Caracteristicas principales de un Boreau de Créditos y el uso de los modelos de regresiön

Adán Noel Duarte Candia

Universidad Nacional de Asunciön

Facultad de Ciencias Exáctas y Naturales

Departamento de Matemática

San Lorenzo - Paraguay

Setiembre - 2017

Índice

Objetivo General

- Analizar las características principales de un Boreau de Créditos y el uso de los modelos de regresión.

Objetivos Específicos

- Conceptualizar el boreau de crédito y el credit scoring.
- Definir un modelo de score.
- Citar los distintos tipos de modelos de score según la aplicación de los mismos.
- Describir las metodologías estadísticas utilizadas para obtener un modelo de Score.
- Bosquejar los tipos de variables utilizadas en un modelo de Score.
- Anunciar las aplicaciones de los modelo de score.
- Mostrar la metodológica del modelo proibit y logit.

Introducción

Un Boreau de Credito es una empresa que en general se constituye como una Sociedad de Información Crediticia orientada a integrar información sobre el comportamiento crediticio de las personas que solicitan un crédito y funciona en base a los reportes de las empresas que otorgan un crédito. Es el punto de referencia de hoy en día para el otorgamiento de créditos a nivel mundial. Las más antigüas del mundo son *Equifax*, *TransUnión* y *Experian*. Con toda la información que dispone un Boreau de Crédito, el análisis que las entidades que otorgan crédito pueden realizar es objetiva e imparcial, ademas es automatizable, parametrizable y permite a los que otorgan un crédito tomar decisiones en relación al riesgo de crédito que quieren tomar con el total conocimiento del perfil del cliente que ha solicitado el financiamiento. Esto también es muy cierto en el ramo de personas morales, donde además se puede observar, por ejemplo, si el cliente potencial tiene compromisos como obligado solidario de algún crédito y si este tiene accionistas que el otorgante de crédito también necesite conocer para tomar una decisión bien informada. Una de las herramientas que se genera mediante la cantidad de información que dispone un Boreau de Crédito es el Score de riesgo crediticio. El score de riesgo o credit coring fueron introducidas en los años 70, y el uso de técnicas de este estilo se pudo generalizar recién en los años 90 gracias al desarrollo de mejores recursos estadísticos y principalmente recursos computacionales. Hoy en día prácticamente todas las entidades financieras y comerciales emplean un score al menos para originar sus financiaciones. Dada su importancia en el proceso de gestión del crédito, en esta monografia se clarifican algunos aspectos asociados a los Boreau de Créditos: qué son?, cuáles son las principales a nivel mundial?, cuáles son a nivel local?, como así también lo relacionado a los modelos de Score: qué son?, qué técnicas se pueden usar para construirlos? y cuáles son más convenientes?, qué tipos de variables emplean?, qué aplicaciones se han desarrollado a partir de ellos y, sobre todo, cómo funcionan y como deben interpretarse sus resultados.

1. El Boreau de Credito y el Credit Scoring

El Boreau de Credito es una empresa que en general se constituye como una Sociedad de Información Crediticia orientada a integrar información sobre el comportamiento crediticio de las personas que solicitan un crédito y funciona en base a los reportes de las empresas que otorgan un crédito.Es el punto de referencia de hoy en día para el otorgamiento de créditos a nivel mundial, entre las principales empresas que funcionan como boreau de créditos aparecen Equifax que es una multinacional que opera desde 1899 en el mundo de la información comercial y de crédito, servicios de marketing, soluciones a medida y sistemas de información. Actualmente Equifax opera en 24 países y alrededor de unos 50 países utilizan la información que proporciona la empresa. Desde el 2013 Equifax hace presencia en Paraguay a partir de la adquisición de *Informconf*[1]. La siguiente es TransUnión que opera desde 1968 en el mundo de la información comercial y de crédito y tiene presencia en mas de 30 países. Otra empresa de Boreau es Experian fundada en Nottingham (Inglaterra) en 1980 para proporcionar servicios de información crediticia como soporte al negocio de venta por correo del Grupo GUS, actualmente se encuentra actuando en más de 60 países.

El Boreau de Crédito no pone una calificación a las personas, simplemente recogen los reportes y los agrupan bajo tu nombre, ya armado el reporte lo pone a disposición de cualquier entidad financiera o comercial que lo solicite.

Toda persona o empresa puede solicitar su Reporte de Crédito a través de las diversas opciones que el Boreau de Crédito pone a disposición para tal fin.

El Reporte de Crédito es un documento que muestra el historial crediticio de una persona, detallando información de los créditos que tiene o ha tenido, la fecha de apertura, saldo, etc. y la forma como estos han sido pagados en el tiempo; para el caso de *Informconf* presentando la cédula vigente en cualquiera de sus oficinas dicho reporte es proporcionado sin costo alguno.

El Reporte de Crédito que brinda *Informconf* a las personas o empresas especifica el nombre de las entidades de crédito que reportan la información crediticia y de aquellos que han consultado el historial durante los últimos 36 meses en relación con la fecha de emisión.

El Boreau de Crédito no pone incumplimientos comerciales en los reportes, pero si puede quitarlos; si se hace un reclamo al Boreau, ellos le generan un pedido de información a la entidad que te ha reportado con el supuesto incumplimiento, si esta no responde en un plazo considerado(entre una semana o un mes) se entiende como aceptada tu reclamo y el Boreau te la quita.

La utilización de modelos de Credit Scoring o simplemente Score para la evaluación del riesgo de crédito, es decir, para estimar cual ha de ser la probabilidad de que una persona que obtiene un crédito termine sin pagarlo y ordenar a los clietes y solicitantes de financiamiento en función de su riesgo de incumplimiento, comenzó en los años 70 pero se generalizó recién a partir de los años 90 gracias al desarrollo de mejores recursos estadísticos y principalmente recursos computacionales. Esto se ha debido también a la creciente necesidad por parte de la industria bancaria de hacer más eficaz y eficiente la originación de créditos, y de tener una mejor evaluación del riesgo de sus clientes. Estos modelos generalmente se asocian a lo que se llaman data mining (minería de datos), que son todos aquellos procedimientos que permiten extraer información útil y encontrar patrones de comportamiento en los datos. Por este motivo, la minería de datos es un conjunto de técnicas con origen diverso, pero en general con raíz estadístico matemática computacional.

[1] Otro boreau de crédito que también opera en paraguay es *Criterion* aunque el más popular es *Informconf*

A pesar de la proliferación de los modelos de Score, el juicio del hombre (o juicio del analista) continúa siendo utilizado en la originación de un créditos, en algunos casos expresado como un conjunto de reglas que la entidad aplica de manera sistemática para filtrar solicitudes de créditos. De hecho, en la práctica ambas metodologías coexisten y se complementan, definiendo sistemas híbridos. En general un Score bajo determina la denegación de una solicitud de crédito, mientras que un Score por encima del mínimo exigido por la entidad financiera dispara análisis posteriores con los que la evaluación continúa.

Un aspecto muy importante, pero que por su extensión merece un tratamiento aparte y no se discute aquí, es el de la validación del score. La validación es un proceso por el cual la entidad que opera a crédito, revisa y evalúa de manera periódica, diversos aspectos del modelo, como su diseño, las variables empleadas, la calidad de los datos y otros aspectos cualitativos, como así también su eficacia para ordenar en función del riesgo (poder discriminatorio), la precisión en sus estimaciones de la tasa de mora (calibración), etc.

Para aclarar aspectos del desarrollo y la utilización de los modelos de Score, construimos un ejemplo y mostramos su funcionamiento. Este modelo debe tomarse solamente como un ejemplo con fines didácticos orientado a ilustrar algunas características básicas de los modelos de Score, que de ninguna manera busca establecer o señalar cuáles son las mejores prácticas ni lineamientos que se debe seguir para su construcción.

Las empresas de información crediticia, conocidas como Boreau de Crédito, además de tener información detallada de los deudores del sistema financiero, reciben información de otras fuentes, como por ejemplo los juzgados comerciales y las tiendas minoristas. Con este set de información brindan, además de informes comerciales, servicios de Score, de detección de fraudes y robo de identidad, etc.

2. Modelo de Score

Los métodos o modelos de credit scoring, a veces denominados clasificadores o simplemente Score, son fórmulas matemáticas que de manera automática evalúan el riesgo de crédito de un solicitante de financiamiento o de alguien que ya es cliente de la entidad. El más común tiene una dimensión individual, ya que se enfocan en el riesgo de incumplimiento del cliente, independientemente de lo que ocurra con el resto de la cartera de préstamos. Este es uno de los aspectos en los que se diferencian de las otras herramientas que miden el riesgo de crédito, como son los modelos de cartera y los VaR marginales; estos tienen en cuenta la correlación de la calidad crediticia de los deudores de una cartera de préstamos.

En un primer acercamiento a los modelos de score, se los puede definir como el conjunto de métodos estadísticos utilizados para clasificar a los solicitantes de crédito, o incluso a quienes ya son clientes de la entidad, entre las clases de riesgo *Bueno* y *Malo* [Hand and Henley, 1997]. Anteriormente en los años 70 se basaban en técnicas estadísticas (principalmente en el análisis discriminante), en la actualidad también están basados en técnicas matemáticas, econométricas y de inteligencia artificial. En cualquiera de estos caso, los modelos de credit scoring emplean principalmente la información de la persona contenida en las solicitudes de crédito y/o en fuentes internas y/o externas de información.

El resultado de la evaluación se refleja en la asignación de alguna medida (en general un número de 0 a 1000) que permita comparar y ordenar a las personas que solicitan un crédito en función de su riesgo crediticio y a la vez cuantificarlo. Por lo general, los modelos de Score le asignan a la persona un puntaje, o una calificación, o rating. Algunos métodos los asignan a grupos de personas, en donde cada grupo tiene un perfil de riesgo diferente; sin embargo, en la práctica esto equivale a una calificación (La faja de Informconf es una clasificación en letras, estas letras pueden ser cualquiera de entre $\{A, B, C, D, E, F, G, H, I, J, K, L, M, N, X\}$

donde A es la mejor clasificación y X la peor). A su vez, estos ordenamientos de los clientes permiten obtener estimaciones más concretas del riesgo; lo que se busca con esto es obtener alguna estimación de la probabilidad de incumplimiento asociada a su score, rating o calificación. Esta estimación se puede obtener directamente del score cuando se usan modelos econométricos, o también en función de la tasa de incumplimiento histórica observada en el grupo de clientes con la misma calificación o con score similar.

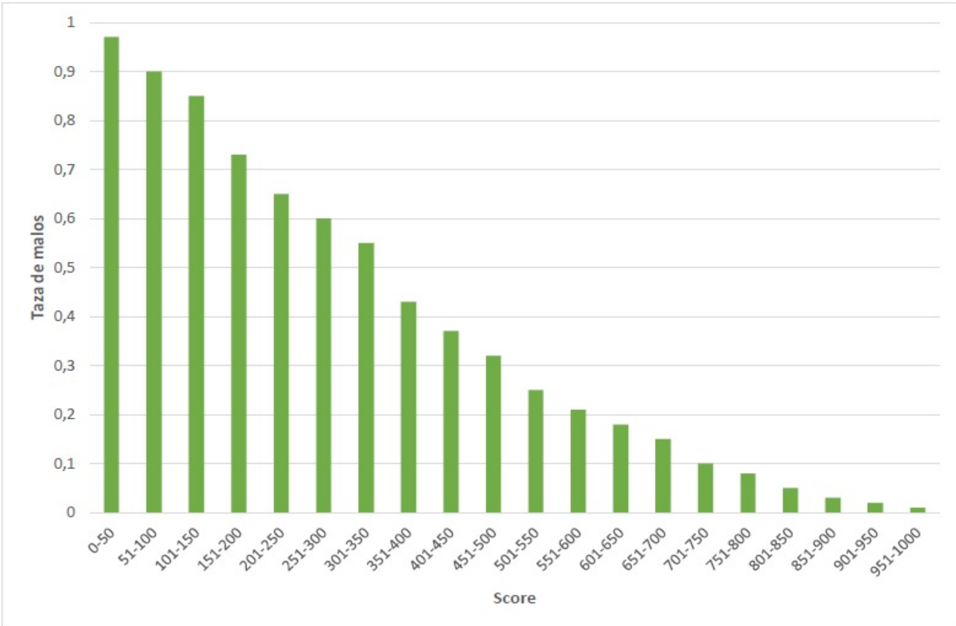

Figura 1: Score y Cuantificación del Riesgo

La figura 1. muestra un ejemplo de la salida de un modelo de credit scoring, que muestra la tasa de incumplimiento histórica asociada a cada rango del score. La relación entre la tasa de incumplimientos y el score se muestra para intervalos del puntaje, ya que es una variable continua, y se observa que el riesgo cae de manera exponencial a medida que el score aumenta. Esto es lo que se espera de las técnicas de credit scoring y sistemas de rating: a mayor score menor riesgo.

Aunque en el ejemplo de la figura 1. la escala del score oscila entre 0 y 1.000, la misma es arbitraria y depende en última instancia de la construcción del modelo (podríamos tener un score que oscila entre 1 y 100 por ejemplo). También podría concebirse un modelo en el cual el riesgo baja a medida que baja el score (al revés que en la figura 1.), pero en la práctica predominan aquellos que presentan una relación inversa entre el score y el riesgo.

