Universidad Nacional Autónoma de México

Facultad de Ciencias

Modelo de Administración del Riesgo de Crédito enfocado en la probabilidad de incumplimiento

T E S I S

Que para obtener el título de
Actuaría

Presenta
Luis Andrés Escareño Hernández.

Director de tesis:
Act. José Antonio Reyes León.

Ciudad Universitaria, Cd. Mx., 2018

Hoja de Datos del Jurado

1. Datos del alumno
Escareño
Hernández
Luis Andrés
55-50-58-35-92
Universidad Nacional Autónoma de México
Facultad de Ciencias
Actuaría
309029005

2. Datos del tutor
Act.
José Antonio
Reyes
León

3. Datos del sinodal 1
Dr.
Yuri
Salazar
Flores

4. Datos del sinodal 2
M. en C.
José Salvador
Zamora
Muñoz

5. Datos del sinodal 3
Act.
Alejandro
Santoyo
Cano

6. Datos del sinodal 4
Act.
Eduardo Selim
Martínez

Mayorga

7.Datos del trabajo escrito.
Modelo de Administración del Riesgo de Crédito enfocado en la probabilidad de incumplimiento
124 p
2018

Agradecimiento

Agradezco a la Universidad Nacional Autónoma de México y a la Facultad de Ciencias por haberme brindado la mejor formación académica posible, la oportunidad de desarrollarme en ámbitos profesionales y personales, y siempre ser mi segundo hogar.

A mis padres por su apoyo incondicional y por ser una guía en cada etapa de mi vida. A quienes les agradezco todas las experiencias que he vivido y siempre incentivarme a seguir adelante.

A mis hermanas por darme un ejemplo de éxito que seguir.

A la Dra. Lizbeth Naranjo por su tiempo y apoyo en todo momento.

Al Dr. José de Jesús Arias por su ayuda en cada etapa de este trabajo y por ser un gran ejemplo a seguir como maestro y amigo.

A la M. en C. Maria Leonor Zamora Maass por compartir su tiempo y conocimientos.

A mis sinodales por darle soporte a este trabajo.

Y a mis amigos y familia por su confianza y apoyo constante.

(JMLJIAAIYCMICYATTLEG)

Índice general

Índice de figuras

Introducción

La probabilidad de incumplimiento es la probabilidad de que un acreditado no cumpla con sus obligaciones de pago en tiempo y forma tal y como lo define la CNBV (Comisión Nacional Bancaria y de Valores). La comisión establece que dicha probabilidad debe de ser calculada con una regresión logística considerando variables referentes al préstamo cuando los días de atraso no superan los 90. Algunas variables que integran esta regresión son el número de días de atraso con respecto al último pago, porcentaje que represente el saldo de crédito hasta cierta fecha y el motivo del crédito.

El objetivo de esta tesis es calcular la probabilidad de incumplimiento sobre una base de datos simulada procedente de una cartera real de cierto banco, utilizando regresión logística. El motivo de utilizar este modelo es por las características de la variable dependiente, ya que es de tipo categórica indicando si los días de atraso de pago del solicitante superan los 90 días o no.

En este trabajo se presenta un modelo con variables diferentes a las establecidas por la CNBV, con la intención de encontrar las características que tienen mayor y menor impacto en el incumplimiento de un crédito. Las razones de utilizar variables diferentes son la ausencia de algunas de ellas en la base original del banco y para proponer nuevas aspectos al momento de calcular la probabilidad de incumplimiento.

Esta tesis está estructurada de la siguiente manera:

- Capítulo 1: Antecedentes del Riesgo
  Se realiza un breve recorrido por las aportaciones de diversas personalidades en el estudio de la probabilidad y estadística, considerando estas dos ramas de las matemáticas elementos que tiene la capacidad de analizar ciertos riesgos.

- Capítulo 2: Administración de Riesgos
  En este capítulo se integran elementos referentes a la administración de riesgos con los desgloses de los diversos tipos de riesgos que existen en la actualidad. Además, se introduce una técnica que mide los factores de riesgo de forma efectiva.

- Capítulo 3: Riesgo de Crédito
  Se presenta un contexto concreto del riesgo de crédito, incluyendo como es que surgió hasta como en la actualidad se implementa la medición de este riesgo en todo el mundo por medio del Comité de Basilea. Además, se anexan las características del sistema financiero mexicano, las normas que las instituciones bancarias toman en cuenta para otorgar los créditos entre otros aspectos referentes a esté riesgo.

- Capítulo 4: Regresión lineal y logística
  Iniciando con la regresión lineal, el objetivo de este capítulo es dar un contexto de cómo funciona una regresión y qué aspectos debe de cumplir, y así poder conocer de forma más profunda la regresión logística, incluyendo los aspectos especiales de está.

- Capítulo 5: Modelo de Regresión Logística
  En este capítulo se presentan las variables que integran el modelo con sus respectiva descripción y codificación. Asimismo, se presentan los resultados producidos por la aplicación de la regresión logística sobre la base simulada, considerando la selección de variables, pruebas de ajuste, análisis de residuos y validación del modelo.

Capítulo 1

Antecedentes del Riesgo

La incertidumbre siempre está presente en cada una de las decisiones que toma cualquier empresa o institución. Determinar todos los elementos que se puedan presentar ante la ocurrencia de un siniestro y sus posibles consecuencias es imposible, por lo que es común que se asocie la incertidumbre con el concepto de riesgo.

La palabra riesgo proviene del árabe *rizq* con una derivación del latín *risicare* que significa atreverse o transitar un sendero peligroso, también se tiene la ramificación del italiano *rischio*, que significa lo que nos depara la providencia o el destino. Estas expresiones dan una sensación más cercana de lo que se piensa cuando hablamos de riesgo, una situación incierta y peligrosa con resultados que no se pueden predecir totalmente.

La presencia del riesgo en cualquier estrategia que determine algún cambio es inevitable, por lo que se han desarrollado múltiples investigaciones para poder medir de manera más precisa la incertidumbre. Por ello dos de las ramas de las matemáticas más significativas que sirven para predecir acontecimientos son la probabilidad y la estadística, dada la estrecha relación que tiene el azar con el riesgo y cómo es posible medir sus efectos en escenarios determinados.

1.1. Contexto de la Probabilidad

En este fragmento se realizará un pequeño recorrido sobre la historia de la probabilidad con el objetivo de entender de una manera más amplia todas las implicaciones que conlleva y cómo se han analizado las diferentes problemáticas.

La probabilidad es una de las más importantes herramientas que se tienen para poder realizar una correcta medición de los riesgos de cualquier proceso de decisión, ya que, a través de la probabilidad es posible medir pérdidas en un contexto de incertidumbre con un razonamiento integral. Por lo que es útil revisar algunos de los aportes más importantes que se han integrado a esta importante rama de las matemáticas y comó ha evolucionado la medición de los riesgos durante este tiempo.

Los orígenes de la probabilidad provienen del Renacimiento, ya que durante este periodo se produjo un cambio en la forma en la que se veía el mundo adoptando las nociones que los griegos y romanos utilizaron años atrás, al incorporar un entorno en el que el azar y la suerte tenían una importante relevancia (Gregoria Mateos, 2002).

Una de las primeras menciones del concepto de probabilidad se atribuye a Girolamo Cardano (1500-1571), quien escribió múltiples estudios sobre análisis de juegos de azar, principalmente los dados. Los avances más importantes de Cardano se encuentran en su libro *Liber de Ludo Alea (Libro de juegos de azar)*, considerado uno de los mejores manuales para un jugador de la época, en el cual se desarrolló el primer progreso teórico en las leyes de la probabilidad. Además incorporó el término probable definiéndolo como eventos en donde cada resultado es incierto, descripción que se relaciona con el concepto de riesgo. Estas primeras aportaciones se pueden complementar con los trabajos de Galileo (1564-1642), en *Sopra le Scoperte dei Dadi (Jugando a los dados)*, centrados en el juego de los dados generando algunos de los primeros aportes en combinatoria al tratar de contabilizar los posibles resultados al tirar tres dados.

Los herederos de los aportes dados por Cardano y Galileo fueron tres grandes personajes que consolidaron las bases de la teoría de la probabilidad: Blaise Pascal (1623-1662), Pierre de Fermat (1601-1665) y Antoine Gombaud (1607-1684), mejor conocido como Chevalier de Mére. La intercomunicación entre estos tres personajes comenzó con el desafío lanzado a Pascal sobre el problema planteado doscientos años atrás por Luca Pacioli, el cual consistía en la forma en la que se debían repartir las apuestas de un juego de azar si éste fuera interrumpido. La correspondencia entre Pascal y Fermat, sobre el problema antes planteado y otros más, resultó en la creación de la Teoría de la Probabilidad.

Siguiendo las bases de Fermat, Gombaud y Pascal, se encuentran los aportes de Christiaan Huygens (1629-1695) presentando el mejor informe en esos años de probabilidad en su libro *De ratiociniis in ludo aleae (Sobre los cálculos en los Juegos de Azar)* elaborado en 1656. En esta obra se integra su aportación más significativa al introducir la noción de esperanza (o valor esperado). Además añadió 14 proposiciones tocando temas como el Problema de los Puntos con dos jugadores, tres jugadores y proposiciones sin demostrar, las cuales, serían resueltas en años subsecuentes por matemáticos como James Bernoulli o Abraham de Moivre.

James Bernoulli, uno de los sucesores del trabajo de Huygens, escribió el libro *Ars Conjectandi (El Arte de la Conjetura)* que está dividido en 4 partes. La primera parte consiste en una continuación de los resultados de Huygens, desarrollando la deducción de la distribución binomial. La segunda parte hace énfasis en las permutaciones, combinatoria, y sucesiones de números racionales (números de Bernoulli). La siguiente parte analiza las dificultades que surgen en los juegos de azar con respecto a su duración proponiendo numerosos problemas con sus respectivas soluciones. La última parte contiene la demostración de la Ley de los grandes números, idea analizada por Cardano sin ninguna demostración, que explica por qué la media de una muestra tomada al azar de una población con un tamaño suficientemente grande tiende a aproximarse al promedio de la población de la que fue tomada.

Unos años después de los descubrimientos de Bernoulli, Abraham de Moivre (1667-1754) incorpora en su obra *The Doctrine of Chances (La Doctrina de las Probabilidades,1718)* el concepto de la función generadora de probabilidad para variables discretas. Además extiende el trabajo de James Bernoulli incrementando el número de eventos de su distribución binomial llegando a una nueva clasificación de fenómenos aleatorios. Esto último resultó en el descubrimiento de la distribución Normal o campana de Gauss como se le llama erróneamente, ya que,

en realidad fue descubrimiento de Moivre en la "Doctrina de las Probabilidades" y ampliada por Laplace años más tarde. Éste descubrimiento es uno de los principales pilares del Teorema Central del Límite, uno de los más importantes de la probabilidad, que permite aproximar una serie de variables aleatorias con características especiales en su esperanza y varianza con la distribución normal estándar.

Años más tarde el discípulo de Moivre, Thomas Bayes (1702-1761) propuso medir la ocurrencia de un evento desconocido entre ciertos límites a partir de la información dada por una muestra. Esto implicó un gran avance en la forma en la que se calculaba la probabilidad en esos años presentando una forma de obtener las proporciones de las posibles causas por las que ocurrió un cierto evento que se ha observado. Por ejemplo, con la información suficiente se tiene la capacidad de efectuar pronósticos sobre qué tan probable es tener cáncer dado que se presentan tumores con sólo la información que contiene la muestra de la frecuencia de tumores presentados en casos comprobados de cáncer. Este aporte por parte de Bayes fue publicado en 1763, dos años después de su muerte, con el nombre de *Essay Towards Solving a Problem in the Doctrine of Chances (Ensayo para la solución de un problema en la doctrina de posibilidades)*.

En 1812, Pierre-Simon Laplace (1749-1827) formalizó la teoría clásica de la probabilidad. Laplace publicó en *Théorie Analytique des Probabilités (Teoría Analítica de las Probabilidades)* la antes mencionada distribución normal. Otra de las obras principales de Laplace fue *Essai Philosophique Sur Les Probabilites (Ensayo Filosófico sobre la Probabilidad)*, en donde recopila los descubrimientos más relevantes sobre probabilidad hasta esos tiempos, como las contribuciones de Bernoulli, de Moivre, de Cardano, entre otros.

Tiempo después, el discípulo de Laplace, Siméon Denis Poisson (1781-1840) presentaría en 1838, el descubrimiento de una nueva función probabilística conocida como distribución Poisson, la cual está incluida en *Recherches sur la Probabilité des Jugements en Matières Criminelles et Matière Civile (Investigación sobre la probabilidad de los juicios en materias criminales y civiles)*. En esta obra Poisson presenta una generalización de la distribución binomial, apoyándose en la Ley de los grandes números, estresando el número de ensayos realizados llegando al hecho de que se aproximaban a la distribución Poisson, con la cual logró justificar sus resultados basándose en datos estadísticos de esos años.

Durante el final del siglo XIX e inicios del XX, se presentaron otras grandes aportaciones por parte de matemáticos provenientes de Rusia, como Pafnuti Chebyshov (1821-1894), quien es conocido por la desigualdad que lleva su nombre descubierta en 1867, Andréi Márkov (1856-1922) conocido por las cadenas (series de eventos) y la desigualdad que llevan su nombre y Aleksandr Liapunov (1857-1918) con sus aportes en ecuaciones diferenciales y en la teoría de la probabilidad. Años más tarde, Andréi Kolmogórov (1903-1987) formaría parte de este grupo con sus aportes basados en el desarrollo de la probabilidad a partir de la Teoría de Conjuntos y su descubrimiento conjunto con Sydney Chapman (1888-1970) en la Ley de Chapman-Kolmogórov.

Como se puede observar, en las contribuciones que se han dado en el campo de la probabilidad se tiene la inquietud de poder predecir de forma cada vez más precisa las consecuencias de ciertos eventos, como puede ser el resultado de un juego de azar o la posibilidad de lluvia en una localidad. Con esto podemos reafirmar que cada descubrimiento que sea incorporado a la teoría de la probabilidad tiene la intención de medir cada vez más la incertidumbre que rodea algún evento y tener la capacidad de poder tomar decisiones más precisas. Es por ello que la probabilidad es una ciencia matemática muy importante que se tiene en la actualidad para poder medir el riesgo con la información adecuada.

1.2. Cimientos de la Estadística

Al conocer los aportes más importantes que se han realizado en la teoría de la probabilidad, es necesario recapitular la historia de otra ciencia matemática que será usada en esta tesis. La estadística se integra con la finalidad de poder proporcionar un panorama en donde sea posible deducir bajo qué principios son regidos los fenómenos que podrían presentarse en un estudio. Con la posibilidad de hacer predicciones basadas en fundamentos matemáticos, con la intención de poder responder las interrogantes que surgen sobre los eventos futuros.

El origen de la palabra estadística según la Real Academia Española proviene del alemán *Statistik* y del italiano *Statista* que en conjunto significan hombre de Estado, debido a que en sus orígenes. La estadística tenía el propósito de medir el nivel económico de la población y poder cobrar los impuestos correctos con respecto a la información registrada, como podrían ser el número de familias, cabezas de ganado, posibles reclutas para las campañas, número de propiedades dedicadas al cultivo, entre otras cosas.

Los primeros usos de la estadística provienen de efectuar registros a la población con el objetivo de recolectar recursos para diferentes propósitos. Un ejemplo es el primer censo encontrado durante el antiguo Egipto en el año 3050 a. de C., en el que se recopiló información básica de la población y el grado de riqueza con el propósito de poder sustentar la edificación de las pirámides. Otro de los primeros censos levantados se elaboró en China en el año 2238 a. de C. con el objetivo de recolectar de forma eficiente los impuestos registrando la información de la población. No obstante, fueron los griegos quienes desarrollaron registros tributarios, sociales y militares de una forma periódica, ya que, de estas anotaciones se calculaban los impuestos, la fuerza militar y cuanta población podía ejercer el derecho del voto.

A pesar de los avances que generaron los egipcios y griegos en sus inicios, se contempla que los romanos fueron la civilización que optimizó estos registros de una forma aún más eficiente y desarrollada:

"Pero fueron los romanos, maestros de la organización política, quienes mejor supieron emplear los recursos de la estadística. Cada cinco años realizaban un censo de la población y sus funcionarios públicos tenían la obligación de anotar nacimientos, defunciones y matrimonios, sin olvidar los recuentos periódicos del ganado y de las riquezas contenidas en las tierras conquistadas. Para el nacimiento de Cristo sucedía uno de estos empadronamientos de la población bajo la autoridad del imperio."(Ruiz Muñoz, 2004)

Excluyendo el trabajo de Sebastián Münster (1488-1552), quien en 1540 realizó un estudio estadístico que recopilaba información sobre el dominio militar, organizaciones sociales, comercio e instituciones políticas, se observó una pérdida en el interés de registrar ciertas características de una población o un terreno durante los años subsecuentes a la caída de Roma.

Sin embargo, la necesidad de contabilizar las defunciones por parte del rey de Inglaterra Enrique VII dio un nuevo impulso a la generación de censos, dado que estos registros eran esenciales para evitar la propagación de la peste bubónica durante los inicios del siglo XVI.

Con la elaboración de estos censos en Inglaterra, en específico con la publicación de los decesos de forma semanal que años después generarían un registro de los nacimientos y fallecimientos por género contenidos en los *Bills of Mortality (Cuentas de Mortalidad)*, propiciaron el

surgimiento del análisis estadístico. En estos censos se consideraba la información proveniente como una herramienta con la capacidad de poder producir predicciones. Específicamente fue John Graunt (1620-1674) quien en 1662 generó pronósticos sobre los nacimientos esperados con una especificación del sexo de los recién nacidos y las muertes que serían generadas por múltiples enfermedades de la época e integró tasas de mortalidad. Además, elaboró un pronóstico de los posibles decesos en la población infante menor de 6 años causados por el creciente contagio de la peste bubónica.

Estos pronósticos se encuentran en *Natural and Political Observations Made upon the Bills of Mortality (Observaciones naturales y políticas hechas a partir de las Cuentas de Mortalidad)*, obra en donde se recopilaron 30 años de encuestas de mortalidad. El propósito era manufacturar algún tipo de mecanismo que identificara la aparición de la peste bubónica y cómo podría propagarse.

Poco tiempo después de los aportes de John Graunt, se encuentra que el análisis estadístico fue usado para romper paradigmas sobre diferentes cosas. Un ejemplo de esto fue el trabajo del alemán Gaspar Neumann (1648-1715), quien probó que la creencia de que los fallecimientos en edades con una terminación en 7 eran considerablemente mayores al resto de los años era totalmente errónea. Esto se logró al analizar múltiples registros de defunción ubicados en las parroquias de la ciudad, probando que la muerte de la población no depende de los dígitos de su edad.

Los procedimientos utilizados por Neumann fueron implementados por Edmund Halley (1656-1742) con la finalidad de realizar un estudio integró sobre las defunciones. Gracias a los resultados de dicho estudio se dio a conocer una de las herramientas más importantes para la rama de los seguros, que son las tablas de mortalidad.

El año en el que la estadística logró obtener su nombre fue en 1760, gracias a Godofredo Achenwall (1719-1772) quien tomó este concepto del italiano *Satista*. En los trabajos de Achenwall se da una visión general de varios países, al describir su forma de generar la agricultura, la elaboración de artefactos hechos a mano y la visión particular del comercio de cada región. No obstante, fue John Sinclair (1754-1835) quien en su libro *Statistical Account of Scotland (Cuenta Estadística de Scotland)* formalizó a la estadística como una rama de las matemáticas con el potencial de generar información sobre ciertas cuestiones y proponer posibles soluciones.

En el año 1805 el matemático francés Adrien-Marie Legendre (1752-1833) registró formalmente el modelo de regresión lineal y el método de mínimos cuadrados en su obra *Nouevelles Méthods Pour la Détermination des Orbites des Cómentes (Nuevos métodos para la determinación de las órbitas de los cometas)*. Estos métodos también son adjudicados de igual manera a Friedrich Gauss (1777 - 1855), quien años antes realizó una aplicación del método en la localización del asteroide Ceres. Además, desarrolló de forma más completa el método de mínimos cuadrados y en qué distribución recaían los errores.

Las aplicaciones de la estadística en esos años fueron aumentando conforme se extendía la difusión de ese concepto, por ejemplo, Adolphe Jacques Quetelet (1796-1874) fue precursor en utilizar la estadística como un mecanismo que estudia la estructura y el funcionamiento de las sociedades humanas. En 1835, Quelet incorporó en dicho análisis diversos elementos que conectaban la teoría de la probabilidad con la estadística, con la intención de medir la variabilidad de ciertos fenómenos que se presentaban en esas épocas en la población. Este análisis está contenido en el libro que lleva por nombre *L'homme et le Développement de ses*

Facultés, ou Essai de Physique Sociale (El hombre y el desarrollo de sus facultades, la prueba física o social).

Durante el siglo XX, la estadística tuvo una restructuración abrumadora por parte de Francis Galton (1822-1911) y Karl Pearson (1857-1936), los cuales, impulsaron la alteración del concepto al cambiar la forma en la que se realizaban los desarrollos estadísticos, proponiendo una rigurosa formalización matemática basada en la teoría de la probabilidad. Además propusieron que la estadística no se restringiera solo en calcular y registrar los datos de cierta población, sino también, en analizar los detalles de la muestra y generar información con una calidad adecuada para tomar mejores decisiones.

Por su parte, Francis Galton desarrolló ciertos conceptos como la regresión a la media y la correlación entre variables. Además fue precursor en la ampliación de la distribución Normal al descubrir la Normal Bi-variada y la construcción de una máquina con la capacidad de comprobar la aproximación de la Distribución Binomial a la Normal, la cual, constaba con una serie de clavos instalados en un tablero vertical y la continua interacción de soltar pelotas desde la parte superior observando en que área caían.

En el caso de Karl Pearson, se encuentra que fue quien estudio la interacción que tenía la estadística con otras ramas de las matemáticas, como el álgebra lineal, la probabilidad o la geometría. Con ayuda de este estudio implemento un soporte más sólido a los conocimientos básicos con la intención de obtener información adicional de los datos en los que se está aplicando el análisis. Además de este avance considerable tenemos que Pearson:

"Introdujo el *método de los momentos* para la obtención de estimadores, el *sistema de curvas de frecuencias* para disponer de distribuciones que pudieran aplicarse a los distintos fenómenos aleatorios, desarrolló la correlación lineal para aplicarla a la teoría de la herencia y de la evolución. Introdujo el *método de la* X^2 para dar una medida del ajuste entre datos y distribuciones, para contrastar la homogeneidad entre varias muestras, y la independencia entre variables. Fundó los *Anales de Eugenesia* y en 1900, junto con Galton y Weldon, fundó la revista *Biometrika* de la que fue editor hasta su muerte. En una descripción autobiográfica decía "*una explicación para mi vida, se debe a una combinación de dos características que he heredado: capacidad para trabajar mucho y capacidad para relacionar las observaciones de los demás*". (Gómez Villegas, 2009)

Durante esos años, se encuentran los últimos aportes que dieron base a la estadística actual dando un mayor contexto a los progresos de Galton y Pearson; William Sealy Gosset (1876-1937), Ronald Fisher (1890-1962) y Jerzy Neyman (1894-1981) terminaron de optimizar las herramientas necesarias para poder realizar un análisis sobre algún fenómeno y presentar información adicional que permita omitir la incertidumbre.

Los avances de Ronald Fisher contemplan el descubrimiento del concepto varianza y suficiencia, el método de máxima verosimilitud y la información de Fisher. Por parte de Gosset estudió la identificación de ciertos comportamientos a los que llamó distribución t-student. Finalmente, Jerzy Neymann desarrolló los intervalos de confianza, la hipótesis nula y la prueba de hipótesis, con la colaboración de Fisher y Pearson. Adicionalmente ideó la prueba FDA (Food and Drug Administration, Agencia de Alimentos y Drogas) con la que se realizan verificaciones a los medicamentos antes de ser colocados en distribución.

Capítulo 2

Administración de Riesgos

La administración de riesgos tiene como objetivo principal evitar o reducir el impacto de cualquier riesgo que pueda exponer a la empresa o institución a una pérdida significativa. La administración de riesgos es indispensable para poder tomar acciones plenamente establecidas y adaptables a los diferentes escenarios posibles para poder identificar, medir y controlar cada una de las posibles pérdidas. Estas acciones no deben ser temporales o eventuales, sino un proceso flexible con bases establecidas que puedan adaptarse y desenvolverse en el transcurso del tiempo de la forma más precisa posible, considerando agentes tanto internos, como externos de la institución o empresa.

Para poder generar una administración de riesgos exitosa es necesario detectar cada uno de los escenarios posibles, determinando así qué factores de riesgo está enfrentando la institución o empresa y generar estrategias que puedan facilitar las decisiones adecuadas sobre el nivel del riesgo:

"Para administrar los riesgos en forma exitosa, se debe contar con las herramientas necesarias que le permitan a la empresa desarrollar un lenguaje común que facilite la comunicación interna y externa, prevenir los riesgos y disminuir la probabilidad de su ocurrencia, detectarlos en caso de que se materialicen, contar con sistemas ágiles y flexibles para responder ante ellos, y con el personal y los recursos de la organización apropiados; generando informes y medir su ocurrencia." (Mejía Quijano, 2006)

La identificación de los riesgos que puede enfrentar una institución o empresa es esencial para poder planificar procedimientos con la capacidad de mitigar los efectos negativos de una situación adversa o incluso poder tomarlos como oportunidades. Para esto, es necesario explicar los tipos de riesgos principales y bajo qué tipo de circunstancias es válido pensar en su materialización. Por ejemplo, en el trascurso de las operaciones correspondientes a cada empresa o institución es posible identificar los efectos simultáneos de dos tipos de riesgos, como podrían ser el riesgo de mercado acompañado de un riesgo de crédito, conceptos que explicaremos más adelante.

Las categorías en las que se generalmente se puede catalogar a los riesgos son las siguientes (De Lara Haro Alfonso):

- Riesgo de Mercado: es el riesgo que se produce por la diferencia de los precios al invertir en una acción en la bolsa o por algún factor cambiario, como puede ser una tasa de interés o tipos de cambio. De manera más formal podríamos decir que es el riesgo que se presenta con un valor presente neto adverso ante la modificación de las variables macroeconómicas a las cuales están expuestos los precios de los instrumentos en un portafolio. Los factores que se puede encontrar en este tipo de riesgo son: el precio, la promoción, la publicidad, las ventas, la competencia o la saturación del mercado.

- Riesgo de Crédito: se define como el riesgo que se puede tener al ocurrir el incumplimiento de la contraparte en una transacción en donde se tiene una promesa de pago. Este riesgo se presenta comúnmente en entidades bancarias en donde se presentan préstamos sujetos ante la posibilidad de que el cliente no pague. Adicionalmente se encuentran las situaciones de insolvencia de emisores de valores, como pueden ser los bonos de deuda, fondos de inversión o en los CDT (Certificado de Depósito a Término).

- Riesgo Operativo: es el riesgo que refleja las pérdidas generadas por deficiencias o irregularidades en los sistemas, procedimientos, recursos humanos, modelos y cualquier factor que pueda causar resultados adversos a la empresa o institución. Este riesgo engloba diversos elementos que pueden pertenecer a diferentes áreas, como puede ser la separación de algún elemento clave en los procesos de la empresa o institución (recursos humanos) o la compra de algún inmueble que no cumpla con las características por las cuales fue adquirido (directivos).

- Riesgo de Reputación: es el riesgo que mide la percepción a las pérdidas generadas por la desintegración de la estima o prestigio adquirido por parte de los consumidores con algún fragmento de la empresa o institución que puede provocar un cambio en las preferencias de los clientes, es decir, la inconformidad de los consumidores con respecto a los errores de la empresa, como pueden ser la mala capacitación de algunos elementos clave, el abuso de los servicios prestados, errores en los sistemas o problemas en algún servicio.

- Riesgo de liquidez: es el riesgo en donde se dificulta la capacidad de un bien o activo de ser transformado en dinero efectivo, y que, puede obstaculizar las operaciones de una empresa o institución. También se consideran los casos en donde se tiene que vender un activo para poder financiar las actividades de la institución o empresa a un precio desfavorable que representa un pérdida en los objetivos futuros, y el caso de tener que pagar las promesas de pago que tenía la empresa o institución con una tasa de descuento desfavorables.

- Riesgo Legal: es el riesgo asociado a las pérdidas generadas al presentarse incumplimientos en las promesas de pago adquiridas por la empresa o institución. Este riesgo considera el escenario en donde no sea posible exigir de forma jurídica el cumplimiento de los compromisos pactados con las contrapartes por algún error de interpretación por parte de la empresa o la omisión de alguna cláusula que evite la existencia de este tipo de huecos legales.

Teniendo en cuenta la clasificación de los diversos riesgos, el siguiente paso necesario para consolidar un proceso exitoso de administración de riesgos que permita la capacidad de predecir las posibles consecuencias que se puedan presentar es la medición de los riesgos. La importancia de la medición de los riesgos radica en tener la capacidad de establecer un modelo que trate de predecir de la manera más certera posible al fenómeno que se está estudiando y observar cómo esté puede evolucionar, considerando que cumpla con los objetivos iniciales, los cuales, deben dirigirse en lo posible a cubrir las necesidades e intereses de los usuarios.

Los factores que pueden perturbar el buen funcionamiento del modelo al trascurso de su operación al grado de errar la precisión de las predicciones necesarias para tomar decisiones, se les conoce como, factores de riesgo. En este contexto podemos decir que existen dos tipos independientes de tipos de factores de riesgos, los cuales son:

- Cuantificables: Son los factores sobre los cuales es posible conformar bases estadísticas, con el objetivo de estudiar su comportamiento durante un cierto tiempo y lograr generar proyecciones. Entre estos factores se encuentra una división con respecto a si es posible la intervención institucional, es decir:
  - Riesgos Discrecionales: Resultado de la toma de una posición en específico con respecto a cierta incertidumbre, como puede ser el riesgo de mercado. La característica principal de este tipo de riesgo es que se comporta de forma independiente del proceso de la institución o empresa.
  - Riesgos No Discrecionales: Este factor es producido por las acciones realizadas por la institución o empresa, es decir, que dependen de las acciones de la empresa.
- No cuantificable: Factores en los cuales se presenta una aleatoriedad amplia que provoca una dificultad alta o imposible para poder ser registrados, por lo cual, no se pueden generar bases estadísticas que respalden alguna proyección.

Una forma eficiente de motivar la prevención de los riesgos y expandir las posibles oportunidades que puedan presentarse es difundir los cuestionamientos clave que cualquier institución o empresa debe plantearse en caso de tener el deseo de constituir una administración de riesgos sólida en cada una de sus operaciones. Algunos de los cuestionamientos son:

- La veracidad de los objetivos.
- Escenarios desfavorables en las operaciones.
- La posibilidad de la aparición de un factor de riesgo y el impacto correspondiente.
- El plan a seguir en caso de que se presenten cada uno de los escenarios desfavorables.
- La distribución de las responsabilidades en caso de ocurrir algún factor de riesgo.

2.1. KPI

Una de las posibles herramientas que se puede emplear para medir los factores de riesgo son los KPI's (*Key Process Indicator*, Indicador Clave de Rendimiento), los cuales son métricas generadas a partir de una medición frecuente y objetiva de las operaciones. Estos indicadores tienen el objetivo de identificar el estado en el que se encuentra un proceso en específico, que sea fundamental para la institución o empresa. Los KPI's tienen la capacidad de generar proyecciones con respecto al comportamiento que se presente en los procesos deseados y encontrar la distribución que más se ajuste a los movimientos del mismo. El propósito de estas proyecciones

es realizar acciones correctivas en el caso de que se hayan presentado factores de riesgo o para realizar planes preventivos en caso de una inminente operación desfavorable.

