12,19	1,84	
B:M	6,8%	9,5%	13,4%	14,3%	17,5%	14,4%	9,3%	11,6%
S^2	323,14							
$X^2_{(6)\ -\ 95\%}$	12,59							

Elaboración propia

Usando el segundo método de agrupamiento de atributos se observa que el aporte de las categorías 'separado' y 'divorciado' es relativamente bajo. Por lo que la aplicación del criterio de Pearson sugiere la agrupación de estos dos atributos en uno nuevo que se denomina 'separado'. Esta nueva conformación de atributos mejora la dependencia de la característica a la clasificación de clientes dada.

Una vez agrupados los atributos de las características, el siguiente paso es descartar las variables que sean malas predictoras o aquellas que son muy dependientes de otras variables. Esto es importante por las mismas razones mencionadas cuando se explicaba la importancia de agrupar atributos. Es decir, si bien se puede haber reducido a 3 el número de atributos por variable, puede tenerse para iniciar el modelo unas 200 características que implicarían unos 600 atributos diferentes. En estos casos el criterio Chi cuadrado puede emplearse para descartar aquellas variables que resulten pobres predictoras.

5.1 Clasificaciones finas, duras, variables cruzadas y árboles de decisión

Las técnicas tradicionales arribas mencionadas pueden ser sintetizadas mediante un único algoritmo de clasificación que permita realizar particiones repetitivas de

subconjuntos de un conjunto inicial X[25] con base en una pregunta en la cual se basa la división del nodo[26] (en el caso del riesgo de crédito la pregunta es si la variable escogida permite discriminar entre la población de buenos y malos[27]). Este es el caso de los árboles de decisión o de clasificación[28], los cuáles son métodos de aprendizaje inductivo supervisado no paramétrico, que destacan por su sencillez e interpretación.

Los árboles de decisión pueden entenderse como la estructura (conjunto de reglas de decisión) resultantes de la partición recursiva del espacio de representación a partir del conjunto de partida. Existen muchos algoritmos para árboles de decisión y su diferencia estriba en la estrategia de podar los árboles, las reglas para particionar los árboles y el tratamiento de valores perdidos[29]. Cuando se emplea el algoritmo CHAID (Chi-squared Automatic Interaction Detection) la regla de partición en este caso es un algoritmo que usa el estadístico Chi cuadrado. En este caso y como vimos en el apartado de selección de atributos y variables, el empleo del estadístico chi cuadrado de Pearson permite en cada nodo evaluar la dependencia de la variable en el nodo respecto a todas las variables incluidas en el modelo, siendo no sólo un ahorro en términos de tiempo, sino una mejora en eficiencia pues cada vez realiza tantas tablas de contingencia como variables existan en el modelo tanto a nivel de variable dependiente con sus covariantes como entre las covariantes (variables cruzadas). Además el algoritmo agrupa las categorías (realiza automáticamente la clasificación gruesa de atributos), permitiendo tener un criterio de selección basado en una regla estadística en lugar del sólo criterio del investigador. En términos simples la ventaja de emplear este algoritmo es que *permite encontrar no sólo el conjunto de variables que mejor particiona a la variable objetivo (bueno, malo), sino que agrupa los atributos dentro de cada variable,*

---

[25] Se define una partición de la siguiente manera: Sea X un conjunto finito y sean x_1, x_2, ...x_p subconjuntos distintos de X tal que x1 U x2 U x3 U....xp sea igual al conjunto X, entonces diremos que $\{x_1, x_2, ...x_p\}$ forman una partición de X

[26] Un nodo es un conjunto de datos. Cuando se inicia la partición de un conjunto de datos el nodo se conoce como inicial. Este muestra el conjunto de información seleccionado y agrupado en función de la variable dependiente u objetivo.

[27] En realidad la pregunta es una hipótesis que testea la independencia entre variables, es decir que busca mostrar si no existe relación de dependencia entre la variable seleccionada y la clasificación de buenos y malos en la población. De existir independencia la variable no debe ser incluida en el modelo.

[28] El término árboles es por la gráfica, la cual parte de un nodo inicial o raíz y a partir del cual se realizan las primeras particiones o divisiones, en cada nodo subsiguiente se sigue el mismo procedimiento hasta llegar a un nodo terminal u hoja. (o sin particiones)

[29] Entre los principales están: C4.5, CHAID, NewId, CART, Arboles Bayesianos y el CN2.

logrando un doble resultado: reducir el número de atributos en cada variable a la vez que determina el conjunto de covariantes con mayor grado de significación[30].

La elaboración de un árbol de decisión se puede resumir en tres pasos:

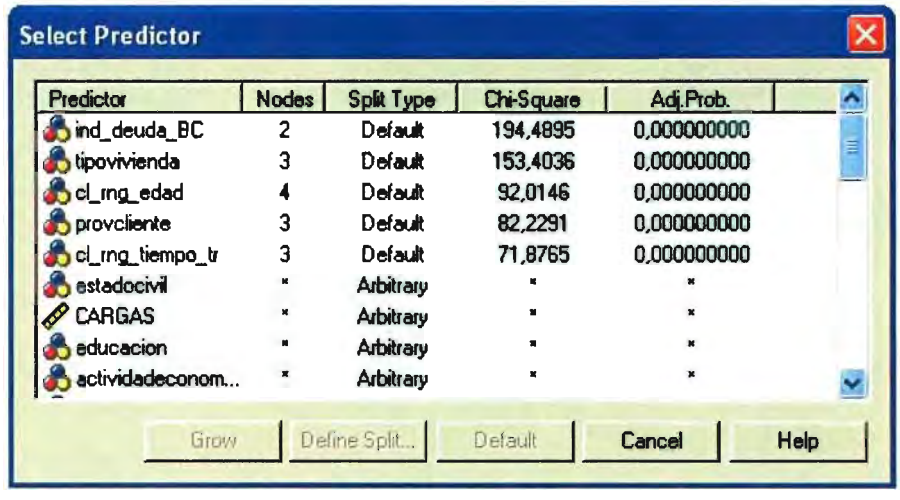
1.- Selección de la partición

2.- Decidir sobre la continuidad de la división del árbol o su parada

3.- La asignación a cada nodo terminal de una clase[31]

Lo primero que se necesita determinar es la variable objetivo a partir de la cual realizaremos las particiones. Como lo que interesa es la discriminación de las poblaciones de buenos y malos, esta se convierte en nuestra variable meta, en donde el bueno y el malo se adaptan a la definición establecida por la institución, por ejemplo bueno es el que tiene 15 días de atraso promedio y 25 de atraso máximo, luego el algoritmo CHAID selecciona la variable que determina la partición, esto constituye el primer paso en la construcción del árbol.

GRUPO DE VARIABLES SELECCIONADAS POR EL ALGORITMO CHAID



Fuente: Institución Financiera Ecuatoriana

Elaboración propia

[30] El anexo 2 "Funcionamiento de los árboles de decisión" muestra como se desarrolla el algoritmo Chi cuadrado.

[31] En la aplicación implementada en este trabajo, no se busca identificar una clase en el sentido estricto de la teoría de los árboles de decisión, sino únicamente buscar las variables más relacionadas con la variable objetivo y su agrupamiento.

Como muestra la imagen adjunta el algoritmo ordena las variables en función del valor del estadístico de prueba chi-cuadrado, mostrando las variables que mayor grado de dependencia muestran con la variable objetivo (bueno malo)[32].

El siguiente paso es particionar el conjunto de partida para cada una de las variables escogidas por el árbol buscando identificar tanto su valor estadístico como el criterio de agrupamiento escogido por el árbol.

VARIABLE NIVEL DE EDUCACIÓN Y AGRUPAMIENTO ESCOGIDO POR CHAID

Node 0	Malo		26,81	11749	
	Bueno		73,19	32070	
Nivel de educación (Adj. P-value=0,0000, Chi-square=154,3355, df=1)					
Primaria, secundaria, técnica	**Node 1**	Malo	27,78	10719	
		Bueno	72,22	27865	
Universitario	**Node 2**	Malo	19,68	1030	
		Bueno	80,32	4205	

Fuente: Institución Financiera Ecuatoriana

Elaboración propia

Para la variable nivel de educación se observa un estadístico de prueba igual a 154,33 y p-value de 0 por lo que la variable muestra dependencia con la variable objetiva, es decir es un buen discriminante. Note ahora que esta variable presenta cuatro atributos para la educación: primaria, secundaria, técnica y universitaria, pero el algoritmo ha reducido el número a dos, el primero conformado por la agrupación: primaria, secundaria y técnica y el segundo con el nivel de educación universitario, este agrupamiento sigue la misma lógica explicada en el apartado selección de características. El árbol está sugiriendo que los clientes con atributo 'educación universitaria' son mejores pagadores que los clientes con cualquier otro nivel educativo pues su porcentaje de malos se ha reducido del 26,81% mostrado en el nodo raíz a un 19.68% en este nodo. El agrupamiento que se puede definir como otro nivel educativo no muestra un cambio en la distribución de buenos y malos de la muestra pero en

[32] Los asteriscos en la columna Chi-square de la imagen indican valores pequeños del estadístico en relación a los cinco más importantes, no significa que la variable no sea significativa.

relación al atributo universitario sí. Siguiendo este ejemplo se debe incluir en el modelo de calificación de clientes una variable (que puede ser dicotómica) que indique si el cliente tiene educación universitaria. (recordando a las variables nominales sería educación: universitario = 1; otra educación = 0)

Adicionalmente el árbol de decisión destaca los mejores cruces de variables que se pueden realizar, pues recordando su definición, en cada nodo vuelve a realizar particiones ordenando las variables por aquellas que presentan un mejor estadístico Chi-cuadrado. Note que esta ventaja es sobre el criterio subjetivo que puede ofrecer el investigador al momento de definir los cruces entre variables, pues la elección de la variable de cruce viene determinada por el valor del Chi-cuadrado.

VARIABLE NIVEL DE EDUCACIÓN Y CRUCES ESCOGIDOS POR CHAID

Node 0	Malo	26,81	11749			
	Bueno	73,19	32070			

Educación (Adj. P-value=0,0000, Chi-square=154,3355, df=1)

Primaria, secundaria, técnica, ninguna	**Node 156**	Malo	27,78	10719		
		Bueno	72,22	27865		

Ciudad (Adj. P-value=0,0000, Chi-square=914,3273, df=2)

Cuenca, Quito	**Node 163**	Malo	21,83	3268
		Bueno	78,17	11700
Guayaquil, Ibarra, Riobamba	**Node 164**	Malo	33,63	7186
		Bueno	66,37	14183
Ambato, Latacunga	**Node 165**	Malo	11,79	265
		Bueno	88,21	1982

Universitario	**Node 157**	Malo	19,68	1030
		Bueno	80,32	4205

Género (Adj. P-value=0,0000, Chi-square=49,5335, df=1)

Masculino	**Node 161**	Malo	23,58	611
		Bueno	76,42	1980
Femenino	**Node 162**	Malo	15,85	419
		Bueno	84,15	2225

Fuente: Institución Financiera Ecuatoriana

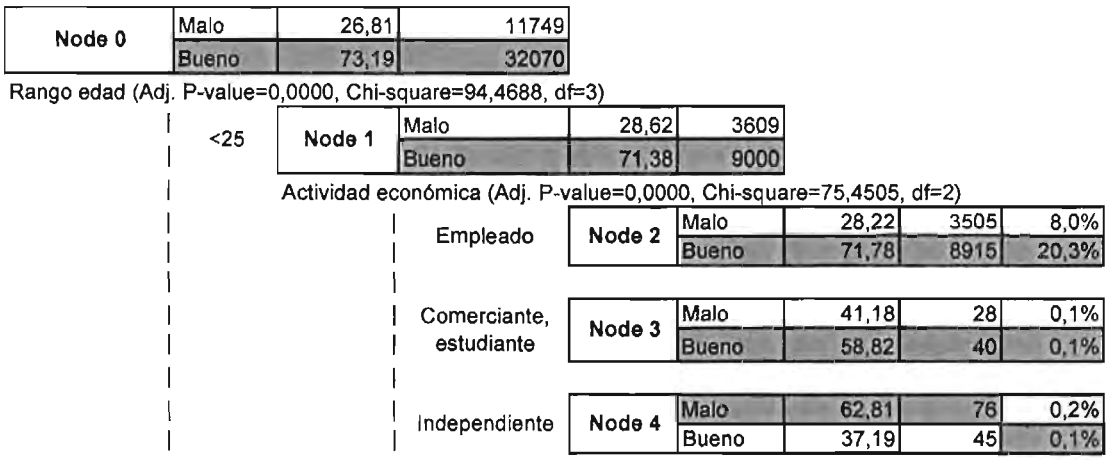
Elaboración propia

En este ejemplo, el agrupamiento realizado por el árbol para el nivel de educación se particiona por la variable ciudad para el caso de la educación primaria,

secundaria, técnica y ninguna y por el género para la variable universitario. El análisis es similar al arriba expuesto. En el caso de los clientes con educación universitaria y de género femenino, el comportamiento de pago es mejor que el de los hombres pues su distribución buenos-malos en relación a la del nodo raíz ha cambiado de 26% a 15% de malos. En el caso de los hombres con educación universitaria esta distribución presenta una pequeña mejora del número de malos pero no muy marcada como para esperar que esta característica sea significativa y de gran valor predictivo en un modelo de calificación crediticia. Siguiendo este ejemplo se debería incluir una variable compuesta o cruzada que indique si el cliente tiene educación universitaria y género femenino.

Es importante mencionar que si bien el uso de árboles de decisión permite al investigador obtener tanto clasificaciones finas, gruesas, como cruces de variables no se debe descuidar la distribución de la población en cada nodo, no se pueden seleccionar atributos para ser incluidos es un modelo si son poco representativos en la población en estudios. Por ejemplo en el cruce de variables 'rango edad' y 'actividad económica' se observa que los clientes con edad inferior a 25 años y con actividad económica comerciante o estudiante muestran una distribución de buenos y malos de 58/41 es decir que estos clientes tienden a ser mucho más malos que el promedio de la población (distribución buenos-malos 73/26). No obstante, se está olvidando revisar el número de clientes asignados a ese nodo respecto al total muestreado es de 68 o tan sólo un 0.15% del total, por lo que, si la muestra ha sido seleccionada aleatoriamente, esta variable creada no tiene representatividad en la población y su inclusión no tiene sentido estadístico.

REPRESENTATIVIDAD DE LA VARIABLE ESCOGIDA POR CHAID



Fuente: Institución Financiera Ecuatoriana

Elaboración propia

Con base en lo expuesto en este apartado se recomienda emplear árboles de decisión como regla de selección, agrupamiento y cruce de variables a incluir en el modelo scoring no sólo por la reducción en tiempo de elaboración de estadísticos sino por su capacidad de análisis de tablas de contingencia (en el caso del algoritmo CHAID) para todas las variables en cada nodo, permitiendo elegir el mejor conjunto de variables predictoras tanto individuales como cruzadas.

Selección del algoritmo scoring

Una vez definido el conjunto de variables que ingresarán al proceso de modelado es necesaria la elección de un algoritmo que permita estimar el comportamiento de pago futuro del cliente. Existen varias técnicas estadísticas para construir scorecards de crédito como el análisis discriminante, las redes neuronales, los algoritmos genéticos, los modelos lineales de probabilidad, los árboles de decisión, los modelos lineales generalizados, etc[33]. En este estudio se opta por el modelo de

[33] Una explicación sucinta de algunas de estas técnicas se puede revisar en el libro de Thomas, Edelman y Crook, 'Credit Scoring and its Applications' en el capítulo 4, páginas 42-61.

regresión logística, técnica estadística que forma parte de los modelos conocidos como lineales generalizados[34].

La regresión logística o modelo logit es una técnica estadística que encuentra la probabilidad de ocurrencia de un evento (en este caso que el cliente sea buen pagador). La estimación de esta probabilidad sigue la siguiente expresión matemática:

$$P(bueno) = \frac{\exp^z}{1+\exp^z}$$

En donde Z es:

$$z = \log\frac{P}{1-p} = \alpha + \beta_1 x_1 + ...\beta_p x_p$$

La forma más sencilla de entender esto es mediante un ejemplo sencillo, en donde se asume que se ha obtenido el valor de z mediante la siguiente ecuación de regresión:

$$\text{Modelo}: z = 0.28 + 0.026 * \text{Edad} - 1.7 * \text{Tiene deuda}$$

Si se acerca un cliente a solicitar un crédito, el analista deberá pedirle información sobre su edad y verificar en una base externa si tiene deuda en el sistema financiero. El cliente entrega esta información: Edad = 35; Sí tiene deuda (Tiene deuda = 1). El puntaje obtenido como calificación para este cliente es:

$$z = 0.28 + 0.026 * 35 - 1.7 * 1 = -0.51$$

$$P(bueno) = \frac{\exp^{-0.51}}{1+\exp^{-0.51}} = 0.38$$

En la práctica se acostumbra a multiplicar este valor de probabilidad por 1000 de tal forma que el puntaje de este cliente es de 380 puntos y dependiendo del puntaje límite establecido para los clientes buenos la institución considerará entregarle o no el

[34] Una explicación detallada de la importancia de los modelos lineales generalizados en la elaboración de modelos de elección binaria se realiza en el anexo 3.- Los modelos lineales generalizados como método estadístico para la construcción de credit scorecards.

crédito. Por ejemplo si la institución ha fijado que los clientes buenos son aquellos que tienen 500 puntos o más, este cliente sería rechazado.

6.- Validación del modelo

6.1.- Análisis de percentiles (Prueba de Kolmogorov-Smirnov)

6.2.- Indice de Gini

6.3.- Rendimiento

Una vez elaborado el modelo estadístico es necesario verificar su fortaleza, para esto está de más recordar la importancia que tienen los estadísticos t y las pruebas de significancia conjunta de los modelos. Sin embargo, en el diseño de modelos scoring existen algunos índices adicionales que son de gran importancia y de frecuente uso: la medida de Kolmogorov-Smirnov, la curva ROC y el índice de Gini[35]

6.1 Prueba Kolmogorov-Smirnov

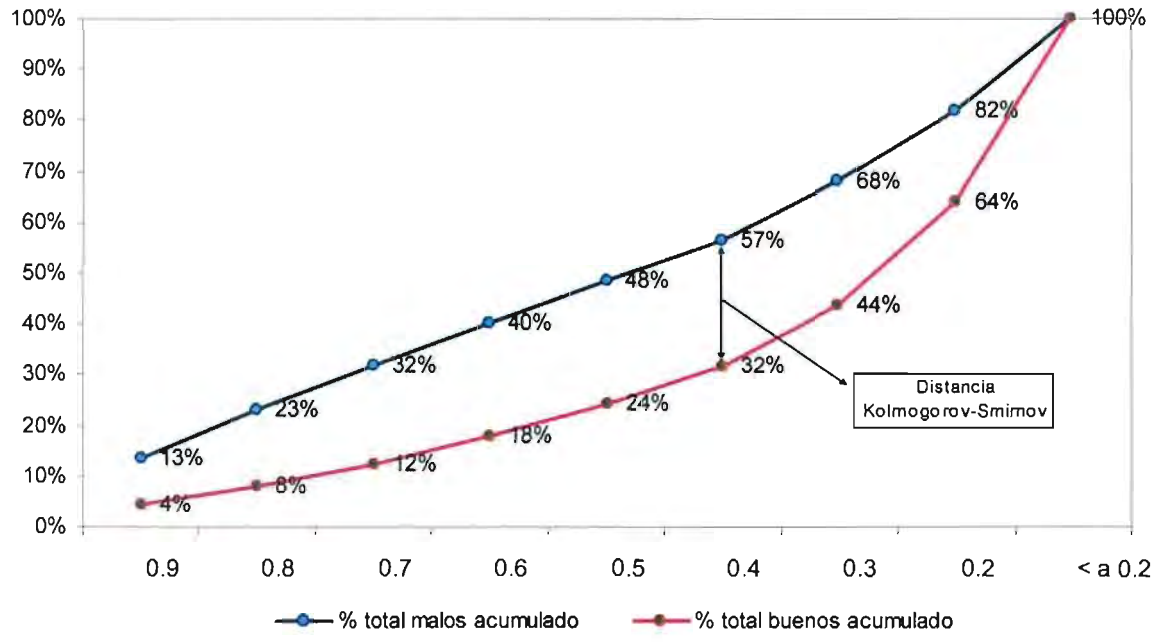
Esta prueba se basa en una medida de separación y consiste en medir cuán distintas son las funciones de de distribución de buenos y malos clientes para cada rango percentil del puntaje score.

En términos formales si **Pb(score)** = $\Sigma_{x<=score}$ **pb(x)** y **Pm(score)** = $\Sigma_{x<=score}$ **pm(x)** entonces el estadístico será:

$$KS = \max_{s} | Pb(score) - Pm(score) |$$

En términos gráficos KS es el largo de la línea punteada medida en el rango score que maximiza la separación entre las dos funciones de distribución.

[35] Una explicación de la importancia de la medición de las curvas ROC se despliega en el anexo 4.- La curva ROC y el coeficiente de GINI.



