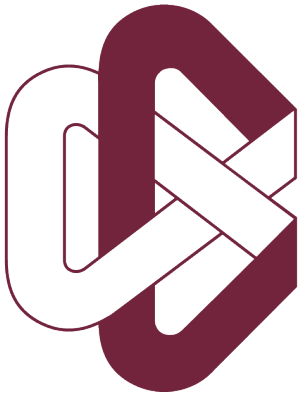

CIMAT

# Centro de Investigación en Matemáticas, A.C.

**Estadística Multivariada**

"Modelando la Probabilidad de Cumplimiento".


Alumnas:

**Alondra Elizabeth Matos Mendoza**
**María Guadalupe López Salomón**


08 de octubre de 2023

# Índice

# 1. Resumen

Con el propósito de distinguir a los clientes confiables, se implementó un modelo de regresión logística que estima la probabilidad de que un solicitante de crédito se clasifique como "bueno" o "malo", dependiendo si representa un riesgo crediticio alto o bajo, en función de su historial y las características más relevantes del prestatario. A partir de estas probabilidades, se calculó un puntaje (score) vinculado a las variables predictoras para denotar el nivel de riesgo. Este método mantiene una nivelación igual para clientes con características (variables) semejantes, de tal forma que se optimiza la evaluación en la determinación de la aprobación de un préstamo y faculta a una empresa para decidir de manera más rápida la aprobación o denegación de solicitudes de crédito. Esta agilidad contribuye a una gestión efectiva del riesgo crediticio.

Es importante destacar que, después de la comparación de varios modelos ajustados utilizando el análisis de devianza, la prueba de la razón de la verosimilitud y el Criterio de Información de Akaike (AIC), el modelo definitivo se evaluó a través de un proceso de validación. En el cual, se realizó un análisis de residuales para verficar la adecuación del modelo; y para medir su desempeño, se calcularon el Perfil de Precisión Acumulada (curva CAP) y la curva Característica Operativa del Receptor (ROC). Además, se obtuvo el índice de Gini y se aplicó la prueba estadística de Kolmogorov-Sminorv para determinar si existen diferencias significativas entre las poblaciones de clientes considerados como buenos y malos.

# 2. Introducción

Dentro de las necesidades más importantes que presentan las instituciones crediticias, en medio de la creciente inestabilidad económica, se encuentran controlar y gestionar el *riesgo de crédito*. El crédito, que consiste en la entrega de fondos por parte de un acreedor a un prestatario, conlleva un riesgo inherente relacionado con el incumplimiento de los términos de pago acordados. Este riesgo, denominado *riesgo de crédito*, tiene un impacto directo en el precio de mercado de las transacciones financieras, ya que se incorpora en los costos de financiamiento [3].

Para abordar las vulnerabilidades que pueden surgir en diversos entornos económicos, es esencial llevar a cabo evaluaciones continuas del riesgo. Por esta razón, las entidades financieras han desarrollado modelos de originación y seguimiento que, además de ser herramientas para medir el riesgo de crédito, se han convertido en elementos clave para definir su estrategia empresarial. Uno de los principales objetivos al crear estos modelos, radica en la necesidad de calcular el capital económico necesario para respaldar las actividades de toma de riesgos en una entidad financiera [2].

Dentro del marco de la gestión de riesgo crediticio, el pronóstico del incumplimiento de los clientes y los cambios en su calificación son de suma importancia. En este contexto, la "probabilidad de incumplimiento" o "default" se convierte en un componente crítico en la evaluación del riesgo de crédito. Ésta última, se refiere a la probabilidad de que un prestatario deje de cumplir con sus obligaciones contractuales, en particular, el pago de una deuda vencida.

De acuerdo con cifras recientes, se ha observado un crecimiento significativo del número de carteras vencidas, lo que resalta la enorme problemática que supone para las instituciones de crédito aprobar préstamos a personas que no van a pagarles. En este contexto, es de sumo interés obtener herramientas y modelos que nos permitan cuantificar el riesgo que se corre al otorgar créditos, una forma de hacerlo, es a través de la medición de la *probabilidad de cumplimiento*.

