BENEMÉRITA UNIVERSIDAD AUTÓNOMA DE PUEBLA

FACULTAD DE CIENCIAS FÍSICO MATEMÁTICAS

Evaluación del Riesgo Crediticio, a través de Credit Scoring mediante Regresión Logística: Un caso de estudio

T E S I S

que para obtener el título de:

LICENCIADO EN ACTUARÍA

presenta:

ESTEFANIA MEZA SALDAÑA

Directores de tesis:

DRA. HORTENSIA REYES CERVANTES
DRA. BLANCA PÉREZ SALVADOR

PUEBLA, PUE. *MAYO 2017*

BENEMÉRITA UNIVERSIDAD AUTÓNOMA DE PUEBLA

FACULTAD DE CIENCIAS FÍSICO MATEMÁTICAS

Evaluación del Riesgo Crediticio, a través de Credit Scoring mediante Regresión logística: Un caso de estudio

T E S I S

que para obtener el título de:

LICENCIADO EN ACTUARÍA

presenta:

ESTEFANIA MEZA SALDAÑA

Directores de tesis:

DRA. HORTENSIA REYES CERVANTES
DRA. BLANCA PÉREZ SALVADOR

PUEBLA, PUE. *MAYO 2017*

Dedico esta tesis con todo mi cariño a:
Dios por su infinito amor.
Mis padres, Alejandro y Enriqueta
que con sus sacrificios pudo haber
sido esto posible.

Agradecimientos

A Dios por la vida, por las bendiciones que ha concedido para mí y para mi familia y por la oportunidad de haberme permitido concluir una meta más en mi vida.
A mis padres por todo su apoyo, y a toda mi familia por su ayuda incondicional.

Mi más sincero agradecimiento a mis directoras de tesis:
Dra. Hortensia Reyes Cervantes, por su apoyo incondicional, por compartir su conocimiento dentro y fuera de las aulas y por supuesto por su tiempo y paciencia para la realización de esta tesis que sin usted no hubiese sido posible.
Dra. Blanca Rosa Pérez Salvador, por compartir su conocimiento que a pesar de la distancia siempre estuvo ahí para cualquier consejo y apoyo.

A los integrantes del jurado que evaluaron este trabajo:
Dr. Francisco Solano Tajonar Sanabria, Dr. Bulmaro Juárez Hernández, M. C. Brenda Zavala López, por su tiempo, comentarios y observaciones y por el interés y disponibilidad para la revisión del presente trabajo.

A la Benemérita Universidad Autónoma de Puebla, especialmente a la Facultad de Físico Matemáticas y la Facultad de Economía, en las cuales nos formamos como profesionistas y personas.

A todos y cada uno de los profesores con los que tuve el privilegio de aprender y convivir.

A mis compañeros(as) y amigos(as) que conocí a través de esta etapa de mi vida, por las alegrías compartidas.

Índice

Introducción

En finanzas, riesgo está relacionado con la posibilidad de que suceda un evento que se convierta en pérdidas para los participantes involucrados. Existen diferentes tipos de riesgo en los mercados financieros, entre ellos se encuentran, el riesgo de mercado, riesgo de operación, riesgo de contraparte y riesgo de crédito, este último es el que se maneja en este trabajo, definiéndolo como caso particular del riesgo de contraparte, cuando el contrato es uno de crédito, y el deudor no puede pagar su deuda por diferentes factores [3].

En la actualidad, los avances tecnológicos han permitido un desarrollo importante en la automatización de la decisión sobre la aceptación o rechazo de una solicitud de crédito a través de modelos analíticos, evitando el otorgamiento bajo criterios ambiguos, estos modelos requieren de información cuantitativa potencialmente útil para su construcción.
La oportunidad de obtener esta información es cada vez más simple, gracias al importante aumento de la capacidad de almacenaje y la disponibilidad de mejores herramientas para el manejo de datos, el proceso de extracción de información relevante a partir de datos disponibles sigue siendo complejo y costoso.

La modelación de la falla financiera, tanto en personas como en empresas, ha sido un problema altamente estudiado en la literatura. Se han desarrollado modelos matemáticos y estadísticos que buscan predecir el desempeño que tendría una persona si se le otorgase un crédito median-

te la asignación de un puntaje estimado a partir de la información del cliente. Este problema se le conoce como *Credit Scoring* [18].
La utilización de modelos de *credit scoring* para la evaluación del riesgo de crédito, es decir, para estimar probabilidades de incumplimiento y ordenar a los deudores y solicitantes de financiamiento en función de su riesgo de incumplimiento se ha desarrollado dentro de las últimas cuatro décadas [2], esto debido al desarrollo de mejores recursos estadísticos y computacionales, además, de la necesidad por parte de la industria bancaria de hacer más eficaz y eficiente la generación de préstamos, y de tener una mejor evaluación del riesgo de su cartera de clientes cada vez es mayor.
Dentro de los diversos métodos estadísticos más comunes para el desarrollo de *Credit Scorings* se encuentran: Análisis discriminante, Modelo de probabilidad lineal, Modelo Logit, Modelos de Programación lineal, Redes Neuronales, Árboles de decisión, entre otros.

Durante las últimas decadas en las grandes ciudades, para los prestamistas, el *scoring* ha sido una de las herramientas más importantes de mayor eficiencia, estos prestamistas clasifican a los prestatarios potenciales sobre la base de historiales de crédito, así como la experiencia y características socio-económicas del prestatario, basándose fundamentalmente en información cuantitativa. Pero experimentos en Bolivia y Colombia sugieren que el *scoring* de las microfinanzas puede mejorar el juicio de riesgo y por lo tanto, reducir costos, el *scoring* puede ser la siguiente innovación tecnológica importante en las microfinanzas [16].

Las instituciones microfinancieras líderes de la región de América Latina y el Caribe están estableciendo estándares de desempeño que nunca se hubieran imaginado antes. Y he aquí la importancia de introducir innovaciones tecnológicas tales como la calificación automatizada del crédito, *scorings*, que preveen el riesgo en función de características cuantificadas, registradas en una base de datos, para las Microfinancieras de mayor tamaño, el *scoring* puede incrementar eficiencia, alcance y sostenibilidad

mediante una mejora en la asignación del tiempo de los agentes de crédito [20].

En México, los riesgos crediticios constituyen en promedio, poco más del 80 % de los activos bancarios sujetos a riesgo. De acuerdo a la Encuesta Nacional de Inclusión Financiera (ENIF) del 2015, casi el 30 % de los adultos en México contaban con un crédito al consumo, (22.1 millones de personas).

La predicción del incumplimiento de un préstamo tiene una utilidad muy práctica. De hecho, la identificación del riesgo de incumplimiento parece ser de suma importancia para los emisores de créditos financieros.

El uso del *Credit Scoring* no está exento de sus limitaciones, a pesar de estas, la mayoría de procesos de aprobación continúan utilizando credit scores.

En este trabajo se desarrolla un modelo estadístico integrado para evaluar un préstamo otorgado por una entidad financiera, mediante el análisis de la información que se tiene de cada uno de los clientes, a través de un Modelo de Regresión Logística, para obtener las características más significativas y poder establecer una regla de aceptación.

Objetivos

1. Aplicar los conocimientos de Estadística y el material de evaluación de riesgo crediticio a una base de datos real.

2. Usar un paquete estadístico que permita dar solución.

3. Tener un criterio de decisión estadística en términos de las variables implicadas para decidir a quienes se les otorga un crédito financiero.

En el presente trabajo se implementa un modelo *credit scoring* a una base de datos de un banco alemán de 1994. Para el desarrollo de este modelo se usa la herramienta estadística de Regresión Logística.
La estructura de este trabajo será implementada en 3 capítulos:

- Capítulo 1: Preliminares.
  Donde se presenta la teoría estadística importante que se necesita para la implementación y desarrollo del Modelo de Regresión Logística.

- Capítulo 2: Credit Score.
  En este capítulo se introduce el concepto de *Credit Scoring* y se hace una síntesis de las ventajas y limitaciones que presenta.

- Capítulo 3: Caso práctico: Análisis de datos.
  La aplicación del método se llevó a cabo en este capítulo, en el cual se hace un resumen del proceso que se realizó y los resultados que se obtuvieron al analizar los datos recopilados mediante la base de datos alemana a través del paquete estadístico SPSS [17].

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

Capítulo 1

Preliminares

En este capítulo se menciona la teoría estadística que se utiliza como base para la aplicación y resolución del caso de estudio en cuestión.

1.1. Escalas de Medición

Los datos generalmente están asociados a la definición de las variables a investigar, pues se relacionan con los conceptos de referencia de la investigación. Un investigador del área social, Stevens en 1946, clasificó los diferentes tipos de escalas que hoy en día conocemos: nominal, ordinal, de intervalo y de razón.

- *Nominal*: Esta escala se utiliza como medida de identificación. Los números son etiquetas que identifican particularidades o clases. Las estadísticas simples se realizan con datos nominales. Un ejemplo es el género, con las opciones etiquetadas "masculino" o "femenino".

- *Ordinal*: Si en una medición se emplea una escala ordinal, los números reflejan el orden de las personas u objetos. Las medidas ordinales se disponen de mayor a menor o viceversa. Las medidas ordinales revelan una propiedad comparable entre ellas, por ejemplo:

qué persona u objeto es mayor o menor, más brillante u obscuro, más duro o blando, que otro, etc.

Pero tales mediciones no dicen cuánto más alto o más fuerte es uno que el otro. Estadísticamente no puede hacerse mucho más con las medidas ordinales, excepto determinar la mediana y los centiles, así como los coeficientes de correlación de los rangos.

- *Intervalo*: La escala por intervalos proporciona números que reflejan las diferencias entre particularidades. En las escalas por intervalos las unidades de medida son iguales. Con los datos, según una escala por intervalos, se pueden utilizar la media aritmética, la desviación típica y el coeficiente de correlación de Pearson. También se pueden emplear la mayor parte de los contrastes de significación o de hipótesis, como son el contraste de la t de Student y el contraste de la F de Snedecor. Las escalas por intervalos muestran que una persona o particularidad es tantas veces mayor o menor, más pesada o ligera, más brillante u oscura, que otra, etc.

- *Razón*: En esta clasificación se tienen todas las propiedades de escala, de intervalo y además existe un punto cero real en su origen, se llama escala de razón. El cero absoluto o natural representa la nulidad de lo que se estudia. Las escalas de razones, en general son medidas de longitud, peso, capacidad, etc. En las escalas de razones los números reflejan razones entre particularidades y los datos obtenidos según tales escalas pueden ser sometidos a cualquier tratamiento estadístico.

1.2. Modelos para Variables de Respuesta Binaria

Las variables dependientes binarias son muy comunes dentro de las ciencias sociales, y a lo largo de la historia varios autores han estudiado modelos que implican este tipo de variables [10].

Ejemplos de Y una variable respuesta binaria: Y indicando el diagnóstico de algún tipo de cáncer (presente o ausente) en un ser humano, la elección del voto hacia algún tipo de partido político (de izquierda o derecha); cada observación tiene uno o dos resultados, la elección de las personas en el uso de transporte público o privado, entre otros.
Existen diversos modelos para el análisis de variables de respuesta binaria, en esta sección se presentan: El modelo de probabilidad lineal, el modelo probit y el modelo logístico.

1.2.1. Modelo de Probabilidad Lineal

Los modelos de Regresión Lineal son técnicas de gran potencia y versatilidad. Los cuales permiten predecir el comportamiento de una variable dependiente en función de una o más variables independientes y así estimar con precisión la capacidad explicativa del modelo, entre otras muchas ventajas. Pero tiene una restricción importante para las ciencias sociales: sólo se puede utilizar con variables dependientes puramente cuantitativas (de intervalo o de razón). El principal interés en un modelo de respuesta binaria radica en la probabilidad de respuesta.
De acuerdo a J. Scott Long [10] la estructura del modelo de probabilidad lineal aplicado a una variable dependiente binaria es la siguiente

$$y_i = \mathbf{x}_i{}^T \boldsymbol{\beta} + \epsilon_i. \tag{1.1}$$

Donde $\mathbf{x}_i$ es un vector de variables explicativas para la observación i-ésima, $\boldsymbol{\beta}$ es un vector de parámetros y ϵ_i es el error del término.
Si se tiene una sola variable independiente, el modelo se puede escribir como,

$$y_i = \alpha + \beta x_i + \epsilon_i. \tag{1.2}$$

La esperanza condicional de y dado x, $E(y|x) = \alpha + \beta x$, se gráfica como una línea recta continua.

Teniendo en cuenta la $E(y|\mathbf{x})$. Cuando y es una variable aleatoria binaria, la esperanza condicional de y es la probabilidad de que el evento ocurra

$$E(y_i) = [1 \times P(y_i = 1)] + [0 \times P(y_i = 0)] = P(y_i = 1).$$

Para el modelo de regresión,

$$E(y_i|\mathbf{x}_i) = [1 \times P(y_i = 1|\mathbf{x}_i)] + [0 \times P(y_i = 0|\mathbf{x}_i)] = P(y_i = 1|\mathbf{x}_i).$$

Por lo tanto, el valor esperado de y dado $\mathbf{x}$ es la probabilidad de $y = 1$ dado $\mathbf{x}$. Por lo que reescribiendo el Modelo de probabilidad lineal queda:

$$P(y_i = 1|\mathbf{x}_i) = \mathbf{x}_i\boldsymbol{\beta}.$$

La interpretación de los parámetros en este caso es: por cada unidad que incremente x_k, el cambio esperado en la probabilidad de que ocurra el evento es β_k manteniendo las variables restantes constantes. Dado que el modelo es lineal, un cambio unitario en x_k siempre resultará en el mismo cambio en la probabilidad.
Dos de las más importantes desventajas del modelo de Regresión Lineal son, que las probabilidades estimadas obtenidas pueden ser menores a cero o mayores que uno y los efectos parciales de cualquier variable explicatoria es constante.
Mientras que la interpretación de los parámetros no cambia al tener una variable de respuesta binaria, varias suposiciones del modelo son quebrantados. Algunos de los problemas que presenta el Modelo de probabilidad lineal para la estimación de $E(Y|X)$ son [10]:

- *Heterocedasticidad*: Si una variable aleatoria binaria tiene media μ, entonces su varianza es $\mu(1-\mu)$, dado que el valor esperado de y dado $\mathbf{x}$ es $\mathbf{x}\boldsymbol{\beta}$, la varianza condicional de y depende de $\mathbf{x}$ de acuerdo a la ecuación:

$$Var(y|\mathbf{x}) = P(y = 1|\mathbf{x})[1 - P(y = 1|\mathbf{x})] = \mathbf{x}\boldsymbol{\beta}(1 - \mathbf{x}\boldsymbol{\beta}).$$

Lo cual implica que la varianza de los errores depende de las $x's$ y no es constante. Dado que el Modelo de probabilidad lineal es heterocedastico, el estimador por mínimos cuadrados ordinarios de β es ineficiente y los errores estándar son sesgados, resultando incorrectas las pruebas estadísticas.