Fuente: Institución Financiera Ecuatoriana

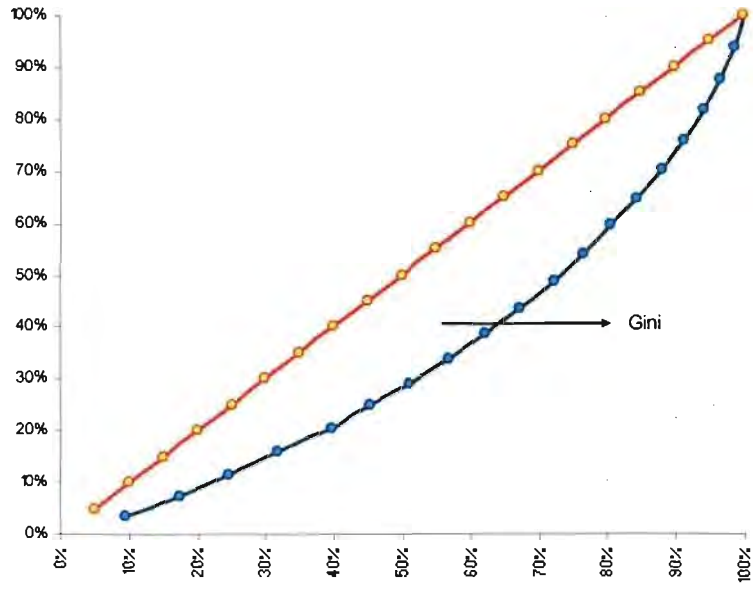
Elaboración propia

Como medida de comparación se sabe por reuniones mantenidas con Equifax (empresa internacional especializada en diseño de modelos scoring) que en El Salvador el modelo alcanza valores de 19 puntos porcentuales. En cambio para la consultora LISIM un buen valor en Colombia para un scoring de aprobación debe estar entre 28 y 35 puntos.

6.2 Curva ROC y el coeficiente de GINI

La curva ROC a diferencia de la medida Kolmogorov-Smirnov muestra la información en una sola curva graficando la distribución acumulada de buenos contra la distribución acumulada de malos para cada rango percentil del puntaje score.

CURVA ROC



Fuente: Institución Financiera Ecuatoriana

Elaboración propia

Esta curva describe la propiedad de clasificación del scorecard en la medida en que cada punto de corte varía. El mejor scorecard posible tendrá un ROC que va a lo largo del eje horizontal antes de subir por el eje vertical. Una curva ROC que pase por la línea roja de la gráfica corresponderá a una en laque la probabilidad para cada score de ser bueno es exactamente igual a la probabilidad de ser malo, por lo que el ratio B:M es el mismo para todos los rangos score. Note que la ocurrencia de esto no es lo más apropiado pues no representa ninguna ventaja en relación a clasificar aleatoriamente dado que se conoce el ratio buenos malos de la población.

Por lo tanto mientras más alejada de la diagonal esté la curva ROC es mejor el scorecard. Si un scorecard tiene una ROC que siempre es más alejada que otra, entonces el primero domina para todos los puntos de corte del score.

Si mientras más lejos de la diagonal se encuentra una curva ROC es mejor el poder de clasificación del modelo, entonces mientras mayor sea el área formada por estas dos curvas mejor también será la clasificación del scorecard.

La medida que recoge el área bajo la curva se conoce como el coeficiente de Gini. Realmente este índice se define como dos veces el área formada por la diagonal y

una curva ROC. Este índice tendrá la propiedad de que en el caso de una perfecta clasificación su valor será de 1, mientras que una clasificación aleatoria con una curva ROC sobrepuesta a la diagonal tendrá un valor de índice igual a 0. ***En sí el Gini da un número resumen del desempeño del scorecard para todos los puntos de corte del score.***

6.3 Rendimiento[36]

Una medida diferente a los índices estadísticos explicados anteriormente es el cálculo del rendimiento promedio por rango de score. El cálculo del rendimiento sigue la lógica empleada en la evaluación de proyectos de inversión, es decir busca resumir en un indicador financiero las diferencias entre alternativas de inversión, en este caso las disimilitudes entre prestar a clientes con bajo puntaje y alto puntaje.

El método empleado para calcular el rendimiento es la tasa interna de retorno TIR) y se espera que en la medida en que se incrementan los puntajes en cada percentil score la TIR también se vaya incrementado, indicando que el modelo scoring está ordenando correctamente pues sugiere, ahora en términos financieros, que si se debe elegir a quien prestar debe ser a los de mayor score o más alto rendimiento.

La formulación del indicador consiste en

1.- Obtener los rendimientos de la cartera por rango de atraso promedio

2.- Para cada rango percentil ponderar los rendimientos por rango de atraso por la participación de clientes en cada rango de atraso.

3.- Obtener el rendimiento por rango percentil como la sumatoria de los rendimientos ponderados en casa rango de atraso.

[36] Esta metodología de evaluación es desarrollo del área de riesgos de la Institución Financiera Ecuatoriana que colaboró con la información para la generación de este modelo.

RENDIMIENTO PARA UNA CARTERA DE CONSUMO RANQUEADA POR PUNTAJE SCORE

Percentil	Score Mín	Score Máx	Atraso Promedio					Rendimiento
			0-4	5-30	31-60	>60	Total	
0% - 5%	76	202	46,4%	25,0%	10,5%	18,2%	100%	15,69%
11% - 15%	215	256	50,2%	24,9%	8,1%	16,8%	100%	17,30%
16% - 20%	266	297	55,2%	20,0%	9,8%	15,1%	100%	18,68%
21% - 25%	308	336	59,0%	20,1%	9,0%	11,8%	100%	21,11%
26% - 30%	345	366	61,5%	20,6%	6,9%	11,0%	100%	22,22%
31% - 35%	371	394	61,7%	20,9%	5,3%	12,0%	100%	21,86%
36% - 40%	399	422	60,7%	23,8%	6,7%	8,8%	100%	23,40%
41% - 45%	424	444	6	Bajo porcentaje de buenos		Alto porcentaje de malos	100%	23,74%
46% - 50%	446	467	7				100%	30,32%
51% - 55%	475	496	63,2%	22,5%	7,0%	7,3%	100%	24,52%
56% - 60%	502	518	66,0%	21,4%	4,7%	7,9%	100%	24,85%
6% - 10%	526	544	60,9%	17,2%	3,0%	9,0%	100%	24,79%
61% - 65%	550	574	7	Alto porcentaje de buenos		Bajo porcentaje de malos	100%	27,54%
66% - 70%	577	597	72,5%	16,7%	5,4%	5,4%	100%	26,99%
71% - 75%	601	634	72,1%	18,5%	5,3%	4,2%	100%	27,66%
76% - 80%	641	666	71,6%	17,6%	5,2%	5,6%	100%	26,77%
81% - 85%	675	703	77,1%	17,3%	1,8%	3,8%	100%	29,00%
86% - 90%	713	749	75,4%	16,1%	5,3%	3,2%	100%	28,63%
91% - 95%	762	817	77,3%	17,9%	3,2%	1,6%	100%	30,10%
96% - 100%	833	1000	86,2%	12,4%	1,3%	0,0%	100%	32,39%
Total general			**66,8%**	**19,6%**	**5,8%**	**7,8%**	**100%**	**24,89%**

Rendimiento creciente

Fuente: Institución Financiera Ecuatoriana

Elaboración propia

En la tabla anterior se observa como se puede elaborar un cuadro de presentación del puntaje score en función del rendimiento generado por la cartera para cada percentil.

Los números son muy explícitos en recoger comportamientos de pagos de clientes por score. Si se observa la distribución total de la población se ve en los extremos que un 66.8% de clientes tiene atraso promedio menor a 4 días, mientras que un 7.8% cae en el rango mayor a 60 días. Se espera que un buen modelo afecte la distribución de clientes por cada rango de atraso promedio para cada score. En la tabla se observa que hacia los puntajes bajos la población decae en 20 puntos porcentuales en relación a la distribución total y en 40 puntos porcentuales respecto al puntaje más alto. De igual forma ocurre en el otro extremo, el porcentaje de malos en el puntaje más alto es 7 puntos porcentuales menor que el de la población y 11 puntos porcentuales menor que la distribución en el puntaje de score más bajo.

Esta medida de rendimiento no sólo permite establecer en términos financieros la capacidad de ordenamiento del modelo sino que determina que estrategias en cuanto a cupos se pueden tomar para diferentes puntos de corte definidos en el modelo o grupos de clientes formados con base en su rentabilidad o pérdida esperada.

7.- Diseño de la scorecard

Una scorecard o tabla de puntaje es una tabla que resume la información sobre los pesos y signos asignados a las características relevantes para identificar el comportamiento de pago de un cliente. En su forma más simple es una tabla en laque se listan las variables seleccionadas y sus atributos con el puntaje asignado a cada uno. Un puntaje negativo indica que la variable castiga al cliente que cumpla con este atributo, mientras que una variable positiva estaría premiándolo. Un mayor o menor valor o peso significa cuán importante es la variable en la calificación.

TABLA SCORECARD

Variables	Atributo	Puntaje
Edad	<25	-20
	26-40	8
	>40	15
Estado civil	Casado	35
	Soltero	-16
Sueldo	>250	12
	<=250	-23

Elaboración propia

8.- Determinación del punto de corte

El punto de corte o cut off es la puntuación mínima admisible para el otorgamiento de un crédito. Siendo calculado en función de la política de riesgo, es decir del grado de exposición que quiera tener la institución una vez evaluada la relación de aceptación y morosidad aceptada.

Se recomienda emplear el criterio de rentabilidad para determinar el punto de corte, fijándolo como el nivel de pérdida que no está dispuesto a asumir la institución financiera.

En la tabla adjunta se muestra la pérdida por rango percentil, si la institución ha tomado como política no prestar a aquel segmento de clientes que tiene una tasa de pérdida inferior a 8.20% entonces se estaría descartando a todos los clientes con un puntaje inferior a 297 puntos (cuadrado rojo), claro que este valor debe ir relacionado con el porcentaje de rechazos que se está asumiendo, recordando que desplazamientos

hacia debajo de los puntos de corte implican una menor tasa de rechazo o una mayor tasa de aceptación, pero a costa de un mayor número de buenos clasificados como malos, es decir a un mayor grado de error.

DETERMINACIÓN DEL PUNTO DE CORTE

Percentil	Score Mín	Score Máx	Atraso Promedio					Rendimiento	Pérdida
			0-4	5-30	31-60	>60	Total		
0% - 5%	76	202	46,4%	25,0%	10,5%	18,2%	100%	15,68%	11,46%
11% - 15%	215	256	50,2%	24,9%	8,1%	16,8%	100%	17,30%	10,51%
16% - 20%	266	297	55,2%	20,0%	9,8%	15,1%	100%	18,68%	9,68%
21% - 25%	308	336	59,0%	20,1%	9,0%	11,8%	100%	21,11%	8,20%
26% - 30%	345	366	61,5%	20,6%	6,9%	11,0%	100%	22,22%	7,52%
31% - 35%	371	394	61,7%	20,9%	5,3%	12,0%	100%	21,86%	7,74%
36% - 40%	399	422	60,7%	23,8%	6,7%	8,8%	100%	23,40%	6,79%
41% - 45%	424	444	61,7%	23,2%	6,6%	8,5%	100%	23,74%	6,59%
46% - 50%	446	467	77,3%	18,8%	2,6%	1,3%	100%	30,32%	2,48%
51% - 55%	475	496	63,2%	22,5%	7,0%	7,3%	100%	24,52%	6,10%
56% - 60%	502	518	66,0%	21,4%	4,7%	7,9%	100%	24,85%	5,90%
6% - 10%	526	544	69,8%	17,2%	3,9%	9,0%	100%	24,79%	5,94%
61% - 65%	550	574	71,2%	17,9%	7,2%	3,7%	100%	27,54%	4,22%
66% - 70%	577	597	72,5%	16,7%	5,4%	5,4%	100%	26,99%	4,57%
71% - 75%	601	634	72,1%	18,5%	5,3%	4,2%	100%	27,66%	4,14%
76% - 80%	641	666	71,6%	17,6%	5,2%	5,6%	100%	26,77%	4,70%
81% - 85%	675	703	77,1%	17,3%	1,8%	3,8%	100%	29,00%	3,31%
86% - 90%	713	749	75,4%	16,1%	5,3%	3,2%	100%	28,63%	3,54%
91% - 95%	762	817	77,3%	17,9%	3,2%	1,6%	100%	30,10%	2,61%
96% - 100%	833	1000	86,2%	12,4%	1,3%	0,0%	100%	32,39%	1,16%
Total general			66,8%	19,6%	5,8%	7,8%	100%	24,89%	5,87%

Fuente: Institución Financiera Ecuatoriana

Elaboración propia

IV.- Aplicación práctica: desarrollo de un modelo scoring de aprobación en una institución financiera de créditos de consumo.

> "En gran medida, la econometría no es una ciencia, definida por un estrecho conjunto de teoremas, sino más bien constituye un enfoque que sólo puede ser apreciado y asimilado en toda su magnitud a través de su empleo cotidiano".
>
> Michael D. Intriligator

Una vez definida la estructura que debe seguirse para desarrollar un modelo scoring se procede a mostrar los resultados obtenidos de la aplicación en una institución financiera ecuatoriana dedicada al otorgamiento de créditos de consumo.

Se sigue la estructura metodológica mostrada en el apartado 3, pero se obvia la explicación de cada numeral pues se desarrollo ampliamente en los capítulos precedentes.

1.- Selección de la muestra

Análisis de cosechas

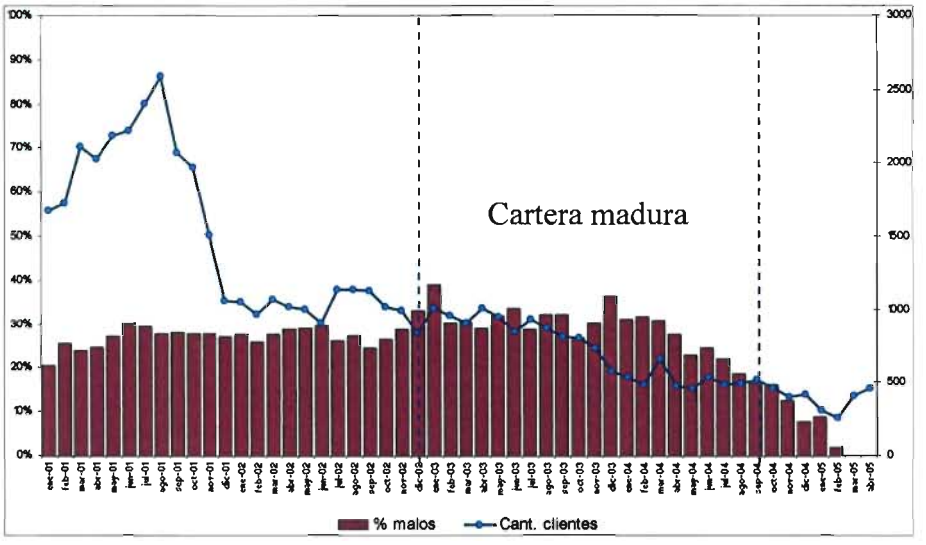
Para la selección de la muestra se construyó el indicador tasa de morosidad definiendo al malo como el cliente que ha incurrido alguna vez, en su primera operación crediticia, en una mora máxima de 30 días.

Bajo esta definición las gráficas y los resultados estadísticos que testean la existencia de una probabilidad uniforme[37] junto con la ventana temporal de estudio elegida se muestra a continuación:

[37] Esta prueba se conoce como Kolmogorov-Smirnov y testea la probabilidad de que una distribución empírica se ajuste a una distribución teórica específica, para el caso de estudio el interés es que la distribución sea uniforme.

Producto: Crédito de consumo

Periodo de madurez de la cartera: diciembre 2002 – abril 2004



Fuente: Institución Financiera Ecuatoriana

Elaboración propia

La tabla debajo muestra los diferentes periodos analizados, en donde LCC indica el nombre del producto; los grupos LCCmalo, LCC05044 y LCC12044, son la muestra total, la muestra entre mayo de 2003 y abril de 2004 y la muestra entre diciembre de 2003 y abril de 2004.

La prueba de Kolmogorov-Smirnov realizada para distintos periodos muestra que la tercera ventana temporal se ajusta mejor a una distribución uniforme (regla empírica que sugiere un buen ajuste cuando la significación asintótica está por encima de 0,5). El periodo comprendido en esta muestra va desde diciembre de 2002 hasta abril de 2004.

Prueba de Kolmogorov-Smirnov para una muestra

		LCCMALO	LCC05044	LCC12044
N		60	34	15
Parámetros uniformes[a,b]	Mínimo	,000	,25	,27
	Máximo	,336	,34	,34
Diferencias más extremas	Absoluta	,351	,154	,196
	Positiva	,033	,120	,091
	Negativa	-,351	-,154	-,196
Z de Kolmogorov-Smirnov		2,716	,895	,759
Sig. asintót. (bilateral)		,000	,399	,612

a. La distribución de contraste es la Uniforme.

b. Se han calculado a partir de los datos.

Paquete estadístico empleado: SPSS versión 11.5

Fuente: Institución Financiera Ecuatoriana

Elaboración propia

2.- Definición de buenos y malos

Una vez seleccionada la ventana temporal o muestra de análisis se procedió a definir el indicador de buenos y malos con base en la matriz atraso promedio atraso máximo.

MATRIZ ATRASO PROMEDIO ATRASO MÁXIMO

Rango Atraso Max	Rango Atraso Prom							Total general
	0	1 - 15	16 - 30	31 - 60	61 - 90	91 - 120	> 120	
0	3.386							3.386
1 - 15	7.878	11.560						19.438
16 - 30	15	5.645	10					5.670
31 - 60		4.886	730	30				5.646
61 - 90		980	1.793	492	3			3.268
91 - 120		103	851	987	47			1.988
> 120		4	276	1.775	1.519	1.119	2.501	7.194
Total general	11.279	23.178	3.660	3.284	1.569	1.119	2.501	46.590

Fuente: Institución Financiera Ecuatoriana

Elaboración propia

En esta matriz se observa que el 61,1% de la población empleada para la construcción de esta tabla se encuentra entre los 15 días de atraso promedio y atraso máximo, la elección de este par de valores fue considerado muy ácido para la institución, pues en su plan comercial tienen planificado realizar campañas agresivas de ventas, por lo que están dispuestos a asumir un mayor nivel de riesgos. Bajo este parámetro de decisión se optó por ampliar la definición de buenos a 30 días de atraso promedio y 60 días de atraso máximo. Además se observó que dado el esquema actual

de entrega de créditos la distribución de la población tiende a ubicarse en rangos de mora bajos por lo que se espera que el scoring refuerce este proceso que de modo no estadístico está arrojando buenos índices de cartera. Bajo esta definición se trabajo con un número de 34.110 buenos y 12.480 malos.

3.- Definición y selección de datos, análisis preliminar de los datos y análisis mutlivariado

Uno de los puntos que se discutían en el apartado 1 era la necesidad de contar con una sólida base de datos. La institución con la que se trabajó cuenta con una base de datos certificada por una empresa consultora internacional, por lo que los datos guardados eran tanto de alta fiabilidad como de alta calidad. Se encontró sin embargo que 2771 clientes carecían de información en alguna variable, como este porcentaje estaba dentro del 5% máximo permitido de error en una base de datos se procedió a descartarlo. Esta eliminación de variables redujo los clientes en la muestra, ahora la base de datos empleada es de 43.819 clientes con 32.070 buenos y 11.749 malos.

La ventaja del empleo de árboles de decisión ayudó a determinar las variables a incluir en el modelo, sus agrupaciones y los mejores cruces. En el anexo 5 se muestran los resultados arrojados por el árbol de decisión usando el algoritmo CHAID y empleando el paquete estadístico SPSS versión 11.5. La tabla adjunta muestra las variables seleccionadas (construidas como dicotómicas y en función de los atributos que mejor discriminaron a la población de buenos y malos)

VARIABLES A INCLUIR EN EL PROCESO DE MODELADO

Variables Simples	Variables cruzadas
Estado civil	Afiliado al IESS y menos de un años de trabajo
Separado	Afiliado al IESS y de 1 a 3 años de trabajo
Soltero	Afiliado al IESS y más de 4 años de trabajo
Casado	Empleado y menos de 25 años de edad
Cargas familiares	Comerciante y menos de 25 años de edad
Menos de dos cargas	Independiente y menos de 25 años de edad
Más de tres cargas	Empleado y edad entre 26 y 40 años
Nivel de educación	Comerciante y edad entre 26 y 40 años
Universitario	Independiente y edad entre 26 y 40 años
Otra	Empleado y edad entre 41 y 45 años
Tipo de vivienda	Comerciante y edad entre 41 y 45 años
Propia	Independiente y edad entre 41 y 45 años
Arrendada	Empleado y edad mayor de 45 años
Otra	Comerciante y edad mayor de 45 años
Actividad económica	Independiente y edad mayor de 45 años
Empleado	Casa propia y menos de 1 año de trabajo
Comerciante	Casa propia y de 1 a 3 años de trabajo
Independiente	Casa propia y más de 4 años de trabajo
Afiliación al IESS	Casa arrendada y menos de 1 año de trabajo
Género	Casa arrendada y de 1 a 3 años de trabajo
Ciudad del crédito	Casa arrendada y más de 4 años de trabajo
grupo1	Otra vivienda y menos de 1 año de trabajo
grupo2	Otra vivienda y de 1 a 3 años de trabajo
grupo3	Otra vivienda y más de 4 años de trabajo
Deuda en Central de riesgo	Separado y con menos de dos cargas
Rango edad	Separado y con más de dos cargas
<25	Soltero y con menos de dos cargas
26-40	Soltero y con más de dos cargas
41-45	Casado y con menos de dos cargas
>45	Casado y con más de dos cargas
Rango sueldo	
<200	
200-300	
>300	
Rango tiempo de trabajo	
<1	
1-3	
>3	
Rango compromiso financiero	
Sin compromiso	
0-20	
>=20	
Rango tiempo de residencia	
<1	
1-3	
3-5	
>5	
Provincia del cliente	
Grupo Provincia 1	
Grupo Provincia 2	
Grupo Provincia 3	

4.- Corrida del modelo empleando un modelo de regresión logística

Seleccionadas y construidas las variables se procedió a correr los modelos estadísticos en búsqueda del que logre un mejor poder de discriminación. Los modelos se construyeron empleando el paquete estadístico SPSS y usando la función regresión logística. Las variables que mejor discriminaron, asumiendo un punto de corte de 0.5, que es el que por default arroja el paquete estadístico, se muestran a continuación junto con la matriz de confusión o de clasificación.