El presente trabajo cuantitativo, se centra en la implementación de un modelo de regresión logística o *logit* para estimar la probabilidad de cumplimiento de la cartera de clientes perteneciente a la entidad financiera LendingClub de San Francisco, California. A través de este modelo, se categorizará a los clientes como *buenos* o *malos*.

# 3. Antecedentes

Dentro de la teoría de la probabilidad, el término *esperado*, siempre se refiere a un *valor esperado* o *valor medio*, y esto también es aplicable en la gestión de riesgos. Si una entidad financiera asigna a cada cliente:

- una *probabilidad de incumplimiento* ($PD$)
- una fracción de pérdida denominada *pérdida en caso de incumplimiento* ($LGD$), que describe la parte de la exposición del préstamo que se anticipa perder en caso de incumplimiento
- y una *exposición en caso de incumplimiento* (EAD) que está en riesgo de perderse en el periodo de tiempo considerado.

La *pérdida esperada*, asociada a cualquier deudor se define en términos de una variable de pérdida [1]:

$$PE = PD * LGD * EAD \tag{1}$$

La probabilidad de incumplimiento o default $PD$, representa la probabilidad prevista para que un cliente deudor se declare insolvente y deje de pagar sus cuotas de amortización. Se calcula a través de la información recibida por las agencias especializadas de rating y scoring.

El ratio de pérdida en caso de incumplimiento $LGD$, es el porcentaje de un préstamo que, una vez impagado y efectuadas las habituales gestiones para su recobro, resulta finalmente incobrable.

La exposición en caso de incumplimiento $EAD$, es otro de los *inputs* o entradas necesarias en el cálculo de la pérdida esperada y el capital, definida como el importe de deuda pendiente de pago en el momento de incumplimiento del cliente.

El *scoring experto* o *estadístico*, es una herramienta que sirve para discriminar los buenos prospectos de los malos prospectos. Consiste de una metodología que es capaz de pronosticar el riesgo futuro por el incumplimiento de pagos en una ventana de tiempo determinada. Está basado en el análisis de dos tipos de datos referentes a los clientes, que pueden ser datos demográficos como: *edad, sexo, ingresos, situación laboral* y datos de buró de crédito como el *número de tarjetas de crédito en mora, historial crediticio* y *comportamieno en cuanto a la morosidad de pagos* [4].

El modelo para discriminar a los buenos clientes, generalmente consiste en una fórmula con parámetro desconocido que se puede estimar con los datos de la institución objetivo, o bien, con información de instituciones externas. Al estimar la probabilidad de cumplimiento de pago, es posible generar un puntaje (*score*) que se le asocia a las variables predictivas para indicar un nivel de riesgo.

El modelo descrito anteriormente, se denomina modelo *scoring* y brinda una estimación del comportamiento promedio de individuos que cumplen con características particulares, proporcionando un margen de decisión crucial en el proceso de otorgamiento de créditos financieros [4].

La eficiencia del modelo *scoring* depende de la representación significativa de clientes "buenos" y "malos", actualización de datos, actualización y ajuste periódico de datos económicos en el país donde se esté aplicando, comportamiento de pago, entre otros, [4].

Para construir un modelo *scoring*, se pueden emplear diferentes técnicas estadísticas. En particular, dentro de este proyecto, se utilizará una regresión logística, a través de la cual, se estimarán los coeficientes de las variables predictoras. Posteriormente, se construirá una *scorecard*, donde se asignarán puntajes en diferentes rangos de las variables predictivas, es decir, de las características con las que cumpla el cliente. De manera que, la tabla de puntajes o *scorecard*, estimará la probabilidad de rechazo de solicitud, a través de los scores obtenidos por cada categoría.

# 4. Planteamiento del problema

Dentro de los constantes desafíos a los que se enfrentan frecuentemente las entidades financieras, se encuentra la necesidad de tener criterios confiables y fidedignos para determinar a quienes pueden otorgar créditos financieros y en qué medida pueden hacerlo.

En este sentido, medir la probabilidad de cumplimiento, es fundamental y de sumo interés para las entidades financieras, pues les permite implementar modelos y estrategias financieras que minimicen el riesgo, al adquirir nuevos clientes, lo que reducirá la pérdida ecónomica debido a una mala decisión.