El motivo principal de los KPI's es encontrar oportunidades de mejora en el interior e exterior de una empresa o institución que con regularidad tienen cierta dificultad para ser definidas y cuantificadas, como puede ser la eficiencia de los empleados en una cierta área. La elección de los KPI's es esencial para poder dar un correcto monitoreo a las actividades que puedan provocar factores de riesgo en diferentes temporalidades, como puede ser, el movimiento de una acción con monitoreo diario o la eficiencia de los empleados con respecto al años anterior.

En el caso de instituciones o empresas con un control extenso de factores de riesgo, se encuentra la constante creación de KPI's en cada rubro de sus operaciones con un seguimiento exhaustivo en la medición de éstos, con el objetivo de medir si las decisiones tomadas operan de forma correcta o si reportan una disminución significativa en los rendimientos.

2.2. *Balanced Scorecard*

El *Balanced Scorecard (Cuadro Integral de Mando)* es un instrumento que da la facilidad de dar un correcto seguimiento a la forma en la que se presentan los principales KPI's de una institución o empresa. Esta herramienta fue introducida en 1992 por Robert Kaplan (1952-) y David Norton (1941-) en la revista *Harvard Business Review*, en donde plasman la motivación de expandir los monitores rígidos en áreas en donde sólo se contemplaban los aspectos financieros como ingresos, utilidades, gastos de producción, entre otros. El propósito de esto es marcar una diferencia con respecto a las otras compañías en el mismo ramo de negocios

La idea general del balanced scorecard consiste en presentar un plan que consiga cumplir con los siguientes campos:

- Capacidad de formular una estrategia precisa y eficiente.
- Medir los alcances de la comunicación en todos los niveles.
- Organizar la designación de responsabilidades para cada área.
- Ajustar los objetivos de la estrategia a la situación financiera.
- Identificar las áreas de oportunidad en cada sector.
- Tener la oportunidad de dar seguimiento a la estrategia planteada.
- Reportar los logros alcanzados de forma frecuente.

La alternativa propuesta por estos autores es colapsar los indicadores en 4 eslabones, en los cuales se especifican las características que deben de tener los KPI?s para solventar los cuestionamientos que puedan generar una ventaja con respecto a la competencia. Las categorías que consideradas son:

1. Perspectiva Financiera.
2. Perspectiva del Cliente.
3. Perspectiva del Proceso/Operación.
4. Perspectiva Humana.

2.2.1. Perspectiva Financiera

En esta categoría se contemplan todos los indicadores que ayudan a consolidar el estatus financiero de la empresa o institución, con el objetivo de que los accionistas observen la evolución de las operaciones y cómo estos avances dan seguimiento a los planteamientos antes generados. Algunos de los indicadores que son usados con más frecuencia miden los siguientes rubros:

- Utilidad Neta.
- Crecimiento de los ingresos.
- Rentabilidad.
- Préstamos.
- Activos.
- Pasivos.
- Gasto de publicidad.
- Efectividad de la inversión.

2.2.2. Perspectiva del Cliente

En esta categoría se contemplan los indicadores que describen la satisfacción del cliente con respecto a los servicios de la empresa o institución, como puede ser, la efectividad de la publicidad, la calidad del servicio, la opinión con respecto al producto, entre otros. El buen funcionamiento de la parte financiera depende fuertemente de que el número de clientes se mantenga estable o aumente. La forma en la que se obtiene la información necesaria para integrar este rubro depende principalmente de las encuestas, en las que se registran las opiniones de los clientes o compradores potenciales. Los rubros que motivan la creación de los indicadores son:

- Calidad del servicio.
- Retención de clientes.
- Nivel de satisfacción con el producto.

- Tiempo de fidelidad.
- Características de la población objetivo.
- Efectividad de la publicidad.
- Frecuencia de compra.
- Características del consumidor.

2.2.3. Perspectiva de Proceso/Operación

Es el análisis de la información proveniente de los procesos internos y externos generados con el propósito de cumplir las necesidades de los consumidores. Esta perspectiva cuenta con subramas especializadas en los procesos de las empresas o instituciones, con el objetivo de catalogar de forma eficiente las acciones realizadas y elevar los procesos a los mejores niveles posibles. Estos procesos se dividen en:

- Operaciones

  Nos indica los mecanismos que integran a la empresa o institución con respecto a sus análisis de calidad e ingeniería. Algunos de los indicadores que constituyen este proceso son:

  - Nivel seis sigma (calidad).
  - Tiempo de los procesos mecánicos.
  - Uso de la capacidad de los almacenes.
  - Tiempo de entrega del servicio o producto.
  - Nivel de solución de defectos.
  - Eficiencia de los equipos.
  - Inactividad de la maquinaria o procesos.

- Medio ambiente y comunidad

  Responde a los procesos que son efectuados de forma indirecta a las operaciones de la empresa o institución al considerar los aspectos ecológicos, limpieza, seguridad y responsabilidad civil, que en los últimos años ha generado una gran atención. Algunos de los indicadores que integran este proceso son:

  - Consumo de energía.
  - Nivel de ahorro debido a iniciativas ecológicas.
  - Reducción y manejo de los residuos.
  - Reciclaje de los productos.
  - Condiciones laborales.
  - Cadena de suministros.

2.2.4. Perspectiva Humana

Proporciona una fuente de información correspondiente al compromiso, capacitación y motivación con la que cuentan los empleados, considerando que éstos son los principales reflectores de los ideales y objetivos de la empresa o institución. Los indicadores más utilizados en esta perspectiva son:

- Ingresos de los empleados.
- Beneficios.
- Satisfacción de los empleados.
- Rotación de los empleados.
- Tiempo de permanencia.
- Competitividad salarial.
- Eficiencia de la capacitación.

Capítulo 3

Riesgo de Crédito

La primera mención del riesgo de crédito proviene del año 1728 a. C. con la creación del Código de Hammurabi, el cual, es una serie de reglas establecidas por el rey de Babilonia con el objetivo de dar a conocer a sus súbditos las características de los delitos y sus correspondientes castigos. Una de sus leyes establecidas es la ley del Talión ("ojo por ojo, diente por diente"). En una fracción de estas leyes se especifica las normas con las que se debía establecer el crédito y de qué forma se debía regular su aplicación. Este inicio fue señalado por John Caouette, Edward Altman y Paul Narayanan en el año 1998 en *Managing Credit Risk (Gestión del Riesgo de Crédito)*.

El desarrollo del riesgo de crédito dependía fuertemente de los bancos, al ser las instituciones en las que se realizaban más préstamos a la población con ciertas características. Es por ello que el manejo de los créditos desde la Edad Media hasta principios del siglo XX fue administrado por los llamados banqueros, los cuales, tenían la habilidad de identificar su entorno para generar ganancias. En otras palabras, tenían la responsabilidad de reconocer los posibles clientes al observar las características que podrían tener y clasificarlos, en posibles deudores o clientes potenciales. Con estas aptitudes, los banqueros tenían la capacidad de calcular cuánto dinero debían prestar, el grado de interés, las garantías que debían exigir y la forma en la que se debía cobrar.

La mayoría los grandes banqueros que generaron fortunas en sus respectivas épocas lograron ubicarse en las altas esferas del poder. El motivo de esta postura se debe a que era el sector de la población que proporcionaba valiosa información sobre todos los proyectos importantes que estaban por realizarse y los posicionaba en una fuente de negocios que en la mayoría de las ocasiones generaban ganancias relevantes.

Durante el estudio del riesgo de crédito se tiene una falta de aportaciones por parte del sector académico, ya que, en cierta medida las características especiales de este riesgo tienen una mayor complejidad en la medición, con respecto a los demás riesgos financieros. Esta complejidad adicional provocó que las opiniones expertas, modelos tradicionales y técnicas modernas en esos años fueran insuficientes en las crisis de los años treinta, finales de los ochenta y principios de los noventa del siglo XX.

3.1. Comité de Basilea

Las economías más importantes del mundo se encontraban inmersas ante la compleja identificación de modelos adecuados para poder medir los riesgos en sus operaciones, por ello, el Banco de Pagos Internacionales generó varias medidas con el objetivo de resarcir esos problemas, ya que:

"Las funciones desarrolladas por el Banco Internacional de Pagos, anticipándose al rol que después de la conferencia internacional de Bretton Woods jugarían instituciones como el Fondo Monetario Internacional y el Banco Mundial, establecieron un papel fundamental en la creación de varios acuerdos de pagos entre diversos países de Europa, la coordinación de la intervención en el mercado del oro y el manejo del dólar. No obstante, en 1973 la caída del sistema de tasas fijas en materia de intereses, la progresiva internacionalización de los mercados financieros y la notable insolvencia de las entidades bancarias Bankhaus Herstaff y Franklin National Bank, constituyeron circunstancias que abocaron la necesidad de una nueva coordinación de los bancos centrales para intercambiar información e intervenir en los mercados." (Ustáriz González, 2003)

El *Comité de Supervisión Bancaria de Basilea (Basel Committee on Banking Supervision)* o mejor conocido como Comité de Basilea fue constituido en 1974 por el Banco de Pagos Internacionales con el propósito de ser una plataforma en la que se pudiera discutir las dificultades que se presentan al realizar una supervisión. Además, el comité tenía la responsabilidad de coordinar el cumplimiento de las normas establecidas con las autoridades de cada país, a causa de que cada nación tiene sus respectivas leyes que se adaptan a sus condiciones y conductas particulares. Este Comité no incorpora normas que deben de ser cumplidas de forma obligatoria con sanciones en caso de presentarse irregularidades, en su lugar es una organización mundial que proporciona recomendaciones a las autoridades correspondientes para que éstas establezcan sus reglamentos de forma adecuada.

El primer acuerdo relevante presentado por el Comité de Basilea fue el *Acuerdo de Capitales de Basilea (Convergencia Internacional de Medidas y Estándares de Capital)*, tras una serie de eventos desafortunados al presentarse condiciones desfavorables para los bancos internacionales en respuesta de la crisis latinoamericana y prácticas desfavorables por parte del mismo sector bancario. El acuerdo fue presentado en 1988 con una aplicación efectiva un año más tarde en los bancos internacionales con mayor desarrollo en cuestiones financieras.

Las ideas presentadas en este acuerdo giran en torno al capital de las instituciones bancarias, al considerarlo como el eje principal que tenía la capacidad de solventar las responsabilidades adquiridas en caso de presentarse eventos desfavorables. La metodología dependía de identificar las fronteras que cada institución bancaria tenía en relación con su capital, ya que, estos límites marcaban el nivel máximo de riesgos que podían ser respaldados dependiendo de sus características, solventando así un correcto funcionamiento del sector bancario.

A pesar de los notables progresos que generó Basilea I con la implementación de estas recomendaciones se encontraron problemas significativos en las valoraciones de los riesgos y el cálculo de la suficiencia, dado a diversas omisiones, como menciona Sierra Núñez:

"En términos concisos, Basilea I define los requerimientos mínimos de capital de un banco en función del riesgo de sus activos y de los riesgos de mercado que afectan a la institución. Sin embargo, la principal limitación del acuerdo de Basilea I es que es insensible a las variaciones

de riesgo y que ignora una dimensión esencial: la de la calidad crediticia y, por lo tanto, la diversa probabilidad de incumplimiento de los distintos prestatarios. Es decir, consideraba que todos los créditos tenían la misma probabilidad de incumplir." (Sierra Núñez, 2011)

Al encontrar que las recomendaciones publicadas en 1988 no generaron el efecto deseado al no acoplarse de forma correcta al transcurso de los años, el Comité de Basilea publicó en 1999 el documento que lleva el nombre de *"A New Capital Adequacy Framework (Un Nuevo Acuerdo de Suficiencia de Capital)"*. Este documento, después de diversas mejoras en 2001 y 2003, dio las herramientas suficientes para poder presentar de forma definitiva un nuevo acuerdo que cumpliera con las expectativas de las instituciones financieras. En 2004 el Comité presentó el convenio que lleva por nombre " *"International Convergence of Capital Measurement and Capital Standards: a Revised Framework (Convergencia Internacional de Medición de Capital y Estándares de Capital: un Marco Revisado)"* o conocido de mejor forma como Basilea II.

El esquema de Basilea II contempla una mayor sensibilidad en la identificación de los riesgos y una base en la que las instituciones financieras puedan apoyarse para mejorar sus controles de riesgos y como solventar las operaciones de cada entidad. Este acuerdo está distribuido en dos bloques principales, la implementación de las características que debe de cumplir una institución para que se considere necesario el cálculo del coeficiente de solvencia y los tres pilares principales en donde se dividen las nuevas recomendaciones coordinadas con los aportes anteriores de Basilea I. Los pilares antes mencionados son los siguientes:

- Requerimientos Mínimos de Capital

  Como su nombre lo indica, el pilar I tiene el propósito de impartir una serie de recomendaciones referentes a la cantidad de recursos necesarios que una institución bancaria debe tener de forma obligatoria con la intención de soportar los riesgos asumidos. Cada riesgo en este acuerdo tiene especificaciones referentes a los modelos estándar que tienen que seguir todas las instituciones financieras y las características que son necesarias en los modelos internos que sean creados por cada entidad con respecto a sus condiciones especiales. Los riesgos que son considerados en este pilar son:

  - Crédito

    El planteamiento para el manejo adecuado de este riesgo se toman en cuenta los elementos establecidos en Basilea I, con la corrección respectiva de las fallas detectadas en el periodo de operación de este acuerdo. Los componentes que son desarrollados para el control del riesgo de crédito son los siguientes (Sierra Núñez):

    - Probability of Default (Probabilidad de Incumplimiento).
    - Loss Given Default (Perdida ante el Incumplimiento).
    - Exposure At Default (Exposición ante el Incumplimiento).

    En términos generales estas tres estrategias dependen de metodologías efectuadas de forma interna y externa, según Sierra Núñez:

    "En lo que respecta al riesgo de crédito, el acuerdo propone tres alternativas para su determinación. El primero de ellos, en su mecánica, es similar a lo establecido en Basilea I (ponderación preestablecida según riesgo para los distintos tipos de activos), pero presenta mejoras que lo hacen más sensible al riesgo e incorpora el uso de clasificaciones externas efectuadas por agencias especializadas. Los otros dos métodos (no consideradas en Basilea I) se basan en mediciones internas realizadas por los propios bancos. " (Sierra Núñez, 2011)

Como dato adicional en el estudio del riesgo de crédito se tiene que su medición en países emergentes representa un soporte fundamental para la estabilidad económica de estos países. En el caso particular de México sabemos que la intervención en la regulación de operaciones de crédito es indispensable para evitar una crisis generalizada en el sector financiero, ya que:

"En México, los riesgos crediticios constituyen, en promedio, más del 80 % de los activos bancarios sujetos a riesgo. Existe un consenso sobre el importante papel que jugaron los créditos en los recientes problemas del sistema bancario mexicano." (Elizondo, 1999)

- Mercado

  En esta rama no se encuentra una actualización por parte del Comité de Basilea, por lo cual, se le considera una reiteración de los aportes propuestos con anterioridad en el acuerdo de Basilea I.

  El contexto en el que se desarrolla esta ramificación responde a la necesidad de tener una reserva suficiente de recursos para dar respaldo a las variaciones del riesgo de mercado. Los eventos que son contemplados son variaciones en (Sierra Núñez):

  - Tasa de interés.
  - Tipos de cambio.
  - Precios y rendimientos de acciones.
  - Materia prima.

- Operativo

  El riesgo operacional fue una incorporación favorable en el acuerdo de Basilea II, ya que, aun considerando los riesgos de mercado y crédito, se tenía presencia de pérdidas relevantes con respecto al personal, los sistemas operativos interno y externo, aspectos tecnológicos y elementos externos como podría ser problemas en los suministros de algún recurso básico, como el agua o la electricidad.

  La presencia de pérdidas que no respondían a los riesgos antes mencionados impulsó una serie de recomendaciones al presenciar una creciente complejidad de las operaciones de cada entidad, tal y como menciona Sierra Núñez:

  "La desregulación y globalización de los servicios financieros, junto con la creciente sofisticación de la tecnología financiera, están haciendo cada vez más diversas y complejas las actividades de los bancos y, por lo tanto, sus perfiles de riesgo. El desarrollo de las prácticas bancarias sugiere que, aparte de los riesgos de crédito, de tipo de interés y de mercado, pueden ser considerados, a efectos de supervisión, otros riesgos, como es el caso del operacional. Por esta razón, el objetivo del Comité es que las instituciones bancarias mantengan el capital necesario para solventar las eventuales pérdidas ocasionadas por el riesgo operativo, más allá del capital mínimo requerido por concepto de la calidad de los activos o por el riesgo de mercado." (Sierra Núñez, 2011)

- Supervisión de la Suficiencia de Capital

  El pilar II está constituido como un manual con el objetivo de dar un entorno mayor a los riesgos que por sus características no puedan encajar con el pilar I, ya que proporciona una serie de instrucciones como consecuencia de la regulación del capital. Algunas de las especificaciones que cubre este pilar son:

  - Procesos de autorización.
  - Medición de avances de las metodologías establecidas.

- Calificación de la calidad de los procesos de identificación de riesgos y los sistemas de control.
- Métodos de validación de los coeficientes de capital.

- Disciplina de Mercado

  La información adecuada es vital para cualquier institución financiera que desee tener las herramientas suficientes para tomar las decisiones correctas de acuerdo al entorno en que se encuentre. Por ello, el tercer pilar establece los requerimientos con respecto a la divulgación que cada entidad bancaria debe de tener con el propósito de implementar una plataforma que sea alimentada por información útil, oportuna y suficiente. El objetivo principal de esta plataforma es otorgar la opción al usuario de generar comparaciones entre las entidades y tener los elementos adecuados para decidir en cuál de ellas puede optimizar sus recursos.

  La generación de la plataforma obliga a las instituciones a revisar de forma frecuente la situación en la que se encuentra la inversión de su capital y la evolución que han tenido los riesgos relacionados con las operaciones de la institución. Con estos elementos es posible identificar situaciones que puedan generar importantes pérdidas, al tener la oportunidad de estudiar las características en las que se encuentra el sector en el que pertenecen, e incluso, pueda utilizarse para reconocer áreas de oportunidad. Los elementos que son presentados por las entidades financieras son:

  - Control de riesgos.
  - Requerimientos de capital.
  - Análisis y descripción de aspectos elementales (capital, utilidades, ingresos, etc).
  - Metodología de los modelos externos e internos.
  - Información operativa.
  - Reportes de resultados.

3.2. Sistema Financiero

El sistema financiero es el encargado de intervenir entre los que tienen los recursos y no los necesitan en un periodo corto y están dispuestos a prestar dichos recursos por una bonificación (tasa de interés) y los que necesitan cierta cantidad de recursos disponibles para aumentar su capital con el propósito de emprender un proyecto lucrativo o cubrir alguna obligación y están dispuestos a cubrir dicha bonificación.

El propósito del sistema es emplear instrumentos financieros que desarrollen un ahorro óptimo por medio de mercados e intermediarios financieros. Los mercados financieros están constituidos con el objetivo de movilizar recursos durante un cierto tiempo mediante plataformas en donde se transfieren activos, las cuales se dividen generalmente en deuda y acciones. Por su parte, los intermediarios financieros tienen el propósito de facilitar las transacciones mediante instituciones de crédito, seguros y fianzas, organismos auxiliares, entre otros.

Los servicios financieros son una parte fundamental en el sistema, como lo explica el Banco de México:

"Los servicios financieros son aquellos otorgados por las distintas organizaciones que conforman el sistema financiero y que facilitan el movimiento del dinero. Entre ellos destacan principalmente los intermediarios financieros. De esta manera, el sistema financiero cumple con sus funciones de intermediar recursos y posibilitar la existencia del sistema de pagos en la economía a través de la prestación de diversos servicios financieros." (Banco de México, 2016)

Las autoridades encargadas de proteger y preservar la estabilidad del Sistema Financiero son (Banco de México):

- La Secretaria de Hacienda y Crédito Público.
- El Banco de México.
- La Comisión Nacional de Seguros y Fianzas.
- La Comisión Nacional Bancaria y de Valores.
- El Instituto para la Protección al Ahorro Bancario.
- La Comisión Nacional del Sistema de Ahorro para el Retiro.
- La Comisión Nacional para la Protección y Defensa de los Usuarios de Servicios Financieros.

3.3. Análisis de crédito tradicional

En la forma contable tradicional, las instituciones bancarias tomaban en cuenta normas que cumplieran con estándares simples para poder otorgar de forma eficiente y rápida los créditos a los clientes solicitantes. En el siguiente diagrama se explica de cómo habitualmente se asignaban los candidatos a recibir el crédito (De Lara Haro, 2014):

El procedimiento que se observa en el diagrama responde a las acciones que analiza una institución bancaria para asignar un crédito, basándose en la 5 C´s siendo éstas la descripción de una serie de características a considerar, las cuales son:

- Carácter

  Es la verificación que se le realiza al solicitante con base en su información personal para llegar a la conclusión de sí es una persona confiable con un "carácter" suficientemente bueno para cumplir con sus obligaciones o una inminente pérdida crediticia. En la verificación que la institución bancaria elabora se toman en cuenta:

  - Datos personales del solicitante (nombre, fecha de nacimiento, RFC, etc).
  - Empleos registrados (compañía, puesto, salario, descripción de la compañía).
  - Créditos bancarios y no bancarios actuales del solicitante (institución bancaria, saldo actual, modalidad de la tarjeta, estado y comportamiento durante la operación).

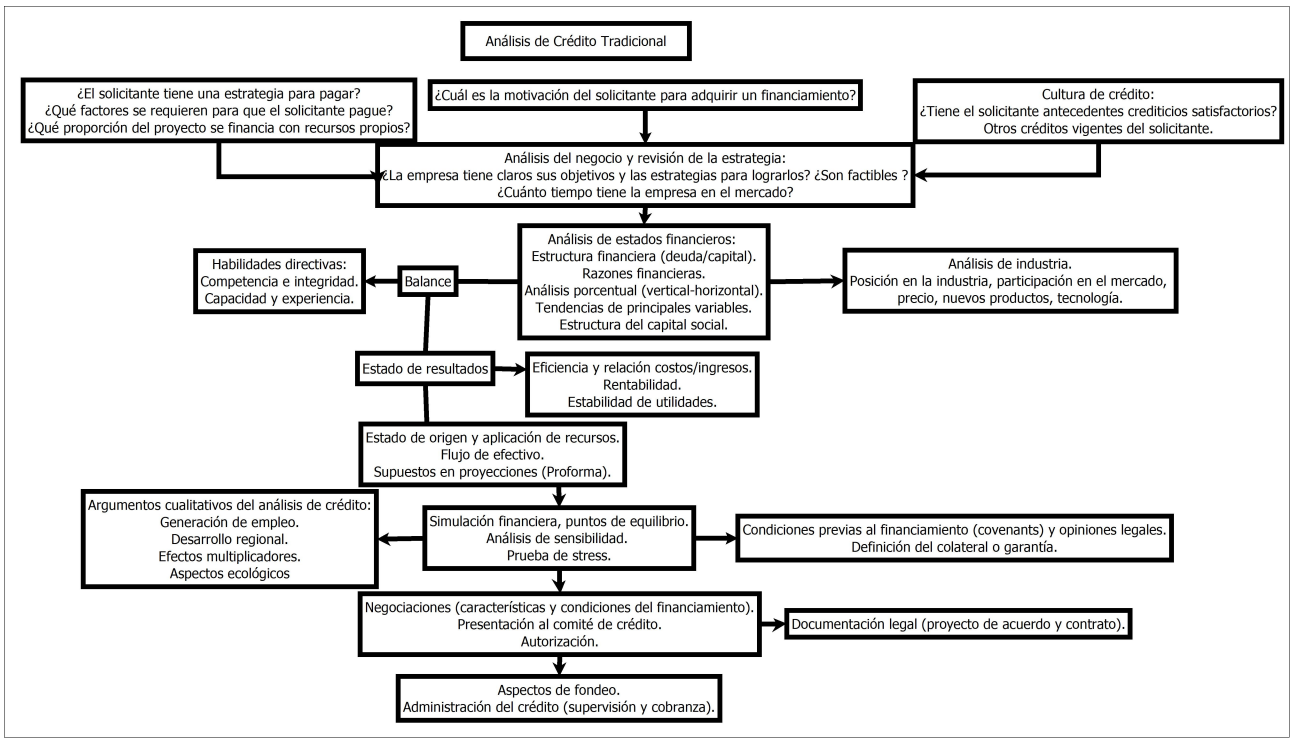


Figura 3.1: Diagrama de Crédito

La modalidad más frecuente que utilizan las instituciones para recolectar la información antes mencionada es usando el Buró de Crédito, el cual, es una base de datos recolectada por una empresa privada desde hace 20 años que registra a los deudores sin importar la institución bancaria a la que pertenezcan.

- Capacidad

  En este proceso se revisa si el solicitante tiene la amplitud de recursos necesarios para poder cumplir con los compromisos futuros que pueda transferir a la institución bancaria. Las características que se pueden tomar en cuenta para poder consolidar un reporte que pronostique la futura situación del interesado son:

  - Sueldo.
  - Años de antigüedad en su presente empleo.
  - Condiciones del sueldo (bonos, comisiones, sueldo base, etc).
  - Prestaciones.
  - Recomendaciones de sus empleos anteriores.
  - Responsabilidades pendientes (préstamos).
  - Fuentes alternas de ingresos.

  La información correspondiente puede dar la ventaja de saber cómo el solicitante tendría los suficientes recursos para pagar sus responsabilidades en caso de adquirirlas y un fragmento del perfil futuro, es decir, una proyección de características que se podrían mantener fijas en una temporalidad cercana, como podría ser los dos años subsecuentes.

- Capital

  Representa la amplitud de recursos que tiene la institución bancaria para otorgar el crédito solicitado. El escenario que debe considerarse es el incumplimiento del solicitante que

genere insuficiencia en el interior de la institución, al tener una dependencia importante por los aportes de ese usuario. El análisis de casos desfavorables debe ser una prioridad por parte de la institución para identificar sus límites correspondientes a cada uno de sus productos en el mercado, además, de tratar de prever la conducta de sus clientes en caso de que éstos adquieran más obligaciones. Los elementos que pueden considerarse en este rubro son:

- Número de solicitantes por producto.
- Conducta promedio de los solicitantes.
- Capital utilizado en cada producto.
- Espacio de tolerancia.
- Deuda promedio de todos los usuarios.
- Aportes promedio de cada mes.
- Ganancias por cobro de intereses.
- Número de incumplimientos en ciertos periodos (bimestres, semestres, meses).

- Colateral

  En los casos en los que se asigna un crédito o cualquier producto al solicitante que signifique una obligación de pago con la institución bancaria emisora, es necesario contemplar la posibilidad de que el usuario incumpla en sus correspondientes pagos. Por lo que es conveniente solicitarle al cliente que entregue como respaldo de su compromiso alguna garantía que sea en cierta medida proporcional al beneficio solicitado. En esta clase de escenarios es esencial contar con las garantías "colaterales", las cuales, deben asegurar el respaldo económico con respecto al insumo solicitado con el propósito de que se tenga la capacidad de recuperar los recursos invertidos.

  Las garantías aseguran una menor exposición al riesgo en caso de que no se realice la operación por parte de la institución bancaria, lo cual, beneficia en el momento de tomar la decisión si tomar el riesgo o rechazarlo. Ciertas garantías que son consideradas en los diferentes niveles de obligaciones de pago son:

  - Avales personales.
  - Inmuebles.
  - Terrenos.
  - Activos del negocio (en caso de tenerlo).
  - Fideicomisos.
  - Cuentas por cobrar.

- Condiciones

  El ambiente en el que se desarrollan las operaciones de una institución bancaria es uno de los factores fundamentales que deben tomarse en cuenta para implementar cualquier tipo de producto a algún sector de la población. El motivo de su relevancia es la poca o nula capacidad de anticiparse ante cualquier eventualidad inesperada produciendo pérdidas inminentes que afectarían la solvencia de la institución. Algunas de las condiciones más comunes que deben de analizarse para plantear alguna estrategia de prevención son:

  - Economía del país.

- Contracciones del mercado.
- Desarrollo del sector.
- Tecnología.
- Modificaciones en la legislación.
- Plazo, tasas de interés y comisiones.

3.4. Componentes del Crédito

Una operación de crédito tradicional está compuesta por varios factores que es indispensable tomar en cuenta para que el título crediticio sea pactado. Los elementos requeridos son (Banco de México):

- Deudor y Acreedor

  La parte inicial en la ejecución de un crédito es tener a una persona que requiera los recursos y otra que tenga la disponibilidad de entregarlos. El nombre que reciben es el siguiente:

  - Deudor: el sujeto o negocio que recibe los recursos.
  - Acreedor: la persona o institución que entrega los recursos acordados.

- Monto de la Operación

  El monto representa la cantidad de recursos que el deudor recibirá al momento de contratarse el préstamo. En algunas ocasiones el monto se entrega por partes dependiendo de la cantidad acordada. Debe considerarse que la cantidad que se le será entregada al deudor no será la misma cantidad que este mismo tendrá que pagar al finalizarse la operación de crédito. En el área de tarjetas de crédito se presenta una variación en el concepto de monto, ya que se le da el nombre de línea de crédito otorgada a una compañía o persona (incluso al gobierno) y representa la cantidad máxima de recursos que la entidad está dispuesta a prestar al deudor.

- Tasa de Interés

  La tasa de interés es el indicador que representa el dinero que el deudor tendrá que pagar por recibir los recursos que solicitó. La tasa se aplica sobre el monto total del préstamo solicitado y deberá pagarse durante o al final de la operación de crédito, según las características especiales del crédito otorgado.

- Denominación de la Moneda

  La denominación más común que se utiliza en una operación de crédito corresponde a la moneda del país en donde se está aplicando el préstamo. Sin embargo, en algunas ocasiones se toma la decisión de pactar los créditos en una moneda diferente, usualmente en dólares americanos, por ejemplo, el caso de las empresas que se dediquen a la importación y exportación de recursos y cuyos cambios de divisa reducen las utilidades de la empresa.

- Comisiones

  Las comisiones son todos aquellos cargos extras que se le cobran al deudor aparte de los efectuados por la tasa de interés con la intención de cubrir los costos y gastos provocados por la realización del crédito, los cuales, son controlados por el Banco de México para prevenir algún abuso imponiendo que sean debidamente informados los deudores antes de efectuar el crédito.

- Garantías

  Las garantías son el instrumento que implementan las entidades bancarias para poder asegurar sus recursos en caso de que el deudor incumpla en sus obligaciones. La propiedad del deudor o de alguna persona que responda por él con el propósito de dar un aval de que se cumplirá de forma correcta con el préstamo solicitado. La relevancia de las garantías es que la entidad bancaria puede vender los artículos de los deudores en caso de que éstos lleguen a incumplir y se puedan cubrir las pérdidas causadas por la falta de pago.

  Las garantías más comunes son (Banco de México):

  - Prendarias

    Los artículos que son recibidos en este tipo de garantía son bienes muebles o artículos que pueden ser movilizados con un valor que puedan asemejarse al del préstamo solicitado. Algunos ejemplos son automóviles, lavadoras, refrigeradores, televisiones, joyas, entre otros.

  - Aval

    Las garantías en donde participa un aval o un tercero que respondería por la obligación adquirida por el deudor en caso de que éste incumpla. Los ejemplos más comunes de avales son familiares del deudor (padres, hermanos, tíos, etc), alguna entidad financiera o el mismo gobierno en casos de mayor relevancia.

  - Hipotecarias

    Son préstamos respaldados por bienes inmuebles o objetos que no pueden ser reubicados como pueden ser departamentos, casas, terrenos, entre otros.