VARIABLES ESCOGIDAS PARA LA MEDICIÓN DE LA PROBABILIDAD DE PAGO DE LOS CLIENTES

Variables	Coeficiente	Significancia
Separado o Unión libre	-0,39	0,000%
Soltero, divorciado o viudo	-0,18	0,001%
Menos de dos cargas familiares	0,35	0,000%
Universitario	0,37	0,000%
Arrienda	-0,43	0,000%
Empleado	0,84	0,000%
Afiliado IESS	0,54	0,000%
Género femenino	0,50	0,000%
Guayaquil, Loja, Manta, Ibarra, Riobamba	-1,82	0,000%
Cuenca o Quito	-0,72	0,000%
< 25 años	-0,23	0,000%
< de 200 usd	-0,22	0,000%
> de 300 usd	0,12	0,600%
< de 3 años de trabajo	-0,56	0,000%
Carga financiera > 20%	-0,43	0,000%
Entre 1-3 años de residencia	-0,13	3,167%
Arrienda y trabaja menos de 1 año	-0,25	1,574%
Casa propia y más de 3 años de trabajo	0,20	0,645%
Soltero, divorciado o viudo y tiene 3 o más cargas familiares	-0,64	0,118%
Constante	0,33	0,260%

Fuente: Institución Financiera Ecuatoriana

Elaboración propia

MATRIZ DE CONFUSIÓN

	Pronosticado		
Observado	Buenos	Malos	% correcto
Buenos	5026	2491	66,86
Malos	2814	4703	62,56
Porcentaje global			64,71

El valor de corte es ,5

Elaboración propia

Los signos de las variables son los esperados y las pruebas de significancia de los coeficientes son validas todas al 95% de confianza por lo que no existe problema en el proceso de elección de las variables. Sin embargo, la observación de la matriz de confusión conmina a la pregunta ¿realmente discrimina este scoring?. Es importante en este punto mencionar que la institución analizada cuenta ya con un proceso de filtro en la aprobación de créditos por lo que las variables externas que pueden incrementar el poder predictivo del modelo se veían restringidas en su aplicación y utilización. Se puede notar que en el proceso de selección de variables sólo se ha incluido la tenencia o no de deuda en central de riesgos, excluyéndose las variables sobre el tipo de deuda que poseen dado que uno de los filtros institucionales es no admitir clientes con calificaciones B, C, D o E.[38]. Por lo tanto esperar que las variables sociodemográficas hablen sobre el comportamiento esperado del cliente por encima del 70% es muy optimista pero poco realista, se debe recordar que en la gráfica 'relación de la seguridad de la predicción y las fuentes de información' mostrada en el apartado 3.3 definición y selección de datos del capítulo 3 se indicaba que contar sólo con información interna sobre el cliente arrojaba seguridades de predicción del 20%. Por esto en un scoring de aprobación el poder discriminatorio siempre será menor que en uno de seguimiento dado la exclusión de variables de comportamiento de pago dentro de la institución o de variables del crédito.

Sin embargo, lejos de buscar salvar con retórica la modelización se presentan a continuación indicadores de rendimiento del modelo como medidas financieras del poder de clasificación del mismo y sus repercusiones sobre el manejo del negocio o la creación de estrategias por grupo de clientes.

5.- Validación del modelo y diseño de la scorecard

Para la presentación de las medidas de validación de un modelo scoring se acostumbra en la práctica a presentar una tabla resumen en donde para cada rango percentil se muestra el score mínimo y máximo obtenido, el número, distribución y

[38] A,B,C,D,E son las calificaciones que otorga el organismo regulador a los clientes según su edad de mora, siendo A cliente al día y E cartera vencida. Además la incursión pionera de la institución en el mercado del crédito de consumo originaba que para la fecha de muestra de los clientes un gran porcentaje no reportara deudas con ninguna otra institución del sistema.

distribución acumulada de los clientes buenos y malos para a partir de esta información construir los indicadores Kolmogorov-Smirnov y Gini.

Adicional a esto en la institución evaluada se presentó una tabla que incluía los rendimientos por rango de score, y la tasa de pérdida a fin de determinar el punto de corte y las estrategias de negocio que se sugieren aplicar.

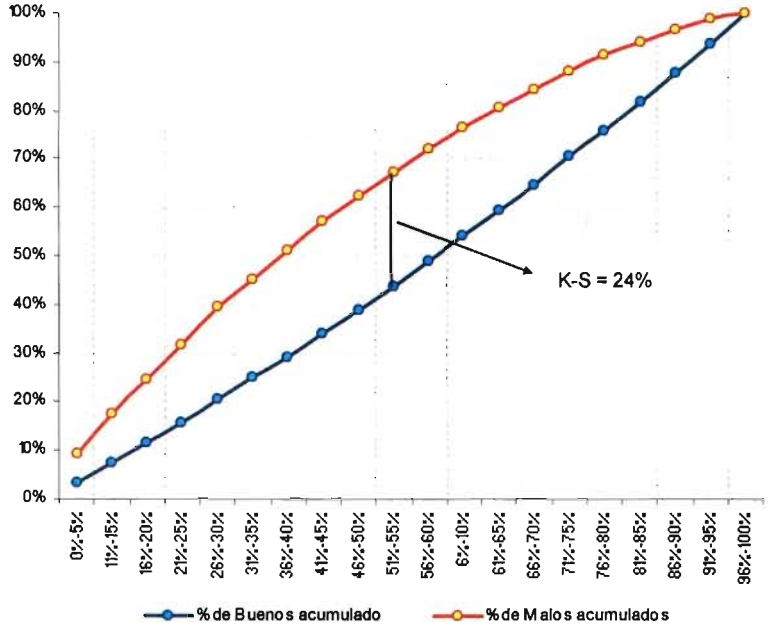
TABLA RESUMEN DE LA VALIDEZ DEL MODELO

Percentil	Score mínimo	Score máximo	# de Buenos	% Buenos	% de Buenos acumulado	# de Malos	% de Malos	% de Malos acumulados	Probabilidad de ser malo	Odds de ser Bueno	K-S	Gini	Total general
0%-5%	47	246	1086	3,38%	3,38%	1105	9,41%	9,41%	50,4%	1	6%	0%	2191
11%-15%	247	285	1294	4,03%	7,42%	941	8,01%	17,41%	42,1%	1	10%	1%	2235
16%-20%	285	328	1310	4,08%	11,51%	837	7,12%	24,54%	39,0%	2	13%	1%	2147
21%-25%	328	353	1369	4,27%	15,77%	861	7,33%	31,87%	38,6%	2	16%	2%	2230
26%-30%	353	379	1539	4,80%	20,57%	913	7,77%	39,64%	37,2%	2	19%	3%	2452
31%-35%	379	407	1370	4,27%	24,85%	663	5,64%	45,28%	32,6%	2	20%	3%	2033
36%-40%	407	436	1369	4,27%	29,11%	680	5,79%	51,07%	33,2%	2	22%	3%	2049
41%-45%	436	457	1549	4,83%	33,94%	689	5,86%	56,93%	30,8%	2	23%	4%	2238
46%-50%	457	488	1546	4,82%	38,77%	629	5,35%	62,29%	28,9%	2	24%	4%	2175
51%-55%	488	515	1598	4,98%	43,75%	568	4,82%	67,10%	26,2%	3	23%	4%	2164
56%-60%	515	542	1668	5,20%	48,95%	595	5,06%	72,17%	26,3%	3	23%	5%	2263
6%-10%	542	574	1618	5,05%	53,99%	511	4,35%	76,52%	24,0%	3	23%	4%	2129
61%-65%	574	595	1718	5,36%	59,35%	467	3,97%	80,49%	21,4%	4	21%	5%	2185
66%-70%	595	632	1714	5,34%	64,70%	469	3,99%	84,48%	21,5%	4	20%	5%	2183
71%-75%	632	658	1824	5,69%	70,38%	415	3,53%	88,02%	18,5%	4	18%	5%	2239
76%-80%	658	689	1761	5,49%	75,87%	381	3,24%	91,26%	17,8%	5	15%	5%	2142
81%-85%	689	724	1863	5,81%	81,68%	328	2,79%	94,05%	15,0%	6	12%	4%	2191
86%-90%	724	771	1887	5,88%	87,57%	310	2,64%	96,69%	14,1%	6	9%	4%	2197
91%-95%	771	821	1974	6,16%	93,72%	236	2,01%	98,70%	10,7%	8	5%	4%	2210
96%-100%	821	956	2013	6,28%	100,00%	153	1,30%	100,00%	7,1%	13	0%	3%	2166
Total general			32070	100,00%		11749	100,00%			3	24%	32%	43819

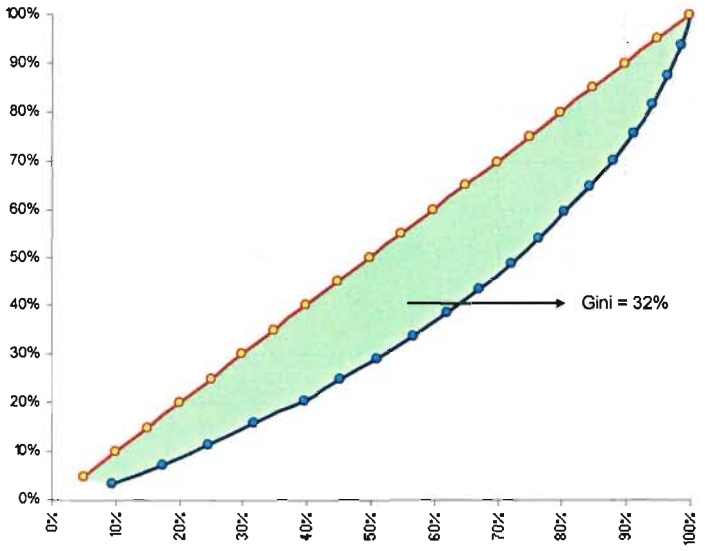
Fuente: Institución Financiera Ecuatoriana

Elaboración propia

CURVA KOLMOGOROV-SMIRNOV



CURVA ROC Y COEFICIENTE DE GINI



Fuente: Institución Financiera Ecuatoriana

Elaboración propia

La evidencia empírica indica que para un modelo de aprobación los niveles mínimos permitidos de K-S van desde el 25% y los de Gini van desde el 35%. Esta crítica que puede surgir se responde por dos vías: la primera es que los resultados arrojados por los modelos en cada institución dependerán siempre de la base de datos que se esté usando, es decir una institución que no aplica filtros en su proceso de selección de clientes y que puede emplear a discreción la información proporcionada por la central de riesgos es probable que alcance niveles por encima de los mínimos establecidos en este tipo de indicadores. Por ejemplo, en un trabajo similar desarrollado en una institución de tarjetas de crédito el autor encontró que el empleo de variables de central de riesgos aumentaba en 7 puntos el índice de Gini y en 5 el indicador de K-S por lo que se llegaba a valores de 39 y 29 respectivamente, esto confirma la hipótesis sostenida en este apartado. La segunda vía es el indicador de rendimientos.

RENDIMIENTOS, PÉRDIDA Y PERFIL DE CLIENTE POR RANGO PERCENTIL

Percentil	0-4	5-30	31-60	61-90	>90	Total general	Rendimiento	Pérdida	Perfil
0%-5%	59%	31%	6%	1%	3%	100%	21,18%	5,63%	Rechazo
6%-10%	64%	28%	4%	1%	3%	100%	22,90%	4,54%	Pérdida:
11%-15%	65%	28%	4%	1%	2%	100%	24,14%	3,76%	4,42%
16%-20%	66%	27%	3%	1%	2%	100%	24,16%	3,75%	
21%-25%	67%	26%	4%	1%	2%	100%	24,27%	3,68%	Cuatro
26%-30%	69%	26%	3%	0%	2%	100%	25,45%	2,93%	Pérdida:
31%-35%	71%	24%	3%	1%	2%	100%	25,23%	3,07%	3,08%
36%-40%	71%	24%	3%	0%	2%	100%	25,89%	2,64%	C/SD = 1,4
41%-45%	73%	23%	3%	0%	1%	100%	26,24%	2,42%	Tres
46%-50%	74%	22%	3%	0%	1%	100%	26,71%	2,12%	Pérdida:
51%-55%	73%	23%	2%	0%	1%	100%	26,79%	2,07%	2,10%
56%-60%	73%	24%	2%	0%	1%	100%	27,22%	1,79%	C/SD = 1,7
61%-65%	76%	21%	2%	0%	1%	100%	27,83%	1,40%	Dos
66%-70%	77%	20%	2%	0%	1%	100%	27,81%	1,41%	Pérdida:
71%-75%	77%	20%	2%	0%	1%	100%	28,16%	1,19%	1,30%
76%-80%	77%	20%	2%	0%	1%	100%	28,17%	1,18%	C/SD = 2,2
81%-85%	79%	18%	1%	0%	0%	100%	28,83%	0,76%	Uno
86%-90%	79%	18%	1%	0%	0%	100%	28,02%	0,64%	Pérdida:
91%-95%	84%	15%	1%	0%	0%	100%	28,56%	0,28%	0,38%
96%-100%	86%	14%	1%	0%	0%	100%	30,22%	-0,14%	C/SD = 2,5
Total general	74%	22%	2%	0%	1%	100%	26,82%	2,05%	

Fuente: Institución Financiera Ecuatoriana

Elaboración propia

La tabla muestra como a medida que mejora la calificación del cliente (rangos percentiles mayores) los rendimientos o su equivalente la pérdida van en aumento o decremento respectivamente esto muestra que el score clasifica bien pues a los clientes que se les otorga calificaciones bajas obtiene tasas de pérdida mayores en relación a los que obtienen calificaciones altas. Por ejemplo la zona roja tiene una pérdida promedio de 4.42% puntos en la tasa de rendimiento, mientras que el grupo verde oscuro muestra indicadores de pérdida de 0.38%. Emplear este tipo de medida en modelos donde los clientes han pasado por filtros en la fase de aprobación del crédito que limitan la utilización de la base de central de riesgos es fundamental para la toma de decisiones. Sin mencionar que la toma de decisiones financieras no puede pasar únicamente por la observación de indicadores estadísticos.

Con base en los indicadores de rendimiento se sugirió definir perfiles de clientes para que el control no sea únicamente el punto de corte sino el monto asignado, de esta forma no se está dejando de prestar por debajo de un límite establecido sino que para cada perfil se está prestando una mayor o menor cantidad en función de la pérdida que se espera reporte el grupo. En este caso la implementación de los rendimientos como

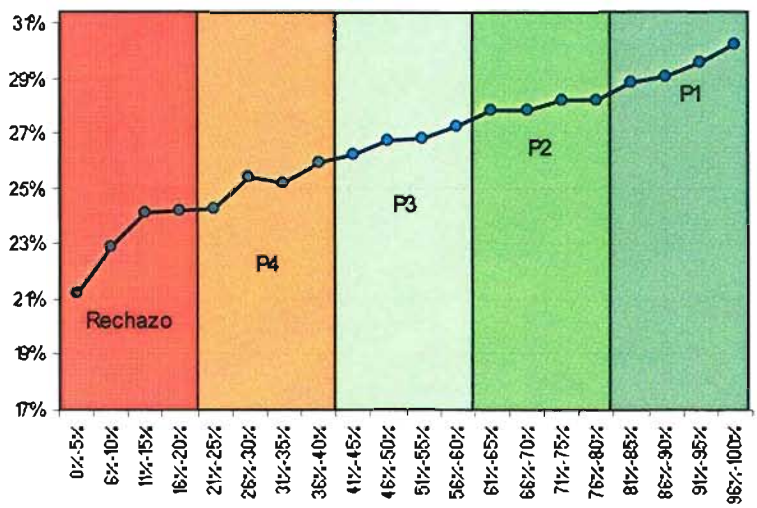
medida permitió a la institución definir estrategias comerciales de colocación en lugar de contraer la tasa de aceptación.

Los perfiles que se definieron fueron cuatro, siendo uno el mejor y cuatro el menos rentable que se desea admitir. El área roja se decidió rechazar:

1. Perfil uno: pérdida 0.38%
2. Perfil dos: pérdida 1.30%
3. Perfil tres: pérdida 2.10%
4. Perfil cuatro: pérdida 3.08%
5. Rechazo: pérdida 4.42%

Esta definición de perfiles es conservadora pues si se observa los clientes considerados para rechazo pierden sólo un 4.42% cifra que puede ser razonable perder en cualquier otra institución financiera, una alternativa de clasificación consiste en no establecer un punto de corte sino controlar el comportamiento de pago de los clientes vía disminuciones en el monto y plazo entregado como crédito. No obstante, estas decisiones pasan por el Directorio de la institución.

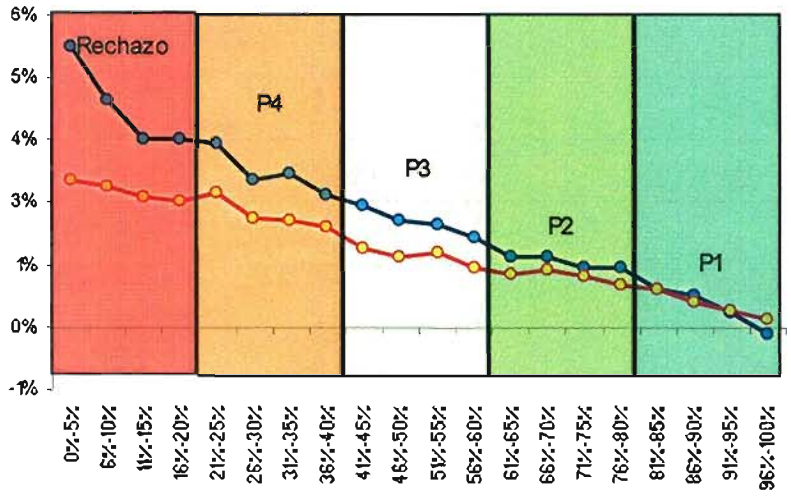
RENDIMIENTOS Y PERFIL DE CLIENTE POR RANGO PERCENTIL



Fuente: Institución Financiera Ecuatoriana

Elaboración propia

PERDIDA, CASTIGOS Y PERFIL DE CLIENTE POR RANGO PERCENTIL



Fuente: Institución Financiera Ecuatoriana

Elaboración propia

Las gráficas muestran la evolución de los rendimientos y las pérdidas por rango percentil del puntaje score se observa como los clientes con mejor calificación o los perfiles uno presentan menores pérdidas y mayores rendimientos, adicional se agregó el porcentaje de créditos castigados en cada perfil para complementar la información sobre el poder de clasificación del modelo. Note que la tasa de castigos tiende a decrecer en la medida en que nos acercamos a los perfiles más altos o rangos score más elevados lo que confirma que el modelo discrimina acertadamente a los clientes en la población.

Presentada las tablas de clasificación y rendimientos y aprobado el modelo por la institución se procedió a elaborar la tabla scorecard[39].

Scorecard			
Constante	0,33	Cuenca o Quito	-0,72
Separado o Unión libre	-0,39	< 25 años	-0,23
Soltero, divorciado o viudo	-0,18	< de 200 usd	-0,22
Menos de dos cargas familiares	0,35	> de 300 usd	0,12
Universitario	0,37	< de 3 años de trabajo	-0,56
Arrienda	-0,43	Carga financiera > 20%	-0,43
Empleado	0,84	Entre 1-3 años de residencia	-0,13
Afiliado IESS	0,54	Arrienda y trabaja menos de 1 año	-0,25
Género femenino	0,50	Casa propia y más de 3 años de trabajo	0,20
Guayaquil, Loja, Manta, Ibarra, Riobamba	-1,82	Soltero, divorciado o viudo y tiene 3 o más cargas familiares	-0,64

[39] Esta tabla se empleó como referencia para la comprensión de los pesos y signos de las variables pues en un modelo logístico los productos coeficientes variables no se suman para la obtención del puntaje final. Se puede revisar el anexo 3 para una mejor comprensión del tema.

La fórmula a implementar en el sistema transaccional es la siguiente:

P(bueno) = (exp(z)/(1+exp(z)))*1000

en donde

Z = 0.33 -0.39*(Separado o Unión libre) – 0.18*(Soltero, divorciado o viudo) + 0.35*(<=2 cargas familiares) + 0.37*(Universitario) – 0.43*(Arrienda) + 0.84*(Empleado) + 0.54*(Afiliado IESS) + 0.50*(Género femenino) – 1.82*(Ciudad UB = Guayaquil, Loja, Manta, Ibarra, Riobamba) – 0.72*(Cuenca o Quito) – 0.23*(<25 años) – 0.22*(sueldo líquido < de 200) + 0.12*(sueldo líquido > de 300) – 0.56*(< de 3 años de trabajo) – 0.43*(carga financiera > 20%) – 0.13*(tiempo de residencia entre 1-3 años) -0.25*(Casa arrendada y menos de 1 año de trabajo) + 0.20*(Casa propia y más de 3 años de trabajo) – 0.64*(soltero, divorciado o viudo y tres o más cargas familiares)

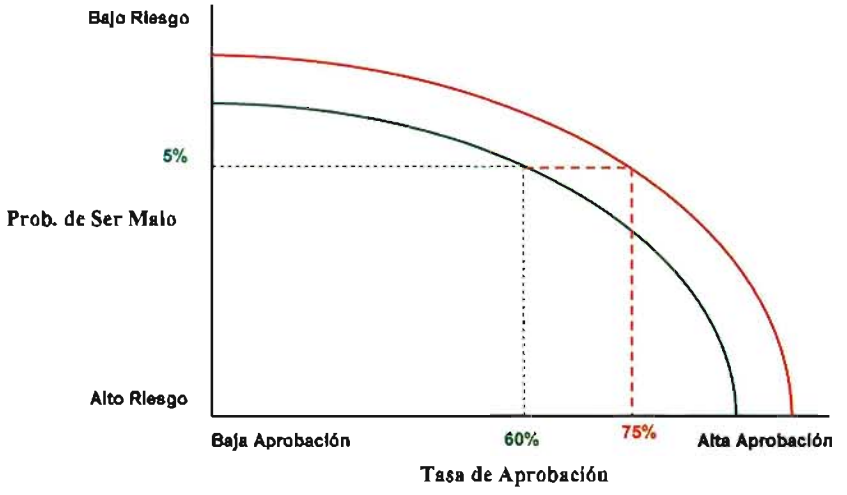
Esta fórmula se espera sea implementada en el sistema transaccional de la institución.