# 5. Pregunta de investigación

- ¿Qué variables tienen un impacto en la probabilidad de que un solicitante de crédito sea aprobado?

# 6. Objetivos

- Identificar el modelo de regresión logística que mejor describa la relación entre las variables en cuestión, con el fin de modelar la probabilidad de cumplimiento $1 - PD$, para determinar si un cliente de LendingClub es candidato a obtener un crédito.
- Desarrollar un modelo de scoring a partir de las estimaciones de los coeficientes del modelo ajustado.

# 7. Marco Teórico

## 7.1. Regresión Logística

Dado que nuestro interés es discriminar a los solicitantes de crédito, como "buenos" y "malos", la regresión logística es una herramienta esencial en la construcción de modelos de clasificación binaria.

Un ámbito crucial donde la regresión logística brilla es en la creación de modelos de scoring crediticio, los cuales, buscan evaluar y mejorar la capacidad predictiva al clasificar a los individuos en dos grupos: "buenos" y "malos", basándose en las características mencionadas. La clave de esta clasificación está en una distribución de probabilidad que separa a la población en estos dos grupos, utilizando un umbral ajustable entre 0 y 1. La probabilidad calculada estima el valor de "y", asignando al individuo a uno de los dos grupos.

En este proceso, no se establecen restricciones rígidas en las variables explicativas. Pueden ser cualitativas o cuantitativas, y variar en su naturaleza. La variable de respuesta, $y$, toma valores de 1 cuando el cliente cumple con la categoría y 0 en caso contrario.

Con la regresión logística se modela la probabilidad de que $y$ sea igual a 1, dados los valores observados de las variables predictoras contenidas en el vector $\mathbf{x_i^T} = [1, x_{i1}, \cdots, x_{in}]$, esto es, $P(y = 1|x)$.

Como estamos interesados en estimar la probabilidad de cumplimiento $1 - PD = PC$, entonces nuestro modelo de regresión logística es de la forma:

$$PC = \frac{e^{\beta_0+\beta_1 x_i+\cdots+\beta_n x_n}}{1 + e^{\beta_0+\beta_1 x_i+\cdots+\beta_n x_n}} \tag{2}$$

donde

$$\beta_i^T = [\beta_1, \beta_2, ..., \beta_n] \tag{3}$$

es el vector de coeficientes asociados a las variables explicativas del modelo y $\beta_0$ es el intercepto del modelo.

Aplicando una transformación de logaritmo del cociente de probabilidades de $(1 - PC)$ y $PC$, podemos realizar la estimación de los coeficientes:

$$ln\left(\frac{PC}{PD}\right) = \beta_0 + \beta_i^T x_i \tag{4}$$

## 7.2. Pruebas estadísticas al modelo logit

En la regresión logística, al igual que en otros modelos estadísticos, es importante realizar pruebas estadísticas para evaluar la calidad del modelo y determinar si las variables predictoras son significativas en la estimación de la variable respuesta. Dentro del conjunto de técnicas para medir la capacidad discriminatoria, tenemos:

### 7.2.1. Devianza

Para evaluar la bondad de ajuste de un modelo en comparación con un modelo de referencia o nulo, podemos emplear el concepto de devianza. La cual es una medida de cuán bien el modelo se ajusta a los datos observados en comparación con un modelo nulo que no tiene variables predictoras.

$$D(\beta) = 2\sum_{i=1}^{n}\left[log\left(1 + e^{x_i^T\beta}\right) - y_i x_i^T\beta\right] \tag{5}$$

que en términos de la logverosimilitud, la devianza de un modelo ajustado se define como:

$$D = 2\left\{l(\hat{\beta}_{max}) - l(\hat{\beta})\right\}\phi \tag{6}$$

donde $\phi$ es el parámetro de escala, $l(\hat{\beta}_{max})$ es la logverosimilitud maximizada del modelo saturado y $l(\hat{\beta})$ es la logverosimilitud maximizada del modelo de interés.

En un modelo saturado, se considera que se tiene un parámetro por cada observación, en este caso $n$ parámetros y se evalúa con $\hat{\mu} = y$. En este modelo, se tiene un número maximal de parámetros $(n)$ que, por supuesto, será el modelo que mejor ajuste a los datos.