  - Fideicomiso

    En las garantías que son respaldadas por un fideicomiso, el deudor entrega un inmueble o mueble, con la particularidad de que es otorgado mediante un fideicomiso. El uso del fideicomiso significa que el deudor entrega el artículo a otra entidad, en este caso el banco, para que éste lo administre en el escenario en donde el deudor caiga en incumplimiento y la entidad tenga la autoridad de poner la garantía a la venta para poder recuperar sus recursos. Con la singularidad de que en este tipo de contratos viene estipulado que no será necesario emprender un juicio al ser propiedad del acreedor en caso de incumplimiento.

3.5. Tipos de Crédito

Es importante conocer las clasificaciones que tienen los créditos en México, con el objetivo de conocer sus características y diferencias respectivas de cada tipo. Los tipos de crédito más usuales son (Banco de México):

- Tarjeta de Crédito

  La tarjeta de crédito es una modalidad en la que se expide un elemento plástico aprobado por la entidad bancaria con la correspondiente línea de crédito en la que se pueden hacer cargos de bienes y servicios. Este tipo de crédito tiene características particulares y por el hecho de ser el más utilizado, se especificarán algunos de sus elementos característicos:

  - Fecha de corte.
  - Pago mínimo.
  - Límite de crédito.
  - Fecha límite de pago.
  - Sobregiro.
  - Modalidades de pago especiales.

  El crecimiento de este tipo de crédito ha provocado que tenga diversos usos, como pueden ser las compras en tiendas de autoservicio, retiro de dinero en cajeros automáticos, compras por Internet, entre otros. No obstante existe una modalidad con el nombre de tarjeta de crédito básica, la cual sólo tiene la opción de ser usada en la adquisición de bienes o servicios con un límite de crédito de 200 días de salario mínimo, con la posibilidad de ser emitida por cualquier entidad bancaria que ofrezcan esta clase de crédito.

  En términos generales este tipo de crédito tiene un formato muy flexible en su uso y pago, sin embargo, en la mayoría de los casos es muy costosa con respecto a los otros tipos de crédito, por la falta de una garantía que respalde los préstamos realizados.

- ABCD

  El crédito ABCD tienen el propósito de ayudar a los usuarios de adquirir bienes duraderos con tiempos de vida útil amplios. Se refiere a artículos que sean bienes muebles o que pueden ser desplazados. Algunos de los artículos más comunes que cuentan como duraderos son:

  - Línea blanca (lavadora, refrigerador, etc).
  - Muebles.
  - Artículos electrónicos (computadoras, televisiones, estéreos, etc).
  - Joyería.
  - Textiles.

- Hipotecario

  El crédito hipotecario es una modalidad orientada a la adquisición de un inmueble, en donde, éste mismo queda de garantía en caso de incumplimiento. Los inmuebles más comunes son terrenos, casas y departamentos con préstamos cuyas coberturas son mayores al 70 % con temporalidad de pago de 10 hasta 30 años regularmente. Las modalidades más comunes de este tipo de crédito son hipotecas para:

  - Construcción.
  - Compra de terreno.
  - Cubrir hipotecas actuales.
  - Préstamos en efectivo.

- Remodelación.

- Automotriz

  Son operaciones de crédito dirigidas a la adquisición de automóviles financiada por la entidad bancaria, con la característica de ser un crédito prendario, es decir, un crédito en donde se deja como garantía una prenda o artículo de valor que podría ser el mismo automóvil en cuestión. Este tipo de préstamo permite que las entidades bancarias puedan realizar algún tipo de acuerdo con empresas de automóviles, con la ventaja para el usuario de tener una alternativa de crédito en caso de ser adquirido el producto.

- Pyme

  Los créditos Pyme van orientados a las pequeñas y medianas empresas que tienen una idea definida de un negocio y la necesidad de capital para poder adquirir instalaciones, maquinaria o algún elemento necesario que esté relacionado con su producción. Los elementos antes mencionados pueden quedar como garantía de pago. La aprobación de préstamos dependen fuertemente de las características de la Pyme, como podrían ser años de antigüedad, nivel de ganancias, ubicación de la empresa, finalidad de la empresa, entre otros.

- Nómina

  El crédito de nómina va orientado a trabajadores que utilizan una cuenta de nómina como depósito de su salario con alguna entidad bancaria. Los préstamos son aprobados con una fuerte certeza de que el deudor tenga la capacidad de pagar y la solvencia para realizar los cargos del préstamo, sin tener la necesidad de que el usuario tenga que realizarlos de forma física.

3.6. Probabilidad de Incumplimiento

Bajo los estatutos de Basilea, las instituciones bancarias deben de calcular la probabilidad de default o de incumplimiento de los deudores, y es necesario especificar cómo se calcula y qué modelos pueden ser utilizados. La probabilidad de incumplimiento es la medición que se realiza para saber qué tan posible es que un deudor incumpla sus obligaciones crediticias, tal y como lo define la CNBV :

"Es la medida de qué tan probable es que un acreditado deje de cumplir con sus obligaciones contractuales. Su mínimo valor es cero, lo cual indicaría que es imposible que incumpla con sus obligaciones, y su máximo valor es uno cuando es seguro que incumpla. Por tipo de crédito, normalmente se estima a partir de la tasa de incumplimiento observada en cada tipo de crédito, que es la proporción de deudores o créditos que dejan de pagar en un periodo de tiempo dado, respecto de los que estaban vigentes en el periodo anterior." (CNBV,2005)

Los modelos que con frecuencia son utilizados para calcular la probabilidad de incumplimiento son (De Lara Haro):

- Modelos Tradicionales

Son los modelos en donde los criterios subjetivos y la experiencia son la base indispensable para el cálculo de la probabilidad de incumplimiento. En la presente tesis se incluye el modelo de este rubro más usado, que es el de las 5 C's.

- Modelos Modernos

  Son los modelos que están basados en la observación de las estimaciones de pérdida esperada y no esperada de los deudores. Los modelos más utilizados en la actualidad son (De Lara Haro, 2014):

  - Redes Neuronales

    Este método intenta establecer un conjunto de redes neuronales artificiales con el propósito de emular el aprendizaje humano, incluyendo métodos estadísticos y de clasificaciones tanto supervisados como no supervisados. El proceso que realiza el modelo es observar las relaciones que se crean entre los datos de entrada y los de salida con el objetivo de crear una simulación de decisiones humanas en donde los datos estén incompletos. Los principales inconvenientes que tiene este modelo son la complejidad, el tiempo y el costo de la creación de las plataformas que puedan tener una interacción avanzada.

  - KMV

    El modelo KMV (Kealhofer, McQuown y Vasicek) es la aplicación de la teoría de decisión que es utilizada en la valuación de acciones, solo que en este caso va aplicada a si el deudor va a cumplir con sus obligaciones o no y la valuación del préstamo solicitado.

  - Econométricos

    Los modelos econométricos tienen la intención de medir la relación que existe entre las variables producidas internamente y las externas al proceso. Además, este tipo de modelos son considerados instrumentos de análisis para tomar decisiones microeconómicas y macroeconómicas. Estos modelos están integrados por regresiones lineales múltiples, modelos logit y probit y discriminantes lineales.

    Mediante los modelos econométricos se tiene la capacidad de encontrar un modelo que calcule la probabilidad mediante un análisis de los individuos:

    "De manera más general, se trata de determinar el conjunto de atributos (razones financieras) que explican el incumplimiento del acreditado y obtener, mediante un modelo, la probabilidad de que dicho acreditado que hoy pertenece al grupo de cartera vigente, con el tiempo pertenezca al grupo de cartera vencida."(De Lara Haro, 2014)

    El modelo que puede encontrar el conjunto de atributos es el siguiente:

    $$P_i = a_0 + a_1 x_1 + ... + a_n x_n$$

    En donde las a's son los coeficientes del modelo y las x's son las razones financieras que se obtienen de los estados financieros del acreditado i. P_i es la probabilidad de incumplimiento del acreditado i, que sólo puede adquirir valores entre cero y uno.

Capítulo 4

Regresión Lineal y Logística

4.1. Regresión Lineal

El propósito de la presente tesis es calcular la probabilidad de incumplimiento para una base de datos crediticia, implementando una regresión logística binaria.

Las variables que se incorporan en una regresión se dividen en dos tipos: dependientes, las cuales tienen una vinculación con respecto a las otras variables en el estudio, y las independientes que sus resultados no dependen de ninguna otra variable.

El nombre que recibe la regresión depende de la cantidad de variables independientes y los resultados que presenta la variable dependiente, por ejemplo, si tenemos una regresión cuya variable dependiente tiene un comportamiento cuantitativo representado por una línea recta con solo una variable independiente, la regresión recibirá el nombre de regresión lineal simple.

El modelo de regresión lineal simple tiene la siguiente forma:

$$Y = \alpha + \beta X + \varepsilon. \tag{4.1}$$

Donde:

- α : Coeficiente intercepto (valor de Y cuando X=0).
- β : Coeficiente de la pendiente.
- ε : Error (Causas no controladas).
- X: Variable independiente.
- Y: Variable dependiente.

Los supuestos que debe de cumplir la regresión lineal para tener predicciones precisas son los siguientes:

1. $E[\varepsilon] = 0$.
2. $E(\varepsilon\varepsilon') = \sigma^2 In$.
3. Número de observaciones mayor al número de variables.

"El supuesto 1) no implica pérdida de generalidad ni supone ninguna restricción, al menos en el caso en que X tiene entre sus columnas una cuyos valores sean constantes (y ésto suele suceder; típicamente, la primera columna está formada por ?unos?). El supuesto 2), bastante más restrictivo, requiere que las perturbaciones sean incorrelacionadas (covarianzas cero) y homoscedásticas (de idéntica varianza)." (Tusell, 2011)

Considerando el cumplimiento de los supuestos tenemos que la esperanza y la varianza quedan de la siguiente manera:

$$\begin{aligned} E[Y|X = x] &= \alpha + \beta x, \\ Var[Y|X = x] &= \sigma^2. \end{aligned} \tag{4.2}$$

Una interpretación de los coeficientes es la siguiente:

" Los parámetros en la esperanza son la intersección α, que es el valor de $E[Y|X = x]$ cuando x es igual a cero, y la pendiente β, que es la tasa de cambio en la $E[Y|X = x]$ para cada cambio en X. Al variar los parámetros, se pueden obtener todas las líneas rectas posibles. En múltiples regresiones, los parámetros son desconocidos y deben ser estimados usando datos. La varianza en (4.2) se supone constante, con un valor σ^2 que generalmente es desconocido. " (Weisberg, 2005)

Regularmente los datos observados y los esperados difieren al momento de aplicar la regresión, por ello a esta diferencia se le llama error aleatorio o estadístico (ε). El error ε depende de parámetros desconocidos, dado que es una variable aleatorio correspondiente a la distancia entre la Y y el $E[Y|X = x]$.

(Weisberg, 2005)

Como se indica anteriormente, para poder obtener los valores correspondientes de α y β es necesario estimarlos con información de la muestra, por lo que se utilizará la estimación de máximo verosimilitud. La función de verosimilitud de una $N[\mu,\sigma^2]$ con y_i variables independientes donde $i = \{1, ..., n\}$ es la siguiente:

$$f(y_i, \theta) = \frac{1}{(2\pi\sigma^2)^{n/2}} e^{-\sum \frac{(y_i-\mu)^2}{2\sigma^2}}. \tag{4.3}$$

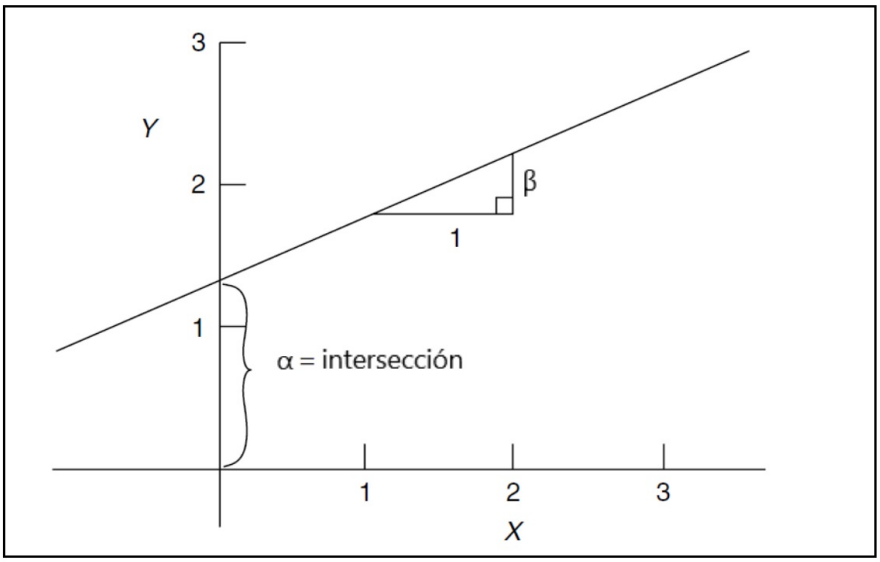


Figura 4.1: Coeficientes de una regresión lineal (Weisberg, 2005)

Sustituyendo las variables que tenemos en (4.2) sustituimos:

$$f(y_i, \theta) = \frac{1}{(2\pi\sigma^2)^{n/2}} e^{-\sum \frac{(y_i - \alpha - \beta x_i)^2}{2\sigma^2}}. \tag{4.4}$$

Ahora, recordando la propiedad de que si z es el punto crítico de $f(y_i, \theta)$, también lo es de $\ln\{f(y_i, \theta)\}$, por lo cual, es suficiente obtener el punto crítico de la logverosimilitud la cual está dada por:

$$\begin{aligned} \ln\{f(y_i, \theta)\} &= \ln\{e^{-\sum \frac{(y_i - \alpha - \beta x_i)^2}{2\sigma^2}}\} - \ln\{(2\pi\sigma^2)^{n/2}\} \\ &= -\sum \frac{(y_i - \alpha - \beta x_i)^2}{2\sigma^2} - \frac{n}{2} * \ln\{2\pi\sigma^2\}. \end{aligned} \tag{4.5}$$

Desarrollando la función de logverosimilitud, podemos calcular las derivadas parciales con respecto a α y a β, respectivamente:

$$\begin{aligned} \frac{\partial}{\partial\alpha}(\ln\{f(y_i, \theta)\}) &= \sum_{i=1}^{n} y_i - n\alpha - \beta \sum_{i=1}^{n} x_i, \\ \frac{\partial}{\partial\beta}(\ln\{f(y_i, \theta)\}) &= \sum_{i=1}^{n} y_i x_i - \alpha \sum_{i=1}^{n} x_i - \beta \sum_{i=1}^{n} x_i^2. \end{aligned}$$

Antes de igualar a cero las derivadas es necesario incluir una serie de fórmulas alternativas

para obtener expresiones más sencillas para las estimaciones de α y β:

$$S_{xx} = \sum_{i=1}^{n} (x_i - \overline{x})^2, \quad S_{yy} = \sum_{i=1}^{n} (y_i - \bar{y})^2, \quad S_{xy} = \sum_{i=1}^{n} (x_i - \bar{x})(y_i - \bar{y}), \tag{4.6}$$

$$\bar{x} = \frac{\sum x_i}{n}, \quad \bar{y} = \frac{\sum y_i}{n}. \tag{4.7}$$

A continuación, se presenta el procedimiento para despejar α:

$$\sum_{i=1}^{n} y_i - n\alpha - \beta \sum_{i=1}^{n} x_i = n\bar{y} - n\alpha - n\beta\bar{x} = 0$$
$$n\bar{y} - n\beta\bar{x} = n\alpha$$

Usualmente se utilizará el símbolo ˆ para representar el estimador de máxima verosimilitud. La estimación de α es la siguiente:

$$\hat{\alpha} = \bar{y} - \hat{\beta}\bar{x} \tag{4.8}$$

Ahora, se prosigue a obtener la estimación de β:

$$\sum_{i=1}^{n} x_i y_i - \alpha \sum_{i=1}^{n} x_i - \beta \sum_{i=1}^{n} x_i^2 = \sum_{i=1}^{n} x_i y_i - (\bar{y} - \beta\bar{x}) \sum_{i=1}^{n} x_i - \beta \sum_{i=1}^{n} x_i^2 = 0$$
$$\sum_{i=1}^{n} x_i y_i - \bar{y} \sum_{i=1}^{n} x_i + \beta\bar{x} \sum_{i=1}^{n} x_i - \beta \sum_{i=1}^{n} x_i^2 = \sum_{i=1}^{n} x_i y_i - n\bar{y}\bar{x} + \beta(\bar{x} \sum_{i=1}^{n} x_i - \sum_{i=1}^{n} x_i^2) = 0$$
$$\sum_{i=1}^{n} x_i y_i - n\bar{y}\bar{x} = \beta(\sum_{i=1}^{n} x_i^2 - n\bar{x^2})$$
$$\frac{\sum_{i=1}^{n} x_i y_i - n\bar{y}\bar{x}}{(\sum_{i=1}^{n} x_i^2 - n\bar{x^2})} = \beta$$

Con la ayuda de las fórmulas adicionales en la parte (4.6) podemos obtener la siguiente fórmula:

$$\hat{\beta} = \frac{S_{xy}}{S_{xx}} \tag{4.9}$$

En donde $\hat{\alpha}$ y $\hat{\beta}$ son los estimadores máximo verosímiles de α y β, respectivamente. En el

gráfico 4.2 se puede ver la recta como el ajuste que se espera obtener y el conjunto de datos resultantes del estudio.

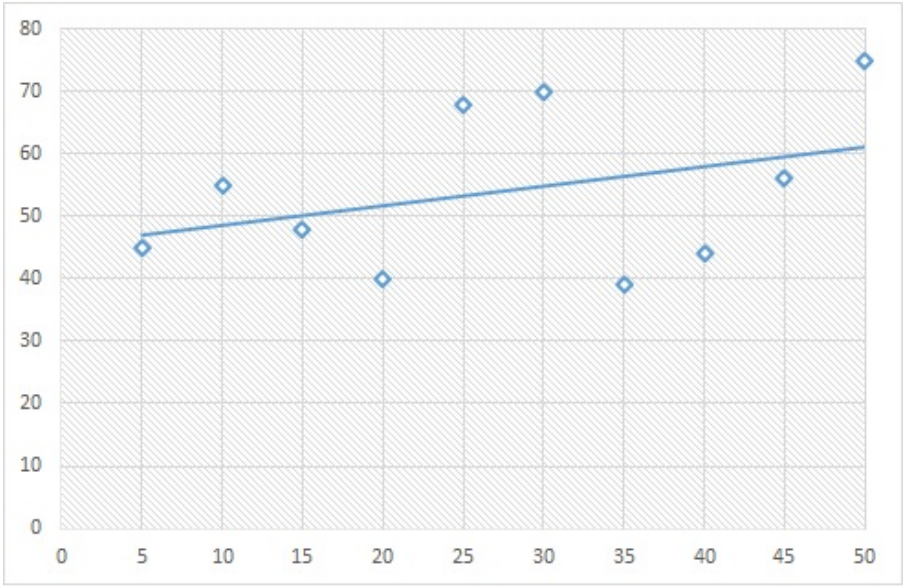


Figura 4.2: Recta de ajuste (Mendenhall; Beaver, 2010)

La variación de la regresión debe ser analizada y dividida con respecto a los factores que integran la regresión. La variación total está dada por:

$$SST = S_{yy} = \sum_{i=1}^{n} (y_i - \bar{y})^2. \tag{4.10}$$

La variación total debe ser separada en dos partes, segun Mendenhall:

- La SSR (suma de cuadrados para regresión) mide la cantidad de variación explicada al usar la recta de regresión con una variable independiente x.
- La SSE (suma de cuadrados de error) mide la variación ?residual? en los datos que no está explicada por la variable independiente x.

$$\begin{aligned} SST &= \sum_{i=1}^{n} (y_i - \bar{y})^2 \\ &= \sum_{i=1}^{n} (\{y_i - \hat{y}_i\} + \{\hat{y}_i - \bar{y}\})^2 \\ &= \sum_{i=1}^{n} (\{y_i - \hat{y}_i\})^2 + \sum_{i=1}^{n} (\{\hat{y}_i - \bar{y}\})^2 + \left[2 \sum_{i=1}^{n} (\{y_i - \hat{y}_i\})(\{\hat{y}_i - \bar{y}\}) = 0\right] \\ &= SSE + SSR \end{aligned} \tag{4.11}$$

En gráfico 4.3 es posible identificar con mayor claridad la interpretación de SSR y SSE para un y_i en específico. La interpretación de SSE es la región que existe entre $\overline{y}$ y la recta de la regresión, mientras que la SSR es la región que hay entre la recta y el valor de y_i.

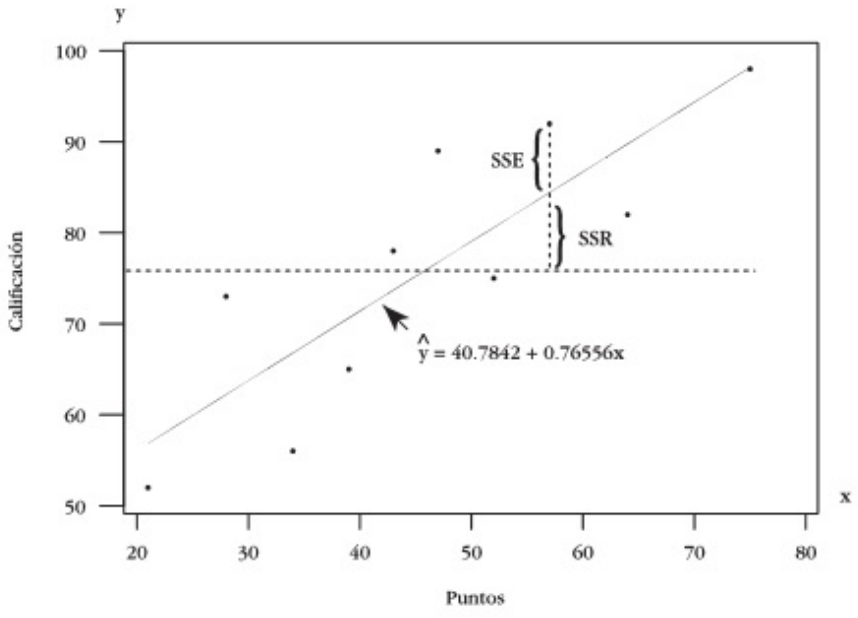


Figura 4.3: Representación SSR Y SSE (Mendenhall; Beaver, 2010)

La forma de calcular la SSE y la SSR es la siguiente:

$$SSR = \sum_{i=1}^{n} (\hat{y}_i - \bar{y})^2, \tag{4.12}$$

$$SSE = SST - SSR. \tag{4.13}$$

Los últimos elementos que conforman a la regresión lineal son el coeficiente de correlación y determinación. El coeficiente de correlación nos da la capacidad de observar el nivel de relación que existe entre las variables en el modelo de regresión. La forma de calcular los coeficientes es el siguiente:

$$R = \frac{S_{xy}}{\sqrt{S_{xx}S_{yy}}} \quad para \quad -1 \leq R \leq 1. \tag{4.14}$$

La interpretación de los valores de R es la siguiente:

- Cuando R es positiva, X aumenta cuando Y aumenta, y viceversa.
- Cuando R es negativa, X disminuye cuando Y aumenta o X aumenta cuando Y disminuye.
- Cuando R es igual a 0, entonces no hay relación lineal aparente entre las variables.
- Cuanto más cercano sea el valor de r a 1 o -1, será más fuerte la relación.

El coeficiente de determinación es el cuadrado del coeficiente de correlación, sin embargo la interpretación difiere, ya que:

"El coeficiente de determinación R^2 se puede interpretar como el porcentaje de reducción en la variación total en el experimento obtenido al usar la recta de regresión, en lugar de ignorar X y usar la media muestral $\bar{y}$ para predecir la variable de respuesta Y."(Mendenhall; Beaver, 2010)

$$R^2 = \left(\frac{S_{xy}}{\sqrt{S_{xx} S_{yy}}} \right)^2. \tag{4.15}$$

4.1.1. Regresión lineal múltiple

La regresión lineal múltiple se representa de la siguiente manera:

$$Y_i = \alpha + \sum_{j=1}^{k} \beta_j X_{ij} + \varepsilon_i. \tag{4.16}$$

En la mayoría de los casos prácticos en donde se aplican regresiones es necesario estudiar múltiples variables con comportamientos no lineales, ya que, los fenómenos económicos o de salud tienen un comportamiento complejo, como podría ser el precio de una acción bursátil en una cierta temporalidad o la causa de una enfermedad con ciertos factores presentes en los afectados.

La esperanza y varianza de una regresión lineal múltiple son las siguientes:

$$\begin{aligned} E[Y|X] &= \alpha + \sum_{j=1}^{k} \beta_j X_{ij}, \\ Var[Y|X] &= \sigma^2. \end{aligned} \tag{4.17}$$

Por cuestiones de comodidad, se define la regresión en términos matriciales de la siguiente manera:

$$Y = \begin{pmatrix} Y_1 \\ Y_2 \\ \dots \\ Y_n \end{pmatrix}, \qquad X = \begin{pmatrix} 1 & x_{11} & \cdots & x_{1k} \\ 1 & x_{21} & \cdots & x_{2k} \\ \vdots & \vdots & \vdots & \vdots \\ 1 & x_{n1} & \cdots & x_{nk} \end{pmatrix}. \tag{4.18}$$

Donde Y es un vector de nX1 y X es una matriz de nX(k+1). Asimismo se definen los

coeficientes B como un vector de (k+1)X1 y el vector de error aleatorio de nX1 de la siguiente forma:

$$\beta = \begin{pmatrix} \alpha \\ \beta_1 \\ \vdots \\ \beta_k \end{pmatrix}, \qquad \varepsilon = \begin{pmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \vdots \\ \varepsilon_n \end{pmatrix}. \tag{4.19}$$

La regresión lineal múltiple en notación matricial queda de la siguiente manera:

$$Y = X\beta + \varepsilon. \tag{4.20}$$

El método de mínimos cuadrados de $\hat{\beta}$ (estimador de β) se utilizar para conocer los coeficientes al minimizar la suma de los cuadrados residuales o *residual sum of squares* (RSS):

$$RSS = \hat{\varepsilon}\prime\hat{\varepsilon} = [Y - X\hat{\beta}]\prime[Y - X\hat{\beta}]. \tag{4.21}$$

Considerando los términos $\hat{\varepsilon}$ y $\hat{Y}$ como:

$$\begin{aligned} \hat{\varepsilon} &= Y - \hat{Y}, \\ \hat{Y} &= X\hat{\beta}. \end{aligned} \tag{4.22}$$

Los símbolos $X\prime$ y X^{-1} se utilizan en esta ocasión para indicar la transpuesta y la inversa de la matriz X respectivamente. La forma de calcular los coeficientes estimados es la siguiente:

$$\hat{\beta} = [X\prime X]^{-1}[X\prime Y]. \tag{4.23}$$

Con el propósito de encontrar los estimadores máximo verosimilitudes, es necesario desarrollar el término RSS de la siguiente manera:

$$\begin{aligned} RSS &= [Y - X\hat{\beta}]\prime[Y - X\hat{\beta}] = [Y\prime - \hat{\beta}\prime X\prime][Y - X\hat{\beta}] = \\ &= Y\prime Y - Y\prime X\hat{\beta} - \hat{\beta}\prime X\prime Y + \hat{\beta}\prime X\prime X\hat{\beta} \\ &= Y\prime Y - 2\hat{\beta}\prime X\prime Y + \hat{\beta}\prime X\prime X\hat{\beta}. \end{aligned}$$

Se busca el estimador máximo verosímil, por lo que se deriva con respecto a $\hat{\beta}$:

$$\frac{\partial RRC}{\partial \hat{\beta}} = -2X\prime Y + 2X\prime X\hat{\beta}. \tag{4.24}$$

Igualando a 0 y despejamos $\hat{\beta}$:

$$\begin{aligned} -2X\prime Y + 2X\prime X\hat{\beta} &= 0 \\ X\prime X\hat{\beta} &= X\prime Y. \end{aligned} \tag{4.25}$$

Por último, se amplía el concepto de coeficiente de determinación a la regresión lineal múltiple. Una definición alternativa de este término es:

"R_2 es también llamado el coeficiente de correlación múltiple dado que es el máximo de la correlación entre Y y cualquier combinación lineal de los términos en la esperanza." (Weisberg, 2005)

El R_2 para el caso múltiple en la regresión lineal es:

$$R^2 = 1 - \frac{RSS}{Y\prime Y}. \tag{4.26}$$

4.2. Indicadores de una regresión

Los indicadores son esenciales para poder medir de forma adecuada los sucesos binarios que se presentan en una regresión logística, de la cual se hablara a detalle en la siguiente sección. Es necesario contextualizar los indicadores que son utilizados con más frecuencia para poder generar un análisis estadístico lo más eficaz posible.

"El *odds* asociado a cierto suceso se define como la razón entre la probabilidad de que dicho suceso ocurra y la probabilidad de que no ocurra; es decir, se trata de un número que expresa cuánto más probable es que se produzca frente a que no se produzca el hecho en ciestión." (Silva Aycaguer; Barroso Utra, 2004)

El termino *odds* también recibe el nombre de momio. Por ejemplo, se puede medir la razón entre qué tan probable es que una persona incumpla en su responsabilidad de pago frente a que no suceda. La fórmula para poder calcular los momios es la siguiente:

$$O(N) = \frac{P(N)}{1 - P(N)}. \tag{4.27}$$

Donde:

- N : Dicho suceso.
- P(N) : Probabilidad de que ocurra.
- O(N) : Momio correspondiente.

Una interpretación de esta fórmula se puede tomar como cuántas veces es más probable que cierto paciente genere una enfermedad comparado a que no presente dicho suceso. La diferencia entre probabilidad y momio es que la probabilidad tiene una escala entre 0 y 1 mientras que los momios van de 0 a infinito. Como forma alternativa es posible ver las probabilidades en términos de momios, a través de la siguiente fórmula:

$$P(N) = \frac{O(N)}{O(N)+1}. \tag{4.28}$$

En el riesgo relativo se tienen dos condiciones ante un evento en común y se mide la razón que existe entre ambas condiciones. La principal diferencia entre la estimación de los momios y el riesgo relativo es que los momios miden la razón entre las probabilidades de cierto evento frente a la ausencia de este y el riesgo relativo es la razón que existe entre dos probabilidades. La fórmula que se utiliza para medir el riesgo relativo o RR es la siguiente:

$$RR = \frac{P_X(N)}{P_Y(N)}. \tag{4.29}$$

La interpretación del riesgo relativo a un evento es el resultado del cociente de la probabilidad de que esté presente el factor X en cierto suceso definido entre la probabilidad del factor Y con la aparición del mismo evento. Con frecuencia, el riesgo relativo es una medida de asociación entre la exposición y el evento en cuestión, como podría ser el riesgo relativo de invertir en Cetes respecto a invertir en una compañía.

Con los conceptos vistos, es posible definir un indicador que incluya ambos conceptos creando una versión alternativa llamada *odds ratio* o cociente de momios. La incorporación de este concepto tiene la intención de ser una herramienta con la capacidad de medir la relación de dos eventos al mismo tiempo, ya que, su cálculo compara dos momios de eventos diferentes, como se puede ver a continuación:

$$OR = \frac{\frac{P_A(N)}{1-P_A(N)}}{\frac{P_{NA}(N)}{1-P_{NA}(N)}} = \frac{O_A(N)}{O_{NA}(N)}. \tag{4.30}$$

Donde:

- A: Presencia del factor A.
- NA: Ausencia del factor A.

- N : Suceso en cuestión.