2.1. Otros tipos de Score

Otros tipos de Score muy utilizados son los Score de Cobranza y los Score de propensión de compra que a diferencia del Score de Riesgo está pensado para ser usado masivamente, es decir, se calcula a toda la cartera de la entidad.

El score de cobranza estima la probabilidad de que una persona que tiene una deuda en alguna entidad pague la misma en un futuro cercano (generalmente que el pago se realice en los siguientes 3 meses). Al calcular el score de cobranza a todas las personas morosas, esto permite ordenar a dichas personas según la probabilidad de cobro, de esta manera la entidad puede establecer diferentes estrategias de gestión de la cobranza, podría por ejemplo la entidad enviar un mensaje de texto como recordatorio de la deuda a todos los deudores con altas probabilidades de cobro y hacer una llamada telefónica al resto, de esta manera la entidad logra reducir los costos que implican la gestión ya que en general un mensaje de texto es más barato que una llamada telefónica.

El Score de propensión de compra estima la probabilidad de que una persona cliente de cierta entidad compre un nuevo producto o mejore un producto que ya posee. El score de propensión de compra es utilizado para campañas de Marketing, al igual que el Score de Cobranza, el de propensión se calcula a todos los clientes de la entidad a quienes se tiene intención de ofrecer un nuevo producto (cross selling) o mejorar un producto existente (up selling) ordenando la población según la probabilidad de compra, de esta manera la entidad será más eficiente al elegir a los clientes con más probabilidad de compra para ofrecer un nuevo producto o mejorar el producto que ya posee; por ejemplo, si tiene una tarjeta de crédito puede ofrecerle otra tarjeta de crédito (cross selling) o bien puede ofrecerle aumentar su línea de crédito (up selling); de esta manera la entidad obtendrá mejores resultados en su campaña de Marketing. (Si no se tuviera el Score de propensión se debería ofrecer a todos los clientes una mejora o un nuevo producto y esto implica más costos para la entidad). El score de propensión generalmente se utiliza en presencia de un score de riego; lo habitual es filtrar a traves del score de riesgo a los mejores clientes y a estos calcularles el score de propensión; de esta manera la entidad evita hacer ofertas a clientes con altos riesgos ya que estos son siempre más propensos de aceptar cualquier tipo de oferta.

2.2. Técnicas empleadas

Para evaluar el riesgo crediticio o la conveniencia de otorgar un crédito, hay una gran variedad de metodologías que están a nuestra disposición (si quisiéramos comparar enfoques alternativos podemos ver [Srinivasan and Kim, 1987], [Mester et al., 1997], [Hand and Henley, 1997] y [Thomas, 2000]): regresión lineal, regresión logística, análisis discriminante, modelos probit, modelos logit, modelos basados en cadenas de Markov, métodos no paramétricos de suavizado, métodos de programación matemática, algoritmos de particionamiento recursivo [2] (árboles de decisión), sistemas expertos, algoritmos genéticos, redes neuronales y, finalmente, el juicio del hombre, es decir, la decisión de un analista acerca de otorgar o no un crédito. En principio el juicio humano presenta la ventaja de ser más eficaz en tratar las excepciones relacionadas con las experiencias pasadas, pero los métodos de Score son más eficientes a la vez que sus predicciones

[2] Los métodos como el análisis discriminante, la regresión lineal y logística y los modelos probit pueden ser considerados como métodos de particionamiento simultáneo, ya que consideran a todas las variables explicativas

más objetivas y consistentes, por lo que pueden calcularse sobre una gran cantidad de solicitudes de crédito en poco tiempo y a un bajo costo para tomar la mejor decisión.

La literatura nos sugiere que todos los métodos de Score arrojan en general resultados similares, por lo que la conveniencia de usar uno u otro método depende de las características particulares del caso. Dentro de los modelos de tipo econométricos, los modelos de probabilidad lineal han caído en desuso por sus desventajas técnicas, en tanto que los modelos de tipo probit, logit y la regresión logística son de usos superiores al análisis discriminante ya que proveen para cada cliente una probabilidad de incumplimiento, en tanto que este sólo clasifica a los clientes en grupos de riesgo. A pesar de que los probit, logit y la regresión logística son, teóricamente, herramientas econométricas más apropiadas que la regresión lineal, ésta arroja estimaciones muy similares a las de los anteriores cuando sus probabilidades estimadas se ubican entre el 20 % y el 80 %. Los modelos no paramétricos y los de inteligencia artificial(como los árboles de clasificación o decisión, las redes neuronales y los algoritmos genéticos) son superiores a los modelos estadísticos cuando se desconoce la probable forma de la relación funcional entre las variables disponibles y la variable a predecir, y cuando se presume que esta relación no es lineal.

Si queremos desarrollar árboles [Orallo et al., 2004], tres algoritmos frecuentemente empleados para construir (entrenar) árboles son ID3, C4.5 y C5: en todos los casos se buscan cuál es la partición óptima de la muestra tal que, dada la variable a predecir (que podría ser el incumplimiento, el cobro o la compra), los distintos grupos o particiones presentan distintos perfiles de riesgo. Los árboles tienen la ventaja de que no requieren la formulación de supuestos estadísticos sobre las distribuciones estadísticas o formas funcionales de las variables. A su vez, presentan la relación entre las variables, los grupos y el riesgo de manera visual, con lo cual (y esto de cara al que va usar el árbol) si el conjunto de variables incluidos en el análisis es reducido, facilita la comprensión de cómo funciona el score.

Los métodos como las redes neuronales y los algoritmos genéticos [Marín and Sandoval, 1995], a pesar de las ventajas que ya mencionamos, son poco intuitivos y en general de difícil implementación. Los modelos que utilizan programación matemática permiten diseñar clasificadores [3] mejor adaptadas a las necesidades de cada entidad crediticia y manejar una gran cantidad de variables; estas se basan en optimizar un criterio objetivo, como por ejemplo la proporción de solicitantes bien clasificados.

Por último están los sistemas expertos que tienen como atractivo la capacidad para justificar sus recomendaciones y decisiones, lo cual puede ser muy importante por cuestiones legales relacionadas al acceso al crédito. En [Srinivasan and Kim, 1987] se compara diversas técnicas y se descubre que los árboles de decisión superan a las regresiones logísticas, mientras que la regresión logística arroja mejores resultados que el análisis discriminante. De hecho, esta superioridad de los árboles se cree está en relación directa a la complejidad de los datos bajo estudio.

El objetivo de cualquier entidad que otorga un crédito es maximizar los beneficios derivados de la intermediación crediticia, lo cual no necesariamente esta relacionado con el riesgo. Es decir, que un solicitante de crédito presente cierto riesgo no necesariamente implica que no conviene otorgarle el crédito que está solicitando. Probablemente un cliente de una entidad que hace una financiación con tarjeta de crédito y que es relativamente riesgoso, es más rentable que uno que no tiene riesgo pero que nunca hizo una financiación con la tarjeta. Por lo tanto, cuando se deba determinar qué solicitudes aceptar y cuáles de estas rechazar, la entidad tiene en cuenta los beneficios que espera de las solicitudes de distinto tipo de riesgo. Por ejemplo, en [Srinivasan and Kim, 1987] se analizan el problema de una empresa comercial que necesita determinar el límite crediticio óptimo para cada uno de sus cliente. Para poder estimarlo, resuelven un problema dinámico

[3] Planillas o programas para asignar un puntaje o rating.

que incluye la evaluación de riesgo del cliente con los potenciales beneficios que de él se pueden derivar y muestran los resultados para distintos métodos de Score. Éstos proveerán distintas estimaciones de riesgo que, metidas en el programa dinámico, genera la posibilidad de obtener estimaciones del límite de crédito óptimo para cada cliente.

Entre todas las metodologías disponibles más utilizadas están, los modelos probit, junto con las regresiones lineal y logística, el análisis discriminante y los árboles de decisión. En [Boyes et al., 1989] y [Greene, 1992] se puede ver que utilizan un probit bivariado para evaluar solicitudes de tarjeta de crédito, teniendo en cuenta no sólo la probabilidad de incumplimiento del deudor, sino también el beneficio que la entidad espera de la utilización de la tarjeta por parte del solicitante. En [Gordy, 2000] se comparar modelos de cartera de riesgo crediticio, utilizando modelos probit para estimar la probabilidad de incumplimiento de cada exposición en la cartera. En [Cheung, 1996] y [Nickell et al., 2001] se pueden ver que utilizan modelos probit ordenados, de los cuales los probit bivariados son un caso particular, para estimar la probable calificación de títulos públicos, en tanto que [Falkenstein et al., 2000] hace una aplicación similar pero para deuda privada.

Aunque los métodos señalados en el párrafo anterior son los que más se utiliza, frecuentemente se emplean de manera combinada. En primer lugar, como se mencionó en la introducción, en general en el sistema financiero estos modelos no se usan de manera mandatoria para aceptar una solicitud de crédito, sino que sus resultados se combinan con revisiones posteriores. En otros casos, previo al cálculo del score se aplican filtros que acotan el universo de solicitantes de créditos a ser evaluados con estos modelos. En ocasiones se combinan diversas metodologías, como por ejemplo en los árboles de regresión: a través de un árbol se particiona la muestra de clientes y luego a los clientes de cada partición se les calcula una regresión logística o modelo probit con distintas características.

2.3. Tipo de variables empleadas

Dependiendo de la aplicación que se haga del score, el tipo de variables utilizadas puede variar significativamente según se trate de modelos para la cartera retail (individuos o pequeñas y medianas empesas), donde generalmente se usan variables socioeconómicas o datos básicos del emprendimiento productivo; o de grandes empresas. En este caso, generalmente se utilizan variables que se extraen de los estados contables, información cualitativa relacionada con la dirección, el sector económico al que corresponde el emprendimiento, proyecciones del flujo de fondos, etc. Para hacer credit scoring para corporaciones, RiskCalc^{TM} de Moody's (ver [Falkenstein et al., 2000]) ha utilizado: activos/IPC, inventarios/costo de mercaderías vendidas, pasivos/activos, crecimiento de los ingresos netos, ingresos netos/activos, prueba ácida, ganancias retenidas/activos, crecimiento en las ventas, efectivo/activos y ratio de cobertura del servicio de la deuda. También se señala que:

- las variables con mayor poder predictivo son ganancias, apalancamiento, tamaño de la empresa y liquidez; y
- si bien la teoría recomienda utilizar ratios de apalancamiento y rentabilidad en un modelo de score, la experiencia sugiere usar ratios de liquidez.

En [Srinivasan and Kim, 1987] se introduce el funcionamiento de distintos modelos para deudas de corporaciones usando activo corriente/pasivo corriente, prueba ácida, patrimonio neto/deuda, logaritmo de los activos, ingresos netos/ventas, ingresos netos/activos. Finalmente, el Z-score que se observa en [Altman, 1968] utiliza: capital de trabajo/activos, ganancias retenidas/activos, EBIT/activos, valor de mercado del patrimonio neto/valor libros de la deuda y ventas/activos.

Dentro de los modelos desarrollados para deudas retail (individuos o pequeñas y medianas empesas), en [Boyes et al., 1989] y [Greene, 1992] se utilizan variables socioeconómicas: edad, estado civil, el tiempo de permanencia en el domicilio actual y en el empleo actual, cantidad de personas que tiene a cargo, el nivel educativo , el tipo de ocupación, si es propietario de la vivienda que habita, gastos mensuales promedio/ingresos mensuales promedio, si tiene o no tarjeta de crédito, cuenta corriente o caja de ahorro, y por supuesto el número de consultas en los bureaus y cómo está calificado en ellos. Dentro de los modelos utilizados en la industria, Fair Isaac Corporation desarrolló uno que es empleado por los tres mayores boreaus de crédito de Estados Unidos de Norteamérica para calcular sus scores. Se trata del FICO credit risk score, que es empleado por los burós Equifax, Experian y Transunion para calcular sus scores: Beacon, Experian/Fair Isaac Risk Model y FICO Risk Score/Classic respectivamente. Estos tres scores tienen una amplia difusión que se utiliza para evaluar solicitudes de crédito y las mismas varían entre un mínimo de 300 puntos y un máximo de 850. Aunque los tres emplean el mismo modelo, una misma persona puede tener distintos score si su información difiere en cada uno de los dichos boreaus de crédito.

El FICO credit risk score utiliza principalmente variables relacionadas al comportamiento de los pagos actual y pasado, y proyecta la idea de que el comportamiento pasado es el mejor predictor que se puede usar para estimar el comportamiento futuro. Los grupos de variables usadas, junto con su incidencia en el score, son: historia de pagos (35 %), largo de historia crediticia (15 %), nuevo crédito (10 %) y tipo de crédito usado (10 %) como así también el monto adeudado (30 %). A diferencia de las aplicaciones con tinte académico, y por motivos legales[4], no utiliza variables como edad, raza, sexo, religión, nacionalidad y estado civil. La ocupación y antigüedad en el empleo, el domicilio, los ingresos, la tasa de interés y el número de consultas realizadas al boreau sobre la información del cliente, por entidades financieras para ofrecer productos preaprobados o para monitoreos. Sin embargo, la cantidad de consultas realizadas en respuesta a solicitudes de crédito sí influye en el score.