En contraste, en un modelo nulo, el predictor lineal $\eta_i = \beta_0 + \beta_1 x_{i1} + \cdots + \beta_k x_{ik}$, es de la forma $\eta_i = \beta_0$ y, básicamente, lo que suponemos es que las variables de respuestas $y_1, \cdots, y_n$ son independientes e idénticamente distribuidas $f(y;\theta)$, donde $\theta$ es un sólo parámetro.

Sin embargo, el modelo de interés, es un modelo intermedio entre el modelo saturado y el modelo nulo.

Para comparar la devianza entre el modelo completo y el modelo nulo podemos aplicar una prueba de razón de verosimilitud, la cual se puede formular de manera que siga una distribución chi-cuadrado:

$$D_0^* - D_1^* \sim \chi^2_{p_1 - p_0} \tag{7}$$

donde $D_0^*$ representa la devianza del modelo nulo, es decir, aquel que no contiene las variables predictoras en cuestión, $D_1^*$ representa la devianza del modelo completo, que incluye las variables predictoras, $p_1$ es

el número de parámetros estimados en el modelo completo que incluye las variables predictoras y $p_0$ es el número de parámetros estimados en el modelo nulo, que generalmente es igual al intercepto.

De modo que, si el modelo completo es significativamente mejor que el modelo nulo para explicar los datos, entonces la diferencia $D_0^* - D_1^*$ sigue una distribución chi-cuadrado con $p_1 - p_0$ grados de libertad.

Bajo $H_0$, la cual implica que ciertas variables predictoras o términos en el modelo no tienen un efecto significativo en la variable de respuesta, se tiene el resultado aproximado:

$$D_0^* - D_1^* \sim \chi_{p_1-p_0}, \quad D_1^* \sim \chi^2_{n-p} \tag{8}$$

además, si $D_0^* - D_1^*$ y $D_1^*$, se consideran asintóticamente independientes, esto implica que:

$$F = \frac{(D_0^* - D_1^*)/(p_1 - p_0)}{D_1^*/(n - p_1)} \sim F_{p_1-p_0, n-p_1} \tag{9}$$

en el límite de muestras grandes, lo cual coincide exactamente en el modelo lineal ordinario.

### 7.2.2. Estadístico de Wald

Para evaluar si los coeficientes de las variables predictoras son significativamente diferentes de cero, podemos utilizar el estadístico de Wald. De manera que, el estadístico de Wald es una prueba que nos indica si las variables predictoras tienen un impacto estadísticamente significativo en la variable de respuesta. La prueba resulta de contrastar la hipótesis nula:

$$H_0 : \beta_i = 0 \tag{10}$$

contra la alternativa

$$H_1 : \beta_i \neq 0 \tag{11}$$

con un estadístico de prueba definido como:

$$W_0 = \frac{\hat{\beta}_i}{s(\hat{\beta}_i)} \tag{12}$$

la expresión (12) corresponde al estadístico de Wald, donde $\hat{\beta}$ es el coeficiente estimado de la variable predictora en cuestión y $s(\hat{\beta}_i)$ es el error estándar del coeficiente estimado.

Note que, bajo el supuesto de que $H_0$ es cierta, sigue una distribución $t$ con $n - p - 1$ grados de libertad y para muestras grandes, se distribuye como una normal estándar.

Si $W_0$ es un valor alejado de cero, se tendrá evidencia de que $H_0$ es falsa, por lo tanto, la región crítica de la prueba es de la forma $|W_0| > t_{\alpha/2}$ para un nivel de significancia adecuado.

Por otro lado, si el verdadero valor del parámetro $\beta_i$ es cero, la variable $x_i$ debe excluirse. Una manera alternativa de escribir la región crítica es usando el $p - value$ donde $p = P(t > |W_0|)$. La región crítica para un nivel de significancia $\alpha$, es de la forma $p < \alpha$.

# 8. Metodología

## 8.1. Descripción de la base de datos

La base de datos consta de $466,284$ créditos otorgados por Lending Club (LC), el cual es un mercado de préstamos en línea que facilita préstamos personales, comerciales y financiamiento de procedimientos médicos.