Con el objetivo de aclarar el uso de los indicadores anteriores, se presentará un ejemplo de cada uno de ellos.

Odds o momios:

$$100 \text{ solicitantes} \quad \left\{ \begin{matrix} 75 & \text{cumplieron} \\ 25 & \text{incumplieron} \end{matrix} \right\} \quad O(C) = \frac{0{,}75}{0{,}25} = 3$$

Riesgo Relativo:

$$100 \text{ solicitantes en el ramo hipotecario} \quad \begin{cases} 90 & \text{cumplieron} \\ 10 & \text{incumplieron} \end{cases}$$

$$100 \text{ solicitantes en el ramo automotriz} \quad \begin{cases} 75 & \text{cumplieron} \\ 25 & \text{incumplieron} \end{cases}$$

$$RR(C) = \frac{0{,}90}{0{,}75} = 1{,}2.$$

Cociente de momios o *odds ratio*:

$$OR(C) = \frac{\frac{0{,}9}{0{,}1}}{\frac{0{,}75}{0{,}25}} = \frac{9}{3} = 3.$$

4.3. Regresión Logística Binaria

En el caso de la regresión logística se tiene que es un tipo de análisis, en donde es necesario que el evento a modelar tenga un número finito de categorías claramente definidas, generalmente binarias, es decir con dos tipos de resultados (ocurrencia y ausencia de una eventualidad). La diferencia principal entre la aplicación de una regresión logística y una lineal es la variable dependiente.

Por lo anterior una regresión logística binaria será la herramienta adecuada para medir la probabilidad de incumplimiento en una base de datos definida, ya que sólo se tienen dos posibles eventos, que son el incumplimiento de un individuo en su responsabilidad de pago y el cumplimiento del mismo.

Es fundamental exponer el método estadístico que será utilizado con cada una de sus características, por lo cual, como primera instancia tenemos que definir la interpretación que tiene la variable dependiente en un modelo binario, la cual es:

$$Y = \begin{cases} 1 & \text{si el evento ocurre} \\ 0 & \text{si el evento no ocurre} \end{cases} \tag{4.31}$$

Es necesario tomar en cuenta la transformación logística de la probabilidad P de que cierto evento ocurra al momento de aplicar la regresión, ya que, se deben de mantener estables las predicciones entre los límites (0,1) y el cumplimiento de la curva logística (curva en forma de S dado el comportamiento de la función logaritmo) entre el riesgo y los niveles de exposición. Tomando en cuenta lo anterior, es necesario definir la variable dependiente de la siguiente manera:

$$Y = \ln\left(\frac{P}{1-P}\right). \tag{4.32}$$

Sustituyendo la fórmula de la regresión con el nuevo término definido, la expresión de la variable dependiente queda de la siguiente forma:

$$\ln\left(\frac{P}{1-P}\right) = \alpha + \beta_1 x_1 + \beta_2 x_2 + ... + \beta_{k-1} x_{k-1} + \beta_k x_k. \tag{4.33}$$

Despejando la función logaritmo de la expresión y tras hacer algunas operaciones algebraicas sencillas, tenemos que:

$$\begin{aligned}
\frac{P}{1-P} =& \exp\left\{\alpha + \beta_1 x_1 + \beta_2 x_2 + ... + \beta_{k-1} x_{k-1} + \beta_k x_k\right\}, \\
P(Y=1) =& \frac{\exp\left\{\alpha + \beta_1 x_1 + \beta_2 x_2 + ... + \beta_{k-1} x_{k-1} + \beta_k x_k\right\}}{1 + \exp\left\{\alpha + \beta_1 x_1 + \beta_2 x_2 + ... + \beta_{k-1} x_{k-1} + \beta_k x_k\right\}} = \\
= & \frac{1}{(1+\exp\{\alpha+\beta_1 x_1+\beta_2 x_2+...+\beta_{k-1} x_{k-1}+\beta_k x_k\})} \frac{1}{\exp\{-\alpha-\beta_1 x_1-\beta_2 x_2-...-\beta_{k-1} x_{k-1}-\beta_k x_k\}}.
\end{aligned}$$

La última expresión puede ser simplificada de la siguiente forma:

$$P(Y=1) = \frac{1}{1 + \exp\left\{-\alpha - \beta_1 x_1 - ... - \beta_k x_k\right\}} = P. \tag{4.34}$$

El método para estimar los coeficientes en una regresión logística es considerablemente más complejo que el método de máxima verosimilitud en una regresión lineal. Los parámetros del modelo deben ser estimados con la información de la muestra, con el objetivo de atribuir una alta probabilidad a los sujetos que cumplen la condición (1 o lo más cercano posible), mientras que en el resto se debe producir una probabilidad baja (es decir, lo más cercano a 0). La verosimilitud del modelo es una medida adecuada para una correcta estimación de coeficientes, como indica Silva Aycaguer y Barroso Utra:

"Una medida razonable para valorar el grado en que el modelo arroja resultados coherentes con la condición que tienen los sujetos de la muestra empleada para su construcción sería el producto de todas las probabilidades (predichas por el modelo) de que los n sujetos de dicha muestra tengan la condición que realmente tienen." (Silva Aycaguer; Barroso Utra, 2004)

Se define P_i como la probabilidad de que el sujeto i caiga en incumplimiento, siendo la generalización de (4.34). Asumiendo que las observaciones son independientes, la verosimilitud del modelo queda de la siguiente manera:

$$V = P_1 P_2 \cdots P_d (1 - P_{d+1})(1 - P_{d+2}) \cdots (1 - P_n). \tag{4.35}$$

En donde se tienen d sujetos con incumplimiento y $n - d$ sin esta condición. Las P_i corresponden a la generalización dEntre más cercano sea el valor de V a 1 más adecuado será el modelo, como indica Silva Aycaguer y Barroso Utra:

"La proximidad de la verosimilitud a 1 expresa cuán eficiente ha sido este recurso para modelar la realidad. Los mejores valores para los parámetros del modelo serán aquellos que hagan que la función de verosimilitud sea lo más grande posible. Las llamadas *estimaciones máximo verosímiles* son, por tanto, aquellos valores de los coeficientes que dan lugar al máximo valor de la verosimilitud del modelo." (Silva Aycaguer; Barroso Utra, 2004)

Para obtener las estimaciones máximo verosímiles es necesario definir algunas variables auxiliares, como es el término λ_i definido de la siguiente manera:

$$\lambda_i = P_i^{Y_i}(1 - P_i)^{1-Y_i}. \tag{4.36}$$

La expresión (4.36) se interpreta de la siguiente manera: si el sujeto i cae en incumplimiento el valor de Y será igual a 1 llegando a la expresión de $\lambda_i = P_i$. La reinterpretación de la verosimilitud del modelo queda de la siguiente manera:

$$V = \prod_{i=1}^{n} \lambda_i. \tag{4.37}$$

Utilizando la logverosimiltud dada su mayor facilidad al momento de hacer operaciones algebraicas:

$$L(V) = \ln(V) = \sum_{i=1}^{n} \{y_i \ln[P_i] + (1 - y_i) \ln[1 - P_i]\}. \tag{4.38}$$

Con el objetivo de encontrar el valor de V que maximice la logverosimilitud, es necesario:

"Para encontrar el valor de $\beta = (\alpha, \beta_0)$ que maximice $L(V)$ nosotros derivamos $L(V)$ con respecto a α y β_0 e igualamos las expresiones resultantes a 0."(Lemeshow; Hosmer, 2000)

Dado que la demostración esta formulada para la regresión logística simple es necesario hacer algunas modificaciones:

"Como en el caso univariado, el ajuste del modelo requiere que obtengamos las estimaciones

del vector $\beta = (\alpha, \beta_0, ..., \beta_k)$. El método de estimación utilizado en el caso múltiple será el mismo que en el caso simple - la máxima verosimilitud. "(Lemeshow; Hosmer, 2000)

Las expresiones quedan de la siguiente manera:

$$\sum_{i=1}^{n} [y_i - P_i] = 0, \tag{4.39}$$

$$\sum_{i=1}^{n} x_{ij} [y_i - P_i] = 0. \tag{4.40}$$

Donde: $i = 1, 2, ..., n$ y $j = 1, ..., k$.

Las expresiones (4.39) y (4.40) son no lineales, por lo que es necesario usar métodos numéricos para poder encontrar las soluciones, lo cual no representa un problema, ya que es posible programar estos métodos fácilmente. Para mayores detalles sobre los métodos disponibles se recomienda consultar McCullagh y Nelder (1983).

La estimación de varianzas y covarianzas será calculada con el método de máxima verosimilitud, utilizando la matriz de derivadas parciales de la logverosimilitud. Las fórmulas de las derivadas parciales son las siguientes:

$$\frac{\partial^2 L(V)}{\partial \beta_j^2} = -\sum_{j=1}^{n} x_{ij} P_i (1 - P_i), \tag{4.41}$$

$$\frac{\partial^2 L(V)}{\partial \beta_j \partial \beta_u} = -\sum_{j=1}^{n} x_{ij} x_{iu} P_i (1 - P_i). \tag{4.42}$$

Donde: $i = 1, 2, ..., n$ y $j, u = 1, ..., k$.

La información de la matriz ($I(\beta)$) es una matriz de dimensión (k+1)X(k+1) cuyas entradas están dadas por las derivadas parciales (4.41) y (4.42). Con la inversa de esta matriz (denominada $\sum(V) = I^{-1}(V)$) es posible calcular las varianzas y covarianzas de los coeficientes que fueron estimados. La notación para la varianza del estimador $\hat{\beta}_j$ es $\sigma^2(\beta_j)$. En el caso de la covarianza tenemos $\sigma^2(\beta_j, \beta_i)$ como la covarianza los estimadores $\hat{\beta}_j$ y $\hat{\beta}_i$.

La forma de calcular la información de la matriz corresponde a la siguiente formula:

$$\hat{I}(V) = X'VX. \tag{4.43}$$

Donde el elemento X es una matriz de dimensiones (n)X(k+1) y V es de dimensiones (n)X(n). La forma matricial de estos elementos es la siguiente:

$$X = \begin{bmatrix} 1 & x_{11} \cdots & x_{1d} \\ 1 & x_{21} & x_{2d} \\ \vdots & \vdots & \vdots \\ 1 & x_{n1} \cdots & x_{nd} \end{bmatrix}, \tag{4.44}$$

$$V = \begin{bmatrix} \hat{P}_1(1-\hat{P}_1) & 0 & \cdots & \cdots & 0 \\ 0 & \hat{P}_2(1-\hat{P}_2) & \cdots & \cdots & 0 \\ 0 & 0 & \hat{P}_3(1-\hat{P}_3) & \cdots & 0 \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ 0 & 0 & \cdots & \cdots & \hat{P}_n(1-\hat{P}_n) \end{bmatrix}. \tag{4.45}$$

Después del cálculo de los estimadores, es necesario evaluar la significancia de las variables en el modelo por medio de fórmulas y pruebas de hipótesis, con el objetivo de determinar que variables no aportan información adicional al modelo. La forma de calcular la significancia del modelo es:

"El principio con la regresión logística es el mismo: Comparar los valores observados de la variable respuesta con los valores predichos obtenidos en el modelo con y sin la variable en cuestión. En la regresión logística, la comparación de los valores observados con los predichos se basa en la función de logverosimilitud definida en la ecuación (4.38). Para comprender mejor esta comparación, es conceptualmente útil si pensamos en un valor observado de la variable respuesta como un valor predicho resultante de un modelo saturado. Un modelo saturado es aquel que contiene tantos parámetros como datos. " (Lemeshow; Hosmer, 2000)

La comparación entre los valores observados y los predichos usando la logverosimilitud está basada en el siguiente cociente:

$$D = -2\ln\left[\frac{Verosimilitud \quad del \quad modelo \quad actual}{Verosimilitud \quad del \quad modelo \quad saturado}\right]. \tag{4.46}$$

El cálculo de la fórmula (4.46) responde a tener una estimador cuya distribución es conocida para poder elaborar una prueba de hipótesis. Sustituyendo la formula (4.38) en (4.46):

$$D = -2\sum_{i=1}^{n}\left[y_i \ln\left(\frac{\hat{P}_i}{y_i}\right) + (1-y_i)\ln\left(\frac{1-\hat{P}_i}{1-y_i}\right)\right]. \tag{4.47}$$

La fórmula (4.47) cumple un papel importante al momento de encontrar el modelo con el mejor ajuste posible:

"El estadístico, D, en la ecuación (4.47) es llamado devianza por algunos autores (por ejemplo, McCullagh and Nelder (1983)), y juega un papel central en algunos enfoques para la estimación de la bondad de ajuste. La devianza para la regresión logística juega el mismo papel que la R_2 en la regresión lineal." (Lemeshow; Hosmer, 2000)

La devianza tiene la capacidad de comparar modelos que difieran en una o más variables. El único requisito es que sean modelos anidados, es decir que el modelo más pequeño esté contenido en el grande. La fórmula de la comparación es la siguiente:

$$G = -2\ln\left[\frac{Verosimilitud \quad sin \quad la \quad variable}{Verosimilitud \quad con \quad la \quad variable}\right]. \tag{4.48}$$

Es necesario agregar algo de notación para expandir la formula (4.48). Dado que las P_i provienen de las ecuaciones (4.34) y (4.38), es posible decir que el factor α es equivalente a $ln(n_1/n_0)$, en donde $n_1 = \sum y_i$ y $n_0 = \sum (1 - y_i)$. La expresión para G queda de la siguiente manera:

$$G = 2\left\{\sum_{i=1}^{n}\left[y_i \ln\left(\hat{P}_i\right) + (1 - y_i)\ln\left(1 - \hat{P}_i\right)\right] - [n_1 \ln(n_1) + n_0 \ln(n_0) - n\ln(n)]\right\}. \tag{4.49}$$

Asumiendo la hipótesis nula de la formula (4.49), los k coeficientes en el modelo son iguales a 0, se tendría que la distribución de G seria de una Ji-Cuadrada con k grados de libertad.

Capítulo 5

Modelo de Regresión Logística

5.1. Descripción de las Variables

Las variables son elementos que pueden cambiar dependiendo del uso requerido, ya que cuentan con la capacidad de adoptar este cambio en diferentes tipos, cualitativas y cuantitativas.

Las variables cualitativas representan, como su nombre lo indica, cualidades o características acumuladas, dependiendo de si son dicotómicas (dos valores) o politómicas (varias categorías). Este tipo de variables se dividen en ordinales, que implementan una escala establecida como por ejemplo: malo, regular y bueno, y nominales, que no conllevan un orden, como el equipo de baloncesto preferido en un cierto estado.

Por su parte, las variables cuantitativas son aquellas que aceptan elementos numéricos divididos en dos tipos, discretas y continuas. Las variables continuas aceptan cualquier valor dentro de un rango, como pueden ser las alturas de los alumnos de un cierto salón, mientras que las variables discretas presentan interrupciones en la escala establecida, como puede ser el número de días que una persona usa su automóvil en un mes.

En el ramo de crédito, existen ciertos aspectos que son fundamentales basándose en la experiencia de actuarios en México, los cuales fueron consultados. Estos aspectos son los siguientes:

- Tasa de interés.
- Plazo.
- Aspectos monetarios.
- Condiciones.

Las variables que conforman este modelo fueron simuladas a partir de una base real de clientes de cierto banco. La simulación y los procedimientos del modelo fueron realizados en el

programa estadístico R-project dada su amplia gama de bibliotecas, la forma intuitiva de su lenguaje y el alto grado de precisión en los cálculos. El código sobre la simulación se encuentra en el Anexo 1 de la presente tesis. La muestra que se toma de la base original es de tamaño 1000, con el propósito de tener p-values altos al momento de simular las variables continuas. El modelo está conformado por 8 variables: 3 continuas y 5 categóricas, las cuales son presentadas a continuación:

- Categóricas

  - Género
    Definición: Describe el género del deudor, es decir, si es masculino o femenino.
    Descripción:

Femenino	Masculino
563	437

    Características:

$$\text{Genero} \begin{cases} 1 - \text{Femenino} \\ 2 - \text{Masculino} \end{cases}$$

  - Tasa de interés
    Definición: Describe el nivel de tasa de interés aplicable al préstamo otorgado.
    Descripción:

0-10	11-20	Mayor a 20
368	196	436

    Características:

$$\text{Tasa Interés} \begin{cases} 1 - \text{De 0 a 10.} \\ 2 - \text{De 11 a 20.} \\ 3 - \text{Mayor a 20.} \end{cases}$$

  - Modalidad de pago
    Definición: Describe la modalidad en que fueron acordados los pagos del crédito.
    Descripción:

Pagos Únicos	Pagos Mensuales	Pagos Periódicos
216	667	117

    Características:

$$\text{Modalidad Pago} \begin{cases} 1 - \text{Pagos únicos.} \\ 2 - \text{Pagos mensuales.} \\ 3 - \text{Pagos periódicos.} \end{cases}$$

- Destino
  Definición: Describe el destino del crédito solicitado. Los destinos registrados son personales, otros créditos al consumo (automotriz, nómina) y otros (hipotecarios, pyme, ABCD, etc).
  Descripción:

Personales	Consumo	Otros
794	154	52

  Características:

$$\text{Destino} \begin{cases} 1 & \text{Personales.} \\ 2 & \text{Consumo.} \\ 3 & \text{Otros.} \end{cases}$$

- Incumplimiento
  Definición: Describe el cumplimiento o incumplimiento del solicitante con respecto al crédito solicitado. Se marca como incumplimiento cuando los días de mora superan los 90. (CNBV)
  Descripción:

Cumplimiento	Incumplimiento
743	257

  Características:

$$\text{Incumplimiento} \begin{cases} 0 & \text{Cumplimiento.} \\ 1 & \text{Incumplimiento.} \end{cases}$$

- Continuas

  - Saldo Insoluto
    Definición:
    Se refiere a la cantidad que se debe, después de los abonos al crédito.
    Descripción:

Mínimo	Mediana	Media	Máximo
393.7	9274	18290	385000

  - Garantía Liquida
    Definición:
    Es la cantidad que deja el deudor como garantía a favor de la institución que otorgó el préstamo, como puede ser, el dinero que se deja en una cuenta de ahorro.
    Descripción:

  - Monto Original
    Definición:

Mínimo	Mediana	Media	Máximo
66.4	3424	8209	366400

El monto original es la cantidad correspondiente a la deuda inicial sin aplicar intereses.
Descripción:

Mínimo	Mediana	Media	Máximo
378.7	11990	20870	293800

- Variable dependiente

  En el caso de la simulación de la variable dependiente se utilizó una metodología diferente, con el propósito de mejorar la precisión del ajuste de la variable simulada con la original teniendo énfasis en los estimadores del modelo. La simulación está conformada de la siguiente manera:

  1. Estimadores:
     Se calculan los estimadores (B´s) aplicando una regresión logística sobre la base original. La regresión se aplica sobre todas las variables especificadas anteriormente. La fórmula utilizada está dada por la ecuación (4.33).
  2. Coeficientes:
     En esta parte se calculan los coeficientes con respecto a cada individuo de la base simulada con respecto a los estimadores del paso anterior. El cálculo de los coeficientes es de la siguiente manera:

$$\eta_i = \widehat{\alpha} + \widehat{B}_1 \acute{X}_{1i} + ... + \widehat{B}_7 \acute{X}_{7i}. \tag{5.1}$$

     En donde:
     $\widehat{\alpha}$: Interceptor sobre la base original.
     $\widehat{B}_i$: Estimadores de la regresión sobre la base original.
     $\acute{X}_{ji}$:Valor de la variable j del individuo i de la base simulada. Con j=1,...,7 y i=1,...,1000
  3. Probabilidad:
     Se calculan las probabilidades de cada individuo utilizando la siguiente formula:

$$P_i = \frac{exp\{\eta_i\}}{1 + exp\{\eta_i\}}. \tag{5.2}$$

  4. Simulación:
     Con las P_i respectivas de cada individuo, procedemos a simular la variable dependiente como se especifica en el Anexo 1.

En el caso específico de la variable *tasa de interés* fue necesario convertirla en categórica, dado que es una variable bi-modal, tema que sobrepasa los alcances de esta tesis. En la variable *monto original* se presenta la misma condición, sin embargo, en el proceso de simulación se alcanza la precisión mínima aceptable, que es 0.05, por lo cual, se decidió conservarla como continua.

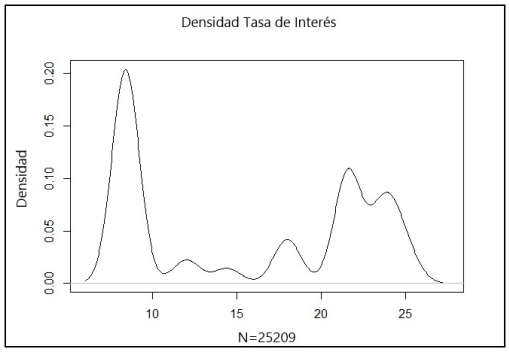


Figura 5.1: Gráfico evidencia tasa de interés

5.2. Resultados del Modelo

En está sección se presentan los resultados de la regresión logística aplicada a esta base con sus respectivas pruebas de ajuste y análisis de residuos. En estos procedimientos se evita utilizar métodos automáticos de selección de variables, dadas las características de la base y los aspectos especiales antes mencionados. La regresión tiene la siguiente expresión:

$$\begin{aligned}\eta_i = \alpha + \beta_1 Genero2_i + \beta_2 Destino2_i + \beta_3 Destino3_i + \beta_4 Modalidad2_i + \\ + \beta_5 Modalidad3_i + \beta_6 Monto_i + \beta_7 Interes2_i + \beta_8 Interes3_i + \\ + \beta_9 Saldo_i + \beta_{10} Garantia_i.\end{aligned} \tag{5.3}$$

$$PI_i = \frac{1}{1 + exp\{-\eta_i\}}. \tag{5.4}$$

En donde:

- Genero2: Tiene valor 1 cuando el género es masculino y 0 en otro caso.
- Destino2: Tiene valor 1 cuando el destino es consumo y 0 en otro caso.
- Destino3: Tiene valor 1 cuando el destino es otros y 0 en otro caso.
- Modalidad2: Tiene valor 1 cuando los pagos son mensuales y 0 en otro caso.
- Modalidad3: Tiene valor 1 cuando los pagos son periódicos y 0 en otro caso.
- Interes2: Tiene valor 1 cuando la tasa de interés está entre 11 % y 20 % y 0 en otro caso.
- Interes3: Tiene valor 1 cuando la tasa de interés es mayor a 20 % y 0 en otro caso.

Los pasos a seguir a partir de esta ecuación serán presentar los coeficientes de cada variable con sus respectivos indicadores, realizar pruebas de bondad de ajuste al modelo, medir el poder predictivo mediante la curva ROC (*A receiver operating characteristic*) y hacer diversas

pruebas estadísticas para observar el nivel de significancia de los estimadores. A continuación, se presentan los resultados de la regresión:

```
Coefficients:
Estimate  Std. Error  z value Pr(>|z|)
(Intercept) -1.989e+01  2.520e+00  -7.892 2.96e-15 ***
Genero2     -8.959e-01  3.388e-01  -2.644  0.00819 **
Destino2     1.833e+00  4.616e-01   3.971 7.17e-05 ***
Destino3    -2.804e+01  3.318e+02  -0.084  0.93267
Modalidad2   8.062e-01  3.729e-01   2.162  0.03060 *
Modalidad3  -3.456e+00  7.949e-01  -4.347 1.38e-05 ***
Monto       -3.901e-04  4.037e-05  -9.663  < 2e-16 ***
Intereses2   1.882e+01  2.364e+00   7.963 1.67e-15 ***
Intereses3   2.066e+01  2.514e+00   8.216  < 2e-16 ***
Saldo        3.239e-04  3.493e-05   9.272  < 2e-16 ***
Garantia     4.812e-05  1.204e-05   3.998 6.39e-05 ***
---
##Términos que indican la signicación de los parámetros.

Signif. codes:  0 "***" 0.001 "**" 0.01 "*" 0.05 "." 0.1 " " 1

#Estimate: estimaciones del modelo.
#Std. Error: dispersión de los datos con respecto a la media.
#z value: coeficiente dividido entre el error. Se utiliza z dado que la distribución
asintótica del parámetro es normal (Cañadas Reche}).
#Pr(>|z|): p-value del contraste con el respectivo nivel de confianza.


Null deviance: 1226.77  on 999  degrees of freedom
Residual deviance:  256.89  on 989  degrees of freedom
AIC: 278.89

#Null deviance: Devianza del modelo nulo.
```

```
#Residual deviance: Devianza del modelo ajustado.
```

Teniendo los estimadores del modelo es necesario:

"Una vez estimado el modelo, partiendo de que se ha realizado un muestreo probabilístico, nos interesa contrastar si los coeficientes estimados son significativamente distintos de 0. Es decir, si una determinada variable explicativa tiene un efecto significativo sobre la respuesta o no. " (Cañadas Reche, 2013)

Para comprobar si los estimadores son significativos se utilizaran pruebas para contrastar hipótesis. Para cada r=1,...,R se contrastarán las siguientes hipótesis:

$$H_0 : \beta_r = 0$$
$$H_1 : \beta_r \neq 0$$

Las pruebas estadísticas que se aplicarán serán el contraste de Wald y el contraste condicional de razón de verosimilitud. El contrate de Wald se basa en la normalidad asintótica de los estimadores. La expresión del contraste es la siguiente:

$$W_r = \frac{\hat{\beta}_r}{\widehat{SE(\beta_r)}}. \tag{5.5}$$

Donde:

- $\hat{\beta}_r$: Estimación para β_r.
- $\widehat{SE(\beta_r)}$: Error estándar de β_r.

Los valores de los estadísticos de Wald (z value) indican que la variable *destino* con $z_{\alpha/2}$=0.084 se encuentran ligeramente debajo del valor absoluto del punto crítico, que es $z_{\alpha/2}$=1.96 con nivel de significancia de 0.05. Para comprobar si el modelo mejora, se retira la variable en cuestión y se usa el contraste condicional de razón de verosimilitud comparando el modelo original y el modelo retirando la variable. La forma del contraste condicional de razón de verosimilitud es la siguiente:

$$G^2_{sin}|G^2_{con} = -2log\frac{V_{sin}}{V_{con}} = -2(L_{sin} - L_{con}) = G^2_{sin} - G^2_{con}. \tag{5.6}$$

Donde:

- V_i: Máxima Verosimilitud del modelo con y sin la variable en cuestión.

- L_i: Máxima Logverosimilitud del modelo con y sin la variable en cuestión.

El objetivo de este contrate es la disminución de devianza entre un modelo a otro. La distribución de este estadístico es ji-cuadrada con grados de libertad igual a la diferencia de la distribución ji-cuadrada de ambos modelos, en otras palabras:

"Si la diferencia entre los dos modelos es en un parámetro, la distribución del estadístico, bajo la hipótesis nula de que ese parámetro es 0, sigue una distribución ji-cuadrada con un grado de libertad. " (Cañadas Reche, 2013)

En R, existe la función Anova del paquete "car" que permite hacer este contraste sobre los parámetros asociados a cada variable del modelo sin la necesidad de generar cada regresión retirando una variable a la vez. Los resultados de esta función son los siguientes:

```
Analysis of Deviance Table (Type II tests)

Response: Incumplimiento
LR Chisq Df Pr(>Chisq)
Genero        7.33  1   0.006787 **
Destino     134.53  2  < 2.2e-16 ***
Modalidad    46.04  2  1.007e-10 ***
Monto       370.40  1  < 2.2e-16 ***
Intereses   489.16  2  < 2.2e-16 ***
Saldo       472.29  1  < 2.2e-16 ***
Garantia     26.13  1  3.186e-07 ***
---
Signif. codes:  0 ?***? 0.001 ?**? 0.01 ?*? 0.05 ?.? 0.1 ? ? 1
```

Se rechaza la hipótesis nula de que la variable *destino* tenga asociado un coeficiente de valor 0, dada la significancia que muestra el p-value. Con esto se puede llegar a la conclusión de que no se encuentran variables con coeficientes no significativos en el modelo antes presentado. Teniendo esto en cuenta podemos proseguir a las pruebas de bondad de ajuste mediante el contraste Hosmer-Lemeshow, la pseudo R_2 de McFadden y la pseudo R_2 de Cox y Snell.

El contraste de Hosmer-Lemeshow consiste en crear 10 grupos donde el estadístico asociado siga una asintóticamente una distribución ji-cuadrada con 8 grados de libertad. :

"Hosmer y Lemeshow (1980) y Lemeshow y Hosmer (1982) proponen agrupar basándose en los valores de las probabilidades estimadas. Supongamos en este caso, que $J = n$. En esta ocasión, pensamos que las n columnas corresponden a los n valores de las probabilidades estimadas, con la primera columna correspondiente al valor más pequeño, y la enésima columna al valor más grande. Se propusieron dos estrategias de agrupación de la siguiente manera: (1)

colapsar la tabla según los percentiles de las probabilidades estimadas y (2) colapsar la tabla en función de los valores fijos de la probabilidad estimada."(Hosmer; Lemeshow, 2000)

En la cita anterior, el valor de J representa los distintos valores de la probabilidad P y n como el número de individuos en la muestra. El objetivo de esta prueba es verificar si el modelo ajusta de forma correcta con respecto a los valores observados. La forma del contraste es la siguiente:

$$HL = \sum_{i=1}^{10} \frac{(O_i - E_i)^2}{E_i}. \tag{5.7}$$

Donde :

- E_i: casos esperados en el grupo i.
- O_i: casos observados en el grupo i.

La forma en las que es posible dividir los datos en grupos son diversas. Algunas de ellas son dividir las probabilidades estimadas en intervalos de igual amplitud o repartir los datos en base a los cuantiles de la distribución creando grupos más homogéneos. En el caso del modelo se utilizó el primer método antes mencionado:

	Incumplimiento=0			Incumplimiento=1		
Intervalos	Observados	Esperados	—Diferencia—	Observados	Esperados	—Diferencia—
(0.001,0.1]	612	614.66	2.66	6	3.33	2.67
(0.1,0.2]	26	27.31	1.31	6	4.68	1.32
(0.2,0.3]	18	17.22	0.78	5	5.77	0.77
(0.3,0.4]	15	12.28	2.72	4	6.71	2.71
(0.4,0.5]	11	9.84	1.16	7	8.15	1.15
(0.5,0.6]	9	6.46	2.54	5	7.53	2.53
(0.6,0.7]	5	7.32	2.32	17	14.67	2.33
(0.7,0.8]	7	5.61	1.39	16	17.38	1.38
(0.8,0.9]	3	3.85	0.85	23	22.14	0.86
(0.9,1]	2	3.39	1.39	203	201.60	1.4

Los resultados del contraste de Hosmer-Lemeshow son los siguientes:

```
> hosmer<-sum((yobs-yexp)^2/yexp)
> hosmer
[1] 8.902631
> p.valuehos<-1-pchisq(hosmer, 8)
```

```
> p.valuehos
[1] 0.3718681
```

El p-value del contrate no muestra evidencias de falta de ajuste en el modelo, al sobrepasar el valor mínimo de 0.05 con 95 % de confianza. Continuando con las medidas de bondad de ajuste, las expresiones para calcular el pseudo R_2 de McFadden y el pseudo de Cox y Snell son las siguientes:

$$R^2_{MF} = 1 - \frac{L(\hat{V})_{Modelo}}{L(\hat{V})_{Interceptor}}, \qquad R^2_{CN} = 1 - \left(\frac{V_{Interceptor}}{V_{Modelo}}\right)^{2/N}. \tag{5.8}$$

Donde:

- $L(\hat{V})_{Modelo}$: logverosimilitud del modelo ajustado.
- $L(\hat{V})_{Interceptor}$: logverosimilitud del modelo con sólo el termino constante.
- $V_{Interceptor}$: máxima verosimilitud del modelo con sólo el termino constante.
- V_{Modelo}: máxima verosimilitud del modelo ajustado.
- N: número de perfiles ajustados.

