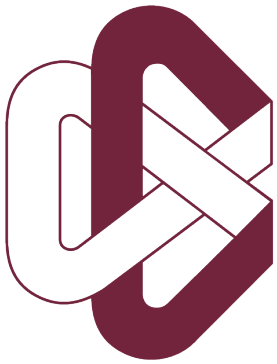

CIMAT

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 β_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 ϕ 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 θ 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 β_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 α, 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. Se puede descargar en: `https://www.kaggle.com/datasets/devanshi23/loan-data-2007-2014/`

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 vivenda ($Home\_ownership$): 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.

- Variables continuas

1. Tasa de interés (int_rate): Tasa de interés del préstamo.
2. Ingreso anual ($annual\_inc$): El ingreso anual proporcionado por el prestatario durante el registro.
3. Meses desde el último incumplimiento de pago ($Mths\_since\_last\_delinq$): El número de meses transcurridos desde la morosidad pasada del prestatario.
4. Meses desde el último registro ($Mths\_since\_last\_record$): El número de meses desde el último registro público.

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. Descripción de los datos

La base de datos alberga un conjunto total de $466,285$ registros, de los cuales aproximadamente el $89\,\%$ se asigna a clientes considerados *buenos*, dejando el porcentaje restante para los clientes catalogados como *malos*.

La Figura 1 visualiza la cantidad de registros ausentes por variable. En particular, *Consultas en los últimos 6 meses* (Inq_last_6mths) y *Cuentas actualmente en mora*(Acc_now_delinq) muestran una incidencia inferior al $0.1\,\%$ de valores NA, siendo la falta de registros atribuible a clientes que no cumplen con esas características. Por lo tanto, es adecuado reemplazar los valores faltantes por ceros.

Por otra parte, las variables Meses desde el último incumplimiento de pago ($Mths\_since\_last\_delinq$) y Meses desde el último registro ($Mths\_since\_last\_record$) exhiben una considerable cantidad de valores NA, los cuales representan el lapso transcurrido desde el último incidente de fraude o manejo inadecuado de la cuenta, así como el número de meses desde el último registro público (como bancarrota, sentencias, juicios, ejecuciones hipotecarias, etc.). Nuevamente, es esencial destacar que la alta prevalencia de valores NA en estas variables no se debe a un error, sino a la inherente naturaleza de las mismas. En escenarios donde estas características no aplican para un individuo, el registro correspondiente se encuentra vacío. En consecuencia, se abordarán de manera independiente los

valores NA en estas variables, reconociendo que su ausencia no denota un error, sino que más bien refleja la inexistencia de ciertas condiciones para determinados clientes.

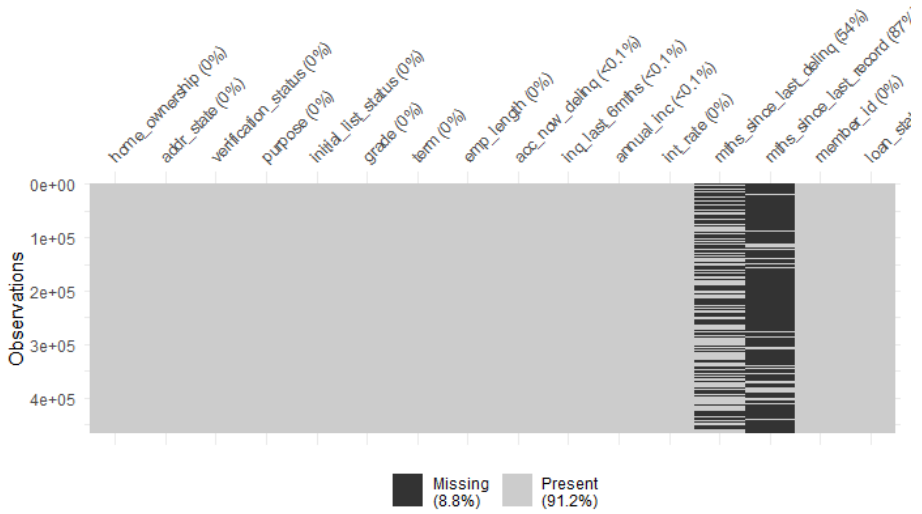

Figura 1: Valores faltantes por variable

Por último, dado que la variable de ingresos cuenta únicamente con 4 valores faltantes, se optó por excluirlos de la base de datos y llevar a cabo un análisis únicamente con los casos completos.

Respecto a la distribución de una de las variables más cruciales en el estudio, la Figura 2 indica que el ingreso presenta una distribución sesgada hacia la derecha. Específicamente, el 95 % de los datos se sitúan por debajo de la marca de $150,000$. Dado que la mayoría de los clientes percibe ingresos inferiores a $150,000$, durante el proceso de categorización se procedió a segmentar el rango de ingresos de 0 a $150,000$ en bins, mientras que otro grupo englobó a aquellos con ingresos superiores.

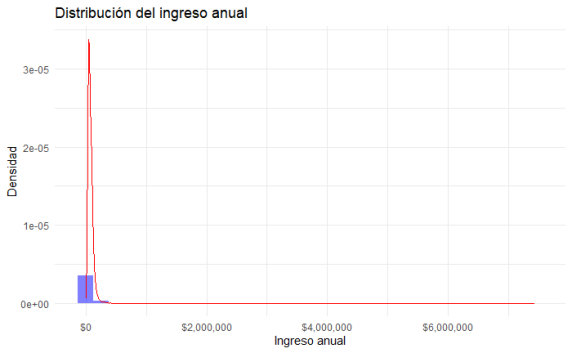

Figura 2: Distribución del ingreso anual

10. Resultados

10.1. Tratamiento de variables

El Coarse Classing consiste en agrupar categorías adyacentes con puntuaciones WOE similares, pues al tener proporciones similares de eventos y no eventos, ambas categorías exhiben un comportamiento similar. Para llevar a cabo esta técnica, se tuvo que cumplir con los siguientes criterios:

- Cada categoría (intervalo) abarca al menos el 5 % de las observaciones. - Las puntuaciones WOE son diferentes para cada categoría, y se deben agrupar aquellas que sean similares. - Las puntuaciones WOE siguen una tendencia monótona, - Los valores faltantes se agrupan de manera independiente.