- *Normalidad*: La distribución normal no describe la distribución de los errores, por lo general es la distribución binomial en la que se basa el análisis de Regresión Logística.

- *Predicciones sin sentido*: Los valores estimados de y en el modelo lineal de probabilidad son negativos o mayores a 1. Dado que la interpretación de $E(y|\mathbf{x})$ como $P(y = 1|\mathbf{x})$, conduce a predicciones de las probabilidades sin ningún sentido.

1.2.2. Modelos Probit y Logit

Para evitar las limitaciones del MPL, se considera una clase de modelos de la forma:

$$P(y = 1|\mathbf{x}) = G(\beta_0 + \beta_1 x_1 + \ldots + \beta_k x_k). \tag{1.3}$$

Donde G es una función que toma valores estrictamente entre cero y uno: $0 < G(z) < 1$, para todos los números reales z. Esto asegura que las probabilidades de respuesta estimadas están estrictamente entre cero y uno.
Existen varias funciones no lineales sugeridas entre ellas, dos de las cuales se encuentran en los modelos: El modelo *Logit* y el modelo *Probit* [19].
En el modelo Logit, G es la función logística

$$G(z) = \frac{exp(z)}{1 + exp(z)}. \tag{1.4}$$

Esta es la función de distribución acumulada para una variable aleatoria logística estándar, la cual está entre cero y uno para todos los números reales z.

En el modelo Probit, G es la función de distribución acumulada normal estándar, la cual se expresa como una integral

$$G(z) = \Phi(z) = \int_{-\infty}^{z} \phi(v)dv. \tag{1.5}$$

Donde $\phi(z)$ es la función de densidad normal estándar con z en los reales.

$$\phi(z) = (2\pi)^{(-1/2)} exp(-z^2/2). \tag{1.6}$$

Esta elección de G asegura que (1.3) está entre cero y uno para todos los valores de los parámetros y variables explicativas.
Las funciones G en (1.4) y (1.5) son ambas funciones crecientes. Cada una crece más rápido, cuando $z \to -\infty$, $G(z) \to 0$, y $G(z) \to 1$, cuando $z \to \infty$.

1.3. Modelo de Regresión Logística

Generalmente, los resultados binarios provienen de una relación no-lineal entre la variable respuesta y las variables independientes del modelo.

La Regresión Logística es un modelo probabilístico, y es una de las técnicas más utilizadas en algunos modelos de Credit Scoring, usando este modelo para la probabilidad de que un sujeto sea merecedor de un crédito. Por ejemplo, para estimar la probabilidad de que un sujeto pague su cuenta a tiempo se pueden utilizar las variables explicativas tales como el tamaño de la cuenta, sus ingresos anuales, ocupación, obligaciones y deudas, porcentaje de la cuenta pagada en tiempo durante el pasado, entre otras características de la historia del aplicante al crédito [1].
Dentro de los principales objetivos del Modelo de Regresión Logística se encuentran:

- Precisar la existencia o ausencia de relación entre una o más variables independientes (x_i) y una variable dependiente dicótomica

(Y),

- Medir el tipo de relación, en caso de que exista y
- Estimar la probabilidad de que se obtenga el suceso definido como "$Y = 1$" en función de los valores de las variables independientes.

La Regresión Logística se basa en la función logística, que expresa una relación entre dos o más variables de forma que a cada elemento de x del conjunto independiente X, le corresponde un único elemento $\pi(x)$ y está representada por:

$$\pi(x) = \frac{1}{1+e^{-x}} = \frac{e^x}{e^x}\frac{1}{1+e^{-x}} = \frac{e^x}{1+e^x}. \tag{1.7}$$

Su gŕafica es una curva S o Sigmoidea, tiene un único punto de inflexión en el que cambia la concavidad y la rapidez del crecimiento, ver la Figura (1.1).

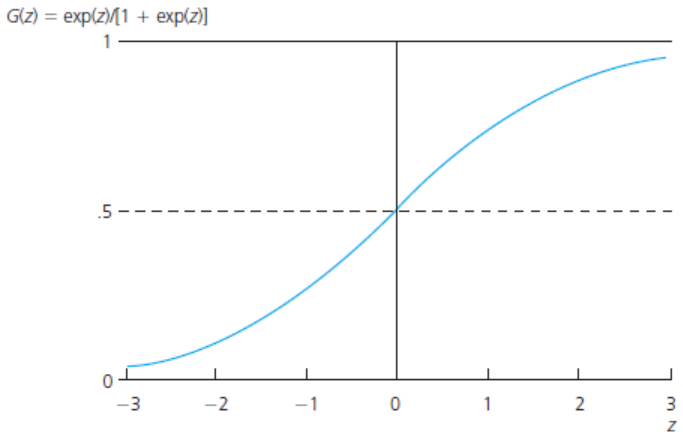

Figura 1.1: Gráfica de la función logística.

1.3.1. Transformación Logit

La transformación logit que proviene de la función logística, es una transformación que tiene ventajas por admitir variables categóricas, además

de tomar valores entre 0 y 1 para la variable dependiente, lo cual se puede asociar a una probabilidad de incumplimiento.
La forma específica del Modelo de Regresión Logística con una sola variable explicativa es

$$\pi(x) = \frac{\exp^{(\beta_0+\beta_1 x)}}{1+\exp^{(\beta_0+\beta_1 x)}}. \tag{1.8}$$

La función logística cuenta con una función inversa llamada *transformación logit* la cual es importante para el desarrollo de la regresión [1]. Obteniendo esta transformación mediante un despeje de variables:
Tomando en cuenta una variación de la ecuación (1.8),

$$\pi(x) = \frac{\exp^{(\beta_0+\beta_1 x)}}{1+\exp^{(\beta_0+\beta_1 x)}} = \frac{1}{1+\exp^{-(\beta_0+\beta_1 x)}}. \tag{1.9}$$

Se obtiene lo siguiente,

$$\begin{aligned}
\pi(x) = \frac{1}{1+\exp^{-(\beta_0+\beta_1 x)}} \\
&\Rightarrow 1+\exp^{-(\beta_0+\beta_1 x)} = \frac{1}{\pi(x)} \\
&\Rightarrow \exp^{-(\beta_0+\beta_1 x)} = \frac{1}{\pi(x)} - 1 \\
&\Rightarrow \exp^{(\beta_0+\beta_1 x)} = \frac{1}{\frac{1-\pi(x)}{\pi(x)}} \\
&\Rightarrow \exp^{(\beta_0+\beta_1 x)} = \frac{\pi(x)}{1-\pi(x)} \\
&\therefore \beta_0+\beta_1 x = \ln\frac{\pi(x)}{1-\pi(x)}.
\end{aligned}$$

Por tanto,

$$logit[\pi(x)] = ln\Big[\frac{\pi(x)}{1-\pi(x)}\Big] = \beta_0+\beta_1 x. \tag{1.10}$$

Al realizar esta transformación, $g(x)$ tiene varias de las propiedades de un modelo de Regresión Lineal. El *logit*, es lineal en sus parámetros, puede ser continua, y su dominio (Valores de x puede estar en un rango de $(-\infty,\infty)$, y codominio (Valores de $\pi(x)$) en el intervalo $(0,1)$, conforme

a la función logaritmo natural, la cual es una función positiva con base en el número racional e, teniendo como único caso posible para que $\frac{\pi(x)}{1-\pi(x)} > 0$, que el numerador y el denominador sean positivos.

1. $\pi(x) > 0$.

2. $1 - \pi(x) > 0 \Rightarrow 1 > \pi(x)$.

Para la función logit y la función logística, cualquier $\pi(x)$ se encuentra dentro del intervalo $(0, 1)$.

Teniendo en cuenta esto, se define la regresión añadiendo un error ϵ y la variable Y, en este caso dicotómica o indicadora de valores cero o uno, donde Y da a $\pi(x)$ una interpretación de probabilidad,

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

Donde ϵ puede tomar uno de dos valores posibles. Si $y = 1$ entonces $\epsilon = 1 - \pi(x)$ con probabilidad $\pi(x)$, y si $y = 0$ entonces $\epsilon = -\pi(x)$ con probabilidad $1-\pi(x)$, por lo que ϵ tiene una distribución con media cero y varianza igual a $\pi(x)[1 - \pi(x)]$.

1.4. Estimación del Modelo de Regresión Logística

Considerando la ecuación (1.8) donde se tiene unicamente una variable dependiente, se debe de desarrollar un método para estimar β_0 y β_1 a partir de una muestra de n observaciones $(y_i, x_i), i = 1, \ldots, n$, donde (y_i, x_i) son las características del i-ésimo individuo de la muestra. En este caso, donde la variable respuesta es dicotómica, se usa el método de máxima verosimilitud para la estimación de los parámetros.
Tomando el valor medio condicionado en estudio:

$$\pi(x_i) = P(Y_i = 1|x_i). \tag{1.11}$$

Donde Y_i es la respuesta asociada a la i-ésima observación, cuya función de densidad es

$$f_i(y_i) = \pi(x_i)^{y_i}(1-\pi(x_i))^{1-y_i}, y_i = 0, 1. \tag{1.12}$$

Y dado que las n observaciones son independientes, la densidad conjunta o la *función de verosimilitud* de $(Y_1, Y_2, \ldots, Y_n)$ queda de la siguiente manera

$$\begin{aligned} l(\beta_0, \beta_1) = f_1(y_1) \times f_2(y_2) \times \cdots \times f_n(y_n) &= \prod_{i=1}^{n} \pi(x_i)^{y_i}(1-\pi(x_i))^{1-y_i} \\ &= \prod_{i=1}^{n} \left(\frac{e^{\beta_0+\beta_1 x_1}}{1+e^{\beta_0+\beta_1 x_1}} \right)^{y_i} \left(1 - \frac{e^{\beta_0+\beta_1 x_1}}{1+e^{\beta_0+\beta_1 x_1}} \right)^{1-y_i}. \end{aligned} \tag{1.13}$$

Este método busca las estimaciones de β_0 y β_1 que maximicen la función de verosimilitud. Para un manejo más fácil de esta ecuación se le aplica logaritmo neperiano, quedando

$$L(\boldsymbol{\beta}) = \ln(l(\beta_0, \beta_1)) = \sum_{i=1}^{n} \left[y_i \ln(\pi(x_i)) + (1-y_i) \ln(1-\pi(x_i)) \right]. \tag{1.14}$$

Para encontrar el valor del vector β que maximiza $L(\beta)$, se deriva $L(\beta)$ con respecto a β_0 y β_1, se igualan las derivadas a 0. Obteniendo las ecuaciones:

$$\sum_{i=1}^{n} [y_i - \pi(x_i)] = 0 \qquad y \qquad \sum_{i=1}^{n} [y_i - \pi(x_i)] x_i = 0. \tag{1.15}$$

Estas expresiones no son lineales en los parámetros β_0 y β_1, por lo que se requieren métodos especiales para su solución [8], utilizando en la actualidad rutinas de programación o paquetes estadísticos, por lo que en este trabajo se utiliza el paquete estadístico SPSS versión 22 [17] para la obtención de los resultados, los valores obtenidos con la solución de las ecuaciones anteriores, se llaman estimadores de máxima verosimilitud y

son denotados por $\hat{\beta}$.

A través de este paquete no solo se obtienen las estimaciones de los coeficientes de regresión, también se tienen sus errores estándar y las covarianzas entre las covariables del modelo.

El próximo paso a seguir es comprobar la significancia estadística de cada uno de los coeficientes de la regresión del modelo, para esto existen dos métodos principales: el estadístico de Wald, el estadístico G de razón de verosimilitud.

El estadístico de Wald: Por definición contrasta la hipótesis de que un coeficiente aislado es distinto de 0, y sigue una distribución normal de media 0 y varianza 1 (Distribución Normal Estándar) [8]. Su valor para un coeficiente en específico viene dado por el cociente entre el valor del coeficiente ($\hat{\beta}_i$) y su correspondiente error estándar $\hat{\sigma}(\beta_i)$.

$$H_0 : \beta_i = 0 \quad vs \quad H_1 : \beta_i \neq 0$$

$$Wald = \frac{\hat{\beta}_i}{\hat{\sigma}(\beta_i)}. \tag{1.16}$$

La obtención de significación indica que dicho coeficiente es diferente de 0 y merece la pena su conservación en el modelo. En modelos con errores estándar grandes, el estadístico de Wald puede proporcionar falsas ausencias de significación (es decir, se incrementa el error tipo II).

El estadístico *G* de razón de verosimilitud: En este método se trata de ir contrastando cada modelo que surge de eliminar cierta cantidad h de variables frente al modelo completo (que incluye las k variables de la muestra). Pudiéndose también aumentar variables con respecto a un modelo inicial que contenga las más significativas.

La valoración se desarrolla mediante el contraste del siguiente juego

de hipótesis:

$$H_0 : \text{Las variables no influyen en el modelo}, \beta_i = 0 \quad \forall i = 1, \ldots, h.$$

$$vs.$$

$$H_1 : \text{Las variables influyen en el modelo}, \beta_i \neq 0 \quad \forall i = 1, \ldots, h.$$

La ausencia de significación implica que el modelo sin la covariable no empeora respecto al modelo completo (es decir, da igual su presencia o su ausencia), por lo que según la estrategia de obtención del modelo más reducido, dicha covariable debe ser eliminada del modelo ya que no aporta nada al mismo.

1.5. Selección de Variables

En la mayoría de los problemas prácticos se tiene un grupo de regresores candidatos, que deberán incluir a todos los factores influyentes, y se debe determinar el subconjunto real de regresores que debe usarse en el modelo. La definición de un subconjunto adecuado de regresores para el modelo es lo que se llama problema de selección de variables.

