UNIVERSIDAD PONTIFICIA CATÓLICA DEL ECUADOR

MAESTRÍA EN SISTEMA DE INFORMACIÓN
MENCIÓN CIENCIA DE DATOS

TITULO

MODELO ESTADÍSTICO DE PUNTAJE DE CREDITO (SCORING CREDIT) PARA LA GESTIÓN DE RIESGO CREDITICO

Autor: Richard Jacho Parrales
Tutor: Alfonso Prado.

ÍNDICE DE CONTENIDO

1. Introducción 6

1.1 Problema 7

1.2 Objetivo general 8

1.3 Objetivos específicos 8

1.4 Justificación 8

2. Marco Teórico 9

2.1 Definición de Riesgo de Crédito 9

2.1.1 Tipos de Riesgo de Crédito 10

2.1.2 Evaluación del Riesgo de Crédito 11

2.1.3 Clasificación de riesgos de créditos 12

2.1.4 Puntaje de Crédito. 13

2.2 Modelo de regresión logística. 16

2.2.1 Estimación del Modelo de Regresión Logística 17

2.2.2 Explicación coeficientes del modelo logístico 20

2.2.3 Test de Razón de Verosimilitud 23

2.2.4 Estadístico de Wald 24

2.2.5 Matriz de confusión 25

2.2.6 Curva Roc 26

3. Metodología de la investigación 27

3.1 Descripción del proceso de modelación. 27

3.2 Análisis exploratorio de las variables 30

3.3 Determinación de la muestra del modelo 36

3.4 Tiempo del periodo de modelización 39

3.5 Variable objetivo 42

3.6 Análisis y selección de las variables independientes 44

3.7 Variables independiente significativas 45

4. Resultados de la investigación 50

4.1 Elaboración del modelo y estadísticos 50

4.2 Regresión Logística 51

4.3 Significancia de los coeficientes 54

4.4 Interpretación de los coeficientes 56

4.5 Estadísticos y análisis de Multicolinealidad 61

4.6 Métricas de rendimiento y evaluación 64

4.7 Matriz de segmento de riesgo del cliente 72

5. Conclusiones y Recomendaciones 76

6. Referencias 80

7. Anexo 82

ÍNDICE FIGURAS

Figura 1 Función Logística 17

Figura 2 Matriz de Confusión 26

Figura 3 Métricas Matriz de Confusión 26

Figura 4 Ciclo Minería de Datos 28

Figura 5 Histograma Saldo Actual 32

Figura 6 Grafica Dispersión Ingreso Promedio Vs Días Atraso 32

Figura 7 Grafica Dispersión Atraso Máximo Y Atraso Promedio 33

Figura 8 Diagrama De Caja Y Bigote Del Monto De Crédito 33

Figura 9 Clientes Según Estado Civil 34

Figura 10 Tipo Clientes Según Estado Civil 35

Figura 11 Chi-Cuadrado Tipo Clientes Según Estado Civil 35

Figura 12 Muestra Estudio 38

Figura 13 Transición de Clientes 40

Figura 14 Correlación de variables predictoras 45

Figura 15 Tipo Clientes Según Números Créditos 47

Figura 17 Tipo Clientes Según Monto Crédito 48

Figura 18 Chi-Cuadrado Tipo Clientes Según Monto Crédito 48

Figura 19 Tipo Clientes Según Saldo Sistema Financiero 49

Figura 20 Chi-Cuadrado Tipo Clientes Según Saldo Sistema Financiero 49

Figura 21 Wilcoxon Tipo Clientes Según Máximo Atraso 49

Figura 22 Curva Roc 66

Figura 23 Matriz De Confusión 68

ÍNDICE TABLAS

Tabla 1 Variable Cuota Promedio 31

Tabla 2 Variable Saldo Actual 31

Tabla 3 Tabla Sobre Atraso Máximo Y Atraso Promedio 43

Tabla 4 Tabla Sobre Porcentaje Atraso Máximo Y Atraso Promedio 43

Tabla 5 Pasos Hacia Atrás (Backward Stepwise) 47

Tabla 6 Modelo de Puntaje Crediticio 51

Tabla 7 Residuos De La Desviaciones 52

Tabla 8 Comparativo Desviaciones 53

Tabla 9 Matriz De Segmento Del Cliente 73

Tabla 10 Matriz De Punto De Corte 75

Tabla 11 Segmento Del Cliente 75

CAPITULO 1

1. Introducción

El riesgo crediticio es una de las principales preocupaciones en el ámbito financiero, y se refiere a la probabilidad de que una empresa o individuo no pueda cumplir con sus obligaciones de pago. El riesgo crediticio puede ser originado por una serie de factores, como la falta de capacidad de pago, la falta de voluntad para pagar, o la incertidumbre económica.

El riesgo crediticio es una preocupación importante para los prestamistas, ya que representa una amenaza para la rentabilidad y la estabilidad financiera.

Por lo tanto, los prestamistas utilizan diversas herramientas para medir y gestionar el riesgo crediticio, como la evaluación de la solvencia de los prestatarios, la evaluación de la calidad de la garantía y la diversificación de la cartera de préstamos.

En la actualidad, el riesgo crediticio se ha convertido en un tema cada vez más importante debido a la creciente complejidad del mercado financiero y la aparición de nuevos productos financieros. Los prestamistas y los prestatarios deben estar preparados para enfrentar y gestionar el riesgo crediticio de manera efectiva para garantizar la estabilidad financiera y el crecimiento económico.

La gestión de riesgo crediticio es una herramienta fundamental para las empresas y los individuos que buscan obtener financiamiento, ya que les permite reducir el riesgo de incumplimiento y mejorar su capacidad de pago.

En este sentido, es importante que los prestamistas y los prestatarios comprendan los fundamentos del riesgo crediticio y utilicen herramientas efectivas para su gestión.

El presente trabajo tiene como objetivo analizar los factores que inciden en el crecimiento de la cartera vencida, que es uno de los principales indicadores del riesgo crediticio en las instituciones financieras.

1.1 Problema

El problema del riesgo crediticio radica en la probabilidad de que un cliente no cumpla con sus compromisos de desembolso, lo que puede generar pérdidas significativas para las instituciones financieras y poner en riesgo su estabilidad financiera. Esto ha llevado a una mayor demanda de modelos y herramientas de evaluación del riesgo crediticio más sofisticados y precisos.

Una empresa con alto riesgo crediticio puede enfrentar varias problemáticas que pueden afectar su estabilidad financiera y su capacidad para obtener financiamiento en el futuro, las empresas con alto riesgo crediticio pueden tener dificultades para obtener financiamiento, ya que los prestamistas pueden ser reacios a otorgar créditos a empresas que presentan mayores probabilidades de incumplimiento.

Si una empresa con alto riesgo crediticio obtiene financiamiento, existe un mayor riesgo de que no pueda cumplir con sus obligaciones de pago. Esto podría afectar su reputación y la capacidad de la empresa para obtener financiamiento en el futuro.

Evaluar adecuadamente el riesgo crediticio de los prestatarios, es posible que otorgue préstamos a personas con una alta probabilidad de no pagar a tiempo o de no pagar en absoluto. Esto podría resultar en un aumento en la cartera vencida.

Además, si relaja sus políticas de crédito y otorga préstamos a prestatarios que en condiciones normales no calificarían para obtenerlos, puede aumentar el riesgo de impago y, por ende, el crecimiento de la cartera vencida.

1.2 Objetivo general

Diseñar un modelo estadístico de puntaje crediticio para la asignación del crédito.

1.3 Objetivos específicos

- Describir en detalle el proceso de construcción de modelos mediante el uso de técnicas de regresión logística.
- Identificar las variables principales que permita tener una alta calidad en la asignación de los créditos.
- Generar una matriz de puntaje que permitan segmentar el nivel de riesgo a cada sujeto de crédito.

1.4 Justificación

Es importante investigar sobre el riesgo crediticio porque permite a los prestamistas, ya sean bancos, cooperativas de crédito u otros tipos de instituciones financieras, determinar la probabilidad de que un prestatario incumpla con sus obligaciones de pago.

Si un prestamista no evalúa adecuadamente el riesgo crediticio de un prestatario, podría otorgar un préstamo a alguien que tiene una alta probabilidad de no pagar a tiempo o de no pagar en absoluto. Esto podría resultar en pérdidas financieras significativas para el prestamista.

Por otro lado, si un prestamista tiene una comprensión clara del riesgo crediticio asociado con un prestatario, puede tomar medidas para mitigar ese riesgo, como requerir un depósito de garantía o establecer un plan de pago más riguroso. Además, también puede ayudar a los prestatarios al permitirles acceder a préstamos que de otra manera podrían no haber sido elegibles para recibir.

La investigación sobre el riesgo crediticio es importante tanto para los prestamistas como para los prestatarios, ya que ayuda a garantizar que los préstamos se otorguen de manera responsable y con un bajo riesgo financiero.

La investigación sobre el riesgo crediticio también ayuda a establecer políticas y procedimientos de préstamo efectivos. Al comprender los diferentes niveles de riesgo asociados con diferentes tipos de préstamos, los prestamistas pueden establecer políticas y procedimientos que minimicen el riesgo de impago y protejan sus activos.

CAPITULO 2

2. Marco Teórico

2.1 Definición de Riesgo de Crédito

El riesgo de crédito se describe al evento de que una entidad financiera o prestamista perciba desgastes económicos debido a la incapacidad de un prestatario para cumplir con sus obligaciones de pago. En otras palabras, es el riesgo de que un prestatario no pueda pagar su deuda a tiempo o de que incumpla completamente con sus pagos. El riesgo de crédito puede ser evaluado tanto para individuos como para empresas.

El riesgo crediticio se refiere a la probabilidad de que un prestatario no cumpla con sus obligaciones de pago. Según Chiang y Zheng (2010), el riesgo crediticio es la probabilidad de pérdida financiera resultante de la incapacidad o falta de disposición del prestatario para cumplir con sus obligaciones de pago. Por su parte, Jorion (2007) define el riesgo crediticio como la probabilidad de que un prestatario no pague a tiempo el principal o los intereses de una deuda.

Diversas formas de riesgo crediticio existen, como el riesgo de falta de pago, el riesgo de concentración, parte contratante y el relacionado con la soberanía. Cada una de estas categorías de riesgo tiene sus propias características distintivas y se somete a evaluación de manera diferenciada.

2.1.1 Tipos de Riesgo de Crédito

Existen varios tipos de riesgo de impago que pueden afectar a las instituciones financieras y a los prestamistas en general. Frank J. Fabozzi este autor “Clasifica el riesgo de crédito en siete categorías: riesgo de impago, riesgo de refinanciamiento, riesgo de dilución, riesgo de contraparte, riesgo de concentración, riesgo de país y riesgo de mercado”. (Frank J. Fabozzi, 2023).

El autor Frank J. Fabozzi define el “Riesgo de impago: se refiere a la posibilidad de que el prestatario no cumpla con sus obligaciones de pago”. (Frank J. Fabozzi, 2023).

El autor Frank J. Fabozzi define el “Riesgo de refinanciamiento: se refiere a la posibilidad de que el prestatario no pueda refinanciar la deuda en el momento del vencimiento”. (Frank J. Fabozzi, 2023).

El autor Frank J. Fabozzi define el “Riesgo de dilución: se refiere a la posibilidad de que la participación del prestamista en el capital de la empresa prestatario se diluya”. (Frank J. Fabozzi, 2023).

El autor Frank J. Fabozzi define el “Riesgo de contraparte: se refiere a la posibilidad de que el prestamista no pueda cumplir con sus obligaciones contractuales”. (Frank J. Fabozzi, 2023).

El autor Frank J. Fabozzi define el “Riesgo de concentración: se refiere a la exposición excesiva de una institución financiera a un sector, industria o prestatario específico”. (Frank J. Fabozzi, 2023).

El autor Frank J. Fabozzi define el “Riesgo de país: se refiere a la posibilidad de que un gobierno incumpla sus obligaciones financieras”. (Frank J. Fabozzi, 2023).

El autor Frank J. Fabozzi define el “Riesgo de mercado: se refiere a la posibilidad de que el valor del activo subyacente en el que se basa el crédito disminuya”. (Frank J. Fabozzi, 2023).

2.1.2 Evaluación del Riesgo de Crédito

La evaluación del riesgo de crédito es un proceso crítico que los prestamistas y las instituciones financieras deben realizar antes de otorgar un préstamo o una línea de crédito a un prestatario. La estimación del riesgo de incumplimiento de crédito involucra diferentes investigaciones de la solvencia del prestatario, su historial crediticio, su capacidad financiera y otros factores que puedan afectar su capacidad para pagar su deuda.

Entre las herramientas utilizadas para evaluar el riesgo de crédito se encuentran las calificaciones crediticias, que son evaluaciones objetivas de la capacidad crediticia de una entidad o individuo. Las calificaciones crediticias son asignadas por agencias de calificación crediticia y se basan en una variedad de factores, como el historial crediticio, la solvencia y la capacidad financiera del prestatario.

Para evaluar el riesgo crediticio, las instituciones financieras utilizan diversos modelos y técnicas. Uno de los más utilizados es el puntaje crediticio, que se basa en un conjunto de variables para determinar la probabilidad de incumplimiento del prestatario. Según Vasicek (2002), los modelos de puntaje crediticio se dividen en dos tipos: modelos de clasificación, que clasifican a los prestatarios en categorías de riesgo, y modelos de regresión, que predicen la probabilidad de incumplimiento en un período determinado. Otras técnicas incluyen el análisis de estados financieros, el análisis de flujo de efectivo y el análisis de capacidad de pago.

2.1.3 Clasificación de riesgos de créditos

La Superintendencia de Bancos es la entidad reguladora se asegura de que las instituciones financieras cumplan con las normativas establecidas la actividad bancaria en muchos países, por lo que su clasificación de riesgos de crédito es muy importante y está ampliamente difundida en el ámbito financiero.

Según la normativa No JB-2004-631 de la Superintendencia de Bancos (2003), los riesgos de crédito se clasifican en cinco categorías, según el nivel de riesgo que representan:

Según la Superintendencia de Bancos el "Riesgo Normal describe como Flujo de efectivo presenta ingresos suficientes para cubrir actividades de negocio, pueden asumir endeudamiento a largo plazo. El microempresario tiene experiencia y capacidad para operar el negocio, el sector económico al que pertenece el negocio es de bajo riesgo. Presenta retrasos en el pago de sus obligaciones de hasta 15 días, tanto en el sistema financiero como con otros acreedores. Rango mínimo de pérdida esperada 1%". (Superintendencia de Bancos , 2003).

Según la Superintendencia de Bancos el "Riesgo Moderado describe como Flujo de efectivo es suficiente para cubrir las actividades de operación, sin embargo, no alcanza a cubrir la totalidad de la deuda. En el último año por lo menos ha presentado un retraso de hasta 30 días debido a la alternabilidad del negocio. La industria y la economía presenta indicadores con un comportamiento estable. El rango mínimo de pérdida esperada es del 5%". (Superintendencia de Bancos , 2003).

Según la Superintendencia de Bancos el "Riesgo Deficiente describe como Flujo de efectivo no alcanza a cubrir totalidad de la deuda, el propietario exterioriza debilidades en la administración del negocio. La producción y la venta del sector económico presenta una tendencia decreciente. En el pago de las obligaciones con en el sistema 18 financiero y acreedores demuestra un retraso de hasta 90 días en por lo menos una cuota; el rango mínimo de pérdida esperada llega al 20%". (Superintendencia de Bancos , 2003).

Según la Superintendencia de Bancos el “Riesgo Dudoso describe como Flujo de efectivo no alcanza a cubrir las actividades de operación viabilidad del negocio en marcha es dudoso, ya dejo de operar, o se encuentra en proceso de quiebra. Presenta por lo menos un retraso de hasta 120 días en el pago de sus cuotas. Se ha iniciado acciones legales por lo dudoso del recaudo; probabilidad mínima de pérdida de hasta el 50%”. (Superintendencia de Bancos , 2003).

Según la Superintendencia de Bancos el “Riesgo Pérdida describe como La posición financiera del negocio están muy debilitadas como consecuencia del sobreendeudamiento y la incapacidad operacional, existe incertidumbre de que permanezca como negocio en marcha; valor de recuperación muy bajo en relación a lo adeudado, patrimonio o garantía remanente es escaso en comparación del valor adeudado. Morosidad superior a 120 días, pérdida esperada del 100%”. (Superintendencia de Bancos , 2003).

Es importante destacar que la clasificación de los riesgos de crédito “se basa en el análisis de la capacidad y la voluntad del deudor de cumplir con sus compromisos crediticios, así como en la evaluación de los factores de riesgo que puedan afectar su capacidad de pago. Además, esta clasificación es un referente importante para la gestión de riesgos crediticios en las entidades financieras y para la toma de decisiones de los inversionistas que evalúan el riesgo crediticio de un deudor”. (Superintendencia de Bancos , 2003).

2.1.4 Puntaje de Crédito.

Credit Scoring es una técnica utilizada en la industria financiera para evaluar el riesgo crediticio de un solicitante de crédito o préstamo. Se trata de un sistema automatizado que utiliza modelos matemáticos y estadísticos para analizar los datos de un solicitante y generar una puntuación que indica el nivel de riesgo crediticio.

