Benemérita Universidad Autónoma de Puebla

FACULTAD DE CIENCIAS FÍSICO MATEMÁTICAS
LICENCIATURA EN ACTUARÍA

CREDIT SCORING: MÁQUINAS DE SOPORTE VECTORIAL
CONTRA LOS MODELOS CLÁSICOS DE REGRESIÓN

TESIS

QUE PARA OBTENER EL TÍTULO DE
LICENCIADO EN ACTUARÍA

PRESENTA
JULIO CESAR CORTES IZASMENDI

DIRECTOR DE TESIS
Dr. FRANCISCO SOLANO TAJONAR SANABRIA

PUEBLA, PUE. 15 DE OCTUBRE DE 2018

Dedicada con mucho cariño a:

Dios por las bendiciones recibidas día a día y por permitirme cumplir uno de mis objetivos.
A mis padres, por el apoyo, consejos y paciencia que me ofrecieron, principalmente a mi madre por sus enseñanzas y motivaciones, este logro también es de ellos.
A mi esposa por apoyarme desde el inicio de mi carrera.
Finalmente a mi hijo, Saúl Itzae, quién es el motivo por el cuál seguir adelante y jamás rendirme.

Agradecimientos

A Dios por las bendiciones recibidas y por haberme dado la oportunidad de cumplir una meta más en mi vida.

A mi madre por ser mi apoyo incondicional, quén ha guiado mis pasos hasta el día de hoy, a mi padre por demostrarme que con trabajo y dedicación se pueden cumplir las metas propuestas, a mi hermano quién también me ha apoyado en los momentos más dificiles.

A mi esposa por acompañarme en este camino y haber creído en mi, lo logramos. A mi hijo Saúl, esa personita especial que se convirtió en mi motor para seguir superandome día a día y a quién siempre amaré.

A mi director de tesis, Dr. Francisco Solano Tajonar Sanabria, por su tiempo, consejos y conocimientos compartidos tanto en los cursos como en la realización de este trabajo, le agradezco haber aceptado ser parte de este trabajo.

A mis sinodales, Dr. Hugo Adán Cruz Suárez, Dr. Fernando Velasco Luna por sus comentarios y observaciones realizadas y a la M. C. Brenda Zavala López por ser una excelente maestra y compartir incondicionalmente sus conocimientos, particularmente le agradezco haberme dado la oportunidad de trabajar con ella mi servicio social, una experiencia que me ayudo a reforzar mis conocimientos. A cada uno de ellos agradezco su disponibilidad y tiempo para la revisión de este trabajo.

A todos los profesores, tanto matemáticos como actuarios, que durante mis años de estudio tuve la dicha de recibir sus conocimientos y me ayudaron a desarrollarme intelectual y personalmente.

A mis amigos y compañeros que conocí en esta bella etapa de mi vida, principalmente a mi amigo Alejandro Marcos Analco, por esas horas de estudio y apoyo para salir siempre adelante, por esas alegrías vividas en cada curso compartido con cada uno de ustedes y sobretodo por esos consejos que me han ayudado profesional y personalmente.

Introducción

El riesgo de crédito es el riesgo de pérdida financiera resultante del incumplimiento o de la calidad crediticia de los emisores de valores, deudores o contrapartes (por ejemplo, en contratos de reaseguro) e intermediarios, a los que la empresa generalmente esta expuesta.

Para una empresa financiera general, el riesgo de crédito es "el riesgo de no recibir los pagos prometidos sobre las inversiones pendientes tales como préstamos y bonos, debido al fallo del prestatario".

Algunos objetivos de interés en las instituciones financieras es el poder captar un mayor número de clientes en su cartera de inversión y/o ahorro, y a su vez poder realizar préstamos y obtener una ganancia en la cobranza de intereses. Lo ideal en éste negocio es que a cualquier persona que se le otorgue un crédito, tenga la capacidad económica de solventar el monto prestado, más los respectivos interéses, dentro del periodo marcado por la institución; sin embargo, ésto no siempre sucede por diversos factores que afectan a la posibilidad de pagar, dígase por desempleo, olvido de pago, situación económica desfavorable, entre otros.

En las instituciones financieras, por ejemplo, se presenta constantemente el concepto de riesgo en varios productos ofrecidos por dicha institución, ya sea en inversiones, créditos hipotecarios, personales o de consumo. En éste ultimo producto supongamos el siguiente escenario: Una persona con características descritas por el conjunto X solicita un crédito de consumo, popularmente llamado, una tarjeta de crédito, ante ésta solicitud el banco tiene cierta incertidumbre en poder otorgar el crédito, pues existen preguntas que no son fáciles de responder en primer instancia. Una de ésas preguntas puede ser si el cliente será un cliente "bueno", o un cliente "malo"; donde la definición de bueno y malo en éste contexto no tiene que ver con la conducta moral de la persona si no mas bien con su conducta financiera ante un préstamo, es decir, con cliente "bueno" se refiere a que el cliente va a pagar el crédito y siempre busca tener un historial limpio ante el banco, entre éste segmento se pueden encontrar varios tipos de clasificación, pues puede

haber personas con un perfecto historial crediticio, es decir, nunca en su vida financiera han tenido atraso de pago ni pago de intereses moratorios, es más paga sus deudas de una manera anticipada, después podemos encontrar aquellos clientes que son regularmente buenos, es decir, que casi siempre pagan a tiempo y en algún momento tienen una deuda pendiente ante la institución y luego estan al corriente en sus pagos en el periodo siguiente. Y así sucesivamente se pueden ir subclasificando los clientes hasta un límite que puede establecer un administrador de riesgos.
De la misma manera ocurre para el caso de clientes malos pues, puede haber clientes malos, ligeramente malos y verdaderamente malos (Morosos), de los cuáles afectan a la cartera de crédito de la institución bancaria y de los cuáles se desea evitar al tener una nueva solicitud de crédito.
Otra pregunta que quizá pueda hacerse un analista o administrador de riesgo es la siguiente: Dado que se le ha otorgado el crédito ¿Cuál es la probabilidad de que ésta persona pague la deuda durante la vida del crédito?. Muchas veces en el ámbito de los créditos generalmente se habla de la probabilidad de incumplimiento a veces abreviada (PD) por sus siglas en inglés (Probability of Default) cuya forma de calcular requiere de aquel análisis económico-estadístico para poder dar una respuesta bajo cierto nivel de confianza.
La consecuencia de tener una cartera con clientes morosos, en su mayoría, puede llevar a tener un nivel de riesgo alto para la institución y en algunos casos llegar a la ruina. Para prevenir dicho escenario se hace uso de la probabilidad, estadística y tecnología para construir modelos matemáticos en los cuáles se pueda obtener una estimación de la probabilidad de incumplimiento y/o calcular un "score" que vaya segmentando a las solicitudes de financiamiento y así tomar decisiones correctas en el otorgamiento de éste.
Basándose en el tratado de basilea II, publicado inicialmente en 2004, dicta en resumen, que una institución financiera debe contar con estrategías de control para el cálculo de riesgos y su supervisión, con lo que el *credit scoring* se adapta fácilmente ante ésta necesidad.

El término *credit scoring* recoge todos los métodos estadísticos que se utilizan para determinar el riesgo asociado con un posible deudor, en otras palabras, estima probabilidades de fallo y ordena a los deudores y solicitantes de financiamiento en función de su riesgo de incumplimiento. Los modelos de *credit scoring*, generalmente se asocian con la minería de datos, que gracias al avance computacional se puede trabajar de manera más eficiente con el banco de datos que puede poseer una institución financiera.

La minería de datos inició durante la década de los 80 y emergió de gran manera en la década de los 90, una definición que se le puede dar es la exploración y análisis de datos con el objetivo de visualizar patrones y relaciones que son difíciles de percibir a simple vista. Es un campo multidisciplinario que incluye diversas áreas, algunas de ellas son : Tecnología en bases de datos, inteligencia artificial y la estadística donde se ocupan herramientas importantes tales como el análisis de regresión (lineal, logística, etc), prueba de hipótesis, intervalos de confianza, etc.

Los algoritmos de minería de datos se clasifican en dos grandes categorías: supervisados o predictivos y no supervisados.
Los algoritmos supervisados o predictivos predicen el valor de una característica de interés de un conjunto de datos. A partir de datos cuyo atributo se conoce, se induce una relación entre dicha etiqueta y otra serie de atributos. Esas relaciones sirven para realizar la predicción en datos cuyo atributo es desconocido.
Algunos métodos que entran dentro de éste bloque son: Máquinas de soporte vectorial (Support Vector Machine), redes neuronales, árboles de decisión, método de consenso (Bagging) y en partícular Bosques aleatorios (Random Forest) y método de potenciación.
Cuando el modelo predictivo que se construyó no da una solución predictiva significativa es conveniente recurrir a modelos de aprendizaje no supervisado, ya que éstos modelos ayudan a descubrir patrones y tendencias en los datos. Algunos de los métodos incluidos en ésta categoría son: segmentación, clustering, análisis de componentes principales, e ntre otros.
La finalidad de éste trabajo es el de comparar las metodologías estudiadas en el análisis econométrico, tales como: Regresión lineal general, Modelos probit, logit, contra la metodología que ofrece la minería de datos cuyas técnicas fueron mencionadas anteriormente. Para cumplir dicho objetivo se utilizará una base de datos que contiene información sobre la aprobación y desaprobación de una tarjeta de crédito de una institución financiera.

La tesis está estructurada en cuatro capítulos. En el primer capítulo se exponen los conceptos relacionados con: riesgo, administración de riesgos, riésgo de crédito y riesgos de crédito. De igual forma se exponen algunos conceptos del instrumento financiero con mayor demanda en una institución financiera, es decir, las tarjetas de crédito, se planea explicar sus características,

su utilización y revisar el comportamiento que llegan a tener los clientes cuando tienen aprobado un crédito con algún banco. En el capítulo dos se revisan los modelos paramétricos (probit y logit) y no paramétricos (máquinas de soporte vectorial) a utilizar, para los modelos no paramétricos se da una exposición más detallada ya que son temas no vistos en la licenciatura. En el capítulo tres se da un breve detalle del concepto *credit scoring* y su utilización. En el capítulo cuatro, se muestra la aplicación de los modelos trabajados con la base de datos ya mencionados. También se expone la utilización de softwares estadísticos para nuestro análisis de datos, el primero de ellos es STATA y el segundo R.
Este trabajo finaliza con las conclusiones obtenidas en el desarrollo del mismo.

Índice general

CREDIT SCORING: MÁQUINAS DE SOPORTE VECTORIAL CONTRA LOS MODELOS CLÁSICOS DE REGRESIÓN.

Julio Cesar Cortés Izasmendi

12 de octubre de 2018

Capítulo 1

Definiciones Básicas

1.1. Riesgo

La palabra riesgo proviene del latín *risicare* que significa "atreverse". En muchas áreas multidisciplinarias se habla del riesgo y existe siempre una aversión ante éste [2]. En la vida cotidiana puede presentarse el riesgo en diferentes formas, ya sea en materia de salud, económica-financiera o a veces en relaciones sociales. En finanzas, el concepto de riesgo se relaciona con la posibilidad de que ocurra un escenario en el cuál se tengan pérdidas para los participantes en los mercados financieros.

1.2. Riesgo bancario

De lo anterior, podemos pensar que el riesgo bancario no es más que la posibilidad de que ocurra un evento cuyo efecto sea negativo para los bancos, es decir, que se tengan pérdidas no deseadas, afectando la estabilidad económica de la institución.
La intermediación bancaria es el proceso por el cual una empresa o varias se especializan en captar depósitos del público para proceder a prestarlos. En México, la banca inició en 1864, a partir del establecimiento en la Ciudad de México de una sucursal de un banco británico: *The Bank of London, Mexico and South America* [15]. Desde entonces el crecimiento de la banca de México crecio de manera rápida, al igual que la necesidad de controlar los riesgos del banco para evitar eventos que llevarán a la bancarrota a la institución. Pues dada la responsabilidad de manejar los ahorros del público

tiene como objetivo principal la protección de dichos activos y evitar tener mala reputación.

Los principales factores que determinan el Riesgo en los bancos se dividen en dos clases [9]:

- **Factores internos:** Que solo dependen de la administración propia de los ejecutivos de cada institución.
- **Factores externos:** Que solo dependen de variables externas a la administración de la institución tales como: situación económica, la tasa de inflación, apreciación o depreciación de la moneda nacional, desastres naturales, entre otros.

1.3. Clasificación de riesgos

Cuando hablamos del riesgo bancario, estamos englobando todos aquellos riesgos que llevan las actividades del banco, ya sea en inversiones o en créditos, por mencionar algunas, pero cualquiera que sea la actividad del banco los riesgos que se enfrentan tienden a ser los mismos.
Existen muchos tipos de riesgo los cuáles los podemos clasificar en:

- **Riesgos de negocio /operativos:** Se puede definir como aquella pérdida obtenida tras la existencia de fallos o falta de adecuación en los procesos de producción, procesos que afectan al personal, procesos de sistemas, procesos de acontecimientos externos que llegan afectar a la empresa. Este tipo de riesgo es un elemento emergente que requiere mayor atención.
- **Riesgo financiero:** La identificación de este tipo de riesgo se debe a la necesidad de seguir atentamente la trayectoria de los capitales aportados dentro del mecanismo empresarial. Este tipo de Riesgo se subdivide en varios tipos de riesgo, algunos de ellos son:
  - **Riesgo de crédito o solvencia:** Se puede definir como las pérdidas obtenidas por el incumplimiento de pago u obligaciones de crédito a las cuáles una persona (física o moral) se comprometío a hacer.

  - **Riesgo de mercado:** Posibilidad de tener pérdidas en los rendimientos esperados de activos invertidos ante un cambio en las condiciones de mercado. Por ejemplo: Variaciones en los precios que se esperaban, en los tipos de interés y en tipos de cambio.
  - **Riesgo de liquidez:** Posibilidad de pérdidas del valor de un activo ante la venta de éste . También se puede definir como la posibilidad de incurrir en pérdidas por no disponer de los recursos suficientes para cumplir con las obligaciones asumidas y no poder desarrollar el negocio en las condiciones previstas.

- **Riesgo de entorno:** Esta ligado a eventos externos que no dependen del control o administración de la empresa. Se subdivide en
  - **Riesgo legal**
  - **Riesgo estratégico**
  - **Riesgo reputacional**
  - **Riesgo país**. [3]

1.4. Créditos

Se entiende como crédito al compromiso realizado entre una persona (física o moral) que otorga capacidad de comprar por adelantado al deudor, que también puede ser física o moral [10]. Un caso particular ante esta definición puede ser los créditos hipotecarios, que otorga el derecho de adquirir una casa con el dinero prestado de una persona o institución bancaria. Ante este préstamo se debe pagar por el otorgamiento de éste, es decir, los intereses, los cuales son la ganancia de que la persona o institución haya limitado el uso de su dinero en la compra de dicha casa.

Ante el acuerdo del crédito se estipulan ciertas condiciones que hacen posible llevar a cabo el trato del préstamo, estas condiciones tiene que ver con los plazos para terminar de pagar la deuda, los montos a pagar, el tipo de interés, etc.

En el desarrollo de éste trabajo vamos a definir al prestamista como una institución bancaria y al prestatario como personas físicas (clientes).

Cuando un cliente tiene la necesidad de comprar un bien o servicio y no cuenta con el capital para solventar dicho valor, se recurre a un prestamista y así, ir pagando la deuda con pagos "más accesibles" que el costo total del bien. Ante ésta solicitud el prestatario tiene la necesidad de evaluar las características de su cliente para poder determinar si otorgar o no el crédito. Estas características o requisitos son propios de la política de cada institución, y aunque existen variables comunes, se siguen desarrollando formas para saber que variables utilizar.

Una vez que los clientes han sido aceptados por la institución bancaria se deben realizar documentos en donde se establezcan la tasa de interés acordada, monto de crédito, plazos y modalidad de pago.

Cuando los clientes no cumplen con la obligación adquirida , el prestamista los empieza a clasificar como clientes morosos, según sus políticas. En algunos casos se registran este tipo de clientes con el fin de que sirva como referencia del comportamiento de los clientes en cada uno de sus créditos adquiridos. En México la información es enviada al buró de crédito, es enviada mensualmente y lleva el registro hasta por un periodo de 24 meses.

1.4.1. Tipos de crédito

- **Créditos de consumo o Créditos comerciales:** Surgen con el fin de cubrir necesidades de consumo de los clientes que no tienen una capacidad económica actual para el cubrimiento total del costo en efectivo. Es decir, se utiliza para el consumo de bienes materiales.

- **Créditos empresariales:** Para financiar las necesidades del capital de trabajo o activos que ayuden a la operación y producción de una empresa.

- **Créditos bancarios:** Son otorgados por instituciones crediticias, típicamente los bancos, mediante la celebración de un contrato. El cliente cuenta con un dinero a disposición y solo paga intereses por la cantidad que utiliza.

1.5. Tarjetas de crédito

La tarjeta de crédito es muy utilizada actualmente por la sociedad y se ha convertido en un instrumento con mayor demanda pero también con mayor control. En la vida actual una tarjeta de crédito puede ayudarnos a cubrir ciertas necesidades que quizá no puedan ser pagadas en efectivo en el momento de adquirir el bien o servicio. Es decir, nos da la capacidad de adqurir objetos sin la necesidad de desembolsar de nuestro capital en ése momento y pagar después. Todo esto es posible siempre y cuando el cliente pague los respectivos interes, por haber utilizado el dinero que pertenece, en este caso, a los bancos. Su uso va desde la compra de alimentos, ropa, etc; hasta pagar viajes, habitación en hoteles, gasolina entre otras actividades diarias del ser humano.