La información que se emplea para hacer un modelo de score para el portafolio retail usualmente se clasifica en positiva y negativa. La información negativa es aquella asociada a los incumplimientos y atrasos en los pagos, mientras que la positiva es la información de los pagos a término y otra información descriptiva de las deudas, como montos de préstamos, tasas de interés y plazo de las financiaciones. La evidencia empírica muestra que la inclusión de la información relacionada al buen comportamiento de pagos del cliente mejora sustancialmente el funcionamiento de estos modelos de score. Por ejemplo, con datos de Argentina, Brasil y México, [Powell et al., 2004] muestra la cuantificación de la mejora en el poder predictivo de estos modelos al incluir la información de tipo positiva respecto a modelos que sólo usan información de tipo negativa, y muestra a demas que su utilización por parte de los que otorgan un crédito facilita bastante el acceso al crédito y mejora la calidad de los portafolios de préstamos de las entidades financieras.

Por último, los modelos de score para microemprendimientos y pequeñas y medianas empresas tienden a combinar información personal del titular del emprendimiento con la información del negocio. Uno de los primeros desarrollos fue el Small Business Scoring Solution que Fair Isaac Corporation introdujo en 1995, en el cual se combinó información de los principales dueños de la empresa y la información del negocio mismo. Dentro de los desarrollos académicos más resaltantes tenemos que en [Miller and Rojas, 2004] se hacen credit scoring de pequeñas y medianas empresas de Brasil, México y Colombia, mientras que en [Milena et al., 2005] se hacen lo mismo para microfinancieras de Nicaragua.

[4] La Consumer Credit Protection Act prohíbe que el credit scoring utilice esta información.

2.4. Aplicaciones

Tanto en el ámbito teórico como en la práctica de la industria crediticia, los modelos de score se pueden emplear para evaluar la calidad crediticia de clientes de todo tamaño: retail (individuos y, pequeñas y medianas empresas) y corporaciones. Sin embargo, la práctica común es evaluar el portafolio retail, mientras que los clientes de corporaciones se evalúan con sistemas de rating. Además de las diferencias en el tipo de variables empleadas para uno y otro cliente, la evaluación de grandes empresas implica revisar aspectos cualitativos de difícil estandarización, por lo tanto, el resultado se expresa como una calificación y no como un score. De todos modos, en [Jennings, 2001] se discute las ventajas de su aplicación en pequeñas y medianas empresas. En el resto del documento se analizan modelos diseñados para la banca minorista exclusivamente.

Las entidades pueden emplear estos modelos en la originación de créditos, es decir, para resolver si se rechaza o continúa el proceso las solicitudes de crédito. En este caso se trata de modelos reactivos. También se emplean para administrar el portafolio de créditos, en cuyo caso se trata de modelos de seguimiento, proactivos o de comportamiento, y se pueden emplear para: administrar límites de tarjetas y cuentas corrientes, analizar la rentabilidad de los clientes, ofrecer nuevos productos o mejorar productos existentes, monitorear el riesgo y detectar posibles problemas de cobranza, entre otras aplicaciones. En el caso de los modelos reactivos, las entidades financieras generalmente determinan un punto de corte para determinar qué solicitudes se aceptan (por tener un puntaje mayor o igual al punto de corte) y cuales no. La fijación del mismo no responde a consideraciones de riesgo necesariamente sino que depende de la tasa de beneficios que desea la entidad y su apetito por el riesgo. A demás, para la misma rentabilidad deseada, una entidad con una mejor gestión de recuperos o un mejor sistema de administración de límites o de alertas tempranas, podría trabajar con menor punto de corte ya que se compensa el riesgo con una menor exposición al mismo. La relación entre la política de crédito de un banco y su manejo del score se esquematiza en la Figura 2.

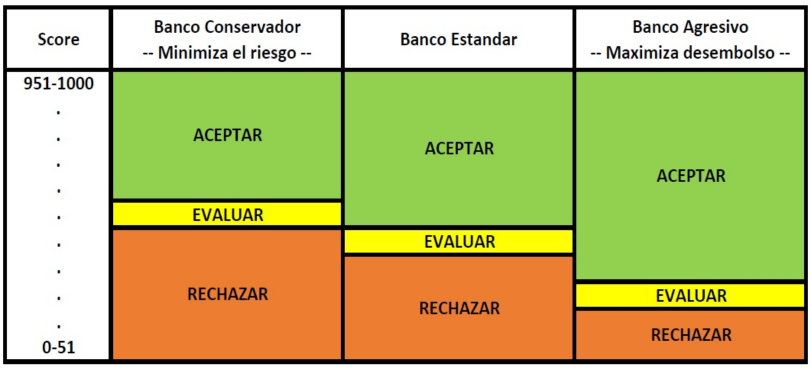

Figura 2: Fijación del punto de corte y política de crédito

En la práctica, sin embargo, la aplicación de un score no es tan directa. Los scores se emplean, en general, de manera mandatoria pero en forma asimétrica: los solicitantes de crédito con un score inferior al punto de corte son rechazados, mientras que aquellos con valores superiores pasan a etapas posteriores de análisis previo al otorgamiento de la financiación pero que también podrían terminar en rechazo.

Por otro lado, en la mayoría de los casos se trata de modelo de score genéricos, es decir, que evalúan la capacidad de repago de un solicitante de crédito sin tener en cuenta las características de la financiación solicitada. Sin embargo, existen muchos desarrollos que apuntan a una mayor precisión en los resultados y están diseñados para solicitantes de un tipo específico de financiación, como ser prendarios o hipotecarios para la vivienda. La dimensión del tipo de financiación que se solicita es de relevante importancia, ya que distintos clientes con idéntica capacidad de repago e historial crediticio probablemente muestren distintos patrones de pago según el tipo, plazo de la financiación solicitada y el sector donde se esté haciendo esta solicitud. Por último, los modelos pueden ser desarrollados con datos externos de la entidad, como es el caso de los scores de boreaus, con datos internos de la entidad o con ambos tipos de datos (internos y externos).

2.5. Un modelo de Score para banca minorista

2.5.1. Descripción de los datos

En general un Boreau puede administrar una base de datos que concentra, mes a mes, millones de datos sobre el grado de cumplimiento en los pagos y la calidad crediticia de todos los clientes del sistema financiero con deudas consolidadas superiores a un monto mínimo (en general suele oscilar alrededor de 50$[5]) con alguna entidad crediticia. Además posee información tales como: datos personales, sector (sistema financiero, comercios, telefónicas, etc.), tipo de deudor (consumo o vivienda, comercial asimilable a consumo y comercial), actividad económica, clasificación de riesgo, entidad acreedora, deuda, tipo de financiación y cobertura con garantías preferidas, entre otras. Por ejemplo, según las normas del Banco Central del Paraguay, todos los deudores del sistema financiero deben ser clasificados en una escala de 1 a 6 en función de su riesgo de crédito, para esto el principio básico es la capacidad de pago futura que tiene cada deudor de sus obligaciones con la entidad. Para la Central de Riesgos del Banco Central del Paraguay, 1 es la mejor calificación y 6, la peor calificación. Es considerado moroso quien ingresa en la categoría 2, cuando existe un retraso en el pago de la deuda de más de 60 días. Desde el día 91 al 120 de retraso, la categoría es 3; del 121 al 180, se ubica la categoría 4; mientras que desde el día 181 al 270, se encuentra la categoría 5, y de 360 y más días de retraso, está la categoría 6, que es la menos favorable.

Cuando un deudor tiene un retraso de entre 1 y 60 días, el cliente no se entera porque las entidades financieras solamente constituyen previsiones sobre esos clientes e ingresan en una subcategoría 1A o 1B.

2.6. Metodología empleada

Cuando al plantear un modelo la variable dependiente toma valores discretos, se emplean modelos de regresión discreta. El caso más simple se da cuando la variable a predecir es binaria, esto es los valores 0 o 1, y se puede estimar con distintos enfoques como el análisis discriminante, el modelo de probabilidad lineal, los modelos de tipo probit y logit o con una regresión logística.

Sea Y una variable aleatoria binaria que toma el valor 1 si ocurre el evento (el deudor paga sus deudas normalmente) y 0 si entra en mora, se cuenta con una muestra aleatoria de n observaciones, Y_i $i : 1, \dots, n$, y se define como Ψ_i al conjunto de información relevante asociado con el individuo i, que se utilizará para explicar a Y_i.

[5]Para el caso de *Informconf* se estable que sea alrededor de 10$

Un modelo de respuesta binaria es un modelo de la probabilidad de ocurrencia del evento Y_i condicional en el conjunto de información Ψ_i :

$$P_i = Pr(Y_i = 1 \mid \Psi_i) \tag{1}$$

Dado que Y_i toma valores 0 y 1, la esperanza de Y_i condicional en Ψ_i es:

$$E(Y_i|\Psi_i) = 1 \times P_i + 0 \times (1 - P_i) = P_i = Pr(Y_i = 1 \mid \Psi_i) \tag{2}$$

Por lo tanto, cuando la variable a explicar es binaria, su esperanza condicional es la probabilidad condicional de ocurrencia del evento.

2.6.1. El modelo de probabilidad lineal

Supóngase que Ψ_i está constituido por un vector columna $\boldsymbol{X_i}$ compuesto por k variables explicativas, incluyendo a la ordenada al origen, $\boldsymbol{\beta}$ es un vector columna que contiene los parámetros correspondientes a las variables explicativas, y que se intenta modelar a la variable Y_i a través del modelo de probabilidad lineal, según la siguiente relación:

$$Y_i = \boldsymbol{\beta^T X_i} + \epsilon_i, \text{ donde } E(\epsilon_i \mid \boldsymbol{X_i}) = 0 \text{ y } E(\epsilon_i) = 0 \tag{3}$$

y usando la ecuación (2) tenemos

$$E(Y_i \mid \boldsymbol{X_i}) = P_i = \boldsymbol{\beta^T X_i} \tag{4}$$

El modelo de probabilidad lineal, como se observa en la ecuación (4), implica estimar un modelo lineal en los parámetros para Y_i. Los valores que se predicen **deberían en su mayoría** ubicarse en el intervalo $[0, 1]$, y de este modo ser interpretados como la probabilidad de que la variable a explicar tome alguno de estos valores.

Mientras que su estimación e interpretación es simple, su utilización no es sencilla por dos problemas en la metodología. Primeramente, como la esperanza condicionada de Y_i es igual a la probabilidad condicionada de ocurrencia del evento (de $Y_i = 1$), ella debe estar restringida al intervalo $[0, 1]$. Sin embargo, el modelo lineal no impone ninguna restricción sobre $\boldsymbol{\beta^T X_i}$, asumiendo indirectamente que la variable dependiente puede tomar cualquier valor. Es por esto que el modelo podría estimar probabilidades negativas o mayores que uno, lo cual carece de significado. A demás, el término de error de este modelo no es homocedástico, ya que la varianza condicional varía según las observaciones, por lo que las estimaciones de $\boldsymbol{\beta}$ no son eficientes.

Para resolver estos inconvenientes se utilizan modelos econométricos, que generalmente se estiman por máxima verosimilitud, que tienen en cuenta la naturaleza discreta de la variable dependiente: se trata de los modelos de respuesta binaria. Ellos utilizan funciones de distribución con el objetivo de limitar las probabilidades estimadas al intervalo $[0, 1]$: las más usadas son la función de probabilidad acumulada normal estándar y **la función logística**. Cuando se usa la normal estándar se trata de un modelo probit, y logit cuando se usa la función logística.

La función logística es la distribución acumulada de la distribución sech^2, la secante hiperbólica al cuadrado. La ventaja de utilizar esta distribución es que tiene una expresión sencilla. Salvo por esta diferencia, que con los recursos computacionales disponibles en la actualidad no es significativa, ambas distribuciones (la normal acumulada y la logística) difieren muy poco y sólo en las colas, teniendo la función logística colas levemente más gordas. Al comparar los resultados que se obtienen con ambas funciones, se debe tener en cuenta que la varianza de la normal estándar es 1, en tanto que la de la distribución sech^2 es $\pi^2/3$. Por tanto, para obtener coeficientes comparables se debe multiplicar a los coeficientes del modelo logit por $\pi/(\sqrt{3})$. Sin embargo, [Amemiya, 1981] propone que lo multipliquemos por 0,625.

2.6.2. El modelo probit y logit

El modelo probit y logit son una de las varias alternativas que se tienen para estimar modelos de respuesta binaria.La idea consiste en usar una función de transformación $F(x)$ que tenga las siguientes propiedades:

$$F(-\infty) = 0, \quad F(\infty) = 1, \quad \text{y } f(x) \equiv \frac{dF(x)}{dx} > 0 \tag{5}$$

$F(x)$ es una función monótona creciente que mapea de la línea real al intervalo $[0, 1]$. Varias funciones de distribución acumulada tienen estas propiedades: la normal, la logística, la de Cauchy y la de Burr, entre otras. Estas distintas alternativas para los modelos de respuesta binaria consisten en una función de transformación $F(x)$ que aplica a una función índice que depende de las variables explicativas del modelo y que tiene las propiedades de una función de regresión, que puede ser lineal o no lineal.