La construcción de un modelo de regresión que sólo incluya un subconjunto de regresores disponibles implica dos objetivos: 1) Se desea que el modelo incluya tantos regresores como sea posible, para que el contenido de información en ellos pueda influir sobre el valor predicho de y. 2) Se desea que el modelo incluya la menor cantidad de regresores posibles, porque la varianza de la predicción $\hat{y}$ aumenta a medida que aumenta la cantidad de regresores. También, mientras más regresores haya en un modelo, los costos de recolección de datos y los de mantenimiento de modelo serán mayores. El proceso de encontrar un modelo que sea un término medio entre los dos objetivos se llama selección de la "mejor ecuación de regresión, [6]."

Existen varios criterios que se pueden aplicar para evaluar los modelos

de regresión de subconjuntos. El criterio que se usará para seleccionar el modelo se debería relacionar con el uso pretendido del modelo.

Con frecuencia se usan ecuaciones de regresión para predecir observaciones en el futuro, o estimación de la respuesta promedio, en general, se desea seleccionar los regresores de tal modo que el error cuadrático medio de la predicción se reduzca al mínimo, esto suele implicar que se deben eliminar del modelo los regresores con efectos pequeños.

1.6. Evaluación del Modelo

Para realizar la verificación del modelo, se utilizarán diversos estadísticos para probar que variables influyen significativamente.

1.6.1. Medidas de Confiabilidad del Modelo

1. **Devianza:** Es similar a la suma de cuadrados del error de la Regresión Lineal y se define como:

$$D = -2\sum_{i=1}^{n}\left(y_i \ln\left(\frac{\hat{p}}{y_i}\right) + (1-y_i)\ln\left(\frac{1-\hat{p}}{1-y_i}\right)\right). \qquad (1.17)$$

   Si D es mayor que una χ^2 con $(n-p)$ grados de libertad para un nivel de significancia dado, entonces se dice que el modelo logístico es confiable.

2. **Prueba de bondad de ajuste de Hosmer- Lemeshov.** En esta prueba se construyen tablas para comparar los resultados de estimación del modelo contra los resultados reales de la muestra, haciendo la clasificación de éxitos y fracasos para ambos casos.

Las hipótesis a contrastar son:

$$H_0 : \hat{\pi}_j = \frac{exp^{(\beta_0+\beta_1 x)}}{1+exp^{(\beta_0+\beta_1 x)}} \quad \forall j = 1, \ldots, J.$$

$$vs.$$

$$H_1 : \hat{\pi}_j \neq \frac{exp^{(\beta_0+\beta_1 x)}}{1+exp^{(\beta_0+\beta_1 x)}} \quad \text{para algún j.}$$

Se define como:

$$\hat{C} = \sum_{k=1}^{g} \frac{(O_k - n'_k \bar{\pi}_k)^2}{n'_k \bar{\pi}_k (1-\bar{\pi}_k)}. \tag{1.18}$$

Donde:

g es el número de grupos,

n'_k es el número total de observaciones en el k-ésimo grupo,

O_k es la suma de las Y en el k-ésimo grupo y

$\bar{\pi}_k$ es el promedio de las π_k en el k-ésimo grupo.

Si el modelo es correcto, la distribución del estadístico $\hat{C}$ es aproximada a la distribución Chi-cuadrada con $g-2$ grados de libertad, $\chi^2(g-2)$ [8].

1.6.2. Estadísticos Influenciales

Existen distintos tipos de residuales que posibilitan constatar si una observación es influyente o no, los residuales son definidos como la diferencia entre los valores observados y los valores ajustados $(y - \hat{y})$.

Dentro de la Regresión Logística existen diversas maneras para poder medir estas diferencias.

Definiendo al valor ajustado para la j-ésima covariable $\hat{y}_j$, como,

$$\hat{y}_j = m_j \hat{\pi}_j = m_j \frac{e^{\hat{g}(x_j)}}{1+e^{\hat{g}(x_j)}}. \tag{1.19}$$

Donde $\hat{g}(x_j)$ es el logit estimado.

1. **Residuales de Pearson:** Definidos como:

$$r_j = r(y_j, \hat{\pi}_j) = \frac{y_j - m_j\hat{\pi}_j}{\sqrt{m_j\hat{\pi}_j(1-\hat{\pi}_j)}}. \tag{1.20}$$

Donde y_j representa el número de veces que $y = 1$ entre las m_j repeticiones de X_j si los valores de la variable respuesta están agrupadas. Si el modelo es correcto, los residuales de Pearson serán variables de media cero y varianza uno que pueden servir para hacer el diagnóstico del modelo.
El estadístico $\chi_0^2 = \sum_{j=1}^{J} r_j^2$ permite realizar un contraste global de la bondad de ajuste. Se distribuye asintóticamente como una χ^2 con $(J - p - 1)$ grados de libertad.

2. **Residuales de devianza** Definidos como:

$$d_j = \pm \left\{ 2 \left[y_j \ln\left(\frac{y_j}{m_j\hat{\pi}_j}\right) + (m_j - y_j)\ln\left(\frac{(m_j - y_j)}{m_j(1-\hat{\pi}_j)}\right)\right]\right\}^{1/2}. \tag{1.21}$$

Donde el signo, + o −, es el mismo al signo de $(y_j - m_j\hat{\pi}_j)$. La distribución que sigue este estadístico es χ^2 con $(J-(p+1))$ grados de libertad.

3. **Pseudo residuales** El paquete estadístico SPSS [17] ofrece valores de dos pseudo residuales: *R-cuadrado de Cox y Snell* y *R-cuadrado de Nagelkerke*, muy comunes dentro de la Regresión Logística, los cuales son análogos al R-cuadrado de una Regresión Lineal.

   - Cox y Snell:

$$R^2 = 1 - \left(\frac{\hat{L}_c}{\hat{L}_0}\right)^{\frac{2}{N}}. \tag{1.22}$$

   Donde:

   - $\hat{L}_c$ es la función log-verosimilitud del modelo evaluado en $\hat{\beta}_0, \hat{\beta}_1, \ldots, \hat{\beta}_k$.

- $\hat{L}_0$ es la función log-verosimilitud del modelo que solo incluye la constante.

- Nagelkerke: Es la versión corregida de Cox y Snell con valor máximo igual a 1.

$$\bar{R}^2 = \frac{R^2}{R^2_{Max}}. \tag{1.23}$$

Donde: $R^2_{Max} = 1 - L(\hat{\beta}_0)^{\frac{2}{N}}$.

1.6.3. Interpretación de los Coeficientes

El modelo logístico con una variable independiente puede ser escrito como:

$$\ln \Omega(x) = \beta_0 + \beta_1 x. \tag{1.24}$$

Donde

$$\Omega(x) = \frac{P(y=1|x)}{P(y=0|x)} = \frac{P(y=1|x)}{1 - P(y=1|x)}. \tag{1.25}$$

Es la probabilidad (odds) del evento dado x, y el $\ln(\Omega(x))$ es el logaritmo de la probabilidad.
Siendo este cociente de probabilidades de las estimaciones más comunes que se usan para la Regresión Logística.
Y en consecuencia,

$$\frac{\partial \ln \Omega(x)}{\partial x_k} = \beta_k. \tag{1.26}$$

Dado que el modelo es lineal, β_k se interpreta de la siguiente manera:
«Para un cambio unitario en x_k, se espera que el logit cambie por β_k, manteniendo todas las demás variables constantes [10].»
Otra forma de verlo es:
Tomando en cuenta las probabilidades de respuesta que se presentan entre los individuos cuando $x = 1$ la probabilidad está definida como $\frac{\pi(1)}{1-\pi(1)}$, y para los individuos con $x = 0$ similarmente, $\frac{\pi(0)}{1-\pi(0)}$. La razón de probabilidades (odds ratio), se define como la razón entre las proba-

bilidades para $x = 1$ y las probabilidades para $x = 0$ es

$$OR = \frac{\frac{\pi(1)}{1-\pi(1)}}{\frac{\pi(0)}{1-\pi(0)}}. \tag{1.27}$$

Que si se sustituye en la expresión del Modelo de Regresión Logística queda:

Variable Respuesta (Y)	Variable independiente (X)	
	x=1	x=0
y=1	$\pi(1) = \frac{e^{\beta_0+\beta_1}}{1+e^{\beta_0+\beta_1}}$	$\pi(0) = \frac{e^{\beta_0}}{1+e^{\beta_0}}$
y=0	$1-\pi(1) = \frac{1}{1+e^{\beta_0+\beta_1}}$	$1-\pi(0) = \frac{1}{1+e^{\beta_0}}$
Total	1	1

Cuadro 1.1: Valores del Modelo de Regresión Logística cuando la variable independiente es dicotómica.

La Razón de probabilidades (Odds Ratio):

$$OR = \frac{\left(\frac{e^{\beta_0+\beta_1}}{1+e^{\beta_0+\beta_1}}\right) \setminus \left(\frac{1}{1+e^{\beta_0+\beta_1}}\right)}{\left(\frac{e^{\beta_0}}{1+e^{\beta_0}}\right) \setminus \left(\frac{1}{1+e^{\beta_0}}\right)} = \frac{e^{\beta_0+\beta_1}}{e^{\beta_0}} = e^{(\beta_0+\beta_1)-\beta_0} = e^{\beta_1}.$$

Así, para la Regresión Logística con una variable independiente dicotómica con valores 1 y 0, la relación entre las razones de probabilidades y el coeficiente de regresión es

$$OR = e^{\beta_1}.$$

Estos cocientes enumeran el número de veces que será más probable que ocurra un éxito del evento correspondiente con cada variable k.

Teniendo así que la razón de Probabilidades es el cociente entre dos probabilidades asociadas (el obtenido tras realizar el incremento y el anterior al mismo).

1.6.4. Valoración de la Capacidad Predictiva del Modelo

Es de interés en la Estadística clasificar a los individuos dependiendo de que si su probabilidad supera un valor de corte π o no, en particular si el valor de la probabilidad estimada excede a π entonces se tendrá una variable igual a 1, de otra forma será igual a 0; el valor más común para π es 0.5.

$$clasificación = \begin{cases} Probabilidad > \pi \Rightarrow y_e = 1 \\ Probabilidad \leq \pi \Rightarrow y_e = 0. \end{cases}$$

La exactitud de una prueba puede definirse en función de su sensibilidad y especificidad diagnósticadas. Siendo necesario seleccionar un punto de corte o valor límite adecuado que permita resumir los resultados en dos grupos.

La sensibilidad de una prueba se define como la probabilidad de obtener un resultado positivo. Y la especificidad de una prueba indica la probabilidad de obtener un resultado negativo.

1. Clasificación.

		Realidad y_0	
		1	0
Modelo y_e	1	VP	FP
	0	FN	VN

Donde:
VP=Valores Verdaderos Positivos.
FP=Falsos Positivos.
FN=Falsos Negativos.
VN=Verdaderos Negativos.

- Sensibilidad = $\frac{VP}{VP+FN}$.
- Especificidad = $\frac{VN}{VN+FP}$.

- Área bajo la curva ROC (Receiver Operating Characteristic) construida para todos los posibles puntos de corte de π para la clasificación de los individuos.

  La curva ROC es un gráfico en el que se observan todos los pares sensibilidad/especificidad resultantes de la variación continua de los puntos de corte en todo el rango de resultados observados. En el eje y de coordenadas se sitúa la sensibilidad o fracción de verdaderos positivos, en el eje x se sitúa la fracción de falsos positivos o (1-especificidad). El área bajo la curva está dentro de un rango de 0 a 1, otorgando una medida de la capacidad del modelo para discriminar entre los sujetos que experimentan el resultado de interés contra los que no lo hacen.

2. Cálculo del área bajo la curva ROC.

   I. Guardar los valores que predice el modelo.

   II. Calcular la U de Mann - Whitney en relación a los esperados. $AUC = 1 - \frac{U}{n_1 n_2}$, donde n_1 y n_2 son los correspondientes números esperados de "1" o "0".

   La prueba U de Mann-Whitney es una prueba no paramétrica para comprobar la heterogeneidad de dos muestras ordinales, donde el estadístico de prueba se construye a partir de la suma de los rangos de una de las muestras, R_i, elegida arbitrariamente.

   $U_i = n_1 n_2 + \frac{n_i(n_i+1)}{2} - R_i$ donde $i = 1, 2$, [9].

3. Elección del punto de corte óptimo.

   - Debe optimizarse la sensibilidad y especificidad, para después elegir un punto de acuerdo a la naturaleza del modelo predictivo.
   - El cambio en el punto de corte corresponde a emplear diferentes constantes en el modelo logístico.

- Con frecuencia la constante estimada, logra una sensibilidad y especificidad máxima, pero puede no ser el caso.
- Una regla general para la curva ROC es [8]:
  a) Si ROC= 0.5 se sugiere no discriminación.
  b) Si $0.7 \leq ROC < 0.8$ se considera discriminación aceptable.
  c) Si $0.8 \leq ROC < 0.9$, se considera discriminación excelente.
  d) Si $ROC \geq 0.9$ se considera discriminación extraordinaria.

Capítulo 2

Credit Score

El *Scoring* es un método que ha venido evolucionando a lo largo de los años y el interés en su aplicación se basa en calificar a individuos de cualquier población con información propia de cada entidad, posibilitando la aplicación en cualquier mercado.
Esta es una técnica de la minería de datos donde el objetivo es hallar patrones y relaciones con el fin de clasificar; siendo este caso una evaluación crediticia para diferenciar entre clientes cumplidos o incumplidos en cuanto a sus obligaciones de pago.

2.1. ¿Qué son los Credit Scoring?

Los *Credit Scoring* son sistemas que ayudan a determinar si se otorga un crédito o no a nuevos solicitantes dentro de una empresa financiera.
Los *Credit Scoring* de acuerdo a Hand and Henley, son procedimientos estadísticos que se utilizan para clasificar a las personas que gestionan y solicitan un crédito, incluyendo a las que ya son clientes de la institución crediticia en cuestión, en los tipos de riesgo "Bueno" y "Malo".
Scoring se refiere al empleo del conocimiento sobre el desempeño y características de préstamos en el pasado para poder así pronosticar el cumplimiento de préstamos en el futuro [15].

2.2. Ventajas y Desventajas del Scoring

2.2.1. Ventajas del Scoring

Cuantifica el riesgo como una probabilidad

Consistencia: En el proceso de análisis se aplica homogéneamente a todas las solicitudes. Dos personas con las mismas características tendrán el mismo pronóstico de riesgo, sin embargo, podrá variar de acuerdo al analista quien hace la evaluación.