Pues bien, la historia de la tarjeta de crédito inicia desde el año 1920 cuando en Estado Unidos, la empresa Western Union, comienza con la entrega de placas de metal, a grupos selectos de sus clientes, que les permitía identificarse y diferir sus pagos, dada ésa innovadora idea, hoteles, tiendas departamentales y empresas de ferrocarriles la copiaron. Con el paso del tiempo fue evolucionando el concepto, el material y la forma de la tarjeta de crédito, pues con la evolución de la tecnología aparecieron tarjetas de crédito con banda magnética, tecnología creada por IBM en el año de 1960, y fueron utilizadas por primera vez en el transporte público de Londres. En el año de 1968, México empezaba con el uso de este instrumento, se llamaba *Bancomático*(Ver figura 1.1) y fue otorgada por *BANAMEX* en afiliación con *Interbank*, hoy *MasterCard*. Al día de hoy, debido al avance tecnológico, se pueden encontrar tarjetas de crédito con un chip integrado, ésto con el fin de evitar fraudes y/o dificultades de pago que se tuvieron con la tecnología pasada.

Si bien la historia de la tarjeta de crédito es un poco extensa, la definición propia de tarjeta de credito es la siguiente:

- **Tarjeta de crédito:** Tarjeta emitida por una entidad bancaria que permite realizar ciertas operaciones desde un cajero automático y la compra de bienes y servicios a crédito, generalmente es de plástico y tiene un microchip o banda magnética en una de sus caras.

Figura 1.1: Primer tarjeta de crédito en México.

1.5.1. Características de una tarjeta de crédito

Como todo instrumento financiero, las tarjetas de crédito tienen sus características para su buen uso y funcionamiento tanto para el cliente como para el emisor del crédito, dichas características son [12]:

- **Línea de crédito:** El banco como emisor del crédito concede al cliente, mediante al acuerdo establecido, una línea de crédito revolvente, es decir, una vez pagando la deuda se vuelve a obtener el dinero prestado para, volver a ocuparlo, hasta un límite de crédito determinado por la misma institución.
- **Periodo:** Es la fecha de inicio y fin que comprende el ciclo en el cuál puede ocuparse la tarjeta. Regularmente oscila entre los 30 y 31 días.
- **Fecha de corte:** Es el día del mes en que termina e inicia un nuevo periodo de registro del uso del plástico.
- **Fecha límite de pago:** Es la fecha en la cuál se tiene que realizar el pago para no caer en morosidad. Generalmente son 20 días naturales a partir de la fecha de corte.
- **Pago mínimo:** Es la cantidad mínima a pagar al banco para no caer en morosidad.
- **Pago para no generar intereses:** Es un monto mínimo que se debe liquidar puntualmente y así evitar el pago de intereses (incluye los pagos mensuales correspondientes a promociones a meses sin intereses).
- **Costo Anual Total (CAT):**De [1] se tiene que es una medida estandarizada del costo del financiamiento, expresado en términos porcentuales anuales que incorpora la totalidad de los costos y gastos inherentes de los créditos que otorgan las instituciones. Es decir, es un

indicador que incorpora en una sola cifra todos los costos relevantes, (intereses, las comisiones y el plazo de pago), en que se incurre al contratar un crédito.

1.6. Cartera de crédito

En [9] se define a la cartera de crédito de un Banco, como el conjunto de prestámos que ha otorgado a sus clientes, y por lo mismo dicha cartera es considerada como parte del Activo de la institución.

Dada la definición anterior podemos observar entonces la importancia que tienen las instituciones financieras en cuidar y administrar correctamente su cartera de crédito, pues al representar un activo, debe generar "rendimientos" o "beneficios'.

1.7. Riesgo de crédito y contraparte

El Banco de México en [2], define la existencia del riesgo de contraparte cuando hay la posibilidad de que una de las partes de un contrato financiero sea incapaz de cumplir con las obligaciones financieras adquiridas, haciendo que la otra parte del contrato tenga pérdidas. El riesgo de cŕedito es un caso particular, cuando el contrato es un crédito, y el deudor no tiene la capacidad económica para saldar la deuda.

Gracias a este tipo de riesgo es donde nace el interés por estudiar modelos probabilísticos-econométricos que midan tal riesgo y se auxilian con el cálculo de las probabilidades de incumplimiento y/o correlaciones entre incumplimientos.

- **Probabilidad de incumplimiento (PD):** Esta medida, explica que tan probable es que un acreditado deje de cumplir con sus obligaciones contractuales.

- **Correlación entre incumplimientos:** Mide la dependencia o grado de asociación entre el comportamiento crediticio de dos deudores.

1.8. Clientes buenos y malos

Antes de llegar a la definición de clientes buenos y malos, es necesario identificar la clasificación de la cartera de crédito en una institución bancaria, es decir, dependiendo del comportamiento de cada cliente estudiar a que clasificación pertenece y después llevarlo a calificarlo como bueno o malo.

1.8.1. Clasificación de la cartera

Realizar una clasificación dentro de una cartera de crédito es de mucha importancia, ya que proporciona información sobre el comportamiento de los clientes según sus pagos, y así poder detectar aquellos clientes morosos. Puede haber distintas clasificaciones dadas las políticas de cada banco pero en general suele haber una métrica estándar. Para este trabajo se utilizará la clasificación expuesta en [10].

- *Current*: Cuando el cliente cumple con sus obligaciones entre la fecha de corte y la fecha límite de pago; no se le cobran intereses moratorios sobre su saldo .

- *Almost current:* Se pueden encontrar clientes en esta clasificación que pagan únicamente el monto mínimo entre la fecha de corte y la fecha límite de pago, el monto neto de su deuda se carga para el siguiente periodo. En ésta clasificación todavía no se puede decir que es un cliente moroso pero tampoco es un cliente "current".

- *Prevent*: El cliente no paga su deuda ni tampoco realiza el pago mínimo en la fecha límite de pago, se empieza a localizar al cliente para recodarle la deuda pendiente, a esto se le llaman gastos de cobranza y se realiza entre la fecha límite de pago y la próxima fecha de corte. Aún no se le cobran intereses moratorios. Si el cliente después de localizarlo paga su deuda entonces pasa al estatus *current*.

- *Bucket 1*: Si durante el periodo de corte se paga menos del mínimo, no se considera este como cumplimiento de la obligación, por lo que avanza a un pago vencido. Es en ésta clasificación donde se le cargan intereses moratorios al monto de la deuda actual y toma el estatus de moroso. En la fecha de facturación se calcula el nuevo saldo y se empieza a contar los días de retraso. Este grupo se envía al apartado de clientes

que tienen de 1 a 29 días de moratoria. Esta información es enviada al buró de crédito. Las mismas acciones son aplicadas a un cliente que no realizó ningún pago.

Así es como se va clasificando los clientes dependiendo de su tiempo de atraso y se puede resumir en la siguiente tabla:

Clasificación	**Tiempo de mora**
Bucket 0 (B_0)	0 días (al corriente)
Bucket 1 (B_1)	1 a 29 días
Bucket 2 (B_2)	30 a 59 días
Bucket 3 (B_3)	60 a 89 días
...	...
Bucket 6 (B_6)	150 a 179 días
Bucket 7 (B_7)	Mayor o igual a 180 días

1.9. Definición de cliente bueno y malo

La clasificación de los clientes depende fuertemente de las políticas de la institución financiera y puede modificarse durante la operación de ésta. A continuación se ilustra una clasificación general que tienen las instituciones bancarias cuando éstas clasifican en función de su tiempo de mora:

Cliente bueno: Es aquella persona que generalmente no tiene adeudo con el banco y siempre o casi siempre está al corriente de sus pagos o por lo menos paga antes de rebasar los 60 días de atraso. Por tanto, podemos clasificar como clientes buenos aquellos que se encuentran en B_0 hasta B_3 .

Cliente intermedio: Regularmente, son aquellos clientes que necesitan ser observados por más tiempo para poder detectar la tendencia del comportamiento en sus pagos y así llegar a una clasificación correcta.

Cliente malo: En este apartado se pueden encontrar aquellas personas que generan pérdidas económicas al banco. Estos clientes no pagaron su cuenta, a pesar de usar técnicas de cobranza. Comúnmente encontramos personas cuya etiqueta es B_4 en adelante. Regularmente una persona que cae en B_7 difícilmente sale de dicho estado para poder convertirse en cliente "current".

En la siguiente Figura se ilustran los estados en los que cualquier cliente está propenso a estar dadas las condiciones de su economía u otros factores:

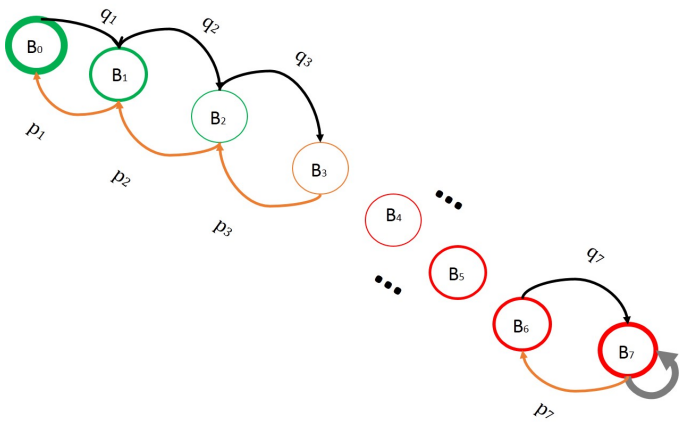


Figura 1.2: Posibles estados de los clientes.

A pesar de que la forma de clasificación expuesta anteriormente se utiliza comúnmente en los bancos existen propuestas capaces de clasificar a los clientes de acuerdo a formas no convencionales, por ejemplo en [11] menciona los modelos de mezclas Poisson que propone una clasificación más robusta, inclusive en [6] se ilustra un enfoque diferente del scoring de crédito de un banco, pues en lugar de predecir si el cliente es bueno o malo predice el número de incumplimientos en un futuro cercano, dando así otra opción para la administración de riesgos de un banco.

Capítulo 2

Modelos paramétricos y no paramétricos para el análisis del riesgo de crédito

En este capítulo se establecen los conceptos necesarios para la construcción de los modelos usados en el presente trabajo.

2.1. Técnicas paramétricas

La estadística paramétrica tiene como objetivo la construcción y uso de procedimientos estadísticos basados en las distribuciones de datos reales, dichas distribuciones estan conformadas por un número finito de parámetros, los cuales con la ayuda de la estadística inferencial se pueden estimar. Existen varios modelos paramétricos en el área de la probabilidad y la estadística, en particular, se está interesado en estudiar aquellos modelos que ayuden a modelar la clasificación de un cliente, ya sea bueno o malo, dependiendo de sus características; que son los modelos "probit" y "logit". Estos modelos tienen como objetivo modelar, la respuesta, binaria o múltiple, de una variable dependiente de una o varias variables independientes. Una de las ventajas de este tipo de modelos es que con ellos se puede calcular la probabilidad de impago para cada cliente y por ello se tiene el interés en dicho trabajo ya que con el cálculo de dichas probabilidades se pretende construir un scoring para el histórico de clientes actual a futuro.

2.1.1. Modelos probit y logit

Las variables binarias en un modelo de regresión son importantes y muy comunes cuando se trata de abordar problemas del tipo clasificación ya sea de clientes, productos, o cualquier otra variable de interés. En éste caso la modelación con regresión logística es una de las herramientas del *credit scorig*, donde se desea ir clasificando e identificando que tipo de cliente se le puede otorgar el crédito y a quien no, otra de las ventajas de trabajar con este modelo es que nos dicta probabilidades dependiendo del cambio de valores de las variables independientes y así explorar estos cambios bajo la teoría de "Margin effect".

Las variables binarias típicamente son codificadas por 0 para un resultado negativo y 1 como un resultado positivo [7], y los ejemplos que son modelados con dichas variables son fácilmente de pensar y están presentes casi en todo momento de la vida, un ejemplo en el ámbito social puede ser en tratar de determinar si una persona votó por "X" o "Y" candidato, formalmente es conocer la probabilidad que tienen los candidatos de ser elegidos por la persona y así, tratar de predecir el voto que realizará dicha persona.

Existen dos modelos que frecuentemente se utilizan para estudiar este tipo de variables, el modelo probit y logit, que entran en la clasificación de modelos no lineales y gracias a esto podemos observar que el cambio de magnitud en la probablidad de que ocurra un evento cuando cambia una de las variables independientes, esta depende también del nivel en la que se encuentran todas las variables independientes, situación que se puede analizar posteriormente en la sección de "Margin Effects".

Para abordar estos modelos, se puede hacer de dos maneras, una es considerando una variable latente que se relaciona con las variables independientes y mediante una ecuación a trozos se define el valor de la variable binaria y dependiendo de los valores que llegará a tener la variable latente, es decir:

$$y_i = \begin{cases} 1, & \text{si} \quad y_i^* > 0, \\ \\ 0, & \text{si} \quad y_i^* \leq 0. \end{cases} \tag{2.1}$$

La otra forma, es usar un cociente de probabilidades llamadas "odds" y linealizar con la función de logaritmo natural, con esto se busca limitar

el rango de probabilidades en el intervalo $[0, 1]$. En este trabajo, vamos a desarrollar la construcción del modelo bajo el esquema de una variable latente.
Supongamos que existe una variable y^* cuyo rango es de -∞ a ∞, cuya relación con las variables independientes está dada por:

$$y^* = \mathbf{x}_i\beta + \epsilon_i \tag{2.2}$$

donde i indica la observación i-ésima,β representa los parámetros a estimar y ϵ es el error aleatorio. La relación que hay con la variable binaria y está descrita por la expresión 2.1 y la idea de considerar dicha variable es bajo el supuesto de un efecto subyacente que genera el estado de la variable binaria Y. Si suponemos que la dimensión de X es 1, entonces la ecuación 2.2 queda descrita de la siguiente manera

$$y^* = \alpha + \beta x_i + \epsilon_i \tag{2.3}$$

donde α representa el intercepto en el eje y y β la pendiente de la recta, así en $\mathbb{R}^2$ podemos visualizar la idea en la siguiente gráfica:
Si desarrollamos

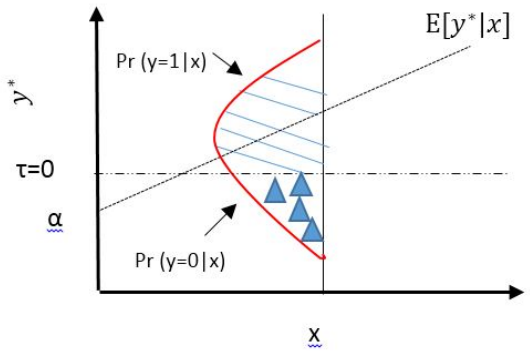


Figura 2.1: Relación entre y^* y $Pr(y = 1|x)$

$$Pr(y = 1|x) = Pr(y^* > 0|x) = Pr(\epsilon > -[\alpha + \beta x_i]|x) \tag{2.4}$$

la última igualdad ocurre por la ecuación 2.3.

Como se observa en la última expresión, $Pr(y = 1|x)$ depende de la distribución de ϵ, de esto se obtienen los modelos probit y logit mencionados anteriormente:

Modelo Probit

En este modelo asumimos la distribución de ϵ como una distribución normal con media igual a cero y $\text{Var}(\epsilon) = 1$, por tanto,

$$Pr(y = 1|x) = \int_{-\infty}^{\alpha+\beta x} \frac{1}{\sqrt{2\pi}} \exp(\frac{-t^2}{2})dt. \tag{2.5}$$

Modelo Logit

Para este caso, se asume que la distribución de ϵ es como una distribución logística, con media igual a 0 y $\text{Var}(\epsilon) = \frac{\pi^2}{3}$, teniendo así,

$$Pr(y = 1|x) = \frac{\exp(\alpha + \beta x)}{1 + \exp(\alpha + \beta x)}. \tag{2.6}$$

2.1.2. Validación del método de clasificación

En data mining, para validar la eficacia del método de clasificación utilizado, se utiliza una base de datos en donde se sabe a qué clasificación pertenece cada registro pero que no fue utilizado para el cálculo de los parámetros del modelo. Se clasifican estos registros de prueba (test) con el modelo estimado y luego se van contando cuántos de estos registros quedaron bien clasificados y cuáles no, esta puede ser una medida de bondad de ajuste llamada matriz de error, que va indicando que tan bueno puede ser la capacidad de predicción del modelo.[10]

2.2. Técnicas no paramétricas

2.2.1. Máquinas de soporte vectorial

Las máquinas de soporte vectorial pertenecen a la familia de los clasificadores lineales puesto que inducen separadores lineales o hiperplanos

cuando el número de características a considerar es mayor o igual a tres.

Recordemos, que todo hiperplano en $\mathbb{R}^n$, se puede expresar como:

$$g(\overrightarrow{x}) =< \overrightarrow{w}, \overrightarrow{x} > +b \tag{2.7}$$

donde $\overrightarrow{w} \in \mathbb{R}^n$, $b \in \mathbb{R}$ y $< ., . >$ indica el producto interno habitual de $\mathbb{R}^n$. Si deseamos una función que clasifique en forma binaria a nuestros datos, se puede obtener definiendo la siguiente función:

$$f(x) = signo(h(x)) = \begin{cases} +1, & si \quad x \geq 0, \\ \\ -1, & si \quad x \leq 0. \end{cases} \tag{2.8}$$

En términos de clasificación se interpreta a $\overrightarrow{x} \in \mathbb{R}^n$ como la representación vectorial de los datos, con una componente real por cada atributo, y al vector $\overrightarrow{w}$ se suele identificar como un vector de pesos. Dicho vector contiene un peso para cada atributo, indicando así la contribución que cada atributo aporta en la clasificación. Finalmente, se denomina al vector b como el sesgo y es quien define el umbral de decisión.