6.- Es posible prestar como locos y obtener beneficios

La pregunta que sigue por defecto al desarrollo del modelo es ¿realmente incrementará el beneficio una discriminación basada en un modelo scoring?. En términos teóricos la contestación es sí, independiente de la estrategia riesgo/aprobación que se escoja.

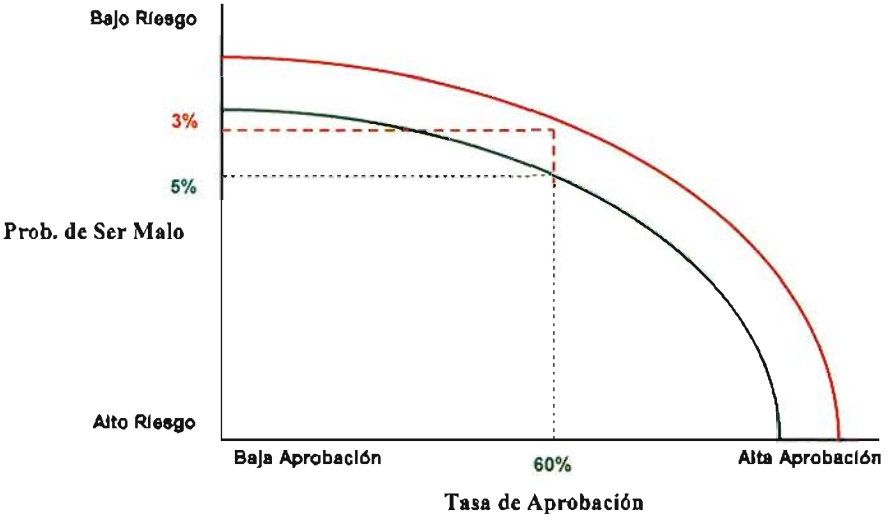
Una vez definido el modelo la directiva de la institución tendrá al menos dos alternativas de elección respecto a los resultados del scoring. La primera es mantener la misma tasa de aprobación pero disminuir el riesgo; la segunda es incrementar la tasa de aprobación pero manteniendo el riesgo. En cualquier instancia el objetivo final es conducir a la institución a obtener una mayor rentabilidad en sus colocaciones.

RELACIÓN MENOR RIESGO IGUAL TASA DE APROBACIÓN



Fuente: Equifax

RELACIÓN IGUAL RIESGO MAYOR TASA DE APROBACIÓN



Fuente: Equifax

Una forma de evaluar la consecución de una mayor rentabilidad producto de la discriminación es simular una estrategia de ventas, basada en la segmentación entregada por el modelo.

MODIFICACIÓN DE CUPOS Y DE RENTABILIDAD

EMPLEANDO SEGMENTACIÓN SCORING

Percentil	Rendimiento	Pérdida	Perfil	Rendimiento Actual	Sueldo Depurado	Cupo sugerido	Rendimiento Simulado
0%-5%	21,18%	5,63%	Rechazo	0,85%	487.614,7	-	
6%-10%	22,90%	4,54%	Pérdida:	0,96%	495.666,5	-	
11%-15%	24,14%	3,76%	4,42%	1,05%	519.481,5	-	
16%-20%	24,16%	3,75%		1,06%	519.543,6	-	
21%-25%	24,27%	3,68%	Cuatro	1,19%	588.472,8	823.862	1,10%
26%-30%	25,45%	2,93%	Pérdida:	1,15%	561.302,5	785.824	1,10%
31%-35%	25,23%	3,07%	3,08%	1,09%	478.329,0	669.661	0,93%
36%-40%	25,89%	2,64%	C/SD = 1,4	1,26%	710.589,1	994.825	1,42%
41%-45%	26,24%	2,42%	Tres	1,28%	536.253,1	911.630	1,32%
46%-50%	26,71%	2,12%	Pérdida:	1,31%	554.079,7	941.935	1,38%
51%-55%	26,79%	2,07%	2,10%	1,39%	581.378,4	988.343	1,46%
56%-60%	27,22%	1,79%	C/SD = 1,7	1,38%	581.955,3	989.324	1,48%
61%-65%	27,83%	1,40%	Dos	1,50%	587.689,4	1.292.917	1,98%
66%-70%	27,81%	1,41%	Pérdida:	1,50%	580.272,0	1.276.598	1,95%
71%-75%	28,16%	1,19%	1,30%	1,55%	583.983,6	1.284.764	1,99%
76%-80%	28,17%	1,18%	C/SD = 2,2	1,51%	565.230,6	1.243.507	1,93%
81%-85%	28,83%	0,76%	Uno	1,60%	594.340,7	1.485.852	2,36%
86%-90%	29,02%	0,64%	Pérdida:	1,64%	579.026,0	1.447.565	2,31%
91%-95%	29,56%	0,28%	0,38%	1,72%	619.751,5	1.549.379	2,52%
96%-100%	30,22%	-0,14%	C/SD = 2,5	1,74%	592.540,1	1.481.350	2,48%
Total general	**26,82%**	**2,05%**		**26,73%**	**11.317.500,2**	**18.167.336**	**27,70%**

Fuente: Institución financiera ecuatoriana

Elaboración propia

La tabla adjunta muestra la pérdida asignada a cada perfil definido por el scoring, su rendimiento actual y en la última columna su rendimiento simulado. La lógica de simulación es muy sencilla. Se puede observar la relación monto sobre sueldo neto o depurado de la población, este ratio es el existente antes de la implementación del modelo scoring (para esta evaluación la relación monto/sueldo neto = C/SD era igual a 2, es decir se entregaba un monto de crédito igual a dos vez su sueldo neto), osea sin segmentación y es la que está rindiendo en promedio un 26.73%. Con esta información como referencia se procede a controlar el monto entregado en función del rendimiento (pérdida) esperado. Se observa en la columna perfil como la relación C/SD es decreciente en la medida en que nos acercamos a los quintiles más bajos de la población, con esta modificación se espera, dado que la rentabilidad también es decreciente cuando se acerca a los quintiles menores, que la cartera segmentada rinda más que la inicial colocada sin discriminación. Los resultados aparecen en la última fila de la columna rendimiento simulado, en donde se observa el incremento en un punto porcentual de rentabilidad (27.70%). Note que la redefinición de la relación C/SD es hacia abajo, es decir se simula con una tendencia mayor a restringir la relación en los puntajes bajos, antes que incrementarla en los altos, por esto el incremento es pequeño;

sin embargo, un crecimiento de un punto porcentual multiplicado por el volumen colocado en una institución de consumo resulta en ganancias significativas para la institución.

Este sencillo ejercicio permite inferir que el empleo de un modelo estadístico que permita segmentar la cartera de clientes no sólo conduce a un mejor control del riesgo de crédito sino que, acompañado de una estrategia comercial lleva a incrementar las tasas de rentabilidad de los créditos colocados, permitiendo en caso de una estrategia de expansión de mercado prestar a más clientes y obtener en promedio mayores retornos de estas colocaciones, que en el caso de basarse en un scoring implícito o subjetivo.

V.- Conclusiones y Recomendaciones

- La banca de consumo por su característica de trabajo, bajos montos de dinero en un volumen alto de operaciones diarias, necesita de mecanismos de aprobación de créditos ágiles y efectivos, que permitan colocar más dinero a un riesgo controlado. Esta particularidad de la banca de consumo hace que el empleo de metodologías estadísticas para todas las fases del crédito sea de gran beneficio pues maneja análisis de variables del tipo multivariado y arroja un número que califica la bondad del cliente superando el criterio subjetivo del analista de crédito y permitiendo definir estrategias de negocios diferenciadas en función del tipo de cliente.

- Uno de los problemas principales que afronta el desarrollo de un modelo scoring es la calidad de la información contenida en la base de datos, en este estudio se contó con una sólida base de datos que permitió no sólo contar con información fiable sino de calidad permitiendo trabajar con una gran cantidad de variables simples y cruzadas sin la pérdida significativa de información.

- El empleo de técnicas estadísticas como los árboles de decisión permiten reducir el tiempo de trabajo en el proceso de selección y definición de variables al mismo tiempo que reducen el criterio subjetivo del investigador en cuanto al cruce de variables que debe realizar y a las agrupaciones o clasificación gruesa de atributos.

- Para determinar la efectividad del modelo estadístico escogido, la institución no sólo debe basarse en los indicadores estadísticos, más aun en un modelo de aprobación en donde por la etapa en la que se encuentra el cliente dentro de la fase del crédito las variables de mayor relevancia son las sociodemográficas, sino en indicadores de rendimiento por percentil score, y con esta lógica se recomienda establecer el punto de corte y la segmentación por grupo de clientes en el caso que se quiera realizar estrategias de colocaciones diferenciadas.

- Para el caso específico de la institución se encontró que las variables que mejor discriminan la población de buenos y malos clientes son: el estado civil, las cargas

familiares, el nivel de educación, el tipo de vivienda, la afiliación a la seguridad social, el género, la edad, la ciudad en laque se entrega el crédito, el tiempo de trabajo, el compromiso o la carga financiera del cliente (definida como el 10% de la deuda que mantiene en central de riesgos sobre el sueldo líquido), y los cruces de variables como tipo de vivienda y tiempo de trabajo, y estado civil y cargas familiares.

- Los indicadores de validación del modelo como el K-S y el Gini se situaron en niveles del 24 y 32 respectivamente, si bien estos índices estadísticos están por debajo de los niveles empíricamente aceptados se recuerda que la base de clientes trabajada no puede incorporar la totalidad de información de central de riesgos, sea por que los clientes pasan por filtros de aprobación automáticos con base en la central de riesgos o por que los clientes que se acercan a la institución para este periodo muestral no habían sido sujetos de crédito en otras instituciones del sistema pues la institución es una de las pioneras en la entrega masiva de créditos de consumo.

- El departamento de riesgos de la institución desarrolló una medida de rendimiento que permite evaluar la rentabilidad o pérdida por rango percentil del puntaje score, con base en esta medida se observó el poder de clasificación del modelo, evidenciándose que clientes con altos puntajes presentan mayores niveles de rentabilidad que aquellos con puntajes bajos.

- Con base en la rentabilidad se establecieron diferentes segmentos de clientes los cuales serán controlados por el monto colocado. Los perfiles de clientes establecidos son cuatro:
    - Perfil uno: pérdida 0.38%
    - Perfil dos: pérdida 1.30%
    - Perfil tres: pérdida 2.10%
    - Perfil cuatro: pérdida 3.08%
    - Rechazo: pérdida 4.42%

- El grupo establecido como rechazo puede ser considerado como parte de la población sujeto de crédito, eliminando el punto de corte en el modelo scoring, no obstante, esta sugerencia queda en manos del Directorio de la institución.

- El empleo de un modelo estadístico que permita segmentar la cartera de clientes no sólo conduce a un mejor control del riesgo de crédito sino que, acompañado de una estrategia comercial lleva a incrementar las tasas de rentabilidad de los créditos colocados, permitiendo en caso de una estrategia de expansión de mercado prestar a más clientes y obtener en promedio mayores retornos de estas colocaciones, que en el caso de basarse en un scoring implícito o subjetivo.

VI.- Bibliografía

AGRESTI, A. (1996) "An Introduction to Categorical Data Analysis", Editorial John Wiley & Sons, INC, pp. 103-129.

CALVACHE, D., CARRANZA, F. (2000) "Diseño y elaboración estadística de un sistema de evaluación para la otorgación de crédito de consumo en una institución financiera". Proyecto previo a la obtención del título de ingeniero matemático mención en estadística, finanzas y gestión empresarial, Quito, Ecuador.

DATA MINING INSTITUTE, "Teoría y Práctica de los Arboles de Decisión"

DOBSON, A. (1993) "An Introduction to Generalized Linear Models", Editorial Chapman & Hall.

MESTER, Loretta J. (1997) "What's the point of Credit Scoring?", Business Review, Set./Oct., pp.3-16, Federal Reserve Bank of Philadelphia.

MONTGOMERY, D., PECK, E., VINING, G. "Introducción al análisis de regresión lineal", Compañía Editorial Continental. Primera edición en español, pp. 309-408

MYERS, R., Montgomery, D., VINING, G. (2002) "Generalized Linear Models With Applications in Engineering and the Sciences", Editorial John Wiley & Sons, INC.

PYNDICK, R., RUBINFELD, D. "Econometría modelos y pronósticos". Editorial Mc Graw Hill. Cuarta Edición

SCHREINER, Mark. (1999) "Un Modelo de Calificación del Riesgo de Morosidad para los Créditos de una Organización de Microfinanzas en Bolivia", www.microfinance.com

SCHREINER, Mark. (2000) "Credit Scoring for Microfinance: Can it work?", Journal of Microfinance, Vol. 2 pp. 105-118.

SCHREINER, Mark. (2001) "Do-It-Yourself Scoring Trees for Microfinance", Tercer Seminario sobre la Banca y Microfinanzas en Latinoamércia y el Caribe, Santo Domingo, Nov 11-12.

SCHREINER, Mark. (2004) "Benefits and Pitfalls of Statistical Credit Scoring for Microfinance", Savings and Development, Vol. 28, No 1, pp. 63-86.

SCHREINER, Mark. (2004) "Scoring Arrears at a Microlender in Bolivia", Journal of Microfinance, Vol. 6, No. 2, pp. 65-68.

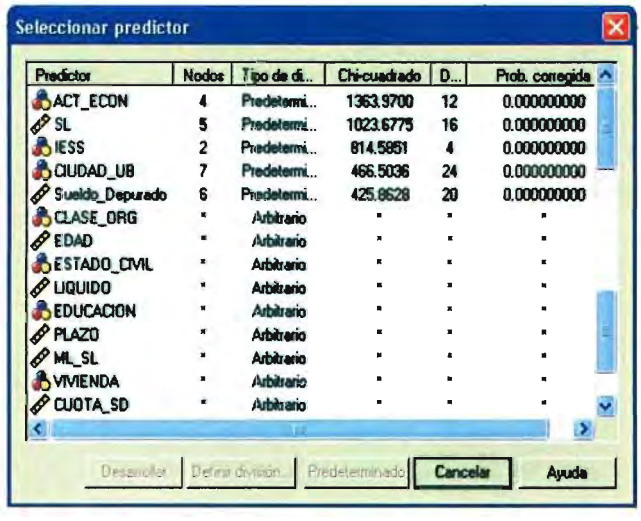
THOMAS, L., Edelman, D., Crook, J. (2002) "Credit Scoring and its applications", Society for Industrial and Applied Mathematics.

Anexo 1.- De estadística, paradojas y resoluciones financieras (comentario al numeral 2 del artículo 1 de la resolución No JB-2004-722)

La determinación de la capacidad de pago de los clientes de una institución financiera no obedece a un criterio univariante. Si bien, inicialmente nos puede interesar estudiar de forma individualizada tanto la relación de dependencia entre una variable, en este caso la buena o mala capacidad de pago del acreditado $\mathbf{Y_{(nx1)}}$, y cada factor de riesgo univariante $\mathbf{X_{(nx1)}}$, como las relaciones entre predictores $\mathbf{X_{i(nx1)}}$ y $\mathbf{X_{j(nx1)}}$. El objetivo final es la obtención de un conjunto "equilibrado" de variables predictoras.

Si se seleccionan separadamente las variables que una a una están más asociadas con el riesgo, es posible que en el conjunto de variables seleccionadas resultante se disponga de información redundante o bien que no tengamos incorporadas variables que de manera conjunta con otras resulten significativas. Se hace necesario entonces realizar el estudio conjunto tomando en cuenta a la vez todos los factores potenciales del riesgo e idealmente todas sus interacciones. Para ello es indispensable hacer uso de técnicas estadísticas multivariantes.

Estas herramientas permiten definir el conjunto de variables predictoras a incorporar para la especificación del modelo de capacidad de pago del cliente. Como muestra de estas interacciones se presenta un grupo de variables, junto con sus coeficientes Chi-cuadrado, empleadas por una institución finanaciera para determinar perfiles crediticios.



Fuente: Base de datos clientes institución financiera.

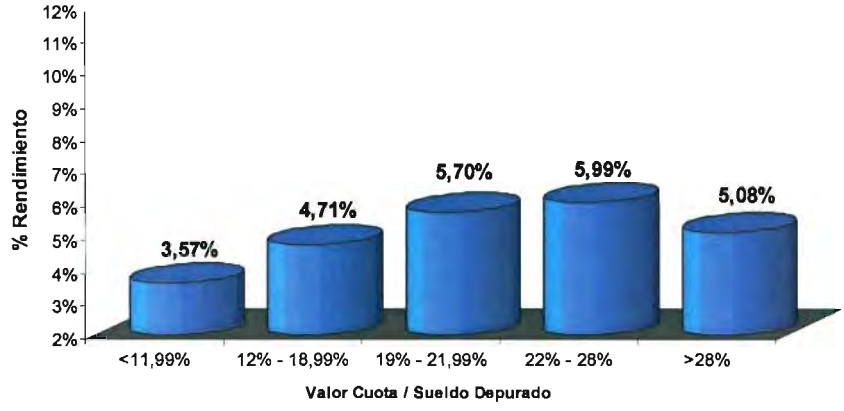
Paquete estadístico SPSS, función Arboles de Decisión algoritmo CHAID.

El valor Chi-cuadrado muestra el grado de dependencia de la variable predictora **X** en relación a la distribución de la variable dependiente **Y.** De este ejemplo se desprenden dos observaciones de importancia: a) no existe una única variable que permita discriminar de forma efectiva entre un cliente con una buena o mala capacidad de pago; b) la relación cuota pagada sobre ingreso neto (CUOTA_SD) muestra para el banco un poder predictivo bajo en relación a otros atributos del cliente.

Es decir, si se tuviese que elegir una única variable para definir la capacidad de pago del cliente, los criterios estadísticos no seleccionarían a esta como relevante. Relegándola, en el momento inicial de elección, al puesto 14 de importancia en cuanto a su poder discriminatorio.

Si el banco basará sus decisiones de límites crediticios únicamente en la proporción del ingreso neto del cliente que debe destinar al pago de una cuota, cometería errores en la aceptación de niveles de riesgos y por ende en el cálculo de sus niveles de pérdidas.

ACTIVIDAD ECONÓMICA: EMPLEADO



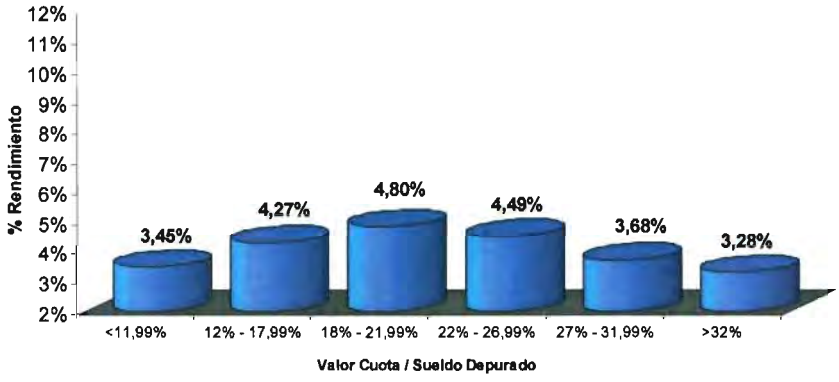
Fuente: Institución Financiera Ecuatoriana

Elaboración propia

La pérdida por rango de relación 'cuota/ingreso neto' que el banco esperaría asumir se describe en el gráfico superior y se estimaría alrededor del 5%. No obstante, la inclusión inicial de una variable de mayor importancia como la actividad económica del acreditado y luego esta particionada por la relación 'cuota/ingreso neto' lleva a niveles disímiles de riesgo.

En el caso de los empleados el riesgo promedio asumido por el banco se reduce en un punto porcentual (4%), mientras que para los comerciantes se incrementa en 4 puntos porcentuales (9%)

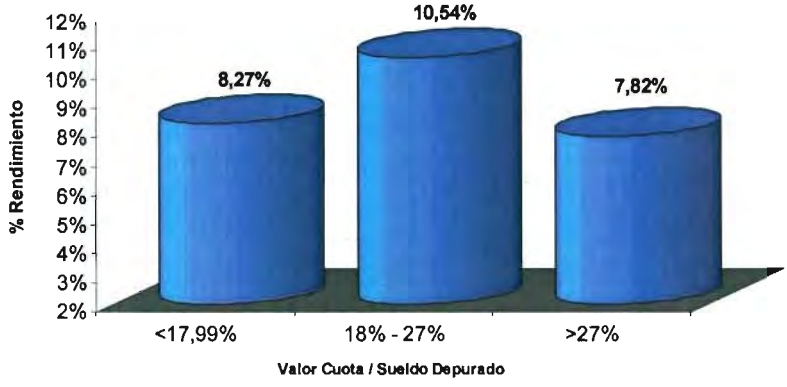
ACTIVIDAD ECONÓMICA: EMPLEADO



Fuente: Institución Financiera Ecuatoriana

Elaboración propia

ACTIVIDAD ECONÓMICA: COMERCIANTE



Fuente: Institución Financiera Ecuatoriana

Elaboración propia

Conclusiones que no son triviales pues han conducido al establecimiento de políticas diferentes con base en análisis estadísticos que son función de los atributos de mayor relevancia en la determinación de la calidad de pago.