Uno de los primeros autores que se ocuparon del puntaje crediticio fue William Sharpe, quien en su libro "Investments" de 1966 presentó la idea de que los modelos estadísticos y matemáticos podrían utilizarse en la toma de decisiones crediticias. En particular, Sharpe propuso el uso de la Regresión Logística para predecir la posibilidad de clientes que no cumplan con el pago en una cartera de préstamos.

El Puntaje de crédito se basa en la colección y evaluación de datos significativo sobre el solicitante, como su historial crediticio, ingresos, empleo, nivel de endeudamiento, historial de pago, entre otros. Esta información es ingresada en un modelo matemático que procesa los datos y genera una puntuación o score, que puede variar de acuerdo a cada institución financiera.

La puntuación Puntaje Crediticio puede ser utilizada por las entidades financieras para determinar si conceden o no el crédito o préstamo solicitado, y en caso afirmativo, establecer las condiciones y términos del mismo, como el plazo y la tasa de interés.

Entre las ventajas del Puntaje Crediticio se encuentran la rapidez y eficiencia de la calificación del riesgo de impago, la objetividad en la toma de decisiones y la reducción del sesgo humano. Sin embargo, también puede presentar algunas desventajas, como la posible exclusión de personas que no cumplen con los requisitos de puntuación, pero que podrían tener capacidad de pago y un buen historial de crédito.

Otro autor importante es John Barron, quien en 1989 publicó el libro "Credit Scoring and Its Applications", en el que se explica cómo se utiliza el puntaje crediticio en la evaluación del riesgo crediticio y se presentan diversos modelos de scoring. Barron y su coautor, Michael Staten, describen cómo el Credit Scoring se ha convertido en una técnica estándar en la evaluación del riesgo crediticio y cómo los prestamistas pueden utilizarla para mejorar la eficiencia y la precisión de sus decisiones crediticias.

Ventajas y desventajas

El puntaje crediticio ofrece varias ventajas en la evaluación de riesgo crediticio:

1. Es una técnica objetiva y basada en datos, lo que significa que se utiliza una metodología estandarizada para evaluar el riesgo crediticio de los prestatarios. Esto reduce la posibilidad de sesgo y subjetividad en el proceso de evaluación.

2. El puntaje crediticio es rápido y eficiente. Los modelos de Credit Scoring pueden procesar grandes cantidades de datos en poco tiempo, lo que permite a los prestamistas evaluar rápidamente el riesgo crediticio de los prestatarios y tomar decisiones informadas sobre la concesión de préstamos.

3. Un sistema de gestión de crédito ayuda a la empresa a administrar y evaluar los riesgos asociados con la concesión de crédito a los clientes.

Sin embargo, el puntaje crediticio también presenta algunas desventajas.

1. Los modelos de puntaje crediticio pueden no tener en cuenta factores individuales, como las circunstancias personales o los imprevistos financieros que puedan afectar la capacidad de pago de un prestatario.

2. El puntaje crediticio puede ser vulnerable a la manipulación de datos. Los prestatarios pueden intentar ocultar información financiera negativa o proporcionar información falsa para mejorar su puntuación de crédito.

2.2 Modelo de regresión logística.

La regresión logística es un método estadístico que se usa para examinar la conexión entre una variable dependiente categórica y una o más variables independientes, ya sean continuas o categóricas. Es un tipo de análisis de regresión que se utiliza comúnmente en el campo de la ciencia social, la biología y la medicina, entre otros.

El modelo de regresión logística se basa en la función logística, que es una función sigmoidea que puede tomar cualquier valor entre 0 y 1. El modelo logístico se usa para calcular la probabilidad de que la variable objetivo sea igual a 1, tomando en cuenta los valores de las variables independientes.

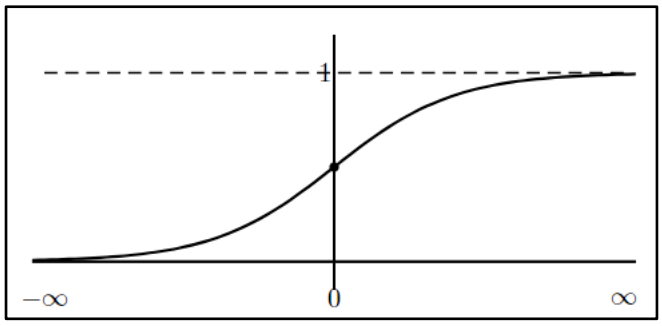
El modelo de regresión logística se puede ajustar a los datos utilizando diversos métodos, incluyendo el método de máxima verosimilitud y el método de mínimos cuadrados. Una vez que el modelo se ha ajustado, se pueden realizar inferencias sobre la correspondencia entre las variables predictoras y la variable predicha.

Según Lara define “El modelo logit o de regresión logística viene a ser la aplicación estadística más usada que arroja probabilidades de ocurrencia de un evento (por ejemplo, la probabilidad de ser impago) previamente establecidos, debido a la mayor facilidad de cálculo y a la mejor interpretación y valoración del modelo en su conjunto. Se trata de un modelo de elección binaria en el que la variable dependiente puede tomar las valores 0 o 1, aplicando la función de distribución logística obtenida a partir de la probabilidad a posteriori aplicada al análisis discriminante mediante el teorema de Bayes”. (Lara, 2010).

$$\Lambda(X\beta) = P_i = P(Y=1|X) = F(Z_i) = \frac{e^{X\beta}}{1+e^{X\beta}} = \frac{1}{1+e^{-X\beta}} \quad (1)$$

De acuerdo a Lara “El modelo de regresión logística o logit puede formularse en termino de probabilidad, a través de la función logística”. (Lara, 2010).

Figura 1 Función Logística



Fuente: Elaboración Propia.

$$P_i = \frac{e^{\beta_0+\beta_1 x_{1i}+\beta_2 x_{2i}+\cdots+\beta_k x_{ki}}}{1+e^{\beta_0+\beta_1 x_{1i}+\beta_2 x_{2i}+\cdots+\beta_k x_{ki}}} = \frac{1}{1+e^{-(\beta_0+\beta_1 x_{1i}+\beta_2 x_{2i}+\cdots+\beta_k x_{ki})}} \qquad i = 1, 2 \ldots n$$

(2)

- Pi = Representa la ocurrencia de un evento probabilístico.
- β0, β1, ..., βk son los coeficientes del modelo de regresión logística para las respectivas variables X1, X2,
- X1, X2, ..., Xk son las variables de entrada.
- *e* = es la base del logaritmo natural aproximadamente 2.71828.

2.2.1 Estimación del Modelo de Regresión Logística

Según Flórez y Rincón “La forma general del modelo logit se puede expresar como”. (Flórez y Rincón, 2002).

$$y_i = E(y_i) + \varepsilon_i$$

(3)

Según Flórez y Rincón “Las observaciones yi son variables aleatorias independientes Bernoulli, con valores esperados”. (Flórez y Rincón, 2002).

$$E(y_i) = \pi_i = \frac{e^{\mathbf{x}'\beta}}{1 + e^{\mathbf{x}'\beta}}$$

(4)

Según Flórez y Rincón “Cada observación sigue una distribución Bernoulli, su distribución será”. (Flórez y Rincón, 2002).

$$f_i(y_i) = \pi_i^{y_i}(1 - \pi_i)^{1-y_i}, i = 1, 2, , 3, \ldots, n$$

(5)

Según Flórez y Rincón “Las observaciones son independientes, la función de verosimilitud será”. (Flórez y Rincón, 2002).

$$L(y_1, y_2, \ldots, y_n, \beta) = \prod_{i=1}^{n} f_i(y_i) = \prod_{i=1}^{n} \pi_i^{y_i}(1 - \pi_i)^{1-y_i}$$

(6)

Según Flórez y Rincón “Al tomar logaritmo a la función de verosimilitud”. (Flórez y Rincón, 2002).

$$LnL(y_1, y_2, \ldots, y_n, \beta) = Ln \prod_{i=1}^{n} f_i(y_i) = \sum_{i=1}^{n} \left[y_i Ln \left(\frac{\pi_i}{1-\pi_i} \right) \right] + \sum_{i=1}^{n} Ln(1-\pi_i)$$

(7)

Como

$$1 - \pi_i = \frac{1}{1 + e^{\mathbf{x}'\beta}} \text{ y } Ln \left(\frac{\pi_i}{1-\pi_i} \right) = \mathbf{x}'\beta,$$

(8)

Según Flórez y Rincón “El logaritmo de la verosimilitud se puede expresar para el modelo de regresión logística”. (Flórez y Rincón, 2002).

$$LnL(\mathbf{y}, \beta) = \sum_{i=1}^{n} y_i \mathbf{x}_i' \beta - \sum_{i=1}^{n} Ln[1 + e^{\mathbf{x}_i'\beta}]$$

(9)

Según Flórez y Rincón “En muchas aplicaciones del modelo se dispone de información repetidas para cada uno de los valores de las variables. Sea yi la cantidad de 1 observados para la i-esima observación y ni la cantidad de ensayos en cada observación, entonces el logaritmo de la verosimilitud se puede presentar”. (Flórez y Rincón, 2002).

$$Ln\, L(\mathbf{y}, \beta) = \sum_{i=1}^{n} y_i \pi_i + \sum_{i=1}^{n} n_i Ln\,(1-\pi_i) - \sum_{i=1}^{n} y_i\, Ln\,(1-\pi_i)$$

(10)

Según Flórez y Rincón “Los estimadores de máxima verosimilitud se pueden obtener mediante un algoritmo de mínimos cuadrados iterativamente re ponderados. Si β es el estimador obtenido, mediante el método iterativo y siendo ciertas las hipótesis del modelo, se puede demostrar que en forma asintótica”. (Flórez y Rincón, 2002).

$$E(\hat{\beta}) = \beta \quad \text{y} \quad V(\hat{\beta}) = (\mathbf{X}'V^{-1}\mathbf{X})^{-1}$$

(11)

Según Flórez y Rincón “El valor estimado del predictor lineal es ηi=xiˆβ, y el valor esperado del modelo de regresión logístico, se suele expresar”. (Flórez y Rincón, 2002).

$$\begin{aligned}\hat{y}_i = \hat{\pi}_i &= \frac{e^{\hat{\eta}_i}}{1+e^{\hat{\eta}_i}} \\ &= \frac{e^{(\mathbf{x}_i'\hat{\beta})}}{1+e^{(\mathbf{x}_i'\hat{\beta})}} \\ &= \frac{1}{1+e^{(-\mathbf{x}_i'\hat{\beta})}}\end{aligned}$$

(12)

2.2.2 Explicación coeficientes del modelo logístico

Según Hosmer y Lemeshow “Señalan que los coeficientes de la regresión logística se pueden convertir en odds ratios para facilitar su interpretación. Un odds ratio es una medida de la asociación entre una variable predictora y el resultado binario”. (Hosmer y Lemeshow, 2000).

El odds ratio se define como la razón entre la probabilidad de que ocurra el resultado binario en el grupo expuesto a la variable predictora y la probabilidad de que ocurra el resultado binario en el grupo no expuesto a la variable predictora. El cálculo del odds ratio obteniendo los coeficientes del modelo logístico se realiza utilizando la fórmula exponencial, aplicada a cada coeficiente:

El OR puede interpretarse como una medida de la fuerza y la dirección de la asociación entre la variable predictora y el resultado binario. Un OR mayor que 1 indica una asociación positiva, es decir, que la variable predictora está asociada con un mayor riesgo de que ocurra el resultado binario. Por el contrario, un OR menor que 1 indica una asociación negativa, es decir, que la variable predictora está asociada con un menor riesgo de que ocurra el resultado binario. Un OR de 1 indica que no hay asociación entre la variable predictora y el resultado binario.

Según Lara describe los siguiente: "El cambio en el logit se interpreta mediante la introducción de una media de asociación, denominada odds ratio o coeficiente de ventaja que se define como la razón de la odds o ventaja de la probabilidad de que ocurra un suceso para un determinado valor de la variable respecto a la odds para otro valor de la misma, indicando de cuanto más probable (o improbable) es que ocurra el suceso de que se analiza entre los individuos que presentan cada categoría de la variable independiente (xj)". (Lara, 2010).

$$\hat{\Psi}_{x_1,x_0} = \frac{\text{odds}(x_1)}{\text{odds}(x_0)} = \frac{\dfrac{\hat{\pi}(x_1)}{1-\hat{\pi}(x_1)}}{\dfrac{\hat{\pi}(x_0)}{1-\hat{\pi}(x_0)}} = e^{\hat{\beta}_j(x_1-x_0)}$$

(13)

Según Lara describe los siguiente: "Se pretende buscar una expresión que venga dada como una función lineal de las variables explicativas. Para ello, la inversa de la función logística, que es el logit o logaritmo de la odds o ventaja de que un suceso ocurra se interpreta como la preferencia de elegir la alternativa uno de la variable respuesta, frente a la alternativa cero". (Lara, 2010).

$$\text{logit}(P_i) = \ln\left[\frac{P_i}{1-P_i}\right] = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + \cdots + \beta_k x_{ki} \quad ,, \quad i = 1, 2 \ldots n$$

(14)

Según Lara describe los siguiente: “Otra formulación del modelo de regresión logista es la que estima la ventaja o preferencia (odds) de un individuo por la categoría uno frente a la cero de la variable dependiente, definiéndose como el cociente entre la probabilidad de que ocurra un acontecimiento y la probabilidad de que no ocurra, que es su complementaria”. (Lara, 2010).

$$\frac{P_i}{1-P_i} = e^{\beta_0+\beta_1 x_{1i}+\cdots+\beta_k x_{ki}} = e^{\beta_0}\cdot e^{\beta_1 x_{1i}}\cdot\ldots\cdot e^{\beta_k x_{ki}} = e^{\beta_0}\left(e^{\beta_1}\right)^{x_{1i}}\ldots(e^{\beta_k})^{x_{ki}} \quad i=1,2\ldots n$$

(15)

En palabras de Kleinbaum y Klein “El OR mide la relación entre dos probabilidades. Si el OR es mayor que 1, entonces la probabilidad de que ocurra el evento en el grupo expuesto es mayor que la probabilidad de que ocurra en el grupo no expuesto”. (Kleinbaum y Klein, 2010).

.

Según Kleinbaum y Klein “Si el OR es menor que 1, entonces la probabilidad de que ocurra el evento en el grupo expuesto es menor que la probabilidad de que ocurra en el grupo no expuesto. Si el OR es igual a 1, entonces no hay diferencia en las probabilidades entre los grupos expuesto y no expuesto”. (Kleinbaum y Klein, 2010).

Según Kleinbaum y Klein “Odds ratio (OR) en la regresión logística se puede expresar como la exponencial del coeficiente de la variable predictora. En otras palabras, el OR se calcula elevando e a la potencia del coeficiente estimado de la variable predictora en el modelo de regresión logística”. (Kleinbaum y Klein, 2010).

Según Hosmer y Lemeshow “La fórmula del odds ratio se puede expresar de la siguiente manera”. (Kleinbaum y Klein, 2010).

OR = e^(coeficiente de la variable predictora)

(16)

2.2.3 Test de Razón de Verosimilitud

El test de razón de verosimilitud es una técnica estadística utilizada en el análisis de modelos de regresión y otras técnicas de modelado estadístico. Es una prueba de hipótesis que se utiliza para comparar la bondad de ajuste de dos modelos diferentes, uno completo y otro reducido. La idea es comparar la probabilidad de que los datos observados sean generados por el modelo completo en comparación con la probabilidad de que sean generados por el modelo reducido.

El test de razón de verosimilitud se basa principalmente en maximizar de la función de verosimilitud que es una dirección de probabilidad de cada observación sean generados por el modelo estadístico. La función de verosimilitud se maximiza en ambos modelos para obtener las estimaciones de los parámetros del modelo.

La estadística de prueba para el test de razón de verosimilitud se calcula como la diferencia de las log-verosimilitudes entre el modelo completo y el modelo reducido:

$$-2ln(\lambda) = -2ln(\frac{L_R}{L_M}) = -2(lnL_R - lnL_M)$$

(17)

$$LR = 2[\log(L(\text{completo})) - \log(L(\text{reducido}))]$$

(18)

Donde:

L(completo) es la verosimilitud del modelo completo
L(reducido) es la verosimilitud del modelo reducido

La estadística de prueba LR sigue una distribución chi-cuadrado con k grados de libertad, donde k es la diferencia en el número de parámetros entre el modelo completo y el modelo reducido.

El valor p se calcula a partir de la distribución chi-cuadrado para determinar si el modelo completo proporciona un mejor ajuste que el modelo reducido. Un valor p pequeño (generalmente menor que 0.05) indica que el modelo completo proporciona un ajuste significativamente mejor que el modelo reducido.

De acuerdo con Blanco “La hipótesis nula será rechazada para el nivel de significación α cuando$-2\ln(\lambda) > \chi 2(p+1-q),\alpha$. Esto es equivalente a que el p valor del contraste sea menor que el nivel de significación fijado”. (Blanco, 2006).

2.2.4 Estadístico de Wald

El estadístico Wald se emplea para evaluar si los coeficientes son estadísticamente significativos en un modelo de regresión logística. Se basa en la idea de que, si el valor estimado del coeficiente es lo suficientemente grande en relación con su error estándar, entonces se puede rechazar la hipótesis nula de que el coeficiente es igual a cero.