La idea de SVM (Support Vector Machine), de margen máximo consiste en seleccionar un hiperplano que maximiza la distancia mínima entre los datos y el hiperplano, además, sólo considera los puntos que están en la frontera de la región de decisión, dichos puntos tienen una cierta incertidumbre de saber a que clase pertenece un registro, a estos puntos o vectores se les conoce como vectores de soporte.

Cabe resaltar que se considera este tipo de modelo no paramétrico, ya que a nivel práctico ha demostrado tener una muy buena capacidad de generalización en numerosos problemas reales.

Desarrollaremos la teoría en dos fases, es decir, primero analizaremos el caso en que los datos son linealmente separables y trataremos de encontrar los parámetros $\overrightarrow{w}$ y b de tal forma que se encuentra el margen máximo posible entre las dos clases, después hablaremos del caso en que los datos no son linealmente separables y los métodos que existen para tratar de encontrar una solución al problema.

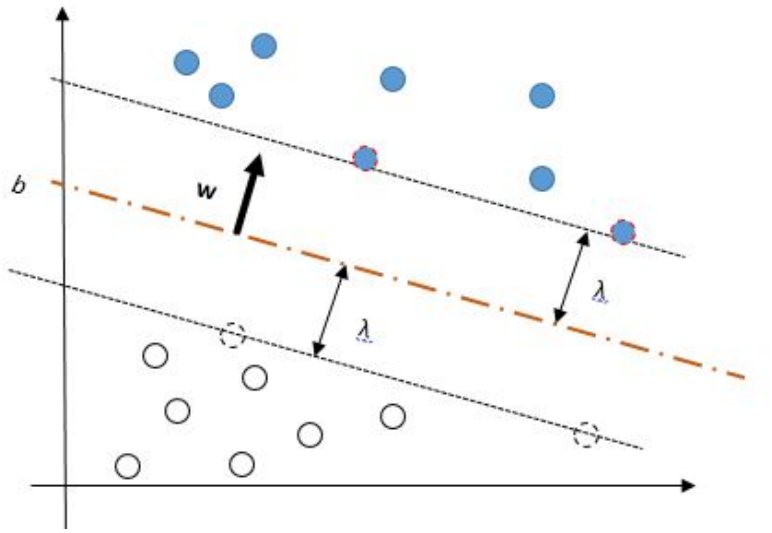


Figura 2.2: Hiperplano (w, b) equidistantes a dos clases, margen geométrico (λ) y vectores de soporte (círculos punteados)

2.2.2. SVM lineales con margen máximo

Es el modelo más sencillo de SVM (Support Vector Machine), pero con condiciones de aplicabilidad más restringidas, pues la hipótesis fundamental es que el conjunto de datos es linealmente separable, sin embargo, contiene ideas subyacentes en la teoría de SVM y es base para todas las demás.

Definición 2.1. *Un concepto (dato) es* ***separable linealmente*** *si existe una función lineal (recta, plano, hiperplano) que separe las dos clases nítidamente.*

Dada la definición anterior, supongamos que tenemos un conjunto de datos linealmente separable, es decir,

$$\forall \overrightarrow{x} \in X \quad \exists h : X \to \mathbb{R} : (h(x) > 0 \text{ si } y = +1 \vee \ h(x) < 0 \text{ si } y = -1)$$

donde y es la clase a la que pertenece el registro, $\overrightarrow{x}$ es aquel vector que representa las características que tiene un registro en nuestra base de datos y X es nuestra base o conjunto de datos.
Considerando el problema de clasificación binaria que contiene N datos de entrenamiento, con cada dato denotado por la tupla $(\overrightarrow{x_i}, y_i)$ para $i =$

$1, 2, ..., N$, donde $\overrightarrow{x_i} = (x_{i1}, x_{i2}, ..., x_{id})^T$ corresponde al conjunto de atributos (variables independientes) del $i - ésimo$ dato. Por convención, sea $y_i \in \{+1, -1\}$, la variable que denota la clasificación perteneciente del registro.

El límite de decisión (hiperplano) de un clasificador lineal puede ser expresado de la siguiente forma

$$\overrightarrow{w} \cdot \overrightarrow{x} + b = 0 \tag{2.9}$$

donde $\overrightarrow{w}$ y b son los parámetros del modelo.

Es fácil observar que cualquier dato que se encuentre localizado sobre el hiperplano de separación satisface la ecuación (2.9). Es decir, si $\overrightarrow{x}_a$ y $\overrightarrow{x}_b$ se encuentran sobre el hiperplano entonces:

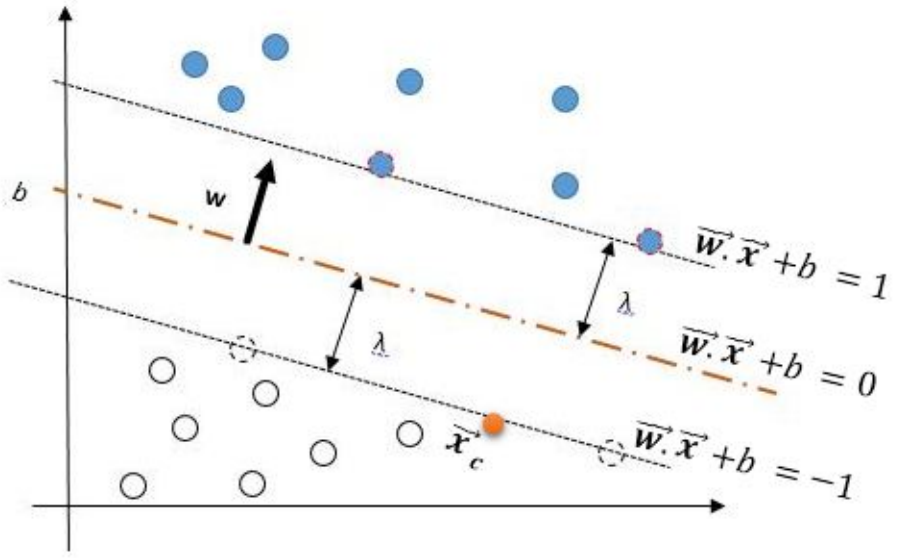


Figura 2.3: Representación de los hiperplanos (w, b) y margen geométrico (λ)

$$\begin{gathered}\overrightarrow{w} \cdot \overrightarrow{x}_a + b = 0 \\ \wedge\ \overrightarrow{w} \cdot \overrightarrow{x_b} + b = 0 \\ \Rightarrow\ \overrightarrow{w} \cdot (\overrightarrow{x_b} - \overrightarrow{x_a}) = 0\end{gathered}$$

Lo que indica que la dirección de $\overrightarrow{w}$ es perpendicular al del hiperplano de separación (Figura 2.3).

Por otro lado, es notable que cuando cualquier punto $\overrightarrow{x_s}$ esta por encima del hiperplano de separación, entonces : $\overrightarrow{w} \cdot \overrightarrow{x_s} + b = k$, para algún $k > 0$, y cuando $\overrightarrow{x_s}$ esta por debajo entonces $\overrightarrow{x_s} + b = k'$, para algún $k' < 0$. Con este razonamiento podemos visualizar a la variable Y como:

$$\forall\ \overrightarrow{z} \in X$$

$$y = \begin{cases} +1, & si \quad \overrightarrow{w} \cdot \overrightarrow{z} + b > 0, \\ \\ -1, & si \quad \overrightarrow{w} \cdot \overrightarrow{z} + b < 0. \end{cases} \tag{2.10}$$

donde X es el conjunto de datos.

En el siguiente teorema, se demuestra como se calcula la distancia λ mejor conocida como el margen geométrico ilustrado en la Figura 2.3, lo cual nos ayudará a plantear el objetivo principal de SVM.

Teorema 2.2. *El margen geométrico λ es igual a $\frac{1}{||\overrightarrow{w}||}$*

Demostración:

Sea $\overrightarrow{x_c}$ un punto perteneciente al hiperplano $h(x) = \overrightarrow{w} \cdot \overrightarrow{x} + b = -1$ y sea $\overrightarrow{x_d}$ un punto en el hiperplano $l(x) = \overrightarrow{w} \cdot \overrightarrow{x} + b = 1$, por tanto, se cumple lo siguiente:

$$H_c : \overrightarrow{w} \cdot \overrightarrow{x_c} + b = -1 \tag{2.11}$$

$$H_d : \overrightarrow{w} \cdot \overrightarrow{x_d} + b = 1 \tag{2.12}$$

Por tanto, si llamamos a d como la distancia entre esos dos hiperplanos, d se interpreta como el margen que tiene nuestro hiperplano de separación, es decir, $d = 2 * \lambda$, para calcular d, restemos la ecuación del hiperplano H_c a la ecuación del hiperplano H_d, obteniendo:

$$\overrightarrow{w} \cdot (\overrightarrow{x_d} - \overrightarrow{x_c}) = 2 \tag{2.13}$$

Por otro lado, recordemos la siguiente propiedad del producto escalar entre vectores:

Sean $\overrightarrow{u}$ y $\overrightarrow{v}$ vectores en $\mathbb{R}^n$, entonces el producto escalar entre vectores puede ser expresado como

$$\overrightarrow{u} \cdot \overrightarrow{v} = ||\overrightarrow{u}||||\overrightarrow{v}|| \cos(\theta) \tag{2.14}$$

donde θ es el ángulo entre dichos vectores.
Si reagrupamos y reordenamos términos en la ecuación anterior, entonces (2.14) puede expresarse como:

$$\overrightarrow{u} \cdot \overrightarrow{v} = (||\overrightarrow{v}|| \cos(\theta))||\overrightarrow{u}|| = v_u||\overrightarrow{u}|| \tag{2.15}$$

donde v_u representa la longitud de $\overrightarrow{v}$ en dirección de $\overrightarrow{u}$ (Figura2.4).
Ahora bien, regresando a la ecuación (2.13) y aplicando (2.14) se obtiene lo siguiente

$$\overrightarrow{w} \cdot (\overrightarrow{x_d} - \overrightarrow{x_c}) = ||\overrightarrow{w}|| \cdot d = 2 \tag{2.16}$$

donde $d = ||(\overrightarrow{x_d} - \overrightarrow{x_c})|| \cos(\alpha)$, es la longitud dirigida de la proyección del vector $(\overrightarrow{x_d} - \overrightarrow{x_c})$ sobre $\overrightarrow{w}$, (Figura 2.5).
De ahí que

$$||\overrightarrow{w}|| * d = 2 \tag{2.17}$$

$$\Rightarrow d = \frac{2}{||\overrightarrow{w}||} \tag{2.18}$$

Pero

$$d = 2 * \lambda \tag{2.19}$$

$$\therefore \lambda = \frac{1}{||\overrightarrow{w}||}.\square \tag{2.20}$$

La fase de entrenamiento para nuestro modelo SVM, incluye la estimación de los parámetros $\overrightarrow{w}$ y b del hiperplano de separación usando los datos de prueba. Dichos parámetros deben ser elegidos de tal forma que se cumplan las siguientes dos condiciones:

$$\overrightarrow{w} \cdot \overrightarrow{x}_i + b \geq 1, \text{si} \quad y_i = 1 \tag{2.21}$$

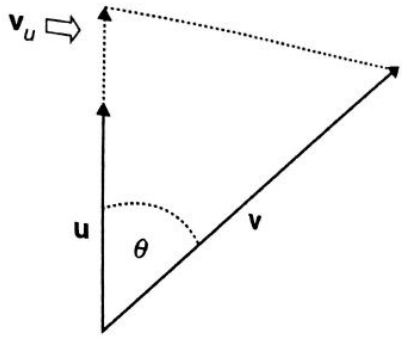


Figura 2.4: Proyección ortogonal del vector v sobre u.

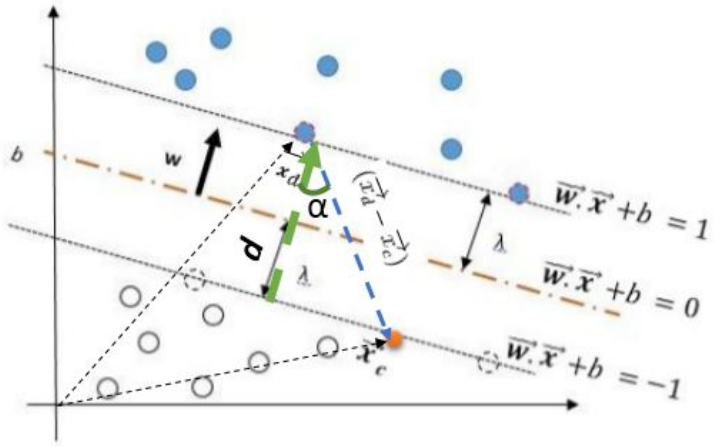


Figura 2.5: Ilustración de la proyección del vector $x_d - x_c$ sobre w.

$$\overrightarrow{w} \cdot \overrightarrow{x}_i + b \leq -1, \text{si} \quad y_i = -1. \tag{2.22}$$

Con ésto se trata de separar lo mejor posible a los datos dependiendo de la clase a la que pertenecen. Ambas inecuaciones se pueden presentar en la siguiente ecuación compacta

$$y_i(\overrightarrow{w} \cdot \overrightarrow{x}_i + b) \geq 1, i = 1, 2, ..., N. \tag{2.23}$$

Otra condición que solicita el modelo es que el margen del hiperplano de separación debe ser máximo. Es decir, se pide maximizar $d = \frac{2}{||\overrightarrow{w}||}$, pero esta función no es diferenciable, no obstante este problema equivale a minimizar la siguiente función:

$$g(\overrightarrow{w}) = \frac{||\overrightarrow{w}||^2}{2}. \tag{2.24}$$

Con esto definimos el siguiente problema

Definición 2.3. *(SVM lineal: Caso separable) La labor de aprendizaje en SVM se puede formalizar como el siguiente problema de optimización con restricciones:*

$$\min_w \frac{||\overrightarrow{w}||^2}{2} \tag{2.25}$$

Sujeto a

$$y_i(\overrightarrow{w} \cdot \overrightarrow{x}_i + b) \geq 1, i = 1, 2, ..., N.$$

Dado que la función objetivo es una función cuadrática y las restricciones son lineales en los parámetros $\overrightarrow{w}$ y b, entonces estamos ante un problema de optimización convexa, el cual puede ser solucionado con el método de **multiplicadores de Lagrange**.
Sea $L(\overrightarrow{w}, \lambda) = \frac{||\overrightarrow{w}||^2}{2} - \sum_{i=1}^{N} \lambda_i(y_i(\overrightarrow{w} \cdot \overrightarrow{x}_i + b) - 1)$ el lagrangiano y λ_i los multiplicadores de Lagrange. A continuación minimizemos al lagrangiano

$$\frac{\partial L}{\partial \overrightarrow{w}} = 0 \Rightarrow \overrightarrow{w} = \sum_{i=1}^{N} \lambda_i(y_i \overrightarrow{x}_i) \tag{2.26}$$

$$\frac{\partial L}{\partial b} = 0 \Rightarrow \sum_{i=1}^{N} \lambda_i y_i = 0. \tag{2.27}$$

Con las condiciones de Karush-Kuhn-Tucker(KKT):

$$\lambda_i \geq 0 \tag{2.28}$$

$$\lambda_i[y_i(\overrightarrow{w} \cdot \overrightarrow{x}_i + b) - 1)] = 0. \tag{2.29}$$

Para simplificar el problema y tener una forma más simple de resolver es utilizando el problema dual del problema original, para lograr dicha transformación sustituyamos las ecuaciónes (2.26)-(2.27) en la expresión del lagrangiano, esto es:

$$L(\overrightarrow{w}, \lambda) = \frac{||\overrightarrow{w}||^2}{2} - \sum_{i=1}^{N} \lambda_i(y_i(\overrightarrow{w} \cdot \overrightarrow{x}_i + b) - 1) \tag{2.30}$$

$$= \frac{||\overrightarrow{w}||^2}{2} - \sum_{i=1}^{N} \lambda_i y_i(\overrightarrow{w} \cdot \overrightarrow{x}_i) - \sum_{i=1}^{N} \lambda_i y_i b + \sum_{i=1}^{N} \lambda_i \tag{2.31}$$

Sustituyendo la ecuación (2.27), se simplifica a la siguiente expresión

$$L(\overrightarrow{w}, \lambda) = \frac{||\overrightarrow{w}||^2}{2} - \sum_{i=1}^{N} \lambda_i y_i(\overrightarrow{w} \cdot \overrightarrow{x}_i) + \sum_{i=1}^{N} \lambda_i. \tag{2.32}$$

Ahora bien, si sustituimos la expresión dada en (2.26), entonces se puede verificar fácilmente (por inducción matemática sobre N), que:

$$\sum_{i=1}^{N} \lambda_i y_i [\sum_{i=1}^{N} \lambda_i (y_i \overrightarrow{x}_i)] \cdot \overrightarrow{x}_i = ||\overrightarrow{w}||^2 = \sum_{i,j} \lambda_i \lambda_j y_i y_j \overrightarrow{x}_i \cdot \overrightarrow{x}_j \tag{2.33}$$

Por tanto, 2.32 se expresa de la siguiente manera:

$$L_D = \sum_{i=1}^{N} \lambda_i - \frac{1}{2} \sum_{i,j} \lambda_i \lambda_j y_i y_j \overrightarrow{x}_i \cdot \overrightarrow{x}_j. \tag{2.34}$$

La ventaja de trabajar con el problema dual es que este problema solo involucra a los multiplicadores de Lagrange y los datos de entrenamiento, situación diferente al problema primal, ya que por su formulación se trabajaba además con los parámetros $\overrightarrow{w}$ y b, escenario en el cual hace más tediosa la labor de encontrar la solución. De igual forma, gracias a los Teoremas de Dualidad dados en el ambiente de optimización, se sabe que si el problema dual tiene solución, por tanto, el problema primal lo tiene y las respectivas soluciones son equivalentes.