Por otro lado, la fijación a priori de un límite de cuota del 50% del ingreso neto mensual está asumiendo una relación monótona creciente entre el valor cuota y el riesgo

de no pago. Relación que no es estrictamente cierta. La distribución mostrada en las gráficas señala un cambio en el nivel de pérdida asumida por el banco para un nivel de la relación cuota/ingreso neto mayor al 27%. Es decir existe un punto de inflexión a partir del cual una mayor relación cuota/ingreso neto genera pérdidas mayores, lo cual resta importancia a la relación implícita planteada por el organismo regulador.

Con esto no se quiere decir que exista una relación parabólica, más bien se resalta la existencia de una relación más compleja que la lineal.

El problema de basar una resolución en un criterio univariante puede agravarse más en caso de la presencia de la **Paradoja de Simpson**. Esta sostiene que la asociación entre dos variables (numéricas o cualitativas) puede cambiar de sentido cuando se controla el efecto de una tercera variable. Para abordar este problema se plantea un sencillo ejemplo teórico.

Se supone la elaboración de un estudio comparativo del efecto de una mayor relación cuota pagada/ingreso neto sobre la capacidad de pago del cliente. Con **A** como la relación cuota pagada/ingreso neto mayor y **B** la menor.

		Cuota/ingreso neto		
		A	**B**	
Capacidad de pago	**Buena**	410	434	844
	Mala	115	91	206
		525	525	1050

Elaboración propia

Del primer resultado obtenido el 21.9% de clientes de A (115/525) muestra una mala capacidad de pago ante una relación cuota pagada/ingreso neto mayor, frente a un 17.3% del grupo B.

Descomponiendo el estudio por rango de edades se encuentra que ahora la proporción de efectos adversos (mala capacidad de pago) en los clientes con menos de 25 años, es 6.2% en el grupo A, frente a 13.3% en el grupo B, diferencia que tiene signo contrario a la observada en el caso agregado.

		Cuota/ingreso neto		
		A	B	
Capacidad de pago	Buena	122	351	473
	Mala	8	54	62
		130	405	535

Elaboración propia

En el otro grupo de clientes de mayor edad se tiene un 27.1% con mala capacidad de paga y una mayor relación (A), frente a 30.8% en el grupo B.

Sólo la inclusión de una tercera variable (edad) la relación entre capacidad de pago y cuota/ingreso neto ha cambiado de signo. En el agregado era mayor la tasa de mala capacidad de pago en el grupo de mayor cuota/ingreso neto; luego de la estratificación por edad, para ambos casos, es menor en el grupo A que en el B.

Lejos de criticar la labor reguladora de la Superintendencia de Bancos y Seguros del Ecuador y convencido que un sistema financiero sano asegura el crecimiento del tejido productivo. Se cree que la fijación de normas cuantitativas debe respaldarse en análisis estadísticos y no en criterios ad hoc que pueden perjudicar el objetivo último que es la construcción de una sólida arquitectura financiera.

Anexo 2.- Funcionamiento de los árboles de decisión usando el algoritmo CHAID

El algoritmo CHAID, es la herramienta de clasificación utilizada en este estudio para la clasificación fina y gruesa de las variables, así como para la determinación de los cruces (sólo dos combinaciones) de las variables que ingresaran al modelado del comportamiento futuro de pago de los clientes.

CHAID se diseñó para identificar las interacciones a incluir en modelos de regresión. Maneja con facilidad las interacciones que tantas dificultades plantean a otras técnicas de modelización. Las interacciones son combinaciones de variables independientes que influyen en el resultado.

Por ejemplo, podría suceder que la rentabilidad fuera la misma para un cliente que realiza escasas transacciones y se encuentra fuertemente endeudado y otro que, por el contrario, realiza numerosas transacciones y, sin embargo, tiene un mínimo saldo en contra; en este caso, estas dos variables independientes (transacciones y saldo en contra) deberían considerarse de forma aislada.

CHAID se la considera como una técnica de explotación de datos (data mining). La explotación de datos consiste en analizar y estudiar grandes masas de datos con el objetivo de descubrir patrones y pautas no triviales desde el punto de vista del aprovechamiento comercial. De ahí su importancia al momento de analizar un conjunto de variables y sus interrelaciones entre sí y la variable objetivo.

Algoritmo Básico Del Árbol de Decisión: CHAID y CHAID exhaustivo

- Aprendizaje.

El aprendizaje implica cambios en el sistema que se adapta para permitir llevar a cabo la misma tarea a partir de las mismas condiciones de un modo más eficiente y eficaz cada vez. En un sistema de Reconocimiento de Formas, y dependiendo del método de aprendizaje se trata de calcular el patrón prototipo o el conjunto de patrones prototipo que caracterizan cada una de las clases a discriminar.

Usualmente se utiliza un *modelo de aprendizaje inductivo* que se puede formular como sigue:

> Una vez establecida la manera de representar el conocimiento y extraído éste, se calcula a partir de un conjunto de entrenamiento el patrón (o conjunto de patrones) prototipo utilizando un algoritmo de aprendizaje. Es necesario un esquema de evaluación que proporciona una medida de bondad del sistema.

- Clasificación.

Consiste en proporcionar nuevos prototipos al sistema, independientes de los utilizados en el aprendizaje para que éste los etiquete utilizando el conjunto de clases disponibles.

- Evaluación.

Toda clasificación lleva aparejada una medida de error, bondad o confianza. Deben proporcionarse mecanismos para evaluar esta bondad. Normalmente se utiliza un conjunto de patrones etiquetados por expertos y no usados en el aprendizaje.

Ejemplo:

En el siguiente ejemplo se muestra la construcción de un árbol de decisión utilizando la herramienta estadística CHAID (Detector Automático de Interacciones Chi cuadrado) buscando la respuesta a una variable categórica Y y 2 variables explicativas, en este caso X1 y X2. La variable Y tiene 4 categorías Y = 1, 2, 3, 4; la variable X1 tiene 4 categorías, X1 = 1, 2, 3, 4; la variable X2 tiene 3 categorías, X2 = 0, 1, 2. Entonces los pasos de construcción son los siguientes.

1. Calcular la distribución de la variable respuesta Y en el nodo raíz.

Categoría	%	n
1	35.00	35
2	8.00	8
3	35.00	35
4	22.00	22
TOTAL	(100,00)	100

2. Para cada variable explicativa X, se encuentra el par de categorías de X que sean por lo menos significativamente diferentes (mayor P-value) con respecto a la distribución de Y dentro de este nodo.

En este ejemplo La variable Y es categórica y se desarrollan una serie de estadísticos CHI CUADRADO.

i. La relación entre X1 y la variable respuesta Y dentro del nodo está dada por la siguiente tabla cruzada

X1/Y	1	2	3	4	Total
1	23	5	19	4	51
2	12	2	15	13	42
3	0	1	0	1	2
4	0	0	1	4	5
Total	35	8	35	22	100

Chi2= 25.63559 g.l.= 9 (p=0.002342955)

g.l. = Grados de Libertad

Como X1 tiene 4 categorías, hay 6 sub tablas cruzadas de 2 x 4 a considerar

X1/Y	1	2	3	4	Total
1	23	5	19	4	51
2	12	12	15	13	42
Total	35	7	34	17	93

Chi2= 9.193281 g.l. = 3 (p=0.02682849)

X1/Y	1	2	3	4	Total
1	23	5	19	4	51
3	0	1	0	1	2
Total	23	6	19	5	53

Chi2= 8.019281 g.l. = 3 (p=0.0456149)

X1/Y	1	2	3	4	Total
1	23	5	19	4	51
4	0	0	1	4	5
Total	23	5	20	8	56

Chi2= 19.72078 g.l. = 3 (p=0.0001939266)

X1/Y	1	2	3	4	Total
2	12	2	15	13	42
3	0	1	0	1	2
Total	12	3	15	14	44

Chi2= 7.23356 g.l. = 3 (p=0.0648145)

X1/Y	1	2	3	4	Total
2	12	2	15	13	42
4	0	0	1	4	5
Total	12	2	16	17	47

Chi2= 4.962482 g.l. = 3 (p=0.1745651)

X1/Y	2	3	4	Total
3	1	0	1	2
4	0	1	4	5
Total	1	1	5	7

Chi2= 3.08 g.l. = 3 (p=0.2143811)

ii. El algoritmo identifica el par de categorías de X1 con el (p-value) y compara ese valor con un valor preespecificado alfa α nivel de significancia (= 0.05; valor de default).

En este ejemplo, es el par definido por las categorías 3 y 4 de X1 ya que (p-value) para esta pareja de categorías (0.2143) es más grande que α, entonces las dos categorías se juntan para formar una sola categoría compuesta dado que estas dos categorías no nos sirven para diferenciar a las clases por lo que se puede agruparlas.

Como resultado, un nuevo juego de categorías de X1 es formado, y reiteramos el análisis de las tablas cruzadas. Ahora se tiene tres tablas cruzadas por analizar.

X1/Y	1	2	3	4	Total
1	23	5	19	4	51
3,4	0	1	1	5	7
Total	23	6	20	9	58

Chi2= 20.25577 g.l. = 3 (p=0.0001502343)

X1/Y	1	2	3	4	Total
2	12	2	15	13	42
3,4	0	1	1	5	7
Total	12	3	16	18	49

Chi2= 6.408565 g.l. = 3 (p=0.09333908)

X1/Y	1	2	3	4	Total
1	23	5	9	4	51
2	12	2	15	13	42
Total	35	7	34	17	93

Chi2= 9.193281 g.l. = 3 (p=0.02682849)

iii. El par de atributos que presenten un (p-value) mayor que alfa $\alpha = 0.05$ son unidos en un solo atributo. En este caso particular; se unirá 2 con la categoría compuesta 3,4. Si existieran más categorías para analizar su representatividad continuaríamos formulando tablas cruzadas, pero ahora solo tenemos dos categorías que representan a la variable X1, el proceso de unir categorías debe parar.

iv. Ahora, el algoritmo calculará un valor de P ajustado para el grupo de categorías de X1 que han sido combinadas y las categorías de Y usando el ajuste de Bonferroni[40]

X1/Y	1	2	3	4	Total
1	23	5	19	4	51
2,3,4	12	3	16	18	49
Total	35	8	35	22	100

Chi2= 13.08861 g.l. = 3 (p=0.004448843)

El valor del estadístico Chi cuadrado del valor p es ajustado usando el multiplicador de Bonferroni. Como X1 es nominal. El ajuste de Bonferroni se calcula como se muestra a continuación:

---

[40] AJUSTE DE BONFERRONI: Técnica estadística que ajusta el nivel de significación en relación al número de pruebas estadísticas realizadas simultáneamente sobre un conjunto de datos. El nivel de significación para cada prueba se calcula dividiendo el error global de tipo I entre el número de pruebas a realizar. El ajuste de Bonferroni se considera conservador.

$$B_{free} = \sum_{i=0}^{r-1} (-1)^i \frac{(r-i)^c}{r!(r-i)!}$$

Donde:

c = Número de las categorías originales de X1 (4)

r = Número de categorías compuestas (2). Se tiene un ajuste en el valor p de 0.0311 (=0.00448843 x 7)

3. Se reiteran los pasos desde (i) hasta (iv) pero reemplazado X1 por X2

i. La tabla cruzada de X2 para Y en el nodo raíz es

X2/Y	1	2	3	4	Total
0	33	8	34	22	97
1	1	0	1	0	2
2	1	0	0	0	1
Total	35	8	35	22	100

Chi2= 2.768778 g.l. = 6 (p=0.8372577)

ii. X2 tiene tres categorías, por lo tanto tendremos solamente tres tablas cruzadas de 2 x 4

X2/Y	1	2	3	4	Total
0	33	8	34	22	97
1	1	0	1	0	2
Total	34	8	35	22	99

Chi2= 2.8881097 g.l. = 3 (p=0.8282962)

X2/Y	1	2	3	4	Total
0	33	8	34	22	97
2	1	0	0	0	1
Total	34	8	34	22	98

Chi2= 1.901759 g.l. = 3 (p=0.5930452)

X2/Y	1	3	Total
0	1	0	1
1	1	1	2
Total	2	1	3

Chi2= 0.75 g.l. = 1 (p=0.3864762)

iii. En este caso, la pareja de categorías con mayor (p-value) son 0 y 1 con un valor de 0.8283. Por lo tanto la combinación de las categorías 0 y 1 forman una sola categoría compuesta. El proceso de combinación de categorías finaliza ya que existen solamente dos categorías de salida.

iv. El valor final p para esta tabla cruzada de X2 y Y es:

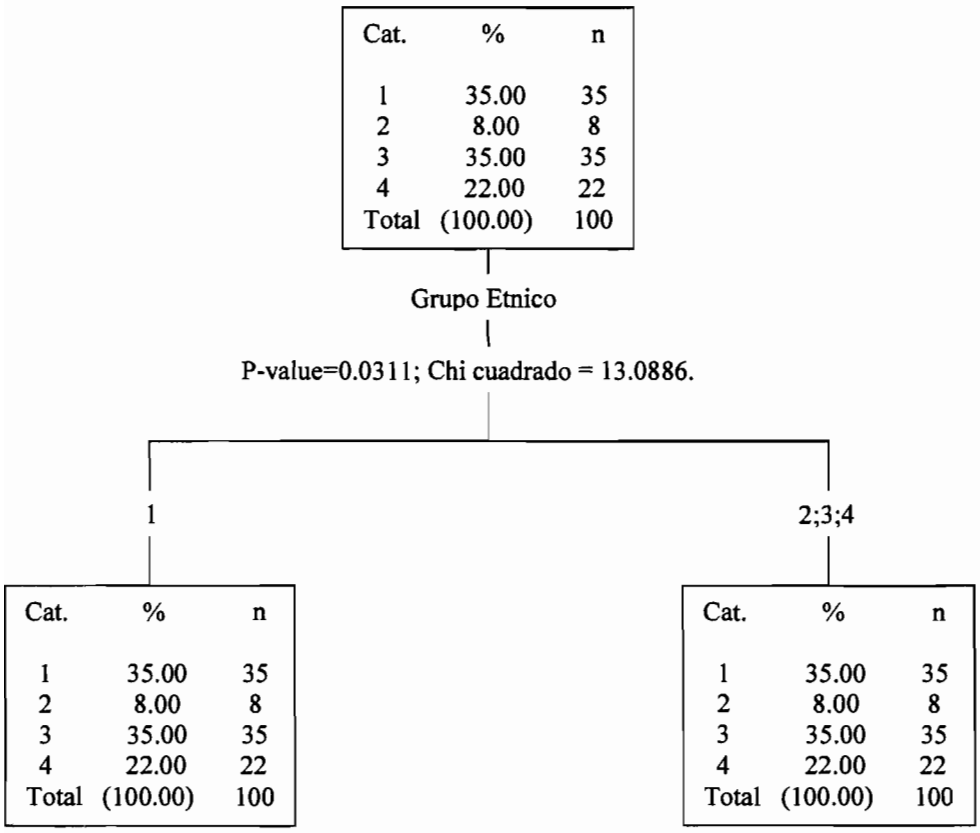
X2/Y	1	2	3	4	Total
0,1	34	8	35	22	99
2	1	0	0	0	1
Total	35	8	35	22	100

Chi2= 1.875902 g.l. = 3 (p=0.5985587)

El valor de p ajustado en este caso es 1 (0.5985587 x 3 es mayor que 1)

3. El paso final consiste en dividir el nodo padre en base a la función de categoría de X1. es decir un sub nodo con 35 clases (X1 = 1) y el otro con las restantes 65 clases (X2 = 1, 2, 3, 4).
4. El crecimiento del árbol continúa hasta que se satisface el criterio de parada.

GRÁFICO DEL ARBOL DE DECISIÓN CONSTRUIDO CON CRITERIO CHAID



Anexo 3.- Los modelos lineales generalizados como método estadístico para la construcción de Credit Scorecards.

Todo modelo de calificación estadística es una formulación matemática que busca otorgar ponderaciones a las diferentes características de un prestatario, un prestamista y del préstamo en sí. El modelo desarrollado entrega una estimación de la probabilidad de que ocurra el evento esperado, generalmente denominado como éxito y que para el caso del riesgo en la aprobación del crédito será la probabilidad de que sea calificado como buen cliente.

Bajo esta lógica el ser buen o mal cliente se puede modelar como una variable binaria[41], es decir que el modelo de elección asume que los individuos se enfrentan con una elección entre dos alternativas, las cuales dependen de características identificables.

Por tanto un modelo adecuado es aquel que permita hacer afirmaciones del tipo: "la probabilidad de que un individuo de 26 años, casado y cuya actividad económica es empleado asalariado sea un mal cliente en el crédito que le voy a otorgar es π[42]".

Es notorio que la modelización busca encontrar el mejor conjunto de covariantes X_i que expliquen a la variable endógena o dependiente y que existe una teoría estadística ampliamente conocida que es el modelo de regresión lineal mútiple. No obstante, las hipótesis de normalidad y de varianza constante (homoscedasticidad) supuestas en este modelo no son sostenibles cuando se intenta modelar una variable de elección del tipo binario.

Los modelos lineales generalizados, GLM por sus siglas en inglés[43], o unificación de los modelos de regresión lineal y no lineal salvan este problema al permitir incorporar distribuciones no normales de la variable respuesta. En este conjunto de modelos la condición requerida es que la distribución de la variable dependiente sea un miembro de la familia exponencial[44].

---

[41] En términos generales la variable es cualitativa o categórica, dentro de este conjunto de variables se encuentran las de tipo binario.

[42] Donde π debe ser un valor numérico comprendido entre 0 y 1, por ejemplo 0.78 o 78%.

[43] Generalized linear model.

[44] Son miembros de la familia exponencial las distribuciones normal, de Poisson, binomial, exponencial y gamma. El modelo lineal con error normal es un caso especial del GLM. Dobson, Annette elabora una muy buena explicación de las distribuciones de la familia exponencial en su libro An Introduction to Generalized Linear Models, págs 26-31. Editorial Chapman & Hall, 1993.

El costo del modelo lineal en la explicación de una variable de respuesta binaria[45]

Un modelo lineal puede expresarse como:

$$Y_i = \alpha + \beta X_i + \varepsilon_i$$

Donde X_i es el valor que toma la variable X para la i-ésima observación y Y_i, al ser una variable de respuesta binaria toma los valores de 0 o 1 de tal forma que:

$$Y_i = \begin{cases} 1 \text{ si es buen cliente (éxito)} \\ 0 \text{ si es mal cliente (fracaso)} \end{cases}$$

ε_i es un término aleatorio distribuido independiente con media 0 y varianza constante e igual a 1.

En este caso el valor esperado de la variable de respuesta vendrá dado por la expresión:

$$E(Y_i) = 1(\pi_i) + 0(1 - \pi_i) = \pi_i$$

ya que la variable respuesta sólo puede tomar dos valores (0 o 1). En este caso π_i representa la probabilidad de que $Y_i = 1$ y $(1 - \pi_i)$ la probabilidad de que $Y_i = 0$.

Sin embargo, un error que rápidamente salta a la vista es la posibilidad de ocurrencia de valores por fuera del intervalo definido para una probabilidad ($0 <= \pi <= 1$), pues el predictor $\alpha + \beta X_i$ puede tomar cualquier valor en el rango de los números reales. Esto exigirá aplicar una restricción al conjunto de la X de tal forma que el modelo sea válido dentro de un subconjunto de R. Usualmente y con el fin de que la variable respuesta sea interpretada como una probabilidad el modelo se escribe como:

$$\pi_i = \begin{cases} \alpha + \beta X_i \ ; 0 < \alpha + \beta X_i < 1 \\ 1 \ ; \alpha + \beta X_i \geq 1 \\ 0 \ ; \alpha + \beta X_i \leq 0 \end{cases}$$

Impuesta la restricción al conjunto de la X queda un problema adicional asociado con la distribución del término de error. Si se reemplaza por 1 y 0 los valores

[45] Se sigue de cerca el desarrollo elaborado por Robert Pindyck y Daniel Rubinfeld en su texto Econometría: Modelos y Pronósticos, págs 313-317. Editorial McGraw-Hill, 2001.

de Y_i en el modelo de regresión lineal y recordando que $\varepsilon_i = Y_i - \alpha + \beta X_i$ se observa que cuando:

$$Y_i = \begin{cases} 1 \Rightarrow \varepsilon_i = 1 - \alpha - \beta X_i \text{ con Probabilidad } \pi_i \\ 0 \Rightarrow \varepsilon_i = -\alpha - \beta X_i \text{ con Probabilidad } 1 - \pi_i \end{cases}$$

Con este resultado ya se puede concluir que el empleo de una variable respuesta binaria imposibilita la normalidad de errores ya que estos sólo podrán tener dos valores: 0 o 1.

Conociendo la distribución de probabilidad del término de error y asumiendo que este tiene media 0 se puede calcular su varianza una vez encontrados los valores de probabilidad.