La fórmula para el estadístico Wald es:

$$H_0)\beta_k = 0$$
$$H_1)\beta_k \neq 0$$

(19)

$$W = (\beta - \beta 0) / SE(\beta)$$

(20)

Donde:

β es el valor estimado del coeficiente
β0 es el valor de la hipótesis de nulidad que se está probando (generalmente cero)
SE(β) es el error estándar del coeficiente

El estadístico Wald sigue una distribución chi-cuadrado con un grado de libertad, por lo que se puede calcular un valor p para determinar la significancia estadística del coeficiente.

Según Blanco "El nivel de significación de un test es un concepto estadístico asociado a la verificación de una hipótesis. Se define como la probabilidad de tomar la decisión de rechazar la hipótesis nula (H0) cuando esta es verdadera (decisión conocida como "Error de tipo I", o "falsos positivos"). La decisión se toma a menudo utilizando el p-valor : si el valor p es inferior a nivel de significación, entonces la hipótesis nula es rechazada". (Blanco, 2006).

2.2.5 Matriz de confusión

La matriz de confusión es una representación tabular que muestra la calidad de las predicciones realizadas por un modelo de clasificación en relación con las clases reales. Es una herramienta descriptiva que permite visualizar los aciertos y los errores del modelo en términos de verdaderos positivos (VP), falsos positivos (FP), falsos negativos (FN) y verdaderos negativos (VN).

- Verdadero positivo (VP): El modelo predijo correctamente que una muestra pertenece a la clase positiva.
- Falso positivo (FP): El modelo predijo incorrectamente que una muestra pertenece al grupo positivo cuando en realidad pertenece al grupo negativo.
- Falso negativo (FN): El modelo predijo incorrectamente que una muestra pertenece a la clase negativa cuando en realidad pertenece a la clase positiva.
- Verdadero negativo (VN): El modelo predijo correctamente que una muestra pertenece a la clase negativa.

Figura 2 Matriz de Confusión

	CLASE PREDICHA	
Clase Real	verdaderos positivos (VP)	falsos negativos (FN)
	falsos positivos (FP)	verdaderos negativos (VN)

Fuente: Elaboración Propia.

Métricas Principales

Existen varias métricas comunes utilizadas para evaluar el rendimiento de un modelo basado en una matriz de confusión. Estas métricas proporcionan información sobre diferentes aspectos de las predicciones del modelo.

Figura 3 Métricas Matriz de Confusión

METRICA	FORMULA	INTERPRETACION
Precisión (Accuracy)	(VP + VN) / (VP + FP + FN + VN)	Mide la proporción de predicciones correctas en relación con el total de predicciones realizadas.
Precisión (Precision)	VP / (VP + FP)	También conocida como valor predictivo positivo, es la proporción de verdaderos positivos en relación con todas las predicciones positivas realizadas.
Exhaustividad (Recall)	VP / (VP + FN)	También conocida como sensibilidad o tasa de verdaderos positivos, es la proporción de verdaderos positivos en relación con todos los casos positivos reales.
Puntuación F1 (F1-Score)	2 * (Precisión * Exhaustividad) / (Precisión + Exhaustividad)	Es una métrica que combina la precisión y la exhaustividad en un solo valor, proporcionando un equilibrio entre ambas medidas.
Especificidad (Specificity)	VN / (FP + VN)	Mide la proporción de verdaderos negativos en relación con todos los casos negativos reales.

Fuente: Elaboración Propia.

2.2.6 Curva Roc

Según Suárez “En una curva ROC, que será definida en el siguiente apartado, el índice de Youden es la distancia vertical máxima entre la curva y la diagonal. Siendo el punto de corte óptimo, aquel en el cual se alcanza el valor de YI”. (Suárez, 2010).

Según Krzanowski “Más tarde, en los años 1970 y 1980, se hizo evidente la importancia de la técnica para la evaluación médica de pruebas y toma de decisiones, y desde entonces se ha visto mucho el desarrollo y uso de la técnica en áreas tales como radiología, cardiología, química clínica y la epidemiología”. (Krzanowski, 2009).

Según Krzanowski “La curva ROC es utilizada para evaluar situaciones en las que el objetivo del modelo es asignar las observaciones a una o más clases. Desafortunadamente, los procedimientos no son perfectos, se cometen errores asignando observaciones a la clase incorrecta por lo que se hace necesario evaluar no solo el comportamiento del modelo y sus variaciones sino también, si es necesario reemplazarlo por otro”. (Krzanowski, 2009).

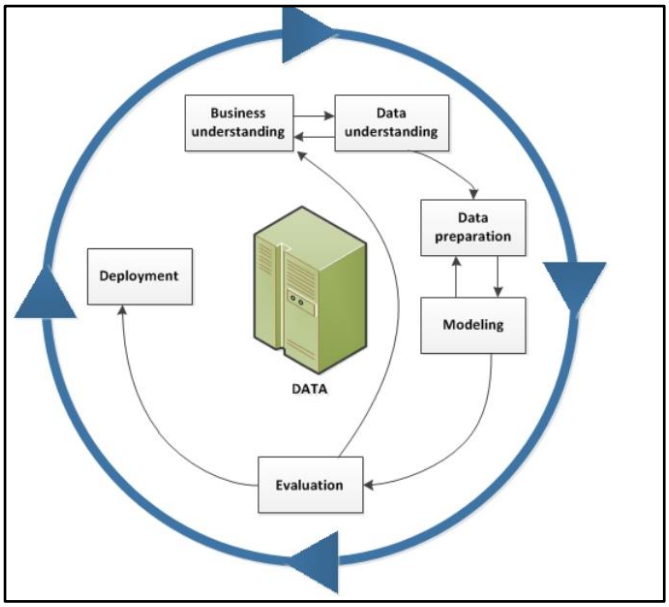
CAPITULO 3

3. Metodología de la investigación

3.1 Descripción del proceso de modelación.

Para lograr un modelo de puntaje crediticio es necesario identificar modelos válidos, usables y claros fundados en un procedimiento que se pueda convertir los datos estadísticos en informes para toma decisiones y, posteriormente, con la información se le aplica el método CRISP-DM, y la implementación de cada paso conduce a algoritmos con mejores resultados. La Figura 4 resume las siguientes fases para implementación de las metas establecidas.

Figura 4 Ciclo Minería de Datos



Fuente: GUÍA CRISP-DM DE IBM SPSS MODELER

Para la metodología utilizaremos el Capítulo 3 que se encuentra las diferentes fases del como la fase del Entendimiento del Negocio, Entendimiento y Preparación de los Datos y para el Capítulo 4 se encuentra el Modelamiento y Evaluación del Modelo

Entendimiento Del Negocio

Se realizará una regresión logística para los clientes de una empresa "C" que dedica a la venta de productos a créditos con el fin de obtener un modelo de puntaje de crédito.

Adicionalmente, es relevante destacar que las operaciones seleccionadas para el modelo de puntaje de crédito de originación corresponden a clientes bancarizados y se actualiza al inicio de cada mes.

La inclusión de variables de la central de riesgos es importante ya que permite utilizar información relacionada con el comportamiento crediticio previo del cliente para explicar su riesgo crediticio actual.

Las variables de central de riesgos pueden incluir el historial de pagos, la cantidad de deudas, el nivel de endeudamiento, la antigüedad crediticia, entre otros. Estas variables son útiles para evaluar la solvencia del cliente que le permita obedecer sus compromisos crediticios y proporcionan una perspectiva más completa del perfil crediticio del solicitante.

Al considerar las diferentes variables del sistema financiero que involucran gran parte del buró de crédito en el modelo de puntaje de crédito de originación, se puede mejorar la precisión en la evaluación del riesgo crediticio y tomar decisiones más informadas sobre la aprobación de los préstamos. Esto ayuda a reducir el riesgo de incumplimiento y las pérdidas asociadas, al tiempo que se promueve una gestión responsable del crédito.

Los modelos de puntaje de crédito pueden aplicarse en diferentes momentos durante del tiempo de un crédito. Según la Superintendencia de Banco "por lo general se clasifican en dos tipos de Puntaje, para nuestro estudio utilizamos Scoring de originación". (Superintendencia de Banco, 2008).

La Superintendencia de Banco define "Scoring de originación Este tipo de modelo se utiliza en la fase de aprobación o evaluación de solicitudes de nuevos créditos. El objetivo principal es evaluar el riesgo crediticio de los solicitantes y determinar si deben ser aprobados o rechazados". (Superintendencia de Banco, 2008)

Se basa en información disponible en el momento de la solicitud, como historial crediticio, ingresos, deudas y otras variables relevantes. El scoring de originación va ayudar a reducir la perdida esperada por riesgo crediticio además de tomar decisiones informadas y eficientes al evaluar la elegibilidad de los solicitantes y asignar límites de crédito adecuados.

La Superintendencia de Banco define "Scoring de comportamiento: Este tipo de modelo se utiliza para dar seguimiento a los clientes que ya están incorporados a la institución financiera. El objetivo principal es evaluar el comportamiento crediticio de los clientes a lo largo del tiempo". (Superintendencia de Banco, 2008)

Se basa en la información actualizada del cliente, como pagos realizados, retrasos en los pagos, utilización de crédito, entre otros factores. El scoring de comportamiento ayuda a monitorear la calidad de la cartera crediticia y a identificar clientes que podrían estar en riesgo de incumplimiento. También se utiliza para tomar decisiones relacionadas con límites de crédito adicionales.

Ambos tipos de modelos de scoring son importantes en la gestión del riesgo crediticio. El scoring de originación permite una evaluación inicial de los solicitantes, mientras que el scoring de comportamiento ayuda a mantener un seguimiento continuo de los clientes existentes. Al combinar estos dos enfoques, las instituciones financieras pueden tomar decisiones más informadas y gestionar mejor su cartera de crédito a lo largo del tiempo.

Entendimiento De Los Datos

3.2 Análisis exploratorio de las variables

El dataset actualmente posee 48591 filas (clientes) y 29 variables(características), el conjunto de datos contiene información de clientes organizada en filas, donde cada fila representa a un cliente individual. Las columnas en el conjunto de datos proporcionan detalles específicos sobre tarjetas de créditos, numero de crédito, numero de pagos, valor del crédito desembolsado, saldo del sistema financiero, ingreso promedio, numero de atraso, entre otras variables de los clientes.

El análisis exploratorio de las variables es una etapa crucial en la modelización de puntaje de crédito, ya que permite comprender mejor las características y la distribución de los datos antes de construir el modelo.

Calcula medidas resumidas como media, mediana, desviación estándar, mínimo, máximo y percentiles para cada variable. Estas estadísticas proporcionan información sobre la tendencia central, la dispersión y la forma de la distribución de los datos.

Tabla 1 Variable Cuota Promedio

Min.	1st Qu.	Median	Mean	3rd Qu.	Max.
11.20	56.71	82.01	93.04	115.61	1150.26

Fuente: Elaboración Propia.

El valor mínimo es 11.20, lo que indica que al menos hay un cliente con valor de la cuota con ese valor. El primer cuartil (1st Qu.) es 56,71, lo que significa que el 25% de los clientes tienen un valor de la cuota igual o menor que 56,71. La mediana es 82.01, lo que indica que el 50% de los clientes tienen un valor de cuota igual o menor que 82.01. La media (Mean) es 93.04, que es el promedio de valor cuota de los clientes. El tercer cuartil (3rd Qu.) es 115.61, lo que significa que el 75% de los clientes tienen un valor de la cuota igual o menor que 115.61. El valor máximo es 1150.26, que es el valor más alto que hay un cliente con valor de la cuota con ese valor.

Tabla 2 Variable Saldo Actual

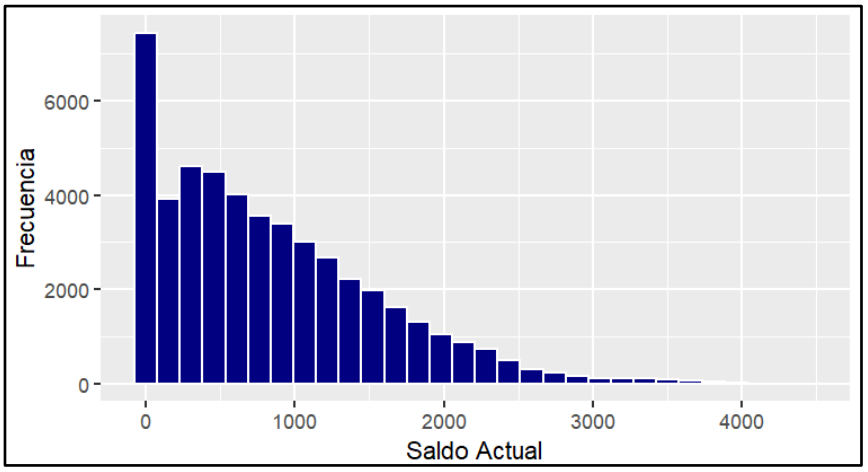
Min.	1st Qu.	Median	Mean	3rd Qu.	Max.
0.0	254.8	677.0	829.4	1254.1	4416.9

Fuente: Elaboración Propia.

El valor mínimo es 0.0, lo que indica que al menos hay un cliente con saldo actual con ese valor. El primer cuartil (1st Qu.) es 254.8, lo que significa que el 25% de los clientes tienen un saldo actual igual o menor que 254.8. La mediana es 677.0, lo que indica que el 50% de los clientes tienen un valor de saldo actual igual o menor que 677.0. La media (Mean) es 829.4, que es el promedio de saldo actual de los clientes. El tercer cuartil (3rd Qu.) es 1254.1, lo que significa que el 75% de los clientes tienen un saldo actual igual o menor que 115.61. El valor máximo es 4416.9, que es el valor más alto que hay un cliente con saldo actual con ese valor.

Crear histogramas o gráficos de densidad para visualizar la distribución de cada variable. Esto puede revelar las observaciones siguen una distribución gaussiana o si hay asimetría o sesgo en los datos.

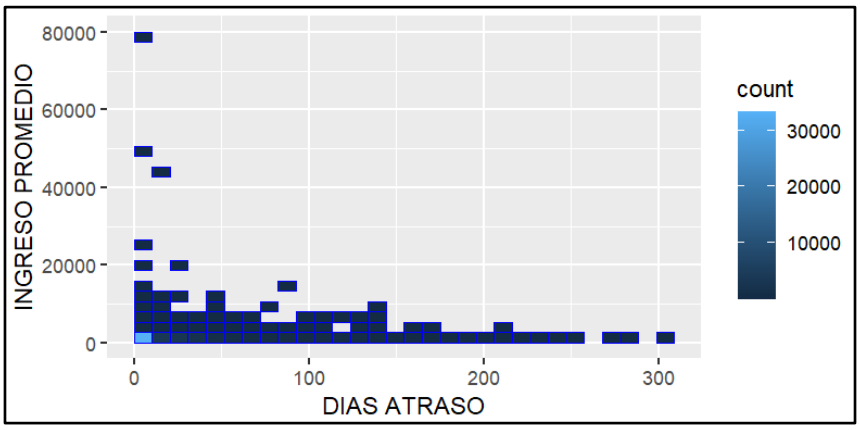
Figura 5 Histograma Saldo Actual



Fuente: Elaboración Propia.

Se utiliza gráficos de dispersión para explorar la relación entre pares de variables. Esto puede ayudar a identificar posibles relaciones lineales, no lineales o patrones en los datos.

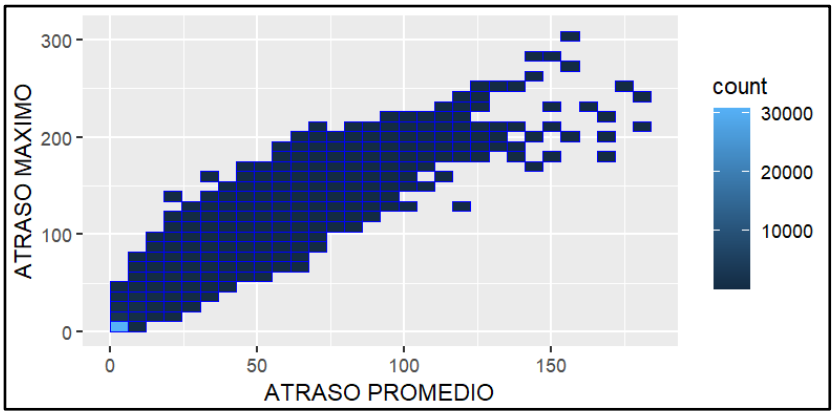
Figura 6 Grafica Dispersión Ingreso Promedio Vs Días Atraso



Fuente: Elaboración Propia.