Los resultados de las pruebas aplicadas al modelo presentado son:

```
#Pseudo McFadden

> RsqrMCFadden<-1-(Regre_Sim$deviance/Regre_Sim$null.deviance)
> RsqrMCFadden
[1] 0.7907186

#Pseudo Cox y Snell

> LRCS<-Regre_Sim$null.deviance - Regre_Sim$deviance
> N<-sum(weights(Regre_Sim))
> RsqrCN<-1-exp(-LRCS/N)
> RsqrCN
[1] 0.6208684
```

Al igual que en el caso del contraste Hosmer-Lemeshow, ambas medidas no presentan evidencia de falta de ajuste en la regresión logística aplicada al modelo. Más adelante se presenta un comparativo entre el modelo original y los modelos retirando cada una de las variables. Teniendo esto en cuenta es posible proseguir con las pruebas de poder predictivo mediante la curva ROC.

La curva ROC es una prueba de poder predictivo mediante modelos de clasificación, ya que:

" La curva ROC es un gráfico de la sensibilidad en función de (1-especificidad) para los posibles puntos de corte π_0. Una curva ROC es más informativa que una tabla de clasificación, ya que resume la potencia predictiva de todos los π_0 posibles. Cuando π_0 se acerca a 0, casi todas las predicciones son $\hat{y} = 1$; entonces, la sensibilidad está cerca de 1, la especificidad está cerca de 0, y el punto de (1-especificidad, sensibilidad) tiene coordenadas cerca de (1,1). Cuando π_0 llega cerca de 1, casi todas las predicciones son $\hat{y} = 0$; Entonces, la sensibilidad está cerca de 0, la especificidad está cerca de 1, y el punto de (1-especificidad, sensibilidad) tiene coordenadas próximas (0,0). La curva ROC suele tener una forma cóncava que conecta los puntos (0,0) y (1,1)." (Agresti, 2007)

Las tablas de clasificación catalogan las observaciones con un punto de corte establecido de la siguiente manera:

	Exito	Fracaso
Exito	Verdaderos Positivos (VP)	Falsos Negativos (FN)
Fracaso	Falsos Positivos (FP)	Verdaderos Negativos (VN)

La curva ROC se establece como la curva que conecta los diversos puntos de corte desde las coordenadas antes mencionadas. A continuación, se presenta un esquema que explica de mejor manera la construcción:

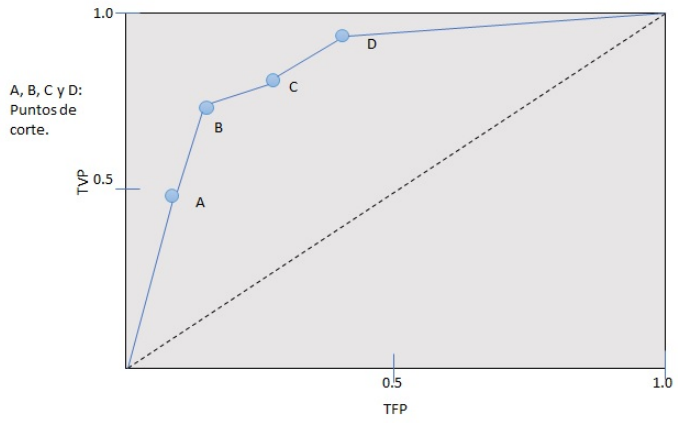


Figura 5.2: Curva ROC

En donde:

$$TFP = \frac{FP}{FP+VN} \quad TVP = \frac{VP}{VP+FN} \tag{5.9}$$

Con ayuda de la biblioteca "ROC" es posible establecer el punto máximo de corte con el que es posible maximizar el porcentaje de individuos correctamente clasificados. La tabla siguiente expresa los diferentes puntos de corte con los datos correspondientes:

Punto	Clasificación (%)	VP	FP	FN	VN
0.3	92.8	641	16	56	287
0.4	93.7	657	23	40	280
0.5	94.2	670	31	27	272
0.6	94.9	682	36	15	267
0.7	94.2	685	46	12	257
0.8	93	692	65	5	238

El código para establecer la curva ROC con el punto de corte máximo es la siguiente:

```
> predigo<-prediction(fitted.values(Regre_Sim),Base_Simulada_Inc$Incumplimiento)
> tab<-performance(predigo,measure = "acc")
> posicionmax<-sapply(tab@y.values,which.max)
> posicionmax
[1] 275
> puntocorte<-sapply(tab@x.values,"[",posicionmax)
> puntocorte
#Punto de corte máximo
0.599068
```

Con base en la estimación del punto de corte máximo se obtiene una clasificación correcta del 94.9 %, lo cual da un margen aceptable de poder predictivo. La tabla de clasificación con este punto de corte es la siguiente:

	Exito	Fracaso
Exito	682	15
Fracaso	36	267

El área debajo la curva que presenta el modelo es la siguiente:

```
[1] "Area under the ROC curve"
```

```
> AUC@y.values
[[1]]
[1] 0.9877078
```

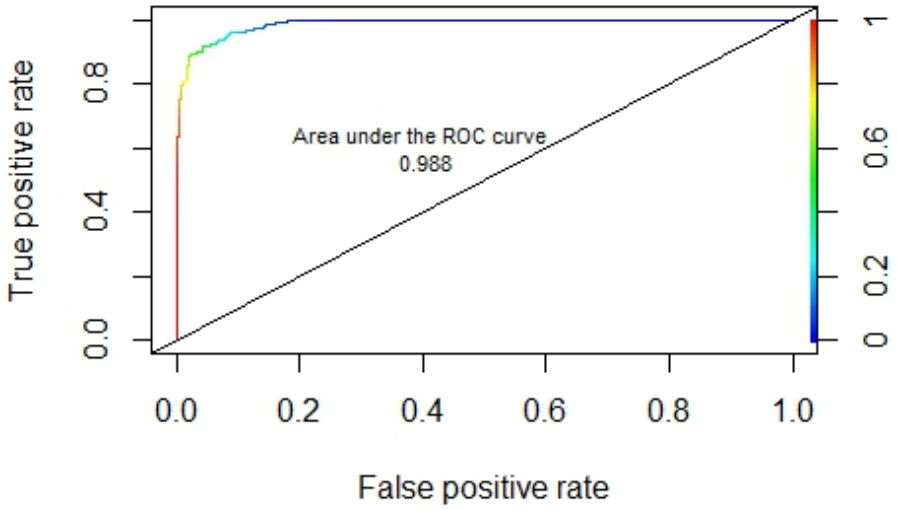


Figura 5.3: Gráfico curva ROC

Se puede concluir que el modelo no tiene evidencias de falta de ajuste debido a los resultados en el contraste Hosmer-Lemeshow, la pseudo R_2 de McFadden y la pseudo R_2 de Cox y Snell, además de un poder predictiva aceptable como lo indica el área debajo de la curva. Con ello se tiene evidencia de que la regresión logística se acopla de forma aceptable a los datos. Aun considerando esto se presenta un comparativo entre el modelo original y los posibles ajustes retirando cada una de las variables:

Variables / Pruebas	Original	Destino	Género	Modalidad de pago	Monto	Tasa de interés	Saldo	Garantía liquida
Curva ROC	0.9877	0.9749	0.9868	0.9828	0.9219	0.8992	0.9038	0.9852
Cox-Snell	0.6208	0.5662	0.6180	0.6030	0.4509	0.3816	0.3920	0.6108
McFadden	0.7907	0.6809	0.7846	0.7530	0.4886	0.3918	0.4056	0.7692
Hosmer-Lemeshow	0.3718	0.02294	0.4372	0.4257	0.0836	6.911e-8	0.7451	0.4270

El modelo original tiene las mejores estimaciones excepto en el contraste Hosmer-Lemeshow, en donde se tiene una diferencia en el p-value aproximada del 5 % al 6& con respecto a los modelos retirando las variables *genero*, *modalidad de pago* y *garantía liquida*. En el caso de las variable *saldo*, la que presenta la máxima diferencia en el p-value, es preferible conservarla dando prioridad a la experiencia del ramo de crédito, además de ser la variable continua que tiene mayor impacto en el calculo de la probabilidad de incumplimiento. Con respecto a las demás variables se decide conservarlas aceptando esta diferencia, considerando que se tiene una mejor estimación en el resto de las pruebas.

Aclarando estos resultados es posible proceder al análisis de residuos, medidas de influencia y pruebas de colinealidad. Estas pruebas son necesarias para validar el modelo con respecto a la falta de ajuste a nivel observación y como afecta al modelo en términos generales.

El análisis de los residuos tiene el objetivo de evaluar la capacidad del modelo y detectar los valores irregulares e influyentes. Los diferentes tipos de residuos que se obtienen en una regresión logística son los siguientes:

1. Respuesta: Son obtenidos al hacer la diferencia entre el valor observado y el estimado del modelo, es decir, $y_r - \hat{\mu}_r$.

2. Pearson: Son los i-ésimos componentes del estadístico ji-cuadrada de Pearson de bondad de ajuste con la siguiente expresión:

$$\varepsilon_{Pi} = \frac{y_i - n_i\hat{p}_i}{\sqrt{n_i\hat{p}_i(1-\hat{p}_i)}}. \tag{5.10}$$

   Son considerados significativos los residuos que superan el valor absoluto 2 (Cañadas Roche).

3. Pearson Estandarizados: Modificación de los residuos de Pearson de forma que su distribución asintótica es normal estándar. La expresión es la siguiente:

$$\varepsilon_{PSi} = \frac{\varepsilon_{Pi}}{(1-h_i)^{1/2}}. \tag{5.11}$$

   En donde el término h_i es el i-ésimo elemento de la matriz siguiente:

$$H = W^{1/2}X(X^tWX)^{-1}X^tW^{1/2} \qquad\qquad W = Diag[n_i\hat{p}_i(1-\hat{p}_i)]. \tag{5.12}$$

4. Devianza: Son los residuos provenientes de la siguiente expresión:

$$\varepsilon_{Di} = sign(y_i - \hat{\mu}_i)(2[y_i log\frac{y_i}{\hat{\mu}_i} + (n_i - y_i)log\frac{n_i - y_i}{n_i - \hat{\mu}_i}]). \tag{5.13}$$

   Donde:

   - *sign*: función signo que extrae el signo de una función real.

5. Devianza estandarizada: Son los residuos de devianza estandarizados dados por la siguiente expresión:

$$\varepsilon_{DSi} = \frac{\varepsilon_{Di}}{(1-h_i)^{1/2}}. \tag{5.14}$$

Los resultados de los residuos que podrían afectar al modelo son los siguientes:

```
#Residuos de Pearson


res_pearson_sig
FALSE  TRUE
984    16
```

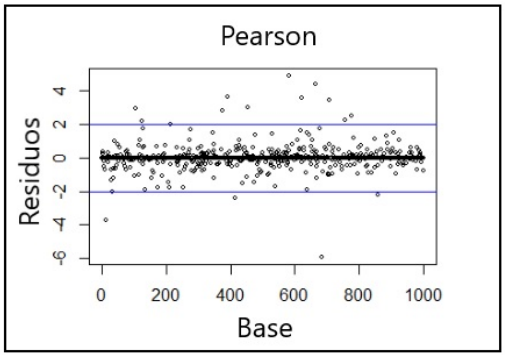


Figura 5.4: Residuos de Pearson

```
#Residuos estandarizados de Pearson


res_stand_sign
FALSE  TRUE
983    17
```

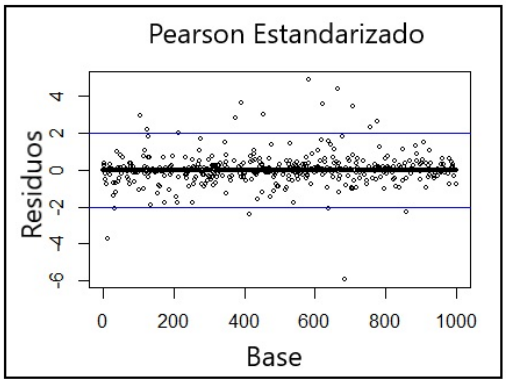


Figura 5.5: Residuos de Estandarizados Pearson

```
#Residuos de Devianza
```

```
res_dev_sign
FALSE  TRUE
990    10
```

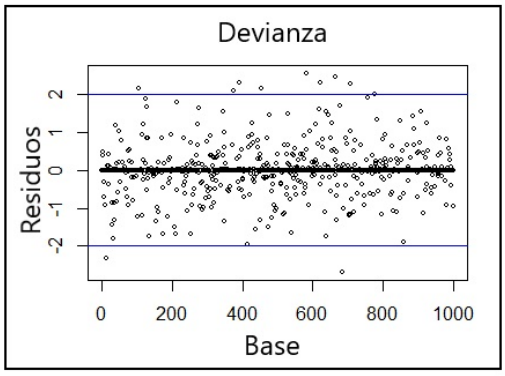


Figura 5.6: Residuos de Devianza

```
#Residuos estandarizados de Devianza


res_dev_sign_estan
FALSE  TRUE
989    11
```

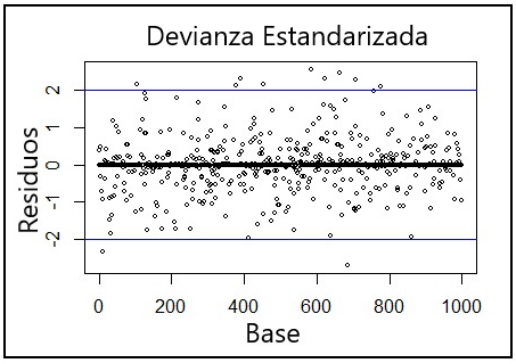


Figura 5.7: Residuos estandarizados de Devianza

El propósito de estas pruebas es conocer que tan homogénea es la base, ya que los valores extremos producen una falta de ajuste en los estimadores. Si una base de datos llegara a tener un alto porcentaje de residuos, seria necesario retirar estas observaciones considerando el ajuste del modelo como prioridad.

Las pruebas anteriores reflejan que los residuos en el modelo no superan el 1.8 % del total de la base simulada, por lo cual se puede asumir que no se presentan problemas significativos de ajuste hasta este momento. La siguiente prueba es la medida de influencia, la cual, detecta las

observaciones que separadas del resto pueden influir en los estimadores del modelo. La prueba de influencia analiza el efecto que estas observaciones tienen en los parámetros del modelo. La prueba de influencia se calcula de la siguiente manera:

$$D_i = \frac{\varepsilon^2_{PSi}}{k+1} \times \frac{h_{ii}}{1-h_{ii}}. \tag{5.15}$$

En donde el termino h_{ii} es el apalancamiento de la i-ésima observación. Los resultados de la aplicación de esta prueba en modelo son los siguientes:

```
> distancia_cook<-cooks.distance(Regre_Sim)
> table(distancia_cook>1)
```

```
FALSE
1000
```

El modelo no presenta problemas de influencia, ya que ninguna observación influye de forma negativa en los estimadores. A continuación, se procede a realizar la prueba de colinealidad al modelo. El objetivo de medir la relación lineal entre los coeficientes es ver la varianza que se produce, ya que, si se tiene una relación fuerte entre los predictores la precisión de los coeficientes estimados baja. La varianza se calcula de la siguiente manera:

$$\hat{Var}(\beta_i) = \frac{\hat{\sigma^2}}{(n-1)s_i^2} \times \frac{1}{1-R_i^2}. \tag{5.16}$$

En donde:
$\hat{\sigma^2}$: Estimación de la varianza de los errores.
s_i^2: Varianza muestral del predictor X_i.
R_i^2:Coeficiente de correlación múltiple de la regresión de X_i sobre los otros predictores.

El término VIF (factor de inflación de la varianza) está determinado por el termino de división derecho de la ecuación (5.16). Con regularidad este término es el que determina la colinealidad de alguna variable en una regresión lineal, sin embargo:

"Tal y como se describe en (Fox, Weisberg, 2011) VIF no es adecuado para modelos en los que haya variables categóricas como predictoras. (Fox, Monnette, 1992) generalizan la noción de inflación de la varianza asociada al incremento de la región de confianza de un conjunto de regresores. Obtuvieron una medida que llamaron *factor de inflación de la varianza generalizado* (GVIF). Si tenemos una variable predictora que tiene p regresores, el valor de $GVIF^{1/2p}$ es una medida de cómo disminuye la precisión de la estimación de los coeficientes debido a la existencia de colinealidad. Al igual que con el factor de inflación de la varianza, un valor próximo a 1 de $GVIF^{1/2p}$ indicaría la ausencia de colinealidad" (Cañadas Reche, 2013)

El término GVIF está definido de la siguiente manera:

$$GVIF = \frac{det(R_{11})det(R_{22})}{det(R)} \tag{5.17}$$

En donde:
R_{11}: Matriz de correlaciones entre el conjunto de regresores que se cree tienen algún problema.
R_{22}: Matriz de correlaciones entre los otros regresores del modelo.
R: Matriz de correlaciones entre todos los regresores del modelo.

El resultado del modelo con respecto a la prueba de colinealidad es el siguiente:

```
GVIF Df    GVIF^(1/(2*Df))
Genero    1.125554  1        1.060921
Destino   1.157912  2        1.037335
Modalidad 1.498392  2        1.106385
Monto     3.636815  1        1.907044
Intereses 6.836504  2        1.616994
Saldo     9.224860  1        3.037246
Garantia  1.291837  1        1.136590
```

Las variables no presentan evidencias de colinealidad fuerte en términos generales. *Saldo* es la variable que tiene mayor índice de colinealidad, sin embargo, se encuentra en los límites aceptables, por lo cual no hay motivos para sacarla del estudio. Además, se observó anteriormente que la mejora al retirar esta variable es mínimo con respecto a las pruebas de ajuste y poder predictivo.

Finalizando el proceso de pruebas, se presenta la validación cruzada con el objetivo de medir si la precisión de los estimadores puede ser útil en el ajuste de otros datos diferentes. Para ello la validación divide en dos la muestra para medir en la primera parte el ajuste del modelo y en la segunda evalúa el ajuste, dado que se tienen observaciones suficientes. La forma en la que se integra esta prueba es:

"En la primera, la llamada *K-Fold cross-validation*, se divide la muestra en k submuestras, de forma que se utilizan k-1 para estimar el modelo y la restante como submuestra de evaluación, este proceso se repite k veces, de forma que cada submuestra es utilizada una vez para evaluar el modelo y k-1 veces para el ajuste. Como medida de validación se suele utilizar la media de las tasas de clasificaciones correctas, o su complementario que sería la tasa clasificaciones incorrectas. " (Cañadas Reche, 2013)

A continuación, se presenta un esquema con el procedimiento de la validación cruzada con el objetivo de aportar una mayor comprensión con 4 submuestras :

La validación cruzada se encuentra incorporada en el paquete "DAAG" (Maindonal; Braun,

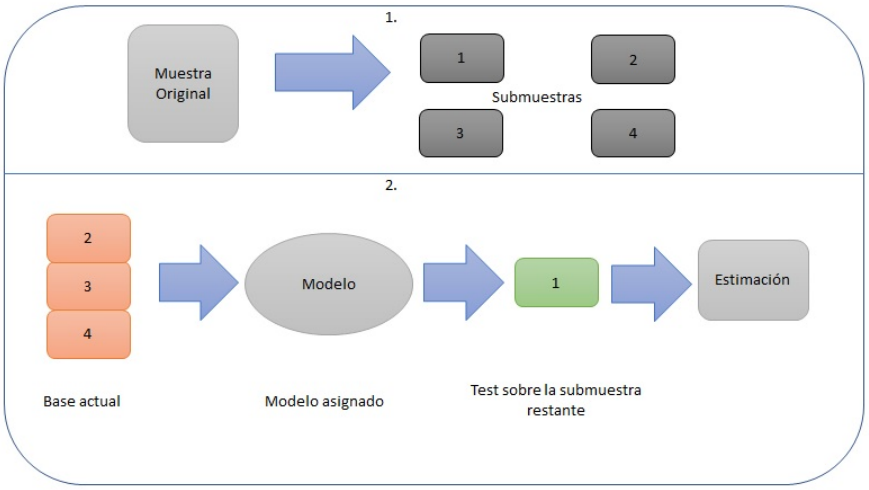


Figura 5.8: Validación Cruzada

2013) mediante la función CVbinary, en donde se obtienen 10 submuestras. Los resultados de aplicar dicho procedimiento:

```
> cruzada<-CVbinary(Regre_Sim)

Fold:  4 5 1 10 9 2 3 7 8 6
Internal estimate of accuracy = 0.942
Cross-validation estimate of accuracy = 0.938
```

Con esta validación tenemos una alta probabilidad de que el modelo presentado pueda aplicarse a otras bases de datos con características similares, ya que el modelo clasifica correctamente al 94 % de los individuos ausentes del ajuste del modelo. En el ejemplo realizado por Cañada Reche se considera aceptable una tasa mayor al 85 %, lo cual valida la probabilidad obtenida en el modelo presentado.

Dado que el modelo no presenta problemas con la selección de variables, precisión en el ajuste, pruebas de residuos, medidas de influencia y pruebas de colinealidad, es posible decir que la regresión aplicada a la base simulada cuenta con los elementos necesarios para tener la capacidad de calcular la probabilidad de incumplimiento de cada individuo.

Considerando lo anterior, se tienen las herramientas suficientes para presentar un análisis sobre las características y variables del modelo. El primer resultado es un análisis del impacto de las variables independientes, para ello es necesario aplicar la función exponencial en los coeficientes del modelo e identificar las variables que nos dan mayor y menos poder al momento de evaluar la probabilidad de incumplimiento. Este análisis diferencia entre variables continuas y categóricas, ya que no cuentan con el mismo peso estadístico.

Los coeficientes obtenidos en la regresión logística se encuentran en la siguiente tabla:

Respecto a la tabla de los coeficientes identificamos que la variable categórica *tasa de interés* es la que tiene mayor poder al momento de evaluar la probabilidad de incumplimiento en este

Variables	Coeficientes	Exp(Coeficientes)
Genero2	-0.8958634	0.408255
Destino2	1.833018	6.252727
Destino3	-28.03809	6.655963e-13
Modalidad2	.8062379	2.239467
Modalidad3	-3.455671	0.03156611
Monto	-0.0003901164	0.99961
Interes2	18.8242	149708400
Interes3	20.65668	935579400
Saldo	0.0003238663	1.000324
Garantía	4.811992e-0.5	1.000048

tipo de variables, sobre todo con tasas mayores al 20 %. En contraste la variable *destino* tiene el efecto menos relevante al momento de la evaluación, seguida por *género* y *modalidad*. Se puede decir que es más probable que se produzcan casos positivos con una tasa mayor al 20 % (20.65668) que con tasas menores a este porcentaje (18.8242), por otra parte, es menor probable que se produzcan casos positivos en destinos diferentes (-28.03809) a consumo y personales. Es necesario aclarar que para realizar este tipo de comparaciones las demás variables quedan fijas. Con la ayuda de este análisis es posible definir qué términos definen en mayor y menor medida el incumplimiento en esta base simulada, dado que:

"Al interpretar los coeficientes de las variables, es imprescindible tener en cuenta cómo se ha definido la variable de respuesta: un coeficiente con signo positivo indica que P(Y=1) crece cuando lo hace la variable, pero el sentido cualitativo de este hecho depende, desde luego, de lo que representen tanto la variable en cuestión como el suceso $Y = 1$." (Silva Aycaguer; Barroso Utra, 2004)

En el caso de las variables continuas, se tiene que *saldo* es la variable con mayor impacto (0.0003238663) al momento de evaluar la probabilidad, seguida por *garantía* (0.00004811992) y *monto* (-0.0003901164).

El segundo resultado significativo es encontrar las características de los individuos que presentan la mayor y menor probabilidad de incumplimiento, lo cual nos puede indicar que condiciones se recomienda analizar al momento de otorgar un crédito. Los resultados son los siguientes:

Variables	Máximo	Mínimo
Genero2	1	0
Destino2	1	0
Destino3	0	0
Modalidad2	0	0
Modalidad3	1	0
Monto	23,382.51	293,807.7
Interes2	0	0
Interes3	0	1
Saldo	384,989	5,111.867
Garantia	808.37773	1,145.044
PI	1	1.980486e-49

Cabe aclarar que las interpretación de las variables continuas se ve afectada por el proceso de simulación, que si bien tiene un grado de precisión aceptable, no remplaza la calidad del modelo en caso de ser aplicado en una base auténtica. La descripción de estos individuos es la siguiente:

Variable	Máximo	Mínimo
Género	Masculino	Femenino
Destino	Consumo	Personales
Modalidad	Pagos periódicos	Pagos únicos
Monto	23,382.51	293,807.7
Interés	0-10	Mayor al 20 %
Saldo	384,989	5,111.867
Garantía	808.3773	1,145.044

Para terminar este capítulo se presenta un árbol de clasificación con el objetivo de complementar los posibles perfiles antes presentados. El árbol de clasificación es una estructura de datos jerárquica que segmenta el espacio de la variable predictora en varias regiones simples. Dichas regiones llevan el nombre de hojas o nodos terminales. Los componentes de un árbol son:

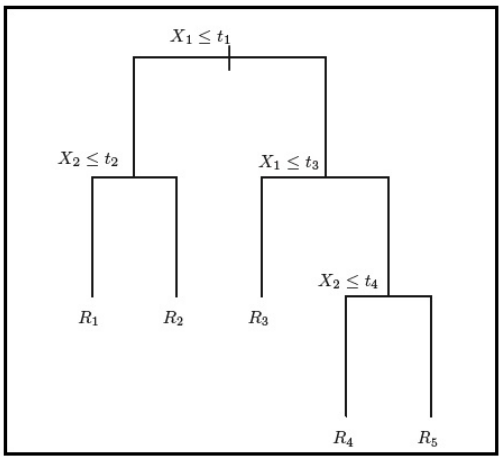


Figura 5.9: Árbol de clasificación (Tibshirani, 2013)

Donde:

- R_i: nodos terminales.
- $X_i \leq t_i$: nodos internos.
- X_i: variable i.

La metodología para construir un árbol de clasificación es la siguiente:

1. Se divide el espacio de predicción, es decir el conjunto de valores posibles $X_1, ..., X_k$ en J regiones distintas y no sobrepuestas $(R_1, ..., R_J)$.

2. Para cada observación en la región R_j, se realiza una predicción basándose en la clase (categoría) más común de observaciones en R_j.

Las regiones se dividen en recuadros de gran dimensión con el propósito simplificar la interpretación de los arboles. El diagrama siguiente muestra un ejemplo de estos recuadros:

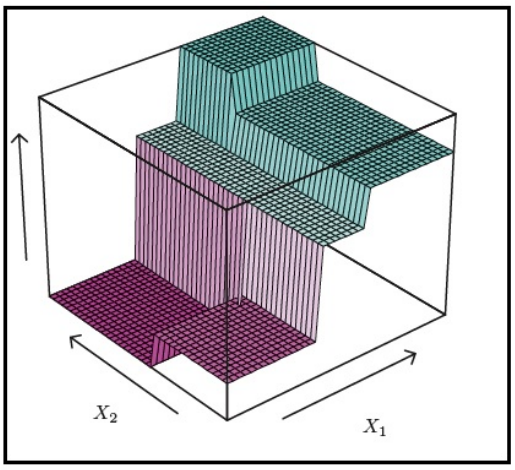


Figura 5.10: División de regiones (Tibshirani, 2013)

La proporción de las observaciones determina el resultado de región, ya que, si se tiene un 80 % de individuos que incumplen en una región, se determina el término 0. Formalmente, las regiones se definen de la siguiente manera:

$$R_1(j, s) = \{X \quad | \quad X_j < s\}. \tag{5.18}$$

Los árboles dependen de calcular las proporciones de forma precisa, por ello existen indicadores que determinan la fracción de observaciones que no pertenecen a la clase más común. Uno de los más utilizados en la práctica es el índice de Gini, ya que mide la sensibilidad del nodo. La expresión de este índice es la siguiente:

$$G = \sum_{k=1}^{K} \hat{p}_{mk}(1 - \hat{p}_{mk}). \tag{5.19}$$

Donde:

- $\hat{p}_{mk}$:proporción de observaciones en la región m que son de la clase k.

El índice se define formalmente como:

"El índice de Gini es una medida de la variación en las k clases. No es difícil ver que el índice de Gini toma un valor pequeño si el termino $\hat{p}_{mk}$ está cerca de cero o uno. Por esta razón, se hace referencia al índice de Gini como una medida de pureza del nodo, es decir, un pequeño

valor que indica si un nodo contiene observaciones provenientes de una sola clase. ” (Tibshirani, 2013)

La selección del árbol de clasificación con las ramificaciones correctas se determina mediante validación cruzada. Aplicando esta prueba en los datos del modelo se obtuvo que el árbol predice correctamente al 88 % de las observaciones en la submuestra de ajuste. En el anexo se encuentran las pruebas y el código correspondiente a la construcción del árbol presentado.

El gráfico del árbol de clasificación es el siguiente:

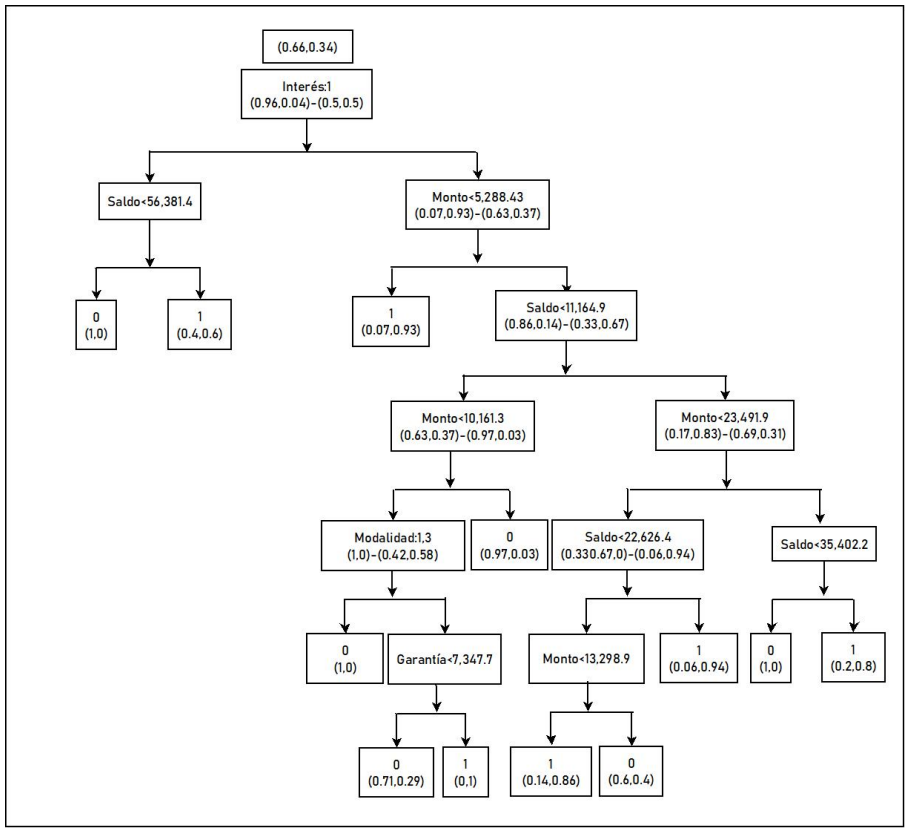


Figura 5.11: Árbol de clasificación

El árbol muestra los perfiles posibles y si cumpliendo con cierta características dicho perfil incumple o no con su responsabilidad de pago. Los nodos internos se interpretan de izquierda a derecha, siendo el cumplimiento de la condición el lado izquierdo y el incumplimiento el lado derecho. Por ejemplo, el nodo [Interés:1] indica en caso positivo una probabilidad cumplimiento alta (0.96) y en caso contrario [Interés:2,3] una proporción equilibrada en la probabilidad (0.5,0.5) entre cumplimiento e incumplimiento. Siguiendo [Interés:1] se tiene el ultimo nodo de esta rama, [Saldo $<$ 56,381.4] interpreta en caso positivo las observaciones que tienen [Interés:1] y [Saldo $<$ 56,381.4] con una probabilidad de cumplimiento del 100 %, sin embargo, en caso [Interés:1] y [Saldo $>$ 56,381.4] una probabilidad predominante de incumplimiento (60 %). Los paréntesis debajo de casa nodo interno expresan las proporciones de cumplimiento e incumplimiento respectivamente.