Sea:

$$E(\varepsilon_i) = (1 - \alpha - \beta X_i)\pi_i + (-\alpha - \beta X_i)(1 - \pi_i) = 0$$

resolviendo para π_i se encuentra que:

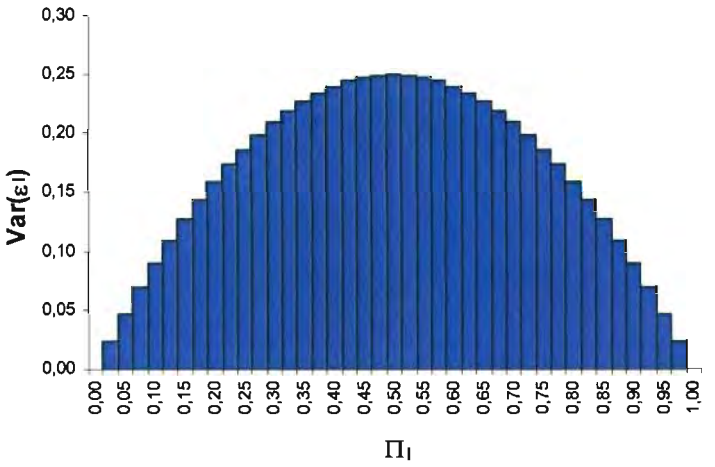
$$(1 - \alpha - \beta X_i)\pi_i + (-\alpha - \beta X_i)(1 - \pi_i) = 0$$
$$(1 - \alpha - \beta X_i)\pi_i + (-\alpha - \beta X_i) - (-\alpha - \beta X_i)\pi_i = 0$$
$$(1 - \alpha - \beta X_i + \alpha + \beta X_i)\pi_i + (-\alpha - \beta X_i) = 0$$
$$\pi_i + (-\alpha - \beta X_i) = 0$$
$$\pi_i = \alpha + \beta X_i$$

y al calcular la varianza se obtiene:

$$Var(\varepsilon_i) = E(\varepsilon_i^2) - E(\varepsilon_i)^2 = E(\varepsilon_i^2) = (1 - \alpha - \beta X_i)^2 \pi_i + (-\alpha - \beta X_i)^2 (1 - \pi_i)$$
$$Var(\varepsilon_i) = (1 - \alpha - \beta X_i)^2 (\alpha + \beta X_i) + (-\alpha - \beta X_i)^2 (1 - \alpha - \beta X_i)$$
$$Var(\varepsilon_i) = (1 - \alpha - \beta X_i)((1 - \alpha - \beta X_i)(\alpha + \beta X_i) + (-\alpha - \beta X_i)^2)$$
$$Var(\varepsilon_i) = (1 - \alpha - \beta X_i)(\alpha + \beta X_i - \alpha^2 - \alpha\beta X_i - \beta X_i \alpha - (\beta X_i)^2 + \alpha^2 + 2\alpha\beta X_i + (\beta X_i)^2)$$
$$Var(\varepsilon_i) = (1 - \alpha - \beta X_i)(\alpha + \beta X_i)$$
$$Var(\varepsilon_i) = \pi_i (1 - \pi_i) = E(Y_i)[1 - E(Y_i)]$$

Note que para valores de π_i cercanos a cero o a uno se tendrán varianzas relativamente pequeñas, mientras que para valores cercanos a ½ los valores de las varianzas serán mayores. Esto muestra que el término de error es heterocedástico, lo que originará pérdida de eficiencia en el estimador de mínimos cuadrados ordinarios.

COMPORTAMIENTO DE LA VARIANZA DEL TÉRMINO DE ERROR EN EL MRL CON VARIABLE RESPUESTA CUALITATIVA



Elaboración propia

Además, se sabe que la varianza de las observaciones está definida por:

$$\sigma_{yi}^2 = \pi_i(1-\pi_i) = E(Y_i)[1-E(Y_i)]$$

note que esta expresión es equivalente a la varianza del término de error y que ambas son funciones de la media, lo que implica que la varianza de las observaciones tampoco es una constante.

Este resultado deriva del hecho de trabajar con una variable respuesta binaria, por lo tanto si la variable dependiente es binaria los términos de error no van a ser normales y no tendrán varianza constante.

Dadas las dificultades que presenta este modelo, es deseable encontrar otro que permita obtener predicciones que caigan para todas las X en el intervalo (0,1), es decir encontrar una transformación que traslade los valores de la covariante X, definida sobre todo el conjunto de los reales a una probabilidad que varía entre 0 y 1, a la vez que se mantiene la propiedad de que para todos los valores de X incrementos en su valor vayan asociados con incrementos (decrementos) en la variable dependiente (función monótona). Es decir, la solución implica asociar a $\pi(x)$ una función $F(\eta)$, donde F es una función de distribución definida sobre los reales. Puesto que la función de distribución F es monótona creciente, la función de probabilidades $\pi(x)$ también lo es.

Este es el caso de los modelos lineales generalizados y particularmente el del modelo de regresión logística. La ventaja de estos modelos es que por construcción no

necesitan establecer únicamente una relación lineal entre la variable independiente y las variables explicativas. Para comprender mejor esto se describe brevemente su funcionamiento.

Todo modelo lineal generalizado tiene tres componentes:

1. Componente aleatorio.- que identifica la variable respuesta Y y asume una distribución de probabilidad para ella.
2. Componente sistemático.- variables explicativas
3. Link o enlace.- describe la relación funcional entre el componente sistemático y el valor esperado del componente aleatorio.

El **componente sistemático** denota el valor esperado de Y:

$$\mu = \mathrm{E(Y)}$$

En un MLG este componente variará de acuerdo a los niveles de las variables explicativas, por ende el componente sistemático de un MLG especifica las variables explicativas que entran en forma lineal como predictores en el lado derecho de la ecuación del modelo.

$$\alpha + \beta_1 X_1 + \beta_2 X_2 + \ldots + \beta_k X_k$$

La **función link o de enlace**, especifica como el valor esperado de Y se relaciona con las variables explicativas en el predictor lineal. Se puede modelar la media μ directamente o como una función monótono $g(\mu)$ de la media.

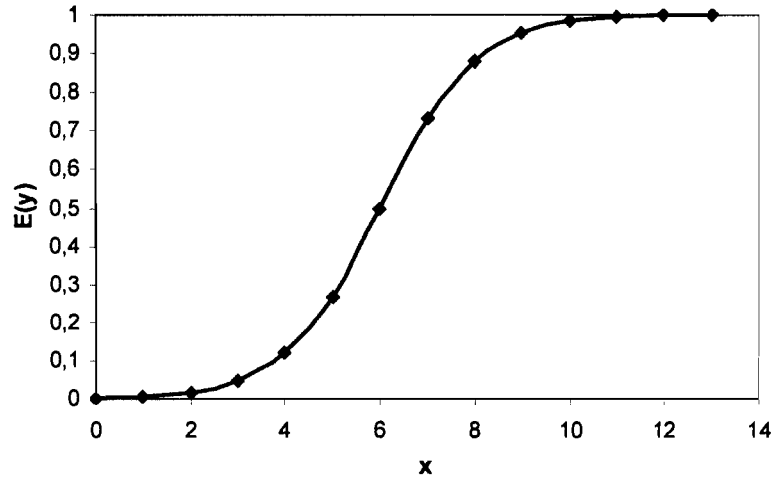
$$g(\mu) = \alpha + \beta_1 X_1 + \beta_2 X_2 + \ldots + \beta_k X_k$$

$\downarrow$

función de enlace

En el caso más simple $g(\mu)$ puede ser igual a μ, el valor esperado de Y, siendo la función link la identidad, que especifica un modelo lineal para la media de la variable respuesta: $\mu = \alpha + \beta X_i + \ldots + \beta X_k$. Pero ya se mostró que la relación entre la variable respuesta y las variables explicativas, en un contexto de probabilidad, no es lineal por lo que es deseable modelar una relación no lineal entre estas. Por ejemplo, una variación en la variable edad del cliente (X), puede tener un menor impacto sobre $\pi(x)$ cuando

este se encuentra cercano a 0 o 1, que cuando esté cerca de la mitad de su rango. En este caso una curva en forma de S (como una función acumulada de distribución) puede recoger mejor esta relación.



Elaboración propia

La de mayor importancia en cuanto a su uso es la distribución logística:

$$F(\eta) = \frac{\exp(\eta)}{1+\exp(\eta)}; \eta \in \mathrm{R}$$

La inversa de esta distribución, función de fractiles, es la curva logit:

$$\eta = \mathrm{logit}(\pi(\mathrm{x})) = \ln\left(\frac{\pi(\mathrm{x})}{1-\pi(\mathrm{x})}\right), \text{donde } \pi(\mathrm{x}) = F(\eta)$$

Note que η es el equivalente a la función de enlace g(μ), mostrándose ahora que la relación entre el componente sistemático y la variable respuesta toma la forma:

$$\eta = \mathrm{logit}(\pi(\mathrm{x})) = \ln\left(\frac{\pi(\mathrm{x})}{1-\pi(\mathrm{x})}\right) = \alpha + \beta_1 X_1 + \beta_2 X_2 + \ldots + \beta_k X_k$$

Permitiendo hallar el valor esperado de la variable respuesta Y como un valor de la función de respuesta logística, es decir como un valor de probabilidad, que es el objetivo en el caso de la medición del riesgo de incumplimiento en un crédito por parte de un cliente.

$$\mathrm{E(Y)} = \frac{\exp(\eta)}{1+\exp(\eta)}$$

$$E(Y) = \frac{\exp(\alpha + \beta_1 X_1 + \beta_2 X_2 + ... + \beta_k X_k)}{1 + \exp(\alpha + \beta_1 X_1 + \beta_2 X_2 + ... + \beta_k X_k)}$$

Tal vez una de las dificultades del modelo de regresión logístico en relación al modelo de regresión lineal es la interpretación de sus coeficientes. No obstante, en la realidad su explicación, más que compleja es simplemente distinta. Esto puede verse en el caso más simple, en que el predictor lineal tiene un solo regresor.

$$\eta(x_i) = \alpha + \beta_1 X_1$$

Para este caso el valor pronosticado en x_i+1 es

$$\eta(x_i + 1) = \alpha + \beta_1 (X_1 + 1)$$

y la diferencia entre las dos predicciones queda como:

$$\eta(x_i + 1) - \eta(x_i) = \alpha + \beta_1 (X_1 + 1) - \alpha - \beta_1 X_1$$
$$\eta(x_i + 1) - \eta(x_i) = \beta_1$$

Recordando que $\eta(x_i)$ es el logaritmo de la predicción cuando la covariante es igual a x_i y $\eta(x_i+1)$ es el logaritmo de la predicción cuando la covariante es igual a x_i +1. Por lo tanto la diferencia entre los dos valores predichos se puede expresar como:

$$\eta(x_i + 1) - \eta(x_i) = \ln\left(\frac{\eta(x_i + 1)}{\eta(x_i)}\right) = \beta_1$$

aplicando antilogaritmos

$$Odds = \left(\frac{\eta(x_i + 1)}{\eta(x_i)} \right) = e^{\beta_1}$$

Pudiéndose interpretar el odds ratio o cociente de ventaja como el aumento estimado en la probabilidad de éxito asociado con un cambio unitario en el valor de la covariante. En términos generales se puede decir que el aumento estimado del odds producto de un cambio de δ unidades en la variable predictora es:

$$e^{\delta\beta_1}$$

Para el caso de la regresión múltiple en el modelo logístico, la interpretación es similar que en el caso univariante, sólo que debemos asumir que las demás variables incluidas en el modelo como explicativas permanecen constantes.

Anexo 4.- La curva ROC y el coeficiente de GINI

Cuando se analiza un modelo scoring uno de los principales referentes suele ser la matriz de confusión o tabla de clasificación, la cual como se mostró anteriormente resume el poder predictivo de la scorecard mediante el cálculo de la tasa de error:

$$\text{Tasa de error del modelo} = \frac{bM + mB}{n}$$

MATRIZ DE CONFUSIÓN

		Valores observados		
		Buenos	**Malos**	**Total**
Valores predichos	**Buenos**	bB	bM	b
	Malos	mB	mM	m
	Total	nB	nM	n

Elaboración propia

No obstante, esta matriz tiene una limitación importante que es el asumir un único punto de corte (cutoff).

Es decir, que la scorecard se está evaluando como si su respuesta fuera del tipo dicotómica, por lo que en principio estamos omitiendo el hecho real que los resultados del scoring se miden en una escala continua. Para superar este problema se debe elegir distintos puntos de corte (cutoff) que permitan una clasificación dicotómica de los valores de prueba según sean superiores o inferiores al valor seleccionado. La diferencia esencial con el caso más simple es que ahora se cuenta no sólo con un único par de valores de sensibilidad y especificidad que definan la exactitud de la prueba[46], sino con un conjunto de pares correspondientes cada uno con un distinto punto de corte[47].

[46] El nombre curva ROC tiene sus orígenes en la estimación de los errores de clasificación en la transmisión y recepción de señales. Luego usando la clasificación como una herramienta en los estudios epidemiológicos se introdujeron dos conceptos adicionales que ahora se emplean en la medición del desempeño del scorecard: la sensibilidad y la especificidad. En epidemiología se introdujeron los términos especificidad y sensibilidad para medir la exactitud de una prueba diagnóstica cuando se emplea una prueba dicotómica (cuyo resultado se puede interpretar directamente como positivo [cumple con la condición de la prueba diagnóstico] o negativo [no cumple con la condición]). La sensibilidad es la probabilidad de clasificar correctamente a un individuo cuyo estado real sea el definido como positivo

Este procedimiento constituye la esencia de las curvas ROC o del análisis ROC (Receiver Operating Charasteristic), cuyo origen y nombre proviene de los estudios aplicados a la detección de señales por radar en donde el objeto era encontrar un operador que pueda con el menor grado de error interpretar correctamente los picos en la pantalla del radar y decidir sobre la presencia o no de un misil.

Para centrar ideas observe la gráfica en la que se ha trazado la distribución de los buenos y de los malos, las cuales muestran un determinado nivel de solapamiento. Si se considera un determinado valor arbitrario de la scorecard al cual se denominará punto de corte, la sensibilidad (proporción de buenos predichos como buenos respecto al total de buenos) y (1-la especificidad) (proporción de malos predichos como buenos respecto al total de malos) se corresponderán respectivamente con el área a la derecha de ese punto bajo la distribución empírica de la población de buenos (áreas clara y oscura) y de la población de malos (área oscura). La curva ROC se obtiene representando para cada posible elección de valor de corte, la proporción de buenos en el eje de las ordenadas (sensibilidad) y la proporción de malos predichos como buenos respecto al total de malos en el eje de las abscisas (1 - especificidad).

---

respecto a la condición que estudia la prueba (conocida como proporción de verdaderos positivos PVP). La especificidad es la probabilidad de clasificar correctamente a un individuo cuyo estado real respecto a la condición que estudia la prueba sea negativo (conocida como la diferencia entre 1 y la proporción de falsos positivos (PFP: la prueba detecta como positivo siendo negativo).

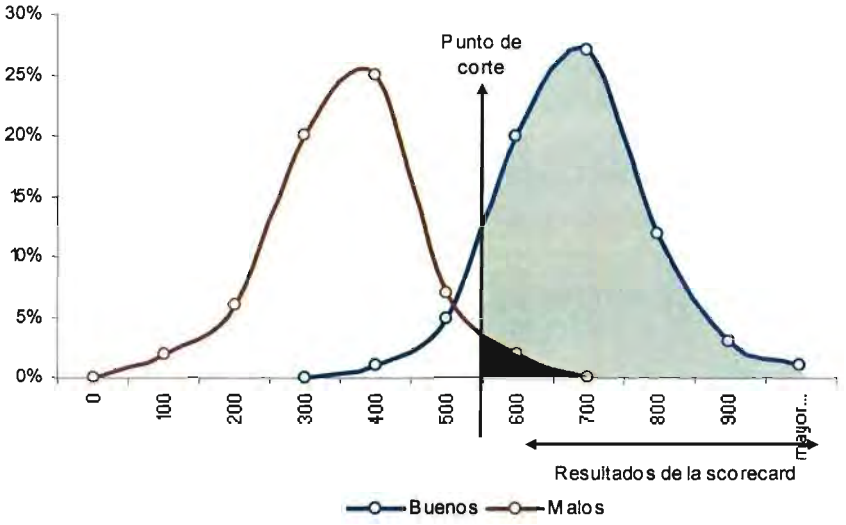
[47] Se puede emplear la matriz de confusión para calcular la sensibilidad y la especificidad. Cuando los datos de una muestra de clientes se clasifican en una tabla de contingencia por el resultado de la scorecard y su estado real respecto al crédito obtenido, es fácil estimar a partir de esta tabla la sensibilidad y la especificidad de la prueba:
La sensibilidad viene expresada como:

$Se = \frac{bB}{bB + mB}$ (proporción de buenos predichos como buenos respecto al total de buenos observados)

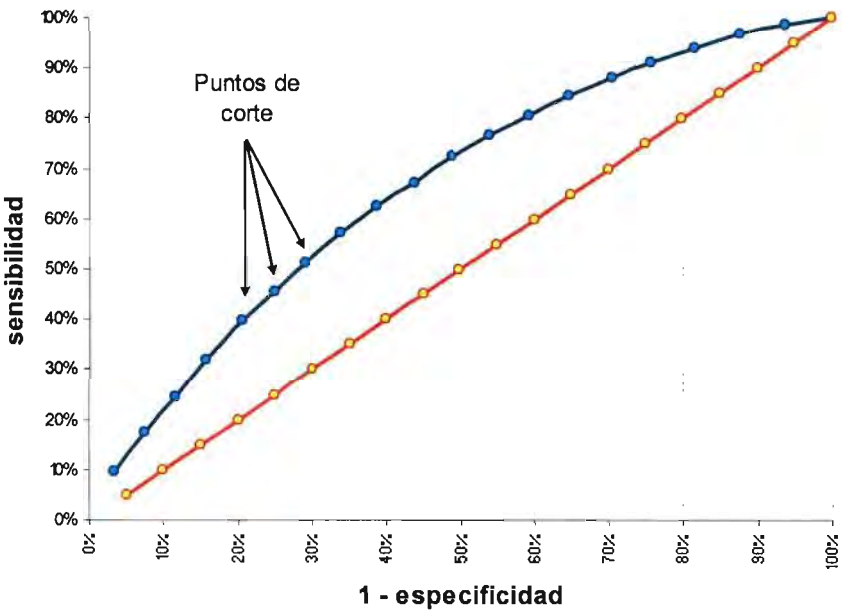
La especificidad viene expresada como:

$Sp = \frac{mM}{mM + bM} = 1 - \frac{bM}{mM + bM}$ (1 – proporción de malos predichos como buenos respecto al total de malos observados)

DISTRIBUCIÓN DE LOS RESULTADOS DE UNA SCORECARD EN LAS POBLACIONES DE CLIENTES BUENOS Y MALOS



CURVA ROC

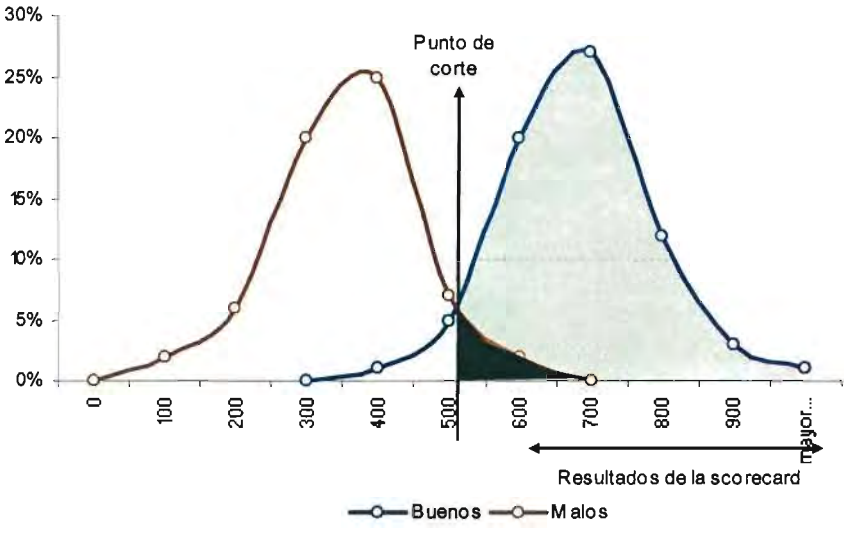


Elaboración propia

Mediante esta representación de los pares (1-especificidad; sensibilidad) obtenidos al considerar todos los puntos de corte de la scorecard, la curva ROC proporciona una representación global de la exactitud del modelo. Note que la curva ROC será necesariamente creciente, propiedad que refleja el compromiso entre

sensibilidad y especificidad: si se modifica el punto de corte para obtener mayor sensibilidad, sólo puede hacerse a expensas de disminuir al mismo tiempo la especificidad.

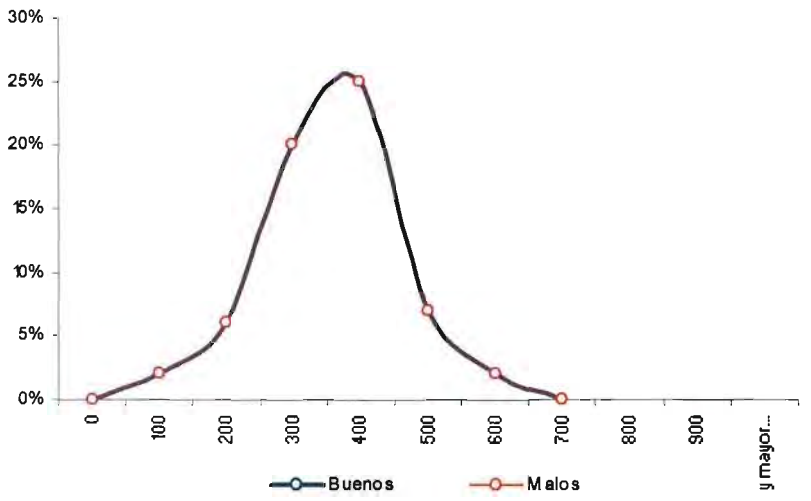
MODIFICACIÓN DEL PUNTO DE CORTE: GANANCIA EN SENSIBILIDAD Y PÉRDIDA EN ESPECIFICIDAD



Elaboración propia

Por otro lado, si el scorecard no permitiera discriminar entre los dos grupos, la curva ROC sería una diagonal, línea de equidad o de 45° (línea roja en la gráfica) y en este caso es importante citar que la ocurrencia de esto no es lo más apropiado en un modelo scoring, pues no representa ninguna ventaja en relación a una clasificación aleatoria dado que se conoce el ratio B:M de la población.