La grafica nos muestra que no tiene una relación entre los días de atraso y el ingreso promedio de los clientes que tiene un crédito. Nos muestra que tenemos valores atípicos en el ingreso que lo vamos a visualizar a continuación

Figura 7 Grafica Dispersión Atraso Máximo Y Atraso Promedio

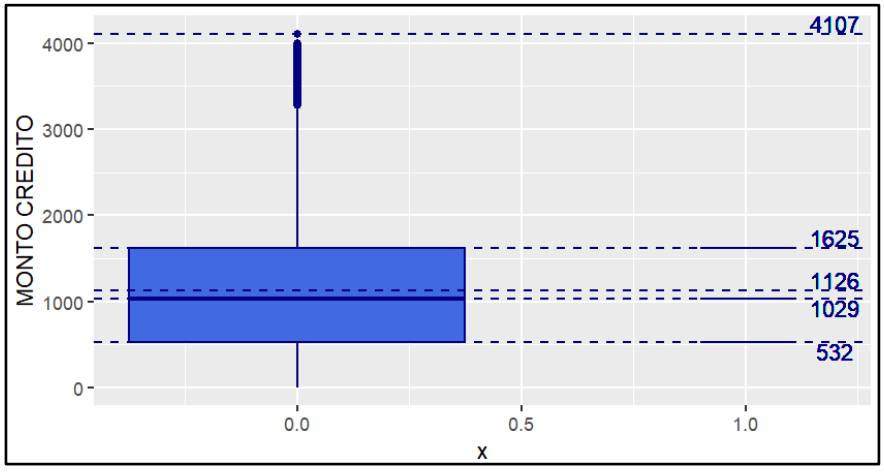


Fuente: Elaboración Propia.

La grafica nos muestra que tiene una relación entre el atraso promedio y el atraso máximo estos ayuda a visualizar la relación que existe entre variables dependiente y evitar la multicolinealidad.

Identifica valores atípicos o extremos en los datos. Los valores atípicos pueden tener un impacto significativo en los modelos de puntaje crediticio, por lo que es importante analizar su presencia y determinar si deben ser excluidos o tratados de manera especial.

Figura 8 Diagrama De Caja Y Bigote Del Monto De Crédito



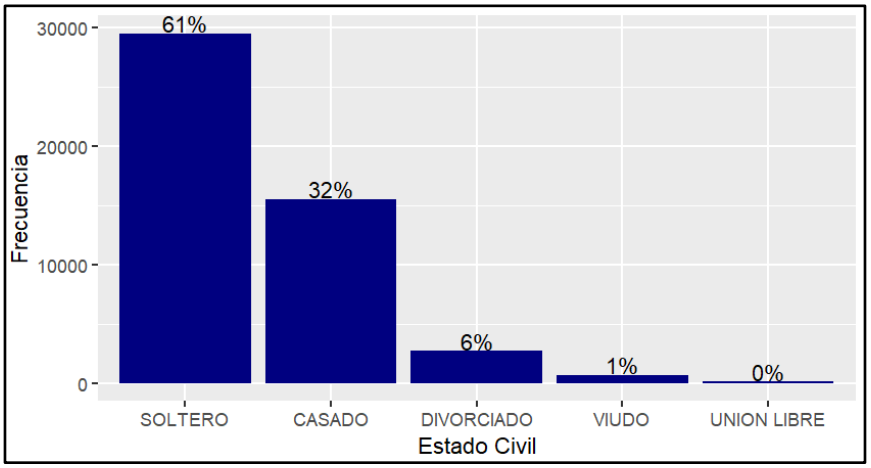
Fuente: Elaboración Propia.

El monto mínimo del crédito es de 0.0, lo que indica que hay casos donde no se otorgó ningún crédito. El primer cuartil (25%) del monto del crédito es de 532.2, lo que significa que el 25% de los créditos tienen un monto menor o igual a este valor. La mediana del monto del crédito es de 1029.1, lo que indica que el 50% de los créditos tienen un monto menor o igual a este valor.

El promedio del monto del crédito es de 1126.2, que representa el valor promedio de los montos de crédito. El tercer cuartil (75%) del monto del crédito es de 1625.3, lo que significa que el 75% de los créditos tienen un monto menor o igual a este valor. El monto máximo del crédito es de 4107.2, indicando el valor más alto observado en los datos.

En las variables categóricas en los datos, se utiliza gráficos de barras para explorar la distribución de categorías y su relación con la variable objetivo.

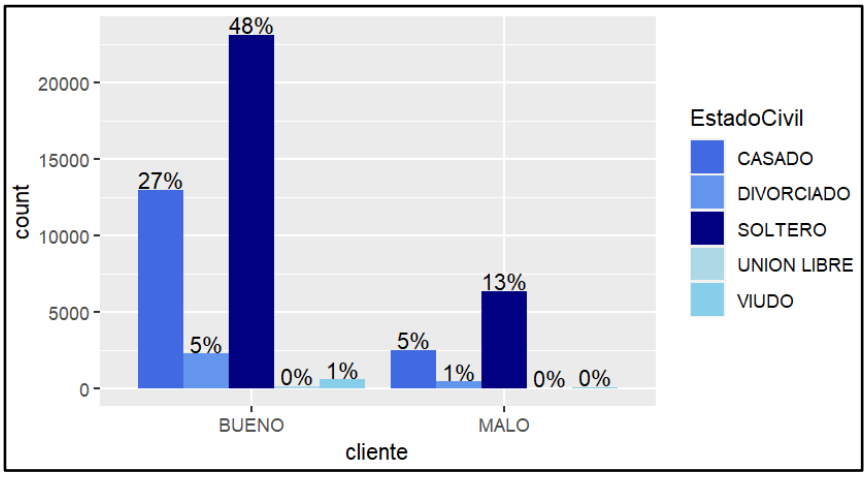
Figura 9 Clientes Según Estado Civil



Fuente: Elaboración Propia.

Como observamos los clientes que tienen un crédito el 61% son solteros, y un 32% estado civil casado y un 8% se encuentra entre divorciados viudo y unión libre.

Figura 10 Tipo Clientes Según Estado Civil



Fuente: Elaboración Propia.

Como observamos los clientes buenos representa un 48% y se encuentra en estado civil soltero, 27% son casados y un 6% se encuentra entre divorciado, soltero y unión libre.

Mientras que los clientes malos representan un 13% y se encuentra en estado civil soltero, 27% son casados y un 6% se encuentra entre divorciado, soltero y unión libre.

Para realizar si la variable objetivo tiene correspondencia con la variable predictoras realizamos unas de chi cuadrado.

Figura 11 Chi-Cuadrado Tipo Clientes Según Estado Civil

```
        Pearson's Chi-squared test

data:  BASE$TipoCliente and BASE$est_civil
X-squared = 216.18, df = 4, p-value < 2.2e-16
```

Fuente: Elaboración Propia.

El resultado del test de chi-cuadrado indica que hay una relación significativa entre "cliente" y "est_civil".

Hipótesis nulidad: No hay asociación entre las variables de estudio.

Hipótesis alternativa: Hay una asociación entre las variables.

Los resultados del test de chi-cuadrado sugieren que hay una relación significativa entre las variables "cliente" y "estado civil". El p-valor muy pequeño menor que 2.2e-16. Indica que la asociación observada no se debe al azar, sino a una verdadera asociación entre las variables.

Sin embargo, es importante tener en cuenta que el test de chi-cuadrado solo indica la presencia de una asociación, pero no proporciona información sobre la magnitud o la dirección de la asociación.

Examinar la correlación entre las variables independientes. Esto se puede hacer calculando coeficientes de correlación. La correlación entre variables puede ayudar a identificar relaciones lineales o redundancias en los datos.

Preparación De los Datos

3.3 Determinación de la muestra del modelo

La selección y consistencia de la muestra en un modelo de puntaje de crédito es un aspecto crucial para garantizar la validez y confiabilidad de los resultados del modelo estadístico.

Es fundamental que la muestra utilizada en el modelo de puntaje de crédito sea representativa de la población objetivo a la cual se desea aplicar el modelo. Esto implica seleccionar una muestra que refleje las características demográficas, socioeconómicas y de comportamiento crediticio de los clientes de la empresa que estamos realizando el estudio a la que se dirigirá el modelo. Una muestra no representativa puede llevar a sesgos y resultados poco confiables.

El tamaño de la muestra debe ser lo suficientemente grande como para garantizar resultados estadísticamente significativos y robustos para esto se tiene una data histórica de los clientes donde nos va a permitir diversas características de acuerdo a su historial de pago del cliente características que nos va definir los mejores y peores clientes de la empresa. Un tamaño de muestra pequeño puede llevar a estimaciones poco precisas y resultados poco confiables. Contar con un tamaño de muestra adecuado para permitir un análisis estadístico sólido y una evaluación efectiva de los coeficientes del modelo.

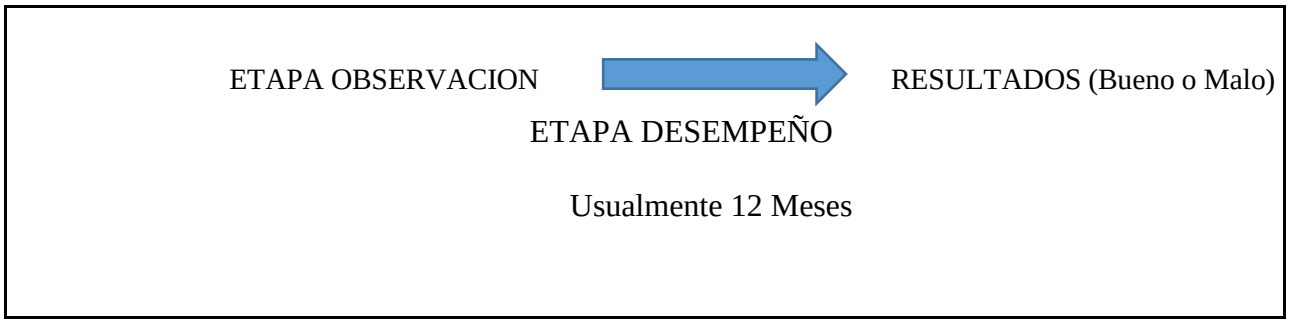
Es importante establecer el período de tiempo adecuado para la muestra. Esto implica seleccionar una ventana de tiempo que sea relevante para el propósito del modelo de puntaje de crédito. Por ejemplo, si se desea predecir el riesgo crediticio en un determinado año, la muestra debe incluir datos históricos de los años anteriores que sean representativos de la situación crediticia de los clientes en ese período.

Es esencial mantener la consistencia temporal en la muestra utilizada. Esto implica utilizar datos de clientes que hayan pasado por un período de tiempo similar en términos de eventos crediticios. Por ejemplo, si se utilizan datos de clientes que han pasado por una crisis económica, es importante asegurarse de que todos los clientes en la muestra hayan experimentado eventos crediticios en ese contexto específico.

Antes de utilizar una muestra en el modelo de puntaje de crédito, es crucial realizar un riguroso control de calidad de los datos. Esto implica verificar la integridad de los datos, identificar y corregir posibles errores. Para los datos atípicos o inconsistentes siempre y cuando no superen el 5% de las observaciones totales en el caso de que supere el valor antes mencionado se realizara un análisis por separado y se analizará la eliminación o conservación de la variable de estudio, y asegurarse de que los datos estén completos y actualizados.

Según Contreras “Usualmente se fija un periodo de 12 meses para un scoring de aprobación. La interrogante es determinar cuáles 12 meses se deben seleccionar para contar con información actualizada que a la vez recoja una madurez adecuada (comportamiento estable)”. (Contreras, 2005).

Figura 12 Muestra Estudio



Fuente: Elaboración Propia.

La etapa de observación y la etapa de desempeño son conceptos diferentes en el contexto de un modelo de puntaje de crédito.

Etapa de observación: La etapa de observación se refiere al período de tiempo durante el cual se recopilan los datos históricos para construir el modelo de puntaje crediticio.

Es la etapa en el cual se obtienen y registran las variables relevantes sobre los clientes, como historial crediticio, ingresos, deudas, entre otros. El objetivo de la etapa de observación es obtener la gran cantidad de datos necesaria para desarrollar el modelo logístico de riesgo crediticio.

Etapa de desempeño: La etapa de desempeño se refiere al período de tiempo en el cual se evalúa el rendimiento y la eficacia del modelo de puntaje de crédito. Es el periodo en el cual se utilizan los datos históricos recopilados durante el periodo de observación y se realizan pruebas para medir la precisión y la capacidad predictiva del modelo.

Durante la etapa de desempeño, se comparan las predicciones del modelo con los resultados reales para evaluar su capacidad para predecir el riesgo crediticio de manera efectiva.

Es importante destacar que la etapa de observación y la etapa de desempeño no necesariamente tienen que ser iguales. En algunos casos, pueden superponerse si se utilizan todos los datos disponibles para construir y evaluar el modelo. Sin embargo, en otros casos, se puede dividir el periodo de observación en una parte para construir el modelo y otra parte para evaluar su desempeño.

El periodo de observación es necesario para recopilar los datos históricos y construir un modelo sólido, mientras que el periodo de desempeño es necesario para evaluar la precisión y la eficacia del modelo en la predicción del riesgo crediticio.

Ambos periodos son importantes para el desarrollo y la validación de un modelo de puntaje crediticio efectivo.

3.4 Tiempo del periodo de modelización

Es fundamental realizar una cuidadosa selección del periodo que refleje de manera precisa la composición actual de la cartera de clientes, mostrando un comportamiento estable. La importancia de esta elección radica en obtener un análisis sólido y representativo de la situación actual de la empresa.

- **Madurez**

Según Bambino Contreras “Se considera a una cartera como madura cuando cuenta con el tiempo suficiente para determinar el comportamiento de pago del cliente; es decir, que pudo ser observada en un periodo igual al periodo de desempeño. Es importante la madurez de la cartera pues permite no calificar como bueno a un cliente malo, ya que si no se cuenta con el tiempo suficiente de observación no se podrá establecer su comportamiento de pago, de forma correcta”. (Bambino Contreras, 2005).

- **b) Estabilidad**

Según Bambino Contreras “Es importante considerar un periodo en el cual la población de clientes presenta un comportamiento estable, para de esta forma alcanzar la madurez adecuada de la cartera. Un indicador comúnmente usado es la tasa de morosidad, la cual es calculada para cada mes de desembolso o colocación (cosecha) dentro del periodo de modelización”. (Bambino Contreras, 2005).

Un indicador clave es la tasa de morosidad.

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

(21)

La tasa de morosidad se define por periodo de cosecha (fecha de venta del producto donde se genera el crediticio) y tiene por esencia mostrar la relación de clientes que no tiene un buen comportamiento de pagos conocimos como clientes malos versus el número total de clientes de forma gráfica y por mes de colocación para indicar un periodo en el que la causa se ha estabilizado. , equivalente a un comportamiento estable de la cartera .

Figura 13 Transición de Clientes

Fuente: Elaboración Propia.

Para el progreso de la tesis se considera una fase de ejecución equivalente a 12 meses que, como ya se mencionó, es la calificación original más común, por lo que, para el período de observación y el período de ejecución de junio de 2021 a junio de 2022, sólo se considerará la cosecha entre enero de 2019 y junio de 2021.

El periodo de modelización debe estar alineado con el objetivo del modelo de puntaje de crédito. Por ejemplo, si el objetivo es predecir el riesgo de incumplimiento en los próximos 12 meses, entonces el periodo de modelización debería abarcar al menos 12 meses de datos históricos para capturar la información relevante.

Es importante tener en cuenta la disponibilidad de datos al determinar el periodo de modelización. Si los datos históricos están disponibles para un periodo de tiempo específico, es recomendable utilizar ese periodo para construir el modelo. Sin embargo, si los datos son limitados o no están disponibles para ciertos periodos, se debe adaptar el periodo de modelización en función de los datos disponibles.

Se debe considerar la estabilidad de los datos a lo largo del periodo de modelización. Si hay cambios significativos en el comportamiento crediticio de los clientes durante ciertos periodos, puede ser necesario ajustar el periodo de modelización para tener una visión más consistente y representativa de la población objetivo.

Es importante realizar una evaluación de la precisión del modelo utilizando técnicas de validación cruzada o de conjunto de datos separados. Esto implica dividir los datos en un conjunto de entrenamiento y un conjunto de prueba para evaluar cómo se desempeña el modelo en datos no utilizados durante la construcción. El periodo de modelización debe permitir realizar esta evaluación adecuadamente.

Es recomendable establecer un plan de actualización periódica del modelo de puntaje de crédito. Esto implica definir un periodo de tiempo después del cual se revisa y actualiza el modelo para adaptarse a los cambios en el comportamiento crediticio de los clientes. El periodo de modelización puede estar vinculado a esta frecuencia de actualización.

La determinación del periodo de modelización en un modelo de puntaje de crédito debe considerar el objetivo del modelo, la disponibilidad de datos, la estabilidad de los datos, la evaluación de la precisión del modelo y la necesidad de actualización periódica. Se recomienda realizar un análisis cuidadoso de estos factores para seleccionar un periodo de modelización adecuado que brinde resultados confiables y relevantes para la gestión del riesgo crediticio.

3.5 Variable objetivo

Al aplicar un modelo de calificación crediticia, es necesario distinguir entre clientes de alto riesgo y clientes de bajo riesgo. Sin embargo, es importante tener en cuenta que clasificar a un cliente como de alto riesgo no significa automáticamente que todos los demás sean de bajo riesgo. Al menos dos categorías más pueden aparecer durante el proceso.

Por otro lado, hay casos "indefinidos" que no pueden clasificarse claramente como buenos o malos debido a información poco clara o incompleta. Por otro lado, existen casos de "experiencia insuficiente", que se refieren a aquellos clientes que tienen un historial de cuenta limitado o corto, lo que dificulta dar una evaluación final de su solvencia.

Cómo definimos lo bueno y lo malo afecta los resultados del puntaje de crédito. Por tanto, su definición requiere, entre otras cosas, del conocimiento de la cartera morosa de los clientes de la empresa, que se encuentra dentro de la gestión de cobranza.