El scoring es explícito: En el scoring estadístico, se conoce y se puede informar el proceso exacto que se utilizó para el pronóstico del riesgo.

Consideración de una amplia gama de factores: Las solicitudes de préstamo se pueden evaluar de manera subjetiva donde se tomaría en consideración ciertas razones financieras y políticas de acuerdo a la institución, pero a diferencia del scoring estadístico, el scoring subjetivo no puede considerar treinta o cincuenta características simultáneamente.

El scoring estadístico puede cuantificar cómo cambiaría el pronóstico de riesgo si una o más variables se modifican ya sea de manera simultánea o de forma aislada. Permitiendo evaluaciones y la administración de riesgo mucho más refinadas.

El scoring estadístico puede probarse antes de usarlo: Una ficha de calificación recién planteada puede probarse para pronosticar el riesgo de los préstamos vigentes en la actualidad, usando únicamente las características conocidas en el momento que se hizo el desembolso. Pudiendo así hacer comparaciones entre el riesgo estimado y el riesgo observado en la práctica, mostrando cómo habría funcionado el scoring si se hubiera aplicado al momento de las solicitudes de los préstamos vigentes.

Revela concesiones mutuas: El scoring muestra lo que el prestamista

puede esperar como consecuencia de implementar diferentes opciones de política, mejorando la administración del riesgo. Por ejemplo, la prueba con información histórica de scoring puede decir a la gerencia de créditos que, de todos los préstamos vigentes alrededor del 8.5 % tienen un riesgo estimado de más del 50 %. De esta manera el scoring indica a la administración que si un prestamista, por ejemplo, adoptara la política de denegar todos los préstamos con más del 50 % de riesgo, se evitarían cierto número de créditos malos.

Por supuesto que el scoring no indica cuál política escoger, pero sí cuales son las probables consecuencias de las diversas opciones, revelando posibles escenarios diferentes a la realidad.

Relación entre el riesgo y las características del prestatario, el préstamo y el prestamista: Por ejemplo, en microfinanzas se tiene el conocimiento de que las mujeres cumplen mejor que los hombres con sus obligaciones financieras. Para un prestamista dado, el scoring:

I. Confirma o desmiente este conocimiento, además de que explica con precisión cuánto más o menos son riesgosas las mujeres.

II. También expone cómo se relaciona el riesgo con el comportamiento del cliente en préstamos anteriores, con el tipo de negocio y ajustes en los términos del contrato de préstamo.

El scoring subjetivo se basa en las creencias que se derivaron de la experiencia y del conocimiento recibido de otras personas, siendo estas correctas o incorrectas, o al menos imprecisas. El scoring estadístico se deriva de las relaciones entre el riesgo y las caraterísticas de los préstamos, a partir de datos históricos de ambos conceptos.

En general el scoring estadístico corrobora la orientación general del juicio subjetivo, por ejemplo, los atrasos en préstamos pasados indican un riesgo mayor de futuros atrasos [15].

2.2.2. Desventajas del Scoring

El scoring estadístico como todos los modelos tiene también varias desventajas. El prestamista que no considere estas desventajas correrá el riesgo de tener un proyecto fracasado por no utilizar de manera adecuada el modelo.
El scoring es una herramienta muy eficaz, pero un mal uso de este puede resultar contraproducente.

La exactitud de los sistemas de scoring sigue siendo una cuestión abierta. La precisión es muy importante en el uso de puntaje de crédito, incluso si el prestamista puede reducir sus costos de evaluar las solicitudes de préstamos mediante el uso de puntajes, si los modelos no son precisos, estos ahorros se consumirán con los préstamos mal realizados. La exactitud de un scoring dependerá del cuidado con el que se desarrolle, los datos sobre los que se basa el sistema deben ser una rica muestra de préstamos bien ejecutados y mal ejecutados, deben estar actualizados y los modelos deben ser reestimados con frecuencia para asegurar que los cambios en las relaciones entre los factores potenciales y el rendimiento del préstamo se capturan.
Si la institución financiera que utiliza el scoring aumenta su grupo de solicitantes mediante la comercialización masiva, debe asegurarse que el nuevo grupo de solicitantes se comporta de manera similar al grupo en que se construyó el modelo. Por lo tanto, el modelo no puede predecir con precisión en el comportamiento de estos nuevos solicitantes.
Deberá tenerse en cuenta no sólo las características de los prestatarios a quienes se les concedió el crédito, sino también de los que fueron denegados, de lo contrario, un "sesgo de selección" en el proceso de aprobación del préstamo podría conducir a un sesgo en los pesos estimados en el modelo de calificación [12].

El Scoring estadístico supone que el futuro será como el pasado

Por ejemplo, un modelo sencillo podría evidenciar que el 10 % de préstamos a agricultores en la base de datos histórica se volvieron

malos y que el 7 % de manufactureros se volvieron malos. Por tanto, si un agricultor aplicara por un préstamo hoy, el modelo pronosticaría un riesgo igual al riesgo histórico. Pero si la base comprende solamente años cuando no hubo sequía, y si este año se da una sequía, el riesgo de los agricultores podría subir astronómicamente. Son necesarios la inteligencia y administración para ajustar el scoring a los cambios en el contexto, la competencia e incluso la política del propio prestamista [15].

El Scoring requiere información de calidad adecuada. Todas las bases de datos tienen información imprecisa o aleatoria, mientras estas perturbaciones no sean demasiado fuertes, el scoring puede captar las señales de riesgo que emiten las características presentes en la base de datos.

El scoring estadístico puede denegar solicitudes pero no puede aprobarlas o modificarlas. A menos que el prestamista tenga información de todas las solicitudes denegadas, el scoring no aplica a toda la población de solicitantes antes de que hayan sido visitados por los analistas de crédito. El scoring compara las solicitudes actuales con las solicitudes históricas que están registradas en la base de datos; en otras palabras, el scoring ignora todos los factores de riesgo que no estén cuantificados ni registrados en la base de datos. Por lo tanto, el scoring no sustituye a los analistas de crédito ni a la evaluación subjetiva personal.

El Scoring funciona con probabilidades, no con certezas. El producto del scoring es un porcentaje, el riesgo pronosticado de que un préstamo se vuelva malo (según la definición del prestamista) antes de que sea cancelado. Aunque el pronóstico es siempre mayor que cero y menor que uno, el riesgo observado en la práctica es siempre cero (no fue malo) o uno (sí fue malo), por lo que el scoring nunca "funciona" para un préstamo dado, solamente funciona en promedio para un grupo de grande de préstamos.

El Scoring estadístico es susceptible al mal uso. El scoring brinda a la administración de la empresa un pronóstico, pero no le indica qué hacer con la información. El abuso más común es el descuido o negligencia, al ignorar el pronóstico y continúan haciendo lo que siempre han hecho, el remedio consiste en la capacitación y seguimiento dentro de la administración.

Otro mal uso es el exceso de anulaciones o excepciones, la decisión de la administración de la empresa de hacer una excepción a la política del uso de scoring. Por ejemplo, si se aprueba un crédito con un pronóstico de 60 % de riesgo de ser malo siendo el umbral de malos del 50 %, dato conocido por los analistas. Es cierto que hay ocasiones que los expertos conocen algo que el modelo ignora, dando como resultado que algunas excepciones son aceptables, sin embargo, hay que dar seguimiento a estas excepciones y comparar su desempeño con lo pronosticado para averiguar quién, en promedio, estaba en lo correcto, los usuarios o el scoring.

Características de los prestatarios, préstamos y prestamistas. La capacidad de pronóstico aumenta con el número de características disponibles. Sin duda, existen rendimientos decrecientes entre mayor sea la cantidad de información, y aún más, el costo marginal de recopilar características adicionales puede ser muy alto [15].

2.3. Modelos Utilizados en el Desarrollo de Sistemas Credit Scoring

Varios métodos estadísticos son usados para desarrollar sistemas de credit scoring incluyendo modelos de probabilidad lineal, modelos logit, modelos probit, modelos de análisis discriminante.

Los primeros tres métodos son técnicas estadísticas estándar para estimar la probabilidad de incumplimiento basada en datos históricos sobre el desempeño del préstamo y las características del prestatario. Estas técnicas difieren en que el modelo de probabilidad lineal asume que hay

una relación lineal entre la probabilidad de incumplimiento y los factores; el modelo logit supone que la probabilidad de incumplimiento es distribuida logísticamente; y el modelo Probit supone que la probabilidad de incumplimiento tiene una distribución normal (acumulativa). El análisis discriminante difiere en que, en lugar de estimar la probabilidad de incumplimiento, divide a los prestatarios en clases de riesgo alto y bajo [12].

Dos métodos más recientes que empiezan a utilizarse para estimar las probabilidades de incumplimiento incluyen, Modelos teóricos del precio de las opciones y la metodología de redes neuronales. Estos métodos tienen el potencial de ser más útiles en el desarrollo de modelos de préstamos comerciales, que tienden a ser más heterogéneos que los préstamos hipotecarios, por lo que los métodos estadísticos tradicionales son más difíciles de aplicar.

La teoría sobre los modelos de precios de opciones comienza con la observación de que la responsabilidad limitada del prestatario es comparable a una opción de venta escrita en los activos del prestatario, con un precio de ejercicio igual al valor de la deuda pendiente, si en algún período futuro, el valor de los activos del prestatario cae por debajo del valor de su deuda pendiente, el prestatario puede incumplir. Los modelos infieren la probabilidad de que una empresa no cumpla con una estimación de la volatilidad de los precios de los activos de la empresa, que generalmente se basa en la volatilidad observada de los precios de las acciones de la empresa.

Las redes neuronales son algoritmos de inteligencia artificial que permiten cierto aprendizaje a través de la experiencia para dicernir la relación entre las características del prestatario y la probabilidad de incumplimiento y determinar qué características son más importantes para predecir el incumplimiento. Es un método más flexible que las técnicas estadísticas habituales, ya que se puede no hacer suposiciones sobre la forma funcional de la relación entre las características y la probabilidad de incumplimiento, o sobre las distribuciones de las variables o errores del modelo, y las correlaciones entre las características no se contabilizan.

Algunos argumentan que las redes neuronales muestran mucha promesa en la puntuación de crédito para los préstamos comerciales, pero otros han argumentado que el enfoque es más ad hoc que el de los métodos estadísticos estándar [11].

Capítulo 3

Caso práctico: Análisis de Datos

3.1. El Sistema Financiero y la Economía Alemana en 1994

En general se entiende que, el sistema financiero de un país está formado por el conjunto de instituciones, mercados y medios, cuyo fin principal es dirigir el ahorro que generan los prestamistas hacia los prestatarios.
El sistema financiero alemán está constituido por el Banco Central que opera en conjunto con once bancos centrales provinciales.
Las entidades de crédito se diferencian entre sí, por su estructura operativa, organización, forma jurídica y/o dimensión, pero en general los bancos independientemente sean entidades privadas, cooperativas o entidades de derecho público, realizan toda clase de operaciones habituales concebibles.
La Banca comercial privada actúa con el carácter propio de entidades universales, tomando depósitos sin límite de importe y a diferentes plazos y conceden créditos de cualquier magnitud, a corto, medio y largo plazo. En los bancos comerciales privados predominan las operaciones

de crédito a corto plazo [14].

3.2. Contexto Histórico

EL sistema financiero y la economía alemana han estado definidos en los últimos años, por el proceso de reunificación.

> "El impacto expansivo del mismo sobre las condiciones económicas, monetarias, financieras y fiscales de Alemania ha sido enorme. Sus efectos son consecuencia de la envergadura del fenómeno y de la forma en que se ha financiado [...]. Los resultados finales se reflejaron en dos variables fundamentales, presiones inflacionarias y desequilibrio de la balanza de pagos [13]."

En 1992 se puso en marcha una nueva estructura del Banco Central Alemán para poder adecuarse a la reunificación de las dos Alemanias existentes en ese momento. Forjandose así la base de la oferta de recursos financieros en el mercado alemán dentro del sector de las familias en esos años.
El endeudamiento de las familias alemanas destacó por su bajo nivel, el comportamiento de las familias alemanas se describió al tener escasa dependencia del crédito bancario, procurando autofinanciar en gran medida sus compras de bienes de consumo duradero e, incluso, de vivienda; en este último caso, a través de la acumulación de depósitos en sociedades de crédito hipótecario entre otras instituciones, con anticipación al momento de la compra de la vivienda, los ahorros financieros netos medios de las familias alemanas durante los años 1985 a 1994, fue estable, siendo un rasgo positivo para la estabilidad de los mercados financieros alemanes y para la eficacia de las políticas macroeconómicas [13].

3.3. Descripción de la Base de Datos

La base de datos German Credit con la que se trabaja en este estudio, consiste en la información de 1000 personas solicitantes de un crédito, contenida en la medición de 20 variables para cada individuo. Cada solicitante ha sido clasificado dentro de una de las dos posibles categorías, "Buen crédito" (700 casos) o "Crédito Malo" (300 casos).