SCORECARD SIN PODER DISCRIMINATORIO



Elaboración propia

La exactitud de la prueba aumenta en la medida en que la curva se desplaza desde la línea de equidad hacia el vértice superior izquierdo. Si la discriminación fuera perfecta (100% de sensibilidad y 100% de especificidad) la curva pasaría por este punto.

Construcción empírica de una curva ROC

El método que se sigue en este estudio para construir una curva ROC es no paramétrico, pues no se hace ninguna suposición sobre la distribución de probabilidad que siguen los resultados de la scorecard y consiste en representar todos los pares de valores (1 - especificidad; sensibilidad) para todos los posibles valores de corte que se pueden considerar con la muestra que se dispone, es decir sustituimos las funciones de densidad por los histogramas obtenidos a partir de la muestra de clientes buenos y malos y construimos la curva ROC a partir de ellos. Se indica a continuación los pasos que pueden seguirse para obtener el gráfico de una curva ROC.

1.- Establecer grupos de igual tamaño (percentiles) para el puntaje obtenido por cliente.

2.- Ordenar la población de buenos y malos con base en los percentiles obtenidos

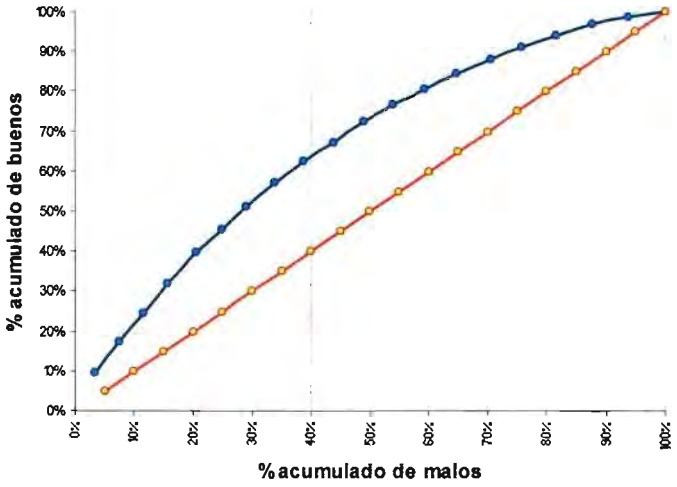
3.- Obtener la distribución acumulada de buenos y malos clientes

4.- Realizar la gráfica de las distribuciones acumuladas y de la recta de 45°

De tal forma de obtener una tabla del siguiente tipo:

Percentil	Score	# de Buenos	% de Buenos acumulado	# de Malos	% de Malos acumulados	Total general
0%-5%	247	1086	3,39%	1105	9,41%	2191
11%-15%	285	1294	7,42%	941	17,41%	2235
16%-20%	328	1310	11,51%	837	24,54%	2147
21%-25%	353	1369	15,77%	861	31,87%	2230
26%-30%	379	1539	20,57%	913	39,64%	2452
31%-35%	407	1370	24,85%	663	45,28%	2033
36%-40%	436	1369	29,11%	680	51,07%	2049
41%-45%	457	1549	33,94%	689	56,93%	2238
46%-50%	488	1546	38,77%	629	62,29%	2175
51%-55%	515	1598	43,75%	566	67,10%	2164
56%-60%	542	1668	48,95%	595	72,17%	2263
6%-10%	574	1618	53,99%	511	76,52%	2129
61%-65%	595	1718	59,35%	467	80,49%	2185
66%-70%	632	1714	64,70%	469	84,48%	2183
71%-75%	658	1824	70,38%	415	88,02%	2239
76%-80%	689	1761	75,87%	381	91,26%	2142
81%-85%	724	1863	81,68%	328	94,05%	2191
86%-90%	771	1887	87,57%	310	96,69%	2197
91%-95%	821	1974	93,72%	236	98,70%	2210
96%-100%	956	2013	100,00%	153	100,00%	2166
Total general		32070		11749		43819

CURVA ROC PARA LOS DATOS DE LA TABLA ANTERIOR



Elaboración propia

El coeficiente de GINI

Como se observa en la construcción de una curva ROC mientras más alejada (hacia arriba y a la izquierda de la línea de 45°) esté la curva, mejor será la exactitud de la scorecard. Esto sugiere que el área bajo la curva ROC se puede emplear como un índice de la exactitud del modelo. La exactitud máxima corresponderá a un valor igual a 0.5 y la mínima a uno de 0

En términos de probabilidad, si X_b y X_m son las dos variables aleatorias que representan los valores de la scorecard en las poblaciones de buenos y malos, puede probarse que el área de la "verdadera" curva ROC (aquella que se obtendría si el tamaño de la muestra fuera infinito y la escala de medida continua) es:

$$\theta = \Pr(X_b > X_m)$$

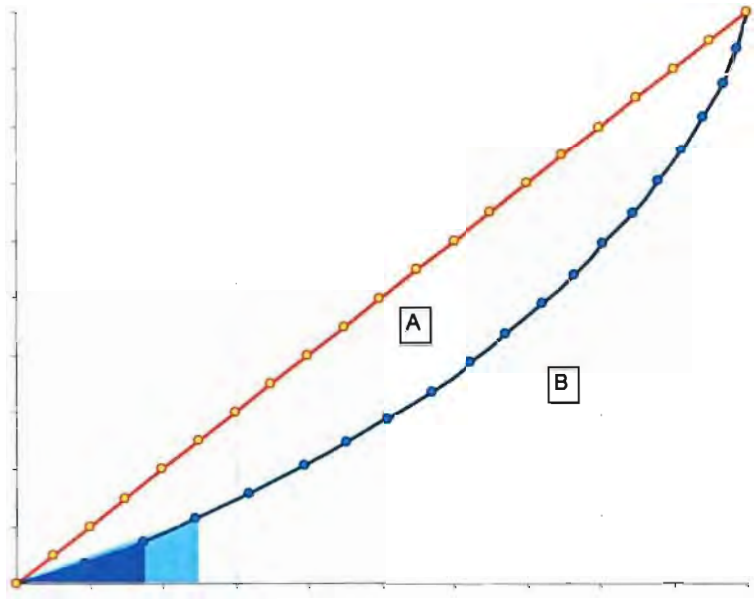
Es decir, la probabilidad de que si se eligen al azar un cliente bueno y uno malo, sea mayor el valor de la scorecard en el bueno que en el malo.

Cuando el método de obtención de la curva es empírico, el área puede calcularse por la regla trapezoidal, es decir como la suma de las áreas de todos los rectángulos y trapecios que se pueden formar bajo la curva. El duplo de esta área es la que se conoce como el coeficiente de GINI.

Este coeficiente tendrá la propiedad de que una perfecta clasificación que va a través del vértice superior tendrá un GINI de 1 mientras que una clasificación aleatoria con un ROC igual a la curva de equidad tendrá un GINI de 0. Por lo tanto, la ventaja de cálculo de este índice es lograr resumir el desempeño del scorecard para todos los puntos de corte (cutoff) del modelo[48].

El coeficiente de Gini es igual al área A, el área dentro de la curva ROC dividida para el área A + B. Para obtener esta medida lo primero que se debe hacer es encontrar el área B. Esto implica calcular el área para una serie de triángulos que se forman de la unión de cada puntaje, de la frecuencia acumulada asociada a ese punto y del punto de origen.

[48] Para facilidad del lector se invierte el sentido del gráfico para mostrar el cálculo del coeficiente.



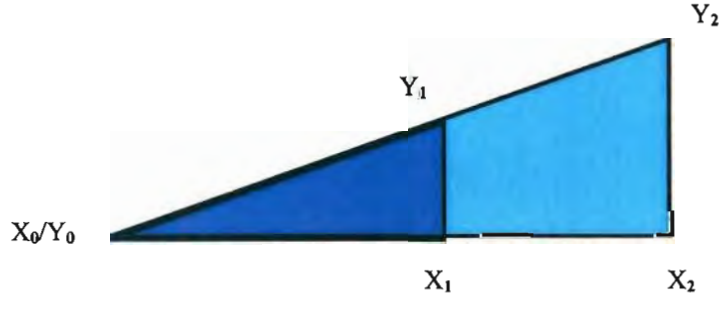
Elaboración propia

Por ejemplo, para el primer triángulo tenemos:

El área para este triángulo se calcula como:

$$\frac{1}{2}\times(x_1 - x_0)\times(y_1 - y_0)$$

El siguiente triángulo es:



y si queremos calcular su área (el área del triángulo turquesa) esta es:

$$\frac{1}{2}\times(x_2-x_0)\times(y_2-y_0)$$

Note sin embargo que ya se ha calculado el área del triángulo rojo por lo que debemos descontarla:

$$\left[\frac{1}{2}\times(x_2-x_0)\times(y_2-y_0)\right]-\left[\frac{1}{2}\times(x_1-x_0)\times(y_1-y_0)\right]$$

o en caso contrario únicamente adicionar el área del trapezoide:

$$\frac{1}{2}\times(y_1+y_2)\times(x_2-x_1)$$

En general el coeficiente de Gini puede calcularse por medio de la siguiente fórmula para el área B:

$$B=\frac{1}{2}\{(x_1-x_0)\times(y_1+y_0)+(x_2-x_1)\times(y_2+y_1)+\cdots+(x_k-x_{k-1})\times(y_k+y_{k-1})\}$$

En donde x_i y y_i son las frecuencias relativas acumuladas de los ejes X y Y; $x_0 = y_0 = 0$ y $x_k = y_k = 100$.

Alternativamente el coeficiente de Gini puede expresarse como:

$$G=\frac{A}{A+B}=\frac{A+B}{A+B}-\frac{B}{A+B}=1-\frac{B}{A+B}$$

En el caso en que $x_k = y_k = 100$, el área

$$A+B=\frac{10000}{2}$$

y por lo tanto

$$G=1-\frac{2B}{10000}$$

Para los datos del caso anterior se tiene:

Percentil	# de Buenos	% Buenos	% de Buenos acumulado	# de Malos	% de Malos	% de Malos acumulados	Total general	mi - mi-1	bi + bi-1	Gini = 1-suma[(mi - mi-1)(bi + bi-1)]
0%-5%	1086	3,4%	3,4%	1105	9,4%	9,4%	2191	9%	3%	0%
6%-10%	1294	4,0%	7,4%	941	8,0%	17,4%	2235	8%	11%	1%
11%-15%	1310	4,1%	11,5%	837	7,1%	24,5%	2147	7%	19%	1%
16%-20%	1369	4,3%	15,8%	861	7,3%	31,9%	2230	7%	27%	2%
21%-25%	1539	4,8%	20,6%	913	7,8%	39,6%	2452	8%	36%	3%
26%-30%	1370	4,3%	24,8%	663	5,6%	45,3%	2033	6%	45%	3%
31%-35%	1369	4,3%	29,1%	680	5,8%	51,1%	2049	6%	54%	3%
36%-40%	1549	4,8%	33,9%	689	5,9%	56,9%	2238	6%	63%	4%
41%-45%	1546	4,8%	38,8%	629	5,4%	62,3%	2175	5%	73%	4%
46%-50%	1598	5,0%	43,7%	566	4,8%	67,1%	2164	5%	83%	4%
51%-55%	1668	5,2%	48,9%	595	5,1%	72,2%	2263	5%	93%	5%
56%-60%	1618	5,0%	54,0%	511	4,3%	76,5%	2129	4%	103%	4%
61%-65%	1718	5,4%	59,4%	467	4,0%	80,5%	2185	4%	113%	5%
66%-70%	1714	5,3%	64,7%	469	4,0%	84,5%	2183	4%	124%	5%
71%-75%	1824	5,7%	70,4%	415	3,5%	88,0%	2239	4%	135%	5%
76%-80%	1761	5,5%	75,9%	381	3,2%	91,3%	2142	3%	146%	5%
81%-85%	1863	5,8%	81,7%	328	2,8%	94,1%	2191	3%	158%	4%
86%-90%	1887	5,9%	87,6%	310	2,6%	96,7%	2197	3%	169%	4%
91%-95%	1974	6,2%	93,7%	236	2,0%	98,7%	2210	2%	181%	4%
96%-100%	2013	6,3%	100,0%	153	1,3%	100,0%	2166	1%	194%	3%
Total general	32070	100,0%		11749	100,0%		43819			32%

Elaboración propia

Anexo 5.- Variables relevantes para la inclusión en el modelo de regresión obtenidas a través de árboles de decisión usando el algoritmo CHAID

Variable: Ciudad

Node 0			
	Malo	26,81	11749
	Bueno	73,19	32070

Ciudad UB (Adj. P-value=0,0000, Chi-square=1098,3787, df=2)

Guayaquil, Ibarra, Manta, Riobamba, Loja	Node 6	Malo	32,88	7874
		Bueno	67,12	16072
Cuenca, Quito	Node 5	Malo	20,76	3585
		Bueno	79,24	13682
Ambato, Latacunga	Node 7	Malo	11,13	290
		Bueno	88,87	2316

Variable: Rango de trabajo

Node 0			
	Malo	26,81	11749
	Bueno	73,19	32070

Rango trabajo (Adj. P-value=0,0000, Chi-square=414,3439, df=2)

<1	Node 29	Malo	34,67	1299
		Bueno	65,33	2448
1-3	Node 30	Malo	28,79	7554
		Bueno	71,21	18681
>3	Node 31	Malo	20,93	2896
		Bueno	79,07	10941

Variable: Estado Civil

Node 0			
	Malo	26,81	11749
	Bueno	73,19	32070

Estado Civil (Adj. P-value=0,0000, Chi-square=312,2349, df=2)

Separado, Unión libre	Node 50	Malo	35,64	1796
		Bueno	64,36	3243
Soltero, divorciado, viudo	Node 49	Malo	27,51	5998
		Bueno	72,49	15808
Casado	Node 48	Malo	23,30	3955
		Bueno	76,70	13019

Variable: Carga Familiar

Node 0	Malo	26,81	11749
	Bueno	73,19	32070

Carga familiar (Adj. P-value=0,0000, Chi-square=25,5678, df=1)

>=3	Node 54	Malo	30,39	1096
		Bueno	69,61	2511
<=2	Node 53	Malo	26,49	10653
		Bueno	73,51	29559

Variable: Educación

Node 0	Malo	26,81	11749
	Bueno	73,19	32070

Educación (Adj. P-value=0,0000, Chi-square=154,3355, df=1)

Primaria, secundaria, técnica	Node 64	Malo	27,78	10719
		Bueno	72,22	27865
Universitario	Node 65	Malo	19,68	1030
		Bueno	80,32	4205

Variable: Vivienda

Node 0	Malo	26,81	11749
	Bueno	73,19	32070

Vivienda (Adj. P-value=0,0000, Chi-square=58,2885, df=1)

Arrendada, propia	Node 76	Malo	28,44	6168
		Bueno	71,56	15516
Familiar, herencia, hipotecada, anticresis	Node 77	Malo	25,21	5581
		Bueno	74,79	16554

Variable: Actividad Económica

Node 0	Malo	26,81	11749
	Bueno	73,19	32070

Actividad económica (Adj. P-value=0,0000, Chi-square=240,5977, df=2)

Independiente, estudiante	Node 80	Malo	55,04	251
		Bueno	44,96	205
Comerciante	Node 79	Malo	42,97	165
		Bueno	57,03	219
Empleado	Node 78	Malo	26,37	11333
		Bueno	73,63	31646

Variable: Seguridad Social (IESS)

Node 0	Malo	26,81	11749
	Bueno	73,19	32070

IESS (Adj. P-value=0,0000, Chi-square=172,1843, df=1)

NO	Node 86	Malo	33,3	2260
		Bueno	66,7	4527
SI	Node 85	Malo	25,62	9489
		Bueno	74,38	27543

Variable: Género

Node 0	Malo	26,81	11749
	Bueno	73,19	32070

Género (Adj. P-value=0,0000, Chi-square=298,4560, df=1)

MASCULINO	Node 89	Malo	29,58	8237
		Bueno	70,42	19608
FEMENINO	Node 90	Malo	21,99	3512
		Bueno	78,01	12462

Variable: Rango Edad del Cliente

Node 0	Malo	26,81	11749
	Bueno	73,19	32070

Rango edad (Adj. P-value=0,0000, Chi-square=94,4688, df=3)

<25	Node 110	Malo	28,62	3609
		Bueno	71,38	9000
26-40	Node 111	Malo	27,04	6631
		Bueno	72,96	17888
41-45	Node 112	Malo	24,41	785
		Bueno	75,59	2431
>45	Node 113	Malo	20,83	724
		Bueno	79,17	2751

Variable: Rango Sueldo

Node 0	Malo	26,81	11749
	Bueno	73,19	32070

Rango sueldo (Adj. P-value=0,0000, Chi-square=100,0059, df=2)

<200	Node 118	Malo	29,58	3533
		Bueno	70,42	8412
200-300	Node 119	Malo	27,14	4701
		Bueno	72,86	12619
>300	Node 120	Malo	24,15	3515
		Bueno	75,85	11039

Variable: Rango Compromiso Financiero

Node 0			
	Malo	26,81	11749
	Bueno	73,19	32070

Rango Compromiso Financiero (Adj. P-value=0,0025, Chi-square=25,7103, df=2)

>=20	Node 135	Malo	29,68	908
		Bueno	70,32	2151
sin compromiso	Node 137	Malo	26,87	9690
		Bueno	73,13	26372
0%-20%	Node 136	Malo	24,5	1151
		Bueno	75,5	3547

Variable: Tiempo de residencia

Node 0			
Malo	26,81	11749	
Bueno	73,19	32070	

Tiempo de residencia (Adj. P-value=0,0000, Chi-square=59,2350, df=3)

<1	Node 13	Malo	27,79	92
		Bueno	72,21	239
1-3	Node 14	Malo	27,48	1297
		Bueno	72,52	3422
3-5	Node 15	Malo	30,2	2283
		Bueno	69,8	5277
>5	Node 16	Malo	25,88	8077
		Bueno	74,12	23132

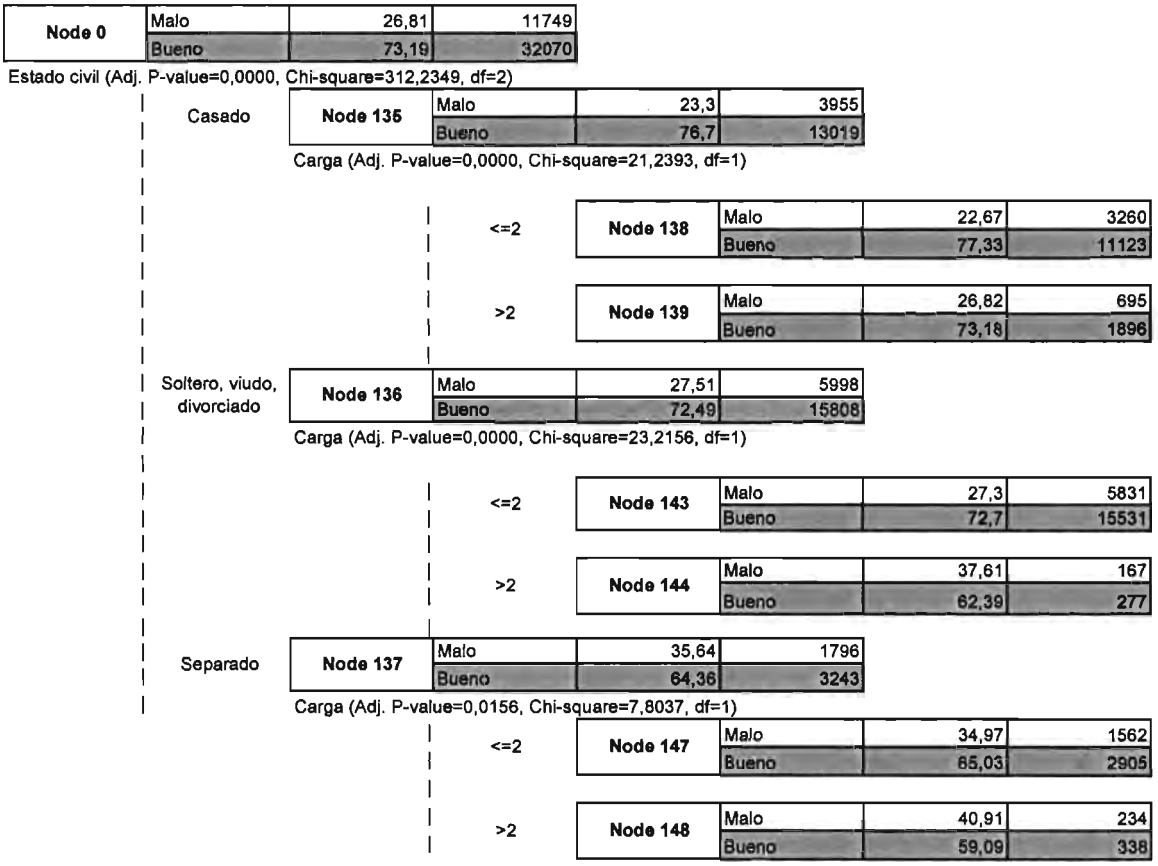
Variable: Provincia del Cliente

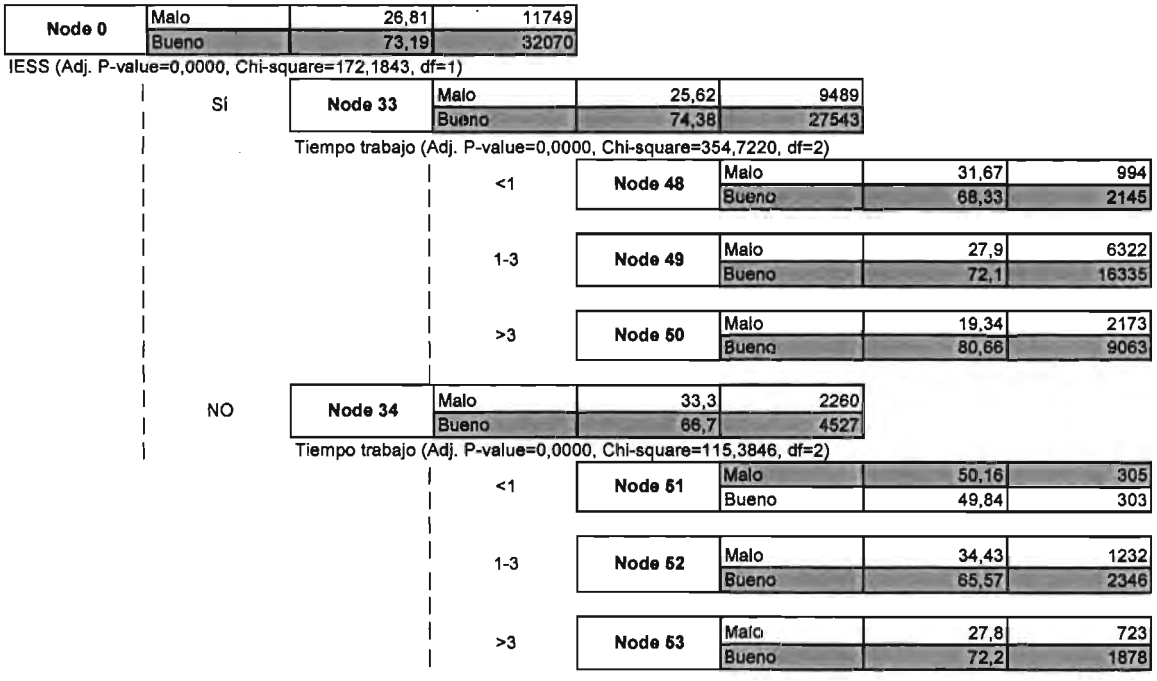
Node 0			
Malo	26,81	11749	
Bueno	73,19	32070	