El concepto de los mejores y peores pagadores tienen como fundamento generalmente en el historial de pago que realiza en un pedido de tiempo además de otras variables como: retraso máximo histórico, retraso promedio, calculadora de retraso (cuántas veces retraso o recurrencia).

Por lo tanto, la identificación es importante a través de la información de la cartera se puede determinar los clientes buenos y malos.

Si hay varios productos, se puede elegir una buena y una mala definición, porque se sabe que los productos del sistema bancarios o financieros no son iguales y sus diferentes tipos pueden influir en la capacidad de pago de los clientes. Por ej. la definición de un préstamo hipotecario puede ser más estricta que la de un producto de electrodomésticos. Es común que los clientes tengan retrasos en los pagos de 30 días, por lo que la definición de los bienes no debe ser tan ácida en este caso (muy pequeño retraso medio y retraso máximo). Aunque al final del día todo depende de los objetivos de mitigación de riesgos que la instalación quiera establecer.

Un método simple para determinar lo bueno y lo malo es una matriz de retraso promedio y retraso máximo, que enumera los intervalos de retraso máximos en filas y los intervalos de retraso promedio en columnas por pares de valores (retraso promedio; máximo). demora)

Tabla 3 Tabla Sobre Atraso Máximo Y Atraso Promedio

# Credito	Atraso Media					
	0 a 30	31 a 60	61 a 90	91 a 120	>120	Total general
Atraso Max						
0 a 30	36.900					36.900
31 a 60	3.562	301				3.863
61 a 90	1.187	1.098	16			2.301
91 a 120	196	1.303	239	1		1.739
> 120	4	766	2.027	920	71	3.788
Total general	41.849	3.468	2.282	921	71	48.591

Fuente: Elaboración Propia.

Tabla 4 Tabla Sobre Porcentaje Atraso Máximo Y Atraso Promedio

# Credito	Atraso Media					
	0 a 30	31 a 60	61 a 90	91 a 120	>120	Total general
Atraso Max						
0 a 30	75,94%					75,94%
31 a 60	7,33%	0,62%				7,95%
61 a 90	2,44%	2,26%	0,03%			4,74%
91 a 120	0,40%	2,68%	0,49%			3,58%
> 120	0,01%	1,58%	4,17%	1,89%	0,15%	7,80%
Total general	86,13%	7,14%	4,70%	1,90%	0,15%	100,00%

Fuente: Elaboración Propia.

Como variable dependiente identificamos a los clientes como buenos cuando tienen un atraso medio de hasta 30 días y un atraso máximo hasta 60 días esto representa 83% clientes buenos y el 17% clientes malos

3.6 Análisis y selección de las variables independientes

El análisis de las variables predictoras de una regresión sobre el puntaje de crédito es fundamental para identificar y comprender el impacto que cada variable tiene en la predicción del riesgo crediticio.

Se recopilan 28 variables independientes. Estos datos pueden incluir información como historial de clientes de la empresa, historial crediticio como ingresos, edad, nivel educativo, estado civil, días atraso, entre otros. Es importante contar con datos de calidad y completos para realizar un análisis preciso.

Se realiza un análisis individual de cada variable independiente. Esto implica examinar la distribución de los valores, identificar valores atípicos o faltantes, calcular estadísticas descriptivas y evaluar su relación con la variable objetivo (por ejemplo, el incumplimiento de pago). Esto se puede hacer mediante tablas de frecuencia, gráficos o medidas de correlación.

Se examina la correlación entre las variables predictoras y objetivo. Esto se puede hacer utilizando medidas de correlación como correlación de Pearson o Spearman. Es necesario identificar las variables que tienen una correlación significativa con la variable objetivo y que podrían ser buenos predictores del riesgo crediticio.

Se examina la presencia de multicolinealidad entre las variables independientes. La multicolinealidad ocurre cuando hay una alta correlación entre dos o más variables independientes, lo que puede dificultar la interpretación y precisión del modelo.

Si se detecta multicolinealidad a través factores de inflación de varianza (VIF), se pueden tomar medidas como eliminar una de las variables altamente correlacionadas o utilizar técnicas de regularización.

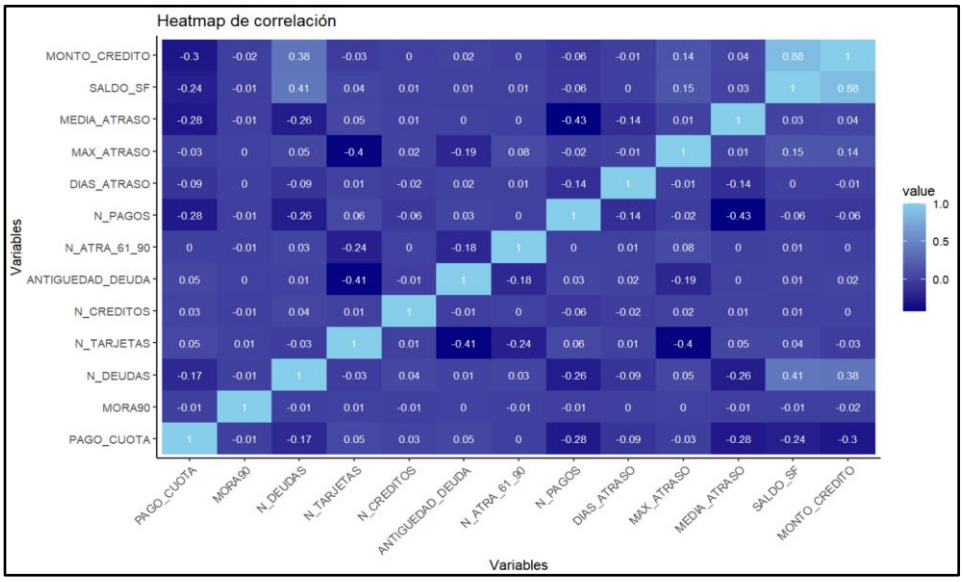
En algunos casos, es necesario realizar transformaciones en las variables de entrada para mejorar la relación con la variable objetivo o cumplir con los supuestos del modelo.

3.7 Variables independiente significativas

Es importante seleccionar variables que tengan una base teórica sólida y se espera que estén relacionadas con el riesgo crediticio. Esto implica comprender el contexto y la industria en la que se aplica el modelo y determinar qué variables pueden influir en la solvencia financiera del cliente.

Se evita incluir variables altamente correlacionadas entre sí, ya que esto puede afectar la interpretación del modelo y conducir a estimaciones inestables. Es recomendable eliminar una de las variables altamente correlacionadas o utilizar técnicas de regularización si se detecta multicolinealidad.

Figura 14 Correlación de variables predictoras



Fuente: Elaboración Propia.

La matriz de correlación de las variables predictoras nos proporciona información sobre si existen variables correlacionadas entre sí y evitar la multicolinealidad tenemos 1 variables como saldo_sf y monto_credito que están correlacionada en un 80% pero el 92% de las variables no presentan problemas de multicolinealidad.

Otra técnica que utilizamos para obtener las variables significativas es la de pasos hacia atrás (Backward Stepwise) es un enfoque para seleccionar un grupo óptimo de variables predictoras en un modelo logístico. Se incluye las 28 variables independientes y, en cada paso, elimina la variable que tiene el menor impacto en el modelo (mayor p-value) según el criterio de evaluación, como el valor AIC (Criterio de Información de Akaike) Este proceso continúa hasta que no hay más variables para elimina y se detiene el proceso.

El AIC es una medida de la calidad del ajuste del modelo que penaliza los modelos con un mayor número de parámetros. Por lo tanto, en el contexto de la técnica de pasos hacia atrás, el objetivo es encontrar el subconjunto más pequeño de variables que proporciona un buen ajuste del modelo y minimiza el AIC.

Es preferible seleccionar un conjunto de variables predictoras más pequeño y más simple en lugar de incluir todas las variables disponibles. Demasiadas variables pueden conducir a un modelo complejo y difícil de interpretar, además de aumentar el riesgo de sobreajuste para ello tenemos según los análisis de correlación de variables predictoras y los resultados Selección de características hacia atrás (Backward Feature Selection). se obtuvieron 13 variables para realizar el modelo logístico

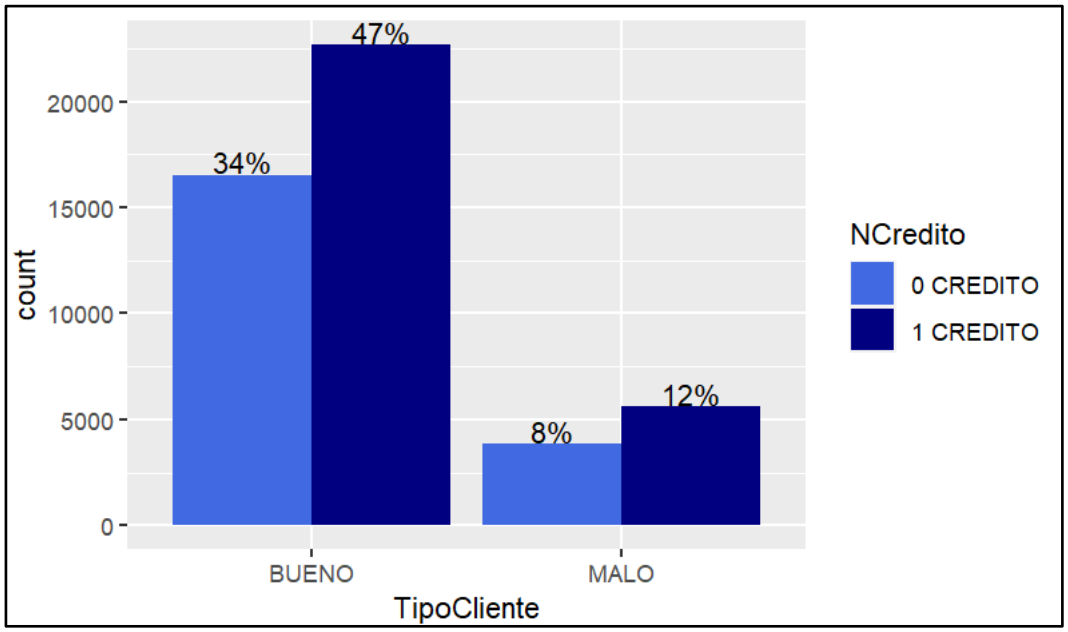
Tabla 5 Pasos Hacia Atrás (Backward Stepwise)

Variables	Deviance	AIC
N_TARJETAS	828,9	854,9
PAGO_CUOTA	830,8	856,8
N_DEUDAS	831,1	857,1
N_CREDITOS	833,1	859,1
DIAS_ATRASO	835,8	861,8
SALDO_SF	853,4	879,4
MAX_ATRASO	857,6	883,6
MONTO_CREDITO	865,2	891,2
ANTIGUEDAD_DEUDA	1129,2	1155,2
N_PAGOS	1191,1	1217,1
MEDIA_ATRASO	1809,4	1835,4
N_ATRA_61_90	5104,4	5130,4
MORA90	5857,3	5883,3

Fuente: Elaboración Propia.

Se evaluó la correlación entre las variables predictoras y la variable objetivo (el incumplimiento de pago). Variables que tengan una correlación significativa con la variable objetivo son más propensas a ser buenos predictores del riesgo crediticio.

Figura 15 Tipo Clientes Según Números Créditos



Fuente: Elaboración Propia.

Figura 16

Chi-Cuadrado Tipo Clientes Según Numero Crédito

```
        Pearson's Chi-squared test with Yates' continuity correction

data:  df$TipoCliente and df$NCredito
X-squared = 7.5357, df = 1, p-value = 0.006049
```

Fuente: Elaboración Propia.

Como el valor p es menor que un nivel de significancia típico (como 0.05 o 0.01), podemos rechazar la hipótesis nula de independencia y concluir que hay una asociación significativa entre las variables.

Figura 17 Tipo Clientes Según Monto Crédito

	0-1000 MC	1001-2000 MC	2001-3000 MC	3000+ MC
BUENO	14397	14189	5031	667
MALO	3043	3569	1448	94

Fuente: Elaboración Propia.

Figura 18 Chi-Cuadrado Tipo Clientes Según Monto Crédito

```
        Pearson's Chi-squared test

data:  tabla_contingencia
X-squared = 108.08, df = 3, p-value < 2.2e-16
```

Fuente: Elaboración Propia.

Dado que el p-value es menor que 0.05 (p-value < 2.2e-16), podemos concluir que hay una asociación significativa entre las variables "rango monto del crédito" y "tipo de cliente". En otras palabras, el monto del crédito y el tipo de cliente están relacionados y no son independientes entre sí.

Figura 19 Tipo Clientes Según Saldo Sistema Financiero

	0-1000 SALDO	1001-2000 SALDO	2001-3000 SALDO	3000+ SALDO
BUENO	21191	10258	2348	413
MALO	3756	3008	804	70

Fuente: Elaboración Propia.

Figura 20 Chi-Cuadrado Tipo Clientes Según Saldo Sistema Financiero

```
        Pearson's Chi-squared test

data:  tabla_contingencia_saldo
X-squared = 460.48, df = 3, p-value < 2.2e-16
```

Fuente: Elaboración Propia.

El resultado del test de chi-cuadrado indica que hay una asociación significativa entre las variables "TipoCliente" y "SALDO_SF" en el conjunto de datos. La estadística de prueba (X-squared) es de 460.48 y los grados de libertad (df) son 3. El valor p (p-value) es muy pequeño, menor que 2.2e-16, lo que indica que la asociación entre estas variables no es el resultado del azar.

Figura 21 Wilcoxon Tipo Clientes Según Máximo Atraso

```
        Wilcoxon rank sum test with continuity correction

data:  MAX_ATRASO by TipoCliente
W = 37280709, p-value < 2.2e-16
alternative hypothesis: true location shift is not equal to 0
```

Fuente: Elaboración Propia.

El resultado del Wilcoxon rank sum test con corrección de continuidad (también conocido como prueba de Mann-Whitney U) indica que hay evidencia estadística significativa para concluir que existe una diferencia en la distribución de la variable "MAX_ATRASO" entre las dos categorías de "TipoCliente".

El valor p del resultado es extremadamente pequeño (p-value < 2.2e-16), lo que significa que la probabilidad de obtener una diferencia tan grande o más extrema en las medianas de "MAX_ATRASO" entre las categorías de "TipoCliente" bajo la hipótesis nula (que no hay diferencia) es prácticamente nula. Por lo tanto, rechazamos la hipótesis nula y concluimos que hay una diferencia significativa entre las medianas de "MAX_ATRASO" para los dos grupos de "TipoCliente".

Considera la importancia práctica de las variables. Algunas variables pueden tener una correlación estadística significativa con la variable objetivo, pero pueden tener poca relevancia práctica o capacidad predictiva en el contexto específico del problema.

CAPITULO 4

4. Resultados de la investigación

4.1 Elaboración del modelo y estadísticos

En este capítulo realizamos el modelo de regresión logística, mediante le paquete Rstudio utilizando el paquete glm se utiliza para ajustar el modelo de regresión logística. La opción family = binomial especifica que se trata de una regresión logística binomial. Para realizar predicciones en nuevos datos utilizando la función predict(), y especificando type = "response" para obtener las probabilidades de clasificación. Finalmente, puedes obtener un resumen del modelo utilizando summary (modelo).

Luego de seleccionar variables independientes que se someten a análisis exploratorio y estadístico realizado en el capítulo anterior, se utilizan métodos de regresión logística para obtener modelos puntuales que permitan el logro de los objetivos.

Además, las variables categóricas se recodifico las variables y se transformó cada categoría en una variable dicotómica

Modelado

4.2 Regresión Logística

Realizamos el modelo en RStudio utilizando la función glm() para ajustar el modelo. Aquí hay algunas consideraciones sobre se incluyen las variables independientes en el modelo de regresión logística. A través de resumen mostrará los coeficientes estimados, los valores p y otras estadísticas relacionadas con el modelo ajustado. Al utilizar en Rstudio la función step, que permite obtener el modelo basado en el criterio de información de Akaike (AIC por sus siglas en inglés), realizando el proceso de eliminación de paso hacia atrás de las variables (Backward Stepwise).

La salida del resumen te proporcionará información sobre la significancia de cada variable independiente y otros detalles del modelo se presenta los resultados del modelo de puntaje crediticio.

Tabla 6 Modelo de Puntaje Crediticio

Coefficients:	Estimate	Std. Error	z value	Pr(>|z|)	Signif. codes
(Intercept)	-2,03011	0,21064	-9,63788	0,00000	***
PAGO_CUOTA	1,39106	0,21710	6,40755	0,00000	***
MORA90	-5,65900	65,82682	-0,08597	0,00093	
N_DEUDAS	3,96280	0,22743	17,42390	0,00000	***
N_TARJETAS	-5,65189	0,19255	-29,35321	0,00000	***
N_CREDITOS	-0,02093	0,06887	-0,30390	0,00761	
ANTIGUEDAD_DEUDA	-6,16268	0,17081	-36,07930	0,00000	***
N_ATRA_61_90	-6,17450	0,15464	-39,92861	0,00000	***
N_PAGOS	2,51438	0,20457	12,29123	0,00000	***
DIAS_ATRASO	2,84561	0,24497	11,61625	0,00000	***
MAX_ATRASO	0,13279	0,00320	41,49786	0,00000	***
MEDIA_ATRASO	2,55441	0,20896	12,22421	0,00000	***
SALDO_SF	0,00280	0,00016	17,24782	0,00000	***
MONTO_CREDITO	-0,00462	0,00015	-30,87950	0,00000	***

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

Fuente: Elaboración Propia.