Para ajustar el modelo que discrimine a los prestatarios confiables, se etiquetó a los clientes como "buenos" o "malos" en función del estatus de su préstamo. De este modo, la variable de respuesta se codificó en valores de 0 y 1, siendo 1 si se trata de un buen cliente y 0 para indicar que es un cliente malo.

Se consideró como un buen cliente a aquellos individuos cuyos prestámos cumplían con alguna de las siguientes condiciones: Totalmente pagado, al corriente, en periodo de gracia, con retraso de 16 a 30 días y aquellos pagados que no cumplían con la política de crédito.

Por otro lado, siendo un mal cliente aquél deudor que causa pérdidas económicas a la compañía, se definió este tipo de clientes como aquellos con préstamos que se caracterizon por: impago (deuda no recuperable), inclumplimiento, retraso de 31 a 120 días y no cumplir con la política de crédito ni de pago.

Excluyendo las variables con exceso de campos sin respuesta, las variables a considerar para desarrollar un modelo de regresión logística que estime la probabilidad de cumplimiento (es decir, la probabilidad de que el cliente sea bueno) son:

- Variables discretas:

1. Grado (*Grade*): Grado de préstamo asignado $\{A, B, C, D, E, F, G\}$.
2. Propiedad de vivienda ($Home\_ownership$): El estado de propiedad de la vivienda proporcionado por el prestatario durante el registro *{Own (Propio). Rent (Renta), Mortgage (Hipoteca), None (Ninguno), Any(Cualquiera), Other (Otro)}*.
3. Estado ($Addr\_state$): El estado proporcionado por el prestatario estadounidense en la solicitud de préstamo.
4. Verificación de estatus ($Verification\_status$): Indica si el ingreso del prestatario fue o no verificado por LC, o si la fuente de ingresos fue verificada *{Verified (Verificado por LC), Not Verified (No verificado por LC), Source Verified (Fuente verificada)}*.
5. Propósito ($Purpose$): Una categoría proporcionada por el prestatario para la solicitud de préstamo. *{car (carro), credit_card (tarjeta de crédito), debt_consolidation (consolidación de la deuda), educational (educación), home_improvement (mejoras para el hogar), house (casa), major_purchase (compras mayores), medical (médico), moving (mudanza), renewable_energy (energía renovable), small_business (pequeños negocios), vacation (vacaciones), wedding (boda), other (otro)}*
6. Estado inicial de la lista ($Initial\_list\_status$): El estado de listado inicial del préstamo. Puede ser entero (Whole) o fraccional (Fractional) *{w, f}*.
7. Plazo ($Term$): Cantidad de tiempo que el prestatario tiene para pagar el préstamo. *{36 months (36 meses), 60 months (60 meses)}*.
8. Duración de empleo ($Emp\_length$): Período de trabajo en años. *{<1,1,2,...,9,10, 10+}*
9. Consultas en los últimos 6 meses ($Inq\_last\_6mths$): El número de consultas en los últimos 6 meses (excluyendo consultas de automóviles e hipotecas)
10. Cuentas actualmente en mora ($Acc\_now\_delinq$): El número de cuentas en las que el prestatario ahora está en mora.
11. Meses desde el último incumplimiento de pago ($Mths\_since\_last\_delinq$): El número de meses transcurridos desde la morosidad pasada del prestatario.

12. Meses desde el último registro (*Mths_since_last_record*): El número de meses desde el último registro público.
13. Meses desde la emisión (*E_mnths_since_issd*): Número de meses que ha transcurrido desde que se expidió el crédito hasta la fecha de corte.
14. Meses desde la última actualización de la línea de crédito (*E_mnths_since_ecrline*): Cantidad de meses transcurridos desde la última modificación de la línea de crédito.

Cabe mencionar que las últimas dos variables discretas fueron creadas a partir de la diferencia entre la fecha de corte (diciembre del 2017) y las fechas registradas en las variables *issue_d* (mes en que se financió el préstamo ) y *earliest_cr_line* (mes en que se abrió la primera línea de crédito reportada por el prestatario).

- Variables continuas

1. Tasa de interés (*int_rate*): Tasa de interés del préstamo.
2. Ingreso anual (*annual_inq*): El ingreso anual proporcionado por el prestatario durante el registro.