Los perfiles que indican incumplimiento (nodo=1) son:

Perfil 1	Perfil 2	Perfil 3	Perfil 4	Perfil 5	Perfil 6
[Interés:1]	[Interés:2,3]	[Interés:2,3]	[Interés:2,3]	[Interés:2,3]	[Interés:2,3]
[56,381<Saldo]	[Monto<5,288.43]	[5,288.43<Monto<10161.3]	[23,491.9<Monto]	[5,288.43<Monto<23,491.9]	[5,288.43<Monto<13,298.9]
		[Modalidad:2]	[35,402.2<Saldo]	[Saldo>11,164.9]	[11,164.9<Saldo<22,626.4]
		[7,347<Garantía]			
		[Saldo<11,164.9]			

Con esto se llega al final del estudio contando con un modelo que mide la probabilidad de incumplimiento de forma satisfactoria y que tiene la capacidad de ser aplicado en otras bases de datos. Además, se tienen las variables con mayor y menor impacto en la probabilidad de incumplimiento con respecto a sí son categóricas o continuas, encontrando en ello las áreas en donde se debe de tener mayor atención al otorgar un crédito. Se indica el perfil de los individuos que tienen la probabilidad de incumplimiento más alta y más baja, además de los posibles perfiles en el árbol de clasificación.

Conclusiones

El análisis presentado en esta tesis corresponde a un estudio realizado en una base de datos simulada procedente de una cartera de clientes de cierto banco en México con una selección de variables y características específicas. El objetivo principal es modelar la probabilidad de incumplimiento con una regresión logística y analizar los resultados. En este análisis se muestran las variables categóricas y continuas que tienen mayor y menor impacto en la probabilidad de incumplimiento y los posibles perfiles de incumplimiento con ayuda del árbol de clasificación.

La selección de variables consideró de forma preferencial la experiencia del sector y aspectos que fueron considerados factibles al momento de la simulación durante el estudio presentado, tomando en cuenta la posibilidad de realizar modificaciones futuras sobre la selección y características de las variables, implicando resultados diferentes.

La simulación realizada fue creada a partir del criterio de la muestra que tuvo el mejor p-value sobre 25,000 muestras generadas en el programa estadístico R-project. La posibilidad de encontrar una muestra con características más cercanas a la cartera original es viable, utilizando métodos de simulación avanzados que sobrepasan los alcances del presente trabajo, sin embargo, en esta ocasión se presenta esta muestra como una posibilidad dentro de la amplia gama de estrategias y criterios. Al procesar la base fue necesario excluir toda información sensible sobre los clientes de la cartera. El estudio de variables bi-modales superan los límites de la presente tesis, por lo cual se integró la variable *monto* sin hacer un tratamiento especial a la segunda moda y modificando de variable continua a discreta *tasa de interés*, dando mayor peso a su importancia con respecto al criterio de actuarios integrantes de bancos nacionales con experiencia en este tipo específico de temas.

Es necesario reiterar la posibilidad de realizar otros modelos con variables diferentes a las presentadas con métodos de simulación y criterios distintos que pueden modificar los resultados presentados, por lo cual, esta tesis se centra en el estudio de esta base de datos simulada con criterios de p-value considerando la permanencia de ciertas variables con deficiencias en precisión dada su importancia con respecto a la experiencia del sector. También se considera la opción de generar modelos con variables dependientes diferentes, como podría ser un modelo estudiando el destino específico de los recursos otorgados o sobre las capas de la tasa de interés del préstamo. Los resultados presentados en esta tesis fueron obtenidos de un análisis adecuado sobre una cartera de clientes suficiente, sin embargo, no representan la estimación de la probabilidad de incumplimiento en términos generales en otras carteras de clientes diferentes, ya que este estudio solo representa una porción de una gran gama de resultados.

Teniendo en consideración lo anterior, la presente tesis calcula la probabilidad de incumplimiento sobre una base de datos simulada de tamaño 1,000 aplicando una regresión logística

sobre las variables *género*, *tasa de interés*, *modalidad de pago*, *destino del crédito*, *saldo insoluto*, *garantía líquida*, *monto original* y *incumplimiento*, siendo la ultima la variable dependiente del estudio. Aplicando la regresión se realizaron pruebas de hipótesis (Wald y razón de verosimilitud) considerando la posible salida de alguna variable, sin encontrar suficiente evidencia para dicha modificación. Se realizaron pruebas de bondad de ajuste (Hosmer-Lemenshow, pseudo R_2 de McFadden) y la medición del poder predictivo mediante las curvas ROC midiendo la correcta clasificación de los individuos que cumplieron e incumplieron encontrando un margen aceptable de precisión sobre la muestra, así como resultados favorables en el análisis de residuos, medidas de influencia, pruebas de colinealidad y validación cruzada.

Realizando este proceso se encontró el perfil con la probabilidad de incumplimiento más baja y alta aplicando el punto de corte optimo correspondiente a la curva ROC. Además, se anexo un árbol de clasificación con posibles perfiles de cumplimiento e incumplimiento. Se concluye la tesis con el análisis de impacto en la variable dependiente con respecto a las demás siendo *tasa de interés* con tasas superiores al 20 % y *saldo* las variables categórica y continua respectivamente con mayor impacto al calcular la probabilidad de incumplimiento. Con ello se da por terminado este trabajo sin omitir el gran agradecimiento a los profesores de la Facultad de Ciencias quienes dieron el apoyo y conocimiento necesarios para concluir este trabajo.

Anexo

---

title: “Simulacion” author: “Luis Andres EscareÃ±o” date: “4 de febrero de 2018” output: pdf_document: default html_document: df_print: paged —

Proceso de Simulación.

Paqueterias:

```
require(goftest)
```

```
## Loading required package: goftest
```

```
require(MASS)
```

```
## Loading required package: MASS
```

```
require(lmom)
```

```
## Loading required package: lmom
```

```
require(fitdistrplus)
```

```
## Loading required package: fitdistrplus
## Loading required package: survival
```

```
require(foreign)
```

```
## Loading required package: foreign
```

```
require(forecast)
```

```
## Loading required package: forecast
```

```
require(rJava)
```

```
## Loading required package: rJava
```

```
require(xlsx)
```

```
## Loading required package: xlsx
## Loading required package: xlsxjars
```

```
require(car)
```

```
## Loading required package: car
```

```
require(gplots)
```

```
## Loading required package: gplots
##
## Attaching package: 'gplots'
## The following object is masked from 'package:stats':
##
##     lowess
```

```
require(ROCR)
```

```
## Loading required package: ROCR
```

```
require(tree)
```

```
## Loading required package: tree
```

```
require(ISLR)
```

```
## Loading required package: ISLR
```

```
require(rpart)
```

```
## Loading required package: rpart
```

```
require(rpart.plot)
```

```
## Loading required package: rpart.plot
```

Extracción de la base.

```
Info<-read.csv("C:/Users/andre/Documents/Tesis/Simulacion/Revision/Info.csv",header = T)
attach(Info)
sample_size <- 1000
```

Se genera una muestra con el propósito de simular los datos de la base, ya que no es posible usar los datos reales por motivos de confidencialidad.

Pruebas de Ajuste de muestra:

```
sample_size <- 1000
```

En esta parte se hacen varias muestras con el propósito de ver cual distribución nos da el mejor p-value posible.

Saldo Insoluto:

```
Plnormal<-NULL
Plexp<-NULL
Plweibull<-NULL
a<-0
b<-0
c<-0

for(i in 1:1000)
{
  Muestra <- Info[sample(nrow(Info), sample_size, replace = FALSE, prob = NULL),]
  Saldo2<-Muestra$Saldo_Insoluto

  AjusteMuestra<-fitdistr(Saldo2,"lognormal")
  a<-ad.test(Saldo2,"plnorm",meanlog=AjusteMuestra$estimate[1],sdlog=AjusteMuestra$estimate[2])
  Plnormal[i]<-a$p.value
  AjusteMuestra1<-fitdistr(Saldo2,"exponential")
  b<-ad.test(Saldo2,"pexp",AjusteMuestra1$estimate)
  Plexp[i]<-b$p.value
  #AjusteMuestraw<-fitdistr(Saldo2,"weibull", lower=c(0,0))
  #c<-ad.test(Saldo2,"pweibull",shape=AjusteMuestraw$estimate[1],scale=AjusteMuestraw$estimate[2])
  #Plweibull[i]<-c$p.value
}

max(Plnormal)
```

```
## [1] 0.777328
```

```r
max(Plexp)
```

```
## [1] 0.02609084
```

```r
max(Plweibull)
```

```
## Warning in max(Plweibull): ningun argumento finito para max; retornando -
## Inf
```

```
## [1] -Inf
```

max(Plweibull)

0.05971072

Dadas las evidencias escogemos la distribución Log-Normal.

Monto:

```r
Plnormal<-NULL
Plexp<-NULL
Plweibull<-NULL
a<-0
b<-0
c<-0


for(i in 1:1000)
{
  Muestra <- Info[sample(nrow(Info), sample_size, replace = FALSE, prob = NULL),]
  Montelus<-Muestra$Monto_Original

  AjusteMuestra<-fitdistr(Montelus,"lognormal")
  a<-ad.test(Montelus,"plnorm",meanlog=AjusteMuestra$estimate[1],sdlog=AjusteMuestra$estimate[2])
  Plnormal[i]<-a$p.value
  AjusteMuestra1<-fitdistr(Montelus,"exponential")
  b<-ad.test(Montelus,"pexp",AjusteMuestra1$estimate)
  Plexp[i]<-b$p.value
  #AjusteMuestraw<-fitdistr(Montelus,"weibull", lower=c(0,0))
  #c<-ad.test(Saldo2,"pweibull",shape=AjusteMuestraw$estimate[1],scale=AjusteMuestraw$estimate[2])
  #Plweibull[i]<-c$p.value
}

max(Plnormal)
```

```
## [1] 0.0461509
```

```r
max(Plexp)
```

```
## [1] 0.001718035
```

```r
max(Plweibull)
```

```
## Warning in max(Plweibull): ningun argumento finito para max; retornando -
## Inf
```

```
## [1] -Inf
```

max(Plweibull)

6e-07

Garantía:

```
Plnormal<-NULL
Plexp<-NULL
Plweibull<-NULL
a<-0
b<-0
c<-0


for(i in 1:1000)
{
  Muestra <- Info[sample(nrow(Info), sample_size, replace = FALSE, prob = NULL),]
  Galantis<-Muestra$Garantia_Liquida

  AjusteMuestra<-fitdistr(Galantis,"lognormal")
  a<-ad.test(Galantis,"plnorm",meanlog=AjusteMuestra$estimate[1],sdlog=AjusteMuestra$estimate[2])
  Plnormal[i]<-a$p.value
  AjusteMuestra1<-fitdistr(Galantis,"exponential")
  b<-ad.test(Galantis,"pexp",AjusteMuestra1$estimate)
  Plexp[i]<-b$p.value
  #AjusteMuestraw<-fitdistr(Galantis,"weibull")
  #c<-ad.test(Galantis,"pweibull",shape=AjusteMuestraw$estimate[1],scale=AjusteMuestraw$estimate[2])
  #Plweibull[i]<-c$p.value
}

max(Plnormal)
```

```
## [1] 0.4177683
```

```
max(Plexp)
```

```
## [1] 1.078453e-05
```

```
max(Plweibull)
```

```
## Warning in max(Plweibull): ningun argumento finito para max; retornando -
## Inf
```

```
## [1] -Inf
```

En el caso de la prueba Weibull se crean Nan´s y saturan el reporte. Por ello se muestra abajo los resultados ejecutando la prueba:

There were 50 or more warnings (use warnings() to see the first 50)

max(Plweibull)

0.02396893

Se decide tomar la distribución Log-normal al tener los p-value más altos en las tres variables.

```
a<-0
b<-0
c<-0
Plsaldo<-NULL
Plmaxsaldo<-0
Plmonto<-NULL
Plmaxmonto<-0
Plgara<-NULL
```

```
Plmaxgara<-0
indsal<-0
indmont<-0
indgara<-0
```

Se crea una secuencia para encontrar las muestras que nos dan más precisión en las variables.

```
gum<-216001

for(i in 1:3000)
{
set.seed(gum+i)
Muestra <- Info[sample(nrow(Info), sample_size, replace = FALSE, prob = NULL),]
Saldo2<-Muestra$Saldo_Insoluto
Monto<-Muestra$Monto_Original
Garan<-Muestra$Garantia_Liquida

AjusteMuestraM<-fitdistr(Monto,"lognormal")
a<-ad.test(Monto,"plnorm",meanlog=AjusteMuestraM$estimate[1],sdlog=AjusteMuestraM$estimate[2])
Plmonto[i]<-a$p.value

AjusteMuestraG<-fitdistr(Garan,"lognormal")
b<-ad.test(Garan,"plnorm",meanlog=AjusteMuestraG$estimate[1],sdlog=AjusteMuestraG$estimate[2])
Plgara[i]<-b$p.value

AjusteMuestraS<-fitdistr(Saldo2,"lognormal")
c<-ad.test(Saldo2,"plnorm",meanlog=AjusteMuestraS$estimate[1],sdlog=AjusteMuestraS$estimate[2])
Plsaldo[i]<-c$p.value


  if(i==1)
  {
    Plmaxmonto<-a$p.value
    Plmaxgara<-b$p.value
    Plmaxsaldo<-c$p.value
  }

  if(Plmaxmonto<Plmonto[i])
  {
    Plmaxmonto<-Plmonto[i]
    indmont<-i
  }

  if(Plmaxgara<Plgara[i])
  {
    Plmaxgara<-Plgara[i]
    indgara<-i
  }

  if(Plmaxsaldo<Plsaldo[i])
  {
    Plmaxsaldo<-Plsaldo[i]
    indsal<-i
  }
```

```
}


Plmaxgara
```

```
## [1] 0.6524036
```

```
indgara
```

```
## [1] 1743
```

```
Plmaxmonto
```

```
## [1] 0.05735863
```

```
indmont
```

```
## [1] 1130
```

```
Plmaxsaldo
```

```
## [1] 0.8263334
```

```
indsal
```

```
## [1] 2995
```

Saldo Insoluto.

Con la muestra que nos genera mejor precisión podemos proceder a hacer las pruebas de comprobación.

```
set.seed(gum+indsal)
Muestra <- Info[sample(nrow(Info), sample_size, replace = FALSE, prob = NULL),]

Saldo2<-Muestra$Saldo_Insoluto

AjusteMuestra<-fitdistr(Saldo2,"lognormal")
ad.test(Saldo2,"plnorm",meanlog=AjusteMuestra$estimate[1],sdlog=AjusteMuestra$estimate[2])
```

```
##
##  Anderson-Darling test of goodness-of-fit
##  Null hypothesis: log-normal distribution
##  with parameters meanlog = 9.11333926487895, sdlog =
##  1.11742639069346
##
## data:  Saldo2
## An = 0.42229, p-value = 0.8263
```

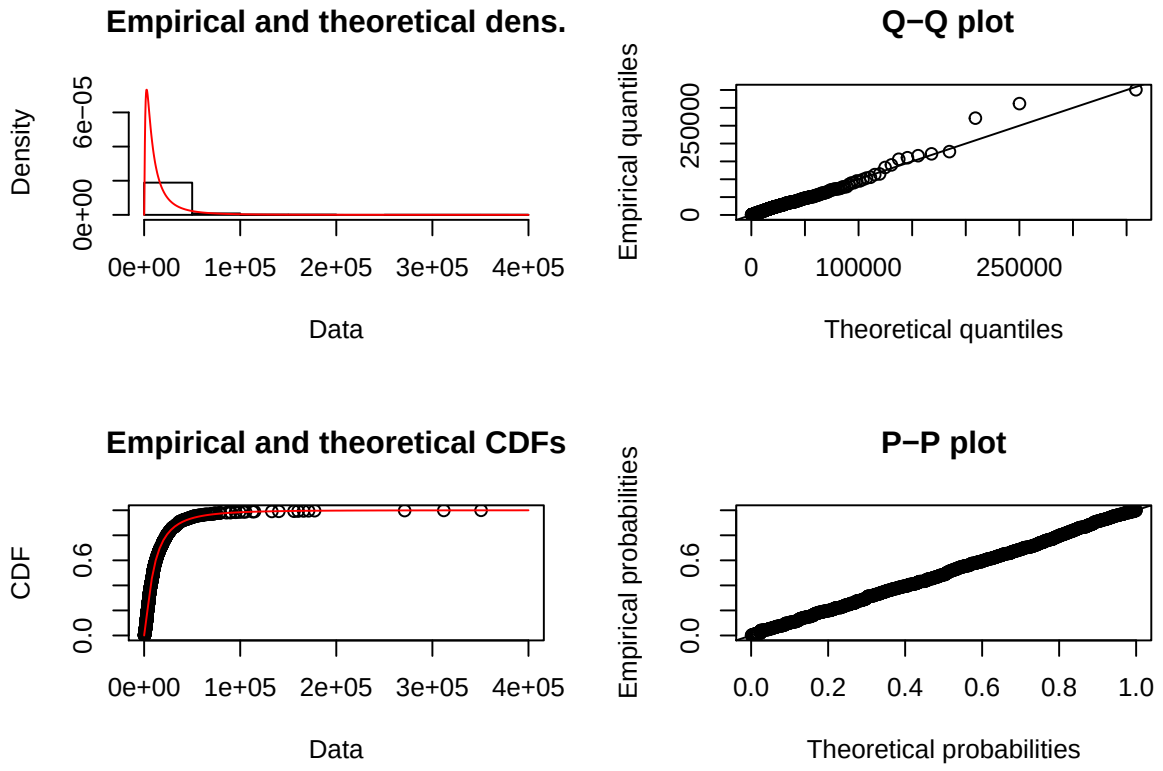
Graficamos:

```
name_variable <- "Saldo_Insoluto"
distribution_selected <- "lnorm"    ### or  "gamma"  or "geom"  or "binom" or "norm" or "lnorm"
distribution_fit <- fitdist(Saldo2, distribution_selected)
summary(distribution_fit)
```

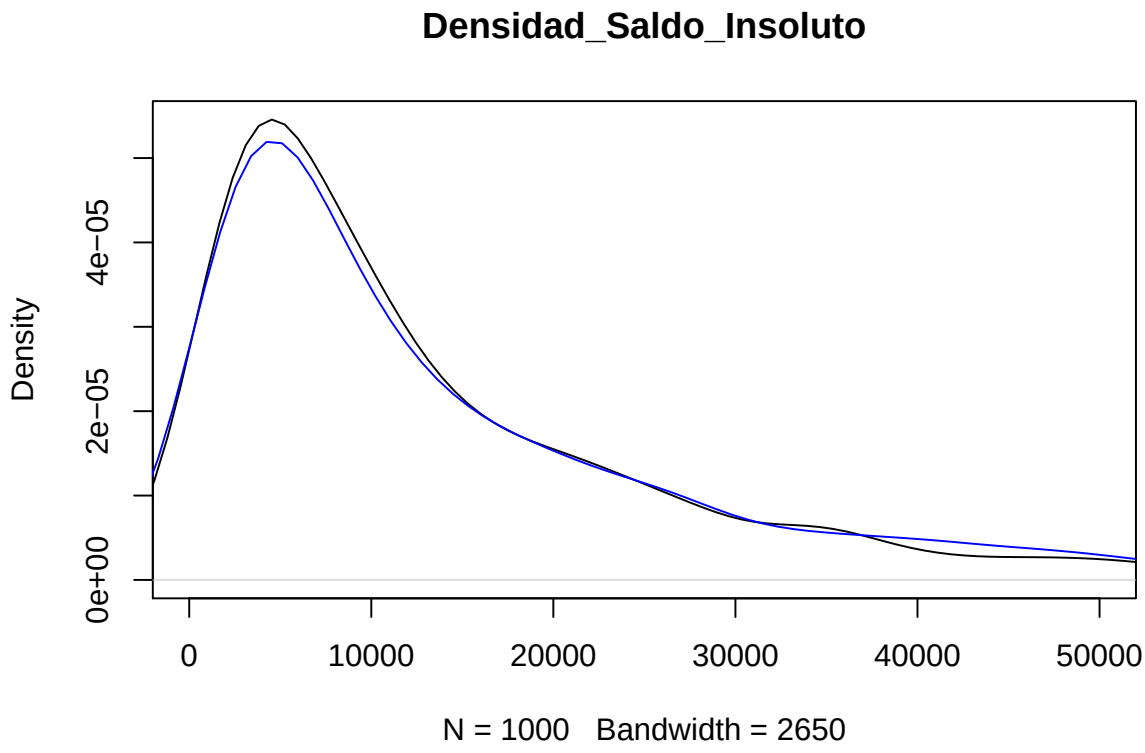
```
## Fitting of the distribution ' lnorm ' by maximum likelihood
## Parameters :
##         estimate Std. Error
## meanlog 9.113339 0.03533613
## sdlog   1.117426 0.02498632
## Loglikelihood:  -10643.31   AIC:  21290.61   BIC:  21300.43
```

```
## Correlation matrix:
##         meanlog sdlog
## meanlog       1     0
## sdlog         0     1
```

```
plot(distribution_fit)
```




```
plot(density(Saldo2),xlim=c(0,50000), main="Densidad_Saldo_Insoluto")
lines(density(rlnorm(sample_size ,meanlog=AjusteMuestra$estimate[1],sdlog=AjusteMuestra$estimate[2])),c
```



Generamos la variable simulada:

```
Simulacion_Saldo <-rlnorm(sample_size ,meanlog=AjusteMuestra$estimate[1],sdlog=AjusteMuestra$estimate[2
```

Procedemos a hacer el mismo proceso con las dos variables restantes.

Garantía liquida.

```
set.seed(gum+indgara)

Muestra <- Info[sample(nrow(Info), sample_size , replace = FALSE, prob = NULL),]

Garantia<-Muestra$Garantia_Liquida

AjusteMuestra<-fitdistr(Garantia,"lognormal")
ad.test(Garantia,"plnorm",meanlog=AjusteMuestra$estimate[1],sdlog=AjusteMuestra$estimate[2])
```

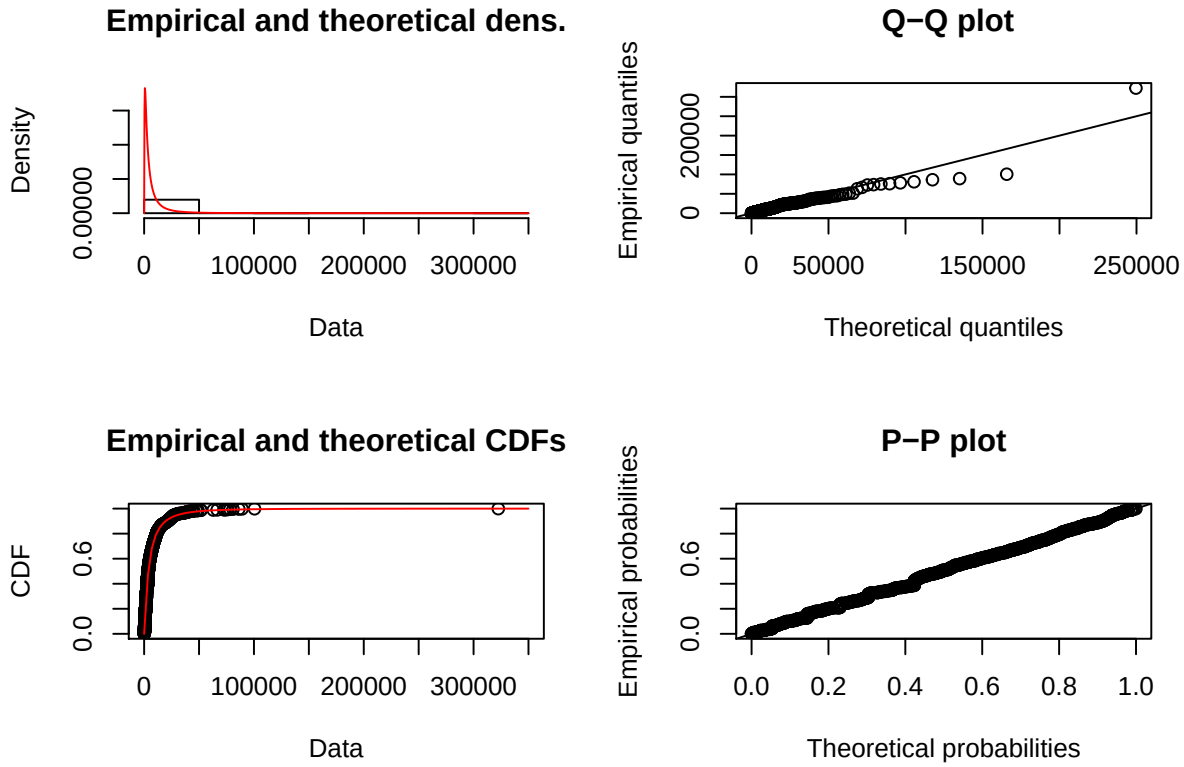
```
##
##  Anderson-Darling test of goodness-of-fit
##  Null hypothesis: log-normal distribution
##  with parameters meanlog = 8.25229419101794, sdlog =
##  1.268917059912
##
## data:  Garantia
## An = 0.59531, p-value = 0.6524
```

```
name_variable <- "Garantia_Liquida"
distribution_selected <- "lnorm"
distribution_fit <- fitdist(Garantia, distribution_selected)
```

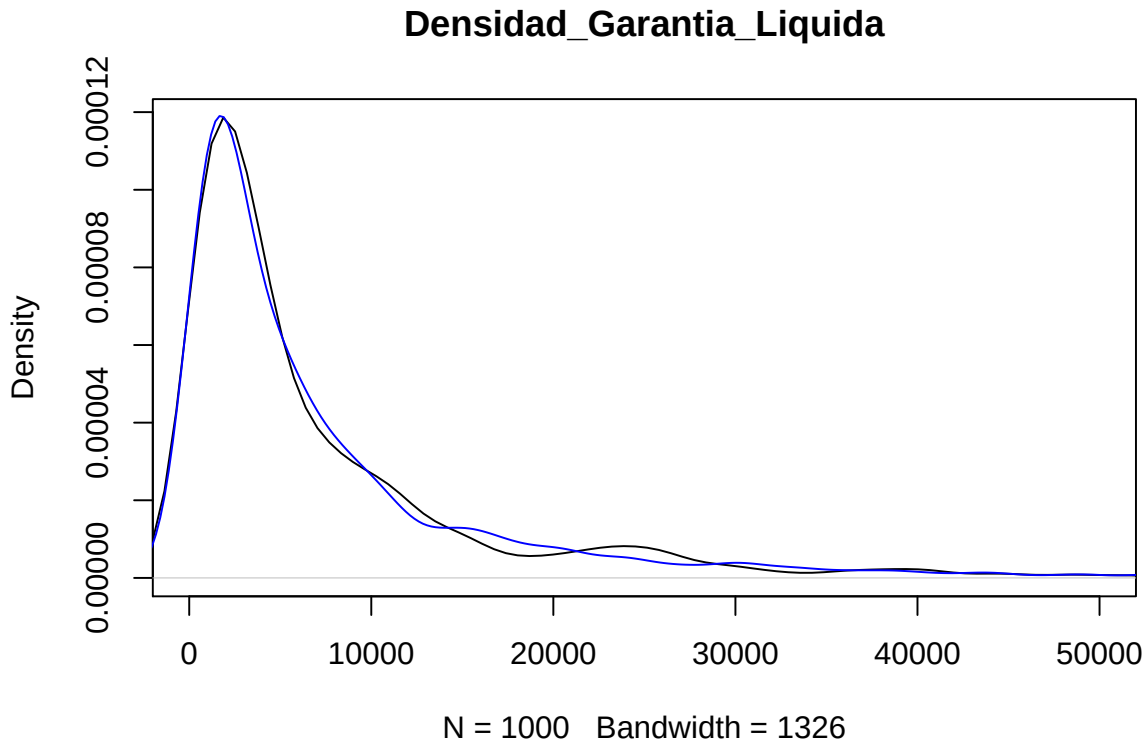
```r
summary(distribution_fit)
```

```
## Fitting of the distribution ' lnorm ' by maximum likelihood
## Parameters :
##         estimate Std. Error
## meanlog 8.252294 0.04012668
## sdlog   1.268917 0.02837377
## Loglikelihood:  -9909.397   AIC:  19822.79   BIC:  19832.61
## Correlation matrix:
##              meanlog        sdlog
## meanlog 1.000000e+00 2.588752e-10
## sdlog   2.588752e-10 1.000000e+00
```

```r
plot(distribution_fit)
```




```r
plot(density(Garantia),xlim=c(0,50000),main="Densidad_Garantia_Liquida")
lines(density(rlnorm(sample_size,meanlog=AjusteMuestra$estimate[1],sdlog=AjusteMuestra$estimate[2])),co
```




```
Simulacion_Garantia <-rlnorm(sample_size,meanlog=AjusteMuestra$estimate[1],sdlog=AjusteMuestra$estimate
```

Monto Original.

```
set.seed(gum+indmont)

Muestra <- Info[sample(nrow(Info), sample_size , replace = FALSE, prob = NULL),]

Monto<-Muestra$Monto_Original

AjusteMuestra<-fitdistr(Monto,"lognormal")
ad.test(Monto,"plnorm",meanlog=AjusteMuestra$estimate[1],sdlog=AjusteMuestra$estimate[2])
```

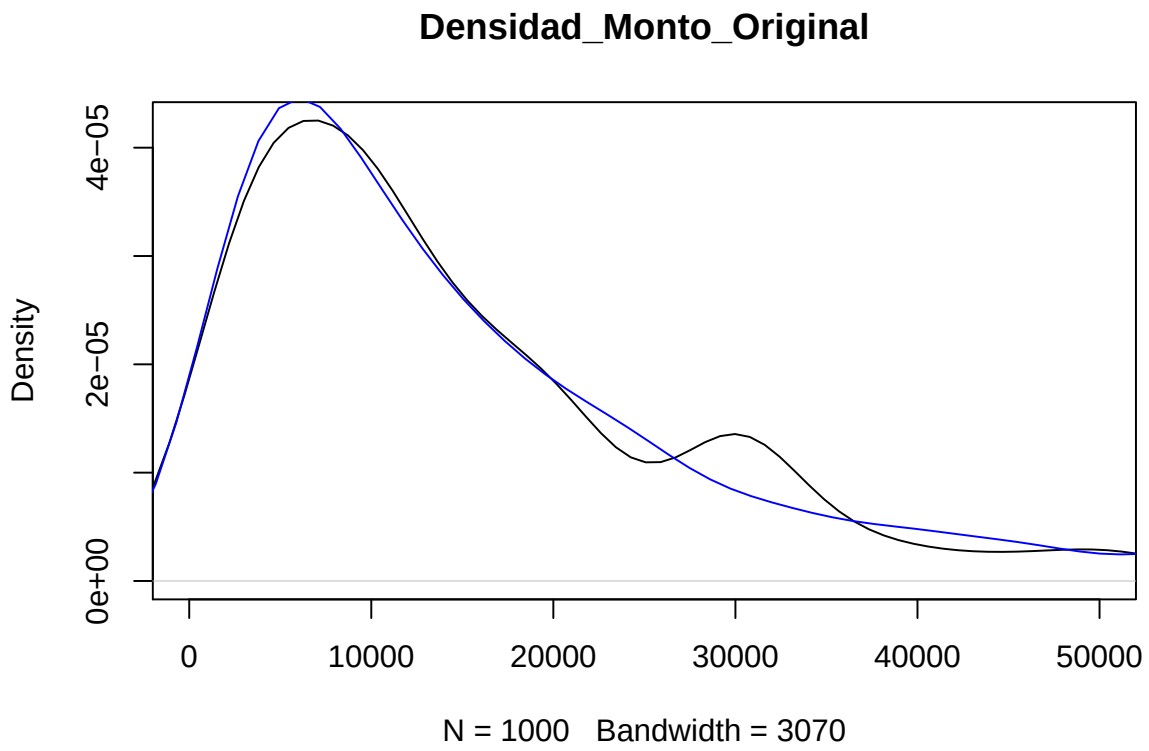
```
##
##  Anderson-Darling test of goodness-of-fit
##  Null hypothesis: log-normal distribution
##  with parameters meanlog = 9.37969373428342, sdlog =
##  1.04313369289156
##
## data:  Monto
## An = 2.3794, p-value = 0.05736
```

```
name_variable <- "Monto_Original"
distribution_selected <- "lnorm"
distribution_fit <- fitdist(Monto, distribution_selected)
summary(distribution_fit)
```

```
## Fitting of the distribution ' lnorm ' by maximum likelihood
```

```
## Parameters :
##         estimate Std. Error
## meanlog 9.379694 0.03298678
## sdlog   1.043134 0.02332508
## Loglikelihood:  -10840.86   AIC:  21685.72   BIC:  21695.54
## Correlation matrix:
##         meanlog sdlog
## meanlog       1     0
## sdlog         0     1
```

```
plot(density(Monto),xlim=c(0,50000),main="Densidad_Monto_Original")
lines(density(rlnorm(sample_size ,meanlog=AjusteMuestra$estimate[1],sdlog=AjusteMuestra$estimate[2])),c
```




```
Simulacion_Monto <-rlnorm(sample_size ,meanlog=AjusteMuestra$estimate[1],sdlog=AjusteMuestra$estimate[2
```

En el caso de la variable Monto no se puede conseguir un alto p-value dado que es una variable bimodal. Este tipo de variable sobrepasan los alcances de este trabajo.