La siguiente es una especificación general para cualquiera de los modelos de respuesta binaria:

$$E\left(Y_i \mid \Psi_i\right) = F\left(h\left(\boldsymbol{X_i}\right)\right), \text{ donde } h \text{ es la función índice} \tag{6}$$

Si bien h puede ser cualquier tipo de función, en general se utiliza una especificación lineal:

$$E\left(Y_i \mid \Psi_i\right) = F\left(\boldsymbol{\beta^T X_i}\right) \tag{7}$$

por lo cual el modelo de respuesta binaria es simplemente una transformación no lineal de una regresión lineal, y si bien $\boldsymbol{\beta^T X_i}$ puede tomar cualquier valor sobre la línea real, $F(\boldsymbol{\beta^T X_i})$ está limitado al intervalo $[0, 1]$. En el modelo probit, la función de transformación $F(x)$ es la función de distribución acumulada normal estándar, y por definición satisface las condiciones impuestas en (5). En este caso, el modelo de respuesta binaria puede escribirse de la siguiente manera:

$$P_i = E\left(Y_i \mid \Psi_i\right) = F\left(\boldsymbol{\beta^T X_i}\right) = \Psi\left(\boldsymbol{\beta^T X_i}\right) \equiv \int_{-\infty}^{\frac{\boldsymbol{\beta^T X_i}}{\sigma}} \frac{e^{\frac{-s^2}{2}}}{\sqrt{2\pi}} ds \tag{8}$$

Cuando se trata de modelos logit, $F(x)$ es la función logística y el modelo de respuesta binaria se escribe como:

$$P_i = E\left(Y_i \mid \Psi_i\right) = F\left(\boldsymbol{\beta^T X_i}\right) = \frac{e^{\boldsymbol{\beta^T X_i}}}{1 + e^{\boldsymbol{\beta^T X_i}}} \tag{9}$$

Los modelos probit y logit pueden ser derivados de otro modelo que introduce una variable no observada o latente y^*, de la siguiente manera. Sea,

$$y_i^* = \boldsymbol{\beta^T X_i} + \epsilon_i \text{ con } \epsilon_i \ \text{ iid}(0, 1) \tag{10}$$

Si bien y^* no se observa, decimos que,

$$Y_i = 1 \text{ si } y_i^* > 0 \text{ y } Y_i = 0 \text{ si } y_i^* \leq 0 \tag{11}$$

Luego, la probabilidad que $Y_i = 1$ viene dada por:

$$\begin{aligned} P(Y_i = 1) &= P\left(y_i^* > 0\right) \\ &= P\left(\boldsymbol{\beta^T X_i} + \epsilon_i > 0\right) \\ &= 1 - P\left(\boldsymbol{\beta^T X_i} + \epsilon_i \leq 0\right) \\ &= 1 - P\left(\epsilon_i \leq -\boldsymbol{\beta^T X_i}\right) \\ &= 1 - F\left(-\boldsymbol{\beta^T X_i}\right) \\ &= F\left(\boldsymbol{\beta^T X_i}\right) \end{aligned} \tag{12}$$

ya que se supone que ϵ_i tiene una distribución que es simétrica. Cuando ϵ_i $N(0,1)$ F es la función de distribución de probabilidades acumuladas normal estándar Φ y se trata del modelo probit, mientras que si F es la función logística se trata de un logit y su densidad también es simétrica alrededor de cero.

Luego, y tomando como ejemplo el modelo logit,

$$\begin{aligned} P(Y_i = 1) &= \frac{e^{\boldsymbol{\beta^T X_i}}}{1 + e^{\boldsymbol{\beta^T X_i}}} \\ P(Y_i = 0) &= \frac{1}{1 + e^{\boldsymbol{\beta^T X_i}}} \end{aligned} \tag{13}$$

2.6.3. Estimación de los modelos logit y probit

La estimación de estos modelos se hace a través de máxima verosimilitud. Con métodos numéricos se buscan los valores de $\boldsymbol{\beta}$ que maximizan la siguiente función logarítmica de verosimilitud:

$$\ell(\boldsymbol{\beta}) = \sum_{i=1}^{n} \left(Y_i \log\left(F\left(\boldsymbol{\beta^T X_i}\right)\right) + (1 - Y_i)\log\left(1 - F\left(\boldsymbol{\beta^T X_i}\right)\right)\right) \tag{14}$$

Las condiciones de primer orden para un máximo en (14) son:

$$\sum_{i=1}^{n} \frac{\left(Y_i - \hat{F}\right)\hat{f}_i X_{ij}}{\hat{F}_i\left(1 - \hat{F}_i\right)} = 0, \qquad j = 1, \ldots, k \text{ donde, } \hat{F}_i \equiv F_i\left(\boldsymbol{b^T X_i}\right) \text{ y } \hat{f}_i \equiv F_i\left(\boldsymbol{b^T X_i}\right) \tag{15}$$

donde $\boldsymbol{b}$ es el vector de estimativos máximo verosímiles. Cuando (15) es globalmente cóncava, satisface las condiciones de primer orden y asegura que el máximo es único. Los modelos probit, logit y otros tipos de modelos de respuesta binaria satisfacen las condiciones de regularidad que son necesarias para que las estimaciones de los parámetros sean consistentes y asintóticamente normales, con la matriz de covarianzas asintótica que no es más que la inversa de la matriz de información.

2.6.4. Derivación de las condiciones de primer y segundo orden para los modelos probit y logit

El método de máxima verosimilitud permite obtener un estimador para un parámetro o conjunto de parámetros. La densidad conjunta para n observaciones que es el producto de las densidades individuales, se denomina función de verosimilitud y es función del vector θ de parámetros desconocidos:

$$f\left(\boldsymbol{X_1}, \boldsymbol{X_2}, \ldots, \boldsymbol{X_n}, \boldsymbol{\theta}\right) = \prod_{i=1}^{n} f\left(\boldsymbol{X_i}, \boldsymbol{\theta}\right) = \mathcal{L}\left(\boldsymbol{\theta}/\boldsymbol{X}\right) \tag{16}$$

Tomando logaritmos,

$$\ln \mathcal{L}\left(\boldsymbol{\theta}/\boldsymbol{X}\right) = \sum_{i=1}^{n} \ln f\left(\boldsymbol{X_i}, \boldsymbol{\theta}\right) = \ln \mathcal{L}\left(\boldsymbol{\theta}\right) \tag{17}$$

$\hat{\theta}$ es el estimador de máxima verosimilitud de θ que permite maximizar (16) y (17) y resuelve la condición de primer orden $\dfrac{d \ln \mathcal{L}\left(\boldsymbol{\theta}\right)}{d\boldsymbol{\theta}} = 0$.

2.6.5. Propiedades de los estimadores de máxima verosimilitud

Si $f\left(\boldsymbol{X_i}, \boldsymbol{\theta}\right)$ cumple ciertas condiciones de regularidad, $\hat{\theta}$ tiene las siguientes propiedades:

❶ **Consistencia:** $\hat{\boldsymbol{\theta}} \xrightarrow{p} \boldsymbol{\theta}$ cuando $n \to \infty$

❷ **Normalidad Asintótica:** $\hat{\boldsymbol{\theta}} \xrightarrow{a} N\left[\boldsymbol{\theta}, \left\{I\left(\boldsymbol{\theta}\right)\right\}^{-1}\right]$, con $I\left(\boldsymbol{\theta}\right) = -E\dfrac{\partial^2 \ln \mathcal{L}}{\partial\boldsymbol{\theta}\partial\boldsymbol{\theta}^T}$

❸ **Eficiencia Asintótica**, y alcanza el límite inferior de *Cramer-Rao* para estimadores consistentes (la mínima varianza alcanzable por un estimador consistente).

❹ **Invarianza:** el estimador de máxima verosimilitud de $\gamma = c(\boldsymbol{\theta})$ es $c(\hat{\boldsymbol{\theta}})$

2.6.6. Derivación de la función de verosimilitud para los modelos probit y logit.

Sabiendo que $Pr(Y_i = 1) = F(\boldsymbol{\beta^T X_i}) = p$ y $Pr(Y_i = 0) = 1 - F(\boldsymbol{\beta^T X_i}) = 1 - p$, si a demás se supone que Y es una realización de un proceso *Bernoulli* se puede escribir,

$$Pr(Y_i = y) = p^y(1-p)^{1-y} \quad \text{o también} \quad Pr(Y_i = y) = F\left(\boldsymbol{\beta^T X_i}\right)^y \left(1 - F(\boldsymbol{\beta^T X_i})\right)^{1-y} \tag{18}$$

Para una muestra independiente de $Y_1, \ldots, Y_n$:

$$\begin{aligned} Pr\left(Y_1 = y_1, Y_2 = y_2, \ldots, Y_n = y_n\right) &= \prod_{i=1}^{n} Pr\left(Y_i = y_i\right) \\ &= \prod_{i=1}^{n} F\left(\boldsymbol{\beta^T X_i}\right)^y \left(1 - F\left(\boldsymbol{\beta^T X_i}\right)\right)^{1-y} \\ &= \prod_{i=1}^{n} F\left(\boldsymbol{\beta^T X_i}\right)^y \prod_{i=1}^{n} \left(1 - F\left(\boldsymbol{\beta^T X_i}\right)\right)^{1-y} = \mathcal{L} \end{aligned} \tag{19}$$

que por definición es una función de verosimilitud. Tomando logaritmos, la función logarítmica de verosimilitud es:

$$\ln \mathcal{L} = \ln \left[\prod_{i=1}^{n} F\left(\boldsymbol{\beta}^{T} \boldsymbol{X_i}\right)^{y} \right] + \ln \left[\prod_{i=1}^{n} \left(1 - F\left(\boldsymbol{\beta}^{T} \boldsymbol{X_i}\right)\right)^{1-y} \right] \tag{20}$$

Si ϵ_i se distribuye idénticamente como una $N(0, \sigma^2)$ sabemos de la ecuación (8) que $\boldsymbol{\beta}$ y σ no están identificados por lo que no se pueden estimar separadamente. Sólo se puede estimar $\boldsymbol{\beta}/\sigma$, por lo que se supone que $\sigma = 1$. Siendo ϵ_i una variable gaussiana, la función a máximizar es

$$\begin{aligned}
\ln \mathcal{L} &= \ln \left[\prod_{i=1}^{n} \Phi\left(\boldsymbol{\beta}^{T} \boldsymbol{X_i}\right)^{y} \right] + \ln \left[\prod_{i=1}^{n} \left(1 - \Phi\left(\boldsymbol{\beta}^{T} \boldsymbol{X_i}\right)\right)^{1-y} \right] \text{ o,} \\
\ln \mathcal{L} &= \sum_{i=1}^{n} y_i \ln \left[\Phi\left(\boldsymbol{\beta}^{T} \boldsymbol{X_i}\right) \right] + \sum_{i=1}^{n} (1 - y_i) \ln \left[\left(1 - \Phi\left(\boldsymbol{\beta}^{T} \boldsymbol{X_i}\right)\right) \right] \qquad (21)
\end{aligned}$$

2.6.7. Condiciones de primer orden para los modelos probit y logit.

$$\begin{aligned}
S(\boldsymbol{\beta}) = \frac{\partial \ln \mathcal{L}}{\partial \boldsymbol{\beta}} &= \frac{\partial \left[\sum_{i=1}^{n} y_i \ln \left(\Phi\left(\boldsymbol{\beta}^{T} \boldsymbol{X_i}\right)\right) \right]}{\partial \boldsymbol{\beta}} + \frac{\partial \left[\sum_{i=1}^{n} (1 - y_i) \ln \left(\left(1 - \Phi\left(\boldsymbol{\beta}^{T} \boldsymbol{X_i}\right)\right)\right) \right]}{\partial \boldsymbol{\beta}} \\
&= \sum_{i=1}^{n} y_i \frac{\phi\left(\boldsymbol{\beta}^{T} \boldsymbol{X_i}\right) \boldsymbol{X_i}}{\Phi\left(\boldsymbol{\beta}^{T} \boldsymbol{X_i}\right)} + \sum_{i=1}^{n} \frac{1 - y_i}{1 - \Phi\left(\boldsymbol{\beta}^{T} \boldsymbol{X_i}\right)} \left[-\phi\left(\boldsymbol{\beta}^{T} \boldsymbol{X_i}\right) \right] \boldsymbol{X_i} \\
&= \sum_{i=1}^{n} \left[\frac{y_i}{\Phi\left(\boldsymbol{\beta}^{T} \boldsymbol{X_i}\right)} - \frac{1 - y_i}{1 - \Phi\left(\boldsymbol{\beta}^{T} \boldsymbol{X_i}\right)} \right] \phi\left(\boldsymbol{\beta}^{T} \boldsymbol{X_i}\right) \boldsymbol{X_i} \qquad (22)
\end{aligned}$$

Resolviendo la resta de la ecuación (22) tenemos:

$$\sum_{i=1}^{n} \left[\frac{y_i - \Phi\left(\boldsymbol{\beta}^{T} \boldsymbol{X_i}\right)}{\Phi\left(\boldsymbol{\beta}^{T} \boldsymbol{X_i}\right) \left[1 - \Phi\left(\boldsymbol{\beta}^{T} \boldsymbol{X_i}\right)\right]} \right] \phi\left(\boldsymbol{\beta}^{T} \boldsymbol{X_i}\right) \boldsymbol{X_i} \tag{23}$$

que es igual a cero cuando se evalúa en los coeficientes estimados b_j.