Se desarrolla una regla de credit scoring para determinar si un nuevo solicitante es "Bueno" o "Malo" cliente, basándose en los valores de una o más variables explicativas resultantes del modelo final. Las variables a considerar son descritas a continuación:

No.	Nombre de la variable	Descripción	Tipo de variable	Descripción en el código
1	Clase	Clasificación de los solicitantes	Categórica (Binaria)	▪ 1 = Malo ▪ 0 = Bueno
2	Balance_ Cuenta	Balance de cuenta	Categórica	▪ 1 = Menor a 0 DM ▪ 2 = Entre 0 y 200 DM ▪ 3 = Mayor o igual a 200 DM ▪ 4 = No tiene cuenta
3	Duración_ Crédito_ Meses	Duración del Crédito contado en meses	Continua	
4	Historia_ Crediticia	Historial crediticio por cada cliente	Categórica	▪ 0 =No tiene créditos tomados o todos los créditos pagados debidamente ▪ 1 =Todos los créditos de este banco pagados debidamente ▪ 2 = Créditos existentes debidamente pagados hasta ahora ▪ 3 = Retraso en el pago en el pasado ▪ 4 =Cuenta crítica
5	Propósito	Propósito por el que se obtuvo el crédito	Categórica	▪ 0 =Carro (Nuevo) ▪ 1 =Carro (Usado) ▪ 2 = Muebles\Equipo ▪ 3 = Radio\Televisión ▪ 4 =Aparatos domésticos ▪ 5 =Reparaciones ▪ 6 =Educación

No.	Nombre de la variable	Descripción	Tipo de variable	Descripción en el código
				▪ 7 =Vacaciones ▪ 8 =Capacitación ▪ 9 =Negocios ▪ 10 =Otros
6	Monto_Crédito	Monto del crédito otorgado	Numérica	
7	Cuenta_Ahorros	Monto de la cuenta de ahorros del cliente	Categórica	▪ 1 =Menor a 100 DM ▪ 2 =Entre 100 y 500 DM ▪ 3 =Entre 500 y 1000 DM ▪ 4 =Mayor a 1000 DM ▪ 5 =Monto desconocido\No tiene cuenta de ahorro
8	Duración_en_trabajo	Años en el trabajo actual	Categórica	▪ 1 =Desempleado ▪ 2 =Menos de 1 año ▪ 3 =Entre 1 y 4 años ▪ 4 =Entre 4 y 7 años ▪ 5 =Mayor o igual a 7 años
9	Tasa_De_Crédito	Tasa de Crédito	Numérica	
10	Género_Edo Civil	Categorías en las que se clasificaron a los clientes de acuerdo a su género y estado civil	Categórica	▪ 1 =Hombre Divorciado\Separado ▪ 2 =Mujer Divorciada\Separada\Casada ▪ 3 =Hombre soltero ▪ 4 =Hombre Casado\Viudo ▪ 5 =Mujer Soltera
11	Otros_Deudores_Fiadores	Tipo de personas que entraron igualmente con el cliente dentro del contrato	Categórica	▪ 1 =Ninguno ▪ 2 =Co-Solicitante ▪ 3 =Fiador
12	Duración_Residencia	Años viviendo en su residencia actual	Numérica	
13	Propiedades	Tipo de propiedades importantes disponibles con las que cuenta el cliente	Categórica	▪ 1 =Bienes raíces ▪ 2 =Contrato Ahorro de Vivienda\Seguro de vida ▪ 3 =Carro (Diferente al del campo de Propósito ▪ 4 =Desconocido\Sin propiedad
14	Edad	Edad a la que el cliente solicitó el crédito	Numérica	

No.	Nombre de la variable	Descripción	Tipo de variable	Descripción en el código
15	Otros_Planes_Pago	Otro tipo de pagos que el cliente realice a la par con el crédito	Categórica	▪ 1 =Bancario ▪ 2 =Tiendas departamentales ▪ 3 =Ninguno
16	Tipo_Vivienda	Tipo de vivienda en la que reside el cliente	Categórica	▪ 1 =Rentada ▪ 2 =Propia ▪ 3 =Libre
17	NoCréditos_ Banco	Número de créditos en este banco	Numérica	
18	Ocupación	Tipo de trabajo en el que se desenvuelve el cliente	Categórica	▪ 1 =Desempleado\Incapacitado\No residente ▪ 2 =Incapacitado- Residente ▪ 3 = Empleado Capacitado\Oficial ▪ 4 = Gerente\Por cuenta propia\Altamente calificado\Oficial
19	Personas_ Dependientes	Número de personas que dependen del cliente	Numérica	
20	Telefono	¿Tiene el cliente número de telefono registrado bajo su nombre?	Binaria	▪ 1 =No ▪ 2 =Si
21	Trabajador Foráneo	Si es o no trabajador foráneo	Binaria	▪ 1 =Si ▪ 2 =No

Cuadro 3.1: Variables de la base de datos German Credit.

3.4. Definición de la Variable Respuesta y las Variables Explicativas

Como se puede observar en la tabla anterior, dentro de la base de datos se encuentran variables socioeconómicas:

- Edad
- Estado civil
- Género
- Cantidad de personas que dependen del acreditado
- Tiempo de permanencia en el actual domicilio
- Tiempo de permanencia en el empleo actual.
- Si es propietario de la vivienda que habita.
- Tipo de ocupación
- Si tiene una cuenta de ahorros y a cuánto asciende.
- Si cuenta con algún teléfono a su nombre.
- Si es trabajador foráneo

Además de variables que describen el comportamiento e historial de cada cliente, entre las que se encuentran:

- Monto del crédito otorgado
- Tipo de crédito otorgado

Se muestran a continuación gráficas y tablas de la relación de algunas variables con respecto a la clasificación de “Buenos” y “Malos”:

	Clase		Total
	Malo	Bueno	
Hombre: Divorciado/soltero	20	30	50
Mujer: Divorciada/Separada/Casada	109	201	310
Hombre:Soltero	146	402	548
Hombre: Casado/Viudo	25	67	92
Total	300	700	1000

Cuadro 3.2: Género-Estado Civil * Clase.

El estado civil combinado con el género (variable ya configurada en la base de datos de esa manera) de acuerdo a clientes "Buenos" y "Malos", se detalla en el Cuadro 3.2, los datos se concentran en las categorías Mujer: Divorciada/Separada/Casada, siendo el doble en clientes "Buenos" y también en Hombre: Soltero siendo en mayor cantidad en los clientes "Buenos".

	Mínimo	Máximo	Media
Duración del Crédito (Mensual)	4	72	20.90
Monto de Crédito	250	18,424	3,271.25
Tasa de Crédito	1	4	2.97
Duración en dirección actual (Años)	1	4	2.85
Edad (Años)	19	75	35.54
Número de créditos en este banco	1	4	1.41
Número de dependientes	1	2	1.15

Cuadro 3.3: Estadísticos descriptivos de las variables de escala.

Se detalla en el Cuadro 3.3 las variables Duración del Crédito donde el crédito más reciente que se tiene es de 4 meses, y el más antiguo de 72 meses; el Monto de Crédito va desde 250 Marcos alemanes hasta 18,424 Marcos alemanes; la Tasa de Crédito del 1 % al 4 %; la variable Duración en dirección actual se encuentra en años siendo el mínimo de 1 año, y con un máximo de 4 años; la edad de los clientes valorada en años se encuentra dentro del intervalo de 19 años a 75 años, con una media de 35.54 años, el Número de créditos en este banco va desde 1 a 4 créditos;

y el número de dependientes con los que cuenta el cliente son de 1 a 2 personas.

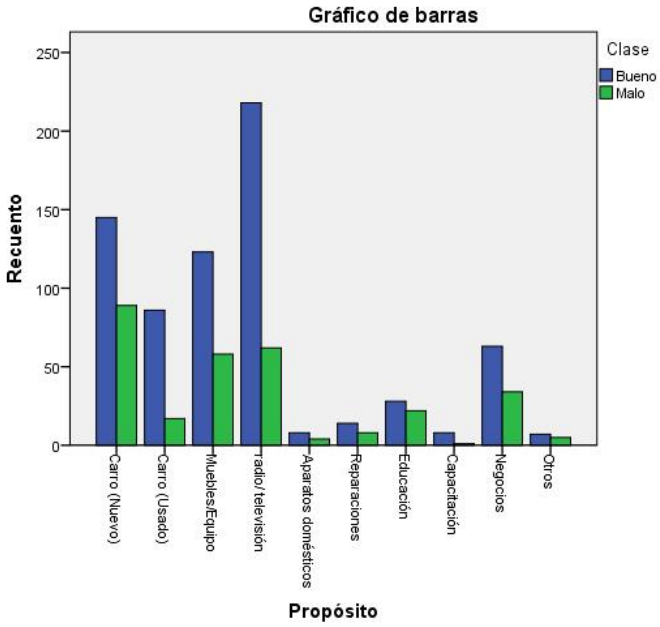

Figura 3.1: Propósito del crédito.

En la Figura 3.1 se observa la característica sobre cual fue el "Propósito del crédito" para cada cliente, de acuerdo a la división de clientes, dentro de los clientes "Buenos" los propósitos se concentran más en la obtención de un Carro (Nuevo), Muebles, y Radio/Televisión y dentro de los "Malos" aunque en menor medida pero sobresale el propósito de Carro (Nuevo).

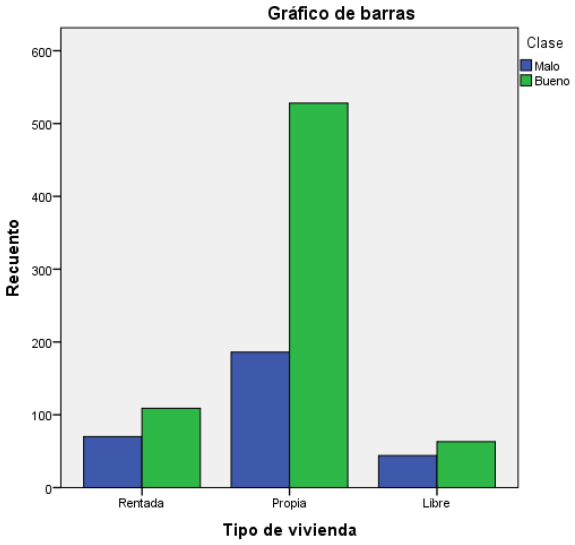

Figura 3.2: Tipo de vivienda*Clase.

En la Figura 3.2 se muestra la variable Tipo de vivienda la cual cuenta con 3 categorías (Rentada, Propia, Libre), teniendo la mayor concentración de los datos en la categoría de Propia para ambas clases, pero con mayor proporción dentro de los clientes "Buenos".

	Clase		Total
	Malo	Bueno	
Desempleado/Incapacitado/No Residente	7	15	22
Incapacitado/Residente	56	144	200
Empleado Capacitado/Oficial	186	444	630
Gerente/Por cuenta propia/Altamente Calificado	51	97	148
Total	300	700	1000

Cuadro 3.4: Tipo de trabajo * Clase.

La variable Tipo de trabajo se presenta en el Cuadro 3.4, esta variable cuenta con 4 categorías, teniendo los datos mayor concentración en la categoría de Empleado Capacitado/Oficial para ambas clases.

	Clase		Total
	Malo	Bueno	
Desempleado	23	39	62
Menos de 1 año	70	102	172
Entre 1 y 4 años	104	235	339
Entre 4 y 7 años	39	135	174
Mayor o igual a 7 años	64	189	253
Total	300	700	1000

Cuadro 3.5: Duración en el trabajo actual * Clase.

En el Cuadro 3.5 se muestra la variable Duración en el trabajo actual, la cual contiene 5 categorías, en los clientes "Malos" se tiene mayor concentración en la categoría de Entre 1 y 4 años, seguido por Menos de un año, dentro de los clientes "Buenos" la mayor concentración de datos está en la categoría Entre 1 y 4 años, pero a esta categoría le sigue la de Mayor o igual a 7 años.

	Clase		Total
	Malo	Bueno	
No tiene créditos tomados/Todos los créditos pagados debidamente	25	15	40
Todos los créditos de este banco pagados debidamente	28	21	49
Créditos existentes debidamente pagados hasta ahora	169	361	530
Retraso en el pago en el pasado	28	60	88
Cuenta crítica/Otros créditos existentes (No en este banco)	50	243	293
Total	300	700	1000

Cuadro 3.6: Estado de pagos anteriores * Clase.

En el Cuadro 3.6 está la variable Estado de pagos anteriores, la cual tiene 5 categorías, y tanto los clientes "Malos" como los "Buenos" se encuentra más de la mitad de ellos dentro de los Créditos existentes debidamente pagados hasta ahora.

	Clase		Total
	Malo	Bueno	
Bancario	57	82	139
Tiendas departamentales	19	28	47
Ninguno	224	590	814
Total	300	700	1000

Cuadro 3.7: Créditos Simultáneos * Clase.

En el Cuadro 3.7 se muestra la variable Créditos Simultáneos que llegasen a tener los clientes, la cual tiene 3 categorías, siendo cerca del 75 % para los clientes "Malos" dentro de la categoría de Ninguno; y más del 75 % para los clientes "Buenos" igualmente para la categoría de Ninguno.

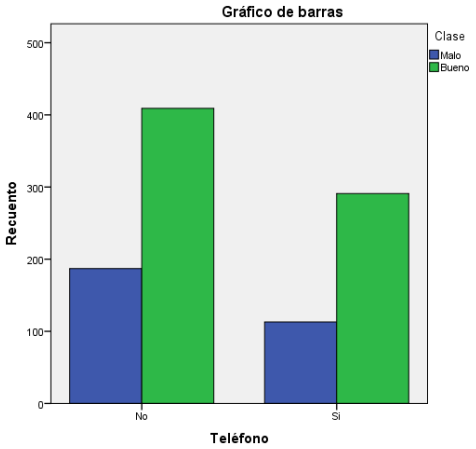

Figura 3.3: Teléfono * Clase.

En la Figura 3.3 se muestra la variable Teléfono, en la cual se especifica si el cliente cuenta con un teléfono bajo su nombre o no, siendo el caso de que predomina para ambas clases el que No cuenta con un teléfono a su nombre.

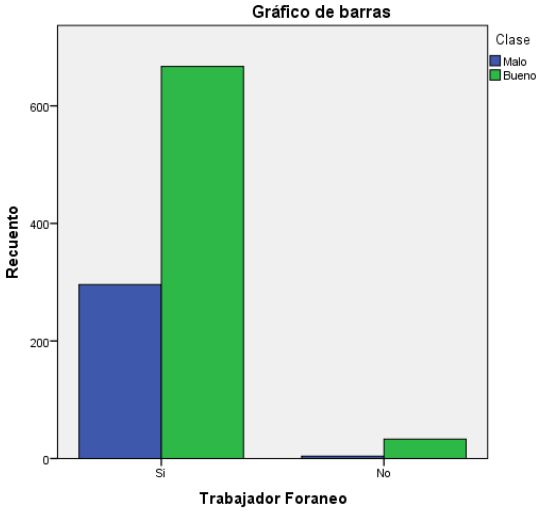

Figura 3.4: Trabajador Foráneo * Clase.

En la Figura 3.4 está la variable Trabajador Foráneo, la cual como su nombre lo indica, detalla si el cliente es o no trabajador foráneo, siendo el caso para esta base de datos de que en su mayoría para ambas clases el que Si sean trabajadores foráneos.

3.5. Selección de Variables Aplicadas al Modelo

Variable Dependiente

La variable dependiente del modelo es llamada *Clase*, la cual es una variable dicotómica, creada para hacer referencia a los clientes cumplidos "Buenos" con valor de 0 y a los clientes incumplidos "Malos" con valor de 1.

Para la construcción del modelo se divide la base en dos partes: Primero se toma una muestra aleatoria del 70 % llamada de *entrenamiento* con la cual se construye el modelo, y con el 30 % restante se conforma la base de *validación* y se usa para evaluar los resultados del modelo obtenido.