Conversión de tasa de interés. Dada la imagen de evidencia de que la tasa de interés en los aspectos de este trabajo no es posible simularla de forma adecuada:

```
for(i in 1:nrow(Info))
{
  Dato<-Tasa_Interes[i]
  if(Dato<=10){
    Tasa_Interes[i]=1
  }
```

```
  if(Dato>10 & Dato<=20){
    Tasa_Interes[i]=2
  }

  if(Dato>20){
    Tasa_Interes[i]=3
  }
}
```

Categóricas.

En el caso de las variables categóricas se utilizó el método de inversión en el caso discreto.

Género.

Generamos las proporciones que tiene la variable y se procede a simular.

```
Prop<-table(Muestra$ï..Genero)

Genero<-c(Prop[1]/nrow(Muestra),Prop[2]/nrow(Muestra))
```

Número de simulaciones.

```
n<-nrow(Muestra)
```

Número de uniformes.

```
m<-nrow(Muestra)
```

Género

```
Simulacion_Genero<-NULL
u<-runif(m)
for(i in 1:n)
{
  j<-1
  r<-Genero[j]

  while(u[i]>=r)
  {
    j<-j+1
    r<-r+Genero[j]
  }
  Simulacion_Genero[i]=j
}
```

En el caso de las variables restantes se sigue la misma metodología.

Destino:

```
Prop2<-table(Muestra$Destino)

Destino<-c(Prop2[1]/nrow(Muestra),Prop2[2]/nrow(Muestra),Prop2[3]/nrow(Muestra))

Simulacion_Destino<-NULL
u<-runif(m)
for(i in 1:n)
{
  j<-1
  r<-Destino[j]
```

```
  while(u[i]>=r)
  {
    j<-j+1
    r<-r+Destino[j]
  }
  Simulacion_Destino[i]=j
}
```

Intereses:

```
for(i in 1:nrow(Muestra))
{
  Dato<-Muestra$Tasa_Interes[i]

  if(Dato<=10){
    Muestra$Tasa_Interes[i]=1
  }

  if(Dato>10 & Dato<=20){
    Muestra$Tasa_Interes[i]=2
  }

  if(Dato>20){
    Muestra$Tasa_Interes[i]=3
  }
}



Prop3<-table(Muestra$Tasa_Interes)

Interes<-c(Prop3[1]/nrow(Muestra),Prop3[2]/nrow(Muestra),Prop3[3]/nrow(Muestra))


Simulacion_Intereses<-NULL
u<-runif(m)
for(i in 1:n)
{
  j<-1
  r<-Interes[j]

  while(u[i]>=r)
  {
    j<-j+1
    r<-r+Interes[j]
  }
  Simulacion_Intereses[i]=j
}
```

Modalidad:

```
Prop5<-table(Muestra$Modalidad_Pago)

Modalidad<-c(Prop5[1]/nrow(Muestra),Prop5[2]/nrow(Muestra),Prop5[3]/nrow(Muestra))
```

```
Simulacion_Modalidad<-NULL
u<-runif(m)
for(i in 1:n)
{
  j<-1
  r<-Modalidad[j]

  while(u[i]>=r)
  {
    j<-j+1
    r<-r+Modalidad[j]
  }
  Simulacion_Modalidad[i]=j
}
```

En esta última fase observamos la proximidad que existe entre las variables categóricas originales y las simulaciones.

```
table(Simulacion_Genero)
```

```
## Simulacion_Genero
##   1   2
## 563 437
```

```
Prop
```

```
##
##   1   2
## 610 390
```

```
table(Simulacion_Destino)
```

```
## Simulacion_Destino
##   1   2   3
## 794 154  52
```

```
Prop2
```

```
##
##   1   2   3
## 803 148  49
```

```
table(Simulacion_Intereses)
```

```
## Simulacion_Intereses
##   1   2   3
## 368 196 436
```

```
Prop3
```

```
##
##   1   2   3
## 387 181 432
```

```
table(Simulacion_Modalidad)
```

```
## Simulacion_Modalidad
##   1   2   3
## 216 667 117
```

```
Prop5
```

```
##
##   1   2   3
## 221 674 105
```

Establecemos la base simulada:

```
Base_Simulada<-data.frame(Saldo=Simulacion_Saldo,Monto=Simulacion_Monto,Modalidad=Simulacion_Modalidad,

summary(Base_Simulada)
```

```
##      Saldo              Monto            Modalidad       Intereses
##  Min.   :   396.7   Min.   :   378.7   Min.   :1.000   Min.   :1.000
##  1st Qu.:  4355.5   1st Qu.:  5981.1   1st Qu.:2.000   1st Qu.:1.000
##  Median :  9274.1   Median : 11987.4   Median :2.000   Median :2.000
##  Mean   : 18287.8   Mean   : 20871.0   Mean   :1.901   Mean   :2.068
##  3rd Qu.: 21222.6   3rd Qu.: 24580.4   3rd Qu.:2.000   3rd Qu.:3.000
##  Max.   :384989.5   Max.   :293807.7   Max.   :3.000   Max.   :3.000
##     Destino          Genero          Garantia
##  Min.   :1.000   Min.   :1.000   Min.   :    66.4
##  1st Qu.:1.000   1st Qu.:1.000   1st Qu.:  1470.3
##  Median :1.000   Median :1.000   Median :  3423.6
##  Mean   :1.258   Mean   :1.437   Mean   :  8209.3
##  3rd Qu.:1.000   3rd Qu.:2.000   3rd Qu.:  8087.3
##  Max.   :3.000   Max.   :2.000   Max.   :366373.5
```

Formación de Y-

En el caso de la variable predictora es necesario usar un método especial para poder armar esta variable.

Comenzamos con colocar las variables categóricas como tipo factor, dado que este tipo es el correcto para estas variables.

```
Base <- Muestra

sapply(Base,class)
```

```
##                X        ï..Genero          Destino   Modalidad_Pago
##        "integer"        "integer"        "integer"        "integer"
##   Monto_Original     Tasa_Interes   Incumplimiento   Saldo_Insoluto
##        "integer"        "numeric"        "integer"        "integer"
## Garantia_Liquida
##        "integer"
```

```
Base$Destino<-factor(Base$Destino)
Base$ï..Genero<-factor(Base$ï..Genero)
Base$Modalidad_Pago<-factor(Base$Modalidad_Pago)
Base$Tasa_Interes<-factor(Base$Tasa_Interes)
Base$Incumplimiento<-factor(Base$Incumplimiento)

sapply(Base,class)
```

```
##                X        ï..Genero          Destino   Modalidad_Pago
##        "integer"         "factor"         "factor"         "factor"
##   Monto_Original     Tasa_Interes   Incumplimiento   Saldo_Insoluto
##        "integer"         "factor"         "factor"        "integer"
## Garantia_Liquida
```

```
##        "integer"
```

```
sapply(Base_Simulada,class)
```

```
##     Saldo     Monto Modalidad Intereses   Destino    Genero  Garantia
## "numeric" "numeric" "numeric" "numeric" "numeric" "numeric" "numeric"
```

```
Base_Simulada$Destino<-factor(Base_Simulada$Destino)
Base_Simulada$Genero<-factor(Base_Simulada$Genero)
Base_Simulada$Modalidad<-factor(Base_Simulada$Modalidad)
Base_Simulada$Intereses<-factor(Base_Simulada$Intereses)

sapply(Base_Simulada,class)
```

```
##     Saldo     Monto Modalidad Intereses   Destino    Genero  Garantia
## "numeric" "numeric"  "factor"  "factor"  "factor"  "factor" "numeric"
```

Aplicamos regresión logística sobre la base original para poder conseguir los coeficientes que se quieren conseguir en la simulación.

```
Regre<-glm(Incumplimiento~ï..Genero+Destino+Modalidad_Pago+Monto_Original+Tasa_Interes+Saldo_Insoluto+Ga
```

```
## Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred
```

```
summary(Regre)
```

```
##
## Call:
## glm(formula = Incumplimiento ~ ï..Genero + Destino + Modalidad_Pago +
##     Monto_Original + Tasa_Interes + Saldo_Insoluto + Garantia_Liquida,
##     family = binomial, data = Base)
##
## Deviance Residuals:
##      Min        1Q    Median        3Q       Max
## -2.01235  -0.38161  -0.00008  -0.00003   2.52545
##
## Coefficients:
##                     Estimate Std. Error z value Pr(>|z|)
## (Intercept)      -1.975e+01  8.283e+02  -0.024  0.98098
## ï..Genero2       -3.068e-01  2.625e-01  -1.169  0.24249
## Destino2          1.809e+00  3.262e-01   5.544 2.95e-08 ***
## Destino3         -1.810e+01  1.896e+03  -0.010  0.99238
## Modalidad_Pago2   2.058e-01  3.339e-01   0.616  0.53775
## Modalidad_Pago3  -3.426e+00  1.167e+00  -2.935  0.00334 **
## Monto_Original   -3.653e-04  4.904e-05  -7.448 9.47e-14 ***
## Tasa_Interes2     1.812e+01  8.283e+02   0.022  0.98254
## Tasa_Interes3     1.991e+01  8.283e+02   0.024  0.98082
## Saldo_Insoluto    3.227e-04  4.467e-05   7.224 5.06e-13 ***
## Garantia_Liquida  4.515e-05  3.489e-05   1.294  0.19568
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 684.63  on 999  degrees of freedom
## Residual deviance: 410.43  on 989  degrees of freedom
## AIC: 432.43
```

```
##
## Number of Fisher Scoring iterations: 19
```

```
Regre$coefficients
```

```
##      (Intercept)       ï..Genero2         Destino2         Destino3
##    -1.974911e+01    -3.068090e-01     1.808632e+00    -1.810478e+01
##  Modalidad_Pago2  Modalidad_Pago3   Monto_Original    Tasa_Interes2
##     2.057519e-01    -3.425721e+00    -3.652849e-04     1.812203e+01
##    Tasa_Interes3   Saldo_Insoluto Garantia_Liquida
##     1.991154e+01     3.226850e-04     4.514889e-05
```

Teniendo los coeficientes procedemos a evaluar el valor de n.

n =B+BX1+...+Bxn donde los B son los coeficientes originales y los x los valores simulados

```
for(i in 1:sample_size)
{
  x1=Base_Simulada$Genero[i]
  x2=Base_Simulada$Destino[i]
  x3=Base_Simulada$Modalidad[i]
  x4=Base_Simulada$Monto[i]
  x5=Base_Simulada$Intereses[i]
  x6=Base_Simulada$Saldo[i]
  x7=Base_Simulada$Garantia[i]

  r1=0
  r2=0
  r3=0
  r4=0
  r5=0
  r6=0
  r7=0
  r8=0
  r9=0
  r10=0

  if(x1==2){r1=1}
  if(x2==2){r2=1}
  if(x2==3){r3=1}
  if(x3==2){r4=1}
  if(x3==3){r5=1}
  if(x5==2){r7=1}
  if(x5==3){r8=1}

  r6=x4
  r9=x6
  r10=x7

  n[i]=Regre$coefficients[1]+Regre$coefficients[2]*r1+Regre$coefficients[3]*r2+Regre$coefficients[4]*r3

}
```

Formación de P. En este caso se utiliza los valores de n para poder formar la p correspondiente a cada individuo.

```
p<-NULL

for(i in 1:sample_size)
{
  p[i]=exp(n[i])/(1+exp(n[i]))
}
```

Formacion de Y

Teniendo los valores correspondientes de P procedemos a hacer una simulación con uniformes (0,1).

```
Incum<-NULL

for(i in 1:sample_size)
{
  if(p[i]>=runif(1)){
    Incum[i]=1
  } else{
    Incum[i]=0
  }
}
```

Integramos la variable dependiente a la base de variables simuladas:

```
Base_Simulada_Inc<-data.frame(Incumplimiento=Incum,Saldo=Simulacion_Saldo,Monto=Simulacion_Monto,Modali

sapply(Base_Simulada_Inc,class)
```

```
## Incumplimiento          Saldo          Monto      Modalidad      Intereses
##      "numeric"      "numeric"      "numeric"      "numeric"      "numeric"
##        Destino         Genero       Garantia
##      "numeric"      "numeric"      "numeric"
```

```
Base_Simulada_Inc$Destino<-factor(Base_Simulada_Inc$Destino)
Base_Simulada_Inc$Genero<-factor(Base_Simulada_Inc$Genero)
Base_Simulada_Inc$Modalidad<-factor(Base_Simulada_Inc$Modalidad)
Base_Simulada_Inc$Intereses<-factor(Base_Simulada_Inc$Intereses)
Base_Simulada_Inc$Incumplimiento<-factor(Base_Simulada_Inc$Incumplimiento)

Regre_Sim<-glm(Incumplimiento~Genero+Destino+Modalidad+Monto+Intereses+Saldo+Garantia, family=binomial,
```

```
## Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred
```

```
summary(Regre_Sim)
```

```
##
## Call:
## glm(formula = Incumplimiento ~ Genero + Destino + Modalidad +
##     Monto + Intereses + Saldo + Garantia, family = binomial,
##     data = Base_Simulada_Inc)
##
## Deviance Residuals:
##      Min        1Q    Median        3Q       Max
## -2.67491  -0.00615  -0.00001   0.00262   2.53389
##
## Coefficients:
##               Estimate Std. Error z value Pr(>|z|)
```

```
## (Intercept) -1.989e+01  2.520e+00  -7.892 2.96e-15 ***
## Genero2     -8.959e-01  3.388e-01  -2.644  0.00819 **
## Destino2     1.833e+00  4.616e-01   3.971 7.17e-05 ***
## Destino3    -2.804e+01  3.318e+02  -0.084  0.93267
## Modalidad2   8.062e-01  3.729e-01   2.162  0.03060 *
## Modalidad3  -3.456e+00  7.949e-01  -4.347 1.38e-05 ***
## Monto       -3.901e-04  4.037e-05  -9.663  < 2e-16 ***
## Intereses2   1.882e+01  2.364e+00   7.963 1.67e-15 ***
## Intereses3   2.066e+01  2.514e+00   8.216  < 2e-16 ***
## Saldo        3.239e-04  3.493e-05   9.272  < 2e-16 ***
## Garantia     4.812e-05  1.204e-05   3.998 6.39e-05 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 1226.77  on 999  degrees of freedom
## Residual deviance:  256.89  on 989  degrees of freedom
## AIC: 278.89
##
## Number of Fisher Scoring iterations: 17
```

Regresión Logística.

Solo se encuentran problemas con destino 3.

```
Anova(Regre_Sim)
```

```
## Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred

## Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred

## Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred

## Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred

## Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred

## Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred

## Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred
## Analysis of Deviance Table (Type II tests)
##
## Response: Incumplimiento
##           LR Chisq Df Pr(>Chisq)
## Genero        7.33  1   0.006787 **
## Destino     134.53  2  < 2.2e-16 ***
## Modalidad    46.04  2  1.007e-10 ***
## Monto       370.40  1  < 2.2e-16 ***
## Intereses   489.16  2  < 2.2e-16 ***
## Saldo       472.29  1  < 2.2e-16 ***
## Garantia     26.13  1  3.186e-07 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Encontramos que no tiene problemas con el Anova la variable destino.

Estimación de valores

```
Valores<-fitted.values(Regre_Sim)
```

Pruebas de ajuste.

Hosmer-Lemshow.

```
ycorte<-fitted.values(Regre_Sim)
ycorte.2<-cut(ycorte, breaks=10, incluide.lowest=TRUE)

table(ycorte.2)
```

```
## ycorte.2
## (-0.001,0.1]    (0.1,0.2]    (0.2,0.3]    (0.3,0.4]    (0.4,0.5]
##          600           39           18           23           21
##    (0.5,0.6]    (0.6,0.7]    (0.7,0.8]    (0.8,0.9]      (0.9,1]
##           17           13           26           31          212
```

```
yinc<-Base_Simulada_Inc$Incumplimiento

yinc<-ifelse(yinc==1,1,0)

yobs<-xtabs(cbind(1-yinc,yinc) ~ ycorte.2)

yobs
```

```
##
## ycorte.2        V1 yinc
##   (-0.001,0.1] 595    5
##   (0.1,0.2]     33    6
##   (0.2,0.3]     13    5
##   (0.3,0.4]     16    7
##   (0.4,0.5]     13    8
##   (0.5,0.6]     12    5
##   (0.6,0.7]      3   10
##   (0.7,0.8]      7   19
##   (0.8,0.9]      3   28
##   (0.9,1]        2  210
```

```
yexp<-xtabs(cbind(1-ycorte, ycorte) ~ ycorte.2)

hosmer<-sum((yobs-yexp)^2/yexp)

hosmer
```

```
## [1] 8.659522
```

```
p.valuehos<-1-pchisq(hosmer, 8)

p.valuehos
```

```
## [1] 0.3718271
```

Pseudo R2.

```
RsqrMCFadden<-1-(Regre_Sim$deviance/Regre_Sim$null.deviance)
RsqrMCFadden
```

```
## [1] 0.7905988
```

Pseudo R2 cox Snell.

```
LRCS<-Regre_Sim$null.deviance - Regre_Sim$deviance
N<-sum(weights(Regre_Sim))
RsqrCN<-1-exp(-LRCS/N)

RsqrCN
```

```
## [1] 0.6208724
```

Curvas ROC

```
prediccion<-ifelse(fitted.values(Regre_Sim)>=0.5,1,0)
table(prediccion)
```

```
## prediccion
##   0   1
## 701 299
```

```
tabli<-table(Base_Simulada_Inc$Incumplimiento,prediccion)
tcc<-100*sum(diag(tabli))/sum(tabli)
tcc
```

```
## [1] 94.2
```

Si se elige punto de corte 0.5 tenemos clasificación correcta en 94.2% de los individuos.

```
predigo<-prediction(fitted.values(Regre_Sim),Base_Simulada_Inc$Incumplimiento)

tab<-performance(predigo,measure = "acc")

posicionmax<-sapply(tab@y.values,which.max)

posicionmax
```

```
## [1] 275
```

```
puntocorte<-sapply(tab@x.values,"[",posicionmax)

puntocorte
```

```
##       49
## 0.599068
```

Con esto obtenemos el punto de corte máximo que es 0.599068.

Medimos el área bajo la curva:

```
AUC<-performance(predigo,"auc")
AUC@y.name
```

```
## [1] "Area under the ROC curve"
```

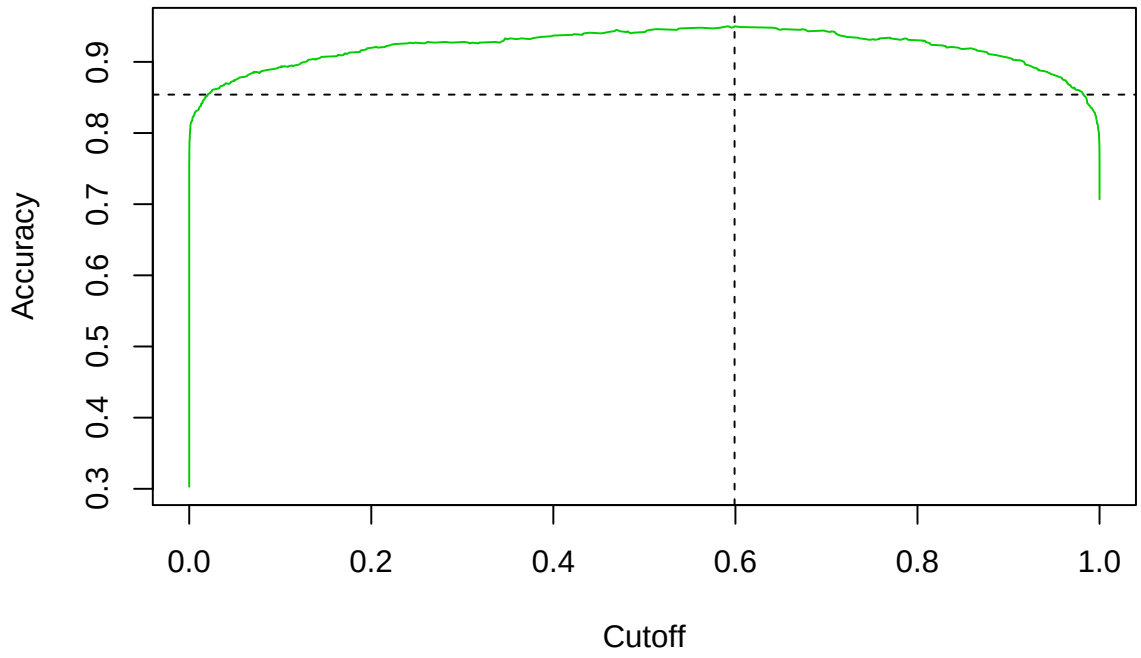
```
AUC@y.values
```

```
## [[1]]
## [1] 0.9877078
```

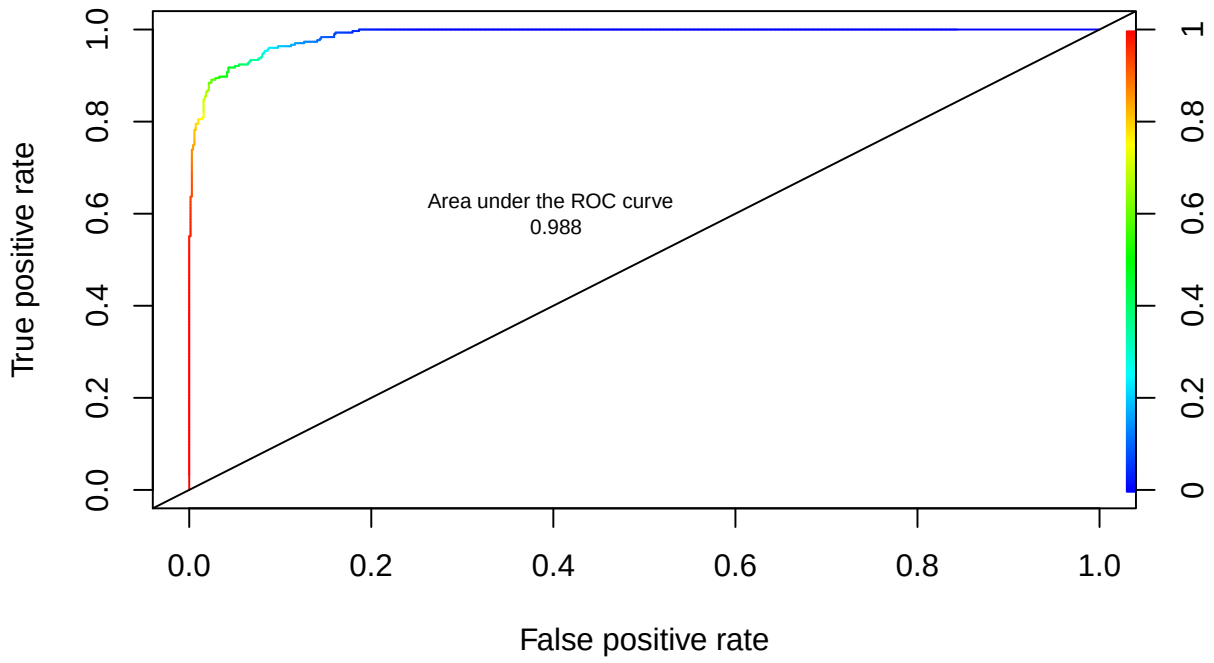
Area bajo la curva de 0.9877078

Graficos curva ROC:

```r
plot(tab,col=3)
abline(h=0.854,lty=2)
abline(v=puntocorte,lty=2)
```




```r
perfi<-performance(predigo,"tpr","fpr")
plot(perfi,colorize=TRUE)
abline(a=0,b=1)
text(0.4,0.6,paste(AUC@y.name,"\n",round(unlist(AUC@y.values),3)),cex=0.7)
```



Curvas ROC con punto critico

```
prediccion<-ifelse(fitted.values(Regre_Sim)>=0.599068,1,0)
table(prediccion)
```

```
## prediccion
##   0   1
## 718 282
```

```
tabli<-table(Base_Simulada_Inc$Incumplimiento,prediccion)
tcc<-100*sum(diag(tabli))/sum(tabli)
tcc
```

```
## [1] 94.9
```

```
table(Base_Simulada_Inc$Incumplimiento,prediccion)
```

```
##    prediccion
##       0   1
##   0 682  15
##   1  36 267
```

Con esto procedemos a encontrar el modelo que tiene el mejor nivel predictivo.

Retiramos Destino al encontrar posible insuficiencia en Z (1.96).

```
Regre_Sim_2<-glm(Incumplimiento~Genero+Modalidad+Monto+Intereses+Saldo+Garantia, family=binomial, data=
```

```
## Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred
```

```r
summary(Regre_Sim_2)
```

```
##
## Call:
## glm(formula = Incumplimiento ~ Genero + Modalidad + Monto + Intereses +
##     Saldo + Garantia, family = binomial, data = Base_Simulada_Inc)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -4.1429  -0.1064  -0.0025   0.0425   2.2379
##
## Coefficients:
##               Estimate Std. Error z value Pr(>|z|)
## (Intercept) -1.063e+01  1.320e+00  -8.052 8.12e-16 ***
## Genero2     -6.038e-01  2.652e-01  -2.277 0.022788 *
## Modalidad2   5.631e-01  3.038e-01   1.853 0.063863 .
## Modalidad3  -2.138e+00  5.607e-01  -3.812 0.000138 ***
## Monto       -2.497e-04  2.463e-05 -10.139  < 2e-16 ***
## Intereses2   1.011e+01  1.237e+00   8.169 3.10e-16 ***
## Intereses3   1.133e+01  1.298e+00   8.734  < 2e-16 ***
## Saldo        1.781e-04  1.787e-05   9.967  < 2e-16 ***
## Garantia     3.113e-05  7.490e-06   4.157 3.23e-05 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 1226.77  on 999  degrees of freedom
## Residual deviance:  391.42  on 991  degrees of freedom
## AIC: 409.42
##
## Number of Fisher Scoring iterations: 9
```

```r
Anova(Regre_Sim_2)
```

```
## Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred

## Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred

## Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred

## Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred
```

```
## Analysis of Deviance Table (Type II tests)
##
## Response: Incumplimiento
##           LR Chisq Df Pr(>Chisq)
## Genero        5.30  1    0.02129 *
## Modalidad    34.59  2  3.086e-08 ***
## Monto       306.27  1  < 2.2e-16 ***
## Intereses   407.66  2  < 2.2e-16 ***
## Saldo       383.95  1  < 2.2e-16 ***
## Garantia     19.93  1  8.023e-06 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Curvas ROC con sim 2

```
prediccion<-ifelse(fitted.values(Regre_Sim_2)>=0.5,1,0)
table(prediccion)
```

```
## prediccion
##   0   1
## 702 298
```

```
tabli<-table(Base_Simulada_Inc$Incumplimiento,prediccion)
tcc<-100*sum(diag(tabli))/sum(tabli)
tcc
```

```
## [1] 92.3
```

Si se elige punto de corte 0.5 tenemos clasificación correcta en 92.3% de los individuos

```
predigo<-prediction(fitted.values(Regre_Sim_2),Base_Simulada_Inc$Incumplimiento)

tab<-performance(predigo,measure = "acc")

posicionmax<-sapply(tab@y.values,which.max)

posicionmax
```

```
## [1] 281
```

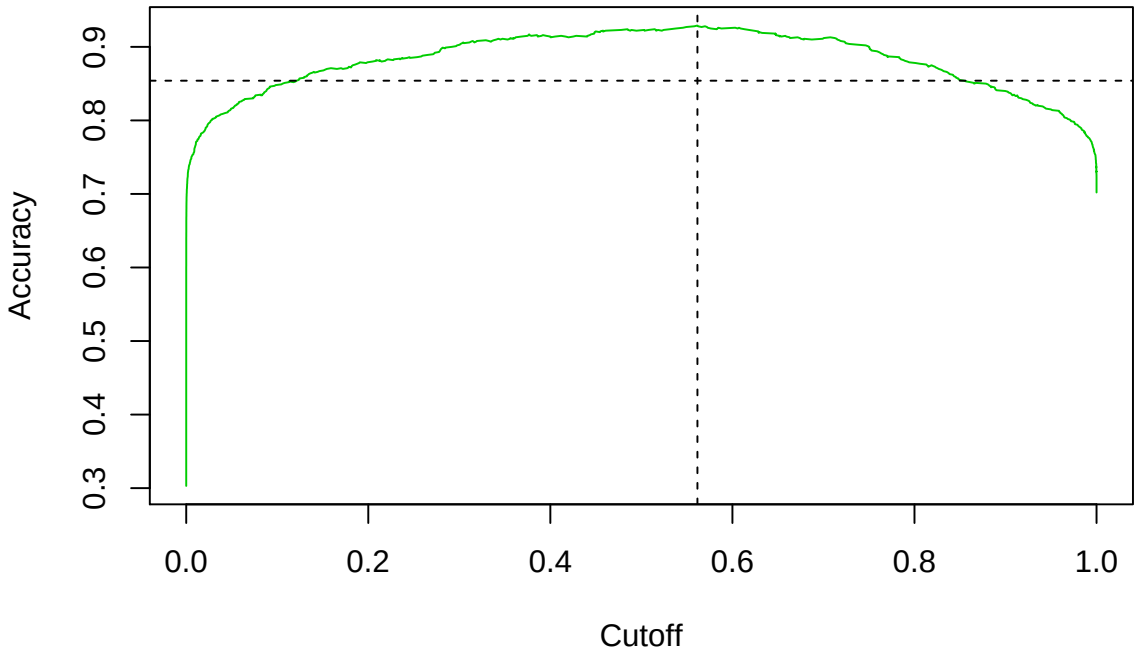
```
puntocorte<-sapply(tab@x.values,"[",posicionmax)

puntocorte
```

```
##       349
## 0.561567
```

Encontramos un punto de corte mayor con 0.561567.

```
plot(tab,col=3)
abline(h=0.854,lty=2)
abline(v=puntocorte,lty=2)
```



Medimos el area bajo la curva:

```
AUC<-performance(predigo,"auc")
AUC@y.name
```

```
## [1] "Area under the ROC curve"
```

```
AUC@y.values
```

```
## [[1]]
## [1] 0.9749468
```

Área bajo la curva de 0.9749468 comparando con el 0.9877078 (original).