2.6.8. Derivación de la matriz de información y las Condiciones de segundo orden para los modelos probit y logit.

$$
\begin{aligned}
\frac{\partial S(\boldsymbol{\beta})}{\partial \boldsymbol{\beta}} &= \frac{\partial}{\partial \boldsymbol{\beta}^{\boldsymbol{T}}}\left\{\sum_{i=1}^{n}\left[\frac{y_i - \Phi\left(\boldsymbol{\beta}^{\boldsymbol{T}}\boldsymbol{X_i}\right)}{\Phi\left(\boldsymbol{\beta}^{\boldsymbol{T}}\boldsymbol{X_i}\right)\left[1-\Phi\left(\boldsymbol{\beta}^{\boldsymbol{T}}\boldsymbol{X_i}\right)\right]}\right]\phi\left(\boldsymbol{\beta}^{\boldsymbol{T}}\boldsymbol{X_i}\right)\boldsymbol{X_i}\right\} \\
&= \sum_{i=1}^{n}\frac{\partial}{\partial \boldsymbol{\beta}^{\boldsymbol{T}}}\left[\frac{y_i - \Phi\left(\boldsymbol{\beta}^{\boldsymbol{T}}\boldsymbol{X_i}\right)}{\Phi\left(\boldsymbol{\beta}^{\boldsymbol{T}}\boldsymbol{X_i}\right)\left[1-\Phi\left(\boldsymbol{\beta}^{\boldsymbol{T}}\boldsymbol{X_i}\right)\right]}\right]\phi\left(\boldsymbol{\beta}^{\boldsymbol{T}}\boldsymbol{X_i}\right)\boldsymbol{X_i} + \sum_{i=1}^{n}\left[\frac{y_i - \Phi\left(\boldsymbol{\beta}^{\boldsymbol{T}}\boldsymbol{X_i}\right)}{\Phi\left(\boldsymbol{\beta}^{\boldsymbol{T}}\boldsymbol{X_i}\right)\left[1-\Phi\left(\boldsymbol{\beta}^{\boldsymbol{T}}\boldsymbol{X_i}\right)\right]}\right]\left[\frac{\partial}{\partial \boldsymbol{\beta}^{\boldsymbol{T}}}\left(\phi\left(\boldsymbol{\beta}^{\boldsymbol{T}}\boldsymbol{X_i}\right)\boldsymbol{X_i}\right)\right] \\
&= \sum_{i=1}^{n}\frac{\partial}{\partial \boldsymbol{\beta}^{\boldsymbol{T}}}\left[\frac{y_i - \Phi\left(\boldsymbol{\beta}^{\boldsymbol{T}}\boldsymbol{X_i}\right)}{\Phi\left(\boldsymbol{\beta}^{\boldsymbol{T}}\boldsymbol{X_i}\right)\left[1-\Phi\left(\boldsymbol{\beta}^{\boldsymbol{T}}\boldsymbol{X_i}\right)\right]}\right]\phi\left(\boldsymbol{\beta}^{\boldsymbol{T}}\boldsymbol{X_i}\right)\boldsymbol{X_i} + \sum_{i=1}^{n}\left[\frac{y_i - \Phi\left(\boldsymbol{\beta}^{\boldsymbol{T}}\boldsymbol{X_i}\right)}{\Phi\left(\boldsymbol{\beta}^{\boldsymbol{T}}\boldsymbol{X_i}\right)\left[1-\Phi\left(\boldsymbol{\beta}^{\boldsymbol{T}}\boldsymbol{X_i}\right)\right]}\right]\boldsymbol{X_i^T}\boldsymbol{X_i} \\
&= \sum_{i=1}^{n}\left[\left(\frac{-\phi(\cdot)\boldsymbol{X_i^T}\Phi(\cdot)(1-\Phi(\cdot)) - (y_i - \Phi(\cdot))\left[\phi(\cdot)\boldsymbol{X_i^T}(y_i - \Phi(\cdot)) + \Phi(\cdot)\left(-\phi(\cdot)\boldsymbol{X_i^T}\right)\right]}{\left[\Phi\left(\boldsymbol{\beta}^{\boldsymbol{T}}\boldsymbol{X_i}\right)\left(1-\Phi\left(\boldsymbol{\beta}^{\boldsymbol{T}}\boldsymbol{X_i}\right)\right)\right]^2}\right)\phi\left(\boldsymbol{\beta}^{\boldsymbol{T}}\boldsymbol{X_i}\right)\boldsymbol{X_i}\right] + \ldots \\
&= \sum_{i=1}^{n}\left[\left(\frac{-\Phi\left(\boldsymbol{\beta}^{\boldsymbol{T}}\boldsymbol{X_i}\right)\left(1-\Phi\left(\boldsymbol{\beta}^{\boldsymbol{T}}\boldsymbol{X_i}\right)\right) - \left(y_i - \Phi\left(\boldsymbol{\beta}^{\boldsymbol{T}}\boldsymbol{X_i}\right)\right)\left[1-2\Phi\left(\boldsymbol{\beta}^{\boldsymbol{T}}\boldsymbol{X_i}\right)\right]}{\left[\Phi\left(\boldsymbol{\beta}^{\boldsymbol{T}}\boldsymbol{X_i}\right)\left(1-\Phi\left(\boldsymbol{\beta}^{\boldsymbol{T}}\boldsymbol{X_i}\right)\right)\right]}\right)\frac{\phi\left(\boldsymbol{\beta}^{\boldsymbol{T}}\boldsymbol{X_i}\right)^2\boldsymbol{X^T}\boldsymbol{X_i}}{\Phi\left(\boldsymbol{\beta}^{\boldsymbol{T}}\boldsymbol{X_i}\right)\left(1-\Phi\left(\boldsymbol{\beta}^{\boldsymbol{T}}\boldsymbol{X_i}\right)\right)}\right] + \ldots \\
&= \sum_{i=1}^{n}\left[\left(-1 + \frac{(\Phi(\cdot) - y_i)\left[1-2\Phi(\cdot)\right]}{\left[\Phi(\cdot)\left(1-\Phi(\cdot)\right)\right]}\right)\frac{\phi(\cdot)^2\boldsymbol{X^T}\boldsymbol{X_i}}{\left[\Phi(\cdot)\left(1-\Phi(\cdot)\right)\right]}\right] + \sum_{i=1}^{n}\left(\frac{y_i - \Phi(\cdot)}{\Phi(\cdot)\left(1-\Phi(\cdot)\right)}\right)\boldsymbol{X^T}\boldsymbol{X_i} \\
&= \sum_{i=1}^{n}\left\{\left(-1 + \frac{(\Phi(\cdot) - y_i)\left[1-2\Phi(\cdot)\right]}{\left[\Phi(\cdot)\left(1-\Phi(\cdot)\right)\right]}\right)\frac{\phi(\cdot)^2\boldsymbol{X^T}\boldsymbol{X_i}}{\left[\Phi(\cdot)\left(1-\Phi(\cdot)\right)\right]} + \left(\frac{y_i - \Phi(\cdot)}{\Phi(\cdot)\left(1-\Phi(\cdot)\right)}\right)\boldsymbol{X^T}\boldsymbol{X_i}\right\}
\end{aligned}
$$

Para estimar la matriz de información se debe calcular la esperanza matemática de $-\frac{\partial S(\boldsymbol{\beta})}{\partial \boldsymbol{\beta}}$. Al hacerlo, $\left(\Phi\left(\boldsymbol{\beta}^{\boldsymbol{T}}\boldsymbol{X_i}\right) - y_i\right)$ es igual a cero y se obtiene la matriz de información,

$$
I\left(\hat{\boldsymbol{\beta}}\right) = \sum_{i=1}^{n}\frac{\phi\left(\boldsymbol{\beta}^{\boldsymbol{T}}\boldsymbol{X_i}\right)^2\boldsymbol{X_i^T}\boldsymbol{X_i}}{\Phi\left(\boldsymbol{\beta}^{\boldsymbol{T}}\boldsymbol{X_i}\right)\left(1-\Phi\left(\boldsymbol{\beta}^{\boldsymbol{T}}\boldsymbol{X_i}\right)\right)} \tag{24}
$$

cuya inversa es la estimación de las varianzas y covarianzas asintóticas de los parámetros estimados.

2.7. Modelos multinomiales o de respuesta múltiple

Cuando Y toma más de dos valores (es policotómica), el método que se debe emplear para estimar los parámetros depende de la naturaleza de Y. Supongamos que Y es el resultado de un sistema de *rating* que asigna calificaciones de riesgo. En este caso Y no sólo es cualitativa multinomial sino que tiene un orden inherente, por lo tanto es una variable ordinal y los valores que puede tomar tienen un ordenamiento que implican una jerarquía inherente. Para este caso para la estimación se emplea un probit o logit ordenado.

2.7.1. El modelo probit y logit ordenado

Sea y^* un índice no observado o latente relacionada con la calidad crediticia. Se supone que es función lineal de ciertas variables explicativas que están contenidas en el vector $\boldsymbol{X_i}$, a las cuales les corresponden coeficientes del vector $\boldsymbol{\beta}$, y de un término de error con distribución normal. La expresión para y^* viene dada

$$y_i^* \quad = \quad -\boldsymbol{\beta X_i} + \epsilon_i \tag{25}$$

El índice y^* está oculto detrás de la calificación asignada. Si tuviéramos un esquema de calificación estructurado en 5 ratings, se puede pensar que las distintas calificaciones están definidas por rangos de y^* de la siguiente manera,

$$\begin{array}{ll} Y_i = 1 & \text{si } y_i^* \leq 0, \\ Y_i = 2 & \text{si } 0 < y_i^* \leq \lambda_2, \\ Y_i = 3 & \text{si } \lambda_2 < y_i^* \leq \lambda_3, \\ Y_i = 4 & \text{si } \lambda_3 < y_i^* \leq \lambda_4, \\ Y_i = 5 & \text{si } \lambda_4 < y_i^* \end{array} \tag{26}$$

Los λ son los puntos de cortes que definen los rangos de las distintas calificaciones, que se estiman en conjunto con el vector $\boldsymbol{\beta}$, estableciendo la restricción[6] de que $0 < \lambda_2 < \lambda_3 < \lambda_4$. La probabilidad de obtener la mejor calificación, 1, es:

$$P(Y_i = 1) = P\left(y_i^* \leq 0\right) = P\left(\epsilon_i - \boldsymbol{\beta^T X_i} \leq 0\right) = P\left(\epsilon_i \leq \boldsymbol{\beta^T X_i}\right) \tag{27}$$

Como ϵ_i $N(0,1)$, tenemos que

$$P\left(\epsilon_i \leq \boldsymbol{\beta^T X_i}\right) = \Phi\left(\boldsymbol{\beta^T X_i}\right). \tag{28}$$

La probabilidad asociada a las demás calificaciones se calcula como se ve a continuación.

$$\begin{array}{lll} P(Y_i = 1) & = & F\left(\boldsymbol{\beta^T X_i}\right) = \Phi\left(\boldsymbol{\beta^T X_i}\right) \\ P(Y_i = 2) & = & F\left(\lambda_2 + \boldsymbol{\beta^T X_i}\right) - F\left(\boldsymbol{\beta^T X_i}\right) \\ P(Y_i = 3) & = & F\left(\lambda_3 + \boldsymbol{\beta^T X_i}\right) - F\left(\lambda_2 + \boldsymbol{\beta^T X_i}\right) \\ P(Y_i = 4) & = & F\left(\lambda_4 + \boldsymbol{\beta^T X_i}\right) - F\left(\lambda_3 + \boldsymbol{\beta^T X_i}\right) \\ P(Y_i = 5) & = & 1 - F\left(\lambda_4 + \boldsymbol{\beta^T X_i}\right) \end{array} \tag{29}$$

[6]Para obtener probabilidades positivas.