Para grandes conjuntos de datos , el problema dual de optimización puede ser

solucionado usando las técnicas numéricas como la programación cuadrática, un tema ajeno al objetivo de este trabajo. Una vez encontrados los valores de $\lambda_i's$, los podemos usar en las ecuaciones (2.26) y (2.29), para lograr encontrar los valores óptimos de $\overrightarrow{w}$ y b. La ecuación de nuestro hiperplano de separación puede ser expresado como sigue:

$$\sum_{i=1}^{N} \lambda_i y_i (\overrightarrow{x}_i \cdot \overrightarrow{x}) + b = 0 \tag{2.35}$$

donde b, puede ser encontrada usando la ecuación (2.29) y usando los vectores de soporte, es decir, aquellos vectores que cumplen (2.11) o (2.12).
Existe una variante en el modelo construido anteriormente y esto es cuando nuestra hipótesis de que los datos son linealmente separables es falsa, para el estudio más a detalle de este tipo de variante se puede consultar [13].

2.2.3. SVM no lineales

En esta sección se presentará una breve exposición de otra variante que tiene este tipo de modelo no paramétrico, y es cuando nuestros datos no poseen un hiperplano de separación lineal. El truco que se utiliza en estos casos es la transformación de los datos de su espacio de coordenadas originales en $\overrightarrow{x}$ a un nuevo espacio $\Phi(\overrightarrow{x})$, para que entonces se pueda utilizar un hiperplano de separación lineal en el espacio transformado.
Para el aprendizaje de nuestro modelo SVM no lineal se tiene la siguiente definición.

Definición 2.4. *(SVM No lineal) La labor de aprendizaje en SVM se puede formalizar como el siguiente problema de optimización con restricciones:*

$$\min_{w} \frac{||\overrightarrow{w}||^2}{2} \tag{2.36}$$

Sujeto a

$$y_i(\overrightarrow{w} \cdot \Phi(\overrightarrow{x}_i) + b) \geq 1, i = 1, 2, ..., N.$$

Es notable la similitud que existe entre la Definición anterior y la Definición 2.3, aunque la principal diferencia entre ambas definiciones es que para el caso de SVM lineales, ocupamos los datos en su forma original, $\overrightarrow{x}$, en cambio para SVM no lineales utilizamos el espacio de transformación de nuestros datos $\Phi(\overrightarrow{x})$.

Siguiendo el enfoque utilizado para SVM lineales, se puede obtener el Lagrangiano dual para el problema de optimización con restricciones

$$L_D = \sum_{i=1}^{N} \lambda_i - \frac{1}{2} \sum_{i,j} \lambda_i \lambda_j y_i y_j \Phi(\overrightarrow{x}_i) \cdot \Phi(\overrightarrow{x}_j). \tag{2.37}$$

Una vez encontrados los valores de $\lambda_i's$ usando las técnicas de programación cuadráticas, los parámetros $\overrightarrow{w}$ y b se pueden derivar usando las siguientes ecuaciones

$$\overrightarrow{w} = \sum_{i=1}^{N} \lambda_i (y_i \Phi(\overrightarrow{x}_i)) \tag{2.38}$$

$$0 = \lambda_i [y_i (\sum_{i=1}^{N} \lambda_i (y_i \Phi(\overrightarrow{x}_i)) \cdot \Phi(\overrightarrow{x}_i) + b) - 1)]. \tag{2.39}$$

Es notable que en la última ecuación dada anteriormente involucra el producto interno entre dos vectores pero ahora en el espacio transformado, tal cálculo puede llegar a ser bastante engorroso, sin embargo, existe un método innovador conocido como el **truco del kernel**, método que solo sera mencionado en forma superficial, ya que para los detalles son necesarios algunos conocimientos de la teoría de los espacios de **Hilbert**, y teoría de la medida.

Truco del Kernel

Para ilustrar este método, se usará un ejemplo elaborado en [13] y basado en el espacio $\mathbb{R}^2$.
En un problema SVM no lineal, supongamos que se desean transformar los datos con la función $\Phi(\overrightarrow{x}) = (x_1^2, x_2^2, \sqrt{2}x_1, \sqrt{2}x_2, 1)$, luego entonces, el producto punto entre dos vectores $\overrightarrow{u}$ y $\overrightarrow{v}$, en el espacio transformado por Φ está dado por:

$$\Phi(\overrightarrow{u}) \cdot \Phi(\overrightarrow{v}) = (u_1^2, u_2^2, \sqrt{2}u_1, \sqrt{2}u_2, 1) \cdot (v_1^2, v_2^2, \sqrt{2}v_1, \sqrt{2}v_2, 1) \tag{2.40}$$

$$= u_1^2 v_1^2 + u_2^2 v_2^2 + 2u_1^2 v_1^2 + 2u_2^2 v_2^2 + 1 \tag{2.41}$$

$$= (\overrightarrow{u} \cdot \overrightarrow{v} + 1)^2. \tag{2.42}$$

El ejemplo anterior da un bosquejo de que el producto interno en el espacio transformado puede ser expresado en términos de una función "similar" en el espacio original, es decir,

$$K(\overrightarrow{u}, \overrightarrow{v}) = \Phi(\overrightarrow{u}) \cdot \Phi(\overrightarrow{v}) = (\overrightarrow{u} \cdot \overrightarrow{v} + 1)^2. \tag{2.43}$$

La función K, la cuál se calcula ocupando las coordenadas en el espacio original es conocida como la función "Kernel". Una función "Kernel" puede ser interpretada como un tipo de medida de similitud entre los objetos de entrada y puede verse como una transformación no lineal, es decir que implica una correspondencia hacia un espacio de mayor dimensión, posiblemente infinita. Cuando se encuentra un producto escalar en función del espacio de entrada entonces a este producto escalar se le denomina "Kernel", y el espacio de mayor dimensión (espacio de características) es un espacio de Hilbert (Reproducing Kernel Hilbert Space, RKHS). Dicha función ayuda a incrementar la posibilidad de que existe separabilidad lineal entre los datos en ese nuevo espacio.
Si se encuentra una transformación no lineal $\phi(\overrightarrow{x})$ a un espacio de mayor dimensionalidad provisto de un producto escalar que puede ser expresado como (kernel):

$$K(\overrightarrow{x_i}, \overrightarrow{x_j}) = \phi(\overrightarrow{x_i})^T \phi(\overrightarrow{x_j})$$

entonces se puede construir una versión no lineal del mismo algoritmo donde la transformación no lineal es ϕ. A esto se le es conocido como el "truco de los kernels".[8]
Algunos ejemplos de funciones kernel ocupadas comúnmente son:

$$K(\overrightarrow{x}, \overrightarrow{y}) = (\overrightarrow{x} \cdot \overrightarrow{y} + 1)^p \tag{2.44}$$

$$K(\overrightarrow{x}, \overrightarrow{y}) = (e^{\frac{-||\overrightarrow{x} - \overrightarrow{y}||^2}{2\sigma^2}}) \tag{2.45}$$

$$K(\overrightarrow{x}, \overrightarrow{y}) = tanh(k\overrightarrow{x} \cdot \overrightarrow{y} - \delta) \qquad (Sigmoidal) \tag{2.46}$$

$$K(\overrightarrow{x}, \overrightarrow{y}) = e^{-\frac{||\overrightarrow{x} - \overrightarrow{y}||}{\sigma}} \qquad (Laplaciano) \tag{2.47}$$

Es importante decir que una función kernel también debe de verificar el teorema de Mercer, que en modo matemático se expresa de la siguiente forma

$$\int_{u,v} K(\overrightarrow{u}, \overrightarrow{v}) g(\overrightarrow{u}) g(\overrightarrow{v}) du dv > 0$$

para toda función g de cuadrado integrable.

Capítulo 3

Credit Scoring

La utilización de modelos de *credit scoring* comenzó en los años 70's con el objetivo de la evaluación del riesgo de crédito, es decir, con la estimación de probabilidades de fallo o "default" y así, ir clasificando a los clientes para el otorgamiento de un crédito, esta idea se generaliza a partir de los 90's, gracias al desarrollo de la tecnología, recursos computacionales y la creciente necesidad de la industria bancaria para hacer más eficiente el otorgamiento de créditos y mejorar su evaluación del riesgo de su portafolio.
Estos modelos se asocian con lo que hoy en día llamamos "data mining" (minería de datos), que engloba aquellos procedimientos que permiten la extracción de información útil y así encontrar patrones de comportamiento de los datos.
Cabe mencionar que a pesar de el desarrollo de modelos de *credit scoring*, el juicio humano continua siendo utilizado en el análisis para otorgamiento de créditos, muchas veces ambas metodologías coexisten y complementan formando así sistemas híbridos.
En esta sección se abordará la definición de *credit scoring* y sus procesos utilizados para posteriormente utilizarlos en nuestro problema de origen: encontrar un buen clasificador de clientes buenos y malos en el otorgamiento de un crédito de consumo.

3.1. ¿Qué es el credit scoring?

Los modelos de *credit scoring* son algoritmos que de manera automática evalúan el riesgo al momento de una solicitud de crédito de un solicitante,

esto aplica tanto para personas físicas y morales. Dicha evaluación se hace de forma individual, es decir, solo se hace el análisis respecto al riesgo de incumplimiento del individuo o empresa, independientemente de lo que suceda con el resto de la cartera de préstamos.

Inicialmente, en los años $70's$, los modelos de *credit scoring* se basaban en técnicas estadísticas, principalmente con la técnica desarrollada por el estadístico Sir Ronald Aylmer Fisher, el análisis discriminante, actualmente los modelos se estan basando en técnicas matemáticas econométricas y de inteligencia artificial.

El resultado de la evaluación se ve reflejada con la asignación de una medida que permite ir ordenando a los evaluados en función de su riesgo y así al final generar un puntaje o "score". En general, el objetivo es estimar la probabilidad de incumplimiento del deudor (PD), asociada a su score, rating o clasificación obtenida.

Cabe mencionar que los modelos de "credit scoring" requieren de dos elementos fundamentales:

- **Información histórica:** Las instituciones bancarias cuentan con base de datos donde contienen el comportamiento de sus clientes y son almacenados aproximadamente cinco años.
- **Análisis estadístico:** Las personas que se encuentran en el área mencionada tienen como labor principal utilizar la información histórica para identificar, mediante algoritmos estadísticos, el comportamiento de los clientes y con ellos poder determinar probabilidades de ocurrencia de eventos futuros.

Entre las metodologías disponibles, los modelos probit/logit junto con regresiones lineales, el análisis discriminante y los árboles de decisión se encuentran entre los métodos más usados en la industria para desarrollar modelos de ésta índole. Varios autores en la literatura que habla de estas metodologías utilizan mayoritariamente los modelos probit ya que toman en cuenta la probabilidad de default del deudor o bien de cada exposición en la cartera del banco. Que dichas metodologías se ocupen frecuentemente no quiere implicar que no puedan combinarse, de hecho muchos analistas de riesgos combinan técnicas para tener mayor argumentos en su clasificación

final ante las solicitudes de crédito. Citando el ejemplo dado en [5] cuando son utilizados los árboles de regresión: a través de un árbol se segmenta la muestra de deudores y luego a los deudores de cada segmento se les estima con una regresión logística o modelo probit con distintas características.

3.1.1. Variables empleadas

Las variables a utilizar para hacer modelos de *credit scoring* varia dependiendo a quien se le va a realizar el estudio, cuando son para empresas (PyMEs) o grandes corporativos se utilizan frecuentemente variables socioeconómicas las cuáles pueden ser extraídas de los estados contables, proyecciones del flujo de fondos, etc. Se tienen más procedimientos para realizar este tipo de estudios que se pueden consultar en [5]. Retomando los modelos que son de nuestro interés para este trabajo que son para individuos utilizamos variables socioeconómicas tales como: edad, estado civil, cantidad de personas a cargo, tiempo de permanencia en el domicilio actual y en el empleo actual, nivel educativo, si es propietario de la vivienda que habita, gastos mensuales promedio al igual que sus ingresos, tipo de ocupación, si tiene tarjetas de crédito y finalmente el número de consultas en el *credit bureaus*, es decir, buró de crédito y como esta su calificación en dicho sistema. Pueden existir más variables pero las mencionadas son las que históricametne han sido significativas en los modelos usados por bancos [5], sin embargo, todo va a depender de la información con la que cuente el investigador y la significancia que le muestre cada modelo paramétrico usado (Logit/probit).

3.1.2. Matriz de riesgo

Una matriz de riesgo es una herramienta de control que es utilizada para identificar aquellas actividades importantes de una institución financiera, donde se analizan los riesgos (tipos y nivel) ligadas a estas actividades. Sirve, de igual forma, para la gestión y administración de riesgos financieros, operativos y estratégicos que tiene la organización. Esta herramienta permite presentar gráficamente el impacto (severidad) y frecuencia (probabilidad de ocurrencia) de los riesgos, de esta manera se convierte en una guía visual que facilita asignar prioridades de atención de determinados riesgos.

Una arquitectura simple de una matriz de riesgo puede contener sólo cuatro cuadrantes:

- Alto impacto / alta probabilidad de ocurrencia
- Bajo impacto / alta probabilidad de ocurrencia
- Alto impacto / baja probabilidad de ocurrencia
- Bajo impacto / baja probabilidad de ocurrencia

Gráficamente se pueden apoyar de los colores para ir identificando los casos según el tipo de riesgo de cada actividad, por ejemplo "alto impacto con alta probabilidad de ocurrencia" se puede identificar de color rojo, y la clasificación de "bajo impacto con baja probabilidad de ocurrencia" de color verde.

Capítulo 4

Aplicación

Para la aplicación de las teorías mostradas anteriormente, se va a utilizar una base de datos de personas que solicitaron una tarjeta de crédito y donde la empresa que otorgaba créditos la calificaba como "Aprobado" y "No aprobado", por confidencialidad solo se mostraran aquellas variables que son útiles en el objetivo de éste trabajo omitiendo el nombre, fecha de nacimiento, dirección y teléfono.
Para poder trabajar la base de datos se tuvo que someter a un proceso de validación, donde se limpian aquellos registros duplicados, con valores para la variable edad muy atípicos, dígase personas de 80 a 130 años, y personas cuyo salario superaba los $\$150,000$ pesos al mes, todos estos errores se deben a una mala captura de información al momento de la solicitud del cliente. Respecto a la variable del ingreso se tomó como límite dicha cantidad ya que las personas que superaban a ese ingreso se tenía la incertidumbre de que existiera mala captura de los datos y también para tener una base de datos lo más homogénea posible.

4.1. Archivos requeridos

Para evitar análisis erróneos en este trabajo se requiere que la base de datos tenga una estructura específica, para así, pueda ser funcional en la mayoría de los softwares estadísticos. La mayoría de paqueterias y softwares trabajan con bases de datos cuya estructura sea igual a la de una matriz, es decir, cada columna representa una variable y cada fila el valor de ésta.

En este trabajo se utilizará el software llamado *R*, donde para efectos de cálculos se necesita trabajar con archivos de **EXCEL** en formato *csv*, de igual manera se trabajará con el software de *STATA*, que acepta archivos en formato *xlsx*.