Se realizó una partición aleatoria de la base de datos en un conjunto de entrenamiento (80 %) y un conjunto de prueba (20 %) para la creación y evaluación del modelo, respectivamente.

## 8.2. Tratamiento de las variables independientes

Para elegir las características que poseen un mayor valor de predicción a nivel global, se utilizó como criterio el Valor de Información (IV), el cual es una función que depende de la proporción de buenos y malos clientes en los atributos de cada característica.

$$IV = \sum_{i=1}^{k} (\,\%buenos_i - \,\%malos_i) WOE_i$$

donde $k$ es el número de bins de la variable en cuestión.

El IV se encuentra entre 0 y 1. Cuanto mayor sea el IV, mayor será la contribución de la variable independiente al modelo.

Para calcular el IV en variables continuas, se realizó una categorización de las mismas agrupando los valores de una variable de tal manera que cada intervalo contenga un porcentaje igual de observaciones.

Después, se aplicó la técnica de agrupación amplia (coarse classing) a cada variable, la cual implica crear nuevas categorías utilizando las categorías iniciales disponibles. Los atributos se formaron considerando la proporción de clientes buenos y malos utilizando una medida llamada WOE (Weight of Evidence).

$$WOE_i = ln \frac{\text{Proporción de buenos que caen en la categoría } i}{\text{Proporción de malos que caen en la categoría } i}$$

Aplicando el refinamiento de clasificación (fine classing) utilizando el WOE, las categorías con valores de WOE similares se agruparon en un mismo atributo.

Para las variables *mths_since_last_delinq* y *mths_since_last_record*, que por naturaleza tenían muchos registros NA, se incluyó una categoría especial para representar estos NA's.

## 8.3. Ajuste del modelo mediante regresión logística

Para representar cada categoría nueva, se utilizaron variables ficticias (dummy) que toman el valor de 1 si el cliente posee esa característica y 0 de lo contrario.

Finalmente, utilizando las variables dummy, se ajustó el modelo de predicción para los nuevos clientes mediante regresión logística, tomando en consideración las variables con suficiente poder de predicción. Es fundamental destacar que, con el fin de evitar la multicolinealidad entre las variables predictoras en el modelo, se aplicaron restricciones de identificabilidad. Esto se logró tomando por defecto la categoría con el menor WOE de cada variable.

### 8.3.1. Comparación de modelos

El modelo saturado ajustado se comparó con otros modelos más sencillos, obtenidos al eliminar las variables que no resultaron significativas (a un nivel de significancia del 5 %), basándose en la proporción de variables dummy que demostraron ser relevantes.

Para realizar la comparación entre dos modelos, se aplicó la prueba de la razón de la verosimilitud y se utilizó el Criterio de Información de Akaike (AIC), el cual es una metodología para elegir modelos minimizando la estimación de la divergencia de Kullback-Leibler entre el modelo ajustado y el modelo verdadero. Se opta por el modelo con el AIC más bajo.

## 8.4. Evaluación del modelo

### 8.4.1. Análisis de residuales

Para realizar el diagnóstico del modelo y verificar la adecuación del mismo, se utilizaron las siguientes gráficas de los residuales de devianza:

- La gráfica de residuales contra predichos, empleada para examinar la presencia de alguna tendencia sobre la media de los residuales, lo cual es señal de dependencia entre las variables predictoras.
- La gráfica de ubicación-escala, la cual sugiere que la varianza de los residuales es constante en caso de no observarse algún patrón.
- La gráfica de probabilidad normal (QQ-Plot), utilizada para observar si los residuales se ajustan aproximadamente a una línea recta, lo cual sugiere que siguen una distribución normal. Si el modelo es correcto, se espera que los errores se comporten como $N(0, 1)$.
- La gráfica de residuales estandarizados contra los puntos de apalancamiento, usada para detectar la presencia de observaciones influyentes mediante la distancia de Cook.

### 8.4.2. Validación del método de clasificación

Las técnicas utilizadas para medir el desempeño del modelo fueron la curva ROC y la curva CAP.