Por lo cual podemos ver que no mejora el poder predictivo del modelo, por lo cual se considera que es mejor conservar la variable que retirarla.

Análisis de residuos.

Residuos de Pearson:

```
res_pearson<-residuals(Regre_Sim,type = "pearson")
res_pearson_sig<-abs(res_pearson)>2
table(res_pearson_sig)
```

```
## res_pearson_sig
## FALSE  TRUE
##   984    16
```

```
res_orde<-sort(abs(res_pearson[res_pearson_sig]),decreasing = TRUE)
head(res_orde)
```

```
##      685      583      663       12      390      623
## 5.898047 4.877144 4.396680 3.730403 3.622463 3.565399
```

Residuos estandarizados de pearson:

```
res_stand<-rstandard(Regre_Sim,type = "pearson")

res_stand_sign<-abs(res_stand)>2
table(res_stand_sign)
```

```
## res_stand_sign
## FALSE  TRUE
##   983    17
```

Residuos de la devianza:

```
res_dev<-residuals(Regre_Sim,type = "deviance")
res_dev_sign<-abs(res_dev)>2
table(res_dev_sign)
```

```
## res_dev_sign
## FALSE  TRUE
##   990    10
```

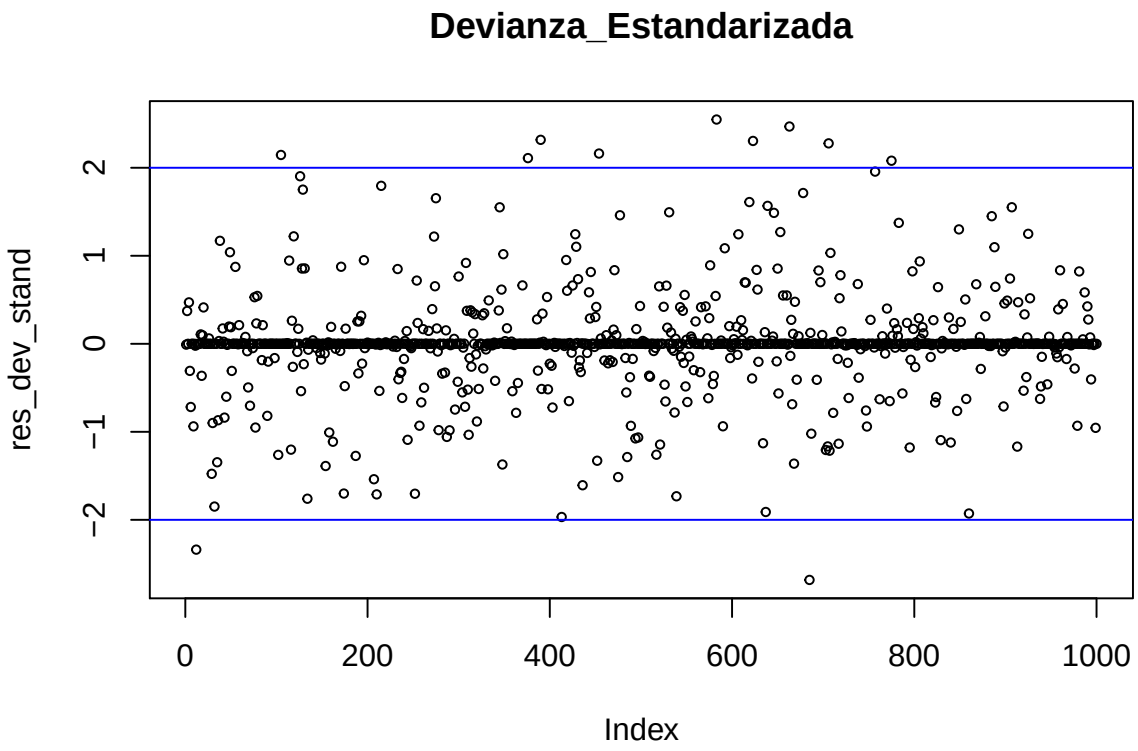
Residuos estandarizados de Devianza:

```
res_dev_stand<-rstandard(Regre_Sim,type = "deviance")
table(abs(res_dev_stand)>2)
```

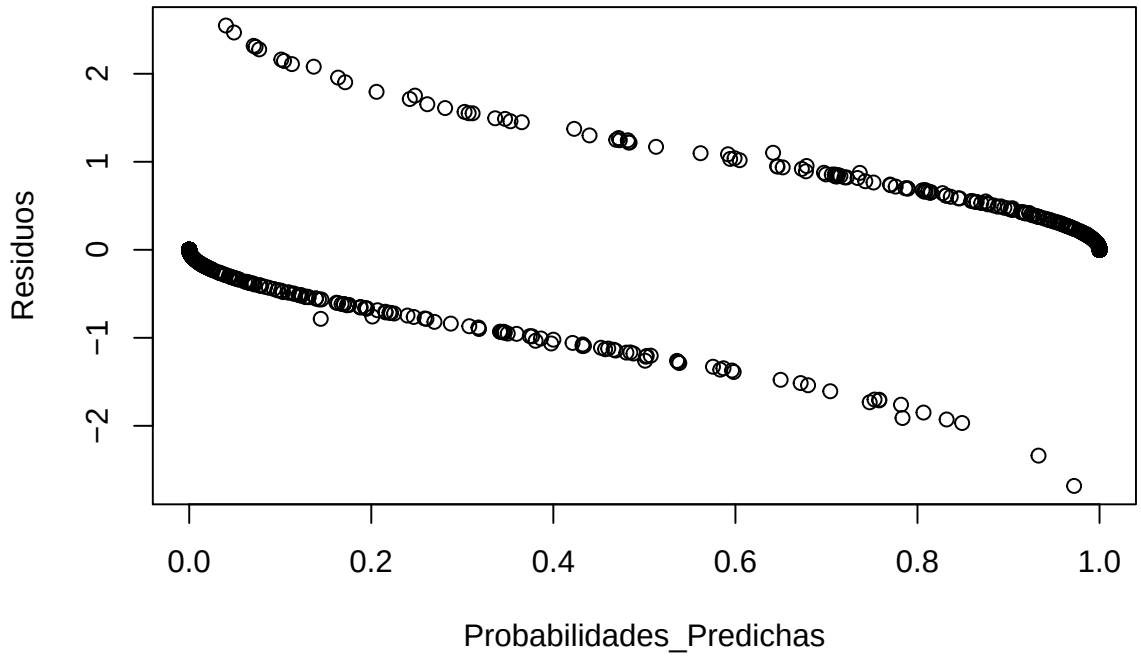
```
##
## FALSE  TRUE
##   989    11
```

Graficas:

```
plot(res_dev_stand,cex=0.6,main="Devianza_Estandarizada")
abline(h=c(-2,2),col="blue")
```




```
plot(fitted.values(Regre_Sim),res_dev_stand,xlab="Probabilidades_Predichas",ylab="Residuos")
```



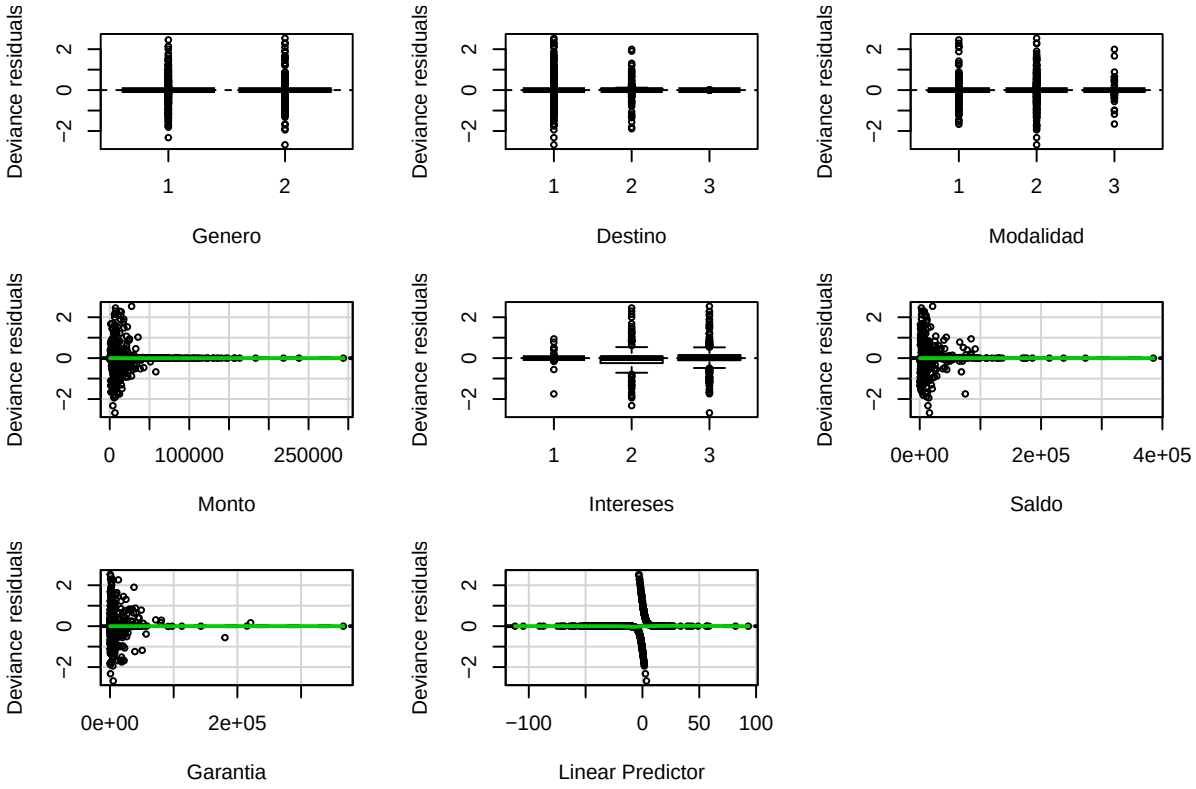

```
#plot(jitter(Base_Simulada_Inc$Incumplimiento), fitted.values(Regre_Sim),cex=0.5, #xlab="Valores_Observ

residualPlots(Regre_Sim,type="deviance",cex=0.6)
```

```
## Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred

## Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred

## Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred
```




```
##           Test stat Pr(>|t|)
## Genero           NA       NA
## Destino          NA       NA
## Modalidad        NA       NA
## Monto         0.005    0.942
## Intereses        NA       NA
## Saldo         0.496    0.481
## Garantia      0.103    0.748
```

Medidas de influencia.

Distancia de cook.

```
distancia_cook<-cooks.distance(Regre_Sim)
table(distancia_cook>1)
```

```
##
## FALSE
##  1000
```

Obtenemos que los datos no cuentan con datos influyentes que puedan afectar los estimadores. Colinealidad.

```
vif(Regre_Sim)
```

```
##                GVIF Df GVIF^(1/(2*Df))
## Genero     1.125554  1        1.060921
## Destino    1.157912  2        1.037335
## Modalidad  1.498392  2        1.106385
## Monto      3.636815  1        1.907044
```

```
## Intereses 6.836504  2        1.616994
## Saldo     9.224860  1        3.037246
## Garantia  1.291837  1        1.136590
```

```
cor.test(Base_Simulada_Inc$Monto,Base_Simulada_Inc$Saldo)
```

```
##
##  Pearson's product-moment correlation
##
## data:  Base_Simulada_Inc$Monto and Base_Simulada_Inc$Saldo
## t = -2.1026, df = 998, p-value = 0.03575
## alternative hypothesis: true correlation is not equal to 0
## 95 percent confidence interval:
##  -0.127877584 -0.004436157
## sample estimates:
##         cor
## -0.06641098
```

Podemos encontrar que existe colinealidad aceptable en la variable Saldo.

Validación Cruzada.

```
require(DAAG)
```

```
## Loading required package: DAAG
## Loading required package: lattice
##
## Attaching package: 'DAAG'
## The following object is masked from 'package:car':
##
##     vif
## The following object is masked from 'package:survival':
##
##     lung
## The following object is masked from 'package:MASS':
##
##     hills
```

```
cruzada<-CVbinary(Regre_Sim)
```

```
## Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred
##
## Fold:  4
## Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred
##  5
## Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred
##  1
## Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred
##  10
## Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred
```

```
##  9
## Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred
##  2
## Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred
##  3
## Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred
##  7
## Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred
##  8
## Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred
##  6
## Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred
##
## Internal estimate of accuracy = 0.942
## Cross-validation estimate of accuracy = 0.938
```

```
Regre_Sim$coefficients
```

```
##   (Intercept)       Genero2      Destino2      Destino3    Modalidad2
## -1.989147e+01 -8.958634e-01  1.833018e+00 -2.803809e+01  8.062379e-01
##    Modalidad3         Monto    Intereses2    Intereses3         Saldo
## -3.455671e+00 -3.901164e-04  1.882420e+01  2.065668e+01  3.238663e-04
##      Garantia
##  4.811992e-05
```

Calcular la probabilidad con los coeficientes.

```
Regre_Sim$coefficients
```

```
##   (Intercept)       Genero2      Destino2      Destino3    Modalidad2
## -1.989147e+01 -8.958634e-01  1.833018e+00 -2.803809e+01  8.062379e-01
##    Modalidad3         Monto    Intereses2    Intereses3         Saldo
## -3.455671e+00 -3.901164e-04  1.882420e+01  2.065668e+01  3.238663e-04
##      Garantia
##  4.811992e-05
```

```
probain<-NULL

for(i in 1:sample_size)
{
  x1=Base_Simulada_Inc$Genero[i]
  x2=Base_Simulada_Inc$Destino[i]
  x3=Base_Simulada_Inc$Modalidad[i]
  x4=Base_Simulada_Inc$Monto[i]
  x5=Base_Simulada_Inc$Intereses[i]
  x6=Base_Simulada_Inc$Saldo[i]
  x7=Base_Simulada_Inc$Garantia[i]

  r1=0
  r2=0
```

```
  r3=0
  r4=0
  r5=0
  r6=0
  r7=0
  r8=0
  r9=0
  r10=0

  if(x1==2){r1=1}
  if(x2==2){r2=1}
  if(x2==3){r3=1}
  if(x3==2){r4=1}
  if(x3==3){r5=1}
  if(x5==2){r7=1}
  if(x5==3){r8=1}

  r6=x4
  r9=x6
  r10=x7

  probain[i]=1/(1+exp(-(Regre_Sim$coefficients[1]+Regre_Sim$coefficients[2]*r1+Regre_Sim$coefficients[3]

}

maximo<-max(probain)
minimo<-min(probain)
indmaxi<-NULL
indmini<-NULL
for(i in 1:sample_size)
{
  if(probain[i]==maximo){indmaxi[1]=i}
  if(probain[i]==minimo){indmini[1]=i}
}


indmaxi
```

```
## [1] 975
```

```
indmini
```

```
## [1] 513
```

```
gta<-indmaxi


Base_Simulada_Inc$Genero[gta]
```

```
## [1] 2
## Levels: 1 2
```

```
Base_Simulada_Inc$Destino[gta]
```

```
## [1] 2
## Levels: 1 2 3
```

```
Base_Simulada_Inc$Modalidad[gta]
```

```
## [1] 3
## Levels: 1 2 3
```

```
Base_Simulada_Inc$Monto[gta]
```

```
## [1] 23382.51
```

```
Base_Simulada_Inc$Intereses[gta]
```

```
## [1] 1
## Levels: 1 2 3
```

```
Base_Simulada_Inc$Saldo[gta]
```

```
## [1] 384989.5
```

```
Base_Simulada_Inc$Garantia[gta]
```

```
## [1] 808.3773
```

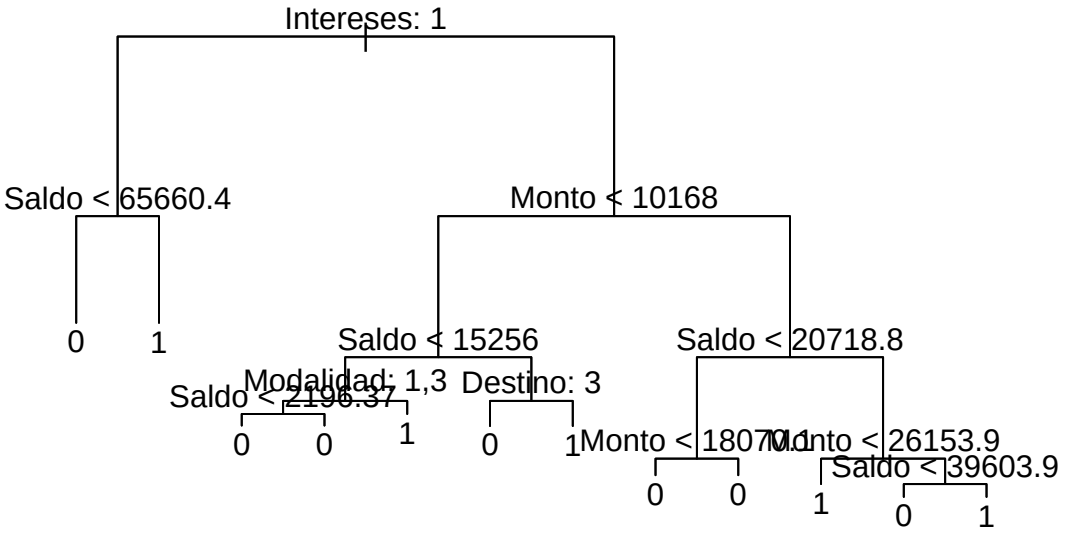
Árbol de clasificación.

Código de creación de árboles de clasificación:

```
tree.inc=tree(Incumplimiento~Genero+Destino+Modalidad+Monto+Intereses+Saldo+Garantia, data=Base_Simulada
summary(tree.inc)
```

```
##
## Classification tree:
## tree(formula = Incumplimiento ~ Genero + Destino + Modalidad +
##     Monto + Intereses + Saldo + Garantia, data = Base_Simulada_Inc)
## Variables actually used in tree construction:
## [1] "Intereses" "Saldo"     "Monto"     "Modalidad" "Destino"
## Number of terminal nodes:  12
## Residual mean deviance:  0.429 = 423.8 / 988
## Misclassification error rate: 0.096 = 96 / 1000
```

```
plot(tree.inc)
text(tree.inc,pretty = 1)
```




```
tree.inc
```

```
## node), split, n, deviance, yval, (yprob)
##       * denotes terminal node
##
##  1) root 1000 1227.000 0 ( 0.697000 0.303000 )
##    2) Intereses: 1 368  143.700 0 ( 0.951087 0.048913 )
##      4) Saldo < 65660.4 347    0.000 0 ( 1.000000 0.000000 ) *
##      5) Saldo > 65660.4 21   17.220 1 ( 0.142857 0.857143 ) *
##    3) Intereses: 2,3 632  870.000 0 ( 0.549051 0.450949 )
##      6) Monto < 10168 295  350.600 1 ( 0.281356 0.718644 )
##       12) Saldo < 15256 186  252.300 1 ( 0.413978 0.586022 )
##         24) Modalidad: 1,3 66   88.500 0 ( 0.606061 0.393939 )
##           48) Saldo < 2196.37 12    0.000 0 ( 1.000000 0.000000 ) *
##           49) Saldo > 2196.37 54   74.790 0 ( 0.518519 0.481481 ) *
##         25) Modalidad: 2 120  148.300 1 ( 0.308333 0.691667 ) *
##       13) Saldo > 15256 109   46.460 1 ( 0.055046 0.944954 )
##         26) Destino: 3 6    5.407 0 ( 0.833333 0.166667 ) *
##         27) Destino: 1,2 103   11.260 1 ( 0.009709 0.990291 ) *
##      7) Monto > 10168 337  352.200 0 ( 0.783383 0.216617 )
##       14) Saldo < 20718.8 256  130.300 0 ( 0.929688 0.070312 )
##         28) Monto < 18070.1 104   89.300 0 ( 0.846154 0.153846 ) *
##         29) Monto > 18070.1 152   21.300 0 ( 0.986842 0.013158 ) *
##       15) Saldo > 20718.8 81  101.700 1 ( 0.320988 0.679012 )
##         30) Monto < 26153.9 50   32.510 1 ( 0.100000 0.900000 ) *
##         31) Monto > 26153.9 31   38.990 0 ( 0.677419 0.322581 )
##           62) Saldo < 39603.9 18    7.724 0 ( 0.944444 0.055556 ) *
```

```
##           63) Saldo > 39603.9 13   16.050 1 ( 0.307692 0.692308 ) *
```

Creación del mejor árbol de clasificación posible con validación cruzada.

```
set.seed(300000)
train=sample(1:nrow(Base_Simulada_Inc),200)
tree.sam=Base_Simulada_Inc[-train,]
Probarb=Base_Simulada_Inc$Incumplimiento[-train]
tree.inc2=tree(Incumplimiento~Genero+Destino+Modalidad+Monto+Intereses+Saldo+Garantia, data=Base_Simula
tree.pred=predict(tree.inc2,tree.sam,type = "class")
table(tree.pred,Probarb)
```

```
##          Probarb
## tree.pred   0   1
##         0 507  33
##         1  57 203
```

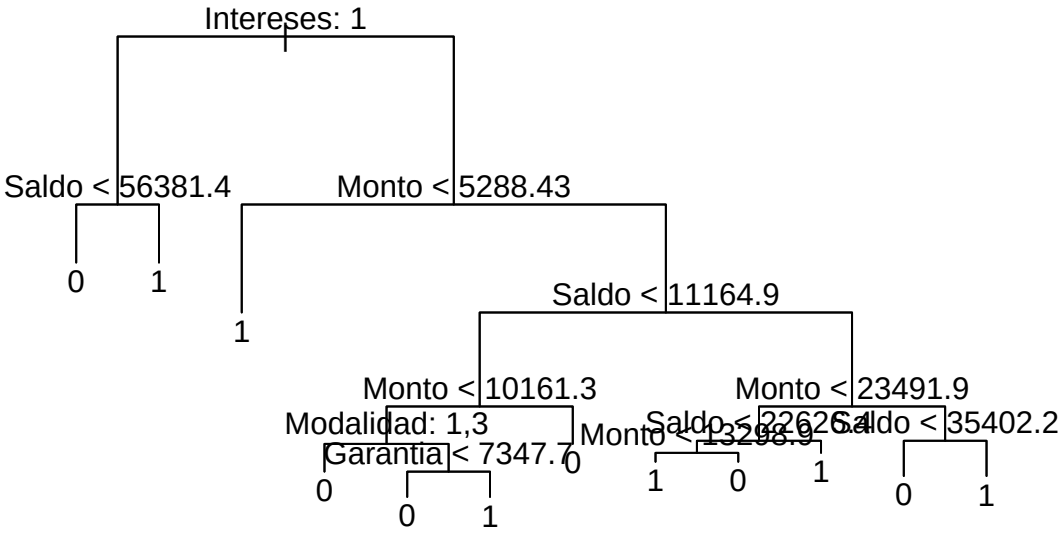
(507+203)/800=88% de clasificación correcta.

```
set.seed(300006)
cv.arbol=cv.tree(tree.inc2,FUN=prune.misclass)
cv.arbol
```

```
## $size
## [1] 15 12 10  9  6  5  4  1
##
## $dev
## [1] 34 30 32 33 33 37 38 63
##
## $k
## [1]      -Inf  0.000000  0.500000  1.000000  1.666667  3.000000  5.000000
## [8] 13.333333
##
## $method
## [1] "misclass"
##
## attr(,"class")
## [1] "prune"         "tree.sequence"
```

Se encuentra que el arbol con size=12 tiene el menor error de validación cruzada.

```
prune.arbol=prune.misclass(tree.inc2, best=12)
plot(prune.arbol)
text(prune.arbol,pretty =6 )
```




```
prune.arbol
```

```
## node), split, n, deviance, yval, (yprob)
##       * denotes terminal node
##
##   1) root 200 255.100 0 ( 0.66500 0.33500 )
##     2) Intereses: 1 72  24.940 0 ( 0.95833 0.04167 )
##       4) Saldo < 56381.4 67   0.000 0 ( 1.00000 0.00000 ) *
##       5) Saldo > 56381.4 5   6.730 1 ( 0.40000 0.60000 ) *
##     3) Intereses: 2,3 128 177.400 1 ( 0.50000 0.50000 )
##       6) Monto < 5288.43 30  14.700 1 ( 0.06667 0.93333 ) *
##       7) Monto > 5288.43 98 128.900 0 ( 0.63265 0.36735 )
##        14) Saldo < 11164.9 56  45.930 0 ( 0.85714 0.14286 )
##          28) Monto < 10161.3 19  25.010 0 ( 0.63158 0.36842 )
##            56) Modalidad: 1,3 7   0.000 0 ( 1.00000 0.00000 ) *
##            57) Modalidad: 2 12  16.300 1 ( 0.41667 0.58333 )
##             114) Garantia < 7347.7 7   8.376 0 ( 0.71429 0.28571 ) *
##             115) Garantia > 7347.7 5   0.000 1 ( 0.00000 1.00000 ) *
##          29) Monto > 10161.3 37   9.195 0 ( 0.97297 0.02703 ) *
##        15) Saldo > 11164.9 42  53.470 1 ( 0.33333 0.66667 )
##          30) Monto < 23491.9 29  26.660 1 ( 0.17241 0.82759 )
##            60) Saldo < 22626.4 12  15.280 1 ( 0.33333 0.66667 )
##             120) Monto < 13298.9 7   5.742 1 ( 0.14286 0.85714 ) *
##             121) Monto > 13298.9 5   6.730 0 ( 0.60000 0.40000 ) *
##            61) Saldo > 22626.4 17   7.606 1 ( 0.05882 0.94118 ) *
##          31) Monto > 23491.9 13  16.050 0 ( 0.69231 0.30769 )
```

```
##             62) Saldo < 35402.2 8   0.000 0 ( 1.00000 0.00000 ) *
##             63) Saldo > 35402.2 5   5.004 1 ( 0.20000 0.80000 ) *
```

Se muestran las proporciones del árbol. Con el símbolo * se indica a los nodos terminales.

Por último se muestra el árbol que se encuentra en el capítulo del modelo de regresión logística:

```
tree.pred2=predict(prune.arbol,tree.sam,type = "class")
table(tree.pred2,Probarb)
```

```
##           Probarb
## tree.pred2   0   1
##          0 507  33
##          1  57 203
```

(507+203)/800=88% de clasificación correcta.

-

Bibliografía

[1] Agresti, A. *An introduction to categorical data analysis.* Wiley, 2007.

[2] Alayón Fumero, J., and Pérez Roger, J. *Análisis y medición del riesgo de crédito.* Universidad de la Laguna, 2014.

[3] Baxter, K. *Administración de riesgo.* Trillas, 2010.

[4] Cañadas Reche, J. L. *Regresión logística: Tratamiento computacional en R.* Universidad de Granada, 2013.

[5] De Lara Haro, A. *Medición y control de riesgos financieros*, 3 ed. Limusa, 2014.

[6] Elizondo, A., and López Romero, C. *El riesgo de crédito en México: una evaluación de modelos recientes para cuantificarlo.* Gaceta de Economía, 1999.

[7] Fernéndez Hatre, A. *Indicadores de gestión y cuadro de mando integral.* Instituto de Desarrollo Económico del Principado de Asturias, 2004.

[8] G. Kleinbaum, D., and Klein, M. *Logistic Regression*, 3 ed. Springer, 2010.

[9] García Sánchez, M., and Sánchez Barradas, C. *Riesgo de crédito en México: aplicación del modelo Credit Metrics.* Universidad de las Américas Puebla, 2005.

[10] Gómez Villegas, M. A. *Karl Pearson, el Creador de la Estadística Matemática.* Universidad Complutense de Madrid, 2009.

[11] Hosmer, D. W., and Lemeshow, S. *Applied logistic regression.* Wiley & Sons, 2000.

[12] J. Sheather, S. *A Modern Approach to Regression with R.* Springer, 2009.

[13] M. Ross, S. *Introducción a la estadística.* Reverte, 2007.

[14] Márquez Diez-Canedo, J. *Una nueva visión del riesgo de crédito*, 2 ed. Limusa, 2009.

[15] Mateos, G., Morales, A., et al. *Historia de la Probabilidad y de la Estadística.* A.H.E.P.E., 2002.

[16] Mejía Quijano, R. C. *Administración de riesgos: Un enfoque empresarial.* Universidad de EAFIT, 2006.

[17] Mendenhall, W., Beaver, R., and Beaver, B. *Introducción a la probabilidad y estadística*, 13 ed. Cengage Learning, 2010.

[18] Meza Orozco, J. d. J. *Matemáticas Financieras aplicadas.* Ecoe, 2011.

[19] Rincon, L. *Curso intermedio de probabilidad*. Departamento de Matemáticas, 2007.

[20] Ruiz Muñoz, D. *Manual de Estadística*. Eumed, 2004.

[21] Ruiz Muñoz, D. *Manual de Estadística*. Eumed, 2004.

[22] Sierra Nuñez, L. *La regulación internacional de Basilea y su evolución en el Sistema Financiero Mexicano*. UNAM, 2011.

[23] Silva Aycaguer, L. C. *Regresión logística*. La Muralla, 2004.

[24] Tibshirani, R., Hastie, T., et al. *An Introduction to Statistical Learning*. Springer, 2013.

[25] Tusell, A. *Análisis de Regresión. Introducción Teórica y Práctica basada en R*. Bilbao, 2011.

[26] Ustáriz González, L. H. *El comité de Basilea y la supervisión bancaria*. Pontificia, 2003.

[27] Vergara Schmalbach, J. C., and Quesada Ibarguen, V. M. *Estadística Básica con aplicaciones en MS Excel*. Grupo Métodos Cuantitativos de Gestión, 2009.

[28] Weisberg, S. *Applied Linear Regression*, 3 ed. Wiley, 2005.

[29] Zamora Maass, M. L. *Modelo de regresi"on logística en Credit Scoring para el sector rural en México de 2002 a 2011*. UNAM, 2013.

Artículos

- Universidad Autónoma de Nayarit. Memoria del XXI Coloquio Mexicano de Economía Matemática y Econometría Tomo II. 2011.
- Banco de México. Divulgación: Sistema Financiero. [Recurso en línea].
- Banco de México. Divulgación: Glosario. [Recurso en línea].
- Banco de México. Definiciones Básicas de Riesgos. México 2005 [Recurso en línea].
- Banco de Pagos Internacionales. Aplicación de Basilea II: aspectos prácticos. Suiza 2004. [Recurso en línea].
- Condusef. Cuanto pagas de intereses. [Recurso en línea].
- CNBV. Glosario de términos: Portafolio de Información. [Recurso en línea].

Sitios Web

- http://expasion.com/diccionario-economico/modelo-econometrico.html
- www.banxico.org.mx/
- http://www.gob.mx/cnbv
- http://www.gob.mx/condusef