4.2. Segmentación de clientes en nuestra base de datos

Para poder aplicar las técnicas comentadas en los capítulos anteriores es necesario tener una muestra de registros que serán utilizados para el entrenamiento de nuestro modelo y otra muestra que contenga los registros con los cuáles se va a poner a prueba el modelo predictivo y así determinar, que tan bueno es o no para clasificar, a nuestros prospectos en adquirir la tarjeta de crédito, en "Aprobado" o "No aprobado". Para éste mecanismo se utilizó un muestreo aleatorio estratificado, ya que en nuestros datos existen dos categorías (estratos) a estudiar. El tamaño de la muestra que conforma la base de datos de entrenamiento de los modelos se calculo mediante a las recomendaciones de autores en el área de minería de datos, donde se indica que del total de datos el 70 % sea tomado como datos de entrenamiento y el resto para los datos de "testing". En nuestra base de datos se tiene un total de 13,735 registros, de ahí que nuestro n para este tipo de muestreo es igual a 9615 y para probar nuestros modelos tendremos un total de 4120 registros. Ahora bien, para poder hacer el muestreo aleatorio mencionado, vamos a tomar una muestra representativa de cada uno de nuestros estratos, esto es, para el estrato con la categoría correspondiente a "NO" tomaremos el siguiente tamaño de muestra:
Sea N el total de datos y N_1 el total de datos con la categoria "NO", por tanto, la muestra a tomar es

$$\frac{N_1}{N}n = \frac{11,361}{13,735}(9,615) = 7,953$$

Del mismo modo, para nuestro estrato cuya etiqueta es "SI" y donde N_2 corresponde al total de datos con dicha categoría, se obtiene el siguiente tamaño de muestra

$$\frac{N_2}{N}n = \frac{2374}{13,735}(9,615) = 1,662$$

Una vez que se sabe el número de elementos a tomar de muestra para cada estrato, se procede a hacer un muestreo aleatorio simple sin reemplazo, con ayuda de la generación de números aleatorios en **EXCEL** y una vez generados solo queda ordenarlos de menor a mayor y así tomar los primeros 7,953 registros, dado que no estaban originalmente ordenados y etiquetados por un ID, solo es cuestión de hacer match con el mismo ID y así tener nuestros 7953 registros de entrenamiento, en la Tabla 4.1 se muestra lo comentado

Tabla 4.1: Base de datos (antes del muestreo aleatorio estratificado)

A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R	S	T	U	V	W	X	Y	Z	AA	
ID	I	F	H	I	I	M	c	EDAD	EDAD2	GENERO	VEP	NDEP	QIM	LNQIM	EDUC	ZM	BBVA	BMX	BNORTE	SANTA	HSBC	IWMT	SCTBNK	AMEX	DEPA	OTROS	Y_R
1	I		0	I	(	(	Qu	60	3600	M	NO	0	12000	9.3926619	12	SI	SI	NO	NO	NO	NO	NO	NO	NO	NO	NO	NO
2	I		S	I	T	e		41	1681	H	NO	2	30000	10.31	17	NO	NO	NO	NO	SI	NO	NO	NO	NO	NO	NO	NO
3	I		0	I	I	T	o	31	961	M	NO	0	15000	9.6158055	17	SI	NO	SI	NO	NO	NO	NO	NO	NO	NO	NO	NO
4	S		S	I	I	z	t	40	1600	M	NO	2	7000	8.85	14	NO	NO	NO	NO	SI	NO	NO	NO	NO	NO	NO	NO
5	(		S	(	C	h		28	784	H	NO	0	15000	9.62	17	SI	NO	NO	NO	SI	NO	NO	NO	NO	NO	NO	NO
6	E		(	I	N	e		24	576	H	NO	1	8000	8.99	17	NO	NO	NO	NO	NO	NO	NO	NO	NO	SI	NO	NO
7	(		I	I	N	a		27	729	H	NO	0	35000	10.46	17	NO	SI	NO	NO	NO	NO	NO	NO	NO	NO	NO	NO
8	(		S	(	S	o		49	2401	H	NO	0	17000	9.74	17	NO	NO	NO	NO	SI	NO	NO	NO	NO	NO	NO	NO
9	I		0	I	(	(	Qu	42	1764	M	SI	1	9000	9.1049799	17	SI	NO	NO	SI	NO	NO	NO	NO	NO	NO	NO	NO
10	(		0	I	I	z	t	25	625	H	SI	2	7000	8.8536654	9	NO	NO	NO	NO	NO	NO	NO	NO	NO	SI	NO	NO
11	(		1	I	I	N	e	30	900	H	NO	1	30000	10.308953	17	NO	SI	NO	NO	NO	NO	NO	NO	NO	NO	NO	SI
12	I			I	\	M	e	36	1296	M	NO	2	50000	10.82	17	SI	NO	SI	NO	NO	NO	NO	NO	NO	NO	NO	NO
13	I			I	(	T	u	29	841	H	NO	0	15000	9.62	17	SI	NO	NO	SI	NO	NO	NO	NO	NO	NO	NO	NO
14	I			I	I	C	u	40	1600	H	SI	3	19000	9.85	17	SI	SI	NO	NO	NO	NO	NO	NO	NO	NO	NO	NO
15	(			I	I	C	o	24	576	H	NO	0	38000	10.55	17	NO	NO	SI	NO	NO	NO	NO	NO	NO	NO	NO	NO
16	I		1	I	I	M	e	32	1024	H	SI	4	15000	9.6158055	12	NO	NO	NO	NO	NO	SI	NO	NO	NO	NO	NO	NO

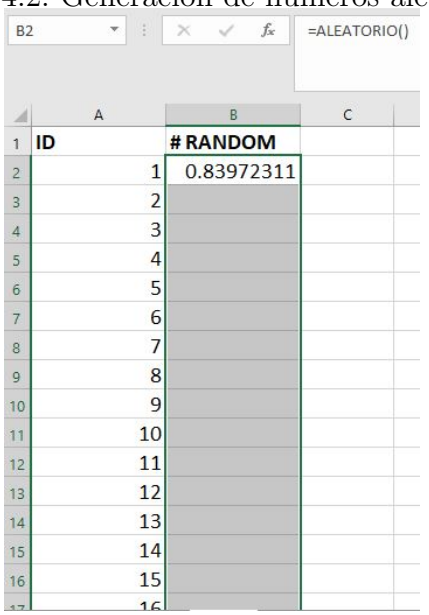
El ID, la variable que aparece marcada de color amarillo, es el número que nos va ayudar a determinar que registro entra a la muestra de entrenamiento y cual a la de prueba. A continuación se muestra el significado de cada variable en nuestra base de datos

1. **EDAD:** Edad del registro a la hora de hacer su solicitud
2. **EDAD2:** Edad del registro al cuadrado
3. **GENERO:** Género al que pertenece el registro
4. **VEP:** Vive en pareja
5. **NDEP:** Número de dependientes económicos
6. **QIM:** Ingreso mensual
7. **LNQIM:** Logaritmo natural del ingreso mensual

8. **EDUC:** Años de educación del registro
9. **ZM:** Vive en zona metropolitana
10. **BBVA:** El registro tiene una tarjeta de crédito del banco BBVA Bancomer
11. **BMX:** El registro tiene una tarjeta de crédito del banco Banamex
12. **BNORTE:** El registro tiene una tarjeta de crédito del banco Banorte
13. **SANTA:** El registro tiene una tarjeta de crédito del banco Santander
14. **HSBC:** El registro tiene una tarjeta de crédito del banco HSBC
15. **IWMT:** El registro tiene una tarjeta de crédito del banco Walmart (INBURSA)
16. **SCTBNK:** El registro tiene una tarjeta de crédito del banco Scotiabank
17. **AMEX:** El registro tiene una tarjeta de crédito del banco American express
18. **DEPA:** El registro tiene una tarjeta de crédito departamental
19. **OTROS:** El registro tiene una tarjeta de crédito diferente a los mencionados anteriormente
20. **Y_R:** Variable que indica si la solicitud del registro fue o no aprobada.

Una vez generados los números aleatorios con la fórmula de **EXCEL** observada en la Tabla 4.2 , copiamos y pegamos como valores dichos números para evitar cambios al momento de ordenar los números, una vez ordenados los números de menor a mayor tomamos los primeros 7,953 registros, y utilizando el ID, hacemos cruce con los ID originales de nuestra base de datos, con eso tenemos la muestra que conforma nuestra base de aprendizaje. Del mismo modo se procede para extraer los 1,662 registros cuya clasificación es "SI". Con esto se puede conformar una nueva base de datos que contienen nuestros datos de entrenamiento, para una mejor estructura de nuestros datos, se procede a mezclar todos los registros cuyo proceso es similar a la extracción de muestras, es decir, una vez teniendo lista nuestra base de datos con todas las variables a trabajar, procedemos a crear una columna donde contengan números aleateorios generados por la misma fórmula de

Tabla 4.2: Generación de números aleatorios



	A	B	C
1	ID	# RANDOM	
2	1	0.83972311	
3	2		
4	3		
5	4		
6	5		
7	6		
8	7		
9	8		
10	9		
11	10		
12	11		
13	12		
14	13		
15	14		
16	15		

Tabla 4.3: Ordenación de los números aleatorios y selección de los primeros 7,953 registros

A	B
ID	# RANDOM
1971	0.00014398
8790	0.00018829
6833	0.00027472
9895	0.0003318
8740	0.00041124
3954	0.00065497
6090	0.00085156
6587	0.00089462
3240	0.00091595
8726	0.00105902
7228	0.00106453
5015	0.00113678
4588	0.00116623

Tabla 4.4: Base de datos de entrenamiento

A	B	C	D	E	F	G	H	I	J	K	L	M	N	C	P	Q	R	S	T	U	V	W	X	Y	Z	AA	AB
ID	RI	FD	HD	BA	EC	M	cp	EDAD	EC	GENERO	VEP	NDEP	QIM	L	EDUC	ZM	BBVA	BMX	BNORTE	SANTA	HSBC	IWMT	SCTBNK	AMEX	DEPA	OTROS	Y_R
9597	G(	##	12:	BA	Qu	Be	#	36	#	M	NO	2	16000		17	NO	NO	SI	NO	NO	NO	NO	NO	NO	NO	NO	SI
9598	N/	##	##	BA	Sa	Sa	#	25	#	M	SI	1	15000		17	SI	NO	SI	NO	NO	NO	NO	NO	NO	NO	NO	NO
9599	SA	##	##	HS	Ja	Sa	#	58	#	M	NO	0	30000		17	NO	NO	NO	NO	NO	SI	NO	NO	NO	NO	NO	NO
9600	VI	##	09:	HS	Sc	He	#	39	#	H	SI	1	30000		17	NO	NO	NO	NO	NO	SI	NO	NO	NO	NO	NO	NO
9601	SI	##	09:	SA	Ta	Re	#	35	#	M	SI	3	16000		9	SI	NO	NO	NO	SI	NO	NO	NO	NO	NO	NO	NO
9602	RA	##	##	BA	Di	Tl	#	43	#	M	SI	1	7500		9	NO	NO	SI	NO	NO	NO	NO	NO	NO	NO	NO	NO
9603	CA	##	##	BA	Yu	M	#	40	#	H	SI	2	12000		12	SI	SI	NO	NO	NO	NO	NO	NO	NO	NO	NO	NO
9604	R(	##	##	BA	M	Ni	#	24	#	M	SI	2	2500		12	NO	SI	NO	NO	NO	NO	NO	NO	NO	NO	NO	NO
9605	P(	##	##	SE	Yu	M	#	28	#	M	NO	1	8000		17	SI	NO	NO	NO	NO	NO	NO	NO	NO	NO	SI	NO
9606	LI	##	##	BA	Di	Iz	#	31	#	H	NO	1	20000		14	NO	NO	SI	NO	NO	NO	NO	NO	NO	NO	NO	NO
9607	M	##	##	IN	Di	Ve	#	27	#	H	NO	1	14000		14	NO	NO	NO	NO	NO	NO	SI	NO	NO	NO	NO	NO
9608	C(	##	##	AT	Ba	Ti	#	26	#	H	NO	0	18000		12	SI	NO	NO	NO	NO	NO	NO	NO	NO	SI	NO	NO
9609	VA	##	##	BA	Ta	Ta	#	43	#	M	NO	2	25000		17	SI	NO	SI	NO	NO	NO	NO	NO	NO	NO	NO	NO
9610	AA	##	##	BA	Di	M	#	40	#	M	SI	3	70000		17	NO	NO	NO	SI	NO	NO	NO	NO	NO	NO	NO	NO
9611	O(	##	##	SA	Ja	Gi	#	46	#	H	NO	3	15000		9	SI	NO	NO	NO	SI	NO	NO	NO	NO	NO	NO	NO
9612	AA	##	##	BA	Qu	Qu	#	31	#	M	SI	1	35000		17	SI	SI	NO	NO	NO	NO	NO	NO	NO	NO	NO	NO
9613	CA	##	##	HS	Ch	Tu	#	25	#	H	NO	0	15000		17	SI	NO	NO	NO	NO	SI	NO	NO	NO	NO	NO	NO
9614	BE	##	##	BA	M	Tc	#	29	#	M	SI	2	30000		17	SI	SI	NO	NO	NO	NO	NO	NO	NO	NO	NO	SI
9615	AA	##	##	SE	Di	Be	#	44	#	M	NO	1	15000		17	NO	NO	NO	NO	NO	NO	NO	NO	NO	NO	SI	SI

EXCEL y así re-ordenamos de menor a mayor nuestros datos. Así nuestro archivo se ve de la siguiente forma

Es preciso comentar que, esta base de datos y su estructura sera utilizada para la aplicación del software libre *R*, puesto que con esta estructura trabajan los algoritmos de procesamiento de datos, caso distinto ocurre para el software *STATA*, donde las variables clasificadas como "SI" serán sustituidas por "1" y "NO" con "0', con ellos la base de datos solo contiene variables númericas donde *STATA* toma a 1 y 0 como variables dummies y no como números.

Tabla 4.5: Base de datos para STATA

ID	RI	FD	HD	BA	EC	MI	cp	EDAD	GENERO	VEP	NDEP	QIM	EDUC	ZM	BBVA	BMX	BNORTE	SANTA	HSBC	IWMT	SCTBNK	AMEX	DEPA	OTROS	*Y_R*
1	RA	##	03	BA	CC	Qu	##	60	1	0	0	12000	12	1	1	0	0	0	0	0	0	0	0	0	*0*
2	F(	##	#	SA	M	Te	##	41	0	0	2	30000	17	0	0	0	0	1	0	0	0	0	0	0	*0*
3	H	##	01	BA	M	To	##	31	1	0	0	15000	17	1	0	1	0	0	0	0	0	0	0	0	*0*
4	SA	##	#	SA	D	Izt	##	40	1	0	2	7000	14	0	0	0	0	1	0	0	0	0	0	0	*0*
5	G	##	#	SA	CC	Ch	##	28	0	0	0	15000	17	1	0	0	0	1	0	0	0	0	0	0	*0*
6	B/	##	#	CE	M	Ne	##	24	0	0	1	8000	17	0	0	0	0	0	0	0	0	0	1	0	*0*
7	O	##	#	BA	M	Na	##	27	0	0	0	35000	17	0	1	0	0	0	0	0	0	0	0	0	*0*
8	C/	##	#	SA	CC	So	##	49	0	0	0	17000	17	0	0	0	0	1	0	0	0	0	0	0	*0*
9	H	##	02	BA	CC	Qu	##	42	1	1	1	9000	17	1	0	0	1	0	0	0	0	0	0	0	*0*

Finalmente, nuestra base de datos de test, sera aquella cuyos registros no pertenecieron a la base de entrenamiento, cuyo total es 4,120 registros.

Antes de empezar a mostrar el proceso realizado en cada software, se hace mención de tres métodos a experimentar en cada una de las técnicas,

la primera de ellas es utilizando todas las variables disponibles en nuestra base de datos, la segunda utilizando solo algunas de éstas y finalmente solo ocupando aquellas que impactan de manera significativa en los modelos de regresión probit y logit.

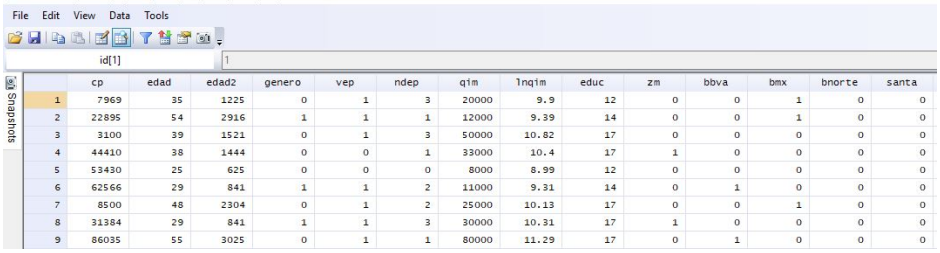
4.3. Uso de las técnicas paramétricas

Modelo paramétrico 1

Modelo Logit

En esta sección se hará uso del software llamado *STATA*, en donde se va a ingresar la base de datos a trabajar. En la sección donde se presentó el tema de Máquinas de Soporte Vectorial, se habló de una base de entrenamiento para el modelo, sin embargo, en este tipo de modelos no se trabaja de dicha forma, solo se ingresará la base de datos de prueba (Test) para poder calificar los resultados de pronóstico de estos modelos. Una vez que ingresamos los datos al software debe aparecer la información como se muestra en la siguiente Tabla.

Tabla 4.6: Base de datos en STATA


File Edit View Data Tools

id[1] 1

	cp	edad	edad2	genero	vep	ndep	qim	lnqim	educ	zm	bbva	bmx	bnorte	santa
1	7969	35	1225	0	1	3	20000	9.9	12	0	0	1	0	0
2	22895	54	2916	1	1	1	12000	9.39	14	0	0	1	0	0
3	3100	39	1521	0	1	3	50000	10.82	17	0	0	0	0	0
4	44410	38	1444	0	0	1	33000	10.4	17	1	0	0	0	0
5	53430	25	625	0	0	0	8000	8.99	12	0	0	0	0	0
6	62566	29	841	1	1	2	11000	9.31	14	0	1	0	0	0
7	8500	48	2304	0	1	2	25000	10.13	17	0	0	1	0	0
8	31384	29	841	1	1	3	30000	10.31	17	1	0	0	0	0
9	86035	55	3025	0	1	1	80000	11.29	17	0	1	0	0	0

Es decir, que los valores numéricos realmente sean numéricos inclusive nuestras variables dummy (0 y 1), esto por requisito del software. Como segundo paso se van a utilizar los códigos construidos para poder obtener el modelo deseado, dichos códigos se presentaran en el Apéndice A.
Iniciamos a constuir el modelo con la siguiente combinación de variables explicativas

$$\begin{aligned} Y_r &= EDAD + GENERO + NDEP + LNQIM + EDUC + BBVA \\ &+ BMX + SANTA + HSBC + DEPA \end{aligned}$$

Una vez ejecutado el código, el software nos muestra las estadísticas de resumen clásicas en un modelo de regresión, que posteriormente se analizarán, y también las probabilidades de predicción de que ocurra el evento de interés es decir, que $Y_r = 1$. Con dichas probabilidades, es necesario desarrollar un "punto de corte" que ayude a determinar la variable en la que se esta interesado, es decir, la variable de pronóstico, para que al final se comparen los resultados obtenidos por el modelo y el valor real que tiene la variable, con ello se medirá la eficiencia del modelo.