Derivando la función de probabilidad condicional respecto de la variable k, vemos que existen tantos efectos marginales como posibles niveles de calificación, en este caso cinco:

$$\begin{aligned}
\frac{\partial P(Y_i = 1)}{\partial X_{ik}} &= \left[f\left(\boldsymbol{\beta}^{\boldsymbol{T}}\boldsymbol{X_i}\right)\right]\beta_k \\
\frac{\partial P(Y_i = 2)}{\partial X_{ik}} &= \left[f\left(\lambda_2 + \boldsymbol{\beta}^{\boldsymbol{T}}\boldsymbol{X_i}\right) - f\left(\boldsymbol{\beta}^{\boldsymbol{T}}\boldsymbol{X_i}\right)\right]\beta_k \\
\frac{\partial P(Y_i = 3)}{\partial X_{ik}} &= \left[f\left(\lambda_3 + \boldsymbol{\beta}^{\boldsymbol{T}}\boldsymbol{X_i}\right) - f\left(\lambda_2 + \boldsymbol{\beta}^{\boldsymbol{T}}\boldsymbol{X_i}\right)\right]\beta_k \\
\frac{\partial P(Y_i = 4)}{\partial X_{ik}} &= \left[f\left(\lambda_4 + \boldsymbol{\beta}^{\boldsymbol{T}}\boldsymbol{X_i}\right) - f\left(\lambda_3 + \boldsymbol{\beta}^{\boldsymbol{T}}\boldsymbol{X_i}\right)\right]\beta_k \\
\frac{\partial P(Y_i = 2)}{\partial X_{ik}} &= \left[1 - f\left(\lambda_4 + \boldsymbol{\beta}^{\boldsymbol{T}}\boldsymbol{X_i}\right)\right]\beta_k
\end{aligned} \tag{30}$$

Combinando (26) y (27) reescribimos los rangos que definen las calificaciones como sigue:

$$\begin{array}{ll}
Y_i = 1 & \text{si } y_i^* = -\boldsymbol{\beta}^{\boldsymbol{T}}\boldsymbol{X_i} + \epsilon_i \leq 0, \\
Y_i = 2 & \text{si } 0 < y_i^* = -\boldsymbol{\beta}^{\boldsymbol{T}}\boldsymbol{X_i} + \epsilon_i \leq \lambda_2, \\
Y_i = 3 & \text{si } \lambda_2 < y_i^* = -\boldsymbol{\beta}^{\boldsymbol{T}}\boldsymbol{X_i} + \epsilon_i \leq \lambda_3, \\
Y_i = 4 & \text{si } \lambda_3 < y_i^* = -\boldsymbol{\beta}^{\boldsymbol{T}}\boldsymbol{X_i} + \epsilon_i \leq \lambda_4, \\
Y_i = 5 & \text{si } \lambda_4 < -\boldsymbol{\beta}^{\boldsymbol{T}}\boldsymbol{X_i} + \epsilon_i
\end{array} \tag{31}$$

Como $\epsilon_i\ N(0,1)$, $\left(y_i^* + \boldsymbol{\beta}^{\boldsymbol{T}}\boldsymbol{X_i} = \epsilon_i\right)\ N(0,1)$ o también $\left(y_i^* = \epsilon_i - \boldsymbol{\beta}^{\boldsymbol{T}}\boldsymbol{X_i}\right)\ N(-\boldsymbol{\beta}^{\boldsymbol{T}}\boldsymbol{X_i}, 1)$.

Los cambios en la probabilidad de tener las distintas situaciones frente a un cambio en alguna de las variables explicativas está asociado a un desplazamiento de la distribución del *score*. Según la última ecuación el *score* tiene una distribución normal con media $-\boldsymbol{\beta}^{\boldsymbol{T}}\boldsymbol{X_i}$ que se mueve sobre los puntos de cortes frente a los cambios en el *score*. Por ejemplo, el cambio en una variable que aumenta el *score* de un individuo mueve su distribución a la izquierda, aumentando la probabilidad de que tenga $Y_i = 1$.

2.7.2. Derivación de la función de verosimilitud para los modelos probit y logit ordenado

Cuando la variable a predecir es policotómica y ordinal, estimar el modelo probit ordenado supone que Y tiene una distribución multinomial (una generalización de la binomial para más de dos categorías). Para este caso la función de verosimilitud se obtiene como,

$$\begin{aligned}
&Pr(Y_1 = y_1, Y_2 = y_2, \ldots, Y_n = y_n) = \prod_{i=1}^{n} Pr(Y_i = y_i) = \mathcal{L} \\
&\mathcal{L} = \prod_{y_i=1} P(Y_i = 1) \prod_{y_i=2} P(Y_i = 2) \prod_{y_i=3} P(Y_i = 3) \prod_{y_i=4} P(Y_i = 4) \prod_{y_i=5} P(Y_i = 5)
\end{aligned} \tag{32}$$

Tomando logaritmos tenemos,

$$
\begin{aligned}
\ell(\beta,\alpha_2,\alpha_3,\alpha_4) &= \sum_{y_i=1}\log\left(P\left(Y_i=1\right)\right)+\sum_{y_i=2}\log\left(P\left(Y_i=2\right)\right)+\sum_{y_i=3}\log\left(P\left(Y_3=1\right)\right)+\\
&\quad \sum_{y_i=4}\log\left(P\left(Y_i=4\right)\right)+\sum_{y_i=5}\log\left(P\left(Y_i=5\right)\right)\\
\ell(\beta,\alpha_2,\alpha_3,\alpha_4) &= \sum_{y_i=1}\log\left(\Phi\left(\boldsymbol{\beta}^{\boldsymbol{T}}\boldsymbol{X_i}\right)\right)+\sum_{y_i=2}\log\left(\Phi\left(\lambda_2+\boldsymbol{\beta}^{\boldsymbol{T}}\boldsymbol{X_i}\right)-\Phi\left(\boldsymbol{\beta}^{\boldsymbol{T}}\boldsymbol{X_i}\right)\right)+\\
&\quad \sum_{y_i=3}\log\left(\Phi\left(\lambda_3+\boldsymbol{\beta}^{\boldsymbol{T}}\boldsymbol{X_i}\right)-\Phi\left(\lambda_2+\boldsymbol{\beta}^{\boldsymbol{T}}\boldsymbol{X_i}\right)\right)+\\
&\quad \textstyle\sum_{y_i=4}\log\left(\Phi\left(\lambda_4+\boldsymbol{\beta}^{\boldsymbol{T}}\boldsymbol{X_i}\right)-\Phi\left(\lambda_3+\boldsymbol{\beta}^{\boldsymbol{T}}\boldsymbol{X_i}\right)\right)+\displaystyle\sum_{y_i=5}\log\left(1-\Phi\left(\lambda_4+\boldsymbol{\beta}^{\boldsymbol{T}}\boldsymbol{X_i}\right)\right)
\end{aligned}
\tag{33}
$$

Al resolver con métodos numéricos las condiciones de primer orden se obtienen las estimaciones de los parámetros $\boldsymbol{\beta}$ y los puntos de cortes. Con estas estimaciones se evalúa la matriz de derivadas parciales segundas. Esta matriz, con el signo invertido, es la matriz de información, y la inversa de la matriz de información da la estimación de las varianzas y covarianzas asintóticas de los parámetros estimados.

2.8. Interpretación del modelo

En el contexto de los *Modelos de Score* se puede asociar $\boldsymbol{\beta}^{\boldsymbol{T}}\boldsymbol{X_i}$ con la calidad crediticia del individuo. Cambiando la denominación por Z_j, esta variable representa la calidad crediticia del individuo, que se puede suponer que es el resultado de una función lineal de sus parámetros como se muestra a continuación:

$$Z_j = \beta_0+\beta_1X_{1j}+\beta_2X_{2j}+\ldots+\epsilon_j \tag{34}$$

Las estimaciones de los parámetros β_i se obtienen por métodos de máxima verosimilitud como se explicó en los párrafos anteriores, y las variables X_{ij} contienen la información de los clientes. Después de obtener las estimaciones b_i, el modelo empírico con el que trabajará el analista de riesgo es,

$$z_j = b_0+b_1X_{1j}+b_2X_{2j}+\ldots+b_5X_{5j} \tag{35}$$

cuando se trata de un modelo que utiliza cinco variables ($i = 5$). La variable z_j es el *Z-score* estimado del cliente, esto es, una medida de su calidad crediticia que se obtiene a partir de los parámetros estimados y de su propia información. Este *Z-score*, aplicado a las funciones de distribución de probabilidades acumuladas normal o logística, permite conocer la probabilidad de incumplimiento y en consecuencia el riesgo asociado que tiene cada cliente.

2.8.1. Relación entre el Z-score y el riesgo

Después de haber definido al *Z-score*, se puede reescribir la ecuación (13) como,

$$
\begin{aligned}
P\left(Y_i=1\right) &= \frac{e^{Z_i}}{1+e^{Z_i}}\\
P\left(Y_i=0\right) &= \frac{1}{1+e^{Z_i}}
\end{aligned}
\tag{36}
$$

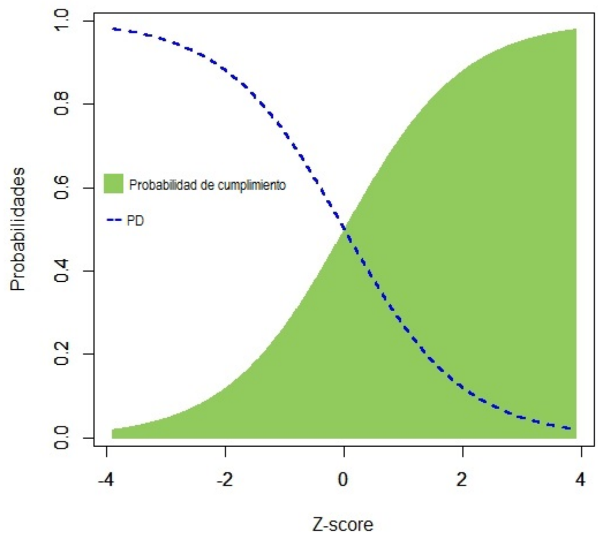

Figura 3: Relación entre el Z-score y el riesgo

donde hay que notar que cambios en Z_i implican cambios en la probabilidad de incumplimiento del individuo. Como se observa en la Figura 3, la relación entre el score y el riesgo (la PD) no es lineal, por lo que el cambio en el riesgo que provenga de un cambio en el Z-score depende de los valores que este último tome. Para valores del Z-score muy bajos, un aumento en el mismo trae consigo una rápida subida en la probabilidad de cumplimiento y una rápida disminución de la PD, mientras que para valores del Z-score altos, una mejora en el mismo hace que la probabilidad de cumplimiento aumente poco y esto hace que se genere una leve caída en el riesgo. Es decir, cuanto mayor es el Z-score, menor es la caída en el riesgo que se deriva de un aumento en el primero.

2.8.2. Efectos Marginales

Para los β_i (o su estimación b_j) no se tienen una interpretación directa como en mínimos cuadrados ordinarios, ya que solamente representan el efecto que un cambio en X_{ij} produce sobre el Z-score del individuo, a demás, su signo muestra si la relación con la PD es directa o inversa. Sin embargo, para ***cuantificar*** el efecto de X_{ij} sobre la PD necesitamos computar su ***efecto marginal***.

Dado que F es una función no lineal, cualquier cambio en los valores de cualquiera de las variables explicativas (si bien afectan linealmente a la función índice) tienen un efecto no lineal sobre la probabilidad estimada de ocurrencia del evento. El efecto marginal de X_{ij} indica el cambio en la probabilidad de ocurrencia del evento (este evento es el cumplimiento de las obligaciones) para el individuo i, ante un pequeño cambio en el valor de la variable X_{ij}.

Como, $E(Y_i \mid \Psi_i) = P_i = F(Z_i)$ tenemos que:

$$\frac{\partial P_i}{\partial X_{ij}} = \frac{\partial F\left(Z_j\right)}{\partial X_{ij}} = \frac{dF\left(Z_j\right)}{dZ_j}\frac{\partial Z_j}{\partial X_{ij}} = f\left(Z_j\right)\frac{\partial Z_j}{\partial X_{ij}} = f\left(Z_j\right)\beta_{ij} \tag{37}$$

En este caso, el evento se define como el cumplimiento normal de las obligaciones, por lo que la ecuación (37) muestra cómo cambia la probabilidad de cumplir con el pago en respuesta a un pequeño cambio en X_{ij}. Sin embargo, como $PD_i = 1 - P_i$,

$$\frac{\partial PD_i}{\partial X_{ij}} = -\frac{\partial F\left(Z_j\right)}{\partial X_{ij}} = -\frac{dF\left(Z_j\right)}{dZ_j}\frac{\partial Z_j}{\partial X_{ij}} = -f\left(Z_j\right)\frac{\partial Z_j}{\partial X_{ij}} = -f\left(Z_j\right)\beta_j \tag{38}$$

La expresión en la ecuación (38) logra cuantificar cuánto cambia la PD del individuo i ante cambios pequeños en la variable continua X_{ij}. Observando de derecha a izquierda, la primer derivada parcial (o β_j) muestra el efecto de un cambio en X_{ij} sobre el Z-score del cliente i, mientras que la segunda muestra el efecto de un cambio en el Z-score sobre la PD, que se tiene al evaluar la función de densidad en Z_j. Empleando la función de densidad de una variable con distribución logística y reemplazando Z_j y β_j por sus estimaciones (z_j y b_j respectivamente), la estimación del efecto marginal viene dada por,

$$\frac{\partial \hat{PD}_i}{\partial X_{ij}} = -\frac{e^{z_j}}{(1+e^{z_j})^2}b_j \tag{39}$$

Como el cociente es positivo, un valor negativo de b_j producirá que aumentos en X_{ij} arrojen un Z-score bajo y a su vez aumenten la PD.