El análisis de los coeficientes proporciona información sobre la relación entre las variables predictoras y la variable resultados en el modelo logístico.

Intercepto (Intercept): El intercepto se estima en -2,030. Este valor representa el logaritmo de la odds (probabilidad logarítmica) de que la variable dependiente tome el valor 1 (Buen Cliente) cuando todas las variables independientes son iguales a cero.

Los coeficientes de las variables de entrada indican cómo cada variable independiente contribuye al logaritmo de las odds de que la variable predictora sea el valor 1 (Buen Cliente), después de controlar las otras variables en el modelo.

Para cada coeficiente, también se proporciona el Error Estándar (Std. Error), el valor z (z value) y el valor p (Pr(>|z|)) asociado. El valor z y el valor p indican la significancia estadística del coeficiente. Un valor p menor que 0.05 se considera generalmente como estadísticamente significativo.

En la columna Signif. codes, se proporcionan códigos para indicar el nivel de significancia de cada coeficiente. Los códigos más comunes son " ($p < 0.001$), " ($p < 0.01$), " ($p < 0.05$), '.' ($p < 0.1$), y ' ' ($p > 0.1$).

Tabla 7 Residuos De La Desviaciones

Min	1Q	Median	3Q	Max
-2.8083	-0.0672	-0.0258	-0.0053	5.3750

Fuente: Elaboración Propia.

Los residuos de la desviación (Deviance Residuals) es una métrica de cómo las observaciones individuales se ajustan al modelo de regresión logística:

- El valor mínimo de los residuos de desviación es -2.8083.
- El primer cuartil (25%) de los residuos de desviación es -0.0672.
- La mediana (50%) de los residuos de desviación es -0.0258.
- El tercer cuartil (75%) de los residuos de desviación es -0.0053.
- El valor máximo de los residuos de desviación es 5.3750.

Estos valores describen la variabilidad y distribución de los residuos de desviación en el modelo de regresión logística. Un valor inferior a cero nos proporciona que las observaciones son menores que el valor esperado según el modelo, mientras que un valor positivo indica que el valor observado es mayor que el valor esperado.

Los residuos de la desviación tienen valores entre -2.8083 y 5.3750. En general, valores de residuos cercanos a cero indican un buen ajuste del modelo, lo que significa que las predicciones del modelo se acercan bastante a los valores reales observados.

Tabla 8 Comparativo Desviaciones

Null deviance:	47817.4 on 48590 degrees of freedom
Residual deviance:	5905.2 on 48577 degrees of freedom
AIC: 5933.2	

Fuente: Elaboración Propia.

La salida obtenida muestra los valores de la deviance nula (Null deviance), los residuos de la desviación (Residual deviance) y el criterio de información de Akaike (AIC)

Deviance nula (Null deviance): La deviance nula es una medida de cuánto mejor se ajusta el modelo completo en comparación con un modelo nulo que no incluye ninguna variable independiente. Un valor alto de la deviance nula indica que el modelo completo es significativamente mejor que el modelo nulo. En tu caso, la deviance nula es 47817.4.

Deviance residual (Residual deviance): La deviance residual mide cuánto error aún queda en el modelo después de haber ajustado las variables independientes. Un valor bajo de la deviance residual indica que el modelo se ajusta bien a los datos. la deviance residual es 5905.2.

Grados de libertad (Degrees of freedom): Los grados de libertad representan la cantidad de información disponible para estimar los parámetros del modelo. En el caso de la deviance nula, los grados de libertad son 48.590, mientras que en la deviance residual, son 48.577.

La deviance nula, con un valor de 47817.4, representa la diferencia de deviance entre el modelo completo (con todas las variables independientes incluidas) y un modelo nulo que no incluye ninguna variable independiente. Cuanto mayor sea la deviance nula, mayor es la diferencia y mejor se ajusta el modelo completo en comparación con el modelo nulo. Esto sugiere que el modelo completo está proporcionando una mejora significativa en la explicación de los datos en comparación con el modelo nulo.

La deviance residual, con un valor de 5905.2, representa la diferencia de desviación entre el modelo ajustado y los datos observados. Cuanto menor sea la desviación residual, mejor será el ajuste del modelo a los datos. La deviance residual indica que el modelo logra explicar una parte sustancial de la variabilidad en los datos, aunque aún existe cierto error residual.

El criterio de Akaike (AIC), con un valor de 5933.2, es una medida que combina la calidad de ajuste del modelo y la simplicidad del modelo. Un valor más bajo de AIC indica un mejor ajuste y una mayor parsimonia del modelo el valor de AIC indica que el modelo estadístico tiene una un ajuste razonable a los datos.

4.3 Significancia de los coeficientes

Para realizar una prueba de hipótesis sobre los coeficientes de las variables en un modelo de regresión logística, se puede utilizar el valor z y el valor p asociados a cada coeficiente.

Aquí hay una descripción general de cómo hacer unas pruebas estadísticas para determinar si los coeficientes estimados en un modelo son significativamente diferentes de cero o no

Definir la hipótesis nulidad y la hipótesis alternativa para cada coeficiente de interés. La hipótesis nula generalmente asume que el coeficiente es igual a cero, mientras que la hipótesis alternativa afirma que el coeficiente es diferente de cero (Ronald A. Fisher, 1925)

Calcula el valor z: El valor z se obtiene dividiendo el coeficiente estimado por su error estándar. Por ejemplo, para el coeficiente "ATRASO_MAXIMO" con un valor estimado de 0,133 y un error estándar de 0,003, el valor z sería 0,133 / 0,003 = 41,498.

Calcula el valor p: El valor p se utiliza para medir si los coeficientes son estadísticamente significativos. Representa la probabilidad de observar un valor igual o más extremo que el coeficiente estimado bajo la hipótesis nulidad. Un valor p bajo (generalmente menor que 0.05) sugiere evidencia en contra de la hipótesis nulidad. Puedes obtener el valor p directamente de la salida del modelo.

Compara el valor p con un nivel de significancia predefinido (por ejemplo, 0.05). Si el valor p es menor que el nivel de significancia, rechazar la hipótesis nula y aceptas la hipótesis alternativa. Si el valor p es mayor que el nivel de significancia, no tienes suficiente evidencia para rechazar la hipótesis nula.

Es importante tener en cuenta que la interpretación de los resultados y la toma de decisiones dependen del contexto y los objetivos del estudio. Además, tener en cuenta que realizar múltiples pruebas de hipótesis puede aumentar el riesgo de errores de tipo I (rechazar incorrectamente la hipótesis nula). En esos casos, se pueden aplicar ajustes, como el ajuste de Bonferroni, para controlar el nivel global de significancia.

En la salida obtenida, los coeficientes estimados para la mayoría de las variables independientes son estadísticamente significativos. Esto se indica por el valor p asociado (Pr(>|z|)) que se muestra en la columna "Pr(>|z|)".

El valor p es < menor que 0.05 (indicado por "***"), lo que proporciona una significancia estadística a un nivel de confianza del 95%. Esto significa que hay evidencia suficiente para rechazar la hipótesis nula de que el coeficiente correspondiente a cada variable es cero, lo que sugiere que estas variables tienen un efecto significativo en la variable dependiente.

4.4 Interpretación de los coeficientes

Los análisis de los coeficientes y los signos de los parámetros de las variables de la regresión logística se refieren a la dirección y magnitud de los coeficientes estimados para cada variable independiente.

Observamos la mayoría de los coeficientes tienen signos consistentes. Por ejemplo, los coeficientes para las variables que tienen signos positivos, lo que indica una relación positiva con la variable dependiente. Esto implica que un aumento en estos predictores se asocia con un mayor logaritmo de las odds de la variable dependiente.

La interpretación de los coeficientes debe realizarse considerando el contexto y las características de los datos. Además, tener en cuenta que la significancia estadística de los coeficientes también es importante para evaluar su confiabilidad.

Se observa una consistencia de signos en la mayoría de las variables, con coeficientes positivos para la mayoría de los predictores significativos. Sin embargo, hay algunas variables con coeficientes negativos o no significativos, lo que puede indicar una relación más débil o no significativa con la variable dependiente en el contexto del modelo y los datos utilizados.

Para calcular los odds ratio y analizar cada coeficiente en tu modelo de regresión logística, puedes utilizar la siguiente fórmula:

$$\text{Odds Ratio} = e^{\text{Coeficiente}}$$

$$e=2.718$$

(22)

La función exponencial (exp) se aplica al coeficiente estimado para obtener el odds ratio correspondiente. El odds ratio representa cómo cambia la odds (probabilidad logarítmica) de que la variable predictora tome el valor 1(clientes buenos) por cada unidad de cambio en la variable independiente, manteniendo constantes las demás variables en el modelo.

Al analizar cada coeficiente, los odds ratios proporcionan una medida de la asociación entre la variable independiente correspondiente y la variable dependiente, considerando los otros predictores en el modelo.

Un odds ratio mayor que 1 indica que un incremento en la variable independiente asocia con un aumento en las odds (probabilidad logarítmica) cuando la variable dependiente presenta un resultado positivo 1 (Cliente Bueno).

Para la variable " DIAS_ATRASO" con un odds ratio de 17.2149782, se puede decir que cada unidad adicional en " DIAS_ATRASO" se asocia, en promedio, con un aumento de aproximadamente 17.21 veces en las odds cuando la variable dependiente presenta un resultado positivo 1 (Cliente Bueno), manteniendo constantes las otras variables en el modelo.

Por otro lado, un odds ratio menor que 1 indica que un aumento en la variable predictora se relaciona con una reducción en las odds cuando la variable dependiente presenta un resultado positivo 1 (Cliente Bueno). Por ejemplo, para la variable " N_TARJETAS" con un odds ratio de 0.0035254, se puede decir que cada unidad adicional en " N_TARJETAS" se asocia, en promedio, con una disminución de aproximadamente 99.65% en las odds cuando la variable dependiente presenta un resultado positivo 1 (Cliente Buenos), si las otras variables predictoras se mantienen constante.

MAX_ATRASO: esta ok remplazar las demás

- El coeficiente estimado es 0.1327893.
- El odds ratio correspondiente es 1.1429483.
- Interpretación: Un incremento de cada unidad adicional en el "ATRASO_MAXIMO" se relaciona, en promedio, con un incremento de aproximadamente un 14% en las odds cuando la variable predictora presenta un resultado positivo 1 (Cliente Buenos), si las otras variables predictoras se mantienen constante.

SALDO_SF:

- El coeficiente estimado es 0.0028016.
- El odds ratio correspondiente es 1.0028054.
- Interpretación: Un incremento de cada unidad adicional en el "SALDO_ACTUAL" se relaciona, en promedio, con un incremento de aproximadamente un 0,28% en las odds cuando la variable predictora presenta un resultado positivo 1 (Cliente Buenos), si las otras variables predictoras se mantienen constante.

MONTO_CREDITO:

- El coeficiente estimado es -0.0046222.
- El odds ratio correspondiente es 0.9953864.
- Interpretación: Un incremento de cada unidad adicional en el "MONTO_PROMEDIO" se relaciona, en promedio, con un incremento de aproximadamente un 0,46% en las odds cuando la variable predictora presenta un resultado positivo 1 (Cliente Buenos), si las otras variables predictoras se mantienen constante.
-

ANTIGUEDAD_DEUDA:

- El coeficiente estimado es -6.1626799.
- El odds ratio correspondiente es 0.0020657.
- Interpretación: Un incremento de cada unidad adicional en la "ANTIGUEDAD_DEUDA" se relaciona, en promedio, con un incremento de aproximadamente un 99,79% en las odds

cuando la variable predictora presenta un resultado positivo 1 (Cliente Buenos), si las otras variables predictoras se mantienen constante.

N_ATRA_61_90:

- El coeficiente estimado es -6.1744981.
- El odds ratio correspondiente es 0.0019995.
- Interpretación: Un incremento de cada unidad adicional en el " N_ATRA_61_90" se relaciona, en promedio, con un incremento de aproximadamente un 99,80% en las odds cuando la variable predictora presenta un resultado positivo 1 (Cliente Buenos), si las otras variables predictoras se mantienen constante.

N_PAGOS:

- El coeficiente estimado es 2.5143799.
- El odds ratio correspondiente es 12.3571943.
- Interpretación: Un incremento de cada unidad adicional en el " N_PAGOS" se relaciona, en promedio, con un incremento de aproximadamente un 12,36 veces en las odds cuando la variable predictora presenta un resultado positivo 1 (Cliente Buenos), si las otras variables predictoras se mantienen constante.

PAGO_CUOTA:

- El coeficiente estimado es 1.3910574.
- El odds ratio correspondiente es 4.0155042.
- Interpretación: Un incremento de cada unidad adicional en el " PAGO_CUOTA" se relaciona, en promedio, con un incremento de aproximadamente un 4,02 veces en las odds cuando la variable predictora presenta un resultado positivo 1 (Cliente Buenos), si las otras variables predictoras se mantienen constante.

MEDIA_ATRASO:

- El coeficiente estimado es 2.5544140.
- El odds ratio correspondiente es 12.8665701.
- Interpretación: Un incremento de cada unidad adicional en la " MEDIA_ATRASO" se relaciona, en promedio, con un incremento de aproximadamente un 12,87 veces en las odds cuando la variable predictora presenta un resultado positivo 1 (Cliente Buenos), si las otras variables predictoras se mantienen constante.

N_DEUDAS:

- El coeficiente estimado es 3.9627955.
- El odds ratio correspondiente es 52.5093472.
- Interpretación: Un incremento de cada unidad adicional en el " N_DEUDAS" se relaciona, en promedio, con un incremento de aproximadamente un 52,51 veces en las odds cuando la variable predictora presenta un resultado positivo 1 (Cliente Buenos), si las otras variables predictoras se mantienen constante.

N_CREDITOS:

- El coeficiente estimado es -0,020929.
- El odds ratio correspondiente es 0,9792881.
    - Interpretación: Un incremento de cada unidad adicional en el "N_DEUDAS" se relaciona, en promedio, con un incremento de aproximadamente un 0,97 veces en las odds cuando la variable predictora presenta un resultado positivo 1 (Cliente Buenos), si las otras variables predictoras se mantienen constante.

MORA90:

- El coeficiente estimado es -5,658997.
- El odds ratio correspondiente es 0,0034860.
    - Interpretación: Un incremento de cada unidad adicional en la " MORA90" se relaciona, en promedio, con un incremento de aproximadamente un 0,003veces en las odds

cuando la variable predictora presenta un resultado positivo 1 (Cliente Buenos), si las otras variables predictoras se mantienen constante.

4.5 Estadísticos y análisis de Multicolinealidad

Cuando las variables predictoras en un modelo de regresión no son independientes entre sí, se produce multicolinealidad, lo que puede tener varios efectos negativos en la estimación de los parámetros y en la interpretación del modelo (O'Brien, 2007).

Prediccion de coeficientes con signos negativos: a multicolinealidad puede causar que los coeficientes estimados tengan signos opuestos a los esperados. Esto ocurre porque la correlación entre las variables predictoras dificulta la atribución adecuada de la influencia de cada variable en la variable dependiente.

Coeficientes con magnitud inverosímil: La multicolinealidad puede resultar en coeficientes estimados con magnitudes inverosímiles o poco confiables. Los coeficientes pueden volverse demasiado grandes o demasiado pequeños debido a la influencia combinada de las variables correlacionadas.

Sensibilidad a pequeños cambios en los datos: La multicolinealidad hace que el modelo sea muy sensible a pequeñas variaciones en los datos de entrada. Cambios mínimos en las observaciones pueden generar grandes cambios en los coeficientes estimados, lo que dificulta la estabilidad y la interpretación confiable del modelo.

Dificultad en la solución numérica del modelo: En casos extremos de multicolinealidad, el modelo puede volverse numéricamente inestable y es posible que no se obtenga una solución adecuada. Esto puede provocar errores o dificultades en la estimación de los parámetros.

Para abordar la multicolinealidad, es recomendable realizar diagnósticos adecuados, como el cálculo de los factores de inflación de la varianza (VIF) o los factores de inflación generalizados de la varianza (GVIF), y tomar medidas como eliminar variables altamente correlacionadas o utilizar técnicas de selección de variables.

El factor de inflación generalizada de la varianza (GVIF) es una medición que evalúa la multicolinealidad en un modelo de regresión. Mide cuánto aumenta la varianza de los coeficientes estimados debido a la correlación entre las variables predictoras. Un GVIF menor a 10 indica un menor multicolinealidad.

Análisis de los resultados del factor de inflación de la varianza (VIF) para cada variable del modelo:

- **MAX_ATRASO:** El VIF de 5.760129 indica que existe una moderada correlación entre esta variable y las demás variables en el modelo. No hay una preocupante multicolinealidad asociada a esta variable.

- **SALDO_SF:** El VIF de 6.925818 indica que hay cierta correlación entre esta variable y las demás variables, pero no llega a niveles preocupantes de multicolinealidad.

- **MONTO_CREDITO:** El VIF de 8.265239 muestra cierta correlación entre esta variable y las demás, pero no alcanza niveles preocupantes de multicolinealidad.