Para determinar la variable pronóstico, vamos a utilizar un punto de corte (Cut off o Umbral) con el cual, dependiendo el "score" o probabilidad obtenida por el modelo, se decidirá si la solicitud de crédito es positiva o negativa.

Para encontrar dicho punto de corte usaremos la siguiente regla:
"Consideremos que la proporción del atributo de interés es k $(0 < k < 1)$ y de no interés es $1 - k$. Entonces, supongamos que $k < 1 - k$, entonces elegimos al atributo $I = k$".

Para nuestro caso particular, sea k:=Proporción de registros que son aceptados para darles crédito ($Y =$"1"), y 1-k:=($Y =$"0"), es decir,

$$\begin{aligned} k &= \frac{2374}{13,735} = 0.1728431 \\ 1-k &= \frac{11,361}{13,735} = 0.8271568 \end{aligned}$$

Por tanto, elegimos a nuestro umbral como $I = 0.1728431$, y nuestra decisión queda expresada de la siguiente manera:

$$\sigma(x) = \begin{cases} 1 & ,\text{si} \quad P(Y=1|x) > I, \\ \\ 0 & ,\text{si} \quad P(Y=1|x) \leq I. \end{cases}$$

Por tanto, si codificamos a 1 como "SI" y a 0 como "NO", entonces utilizando la fórmula de Excel "=SI()", se obtiene la variable llamada "Pronóstico" tal y como se muestra en la Tabla 4.7

Tabla 4.7: Tabla de probabilidades para el modelo LOGIT

MODELO LOGIT				
ID	**P(y=0)**	**P(Y=1)**	**Pronostico**	**Y_R**
1	0.8507015	0.1492985	0	0
2	0.6380719	0.3619281	1	1
3	0.8159087	0.1840913	1	0
4	0.9475228	0.0524772	0	0
5	0.9735238	0.0264762	0	0
6	0.8557936	0.1442064	0	0
7	0.7230306	0.2769694	1	0
8	0.8459171	0.1540829	0	0
9	0.7865731	0.2134269	1	0
10	0.9117523	0.0882477	0	0
11	0.8749657	0.1250343	0	0
12	0.867421	0.132579	0	0
13	0.8429797	0.1570203	0	0
14	0.8191327	0.1808673	1	1
15	0.8610076	0.1389924	0	0

Ahora bien, se necesita un método el cuál se pueda comparar los resultados obtenidos para cada kernel a utilizar, y así elegir al mejor modelo que predice con menor error la variable de interés. Para ello utilizaremos una matriz llamada "Tabla de clasificación" o a veces llamada "Matriz de error", la cual tiene la siguiente forma:

Tabla 4.8: Tabla de clasificación

	Y_r (REAL)	
Y_r (PRONOSTICO)	**0**	**1**
0	VERDADEROS NEGATIVOS	FALSO NEGATIVO
1	FALSO POSITIVO	VERDADEROS POSITIVOS

La tabla anterior nos indicará la capacidad de predicción de nuestro modelo, tomando en cuenta variables que se pueden definir de manera inmediata al observar la tabla, es decir, si tomamos como $\omega :=$ "Precisión del modelo"

como la proporción del número total de predicciones correctas respecto al total, $a :=$ "Verdaderos negativos", $b :=$ "Falsos negativos", $c :=$ "Falsos positivos" y a $d :=$ "Verdaderos positivos", entonces tenemos lo siguiente

$$\omega = \frac{a+d}{(a+b+c+d)}$$

También podemos definir varios cocientes que se pueden interpretar de algún modo tal que ayude a saber que tan potente es nuestro modelo al pronosticar nuestra variable de interés. Por ejemplo:

- Sensibilidad := "Probabilidad de que el modelo clasifique como 1 a la variable Y dado que el valor real es $Y = 1$", es decir,

$$Pr(Y_r(Pronóstico) = 1 | Y_r(Real) = 1) = \frac{d}{b+d}$$

- Especificidad:="Probabilidad de que el modelo clasifique como 0 a la variable Y dado que el valor real es $Y = 0$", es decir, esto esta de acuerdo con lo de abajo

$$Pr(Y_r(Pronóstico) = 0 | Y_r(Real) = 0) = \frac{a}{c+a}$$

Existen más probabilidades por calcular, pero solo tomaremos en cuenta los expuestos anteriormente, y con ellos observaremos los resultados que da cada modelo para después elegir el que más convenga según nuestro interés.
Para el caso del modelo logit se obtiene la Tabla 4.9, que fue realizada con los datos contenidos en el archivo de prueba (test), que se ocupará para los demás modelos a probar.

Tabla 4.9: Tabla de clasificación para el modelo logit

MODELO LOGIT						
Cuenta de ID	**Real**					
Pronóstico	**0**	**1**	**Total general**	59.61%		
0	1986	242	2228	% Deteccion de 0	58.27%	Specificity (Pr(-|~D))
1	1422	470	1892	% Detección de 1	66.01%	Sensitivity (Pr(+| D))
Total general	**3408**	**712**	**4120**			

Se puede observar que la precisión del modelo ω es casi 0.6, es decir, empieza a ser un modelo con capacidad de predicción "relativamente" bueno, y usamos "relativo" ya que si analizamos las demás proporciones para la detección de 1, es buena pues esta arriba del 50 % y además esta dentro de los porcentajes que se han aceptado en diferentes estudios econométricos.

Modelo Probit

De la misma forma realizamos la modelación para nuestro modelo probit con las mismas variables, y se observará los porcentajes de precisión para saber cual de estos dos modelos paramétricos se puede utilizar.
Una vez que se ha ejecutado el código en el software se construye la misma tabla que contiene la variable pronóstico cuya construcción esta basada con el mismo umbral ocupado en el modelo logit $I = 0.1728431$, así obtenemos lo siguiente:

Tabla 4.10: Tabla de probabilidades para el modelo probit

MODELO PROBIT				
ID	**P(y=0)**	**P(Y=1)**	**Pronostico**	**Y_R**
1	0.847248	0.152752	0	0
2	0.6448901	0.3551099	1	1
3	0.8149438	0.1850562	1	0
4	0.9484064	0.0515936	0	0
5	0.9785753	0.0214247	0	0
6	0.8554508	0.1445492	0	0
7	0.7237657	0.2762343	1	0
8	0.8454183	0.1545817	0	0
9	0.7842312	0.2157688	1	0
10	0.9129367	0.0870633	0	0
11	0.8748352	0.1251648	0	0
12	0.8626043	0.1373957	0	0
13	0.8414248	0.1585752	0	0
14	0.8155309	0.1844691	1	1

Con estos resultados obtenidos, se resumen en la matriz de error de este modelo cuya forma es:

Tabla 4.11: Tabla de clasificación para el modelo probit

MODELO PROBIT						
Cuenta de ID	**Real**					
Pronóstico	**0**	**1**	**Total general**	58.71%		
0	1936	229	2165	% Deteccion de 0	56.81%	Specificity (Pr(-|~D))
1	1472	483	1955	% Detección de 1	67.84%	Sensitivity (Pr(+| D))
Total general	**3408**	**712**	**4120**			

De inmediato se observa que la precisión para este modelo ω es menor que la del modelo logit por 0.9 %, pero para detección de 1 (Sensibilidad) es mejor que el modelo logit, por 1.83 %, por tanto, para saber que modelo debemos utilizar dependerá del interés del investigador o responsable en hacer este tipo de análisis. Para este caso en particular, se tiene interés en pronósticar a aquellos registros con variable de respuesta "SI", por tanto, podemos concluir que el modelo paramétrico que proporciona mejores resultados es el modelo probit, tomando en cuenta las variables explicativas que se ocuparón. A continuación se muestran los resultados obtenidos para las siguientes dos combinaciones de variables que se utilizarón

Modelo paramétrico 2

$$Y_r = EDAD + GENERO + EDUC + BMX + SANTA + DEPA$$

Tabla de clasificación para

- Logit

Tabla 4.12: Tabla de clasificación para el modelo logit

MODELO LOGIT						
Cuenta de ID	**Real**					
Pronóstico	**0**	**1**	**Total general**	59.49%		
0	1981	242	2223	% Deteccion de 0	58.13%	Specificity (Pr(-|~D))
1	1427	470	1897	% Detección de 1	66.01%	Sensitivity (Pr(+| D))
Total general	**3408**	**712**	**4120**			

- Probit

Tabla 4.13: Tabla de clasificación para el modelo probit

MODELO PROBIT						
Cuenta de ID	**Real**					
Pronóstico	**0**	**1**	**Total general**	58.50%		
0	1928	230	2158	% Deteccion de 0	56.57%	Specificity (Pr(-|~D))
1	1480	482	1962	% Detección de 1	67.70%	Sensitivity (Pr(+| D))
Total general	**3408**	**712**	**4120**			

Igual que en el caso anterior, gana en la Sensibilidad el modelo PROBIT, aunque cabe mencionar que el porcentaje fue menor que en el modelo 1, pues se han quitado variables que influyen en la construcción de probabilidades y por consiguiente en la construcción de la variable Pronóstico.

Modelo paramétrico 3

$$\begin{aligned} Y_r &= EDAD + GENERO + VEP + NDEP + QIM + EDUC + ZM \\ &+ BBVA + BMX + BANORTE + SANTA + HSBC \\ &+ IWMT + SCTBNK + AMEX + DEPA \end{aligned}$$

Tabla de clasificación para

- Logit

Tabla 4.14: Tabla de clasificación para el modelo LOGIT

MODELO LOGIT						
Cuenta de ID	**Real**					
Pronóstico	**0**	**1**	**Total general**	59.37%		
0	1977	243	2220	% Deteccion de 0	58.01%	Specificity (Pr(-|~D))
1	1431	469	1900	% Detección de 1	65.87%	Sensitivity (Pr(+| D))
Total general	**3408**	**712**	**4120**			

- Probit

Tabla 4.15: Tabla de clasificación para el modelo PROBIT

MODELO PROBIT						
Cuenta de ID	**Real**					
Pronóstico	**0**	**1**	**Total general**	58.59%		
0	1940	238	2178	% Deteccion de 0	56.92%	Specificity (Pr(-|~D))
1	1468	474	1942	% Detección de 1	66.57%	Sensitivity (Pr(+| D))
Total general	**3408**	**712**	**4120**			

Para este modelo se ocuparón todas las variables explicativas disponibles y vemos el modelo PROBIT vuelve a dar mejores resultados, aunque lo interesante es que con un porcentaje más bajo que en el primer modelo aún utilizando todas las variables disponibles en nuestra base de datos, lo que

indica que no todas las variables son explicativas, esto tiene fundamento con la tabla de regresión que se obtiene después de ejecutar nuestro código y que se muestran en el Apéndice A.

4.4. Clasificación por máquinas de soporte vectorial

Modelo SVM 1

Empezamos con abrir el programa de *R* llamado *RStudio*, y se empiezan a cargar tanto la base de datos de entrenamiento y de prueba del modelo, cuyos comandos aparecen en el Apéndice B, posteriormente se carga una librería de RStudio llamada *e1071*, que es aquella que contiene los algoritmos necesarios para utilizar las máquinas de soporte vectorial, cuando el software carga el paquete solo basta aplicar un comando especificando, la variable de respuesta y sus independientes así como el tipo de kernel a utilizar, para éste caso ultilizaremos el *sigmoidal* y el Laplaciano. Para cada uno de ellos, generamos la probabilidad de que el registro obtenga un "NO" y con ello podemos obtener la probabilidad de que obtenga un "SI".
Comenzamos con utilizar el modelo *sigmoidal* utilizando la siguiente expresión:

$$\begin{aligned} Y_r &= EDAD + GENERO + NDEP + LNQIM + EDUC + BBVA \\ &+ BMX + SANTA + HSBC + DEPA \end{aligned}$$

Con Y_r la variable de respuesta cuyo posibles valores son "SI" o "NO". Una vez que nuestro modelo ha sido entrenado con los datos de aprendizaje, procedemos a generar la probabilidad (estimada por el modelo) de que un registro con valores particulares para cada variable explicativa obtenga un "SI" o "NO" en la solicitud de su crédito. A continuación mostramos la parte de la tabla que contiene dichas probabilidades.

En la Tabla 4.16, se observan columnas correspondientes al número de registro (ID), Probabilidad de obtener un "NO" (P(Y=0)), Probabilidad de obtener un "SI" (P(Y=1)), el pronóstico de la variable Y que da el modelo (Pronóstico) y finalmente la clasificación real que tuvo el registro al solicitar su tarjeta de crédito.

Tabla 4.16: Tabla de probabilidades para el modelo Sigmoidal

MODELO SIGMOIDAL				
ID	P(y=0)	P(Y=1)	Pronostico	Y_R
1	0.8273556	0.1726444	0	0
2	0.83048502	0.169515	0	1
3	0.84988549	0.1501145	0	0
4	0.8495061	0.1504939	0	0
5	0.85464925	0.1453508	0	0
6	0.83861837	0.1613816	0	0
7	0.83744272	0.1625573	0	0
8	0.83951522	0.1604848	0	0
9	0.75720035	0.2427996	1	0
10	0.85407642	0.1459236	0	0
11	0.84287018	0.1571298	0	0
12	0.83458227	0.1654177	0	0
13	0.82856227	0.1714377	0	0
14	0.83866756	0.1613324	0	1
15	0.8479103	0.1520897	0	0

Para determinar la variable pronóstico, se va a utilizar el punto de corte con el cual se trabajó para el modelo paramétrico usando el mismo procedimiento. Por tanto, elegimos a nuestro umbral como $I = 0.1728431$, y nuestra decisión queda expresada de la siguiente manera:

$$\sigma(x) = \begin{cases} SI \quad , si \quad P(Y = 1|x) > I, \\ \\ NO \quad , si \quad P(Y = 1|x) \leq I. \end{cases}$$

Por tanto, si codificamos a 1 como "SI" y a 0 como "NO", entonces utilizando la fórmula de Excel "=SI()", se obtiene la variable llamada "Pronóstico" tal y como se muestra en la Tabla 4.16
En las siguientes tablas se muestran los resultados obtenidos tanto del modelo *sigmoidal* como el *laplaciano*.

Tal y como muestra la Tabla 4.17, respecto a la precisión del modelo, se tiene un mayor porcentaje utilizando el kernel *sigmoidal* que usando el *laplaciano*, y de hecho aunque el modelo de Laplace tiene un valor mayor en la sensibilidad, podemos ver que tiene resultados muy malos para el pronóstico de la variable Y_r, puesto que dichos valores son muy cercanos al 0.5, probabilidad que se tiene al lanzar una moneda al aire, es decir, realizar

Tabla 4.17: Tabla de clasificación para el modelo *sigmoidal*

MODELO SIGMOIDAL						
Cuenta de ID	**Real**					
Pronóstico	**0**	**1**	**Total general**	64.85%		
0	2369	409	2778	% Deteccion de 0	69.51%	Specificity (Pr(-|~D))
1	1039	303	1342	% Detección de 1	42.56%	Sensitivity (Pr(+| D))
Total general	**3408**	**712**	**4120**			

Tabla 4.18: Tabla de clasificación para el modelo *laplaciano*

MODELO LAPLACEDOT						
Cuenta de ID	**Real**					
Pronóstico	**0**	**1**	**Total general**	53.69%		
0	1819	319	2138	% Deteccion de 0	53.37%	Specificity (Pr(-|~D))
1	1589	393	1982	% Detección de 1	55.20%	Sensitivity (Pr(+| D))
Total general	**3408**	**712**	**4120**			

dicho procedimiento con el modelo de Laplace es "casi equivalente", a lanzar una moneda al aire y así decidir según el resultado de la moneda asignarle un valor a Y_r. Por tanto, el modelo que da mejores resultados para pronósticar la variable de interés es el del kernel *sigmoidal*.

Ahora bien basta con analizar los datos obtenidos por dicho modelo, es decir, la especificidad y la sensibilidad, y así determinar como pronostica nuestro modelo.

Para la sensibilidad, es decir, la probabilidad de que el modelo pronostique que el registro va a aprobar su solicitud de crédito, dado que realmente fue aprobado, es del $0.4256 = \frac{266}{625}$, cuya interpretación es que para el valor $Y_r = 1$, de 625 casos solo llega a acertar o pronosticar correctamente 266 casos, algo quizá no tan favorable, pero también recordemos la tasa de aprobación que se tiene históricamente, es decir, hubo pocos registros cuya decisión fue positiva en la solicitud de su crédito y por el contrario, más personas que fueron rechazadas en su solicitud, lo que influye en la forma en la que aprendió a clasificar nuestro modelo, entonces analizando la especificidad, vemos que tenemos un valor de $0.6951 \approx \frac{139}{200}$, que si se interpreta como en el caso de la sensibilidad, entonces se tiene que de cada 200 casos 130 registros son detectados o se pronóstica que van a ser rechazados o bien que el modelo tiene una probabilidad de $1 - 0.6951 = 0.3049$ de equivocarse en clasificar a un registro como rechazado y que no lo fuese, así que en términos de especificidad

conviene utilizar este modelo y así ir "depurando" a aquellas solicitudes que nuestro modelo pronóstique como "NO" e ir atendiendo las restantes, es decir, para nuestros datos, de las 4,120 solicitudes nuestro modelo va a depurar 2778, con una incertidumbre de 0.3049 de que se equivoque en rechazar a aquellos registros que posiblemente tengan un perfil de aceptación en su solicitud, quedando solo con 1,342 solicitudes.