- **Curva CAP (Cumulative Accuracy Profile)**

  La curva CAP es una gráfica que muestra la acumulación de porcentajes de clientes en el eje $x$ y el porcentaje acumulado de clientes en riesgo en el eje $y$. Es relevante para comparar el modelo ajustado con un modelo ideal y uno que clasifica aleatoriamente. Además, permite fácilmente contrastar la proporción de malos rechazados contra la proporción del total de rechazados.

En el caso del modelo logístico que genera probabilidades de cumplimiento, se ordenan los registros de la muestra de prueba en orden ascendente según estas probabilidades. Para calcular la curva CAP, se toma una fracción $x$ de los registros y se calcula el porcentaje de clientes en riesgo que tienen una probabilidad igual o menor a la máxima probabilidad de esa fracción $x$.

En un modelo ideal, la acumulación de frecuencias debería alinearse perfectamente con la frecuencia de clientes en riesgo, lo que daría lugar a una curva CAP lineal que alcanza 1 y se mantiene constante. Esto indica que el modelo detecta de manera correcta a todos los clientes en riesgo. Por otro lado, en un modelo aleatorio sin capacidad de discriminación, la proporción $x$ de registros con baja probabilidad abarcaría cerca del $x$ por ciento del total de registros de prueba.

- **Curva ROC (Receiver operating characteristic)**

  La curva ROC es un gráfico que muestra la tasa de verdaderos positivos en función de la tasa de falsos positivos en varios puntos de discriminación (también conocidos como umbrales o puntos de corte). Cada punto en la curva representa un valor específico obtenido para un determinado punto de corte.

  El área bajo la curva ROC, conocido como AUC (Area Under the Curve), mide el poder discriminatorio del modelo. Si el AUC es 1, el modelo tiene una discriminación perfecta. Si es 0.5, el modelo no tiene capacidad discriminatoria en absoluto y predice al azar. Si es 0, el modelo hace predicciones erróneas al invertir ambas clases.

### 8.4.3. Prueba de diferencias de dos poblaciones

Se calculó el Coeficiente de Gini y se utilizó la prueba K-S para asegurar que la clasificación del modelo ajustado, determinada por un punto de corte predefinido en el rango de cero a uno, esté asociada con una distribución de probabilidad que distingue a la población en dos grupos distintos. Se espera que el modelo de regresión logística asigne a cada individuo a un grupo, donde la probabilidad estimada para los buenos clientes debería ser cercana a uno y para los malos clientes cercana a cero.

- **Coeficiente de Gini con observaciones agrupadas**

  El coeficiente de Gini evalúa la eficiencia de un modelo en comparación con una clasificación aleatoria. Varía entre -1 y 1, donde valores negativos indican clasificaciones invertidas, y valores positivos cercanos a uno indican un modelo altamente efectivo en la distinción entre clases.

  El coeficiente de Gini se calcula como:

$$GINI = \frac{AUC - 0.05}{0.5} = 2 * AUC - 1$$

- **Test de Kolmogorov-Smirnov (Prueba K-S)**

  Se aplicó la prueba de Kolmogorov-Smirnov para examinar la hipótesis nula de que, de acuerdo con la predicción del modelo, la distribución poblacional es igual tanto para la clase positiva (clientes buenos) como para la clase negativa (clientes malos).

  El valor del estadístico de prueba se obtiene como la mayor diferencia absoluta entre las distribuciones empíricas, buscando detectar las discrepancias entre las frecuencias relativas acumuladas de las dos muestras de estudio.

# 9. Resultados

# 10. Conclusión

# Referencias

[1] Bluhm, C., Overbeck, L., & Wagner, C.(2010) *Introduction to Credit Risk Modeling*[Second Edition, CRC Press.]

[2] Chatterjee, S. (2016) *Modelos del riesgo de Crédito*[CEMLA, Handbook, núm 34, del Centre for Central Banking Studies, Banco de Inglaterra]

[3] Lenin, A., & Ayús, T. (2016) *El método popperiano en la estimación de la probabilidad de incumplimiento de un deudor* [Revista de Investigaciones de la Escuela de Administración y Mercadotecnia del Quindío EAM]

[4] Nieto, M., S. (2010) *Crédito al Consumo: La Estadística aplicada a un problema de Riesgo crediticio*[Proyecto de Tesis, Universidad Autónoma Metropolitana]