- **N_TARJETAS:** El VIF de 1.161826 muestra que no existe una correlación significativa entre el número de tarjetas y las otras variables en la regresión logística.

- **ANTIGUEDAD_DEUDA:** El VIF de 1.325003 indica que existe cierta correlación entre esta variable y las demás variables, pero no llega a niveles preocupantes de multicolinealidad.

- **N_ATRA_61_90:** El VIF de 4.520555 muestra que no existe una correlación significativa entre esta variable y las otras variables en la regresión logística.

- **N_PAGOS:** El VIF de 8.825239 indica que hay cierta correlación entre esta variable y las demás variables, pero no alcanza niveles preocupantes de multicolinealidad.

- **DIAS_ATRASO:** El VIF de 2.995656 muestra cierta correlación entre esta variable y las demás, pero no llega a niveles preocupantes de multicolinealidad.

- **PAGO_CUOTA:** El VIF de 5.500989 indica que existe cierta correlación entre esta variable y las demás variables, pero no llega a niveles preocupantes de multicolinealidad.

- **MEDIA_ATRASO:** El VIF de 8.489761 muestra que no existe una correlación significativa entre esta variable y las otras variables en la regresión logística.

- **N_DEUDAS:** El VIF de 5.248334 muestra cierta correlación entre esta variable y las demás, pero no alcanza niveles preocupantes de multicolinealidad.

- **N_CREDITOS:** El VIF de 1.005430 muestra cierta correlación entre esta variable y las demás, pero no alcanza niveles preocupantes de multicolinealidad.

- **MORA90:** El VIF de 1.000020 muestra cierta correlación entre esta variable y las demás, pero no alcanza niveles preocupantes de multicolinealidad.

Las variables presentan cierta correlación con las demás, pero no se detecta una multicolinealidad preocupante en el modelo. Esto implica que los coeficientes estimados y la interpretación de las variables son confiables.

Evaluación Del Modelo

4.6 Métricas de rendimiento y evaluación

Realizamos una función Prueba_Modelo diseñada para medir un modelo logístico utilizando varios indicadores de rendimiento. Además, se utilizó un 30% para realizar las pruebas y un 70% entrenamiento. A continuación, se describe brevemente cada uno de los indicadores que se calculan dentro de la función.

Coeficiente de Gini: La medida de Gini es utilizada para cuantificar el grado de desigualdad que se utiliza comúnmente en el contexto de la clasificación. En este caso, se calcula el coeficiente de Gini para evaluar la capacidad predictiva del modelo de regresión logística.

Matriz de confusión: Se procede a realizar el cálculo de la matriz de confusión, que proporciona información sobre la cantidad de verdaderos negativos, verdaderos positivos, falsos negativos y falsos positivos obtenidos por el modelo..

Promedio de predicción: Se calcula el promedio de la variable de predicción para evaluar la precisión general del modelo.

Área bajo la curva ROC: Se calcula el área bajo la curva ROC (AUC-ROC) para evaluar la capacidad discriminativa del modelo.

Kolmogorov-Smirnov (KS): Se calcula el estadístico de Kolmogorov-Smirnov (KS) para evaluar la diferencia máxima entre las tasas de falsos positivos y verdaderos positivos a lo largo de diferentes umbrales de clasificación.

Prueba de Hosmer-Lemeshow: Se realiza la prueba de Hosmer-Lemeshow para evaluar el ajuste del modelo.

La función devuelve una lista que contiene los resultados de estos indicadores, así como otros objetos como la matriz de confusión, la curva ROC y los datos utilizados en el análisis.

- **Curva Roc**

Se crea una función TEST_MODELO para obtener 7 métricas de rendimiento entre una de ellas la Curva de Característica Operativa del Receptor (ROC) Para obtener el valor del área bajo la curva se lo realiza mediante el fragmento de código que se encuentra dentro de la función que presenta a continuación:

```
data.roc ← prediction(data %>%
                        select(probabilidades) %>%
                        pull(), data$default)
```

(23)

```
area_roc ← slot(performance(data.roc,"auc"), "y.values")[[1]]
```

(24)

Para realizar la consulta valor de área bajo la curva ROC (AUC):

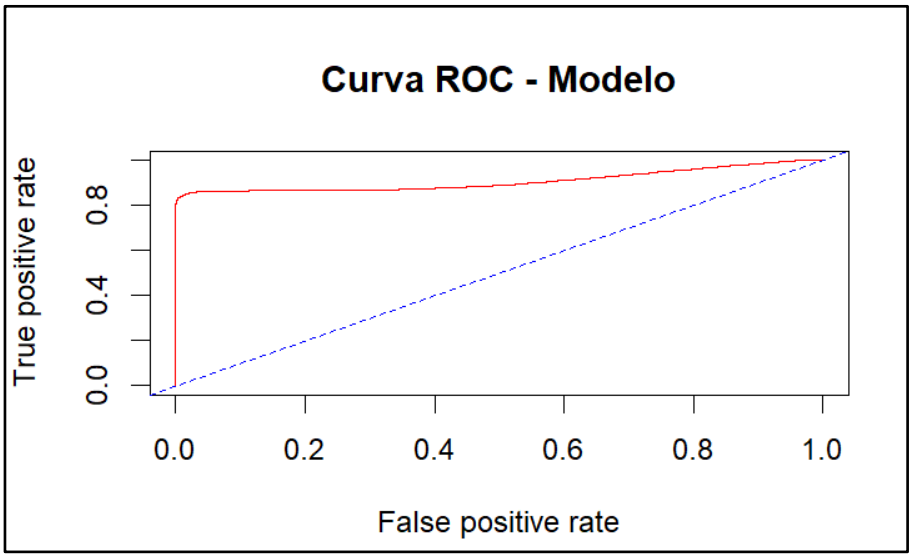
```
Modelo_Regre ← Prueba_Modelo(Regresion_Logistica,Base_Modelo %>%
                               rename(default = OBJETIVO),0.3)
Modelo_Regre[3:3]
```

(25)

Con un valor de área bajo la curva ROC (AUC) de 0.9064342, podemos hacer el siguiente análisis del gráfico ROC.

El gráfico muestra un rendimiento excepcionalmente bueno del modelo de regresión logísti.ca. La curva ROC está muy cerca de la esquina superior izquierda del gráfico, lo que indica una alta sensibilidad (tasa de verdaderos positivos) y una baja tasa de falsos positivos.

Figura 22 Curva Roc



Fuente: Elaboración Propia.

Un valor de AUC cercano a 1 indica que el modelo tiene una capacidad sobresaliente para distinguir entre las clases positiva y negativa. Esto sugiere que el modelo es altamente efectivo en la clasificación de las observaciones correctamente. En este caso, un AUC de 0.9064342 indica una alta precisión y un buen poder de discriminación del modelo.

El hecho de que la curva ROC esté bastante cerca de la línea de referencia (línea punteada azul) indica que el modelo tiene un rendimiento significativamente mejor que un clasificador aleatorio. La distancia entre la curva ROC y la línea de referencia también puede indicar la robustez y la calidad del modelo.

- **Promedio De Predicción**

Se crea una función TEST_MODELO para obtener 7 métricas de rendimiento entre una de ellas el Promedio De Predicción calcula la proporción media de predicciones correctas del modelo de regresión logística se lo realiza mediante el fragmento de código que se encuentra dentro de la función que presenta a continuación:

```
media_prediccion ← mean(data %>%
                          select(prediccion) %>%
                          pull() == default_mod)
```

(25)

Para realizar la consulta el Promedio De Predicción calcula la proporción media de predicciones correctas del modelo de regresión logística:

```
Modelo_Regre ← Prueba_Modelo(Regresion_Logistica,Base_Modelo %>%
                              rename(default = OBJETIVO),0.3)
Modelo_Regre[1:1]
```

(26)

El valor de "PROMEDIO DE PREDICCION" de 0.9460188 indica el promedio de predicciones correctas del modelo de regresión logística. El valor de 0.9460188 significa que, en promedio, el modelo clasifica correctamente aproximadamente el 94.60% de las observaciones en la muestra de datos utilizada. Esto implica que el modelo tiene una alta capacidad de pronosticar correctamente (en este caso, la variable "Clientes Buenos") en la mayoría de las observaciones.

- **Coeficiente De Gini**

En el caso de la regresión logística, el coeficiente de Gini mide la desigualdad en la distribución de las predicciones del modelo. Un valor de 0.2019414 indica una ligera desigualdad en la distribución de las predicciones del modelo. Cuanto más cercano a 0 es el valor del coeficiente de Gini, menor es la desigualdad y mejor es la capacidad del modelo para clasificar correctamente las observaciones

- **Matriz De Confusión**

Se crea una función TEST_MODELO para obtener 7 métricas de rendimiento entre una de ellas la Matriz De Confusión se lo realiza mediante el fragmento de código que se encuentra dentro de la función que presenta a continuación:

```
data ← data %>%
  mutate(probabilidades = predict(logit_regre, type = "response"),
         prediccion = ifelse(probabilidades > prob_RL, 1, 0))
```

(27)

```
matriz_confusion ← confusionMatrix(table(data %>%
                                         select(prediccion) %>%
                                         pull(), default_mod))
```

(28)

Desarrollo de la Tabla De Confusión:

```
Modelo_Regre ← Prueba_Modelo(Regresion_Logistica,Base_Modelo %>%
                             rename(default = OBJETIVO),0.3)
Modelo_Regre[6:6]
```

(29)

Figura 23 Matriz De Confusión

```
$`MATRIZ DE CONFUSION - INDICADORES`
Confusion Matrix and Statistics

   default_mod
       0     1
 0 37864  1336
 1  1298  8093

               Accuracy : 0.9458
                 95% CI : (0.9437, 0.9478)
    No Information Rate : 0.806
    P-Value [Acc > NIR] : <2e-16

                  Kappa : 0.8264

 Mcnemar's Test P-Value : 0.471

            Sensitivity : 0.9669
            Specificity : 0.8583
         Pos Pred Value : 0.9659
         Neg Pred Value : 0.8618
             Prevalence : 0.8060
         Detection Rate : 0.7792
   Detection Prevalence : 0.8067
      Balanced Accuracy : 0.9126
```

Fuente: Elaboración Propia.

Donde 0 = Clientes buenos

Donde 1 = Clientes malos

En la matriz de confusión, los valores de la columna "Tipo Clientes Buenos o Malos" representan las clases reales (Buenos, Malos), y los valores de las filas representan las clases predichas por el modelo.

Verdaderos positivos (VP): Representa la cantidad de observaciones en los cuales el modelo logístico acertó al predecir el grupo de clase positiva (Buenos), coincidiendo con la clase real que también es positiva. En esta situación, el valor obtenido es 37879.

Falsos positivos (FP): Representa la cantidad de observaciones en los cuales el modelo logístico predijo incorrectamente el grupo de clase positiva (Buenos), pero la clase real fue negativa. En este caso, el valor es 1336.

Falsos negativos (FN): Representa la cantidad de observaciones en los cuales el modelo logístico pronostico incorrectamente los grupos de clase negativa (Malos), pero la clase real fue positiva. En este caso, el valor es 1298.

Verdaderos negativos (TN): Representa la cantidad de observaciones en los cuales el modelo logístico pronostico correctamente los grupos de clase negativa (Malos) y la clase real también fue negativa. En este caso, el valor es 8093.

Estos indicadores de la matriz de confusión se utilizan para calcular diversas métricas de evaluación del modelo, como la precisión, el recall, la tasa de verdaderos positivos (TPR), la tasa de falsos positivos (FPR) y la puntuación F1, entre otras.

- **Métricas De Evaluación**

Accuracy (Exactitud): El valor de 0.946 indica la proporción de predicciones correctas realizadas por el modelo sobre el total de observaciones. En otras palabras, el modelo clasifica correctamente aproximadamente el 94.6% de las observaciones.

95% CI (límites de Confianza): Indica que los intervalos límites de confianza del 95% para la exactitud del modelo. En este caso, el intervalo de confianza está entre 0.944 y 0.948. Esto significa que podemos tener un 95% de confianza de que la exactitud real del modelo se encuentra dentro de este rango.

No Información Rate (Tasa de No Información): El valor de 0.806 indica la proporción de la clase más frecuente en los datos de entrenamiento. Es decir, si el modelo simplemente predijera la clase más frecuente en todos los casos, tendría una exactitud del 80.6%. Este valor se utiliza como punto de referencia para evaluar el rendimiento del modelo.

P-Value [Acc > NIR] (Valor p [Exactitud > Tasa de No Información]): El valor de < 2.2e-16 indica el valor p para evaluar si la exactitud del modelo es significativamente mayor que la tasa de no información. Un valor de p muy pequeño (< 0.05) indica que la exactitud del modelo es significativamente mejor que simplemente predecir la clase más frecuente.

Kappa: El valor de 0.827 indica el coeficiente de Kappa, que es una medida de concordancia entre las predicciones del modelo y las clases reales ajustado por el azar. Un valor de Kappa cercano a 1 indica una fuerte concordancia, mientras que un valor cercano a 0 indica una concordancia al azar. En este caso, el valor de Kappa es 0.827, lo cual indica una concordancia muy alta entre las predicciones del modelo y las clases reales.

- **Métricas De Rendimiento**

Sensibilidad (Sensitivity): El valor de 0.9672 indica la proporción de casos positivos reales (Buenos Clientes) que el modelo clasifica correctamente como positivos. En otras palabras, la sensibilidad del modelo es del 96.72%.

Especificidad (Specificity): El valor de 0.8579 indica la proporción de casos negativos reales (Malos Clientes) que el modelo clasifica correctamente como negativos. En otras palabras, la especificidad del modelo es del 85.79%.

Valor Predictivo Positivo (Pos Pred Value): El valor de 0.9658 indica la proporción de casos clasificados como positivos por el modelo que son realmente positivos. En otras palabras, el valor predictivo positivo del modelo es del 96.58%.

Valor Predictivo Negativo (Neg Pred Value): El valor de 0.8631 indica la proporción de casos clasificados como negativos por el modelo que son realmente negativos. En otras palabras, el valor predictivo negativo del modelo es del 86.31%.

Prevalencia (Prevalence): El valor de 0.8060 indica la proporción de casos positivos en los datos de entrenamiento. Es decir, la prevalencia de la clase positiva (Buenos Clientes) en los datos es del 80.60%.

Tasa de Detección (Detection Rate): El valor de 0.7795 indica la proporción de casos positivos reales (Buenos Clientes) que el modelo clasifica correctamente como positivos. Es similar a la sensibilidad, pero se calcula en función de la prevalencia.

Prevalencia de Detección (Detection Prevalence): El valor de 0.8071 indica la proporción de casos clasificados como positivos por el modelo en relación con la prevalencia total. Es similar al valor predictivo positivo, pero se calcula en función de la prevalencia.

Precisión Balanceada (Balanced Accuracy): El valor de 0.9126 es el promedio de sensibilidad y especificidad. Representa una medida general del rendimiento del modelo que tiene en cuenta tanto la capacidad de detectar los casos positivos como de clasificar correctamente los casos negativos.

4.7 Matriz de segmento de riesgo del cliente

La matriz de segmento de riesgo de cliente es una herramienta que permite segmentar a los clientes en diferentes categorías de riesgo en función de su puntuación en el modelo. Cada categoría representa un grupo de clientes con características similares en términos de su perfil de riesgo.

Permite identificar y clasificar a los clientes en diferentes niveles de riesgo. Esto facilita la toma de decisiones y estrategias para gestionar el riesgo crediticio, como asignar límites de crédito o implementar medidas de control más estrictas para los clientes de mayor riesgo.

Cada segmento de riesgo puede requerir diferentes enfoques de servicio. Por ejemplo, los clientes de bajo riesgo pueden recibir ofertas de productos más atractivas o beneficios adicionales, mientras que los clientes de alto riesgo pueden requerir una mayor supervisión y seguimiento en términos de sus pagos y comportamiento crediticio.

La matriz de segmento de riesgo de cliente también puede ser útil en el diseño de estrategias de cobranza. Los clientes en diferentes segmentos de riesgo pueden requerir diferentes enfoques y acciones para garantizar el cumplimiento de los pagos. Por ejemplo, los clientes de mayor riesgo pueden requerir un seguimiento más cercano y acciones de cobranza más proactivas.