2.9. Ejemplo de un modelo logit

Para este ejemplo tomamos una base de datos simulada para recrear 1000 clientes de un banco[7] clasificados en *bueno = 1* o *malo = 0* y guardada en la variable *Credit*. Para predecir la probabilidad de que un cliente sea *malo* en un futuro se dispone de informaciones como *El saldo actual de la cuenta corriente*, *Pagos de créditos anteriores*, *Porpósito del crédito*, *Cantidad de créditos en el banco*, *Valor de la caja de ahorro*, *Antigüedad laboral*, *Cuota en % de los ingresos disponibles*, *Estado civil y sexo* entre otras variables. Para estimar el valor de Z dado por la ecuación (34) usamos la regresión logística[8](el modelo logit), el software $r - project$ y la función *glm* obteniendo el siguiente resultado

```
Call:
glm(formula = Credit ~ +balance_credit_acc + duration + moral +
```

[7]tomados de `http://www.learnanalytics.in/datasets/Credit_Scoring.zip`

[8]Podriamos definirla como el conjunto de reglas matemáticas y estadísticas que permite predecir un hecho futuro (que el cliente sea malo) a partir de hechos recientes (cantidad de consultas al boreau, la cantidad de morosidades, cantidad de créditos anteriores, antigüedad laboral, etc.)

```
    verw + hoehe + sparkont + beszeit + rate + famges + buerge +
    wohnzeit + verm + alter + weitkred + wohn + bishkred + beruf +
    pers + telef + gastarb, family = binomial, data = datos)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-2.7232  -0.6838   0.3786   0.6931   2.3268

Coefficients:
                      Estimate Std. Error z value Pr(>|z|)
(Intercept)          -6.493e-01  1.015e+00  -0.639 0.522513
balance_credit_acc2     3.783e-01  2.189e-01   1.728 0.083918 .
balance_credit_acc3     9.630e-01  3.703e-01   2.600 0.009315 **
balance_credit_acc4     1.752e+00  2.339e-01   7.491 6.85e-14 ***
duration             -2.833e-02  9.404e-03  -3.012 0.002594 **
moral1               -1.852e-01  5.687e-01  -0.326 0.744723
moral2                6.038e-01  4.432e-01   1.362 0.173054
moral3                9.735e-01  4.789e-01   2.033 0.042085 *
moral4                1.512e+00  4.462e-01   3.389 0.000701 ***
verw1                 1.629e+00  3.769e-01   4.322 1.54e-05 ***
verw2                 7.287e-01  2.629e-01   2.772 0.005570 **
verw3                 8.507e-01  2.482e-01   3.427 0.000609 ***
verw4                 4.994e-01  7.667e-01   0.651 0.514860
verw5                 1.943e-01  5.541e-01   0.351 0.725836
verw6                -1.163e-01  3.962e-01  -0.293 0.769167
verw8                 1.915e+00  1.170e+00   1.637 0.101633
verw9                 6.524e-01  3.362e-01   1.940 0.052332 .
verw10                1.387e+00  7.794e-01   1.780 0.075103 .
hoehe                -1.200e-04  4.489e-05  -2.672 0.007536 **
sparkont2             3.730e-01  2.914e-01   1.280 0.200586
sparkont3             3.580e-01  4.006e-01   0.894 0.371527
sparkont4             1.422e+00  5.339e-01   2.663 0.007748 **
sparkont5             9.648e-01  2.648e-01   3.643 0.000269 ***
beszeit2             -1.038e-02  4.392e-01  -0.024 0.981141
beszeit3              2.868e-01  4.195e-01   0.684 0.494123
beszeit4              8.251e-01  4.580e-01   1.802 0.071615 .
beszeit5              2.415e-01  4.214e-01   0.573 0.566551
rate2                -2.812e-01  3.085e-01  -0.912 0.361976
rate3                -6.389e-01  3.404e-01  -1.877 0.060498 .
rate4                -9.390e-01  3.033e-01  -3.096 0.001961 **
famges2               2.699e-01  3.880e-01   0.696 0.486693
famges3               8.215e-01  3.803e-01   2.160 0.030742 *
famges4               3.713e-01  4.558e-01   0.815 0.415306
buerge2              -4.328e-01  4.129e-01  -1.048 0.294598
```

```
buerge3              9.194e-01  4.263e-01   2.156 0.031047 *
wohnzeit2           -7.444e-01  2.999e-01  -2.482 0.013060 *
wohnzeit3           -5.068e-01  3.359e-01  -1.509 0.131306
wohnzeit4           -3.694e-01  3.031e-01  -1.219 0.222914
verm2               -2.654e-01  2.545e-01  -1.043 0.297120
verm3               -1.662e-01  2.371e-01  -0.701 0.483317
verm4               -7.113e-01  4.238e-01  -1.678 0.093276 .
alter                1.258e-02  9.273e-03   1.357 0.174865
weitkred2           -3.507e-02  4.264e-01  -0.082 0.934448
weitkred3            4.040e-01  2.474e-01   1.633 0.102437
wohn2                4.742e-01  2.363e-01   2.007 0.044755 *
wohn3                6.108e-01  4.795e-01   1.274 0.202768
bishkred2           -3.896e-01  2.465e-01  -1.580 0.113996
bishkred3           -3.155e-01  5.995e-01  -0.526 0.598763
bishkred4           -3.676e-01  1.075e+00  -0.342 0.732389
beruf2              -5.306e-01  6.795e-01  -0.781 0.434870
beruf3              -5.397e-01  6.553e-01  -0.824 0.410159
beruf4              -4.318e-01  6.661e-01  -0.648 0.516873
pers2               -2.607e-01  2.513e-01  -1.038 0.299474
telef2               2.677e-01  2.025e-01   1.322 0.186227
gastarb2             1.399e+00  6.223e-01   2.248 0.024594 *
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 1221.73  on 999  degrees of freedom
Residual deviance:  892.96  on 945  degrees of freedom
AIC: 1003

Number of Fisher Scoring iterations: 5
```

Tomando las variables que sean significativas[9] al 5 % tenemos que:

$Z = 0{,}9735281824 \times moral3 + 0{,}8215062770 \times famges3 + 0{,}9193954333 \times buerge3 - 0{,}7443610334 \times wohnzeit2 + 0{,}4742480631 \times wohn2 + 1{,}3988588259 \times gastarb2 + 0{,}9629526445 \times balance\_credit\_acc3 - 0{,}0283268096 \times duration + 0{,}7287271411 \times verw2 - 0{,}0001199527 \times hoehe + 1{,}4215951645 \times sparkont4 - 0{,}9389707870 \times rate4 + 1{,}7519424391 \times balance\_credit\_acc4 + 1{,}5121924512 \times moral4 + 1{,}6290689480 \times verw1 + 0{,}8507062365 \times verw3 + 0{,}9647584990 \times sparkont5$.

Luego, la probabilidad de que un cliente sea bueno viene dado por:

$$Pr = \frac{e^Z}{1 + e^Z},$$

[9]Son las variables que tienen por lo menos un asterisco (*)

como $0 \leq Pr \leq 1$ el score se obtiene multiplicando por 1000 la probabilidad anterior, es decir, score = $1000 \times Pr$

2.9.1. Tabla de Comportamiento

La tabla de comportamiento es una de las maneras en que se presentan los resultados de un modelo de score. Esta tabla permite pensar en la mejor opción al momento de definir un punto de corte (esto es, a partir de que puntuación de score vamos a rechazar una solicitud de crédito), a demás, presenta medidas de la bondad de ajuste del modelo, como así también la distribución de los buenos y malos.

Tramos	Score medio	%Total	%AcTotal	%Bueno	%AcBueno	%DesBueno	%Malo	%AcMalo	%DesMalo	Tasa Malo	TasaAcMalo	KS
$[975, 1000]$	986.88	9.7	9.7	13.43	13.43	100.00	1.00	1.00	100.00	3.09	3.09	12.43
$[947, 975)$	960.56	10.1	19.8	13.43	26.86	86.57	2.33	3.33	99.00	6.93	5.05	23.52
$[906, 947)$	925.85	10.1	29.9	13.57	40.43	73.14	2.00	5.33	96.67	5.94	5.35	35.10
$[845, 906)$	875.16	10.1	40.0	11.29	51.71	59.57	7.33	12.67	94.67	21.78	9.50	39.05
$[778, 845)$	810.96	10.0	50.0	11.00	62.71	48.29	7.67	20.33	87.33	23.00	12.20	42.38
$[684, 778)$	730.01	9.9	59.9	10.71	73.43	37.29	8.00	28.33	79.67	24.24	14.19	45.10
$[549, 684)$	620.48	10.0	69.9	9.00	82.43	26.57	12.33	40.67	71.67	37.00	17.45	41.76
$[429, 549)$	495.68	10.1	80.0	8.00	90.43	17.57	15.00	55.67	59.33	44.55	20.88	34.76
$[302, 429)$	361.50	10.0	90.0	6.14	96.57	9.57	19.00	74.67	44.33	57.00	24.89	21.90
$[0, 302)$	184.81	10.0	100.0	3.43	100.00	3.43	25.33	100.00	25.33	76.00	30.00	0.00
Total	94.76	100.0	NA	100.00	NA	NA	100.00	NA	NA	NA	NA	45.10

Cuadro 1: Tabla de Performance Score de Riesgo

En el cuadro (1) agrupamos los mil clientes y los ordenamos según el score en 10 tramos, donde en cada uno se tiene aproximadamente 10 % de los casos. A demás se observa como se distribuyen los buenos y los malos; en este ejemplo tenemos una tasa de malo del 30 % (esto es, la penúltima fila de la columna "TasaAcMalo"). La columna "Tasa Malo" es el porcentaje de malos que cae en cada tramo de score. Notemos que el mejor tramo $[975, 1000]$ tiene una tasa de malo del 3,09 %, mientras que el peor tramo $[0, 302]$ tiene una tasa del 76,00 % (25 veces más que el mejor tramo) y que la tasa de malo va creciendo a medida que disminuye el score. Esto está en concordancia con los resultados esperados, es decir, a mayor score menor probabilidad de incumplimiento.

Supongamos que estos datos corresponden a una entidad en particular y que la misma usualmente rechaza la mitad de las solicitudes de crédito que recibe. Bajo estas condiciones dicha entidad podría usar el score para decidir que mitad de las solicitudes rechazar y lo más razonable es rechazar a la mitad de los casos peor scoreados que según el cuadro (1) serian todas las solicitudes que tengan un score menor a 778. Por tanto, la entidad puede establecer como punto de corte 778 y con esto mantendría el porcentaje habitual de solicitudes rechazadas pero la tasa total de malo pasaría de 30 % a 12,20 % obteniéndose una reducción de más de la mitad de la tasa de malo. Otra posibilidad sería buscar por ejemplo en que tramo se tiene una tasa de malos acumulada 10 puntos porcentuales menor a la tasa total que para nuestro caso es el tramo $[429, 549)$ y establecer como punto de corte 429. Con estas condiciones la entidad estaría rechazando sólo el 20 % que implica hacer negocios con más clientes pero con una tasa de mora menor a la actual. Otra forma de establecer un punto de corte es considerando el porcentaje acumulado de buenos (%AcBueno) o el porcentaje desacumulado de malos (%DesMalo). Si el punto de corte de la entidad es 778, la entidad estaría trabajando con el 50 % de las solicitudes que se espera contenga al 62,71 % de los buenos y estaría rechazando al 79,67 % de los malos. Si el punto de corte es 429 la entidad rechazaria el 20 % de las solicitudes quedándose con el 90,43 % de los buenos y rechazando al 44,33 % de los malos. Estas son algunas de las formas en que las entidades de crédito y que según su apetito de riesgo eligen el punto de corte para el score.

Finalmente tenemos la columna *KS* que es la prueba estadística de *Kolmogóriv - Smirnov* que determina la bondad de ajuste de dos distribuciones que para nuestro caso son las distribuciones de buenos y malos. El *KS* varia de 0 a 100 donde 100 significa que hay una separación total de las distribuciones y 0 significa que las dos distribuciones son idénticas. Cuanto más alto es el *KS* mejor es el modelo ya que esto implica que el modelo logra separar en gran medida a los buenos de los malos. La experiencia sugiere que en poblaciones con poca información crediticia el *KS* varia entre 15 y 25, mientras que en poblaciones con mucha información de crédito varia entre 30 y 70.

Para modelos de respuesta binaria el *KS* de cada tramos de score es el valor absoluto de la diferencia entre el porcentaje acumulado de buenos y el porcentaje acumulado de malos, luego el *KS* final del modelo es el máximo entre todos estos KS's. El modelo de nuestro ejemplo tiene un KS = $45.10\,\%$ y el código usado aparece en el anexo.

Conclusión

Un Boreau de Crédito es una empresa orientada a integrar información sobre el comportamiento crediticio de las personas que solicitan un crédito y funciona en base a los reportes de las empresas que otorgan un crédito. Esta información es resumida mediante un credit scoring que luego se utiliza para analizar el riesgo crediticio de un solicitante. La utilización del credit scoring o score comenzó en los años 70 pero se generalizó a partir de los años 90.

Un modelo de score o simplemente score son algoritmos que de manera automática evalúan el riesgo de crédito de un solicitante de financiamiento o de alguien que ya es cliente de la entidad. En general tienen una dimensión individual, ya que se enfocan en el riesgo de incumplimiento del individuo o empresa, independientemente de lo que ocurra con el resto de la cartera de préstamos.