Variables independientes

Las variables independientes seleccionadas de acuerdo a las características de la base de datos de entrenamiento, son:

1. Duración de crédito:
   Esta variable se refiere a los meses que hasta el momento de la conformación de la base ha estado activo el crédito. Es de caracter cuantitativo.

2. Monto de crédito:
   Variable numérica que expresa el monto total del crédito otorgado.

3. Tasa de crédito:
   Variable numérica que expresa la tasa del crédito.

4. Balance de cuenta:
   Variable categórica que expresa el estado de la cuenta corriente de cada cliente, tiene cuatro categorías:

   - 1 = Menor a cero DM.
   - 2 = Entre 0 y 200 DM.
   - 3 = Mayor o igual a 200 DM.
   - 4 = No existe cuenta.

5. Historia Crediticia: Cuenta con 5 categorías en donde se muestra si el cliente ha tenido otros créditos.

   - 0 = No tiene créditos tomados/Todos los créditos pagados debidamente.
   - 1 = Todos los créditos de este banco pagados debidamente.
   - 2 = Créditos existentes debidamente pagados hasta ahora.
   - 3 = Retraso en el pago en el pasado.
   - 4 = Cuenta crítica/Otros créditos existentes (No en este banco)

6. Propósito: Tiene 11 categorías en las cuales se clasificó el propósito por el cual el cliente solicitó un crédito.
   - 0 = Carro (Nuevo).
   - 1 = Carro (Usado).
   - 2 = Muebles/Equipo.
   - 3 = Radio/Televisión.
   - 4 = Aparatos domésticos.
   - 5 = Reparaciones.
   - 6 = Educación.
   - 7 = Vacaciones.
   - 8 = Capacitación.
   - 9 = Negocios.
   - 10 = Otros.
7. Cuenta de ahorros: Cuenta con 5 categorías:
   - 1 = Menor a 100 DM.
   - 2 = Entre 100 y 500 DM.
   - 3 = Entre 500 y 1000 DM.
   - 4 = Mayor a 1000 DM.
   - 5 = Monto Desconocido /No tiene cuenta de ahorro.
8. Duración en el trabajo:
   - 1 = Desempleado.
   - 2 = Menos de 1 año.
   - 3 = Entre 1 y 4 años.
   - 4 = Entre 4 y 7 años.
   - 5 = Mayor o igual a 7 años.
9. GéneroEdoCivil:

- 1 = Hombre: Divorciado/Separado.
- 2 = Mujer: Divorciada/Separada/Casada.
- 3 = Hombre: Soltero.
- 4 = Hombre: Casado/ Viudo.
- 5 = Mujer: Soltera.

10. Otros planes de pago: Tipo de crédito simultáneo.

    - 1 = Bancario.
    - 2 = Tiendas departamentales.
    - 3 = Ninguno.

3.6. Estimación del Modelo en SPSS

Para construir el modelo se utiliza la base de entrenamiento, teniendo como variable dependiente a la variable *Clase*, para los clientes incumplidos está la etiqueta de *Malos* con un valor de 1, y para los clientes cumplidos como *Buenos* con el valor de 0; incluyendo la lista de variables independientes.
Se selecciona un método para la introducción de variables en el modelo, por lo cual el programa ofrece diferentes casos: Método hacia adelante (forward), hacia atrás (backward) o de inclusión total (enter), donde se coloca la totalidad de las variables; estos métodos se eligen bajo dos criterios: Bajo el estadístico de Wald o Devianza (LR).
Para el modelo en estudio se seleccionó el método Backward: LR para encontrar un modelo que tuviera variables con nivel de significancia menor al 5 %, este método inicia incluyendo todas las variables, en este caso 10 variables.

3.6.1. Ajuste del Modelo

El ajuste del modelo fue evaluado con el estadistístico Hosmer-Lemeshow. Esta prueba se usa para evaluar la hipótesis nula de proximidad entre la probabilidad de los valores observados contra la probabilidad de los valores estimados en cada paso de cambio del modelo.
La Figura 3.5, muestra en cada escalón los valores obtenidos, se observan tres escalones lo que representa que durante la construcción del modelo hubo 3 casos en los que hubo una inclusión y/o eliminación de variables afectando al modelo.
En este caso se obtuvo un nivel de bondad de ajuste del 67.5 % en el tercer y último escalón. Con ello podemos decir que tiene un buen ajuste el modelo.

Prueba de Hosmer y Lemeshow

Escalón	Chi-cuadrado	gl	Sig.
1	2.772	8	.948
2	4.046	8	.853
3	5.750	8	.675

Figura 3.5: Prueba de Hosmer y Lemeshow.

3.6.2. Poder Predictivo

El poder predictivo del modelo es la capacidad que tiene de predecir la variable dependiente; sustentado en los valores de las variables independientes.
Uno de los estadísticos que evalúan el poder predictivo es la R^2, en este caso el paquete estadístico ofrece dos tipos de R^2 análogos del Modelo de Regresión Lineal.

Resumen del modelo

Escalón	Logaritmo de la verosimilitud -2	R cuadrado de Cox y Snell	R cuadrado de Nagelkerke
1	641.436[a]	.263	.373
2	642.168[a]	.262	.372
3	648.947[a]	.255	.362

a. La estimación ha terminado en el número de iteración 5 porque las estimaciones de parámetro han cambiado en menos de .001.

Figura 3.6: Tabla de R^2 del modelo.

En la Figura 3.6 se detalla que en el modelo se tuvieron 3 cambios y finalizando con 5 iteraciones. El último valor de determinación fue de 0.362, explicando entre el 0.255 y el 0.362 de la variable dependiente, es decir, la variables *Clase* queda explicada en un rango entre 25.5 % y el 36.2 % por las variables explicativas del modelo.

3.6.3. Clasificación

La tabla de clasificación indica paso a paso la clasificación de clientes cumplidos (Buenos) e incumplidos (Malos). En ella se pueden ver el total de proporciones correctamente clasificadas en cada uno de los grupos. En este caso, como se ve en la Figura 3.7, se obtuvo un total de 79.1 % de clasificaciones correctas con un punto de corte óptimo de 0.55 el cual pudo ser modificado, sin embargo resultó ser óptimo por mantener una clasificación de clientes incumplidos mayor al 90 %.

Tabla de clasificación[a]

			Pronosticado		
			Clase		Corrección de porcentaje
	Observado		Bueno	Malo	
Paso 1	Clase	Bueno	454	36	92.7
		Malo	118	92	43.8
	Porcentaje global				78.0
Paso 2	Clase	Bueno	457	33	93.3
		Malo	120	90	42.9
	Porcentaje global				78.1
Paso 3	Clase	Bueno	455	35	92.9
		Malo	111	99	47.1
	Porcentaje global				79.1

a. El valor de corte es .550

Figura 3.7: Tabla de Clasificación.

Para comprobar que el punto de corte de 0.55 fue óptimo se obtuvieron las clasificaciones en caso de que este punto tuviera valores alternativos. La especificidad y la sensibilidad fueron utilizadas para el cálculo, ya que muestran las proporciones de clasificación.

Punto de corte	Pasos	Sensibilidad	Especificidad	1−Especificidad
0.55	Paso 1	92.65 %	43.81 %	56.19 %
	Paso 2	93.27 %	42.86 %	57.14 %
	Paso 3	92.86 %	47.14 %	52.86 %
0.5	Paso 1	90.61 %	52.38 %	47.62 %
	Paso 2	90.20 %	52.86 %	47.14 %
	Paso 3	89.80 %	50.48 %	49.52 %
0.45	Paso 1	86.73 %	56.67 %	43.33 %
	Paso 2	86.33 %	57.62 %	42.38 %
	Paso 3	86.53 %	56.67 %	43.33 %
0.40	Paso 1	83.27 %	63.33 %	36.67 %
	Paso 2	83.47 %	63.33 %	36.67 %
	Paso 3	83.27 %	62.86 %	37.14 %

Cuadro 3.8: Valores de la Sensibilidad y Especificidad obtenidas en cada punto de corte evaluado.

De acuerdo al Cuadro 3.8 el punto de corte 0.40 sería el adecuado si se buscara que la clasificación correcta de clientes cumplidos fuera mayor al 60 % aunque la sensibilidad fuera menor al 85 %.

En cambio si se busca una cantidad de clientes cumplidos (Especificidad igual al 50 %) para tomar una cantidad aceptable de oportunidades posibles y una sensibilidad menor al 90 % para la correcta clasificación de clientes incumplidos, el valor del corte adecuado sería del 0.5.

Por lo que se eligió el punto de corte de 0.55 ya que aunque se arriesga la clasificación correcta por debajo del 50 % de incumplimientos, no importando las posibles ganancias que no serían tomadas por la proporción de clasificación correcta de estos pero se está asegurando una correcta clasificación de clientes incumplidos al ser mayor del 90 % la Sensibilidad.

Lo ideal sería conseguir un punto medio, para minimizar la proporción de pérdidas en ambos casos y dando prioridad a mantener la menor cantidad de clientes incumplidos clasificados incorrectamente.

3.6.4. Poder Discriminatorio

Es la capacidad que tiene el modelo para poder clasificar de manera correcta a los préstamos.
La curva ROC (Receiver Operating Characteristic) brinda una representación gráfica del poder discriminatorio de un sistema de scoring, su gráfica se muestra en la Figura 3.8.

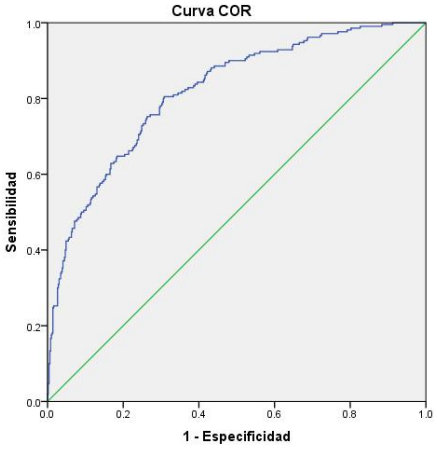

Figura 3.8: Gráfica de curva ROC.

Se obtuvo un área bajo la curva igual a 0.817 como lo muestra la Figura 3.9, esta área significa que para dos préstamos, uno seleccionado aleatoriamente del grupo de malos y otro elegido al azar del grupo de buenos, el préstamo malo presentará un riesgo mayor al bueno el 81.7 % de las veces.

Área bajo la curva

Variable(s) de resultado de prueba: Probabilidad pronosticada

Área	Error estándar[a]	Significación asintótica[b]	95% de intervalo de confianza asintótico	
			Límite inferior	Límite superior
.817	.017	.000	.783	.851

a. Bajo el supuesto no paramétrico

b. Hipótesis nula: área verdadera = 0,5

Figura 3.9: Área bajo la curva ROC.

Y de acuerdo a la regla general, con este valor se considera una discriminación excelente.

3.6.5. Interpretación

Por último, ya es posible realizar el cálculo de la probabilidad de incumplimiento a través de la ecuación de Regresión Logística y los valores estimados de sus coeficientes junto con los valores de OR:

Variable	Coeficiente (c)	$OR = Exp(c)$	$Coef = \ln(OR)$
Balance_Cuenta			
Balance_Cuenta(1)	1.645	5.18	1.645
Balance_Cuenta(2)	1.353	3.87	1.353
Balance_Cuenta(3)	0.566	1.76	0.566
Historia_Crediticia			
Historia_Crediticia(1)	1.436	4.20	1.436
Historia_Crediticia(2)	1.657	5.24	1.657
Historia_Crediticia(3)	0.767	2.15	0.767
Historia_Crediticia(4)	0.767	2.15	0.767
Monto_Crédito	0.000	1	0.000
Tasa_Crédito	0.371	1.45	0.371
Propósito			
Propósito(1)	2.002	7.40	2.002
Propósito(2)	0.209	1.23	0.209
Propósito(3)	0.953	2.59	0.953
Propósito(4)	1.024	2.78	1.024

Propósito(5)	2.055	7.81	2.055
Propósito(6)	0.774	2.17	0.774
Propósito(7)	1.721	5.59	1.721
Propósito(8)	-0.05	0.95	-0.05
Propósito(9)	1.141	3.13	1.141
Cuenta_Ahorros			
Cuenta_Ahorros(1)	1.012	2.75	1.012
Cuenta_Ahorros(2)	0.458	1.58	0.458
Cuenta_Ahorros(3)	0.832	2.30	0.832
Cuenta_Ahorros(4)	-0.503	0.60	-0.503
Género_EdoCivil			
Género_EdoCivil(1)	0.548	1.73	0.548
Género_EdoCivil(2)	0.278	1.32	0.278
Género_EdoCivil(3)	-0.438	0.65	-0.438
Duración_Crédito_Meses	0.032	1.03	0.032
Constante	-6.608	0.00135	-6.608

Cuadro 3.9: Coeficientes estimados.

Con el Cuadro 3.9 se puede indicar que la Variable *Propósito* (Variable que se evalúa con variables dummies para indicar el propósito por el cual fue la solicitud del préstamo de cada cliente) es aquella que tiene más poder al momento de la evaluación, sobre todo al tratarse de clientes cuyo propósito son 'Reparaciones', el cual tiene un aumento de 7.81 veces en la probabilidad de incumplimiento. A esta variable le sigue en términos de relevancia la variable *Historia_Crediticia* en la categoría 'Créditos existentes debidamente pagados hasta ahora', que aumenta la probabilidad de incumplimiento 5.24 veces.

3.6.6. Validación

El scoring estadístico tiene la capacidad de ser probado antes de usarse. Este procedimiento expone como funciona el scoring si se aplicara en el presente. La validación se realiza con una muestra no utilizada para construir el modelo. Para validar el modelo se usó una muestra denominada *Muestra de validación* que se extrajo de la muestra original, siendo el 30 % del total de datos, la muestra es aleatoria y se asegura que el 30 % de los datos contiene una proporción similar de buenos y malos como la muestra del 70 %. Cuando se estimaron los coeficientes se aplica el modelo a esta muestra con el mismo punto de corte. Los resultados revelaron una sensibilidad del 92.86 % y una especificidad del 42.22 %, con una clasificación total correcta del 77.7 %

Conclusiones

En la actualidad es importante contar con un basto conocimiento de los riesgos y las diferentes metodologías que existen para su medición, teniendo como fin la mejora en la operación crediticia, dentro de este análisis se logró realizar el Modelo de Regresión Logística de credit scoring, para así divulgar el método, la manera en que se plantea y realiza el modelo, tomando en cuenta sus ventajas al no requerir el supuesto de normalidad y por calcular directamente las probabilidades de incumplimiento.