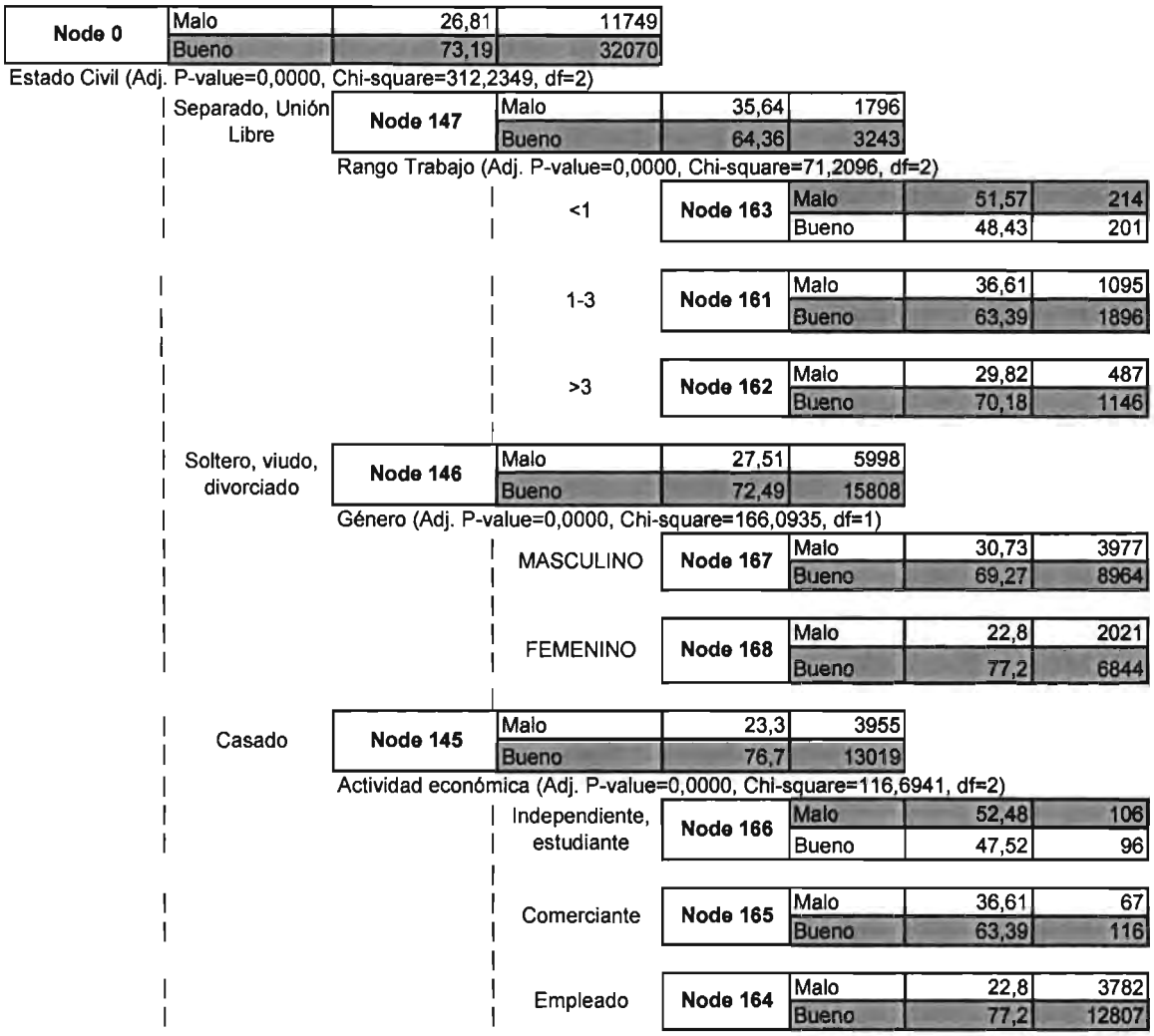
Provincia cliente (Adj. P-value=0,0000, Chi-square=1067,6885, df=2)

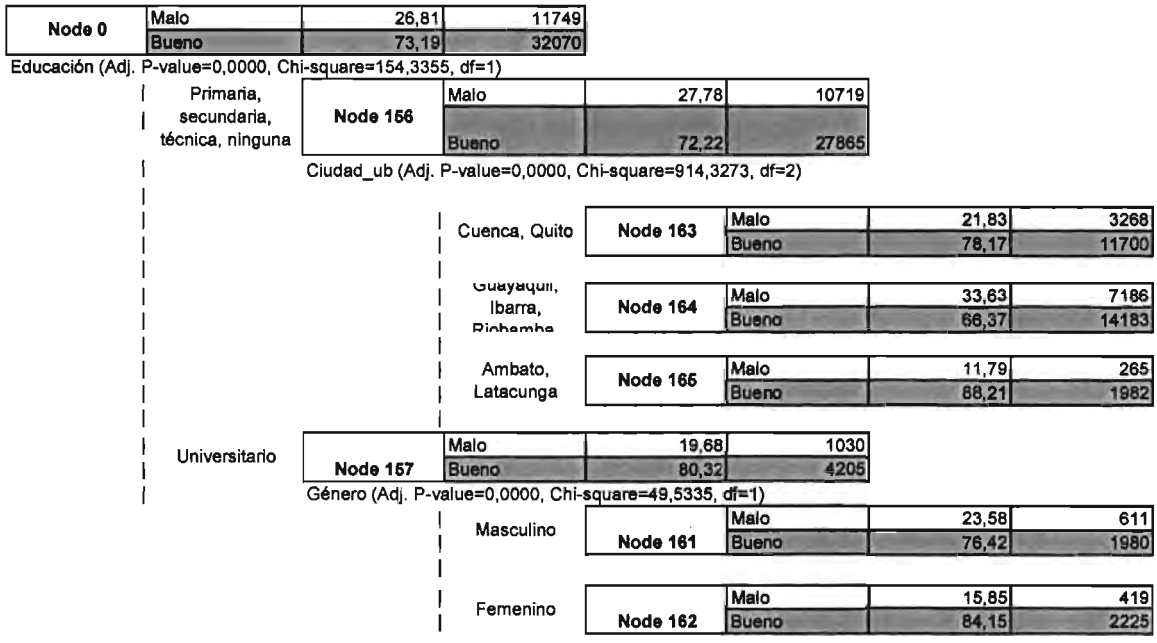
Guayas, Chimborazo, Loja, Cañar	Node 170	Malo	32,89	7884
		Bueno	67,11	16084
Pichincha, Imbabura	Node 171	Malo	20,98	3138
		Bueno	79,02	11818
Azuay, Tungurahua, Zamora, Sucumbíos, Cotopaxi, EL Oro, Galápagos, Bolívar	Node 172	Malo	14,85	727
		Bueno	85,15	4168

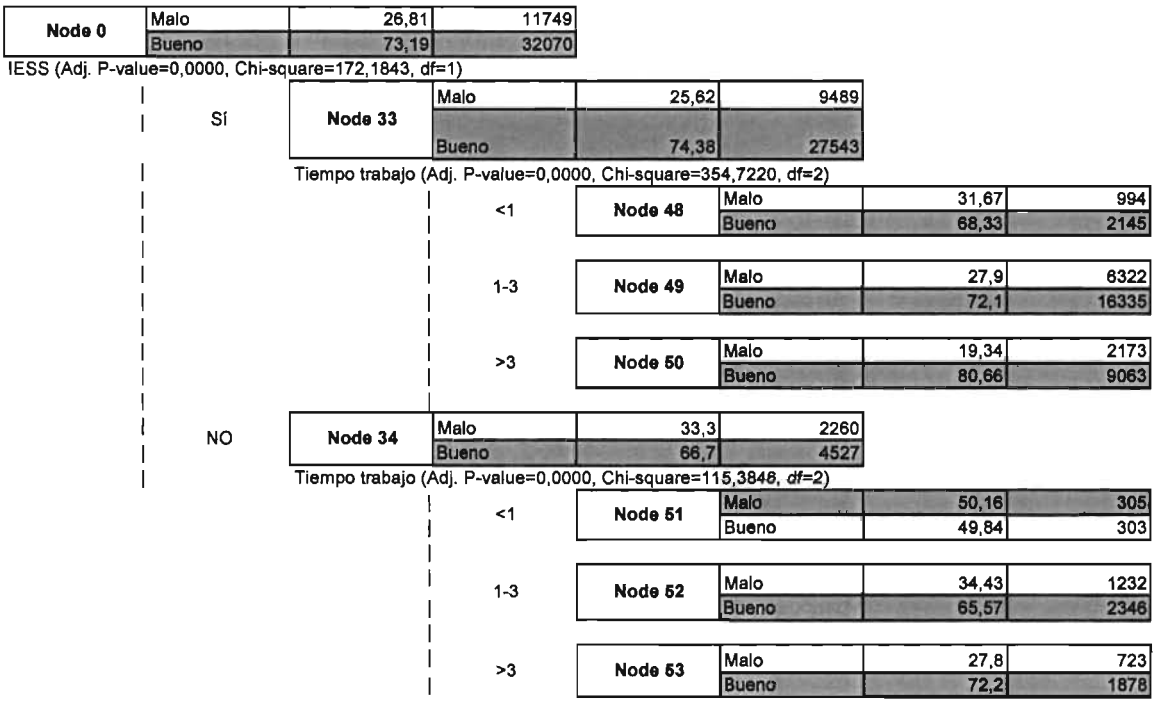
Variables cruzadas

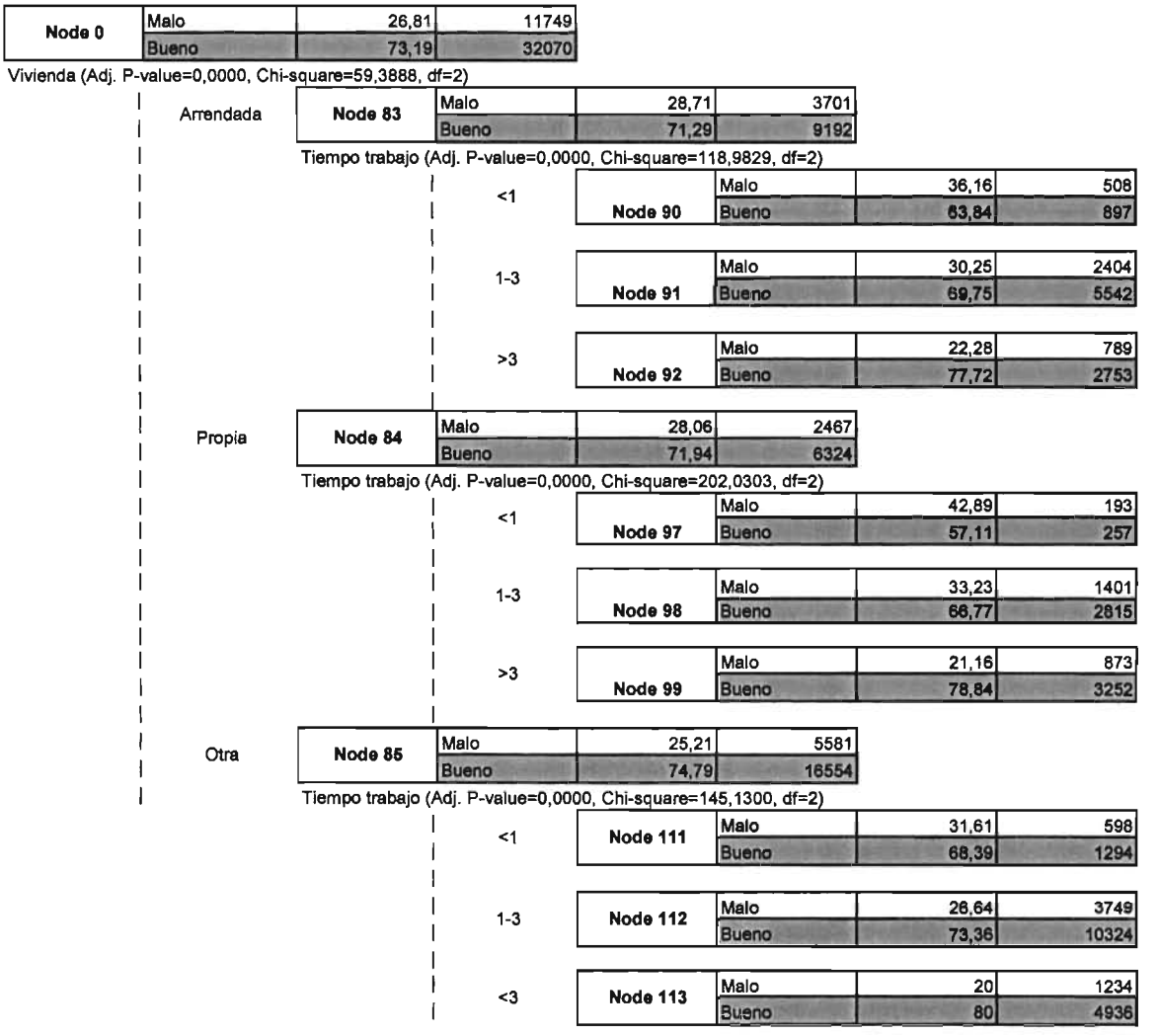





Node 0
Malo 26,81 11749
Bueno 73,19 32070
Estado Civil (Adj. P-value=0,0000, Chi-square=312,2349, df=2)
Separado, Unión Libre
Node 147
Malo 35,64 1796
Bueno 64,36 3243
Rango Trabajo (Adj. P-value=0,0000, Chi-square=71,2096, df=2)
<1
Node 163
Malo 51,57 214
Bueno 48,43 201
1-3
Node 161
Malo 36,61 1095
Bueno 63,39 1896
>3
Node 162
Malo 29,82 487
Bueno 70,18 1146
Soltero, viudo, divorciado
Node 146
Malo 27,51 5998
Bueno 72,49 15808
Género (Adj. P-value=0,0000, Chi-square=166,0935, df=1)
MASCULINO
Node 167
Malo 30,73 3977
Bueno 69,27 8964
FEMENINO
Node 168
Malo 22,8 2021
Bueno 77,2 6844
Casado
Node 145
Malo 23,3 3955
Bueno 76,7 13019
Actividad económica (Adj. P-value=0,0000, Chi-square=116,6941, df=2)
Independiente, estudiante
Node 166
Malo 52,48 106
Bueno 47,52 96
Comerciante
Node 165
Malo 36,61 67
Bueno 63,39 116
Empleado
Node 164
Malo 22,8 3782
Bueno 77,2 12807


Node 0
Malo 26,81 11749
Bueno 73,19 32070
Educación (Adj. P-value=0,0000, Chi-square=154,3355, df=1)
Primaria, secundaria, técnica, ninguna
Node 156
Malo 27,78 10719
Bueno 72,22 27865
Ciudad_ub (Adj. P-value=0,0000, Chi-square=914,3273, df=2)
Cuenca, Quito
Node 163
Malo 21,83 3268
Bueno 78,17 11700
Guayaquil, Ibarra, Riobamba
Node 164
Malo 33,63 7186
Bueno 66,37 14183
Ambato, Latacunga
Node 165
Malo 11,79 265
Bueno 88,21 1982
Universitario
Node 157
Malo 19,68 1030
Bueno 80,32 4205
Género (Adj. P-value=0,0000, Chi-square=49,5335, df=1)
Masculino
Node 161
Malo 23,58 611
Bueno 76,42 1980
Femenino
Node 162
Malo 15,85 419
Bueno 84,15 2225


Node 0
Malo 26,81 11749
Bueno 73,19 32070
IESS (Adj. P-value=0,0000, Chi-square=172,1843, df=1)
Sí
Node 33
Malo 25,62 9489
Bueno 74,38 27543
Tiempo trabajo (Adj. P-value=0,0000, Chi-square=354,7220, df=2)
<1
Node 48
Malo 31,67 994
Bueno 68,33 2145
1-3
Node 49
Malo 27,9 6322
Bueno 72,1 16335
>3
Node 50
Malo 19,34 2173
Bueno 80,66 9063
NO
Node 34
Malo 33,3 2260
Bueno 66,7 4527
Tiempo trabajo (Adj. P-value=0,0000, Chi-square=115,3848, df=2)
<1
Node 51
Malo 50,16 305
Bueno 49,84 303
1-3
Node 52
Malo 34,43 1232
Bueno 65,57 2346
>3
Node 53
Malo 27,8 723
Bueno 72,2 1878


Node 0
Malo 26,81 11749
Bueno 73,19 32070
Vivienda (Adj. P-value=0,0000, Chi-square=59,3888, df=2)
Arrendada
Node 83
Malo 28,71 3701
Bueno 71,29 9192
Tiempo trabajo (Adj. P-value=0,0000, Chi-square=118,9829, df=2)
<1
Node 90
Malo 36,16 508
Bueno 63,84 897
1-3
Node 91
Malo 30,25 2404
Bueno 69,75 5542
>3
Node 92
Malo 22,28 789
Bueno 77,72 2753
Propia
Node 84
Malo 28,06 2467
Bueno 71,94 6324
Tiempo trabajo (Adj. P-value=0,0000, Chi-square=202,0303, df=2)
<1
Node 97
Malo 42,89 193
Bueno 57,11 257
1-3
Node 98
Malo 33,23 1401
Bueno 66,77 2815
>3
Node 99
Malo 21,16 873
Bueno 78,84 3252
Otra
Node 85
Malo 25,21 5581
Bueno 74,79 16554
Tiempo trabajo (Adj. P-value=0,0000, Chi-square=145,1300, df=2)
<1
Node 111
Malo 31,61 598
Bueno 68,39 1294
1-3
Node 112
Malo 26,64 3749
Bueno 73,36 10324
<3
Node 113
Malo 20 1234
Bueno 80 4936

Node 0			
	Malo	26,81	11749
	Bueno	73,19	32070

Rango edad (Adj. P-value=0,0000, Chi-square=94,4688, df=3)

<25	Node 177	Malo	28,62	3609
		Bueno	71,38	9000

Actividad económica (Adj. P-value=0,0000, Chi-square=75,4505, df=2)

Empleado	Node 181	Malo	28,22	3505
		Bueno	71,78	8915
Comerciante, estudiante	Node 182	Malo	41,18	28
		Bueno	58,82	40
Independiente	Node 183	Malo	62,81	76
		Bueno	37,19	45

26-40	Node 178	Malo	27,04	6631
		Bueno	72,96	17888

Actividad económica (Adj. P-value=0,0000, Chi-square=157,3984, df=2)

Empleado	Node 190	Malo	26,54	6380
		Bueno	73,46	17655
Comerciante, estudiante	Node 191	Malo	48,09	113
		Bueno	51,91	122
Independiente	Node 192	Malo	55,42	138
		Bueno	44,58	111

41-45	Node 179	Malo	24,41	785
		Bueno	75,59	2431

Actividad económica (Adj. P-value=0,0097, Chi-square=9,2680, df=2)

Empleado	Node 193	Malo	24,05	752
		Bueno	75,95	2375
Comerciante, estudiante	Node 194	Malo	31,82	14
		Bueno	68,18	30
Independiente	Node 195	Malo	42,22	19
		Bueno	57,78	26

>45	Node 180	Malo	20,83	724
		Bueno	79,17	2751

Actividad económica (Adj. P-value=0,0021, Chi-square=12,3254, df=2)

Empleado	Node 196	Malo	20,49	696
		Bueno	79,51	2701
Comerciante, estudiante	Node 197	Malo	30,95	13
		Bueno	69,05	29
Independiente	Node 198	Malo	41,67	15
		Bueno	58,33	21