Modelo SVM 2

$$Y_r = EDAD + GENERO + EDUC + BMX + SANTA + DEPA$$

Tabla de clasificación para

- KERNEL SIGMOIDAL

Tabla 4.19: Tabla de clasificación para el modelo *sigmoidal*

MODELO SIGMOIDAL						
Cuenta de ID	Real					
Pronóstico	0	1	Total general	58.69%		
0	2052	346	2398	% Deteccion de 0	60.21%	Specificity (Pr(-|~D))
1	1356	366	1722	% Detección de 1	51.40%	Sensitivity (Pr(+| D))
Total general	3408	712	4120			

- KERNEL LAPLACIANO

Tabla 4.20: Tabla de clasificación para el modelo *laplaciano*

MODELO LAPLACEDOT						
Cuenta de ID	Real					
Pronóstico	0	1	Total general	49.76%		
0	1794	456	2250	% Deteccion de 0	52.64%	Specificity (Pr(-|~D))
1	1614	256	1870	% Detección de 1	35.96%	Sensitivity (Pr(+| D))
Total general	3408	712	4120			

Para este modelo cuyas variables de explicación son menos que en el modelo anterior vemos que se ve más afectado al momento de predecir la variable Y, tanto en la precisión del modelo ω como para las demás medidas, es decir, la "Especificidad" y "Sensibilidad". Sin embargo, se observan mejores porcentajes para el kernel *sigmoidal*, cuya mejor función tanto para el modelo anterior como para el modelo actual es el de detectar valores 0.

Modelo SVM 3

$$\begin{aligned} Y_r &= EDAD + GENERO + VEP + NDEP + QIM + EDUC + ZM \\ &+ BBVA + BMX + BANORTE + SANTA + HSBC \\ &+ IWMT + SCTBNK + AMEX + DEPA \end{aligned}$$

Tabla de clasificación para

- KERNEL SIGMOIDAL

Tabla 4.21: Tabla de clasificación para el modelo *sigmoidal*

MODELO SIGMOIDAL						
Cuenta de ID	**Real**					
Pronóstico	**0**	**1**	**Total general**	57.52%		
0	2084	426	2510	% Deteccion de 0	61.15%	Specificity (Pr(-|~D))
1	1324	286	1610	% Detección de 1	40.17%	Sensitivity (Pr(+| D))
Total general	**3408**	**712**	**4120**			

- KERNEL LAPLACIANO

Tabla 4.22: Tabla de clasificación para el modelo *laplaciano*

MODELO LAPLACEDOT						
Cuenta de ID	**Real**					
Pronóstico	**0**	**1**	**Total general**	52.28%		
0	1668	226	1894	% Deteccion de 0	48.94%	Specificity (Pr(-|~D))
1	1740	486	2226	% Detección de 1	68.26%	Sensitivity (Pr(+| D))
Total general	**3408**	**712**	**4120**			

Finalmente, para el último modelo donde fueron utilizadas todas las variables explicativas vemos que hay algunos cambios en beneficio para un modelo y otros cambios que no benefician en los valores anteriormente reportados. Analizando los números para el kernel *sigmoidal* se nota un decremento en la precisión del modelo pero un aumento en la detección de valores 0 (Especificidad) y como es de esperar la sensibilidad se decrementó respecto a los dos modelos anteriores. Sin embargo, para el kernel *laplaciano*, vemos que mejora en la precisión del modelo y esto es gracias a la mejora en la detección de 1 (Sensibilidad) cuyo valor es más alto que en los modelos anteriores e inclusive para todos los modelos realizados con el kernel *sigmoidal*, el cuál

Tabla 4.23: Porcentajes globales obtenidos en cada modelo y regresión

	Modelos	
	Logit	Sigmoidal
Regresión 1	59.61 %	64.85 %
Regresión 2	59.49 %	58.69 %
Regresión 3	59.37 %	57.52 %

era el ganador. Ahora bien, si comparamos el modelo logit con el modelo de máquinas de soporte vectorial se puede resumir en la Tabla 4.23 el % de detección global que obtuvieron para cada conjunto de variables regresoras.

Si solo terminaramos con éste análisis podemos concluir que el mejor modelo que logra modelar correctamente la variable Y es el modelo logit, es decir, el modelo paramétrico, con las hipótesis necesarias que se tiene implícitamente en este modelo, por ejemplo, la normalidad de los datos.
Analizando el enfoque de la detección de las solicitudes con mayor probabilidad de no aprobar tenemos la siguiente información

Tabla 4.24: Porcentajes detección de 0 obtenidos en cada modelo y regresión.

	% Detección de 0	
	Logit	Sigmoidal
Regresión 1	58.27 %	69.51 %
Regresión 2	58.13 %	60.21 %
Regresión 3	58.01 %	61.15 %

Y por tanto, podemos quedarnos con el modelo no paramétrico ya que nos garantiza depurar de una "mejor manera", (con probabilidad mayor del 60 %), aquellas solicitudes que no serán aceptadas por el banco.

Conclusiones

Los modelos de *credit scoring* se emplean generalmente para evaluar individuos y pequeñas y/o medianas empresas. Si bien la aplicación más conocida de este tipo de modelos utilizados en el enfoque del *credit scoring* es en la originación de financiamientos, también las entidades financieras lo utilizan con otros propósitos uno de esos puede ser para el diseño de estrategias de *marketing* para ofrecer productos de manera proactiva y masiva [5]. En este trabajo se utilizarón bases de datos de personas físicas a las cuales les fue ofrecida una tarjeta de crédito por parte del banco y la meta fue reforzar la estrategia para que fuera de manera proactiva y al tener como prospecto que es más probable de que aprobará su solicitud optimizando procesos internos del banco.

De igual manera, se pretendió contrastar modelos paramétricos y no paramétricos en este tipo de problema, lo cual nos da la impresión que todo depende de las variables y estructura de la base de datos a analizar. En forma de resumen se puede concluir lo siguiente:

- Para los modelos paramétricos el modelo que mejor predice de manera global a la variable Y es el modelo logit aunque cabe mencionar que el modelo probit obtuvo mejor porcentaje de detección del valor 1 (Si aprobada) cuyo porcentaje promedio fue de 67.37 %. En este caso dependerá del analista a cargo para determinar que modelo utilizar, aunque un enfoque que se puede utilizar es el ocupar el modelo para "depurar" a los solicitantes cuya probabilidad de que no aprueben sea alta y así optimizar tiempo y servicio para el resto de clientes.

- Para los modelos no paramétricos el kernel que mejor predice de forma global a la variable Y es el *Sigmoidal* aunque es notable el comportamiento

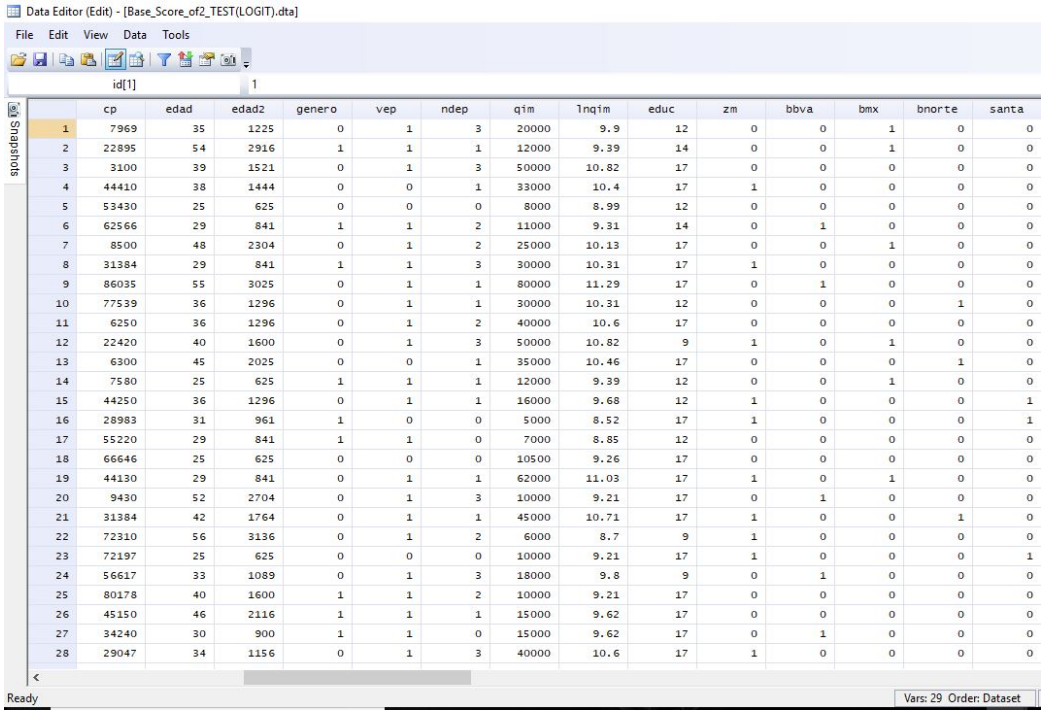
del kernel *Laplaciano* cuando se utilizaron todas las variables disponibles en la base de datos, pues se obtuvo un 68.26 % en la detección de 1, es decir, que la solicitud sea aprobada, como se mencionó anteriormente dependerá del analista y sus criterios para saber que modelo utilizar, aunque por los resultados el kernel sigmoidal obtiene mayor porcentaje de predicción global que el de Laplace.

La conclusión encontrada con el desarrollo de éste trabajo es que el modelo no paramétrico soluciona de una mejor manera la problemática de interés. Existen ventajas y desventajas en éste tipo de modelos pero una de las ventajas que cabe mencionar es que no es necesario la hipótesis de normalidad de los datos, que en varios casos de situaciones reales así sucede, por ende puede ser una opción trabajar con el universo de modelos propuestos en ésta categoría sin dejar de lado la importancia de la estadística clásica donde se fundamentan los modelos no paramétricos.

Apéndice A

En éste apéndice se muestran los registros de los códigos empleados en el trabajo de tesis.

Modelos probit y logit estimado utilizando *STATA*.



Data Editor (Edit) - [Base_Score_of2_TEST(LOGIT).dta]

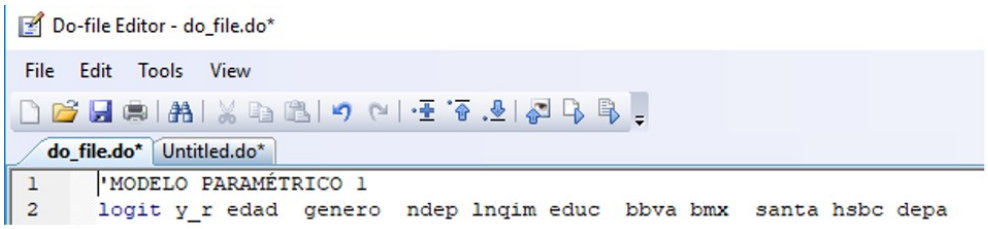
File Edit View Data Tools

id[1] 1

	cp	edad	edad2	genero	vep	ndep	qim	lnqim	educ	zm	bbva	bmx	bnorte	santa
1	7969	35	1225	0	1	3	20000	9.9	12	0	0	1	0	0
2	22895	54	2916	1	1	1	12000	9.39	14	0	0	1	0	0
3	3100	39	1521	0	1	3	50000	10.82	17	0	0	0	0	0
4	44410	38	1444	0	0	1	33000	10.4	17	1	0	0	0	0
5	53430	25	625	0	0	0	8000	8.99	12	0	0	0	0	0
6	62566	29	841	1	1	2	11000	9.31	14	0	1	0	0	0
7	8500	48	2304	0	1	2	25000	10.13	17	0	0	1	0	0
8	31384	29	841	1	1	3	30000	10.31	17	1	0	0	0	0
9	86035	55	3025	0	1	1	80000	11.29	17	0	1	0	0	0
10	77539	36	1296	0	1	1	30000	10.31	12	0	0	0	1	0
11	6250	36	1296	0	1	2	40000	10.6	17	0	0	0	0	0
12	22420	40	1600	0	1	3	50000	10.82	9	1	0	1	0	0
13	6300	45	2025	0	0	1	35000	10.46	17	0	0	0	1	0
14	7580	25	625	1	1	1	12000	9.39	12	0	0	1	0	0
15	44250	36	1296	0	1	1	16000	9.68	12	1	0	0	0	1
16	28983	31	961	1	0	0	5000	8.52	17	1	0	0	0	1
17	55220	29	841	1	1	0	7000	8.85	12	0	0	0	0	0
18	66646	25	625	0	0	0	10500	9.26	17	0	0	0	0	0
19	44130	29	841	0	1	1	62000	11.03	17	1	0	1	0	0
20	9430	52	2704	0	1	3	10000	9.21	17	0	1	0	0	0
21	31384	42	1764	0	1	1	45000	10.71	17	1	0	0	1	0
22	72310	56	3136	0	1	2	6000	8.7	9	1	0	0	0	0
23	72197	25	625	0	0	0	10000	9.21	17	1	0	0	0	1
24	56617	33	1089	0	1	3	18000	9.8	9	0	1	0	0	0
25	80178	40	1600	1	1	2	10000	9.21	17	0	0	0	0	0
26	45150	46	2116	1	1	1	15000	9.62	17	0	0	0	0	0
27	34240	30	900	1	1	0	15000	9.62	17	0	1	0	0	0
28	29047	34	1156	0	1	3	40000	10.6	17	1	0	0	0	0

Ready Vars: 29 Order: Dataset

Tabla 4.25: Base de datos cargada en el software *STATA*




```
. logit y_r edad  genero  ndep lnqim educ  bbva bmx  santa hsbc depa

Iteration 0:   log likelihood = -1896.5292
Iteration 1:   log likelihood = -1798.2619
Iteration 2:   log likelihood = -1792.0261
Iteration 3:   log likelihood =  -1791.988
Iteration 4:   log likelihood = -1791.9879

Logistic regression                               Number of obs   =       4120
                                                  LR chi2(10)     =     209.08
                                                  Prob > chi2     =     0.0000
Log likelihood = -1791.9879                       Pseudo R2       =     0.0551

------------------------------------------------------------------------------
         y_r |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
       edad |   .0267793    .004152     6.45   0.000     .0186415    .0349171
     genero |   .4439413   .0894842     4.96   0.000     .2685556     .619327
       ndep |  -.0206958   .0350509    -0.59   0.555    -.0893944    .0480027
     lnqim |  -.0237078    .064238    -0.37   0.712    -.1496119    .1021963
       educ |   .0834391   .0169811     4.91   0.000     .0501567    .1167214
       bbva |   .1280612   .1291477     0.99   0.321    -.1250638    .3811861
        bmx |   .6535577   .1176395     5.56   0.000     .4229885     .884127
      santa |    .496601   .1489158     3.33   0.001     .2047314    .7884707
       hsbc |   .4022992   .2104437     1.91   0.056    -.0101629    .8147614
       depa |  -1.026873   .2012697    -5.10   0.000    -1.421354   -.6323912
      _cons |  -4.035421    .633083    -6.37   0.000    -5.276241   -2.794601
------------------------------------------------------------------------------
```

Tabla 4.26: Código y resultados para el modelo logit usando el modelo 1.