Otros tipos de Score muy utilizados son los Score de Cobranza y los Score de propensión de compra que a diferencia del Score de Riesgo está pensado para ser usado masivamente, es decir, se calcula a toda la cartera de la entidad.

Existen varias metodologías disponibles para la evaluación del riesgo de crédito como por ejemplo el análisis discriminante, la regresión lineal, la regresión logística, los modelos probit, los modelos logit, los métodos no paramétricos de suavizado, los métodos de programación matemática, los modelos basados en cadenas de Markov, los algoritmos de particionamiento recursivo (árboles de decisión), los sistemas expertos, los algoritmos genéticos, las redes neuronales y, finalmente, el juicio humano, es decir, la decisión de un analista acerca de otorgar un crédito.

El tipo de variables utilizadas varía significativamente según se trate de modelos para la cartera retail (individuos y PyMEs), donde generalmente se usan variables socioeconómicas o datos básicos del emprendimiento productivo, o de grandes empresas donde se utilizan variables extraídas de los estados contables, información cualitativa acerca de la dirección, el sector económico, proyecciones del flujo de fondos, entre otros.

Las entidades emplean estos modelos para resolver solicitudes de crédito. También se emplean para administrar el portafolio de créditos, en cuyo caso se trata de modelos de seguimiento, proactivos o de behavioural scoring, y se pueden emplear para: administrar límites de tarjetas y cuentas corrientes, analizar la rentabilidad de los clientes, ofrecer nuevos productos, monitorear el riesgo y detectar posibles problemas de cobranza, entre otras aplicaciones.

El modelo probit y logit son una de las varias alternativas para estimar modelos de respuesta binaria.La idea consiste en utilizar una función de transformación $F(x)$ que tenga las siguientes propiedades:

$$F(-\infty) = 0, \;\; F(\infty) = 1, \;\; \text{y } f(x) \equiv \frac{dF(x)}{dx} > 0$$

$F(x)$ es una función monótona creciente que mapea de la línea real al intervalo $[0, 1]$. Cuando la función de transformación $F(x)$ es la función de distribución acumulada normal estándar tenemos un modelo probit y cuando $F(x)$ es la función de distribución acumulada logística tenemos un modelo logit.

Finalmente vimos en el ejemplo con datos simulados que mediante la tabla de comportamiento de un score se pueden tomar las mejores decisiones en función de diferentes criterios.

Bibliografía

[Altman, 1968] Altman, E. I. (1968). Financial ratios, discriminant analysis and the prediction of corporate bankruptcy. *The journal of finance*, 23(4):589–609.

[Amemiya, 1981] Amemiya, T. (1981). Qualitative response models: A survey. *Journal of economic literature*, 19(4):1483–1536.

[Boyes et al., 1989] Boyes, W. J., Hoffman, D. L., and Low, S. A. (1989). An econometric analysis of the bank credit scoring problem. *Journal of Econometrics*, 40(1):3–14.

[Cheung, 1996] Cheung, S. (1996). Provincial credit ratings in canada: An ordered probit analysis.

[Falkenstein et al., 2000] Falkenstein, E. G., Boral, A., and Carty, L. V. (2000). Riskcalc for private companies: Moody's default model.

[Fernández, 1995] Fernández, A. J. J. (1995). *Análisis de regresión logística*. Centro de Investigaciones Sociológicas (CIS).

[Gordy, 2000] Gordy, M. B. (2000). A comparative anatomy of credit risk models. *Journal of Banking & Finance*, 24(1):119–149.

[Greene, 1992] Greene, W. H. (1992). A statistical model for credit scoring.

[Hand and Henley, 1997] Hand, D. J. and Henley, W. E. (1997). Statistical classification methods in consumer credit scoring: a review. *Journal of the Royal Statistical Society: Series A (Statistics in Society)*, 160(3):523–541.

[Harrell Jr et al., 2008] Harrell Jr, F. E. et al. (2008). Hmisc: harrell miscellaneous. *R package version*, 3(2).

[Jennings, 2001] Jennings, A. (2001). The importance of credit information and credit scoring for small business lending decisions. In *Proceedings from the Global Conference on Credit Scoring*, pages 5–11. Washington, D. C.

[Marín and Sandoval, 1995] Marín, F. J. and Sandoval, F. (1995). Diseño de redes neuronales artificiales mediante algoritmos genéticos.

[Martínez et al., 2003] Martínez, A. C., Mendoza, H. M. T., and Alvarez, M. M. R. (2003). *Análisis multivariado: un manual para investigadores*.

[Màster, 2013] Màster, T. (2013). The r project for statistical computing.

[Mester et al., 1997] Mester, L. J. et al. (1997). What is the point of credit scoring? *Business review*, 3(Sep/Oct):3–16.

[Milena et al., 2005] Milena, E., Miller, M., and Simbaqueba, L. (2005). The case for information sharing by microfinance institutions: Empirical evidence of the value of credit bureau-type data in the nicaraguan microfinance sector. *The World Bank*.

[Miller and Rojas, 2004] Miller, M. and Rojas, D. (2004). Improving access to credit for smes: an empirical analysis of the viability of pooled data sme credit scoring models in brazil, colombia & mexico. *World Bank working paper*.

[Neter et al., 1996] Neter, J., Kutner, M. H., Nachtsheim, C. J., and Wasserman, W. (1996). *Applied linear statistical models*, volume 4. Irwin Chicago.

[Nickell et al., 2001] Nickell, P., Perraudin, W., Varotto, S., et al. (2001). *Stability of ratings transitions*. Bank of England.

[Orallo et al., 2004] Orallo, H., RAMIREZ, J., QUINTANA, C. R., Orallo, M. J. H., Quintana, M. J. R., and Ramírez, C. F. (2004). *Introducción a la Minería de Datos*. Pearson Prentice Hall,.

[Pineda et al.,] Pineda, S. C. S., Morales, J. C. C., and Velásquez, J. R. Análisis comparativo entre árboles de regresión y clasificación (cart) y regresión logística: aplicación a la caracterización de empresas innovadoras colombianas. *innovación*.

[Powell et al., 2004] Powell, A., Majnoni, G., Miller, M., and Mylenko, N. (2004). *Improving Credit Information, Bank Regulation, and Supervision: On the Role and Desing of Public Credit Registries*, volume 3443. World Bank Publications.

[Sicsú, 2010] Sicsú, A. L. (2010). *Credit Scoring: desenvolvimento, implantação, acompanhamento*. Blucher.

[Siddiqi, 2012] Siddiqi, N. (2012). *Credit risk scorecards: developing and implementing intelligent credit scoring*, volume 3. John Wiley & Sons.

[Srinivasan and Kim, 1987] Srinivasan, V. and Kim, Y. H. (1987). Credit granting: A comparative analysis of classification procedures. *The Journal of Finance*, 42(3):665–681.

[Thomas, 2000] Thomas, L. C. (2000). A survey of credit and behavioural scoring: forecasting financial risk of lending to consumers. *International journal of forecasting*, 16(2):149–172.

Anexo

Código usado en $r-project$ para la obtención del modelo del ejemplo 2.9

```
rm(list=ls())
dir='C:/Users/Documents/Model/DATA'
setwd(dir)
#leemos los datos que están en credit.csv
datos <- read.csv('credit.csv',sep = ',', dec = '.', header = TRUE)
dim(datos)
names(datos)
str(datos)

#convertimos en categoricas las siguientes columnas
vec=c(2,4,5,7:13,15:21)
for (i in vec){
datos[,i]<-as.factor(datos[,i])
}

# Hacemos la regresión y miramos el resultado
regresion<- glm(Credit~+balance_credit_acc+duration+moral+verw+
hoehe+sparkont+beszeit+rate+famges+buerge+wohnzeit+verm+alter+
weitkred+wohn+bishkred+beruf+pers+telef+
gastarb,data=datos,family=binomial)

summary(regresion)

# Coeficientes del modelo
cbind(regresion$coefficients)

# Armamos el modelo final
datos$moral3<-ifelse(datos$moral==3,1,0)
datos$famges3<-ifelse(datos$famges==3,1,0)
datos$buerge3<-ifelse(datos$buerge==3,1,0)
datos$wohnzeit2<-ifelse(datos$wohnzeit==2,1,0)
datos$wohn2<-ifelse(datos$wohn==2,1,0)
datos$gastarb2<-ifelse(datos$gastarb==2,1,0)
datos$balance_credit_acc3<-ifelse(datos$balance_credit_acc==3,1,0)
datos$balance_credit_acc4<-ifelse(datos$balance_credit_acc==4,1,0)
datos$verw2<-ifelse(datos$verw==2,1,0)
datos$sparkont4<-ifelse(datos$sparkont==4,1,0)
datos$rate4<-ifelse(datos$rate==4,1,0)
datos$moral4<-ifelse(datos$moral==4,1,0)
datos$verw1<-ifelse(datos$verw==1,1,0)
```

```
datos$verw3<-ifelse(datos$verw==3,1,0)
datos$sparkont5<-ifelse(datos$sparkont==5,1,0)

datos$Z<-0.9735281824*datos$moral3+0.8215062770*datos$famges3+
0.9193954333*datos$buerge3-0.7443610334*datos$wohnzeit2+
0.4742480631*datos$wohn2+1.3988588259*datos$gastarb2+
0.9629526445*datos$balance_credit_acc3+
-0.0283268096*datos$duration+0.7287271411*datos$verw2+
-0.0001199527*datos$hoehe+1.4215951645*datos$sparkont4+
-0.9389707870*datos$rate4+1.7519424391*datos$balance_credit_acc4+
1.5121924512*datos$moral4+1.6290689480*datos$verw1+
0.8507062365*datos$verw3+0.9647584990*datos$sparkont5

# Probabilidad de bueno
datos$Pr<- exp(datos$Z)/(1+exp(datos$Z))
# Score
datos$score<- round(datos$Pr*1000)
#Resumen del Score
summary(datos$score)
# Min.  1st Qu.  Median  Mean    3rd Qu.    Max.
# 22.0    500.8   777.0   694.8   925.0     998.0

#Armamos una función que calcula la tabla de comportamiento
tabl_comp<-function(x,y,g=10){
#Para instalar el paquete Hmisc ejecutamos install.package('Hmisc')
library(Hmisc)

rango<-cut2(x,g=g)
ordena<-seq(1:length(unique(rango)))
tab_base<-cbind(ordena,table(factor(rango,exclude=NULL),y))
tab_base<-tab_base[order(-tab_base[,1]),2:3]

total<-margin.table(tab_base,1)
total_ac<-cumsum(total)
total_des<-cbind(ordena,cumsum(total[order(-total_ac)]))
total_des<-total_des[order(-total_des[,1]),2]

porc_total<-prop.table(total)*100
porc_ac_total<-cumsum(porc_total)
porc_des_total<-cbind(ordena,cumsum(porc_total[order(-porc_ac_total)]))
porc_des_total<-porc_des_total[order(-porc_des_total[,1]),2]

buenos<-tab_base[,1]
porc_buenos<-prop.table(buenos)*100
```

```
porc_ac_buenos<-cumsum(porc_buenos)
porc_des_buenos<-cbind(ordena,cumsum(porc_buenos[order(-porc_ac_buenos)]))
porc_des_buenos<-porc_des_buenos[order(-porc_des_buenos[,1]),2]

malos<-tab_base[,2]
malos_ac<-cumsum(malos)
malos_des<-cbind(ordena,cumsum(malos[order(-malos_ac)]))
malos_des<-malos_des[order(-malos_des[,1]),2]

porc_malos<-prop.table(malos)*100
porc_ac_malos<-cumsum(porc_malos)
porc_des_malos<-cbind(ordena,cumsum(porc_malos[order(-porc_ac_malos)]))
porc_des_malos<-porc_des_malos[order(-porc_des_malos[,1]),2]

tasa_malos<-tab_base[,2]/total*100
tasa_ac_malos<-malos_ac/total_ac*100
tasa_des_malos<-malos_des/total_des*100

ks<-abs(porc_ac_buenos-porc_ac_malos)

val_medio<-cbind(c(1:length(unique(rango))),tapply(x,factor(rango,
exclude=NULL),mean))
med_score<-mean(x)
val_medio<-val_medio[order(-val_medio[,1]),2]
tab_final<-cbind(val_medio,total,porc_total,porc_ac_total,porc_des_total,
buenos,porc_buenos,porc_ac_buenos,porc_des_buenos,malos,porc_malos,
porc_ac_malos,porc_des_malos,tasa_malos,tasa_ac_malos,ks)
colnames(tab_final)<-c('Score medio','Total','%Total','%AcTotal',
'%DesTotal','Bueno','%Bueno','%AcBueno','%DesBueno','Malo',
'%Malo','%AcMalo','%DesMalo','Tasa_Malo','TasaAcMalo','KS')

total_tab <- cbind(med_score,sum(tab_final[,2]),sum(tab_final[,3]),NA,NA,
sum(tab_final[,6]),sum(tab_final[,7]),NA,NA,sum(tab_final[,10]),
sum(tab_final[,11]),NA,NA,NA,NA,max(tab_final[,16]))

tab_final <- round(rbind(tab_final,total_tab),2)
tab_final
}
datos$target<-1-datos$Credit

tabl_comp(datos$score,datos$target)
```