Para obtener la matriz de segmento del cliente se crea dos nuevas columnas, PROB y SCORE, que contienen las probabilidades normalizadas y los scores discretizados, respectivamente, a partir de las probabilidades predichas por el modelo.

```
PUNTAJE ← MODELO_REGRE$`DATA BASE` %>%
  mutate(PROB = scales::rescale(probabilidades, to = c(0.0001, 0.9999)),
         SCORE = round(scales::rescale(PROB, to = c(999, 1)), 0))
```

(30)

Además, realizamos los scores discretizados basados en las probabilidades normalizadas, agrupa los scores en diferentes categorías definidas por los puntos de corte, cuenta la cantidad de ocurrencias para cada combinación de categoría de score y default (Buenos clientes o Malos Clientes). Luego, presenta el resultado final en un dataframe llamado GRUPO_PUNTAJE.

```
GRUPO_PUNTAJE ← MODELO_REGRE$`DATA BASE` %>%
  mutate(score = round(rescale(probabilidades, to = c(999, 1)), 0),
         grupo_score = cut(score, breaks = c(1,306,514,671,688,721,753,791,850,870,999 )
                           , include.lowest = TRUE)) %>%
  count(grupo_score, default) %>%
  spread(default, n) %>%
  rename(Buenos = `0`, Malos = `1`) %>%
  mutate(Total = Buenos + Malos,
         Porc_Buenos = `Buenos` / Total* 100,
         Porc_Malos = `Malos` / Total * 100,
         Porc_Total = Total / sum(Total) * 100,
         Porc_Acumulado = cumsum(`Buenos`) / sum(`Buenos`) * 100,
         rank = c(1:length(Total))) %>%
  arrange(desc(rank)) %>%
  select(-rank)
```

(31)

Tabla 9 Matriz De Segmento Del Cliente

	grupo_score	Buenos	Malos	Total	Porc_Buenos	Porc_Malos	Porc_Total
	<fct>	<int>	<int>	<int>	<dbl>	<dbl>	<dbl>
1	(870,999]	37360	246	37606	99.3	0.654	77.4
2	(850,870]	181	14	195	92.8	7.18	0.401
3	(791,850]	443	36	479	92.5	7.52	0.986
4	(753,791]	175	11	186	94.1	5.91	0.383
5	(721,753]	129	13	142	90.8	9.15	0.292
6	(688,721]	107	9	116	92.2	7.76	0.239
7	(671,688]	42	3	45	93.3	6.67	0.0926
8	(514,671]	281	194	475	59.2	40.8	0.978
9	(306,514]	393	907	1300	30.2	69.8	2.68
10	[1,306]	51	7996	8047	0.634	99.4	16.6

Fuente: Elaboración Propia.

Podemos observar la tabla que los clientes buenos representan el 80,6% del total de los clientes y se encuentra en un rango score de 870 a 999 el 77.4% podemos decir que estos clientes no, representan perdida para la empresa,

Pero si observamos el 19,4% de los clientes son malos donde el 16,6% se encuentra en un en un puntaje entre 1 a 306 puntos son clientes que representan una perdida para la empresa.

Punto de Corte Matriz de Riesgo

La determinación del punto de corte en una matriz de riego es un paso importante para establecer un umbral que permita clasificar a los solicitantes en diferentes categorías de riesgo.

El punto de corte se define como el valor del puntaje a partir del cual se considera que un solicitante representa un riesgo crediticio aceptable o inaceptable.

Para obtener la matriz de punto de corte se filtra los registros donde MORA120 es igual a 1, calcula los scores discretizados basados en las probabilidades normalizadas, agrupa los scores en diferentes categorías definidas por los puntos de corte, y cuenta la cantidad de ocurrencias para cada combinación de categoría de score y default (Buenos clientes o Malos Clientes). Luego, presenta el resultado final en un dataframe llamado MORA_120_DIAS.

```
MORA_120_DIAS ← MODELO_REGRE$`DATA BASE` %>%
  filter(MORA120 == 1) %>%
  mutate(score = round(rescale(probabilidades,to = c(999, 1)), 0),
         grupo_score = cut(score, breaks =c(1,306,514,671,688,721,753,791,850,870,999 ),
                           include.lowest = TRUE)) %>%
  count(grupo_score,default) %>%
  spread(default,n) %>%
  rename(mora120 = `1`) %>%
  mutate(rank = c(1:length(grupo_score)))  %>%
  arrange(desc(rank)) %>%
  select(-rank)
```

(32)

Luego se une las dos tablas GRUPO_PUNTAJE y MORA_120_DIAS para obtener el porcentaje de perdida

```
PUNTAJE_TOTAL ← GRUPO_PUNTAJE %>%
  left_join(MORA_120_DIAS, by = "grupo_score") %>%
  mutate(Porc_Perdida = coalesce(mora120, 0) / Total * 100,
         Clientes120Dias = coalesce(mora120, 0)) %>%
```

(33)

Tabla 10 Matriz De Punto De Corte

	grupo_score	Buenos	Malos	Total	Porc_Buenos	Porc_Malos	Porc_Total	Porc_Acumulado	Porc_Perdida
	<fct>	<int>	<int>	<int>	<dbl>	<dbl>	<dbl>	<dbl>	<dbl>
1	(870,999]	37360	246	37606	99.3	0.654	77.4	100	0.0106
2	(850,870]	181	14	195	92.8	7.18	0.401	4.60	0.513
3	(791,850]	443	36	479	92.5	7.52	0.986	4.14	0
4	(753,791]	175	11	186	94.1	5.91	0.383	3.01	0
5	(721,753]	129	13	142	90.8	9.15	0.292	2.56	0
6	(688,721]	107	9	116	92.2	7.76	0.239	2.23	0
7	(671,688]	42	3	45	93.3	6.67	0.0926	1.96	0
8	(514,671]	281	194	475	59.2	40.8	0.978	1.85	10.9
9	(306,514]	393	907	1300	30.2	69.8	2.68	1.13	25.9
10	[1,306]	51	7996	8047	0.634	99.4	16.6	0.130	49.0

Fuente: Elaboración Propia.

Existen diversos métodos y enfoques para obtener el punto de corte en la matriz de riesgo, pero el que utilizaremos es el basado en la política y estrategia de la empresa que no pase un 10% de perdida.

Mediante la matriz de punto de corte podemos crear los segmentos de los clientes

Tabla 11 Segmento Del Cliente

Puntaje	Segmento
[870 - 999] [850 - 870]	A
[791 - 850] [753 - 791]	B
[721 - 753] [688 - 721]	C
[671 - 688] [514 - 671]	D
[306 - 514] [1 - 306]	E

Fuente: Elaboración Propia.

Para obtener el punto de corte aquí otros ejemplos de cómo obtener el punto de corte en una matriz de riesgo.

La institución puede establecer sus propias políticas y estrategias de riesgo para determinar el punto de corte. Por ejemplo, pueden establecer un umbral basado en el nivel de riesgo que están dispuestos a aceptar en función de su apetito de riesgo y los objetivos comerciales.

Se puede realizar un análisis de pérdidas y ganancias para evaluar el impacto financiero de diferentes puntos de corte. Esto implica estimar las pérdidas asociadas con los clientes clasificados en diferentes categorías de riesgo y encontrar un equilibrio entre minimizar las pérdidas y maximizar las ganancias.

Se pueden utilizar métricas de evaluación del modelo, como la precisión, Existen diversos métodos y enfoques para obtener el punto de corte en la matriz de riesgo, pero el que utilizaremos es el basado en la política y estrategia de la empresa que no pase un 10% de perdida.

(ROC) el coeficiente de Gini, el valor esperado de pérdida, entre otras, para fijar el punto de corte óptimo. El objetivo es encontrar el punto de corte que maximice la capacidad predictiva del modelo y minimice el error de clasificación.

Se puede realizar un análisis de sensibilidad para evaluar el impacto de diferentes puntos de corte en los resultados de negocio. Esto implica evaluar cómo cambian las tasas de aprobación, las tasas de rechazo, las tasas de incumplimiento y otras métricas clave a medida que se ajusta el punto de corte.

CAPITULO 5

5. Conclusiones y Recomendaciones

El uso de una regresión logística para obtener grupos clientes en la gestión de la cartera de clientes tiene múltiples beneficios en términos de control del riesgo de crédito y rentabilidad. Al implementar un modelo de puntaje, respaldado por análisis estadísticos, se puede tomar decisiones más objetivas y basadas en datos con respecto a la aprobación de créditos.

Acompañado de una estrategia comercial efectiva, el uso de un modelo de puntaje estadístico permite aumentar los rendimientos de los créditos colocados. Esto se debe a que el modelo ayuda a identificar de manera más precisa y consistente a los clientes con menor riesgo crediticio y mayor capacidad de pago. Al dirigir los recursos y esfuerzos de la institución hacia estos segmentos de clientes de bajo riesgo, se pueden obtener retornos más favorables en las colocaciones de crédito.

Además, al contar con un modelo estadístico confiable, la institución tiene la posibilidad de expandir su mercado y prestar a un mayor número de clientes, manteniendo un control riguroso del riesgo. Esto se logra al tener una metodología objetiva y consistente para evaluar a los solicitantes de crédito, lo cual brinda mayor confianza a la institución para ampliar su cartera.

En contraste, basarse en un puntaje implícito o subjetivo puede llevar a decisiones menos precisas y más propensas a errores. Esto puede resultar en tasas de rentabilidad más bajas y mayores riesgos crediticios para la institución.

El uso de metodologías estadísticas, como las scorecards, permite una evaluación ágil y efectiva de los créditos. Que generan un puntaje que refleja el nivel de riesgo del cliente. Esto facilita las mejores decisiones y que garantiza la rapidez en las distintas fases de la aprobación de créditos, permitiendo colocar más dinero a un riesgo controlado.

Además, el análisis estadístico de variables y la calificación objetiva del cliente mediante un puntaje brindan la oportunidad de establecer principales actividades enfocada a la recuperación de la gestión del cobro. De acuerdo a su comportamiento del cliente y sus niveles de riesgo, se pueden diseñar ofertas y condiciones de crédito adaptadas a sus características particulares. Esto ayuda a optimizar la rentabilidad de la cartera de créditos y a ajustar las estrategias de gestión de riesgos de manera más precisa.

Es cierto que la parte principal es la información de calidad que se encuentra en los sistemas de gestión de datos es crucial para la implementación de un modelo logístico efectivo. En el

desarrollo del análisis, contar un sistema de gestión de información consistente fue un factor clave que permitió trabajar con confianza y utilizar una amplia variedad de variables individuales y mixta sin perder información significativa.

La calidad de los datos se refiere a la precisión, integridad, consistencia y actualidad de los datos. Una base de datos confiable y de calidad proporciona una representación precisa de los clientes y su historial crediticio, lo que es fundamental para construir un modelo de scoring preciso y confiable.

La disponibilidad de una gran cantidad de variables individuales y mixtas también es una ventaja significativa. Al contar con más variables relevantes, se pueden capturar mejor las diferentes dimensiones y factores que influyen en el riesgo crediticio. Esto permite un análisis más completo y detallado de los clientes, lo que a su vez puede mejorar la precisión y la capacidad predictiva del modelo de scoring.

Las variables que mayor aportan al desarrollo del modelo estadístico de regresión logística son las variables como Atraso Máximo, Saldo Actual de la Deuda, Monto Crédito, Atraso 0 días, atraso de 1 a 30 días, atraso de 30 a 60 días, y de acuerdo a cada producto que ofrece la empresa desde el producto1 hasta el producto 6.

Los indicadores de Perdida por percentil score permiten evaluar el desempeño del modelo en diferentes segmentos de clientes, clasificados en función de su puntaje de crédito. Al establecer un punto de corte adecuado, es posible crear segmentos de clientes con diferentes niveles de riesgo y aplicar estrategias de colocación diferenciadas en cada segmento.

Se observa que los clientes con altos puntajes en el modelo de scoring tienden a presentar menores niveles de perdida en comparación con aquellos con puntajes bajos. Esto indica que el modelo es efectivo en identificar a los clientes con menor riesgo crediticio y mayor capacidad de pago, lo que se traduce en una menor perdida para la empresa.

En función de la perdida y los puntajes del modelo, se formaron diferentes segmentos de clientes. Perfiles de clientes establecidos o segmento como A, B, C, D, E, F que permitirán medir el nivel de riesgo del cliente.

6. Referencias

Lara Rubio, J. (2010). La Gestión de Riesgos en las Instituciones de Microfinanzas. Editorial de la Universidad de Granada.

David W. Hosmer, Stanley Lemeshow. (2000). Applied Logistic Regression. Editorial Wiley-Interscience.

David G. Kleinbaum & Mitchel Klein. (2010). Introduction to Logistic Regression. Editorial Springer New York, 2010.

Philippe Jorion. (2006). Value at Risk, 3rd Ed.: The New Benchmark for Managing Financial Risk. Editorial McGraw Hill Professional.

Frank J. Fabozzi, Steven V. Mann, Moorad Choudhry. (2003). Measuring and Controlling Interest Rate and Credit Risk 2nd Edición. Editorial Wiley.

Ecuador, J. B. (01 de 22 de 2004). Resolución No JB-2004-631. Obtenido de SBS: http://www.sbs.gob.ec/medios/PORTALDOCS/downloads/normativa/nueva_codificacion/todos/L1_X_cap_I.pdf

Vasicek, O. A. (2002). The Distribution of Loan Portfolio Value. Risk, 15, No. 12.

William F Sharpe (1985). Investments 3rd Edición. Editorial Prentice-Hall.

John Barron, (1989). Credit Scoring and Its Applications. Editorial Society for Industrial and Applied Mathematics.

(Jorge Blanco, 2006) Blanco, Jorge. Introducción al Análisis Multivariado. Instituto de Estadística, Montevideo, Uruguay, 2006.

(Krzanowski y Hand, 2009.) Wojtek J. Krzanowski y David J. Hand, ROC Curves for Continuous Data. Chapman & Hall/CRC by Taylor and Francis Group, LLC, 2009.

Bolivia, Superintendencia de Bancos y Entidades Financieras, Bolivia, e Intendencia de Estudios y Normas. 2008. Guías para la gestión de riesgos. La Paz, Bolivia: Superintendencia de Bancos y Entidades Financieras de Bolivia.

Bambino Contreras, Carlos. 2005. "Prestar como locos y obtener beneficios: ¿es realmente posible? (Un análisis logit multinomial para los determinantes del comportamiento de pago de una cartera de consumo)". Quito: FLACSO. http://repositorio.flacsoandes.edu.ec/handle/10469/61.

O'Brien, Robert. 2007. "A Caution Regarding Rules of Thumb for Variance Inflation Factors". Quality and Quality, 673–90.

Siddiqi, Naeem. 2006. Credit risk scorecards: developing and implementing intelligent credit scoring. Hoboken, N.J: Wiley.

Moral, Irene. 2006. "Modelos de regresión: lineal simple y regresión logística". Revista Seden, el 3 de diciembre de 2006.

Junta Bancaria del Ecuador. (15 de 03 de 2011). Resolución No. JB-2011-1897. Recuperado el 31 de 01 de 2015, de Normas generales para la aplicación del sistema financiero

Flórez, Orlando Moscote, y William Arley Rincón. 2002. "Modelo Logit y Probit: un caso de aplicación", Comunicaciones en Estadística, 5 (diciembre): 123–133.

Roberto González Suárez y Emma Domínguez Alonso. 2010. "Análisis de las curvas receiver-operating characteristic: un método útil para evaluar procederes diagnósticos".

7. Anexo

Código R

Proceso del modelo de regresión logística

```
## Variables Dependiente e Independiente

Base_Modelo <- Base_Modelo %>% select(Variable Dependiente, Variables Independientes )

## Modelo de Regresión Logística
set.seed(10)
Regresion_Logistica <- step(glm(OBJETIVO ~ ., data = Base_Modelo,
                                family = "binomial"))

### Cálculo del VIF
vif_modelo <- vif(Regresion_Logistica)
print(vif_modelo)

### Test Modelo regresión Logística
Prueba_Modelo <- function(logit_regre,data,prob_RL){
  default_mod <- data %>% select(default) %>% pull()
  ### Coeficiente De Gini
  gini_coefficient <- gini(logit_regre$fitted.values, default_mod)
  ### Matriz
  data <- data %>%
    mutate(probabilidades = predict(logit_regre, type = "response"),
           prediccion = ifelse(probabilidades > prob_RL, 1, 0))
  ### Media Prediccion
  media_prediccion <- mean(data %>% select(prediccion) %>% pull() == default_mod)
  ### Matriz De Confusión
  matriz_confusion <- confusionMatrix(table(data %>% select(prediccion) %>% pull(), default_mod))
  ### Data Roc
  data.roc <- prediction(data %>% select(probabilidades) %>% pull(), data$default)
```

```
 ### Roc
 area_roc <- slot(performance(data.roc,"auc"), "y.values")[]1]]
 #### Kolmogorov Smirnov
 ks.data <- performance(data.roc, "tpr", "fpr")
 data.ks <- max(attr(ks.data, "y.values")[[1]] - (attr(ks.data, "x.values")[[1]]))
  ### Hosmer Lemeshow
 hl.data <- hoslem.test(default_mod, fitted(logit_regre), g = 10)
 list("MEDIA DE PREDICCION" = media_prediccion,
     "GINI" = gini_coefficient,
     "ROC" = area_roc,
     "KOLMOGOROV SMIRNOV" = data.ks,
     "HOSMER LEMESHOW" = hl.data,
     "MATRIZ DE CONFUSION" = matriz_confusion,
     "DATA BASE" = data,
     "DATA ROC" = data.roc) }

### Test del modelo
Modelo_Regre <- Prueba_Modelo(Regresion_Logistica,Base_Modelo %>%
                        rename(default = OBJETIVO),0.3)

### Probabilidades del Modelo
Puntaje <- Modelo_Regre$`DATA BASE` %>%
  mutate(PROB = scales::rescale(probabilidades, to = c(0.0001, 0.9999)),
        SCORE = round(scales::rescale(PROB, to = c(999, 1)), 0))

### Segmento de clientes por probabilidades
Puntaje_Total <- Grupo_Puntaje %>%
  left_join(Mora_120_Dias, by = "grupo_score") %>%
  mutate(Porc_Perdida = coalesce(mora120, 0) / Total * 100,
        Clientes120Dias = coalesce(mora120, 0)) %>%
  select(-mora120)
```