Se realizó esta técnica con la ayuda de la base de datos alemana que se encuentra disponible en la red, siendo conformada por una muestra con 1000 observaciones de clientes, con 20 variables originalmente.

De las 20 variables explicativas, se encontró que 10 únicamente eran las más significativas. Y mediante el criterio de selección Backward el mejor modelo ajustado quedó con las siguientes variables:

- Balance de cuenta.
- Historia crediticia.
- Monto de crédito.
- Tasa de crédito.
- Propósito.
- Cuenta de ahorros.
- Género-Estado Civil.

- Duración del crédito.

El criterio de Hosmer-Lemeshov presenta un p-valor de 0.675, concluyendo un buen ajuste. No obstante se obtuvo un bajo poder predictivo, evaluado por una R^2 igual a 0.362.
El área bajo la curva fue de 0.817, y por regla general del poder discriminatorio, se considera que el modelo tiene una discriminación excelente.
La manera de definir el punto de corte fue buscando tener una clasificación correcta de clientes malos mayor al 90 %. La sensibilidad declara que de los 210 préstamos malos en la muestra, el modelo detectó el 92.86 % de ellos.
La validación del modelo fue realizada con el 30 % de la base original. El modelo detectó el 71.7 % de los préstamos malos, y el 77.7 % de los registros de esta muestra fue clasificado correctamente. La discriminación es buena y puede mejorarse jugando con los datos, alternando entre la construcción del modelo y la validación, mejorando la definición de categorías en algunas variables e incluyendo variables que influyan en el riesgo, sugeridas por los expertos.
También haciendo énfasis en que es igual de importante evaluar continuamente el modelo de credit scoring con el fin de revalidar su correcto ajuste con los valores reales, en conjunto con la contribución del conocimiento del experto para considerar todos los aspectos.
El modelo de *credit scoring* depende únicamente de los datos con los que cuenta la entidad en cuestión, las variables que se incluyen en el modelo son propias para la institución por lo que no serán las mismas por completo si se aplica a otra institución.
Cuanto esté dispuesta a correr riesgos la institución dependerá de los objetivos de la misma, por lo que es fundamental considerarlo para aceptar o rechazar a un cliente dependiendo de su probabilidad de incumplimiento.

Apéndice A

Base de datos German Credit

Muestra de 100 observaciones de la Base de datos German Credit.

No.	Clase	Balance de Cuenta	Duración del Crédito(Meses)	Historia Crediticia	Propósito	Monto del Crédito	Cuenta de ahorros	Duración en el trabajo actual	Tasa del Crédito	Género - Edo. Civil	Otros deudores Fiadores	Duración Residencia	Propiedades	Edad (años)
1	0	1	18	4	2	1049	1	2	4	2	1	4	2	21
2	0	1	9	4	0	2799	1	3	2	3	1	2	1	36
3	0	2	12	2	9	841	2	4	2	2	1	4	1	23
4	0	1	12	4	0	2122	1	3	3	3	1	2	1	39
5	0	1	12	4	0	2171	1	3	4	3	1	4	2	38
6	0	1	10	4	0	2241	1	2	1	3	1	3	1	48
7	0	1	8	4	0	3398	1	4	1	3	1	4	1	39
8	0	1	6	4	0	1361	1	2	2	3	1	4	1	40
9	0	4	18	4	3	1098	1	1	4	2	1	4	3	65
10	0	2	24	2	3	3758	3	1	1	2	1	4	4	23
11	0	1	11	4	0	3905	1	3	2	3	1	2	1	36
12	0	1	30	4	1	6187	2	4	1	4	1	4	3	24
13	0	1	6	4	3	1957	1	4	1	2	1	4	3	31
14	0	2	48	3	10	7582	2	1	2	3	1	4	4	31
15	0	1	18	2	3	1936	5	4	2	4	1	4	3	23
16	0	1	6	2	3	2647	3	3	2	3	1	3	1	44

Continúa en la siguiente página.

Cuadro A.1 – *Continuación de la página anterior*

17	0	1	11	4	0	3939	1	3	1	3	1	2	1	40
18	0	2	18	2	3	3213	3	2	1	4	1	3	1	25
19	0	2	36	4	3	2337	1	5	4	3	1	4	1	36
20	0	4	11	4	0	7228	1	3	1	3	1	4	2	39
21	0	1	6	4	0	3676	1	3	1	3	1	3	1	37
22	0	2	12	4	0	3124	1	2	1	3	1	3	1	49
23	0	2	12	4	4	1424	1	4	4	3	1	3	2	26
24	0	1	6	4	0	4716	5	2	1	3	1	3	1	44
25	0	2	11	3	3	4771	1	4	2	3	1	4	2	51
26	0	1	12	2	2	652	1	5	4	2	1	4	2	24
27	0	2	9	4	3	1154	1	5	2	3	1	4	1	37
28	0	4	15	2	0	3556	5	3	3	3	1	2	4	29
29	0	3	42	4	1	4796	1	5	4	3	1	4	4	56
30	0	3	30	4	3	3017	1	5	4	3	1	4	2	47
31	0	4	36	4	0	3535	1	4	4	3	1	4	3	37
32	0	4	36	4	0	6614	1	5	4	3	1	4	3	34
33	0	4	24	2	3	1376	3	4	4	2	1	1	3	28
34	0	1	15	2	0	1721	1	2	2	3	1	3	1	36
35	0	1	6	4	0	860	1	5	1	2	1	4	4	39
36	0	4	12	4	0	1495	1	5	4	3	1	1	1	38
37	0	4	12	4	3	1934	1	5	2	3	1	2	4	26
38	0	4	18	2	1	3378	5	3	2	3	1	1	2	31
39	0	4	24	4	1	3868	1	5	4	2	1	2	3	41
40	0	4	12	4	5	996	5	4	4	2	1	4	1	23
41	0	1	24	2	10	1755	1	5	4	2	3	4	1	58
42	0	4	18	4	0	1028	1	3	4	2	1	3	1	36
43	0	2	24	4	9	2825	5	4	4	3	1	3	4	34
44	0	2	18	2	6	1239	5	3	4	3	1	4	4	61
45	0	4	24	2	9	1258	1	4	4	3	1	1	1	25
46	0	4	24	2	0	1474	2	2	4	4	1	3	1	33
47	0	1	24	4	9	1382	2	4	4	3	1	1	1	26
48	0	4	12	2	0	640	1	3	4	1	1	2	1	49
49	0	3	36	2	3	3919	1	3	2	3	1	2	1	23
50	0	4	9	4	0	1224	1	3	3	3	1	1	1	30
51	0	4	12	4	3	2331	5	5	1	3	2	4	1	49
52	0	4	24	2	1	6313	5	5	3	3	1	4	3	41
53	0	1	12	4	3	385	1	4	4	2	1	3	1	58
54	0	4	12	4	3	1655	1	5	2	3	1	4	1	63
55	0	1	15	2	3	1053	1	2	4	4	1	2	1	27
56	0	4	21	2	3	3160	5	5	4	3	1	3	2	41
57	0	4	36	2	0	3079	5	3	4	3	1	4	1	36
58	0	4	12	4	0	1163	3	3	4	3	1	4	1	44
59	0	4	24	2	1	2679	1	2	4	2	1	1	4	29
60	0	4	48	4	3	3578	5	5	4	3	1	1	1	47
61	0	4	36	3	0	10875	1	5	2	3	1	2	3	45

Continúa en la siguiente página.

Cuadro A.1 – *Continuación de la página anterior*

62	0	1	12	3	0	1344	1	3	4	3	1	2	1	43
63	0	4	6	4	3	1237	2	3	1	2	1	1	2	27
64	0	4	12	2	3	3077	1	3	2	3	1	4	3	52
65	0	4	24	2	3	2284	1	4	4	3	1	2	3	28
66	0	2	12	2	3	1567	1	3	1	2	1	1	3	22
67	0	4	24	3	0	2032	1	5	4	3	1	4	4	60
68	0	2	21	4	2	2745	4	4	3	3	1	2	3	32
69	0	4	30	2	3	1867	5	5	4	3	1	4	3	58
70	0	4	36	2	3	2299	3	5	4	3	1	4	3	39
71	0	4	24	2	2	929	5	4	4	3	1	2	3	31
72	0	3	12	2	3	3399	5	5	2	3	1	3	3	37
73	0	2	9	2	2	2030	5	4	2	3	1	1	3	24
74	0	4	21	4	1	3275	1	5	1	3	1	4	3	36
75	0	4	24	4	0	1940	4	5	4	3	1	4	1	60
76	0	1	21	4	0	1602	1	5	4	4	1	3	3	30
77	0	4	15	2	3	1979	5	5	4	3	1	2	3	35
78	0	4	24	4	0	2022	1	3	4	2	1	4	3	37
79	0	4	36	4	3	3342	5	5	4	3	1	2	3	51
80	0	2	18	2	0	5866	2	3	2	3	1	2	3	30
81	0	3	15	4	1	2360	3	3	2	3	1	2	3	36
82	0	4	15	4	2	1520	5	5	4	3	1	4	2	63
83	0	1	12	2	0	3651	4	3	1	3	1	3	2	31
84	0	4	24	4	1	2346	1	4	4	3	1	3	3	35
85	0	4	36	3	3	4454	1	3	4	2	1	4	1	34
86	0	1	6	4	0	666	4	4	3	2	1	4	1	39
87	0	2	24	3	0	1965	5	3	4	2	1	4	3	42
88	0	2	12	4	0	1995	2	2	4	3	1	1	3	27
89	0	2	30	2	3	2991	5	5	2	2	1	4	3	25
90	0	2	30	0	9	4221	1	3	2	2	1	1	3	28
91	0	1	9	2	3	1364	1	4	3	3	1	4	1	59
92	0	2	18	4	2	6361	1	5	2	3	1	1	4	41
93	0	4	27	4	2	4526	4	2	4	3	1	2	1	32
94	0	2	12	4	3	3573	1	3	1	2	1	1	1	23
95	0	1	9	2	2	2136	1	3	3	3	1	2	1	25
96	0	2	42	4	9	5954	1	4	2	2	1	1	1	41
97	0	4	24	4	2	3777	4	3	4	3	1	4	1	40
98	0	1	15	2	9	806	1	3	4	2	1	4	2	22
99	0	2	24	3	9	4712	5	3	4	3	1	2	2	34
100	0	2	36	3	0	7432	1	3	2	2	1	2	2	54

Cuadro A.1: Base de datos German Credit Parte 1.

No.	Otros Planes de Pago	Tipo de Vivienda	No. Créditos en el banco	Ocupación	No. dependientes	Teléfono	Trabajador foraneo	No.	Otros Planes de Pago	Tipo de Vivienda	No. Créditos en el banco	Ocupación	No. dependientes	Teléfono	Trabajador foraneo
1	3	1	1	3	1	1	1	51	3	2	1	3	1	2	1
2	3	1	2	3	2	1	1	52	3	2	1	4	2	2	1
3	3	1	1	2	1	1	1	53	3	2	4	2	1	2	1
4	3	1	2	2	2	1	2	54	3	2	2	2	1	2	1
5	1	2	2	2	1	1	2	55	3	2	1	3	1	1	2
6	3	1	2	2	2	1	2	56	3	2	1	3	1	2	1
7	3	2	2	2	1	1	2	57	3	2	1	3	1	1	1
8	3	2	1	2	2	1	2	58	3	2	1	3	1	2	1
9	3	2	2	1	1	1	1	59	3	2	1	4	1	2	1
10	3	1	1	1	1	1	1	60	3	2	1	3	1	2	1
11	3	1	2	3	2	1	1	61	3	2	2	3	2	2	1
12	3	1	2	3	1	1	1	62	3	2	2	2	2	1	1
13	3	2	1	3	1	1	1	63	3	2	2	3	1	1	1
14	3	2	1	4	1	2	1	64	3	2	1	3	1	2	1
15	3	1	2	2	1	1	1	65	3	2	1	3	1	2	1
16	3	1	1	3	2	1	1	66	3	2	1	3	1	2	1
17	3	2	2	2	2	1	1	67	3	3	2	3	1	2	1
18	3	1	1	3	1	1	1	68	3	2	2	3	1	2	1
19	3	2	1	3	1	1	1	69	3	2	1	3	1	2	1
20	3	2	2	2	1	1	1	70	3	2	1	3	1	1	1
21	3	1	3	3	2	1	1	71	2	2	1	3	1	2	1
22	1	2	2	2	2	1	1	72	3	2	1	4	1	1	1
23	3	2	1	3	1	1	1	73	3	2	1	3	1	2	1
24	3	2	2	2	2	1	1	74	3	2	1	4	1	2	1
25	3	2	1	3	1	1	1	75	3	2	1	3	1	2	1
26	3	1	1	3	1	1	1	76	3	2	2	3	1	2	1
27	3	2	3	2	1	1	1	77	3	2	1	3	1	1	1
28	3	2	1	3	1	1	1	78	3	2	1	3	1	2	1
29	3	3	1	3	1	1	1	79	3	2	1	3	1	2	1
30	3	2	1	3	1	1	1	80	3	2	2	3	1	2	1
31	3	2	2	3	1	2	1	81	3	2	1	3	1	2	1
32	3	2	2	4	1	2	1	82	3	2	1	3	1	1	1
33	3	2	1	3	1	1	1	83	3	2	1	3	2	1	1
34	3	2	1	3	1	1	1	84	3	2	2	3	1	2	1
35	3	2	2	3	1	2	1	85	3	2	2	3	1	1	1
36	3	2	2	2	2	1	1	86	3	2	2	2	1	2	1
37	3	2	2	3	1	1	1	87	3	1	2	3	1	2	1
38	3	2	1	3	1	2	1	88	3	2	1	3	1	1	1

Continúa en la siguiente página.