Do-file Editor - do_file.do*

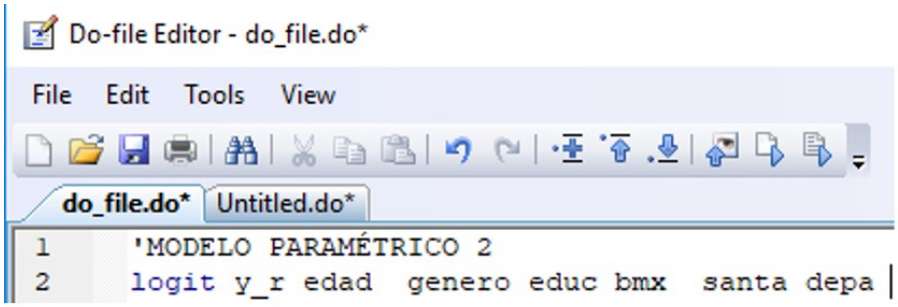
File Edit Tools View

do_file.do* Untitled.do*

```
1    'MODELO PARAMÉTRICO 1
2    probit y_r edad  genero  ndep lnqim educ  bbva bmx  santa hsbc depa
3
```

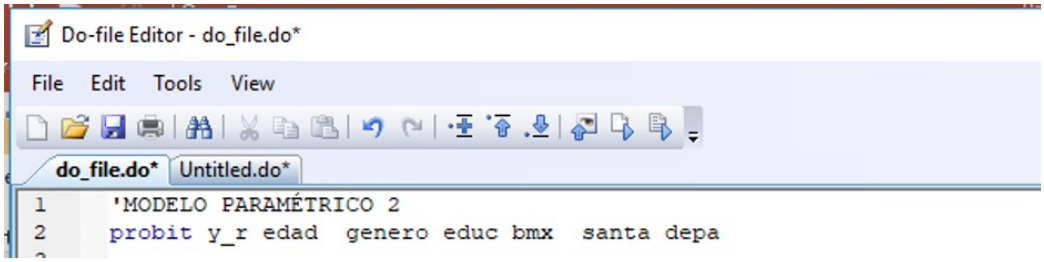
```
. probit y_r edad  genero  ndep lnqim educ  bbva bmx  santa hsbc depa

Iteration 0:   log likelihood = -1896.5292
Iteration 1:   log likelihood = -1794.1708
Iteration 2:   log likelihood =  -1792.358
Iteration 3:   log likelihood = -1792.3562
Iteration 4:   log likelihood = -1792.3562

Probit regression                                 Number of obs   =       4120
                                                  LR chi2(10)     =     208.35
                                                  Prob > chi2     =     0.0000
Log likelihood = -1792.3562                       Pseudo R2       =     0.0549

------------------------------------------------------------------------------
         y_r |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
       edad |    .015074   .0023444     6.43   0.000      .010479     .019669
     genero |   .2478412   .0504703     4.91   0.000      .1489212    .3467612
       ndep |  -.0123151   .0194701    -0.63   0.527     -.0504758    .0258457
      lnqim |  -.0088359   .0361198    -0.24   0.807     -.0796294    .0619577
       educ |   .0448782   .0092746     4.84   0.000      .0267003    .0630562
       bbva |    .070361   .0711264     0.99   0.323     -.0690442    .2097662
        bmx |   .3704429   .0661195     5.60   0.000       .240851    .5000349
      santa |   .2804026   .0840844     3.33   0.001      .1156002    .4452051
       hsbc |   .2223233    .119487     1.86   0.063      -.011867    .4565135
       depa |  -.5242759   .0993155    -5.28   0.000     -.7189307    -.329621
      _cons |  -2.336854     .35364    -6.61   0.000     -3.029976   -1.643732
------------------------------------------------------------------------------

.
end of do-file
```

Tabla 4.27: Código y resultados para el modelo probit usando el modelo 1.




```
. logit y_r edad  genero educ bmx  santa depa

Iteration 0:   log likelihood = -1896.5292
Iteration 1:   log likelihood = -1800.2321
Iteration 2:   log likelihood = -1794.0748
Iteration 3:   log likelihood = -1794.0372
Iteration 4:   log likelihood = -1794.0371

Logistic regression                             Number of obs   =       4120
                                                LR chi2(6)      =     204.98
                                                Prob > chi2     =     0.0000
Log likelihood = -1794.0371                     Pseudo R2       =     0.0540

------------------------------------------------------------------------------
         y_r |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
        edad |   .0259682   .0040189     6.46   0.000     .0180914    .033845
      genero |   .4627019   .0853821     5.42   0.000      .295356    .6300478
        educ |   .0834168   .0161719     5.16   0.000     .0517204    .1151132
         bmx |   .5579028   .0954567     5.84   0.000     .3708111    .7449944
       santa |   .4053561    .132024     3.07   0.002     .1465938    .6641184
        depa |  -1.118881   .1887174    -5.93   0.000    -1.488761   -.749002
       _cons |  -4.174701   .3252099   -12.84   0.000    -4.812101   -3.537301
------------------------------------------------------------------------------

.
```

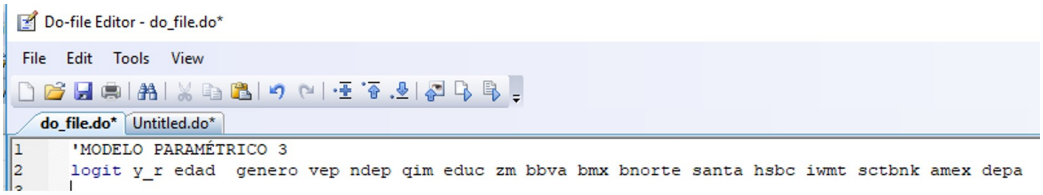
Tabla 4.28: Código y resultados para el modelo logit usando el modelo 2.




```
. probit y_r edad  genero educ bmx  santa depa

Iteration 0:   log likelihood = -1896.5292
Iteration 1:   log likelihood = -1796.1394
Iteration 2:   log likelihood = -1794.3529
Iteration 3:   log likelihood = -1794.3511
Iteration 4:   log likelihood = -1794.3511

Probit regression                                 Number of obs   =       4120
                                                  LR chi2(6)      =     204.36
                                                  Prob > chi2     =     0.0000
Log likelihood = -1794.3511                       Pseudo R2       =     0.0539
```

y_r	Coef.	Std. Err.	z	P>|z|	[95% Conf.	Interval]
edad	.0146643	.0022623	6.48	0.000	.0102303	.0190983
genero	.2577063	.0482221	5.34	0.000	.1631927	.3522198
educ	.0452593	.0088042	5.14	0.000	.0280034	.0625153
bmx	.3178974	.0546608	5.82	0.000	.2107643	.4250305
santa	.2301751	.0753639	3.05	0.002	.0824645	.3778857
depa	-.5756624	.0918709	-6.27	0.000	-.755726	-.3955988
_cons	-2.380474	.1775254	-13.41	0.000	-2.728417	-2.032531

Tabla 4.29: Código y resultados para el modelo probit usando el modelo 2.




```
. logit y_r edad  genero vep ndep qim educ zm bbva bmx bnorte santa hsbc iwmt sctbnk amex depa

Iteration 0:   log likelihood = -1896.5292
Iteration 1:   log likelihood =  -1792.712
Iteration 2:   log likelihood = -1786.1266
Iteration 3:   log likelihood = -1786.0873
Iteration 4:   log likelihood = -1786.0873

Logistic regression                             Number of obs   =       4120
                                                LR chi2(16)     =     220.88
                                                Prob > chi2     =     0.0000
Log likelihood = -1786.0873                     Pseudo R2       =     0.0582

------------------------------------------------------------------------------
         y_r |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
       edad |   .0276363   .0042041     6.57   0.000     .0193964    .0358763
     genero |   .4523598   .0890327     5.08   0.000     .2778588    .6268607
        vep |  -.0429591    .095293    -0.45   0.652    -.2297301    .1438118
       ndep |  -.0147515   .0376634    -0.39   0.695    -.0885704    .0590674
        qim |  -1.46e-06   1.90e-06    -0.77   0.443    -5.18e-06    2.26e-06
       educ |   .0802287   .0167293     4.80   0.000     .0474398    .1130175
         zm |   .1009969   .0870044     1.16   0.246    -.0695286    .2715224
       bbva |   .1644933   .3067452     0.54   0.592    -.4367162    .7657028
        bmx |   .6870078   .3022752     2.27   0.023     .0945593    1.279456
     bnorte |   .0668435   .3331691     0.20   0.841    -.5861561     .719843
      santa |   .5296968   .3155499     1.68   0.093    -.0887696    1.148163
       hsbc |    .441566   .3490513     1.27   0.206    -.2425621    1.125694
       iwmt |  -.3069364   .3405394    -0.90   0.367    -.9743814    .3605086
     sctbnk |    .512785   .4511974     1.14   0.256    -.3715456    1.397116
       amex |   .4448261   .3568943     1.25   0.213    -.2546738    1.144326
       depa |  -.9957111   .3436927    -2.90   0.004    -1.669336   -.3220859
      _cons |  -4.278957   .4435764    -9.65   0.000     -5.14835   -3.409563
------------------------------------------------------------------------------
```

Tabla 4.30: Código y resultados para el modelo logit usando el modelo 3.

Do-file Editor - do_file.do*

File Edit Tools View

do_file.do* Untitled.do*

```
1     'MODELO PARAMÉTRICO 3
2     probit y_r edad  genero vep ndep qim educ zm bbva bmx bnorte santa hsbc iwmt sctbnk amex depa
```

```
. probit y_r edad  genero vep ndep qim educ zm bbva bmx bnorte santa hsbc iwmt sctbnk amex depa

Iteration 0:   log likelihood = -1896.5292
Iteration 1:   log likelihood = -1788.1691
Iteration 2:   log likelihood = -1786.2959
Iteration 3:   log likelihood =  -1786.294
Iteration 4:   log likelihood =  -1786.294

Probit regression                                 Number of obs   =       4120
                                                  LR chi2(16)     =     220.47
                                                  Prob > chi2     =     0.0000
Log likelihood =  -1786.294                       Pseudo R2       =     0.0581

------------------------------------------------------------------------------
         y_r |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
        edad |   .0156324   .0023799     6.57   0.000     .0109679    .020297
      genero |   .2530831   .0502038     5.04   0.000     .1546855    .3514807
         vep |  -.0259066   .0539782    -0.48   0.631    -.1317019    .0798886
        ndep |  -.0088477   .0209966    -0.42   0.673    -.0500003    .0323049
         qim |  -7.48e-07   1.06e-06    -0.71   0.480    -2.82e-06    1.33e-06
        educ |   .0430889   .0091488     4.71   0.000     .0251576    .0610202
          zm |   .0604603   .0488978     1.24   0.216    -.0353775    .1562981
        bbva |   .1097395   .1702494     0.64   0.519    -.2239431    .4434221
         bmx |   .4083165   .1683397     2.43   0.015     .0783767    .7382563
      bnorte |   .0642753    .184248     0.35   0.727    -.2968442    .4253947
       santa |   .3174081   .1761044     1.80   0.071    -.0277501    .6625663
        hsbc |   .2648816   .1955819     1.35   0.176     -.118452    .6482151
        iwmt |  -.1458371   .1862192    -0.78   0.434    -.5108199    .2191458
      sctbnk |    .305385   .2553459     1.20   0.232    -.1950837    .8058537
       amex |   .2661713   .2002165     1.33   0.184    -.1262458    .6585885
       depa |  -.4890222   .1839661    -2.66   0.008    -.8495892   -.1284552
      _cons |  -2.454215   .2455421   -10.00   0.000    -2.935468   -1.972961
------------------------------------------------------------------------------
```

Tabla 4.31: Código y resultados para el modelo probit usando el modelo 3.

Apéndice B

Modelos probit y logit estimado utilizando *R*.

El código usado para los demás modelos de SVM es similar a los que se presentan a continuación, con la única diferencia de las variables utilizadas.

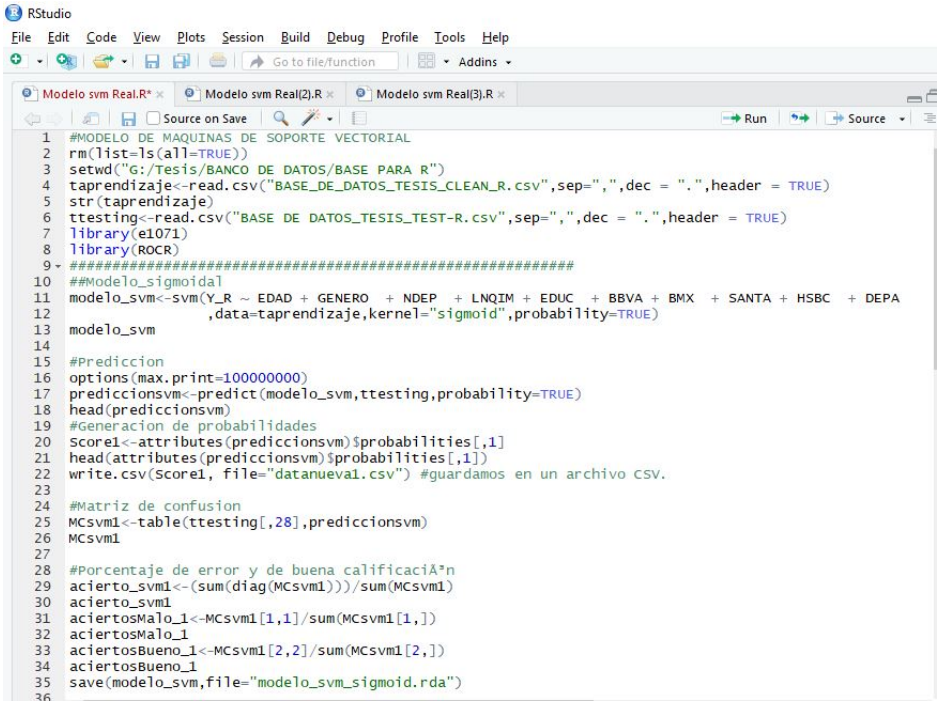


Tabla 4.32: Código y resultados para el modelo *sigmoidal*.

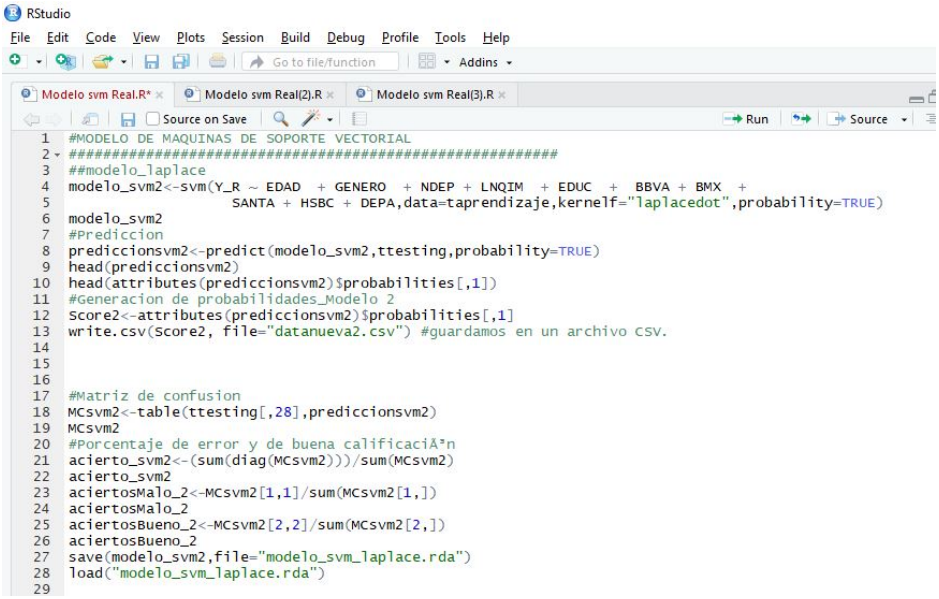


Tabla 4.33: Código y resultados para el modelo *Laplace*.

Bibliografía

[1] Banco de México. *Sistema Financiero - Calculadoras del CAT (Costo Anual Total).* Banxico.org.mx. Consultado el 27 de abril de 2017, desde *http : //www.banxico.org.mx/ waCalculadoraTarjetaCredito/MasInformacion.jsp*

[2] Banco de México. *Definciones básicas de Riesgo.*(Consultado el: 26 de abril de 2017.) Disponible en:*http : //www.banxico.org.mx/sistema − financiero/material − educativo/intermedio/riesgos/ %7BA5059B92 − 176D − 0BB6 − 2958 − 7257E2799FAD %7D.pdf*

[3] Cabeza Lambán, M., & Torra Porras, S. (2007). *El Riesgo en la empresa* (1st ed.). Ithaca, NY: Palisade Corporation.

[4] Espin García, O. & Rodríguez-Caballero. C. (2013). *Metodología para un scoring de clientes sin referencias crediticias.* Cuadernos de Economía, 32(59),XX-XX.

[5] Gutiérrez Girault Matías Alfredo (2007). *Modelos de Credit Scoring - Qué, Cómo, Cuándo y para qué-.*

[6] Karlis, D. Rahmouni, M. (2007). *Analysis of defaulters behavior using the Poisson mixture approach.* Journal of Management Mathematics, 18, 2007.

[7] Long, J. and Freese, J. (2001). *Regression models for categorial dependent variables using Stata.* Texas: Stata, pp.99-136.

[8] Manel Martínez, R. (2018). *Introducción a los métodos Kernel* Arantxa.ii.uam.es. Consultado el 26 Abril 2018, desde *http :*

*//arantxa.ii.uam.es/ − jms/seminarios$_d$octorado/abstracts*2007 − 2008/20080429*MMartinez.pdf*.

[9] Morales Guerra, M.L. (2007). *La administración del riesgo de crédito en la cartera de consumo de una institución bancaria.* Tesis Licenciatura Universidad de San Carlos de Guatemala Facultad de Ciencias Económicas.

[10] Nieto Murillo, S. (2010). *Crédito al consumo: La estadística aplicada a un problema de riesgo crediticio.* (Licenciatura). Universidad Autónoma Metropolitana.

[11] Rodríguez Caballero, C.V. (2009). *La inferencia bayesiana en la administración de riesgos.* Libro colectivo de administración de riesgos financieros. Grupo de investigación de mercados e instituciones financieras. Vol II.

[12] Segovia, A., & García, C. (2012). *Las tarjetas de crédito: úsalas a tu favor.* Profeco.gob.mx. Consultado el 27 de Abril 2017, desde *http : //www.profeco.gob.mx/encuesta/brujula /bruj_*2012/*bol*217*_tarjeta_de_credito.asp*

[13] Tan, P., Steinbach, M. and Kumar, V. (2015). *Introduction to data mining.* Dorling Kindersley: Pearson, pp. 256-276

[14] Torrico Salamanca, S.(2015) *Macro credit scoring como propuesta para cuantificar el riesgo de crédito.* Investigación & desarrollo, 14(2), pp.42-63

[15] Turrent, E.*Historia sintética de la banca en México.* Banco de México. (Consultado el: 26 de abril de 2017.) Disponible en *http : //www.banxico.org.mx/sistema − financiero/material − educativo/basico/ %7BFFF*17467 − 8*ED*6 − 2*AB*2 − 1*B*3*B* − *ACCE*5*C*2*AF*0*E*6 *%7D*.