Cuadro A.2 – *Continuación de la página anterior*

39	3	1	2	4	1	2	1	89	3	2	1	3	1	1	1
40	3	2	2	3	1	1	1	90	3	2	2	3	1	1	1
41	3	2	1	2	1	2	1	91	3	2	1	3	1	1	1
42	3	2	2	3	1	1	1	92	3	2	1	3	1	2	1
43	3	2	2	3	2	2	1	93	2	2	2	2	2	2	1
44	3	3	1	3	1	1	1	94	3	2	1	2	1	1	1
45	3	2	1	3	1	2	1	95	3	2	1	3	1	1	1
46	3	2	1	3	1	2	1	96	1	2	2	2	1	1	1
47	3	2	2	3	1	2	1	97	3	2	1	3	1	2	1
48	3	2	1	2	1	1	1	98	3	2	1	2	1	1	1
49	3	2	1	3	1	2	1	99	1	2	2	4	1	2	1
50	3	2	2	3	1	1	1	100	3	1	1	3	1	1	1

Cuadro A.2: Base de datos German Credit Parte 2

Apéndice B

Funciones de densidad

Distribución Logística: Distribución de Probabilidad

I. La notación común es $X \sim Log(\alpha, \beta)$.
Con:
$\alpha \in (-\infty, +\infty)$, (parámetro de posición)
$\beta > 0$, (parámetro de escala).

II. Su función de densidad es:

$$f(x; \alpha, \beta) = \frac{e^{-(x-\alpha)/\beta}}{\beta(1 + e^{-(x-\alpha)/\beta})^2}$$

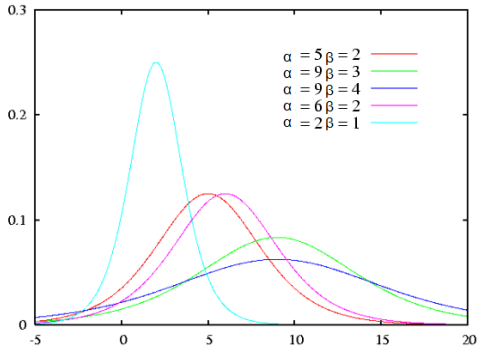

Figura B.1: Gráfica de la Función de densidad Logística.

III. Y la función de distribución es:

$$F(x) = \frac{1}{1 + e^{-(\frac{x-\alpha}{\beta})}}$$

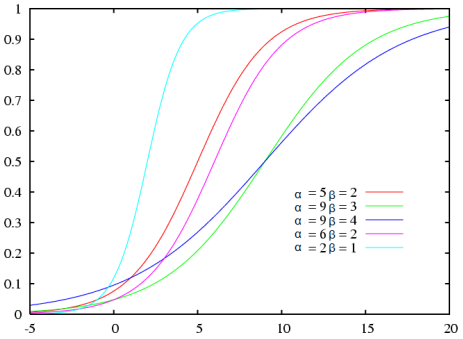

Figura B.2: Gráfica de la Distribución Logística Acumulada.

IV. La media de la función es:

$$E(X) = \alpha$$

V. Y la varianza:

$$Var(X) = \frac{\pi^2 \beta^2}{3}$$

VI. Propiedades:

Si $\alpha = 0$ *y* $\beta = 0.5513$, entonces $Log(0, 0.5513) \sim Normal(0, 1)$.

Si U es una variable uniformemente distribuida en el intervalo $(0, 1)$ $(U \sim Uniforme(0, 1))$, entonces la variable X,

$$X = \ln(\frac{U}{1-U}) \tag{B.1}$$

sigue una distribución logística.

Esta transformación, denominada *logit*, se utiliza para modelar datos de respuesta binaria.

Distribución Normal La notación común es $X \sim N(\mu, \sigma^2)$.

X tiene una distribución normal de probabilidad si y sólo si, para $\sigma > 0$ y $-\infty < \mu < \infty$,

1. Su función de densidad es:

$$f(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}, \quad \text{con} \quad -\infty < x < \infty.$$

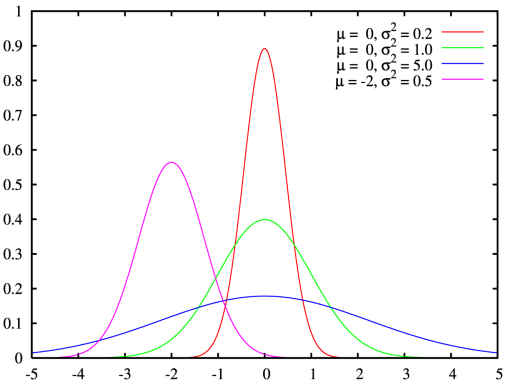

Figura B.3: Gráfica de la función de densidad Normal a diferentes valores.

2. Su función de distribución acumulada se expresa en términos

de una integral:

$$\Phi_{\mu,\sigma^2}(x) = \frac{1}{\sigma\sqrt{2\pi}} \int_{-\infty}^{x} e^{-\frac{(u-\mu)^2}{2\sigma^2}} du, \quad \text{con} \quad -\infty < x < \infty.$$

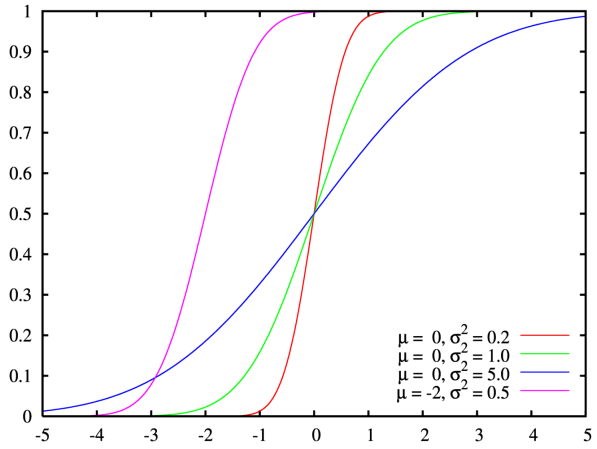

Figura B.4: Gráfica de la Distribución Normal acumulada.

3. Su valor esperado es:

$$E(X) = \mu$$

La moda y la mediana son ambas iguales a la media, μ.

4. Su varianza es:

$$Var(X) = \sigma^2$$

5. Un caso especial de la Función Normal, es la Función Normal Estándar, es decir, aquella cuyos parámetros son $\mu = 0$ y $\sigma = 1$. $X \sim N(0, 1)$.

   - Su función de densidad de probabilidad es:

$$f(x) = \frac{1}{\sqrt{2\pi}} e^{-\frac{x^2}{2}}, \quad \text{con} \quad -\infty < x < \infty.$$

- Su función de distribución acumulada es:

$$\Phi_{0,1}(x) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{x} e^{-\frac{u^2}{2}} du, \quad \text{con} \quad -\infty < x < \infty.$$

Apéndice C

Supuestos del Modelo lineal de probabilidad

Considerando la ecuación del Modelo de probabilidad lineal con una sola variable independiente como:

$$y = \alpha + \beta x + e.$$

Es habitual afirmar las suposiciones del modelo de regresión en términos del error aleatorio del modelo, e.

S. 1. El valor de y, para cada valor de x, es:

$$y = \alpha + \beta x + e.$$

S. 2. El valor esperado del error aleatorio e es:

$$E(e) = 0.$$

Lo cual es equivalente a asumir que:

$$E(y) = \alpha + \beta x.$$

S. 3. La varianza del error aleatorio e es:

$$var(e) = \sigma^2 = var(y).$$

Las variables aleatorias y y e tienen la misma varianza porque ellos difieren solamente por una constante.

S. 4. La covarianza entre cualquier par de errores aleatorios e_i y e_j es:

$$cov(e_i, e_j) = cov(y_i, y_j) = 0.$$

Esta suposición se puede hacer más fuerte asumiendo que los valores de los errores aleatorios e son estadísticamente independientes, en cuyo caso los valores de la variable independiente y son también estadísticamente independientes.

S. 5. La variable x no es aleatoria y debe tomar al menos dos valores diferentes.

S. 6. Los valores de e son normalmente distribuidos alrededor de su media

$$e \sim N(0, \sigma^2).$$

Si los valores de y son normalmente distribuidos y viceversa, [4].

Apéndice D

Residuales de Pearson

La devianza es una de las medidas más utilizadas para ver que tan bien el modelo ajusta los datos, pero existen alternativas, como es el caso del Estadístico Chi-cuadrado.
De acuero a Faraway [7], el estadístico χ^2 de Pearson, tiene la forma general:

$$\chi^2 = \sum_{i=1}^{n} \frac{(O_i - E_i)^2}{E_i}. \tag{D.1}$$

Donde:
O_i es el valor observado y E_i es el valor estimado bajo el modelo propuesto para el caso i.

Para una respuesta binomial, se tiene que para los éxitos, el valor observado es, $O_i = y_i$ y su respectivo valor estimado, $E_i = n_i\hat{\pi}_i$ y para los fracasos para se tiene que $O_i = n_i - y_i$ y $E_i = n_i(1 - \hat{\pi}_i)$ lo cual da como resultado:

$$\chi^2 = \sum_{i=1}^{n} \frac{(y_i - n_i\hat{\pi}_i)^2}{n_i\hat{\pi}_i(1 - \hat{\pi}_i)}. \tag{D.2}$$

Si se definen los *Residuales de Pearson* como:

$$r_i^P = \frac{y_i - n_i\hat{\pi}_i}{\sqrt{Var(\hat{y_i})}}. \tag{D.3}$$

Donde:
$Var(\hat{y}_i) = n_i\hat{\pi}_i(1-\hat{\pi}_i)$.
Los cuales se pueden ver como un tipo de residuales estandarizados, entonces $\chi^2 = \sum_{i=1}^{n}(r_i^P)^2$.

La distribución del estadístico χ^2 bajo la suposición de que el modelo ajustado es correcto en todos los aspectos es una Chi-cuadrada con $(n-(p+1))$ grados de libertad.

Es útil pensar al estadístico χ^2 de Pearson como el resultado de una tabla $2 \times N$. Los renglones de la tabla corresponden a los dos valores de la variable respuesta, $y = 0, 1$. Las N columnas corresponden a las N posibles covariables.

	Subgrupos			
	1	2	...	N
Éxitos	Y_1	Y_2	...	Y_N
Fracasos	$n_1 - Y_1$	$n_2 - Y_2$	...	$n_N - Y_N$
Totales	n_1	n_2	...	n_N

Cuadro D.1: Frecuencias para N distribuciones binomiales.

La estimación del valor esperado bajo la hipótesis de que el modelo logístico es correcto para la celda correspondiente al renglón de $y = 1$ y la $i-$ésima columna es $n_i\hat{\pi}_i$. Y para el renglón $y = 0$ y la $i-$ésima columna es $n_i(1-\hat{\pi}_i)$

Este estadístico, de acuerdo a Dobson [5], es asintóticamente equivalente a la ecuación de la Devianza en (1.17).

$$D = 2\sum_{i=1}^{N}\left[y_i \ln\left(\frac{y_i}{n_i\hat{\pi}_i}\right) + (n_i - y_i)\ln\left(\frac{n_i - y_i}{n_i - n_i\hat{\pi}_i}\right)\right] \tag{D.4}$$

Para probar la relación entre χ^2 y D, se usa la expansión en series de Taylor de $s\ln(s/t)$ para $s = t$, esto es,

$$s \ln \frac{s}{t} = (s-t) + \frac{1}{2}\frac{(s-t)^2}{t} + \dots \tag{D.5}$$

Así,

$$\begin{aligned} D &= 2\sum_{i=1}^{N}\{(y_i - n_i\hat{\pi}_i) + \frac{1}{2}\frac{(y_i - n_i\hat{\pi}_i)^2}{n_i\hat{\pi}_i} + [(n_i - y_i) - (n_i - n_i\hat{\pi}_i)] \\ &\quad + \frac{1}{2}\frac{[(n_i - y_i) - (n_i - n_i\hat{\pi}_i)]^2}{n_i - n_i\hat{\pi}_i} + \dots\} \\ &\cong \sum_{i=1}^{N}\frac{(y_i - n_i\hat{\pi}_i)^2}{n_i\hat{\pi}_i(1-\hat{\pi}_i)} = \chi^2. \end{aligned}$$

Por lo que, la distribución asintótica de D, bajo la hipótesis de que el modelo es correcto es $D \sim \chi^2(N-p-1)$, por consiguiente, aproximadamente $X^2 \sim \chi^2(N-p-1)$.

Bibliografía

[1] Agresti A., *Categorical Data Analysis*, John Wiley Sons, Inc, (1990).

[2] Altman, E. I., Saunders A., *Credit Risk Measurement: Developments over the Last 20 Years*, Journal of Banking and Finance, (1998).

[3] Banco de México, *Definiciones básicas de Riesgos, (2005).*

[4] Carter R., Griffiths W., Lim G., Principles of Econometrics*, John Wiley Sons, Inc, (2011).*

[5] Dobson A. J., Barnett A. G., An Introduction to Generalized Linear Models*, Chapman & Hall/CRC , (2008).*

[6] Draper R., Smith H., Applied Regresion Analysis*, Intersciencie, (1998).*

[7] Faraway J., Extending the Linear Model with R: Generalized Linear, Mixed Effects and Nonparametric Regression Models*, Chapman & Hall/CRC , (2006).*

[8] Hosmer D. & Lemeshow S., Applied Logistic Regression*, John Wiley & Sons, (2000).*

[9] Infante S., Zárate G., Métodos estadísticos: un enfoque interdisciplinario*, Trillas, (1990).*

[10] Long J. S. Regression Models for Categorical and Limited Dependent Variables*, SAGE Publications, Inc., (1997).*

[11] Malhotra D.K., Malhotra R., McLeod R., Artificial Neural Systems in Commercial Lending, *The Bankers Magazine, (1994).*

[12] Mester Loretta J. What's the point of Credit Scoring?,*Business Review, Federal Reserve Bank of Philadelphia, (1997).*

[13] Quirós G, Mercados financieros alemanes, *Banco de España, (1995).*

[14] Sainz A. El sistema bancario en Alemania, *I.D.O.E Universidad de Alcalá, Num. 88, (1994).*

[15] Schreiner M. Benefits and Pitfalls of Statistical Credit Scoring for Microfinance, *Microfinance Risk Management, (2004).*

[16] Schreiner M. Credit Scoring for Microfinance: Can It Work?,*Microfinance Risk Management, (2000).*

[17] SPSS (2010), IBM SPSS Statistics 22 para Windows.

[18] Thomas, L. C. A survey of credit and behavioral scoring: forecasting financial risk of lending to consumers. *International Journal of Forecasting, 149-172,(2000).*

[19] Wooldridge M. Jeffrey, Introductory Econometrics, *Cengage,(2006).*

[20] Women's World Banking, Guidelines based on experience with WWB affiates in Colombia and the Dominican Republic, Vol. 1, (2003).