Para validar la correcta agrupación con WOE, se representaron gráficamente los valores de WOE y se buscó la presencia de linealidad. Las siguientes Figuras 3 y 4 muestran el resultado para las variables Propósito (*Purpose*) y Tasa de interés *(Int_rate)*, pues el mismo procedimiento se llevó a cabo para las variables continuas después de categorizarlas mediante bins (Fine Classing).

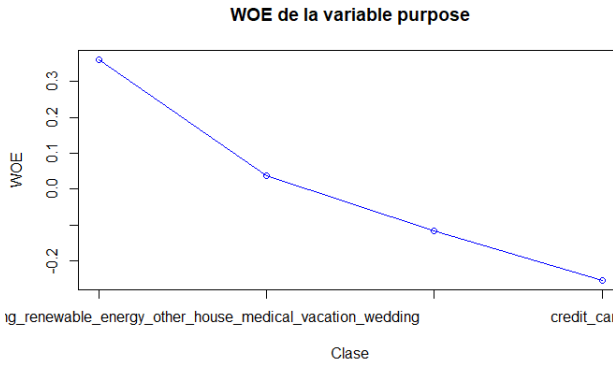

Figura 3: Binning con WOE para a variable Propósito *(Purpose)*

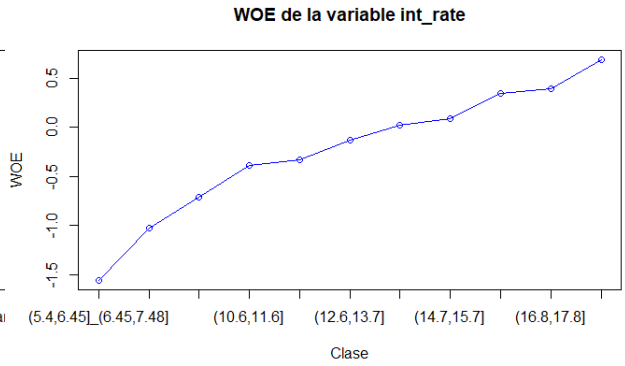

Figura 4: Binning con WOE para a variable Tasa de interés *(Int_rate)*

En modelos de scorecard, es popular emplear una técnica conocida como Valor de Información para la selección de variables. Utilizando esta técnica, se determinó excluir la variable Cuentas actualmente en mora (Acc_now_delinq) debido a que su valor estaba considerablemente por debajo de 0.02.

10.2. Modelo de regresión logística

Después de definir las categorías para cada variable independiente, se procedió a ajustar el modelo de regresión utilizando variables dummy para cada clase, tomando como referencia las categorías con el menor Valor de Peso de la Evidencia (WOE). Los valores estimados de los coeficientes aparecen en la Figura 5, la cual sugiere que todas las variables son significativas a un nivel de confianza del 95 %.

Dado que uno de los tres niveles del factor Verificación de estatus ($Verification_status$) no demostró ser significativo, se realizó una prueba ajustando el modelo sin dicha variable. No obstante, este modelo fue descartado, ya que arrojó un AIC más elevado ($AIC = 242466$) en comparación con el modelo original ($AIC = 242299$).

Cabe mencionar que la numeración de las etiquetas de los coeficientes se corresponde con el orden de las categorías tal como se presentan en el anexo que detalla las categorías reales.

```
Coefficients:
                Estimate Std. Error z value Pr(>|z|)
(Intercept)      3.81071    0.07778  48.993  < 2e-16 ***
d_grade_2        0.52840    0.07365   7.174 7.27e-13 ***
d_grade_3        0.62906    0.07849   8.014 1.11e-15 ***
d_grade_4        0.64351    0.08166   7.880 3.28e-15 ***
d_grade_5        0.35263    0.08298   4.249 2.14e-05 ***
d_home_1        -0.15187    0.01210 -12.553  < 2e-16 ***
d_home_2        -0.00861    0.01999  -0.431 0.666612
d_addr_state_1  -0.12994    0.01111 -11.698  < 2e-16 ***
d_ver_1          0.10967    0.01419   7.731 1.07e-14 ***
d_ver_2          0.11035    0.01298   8.504  < 2e-16 ***
d_pur_1         -0.23600    0.01989 -11.868  < 2e-16 ***
d_pur_2         -0.11182    0.01418  -7.888 3.08e-15 ***
d_pur_3         -0.08299    0.02380  -3.487 0.000488 ***
d_list_2         0.29236    0.01197  24.425  < 2e-16 ***
d_term_1         0.07105    0.01365   5.206 1.93e-07 ***
d_emp_1         -0.13187    0.01804  -7.309 2.70e-13 ***
d_emp_2         -0.04422    0.01433  -3.085 0.002037 **
d_emp_3         -0.10250    0.01806  -5.677 1.37e-08 ***
d_emp_4         -0.07097    0.01730  -4.103 4.09e-05 ***
d_inq_2         -0.12533    0.01291  -9.711  < 2e-16 ***
d_inq_3         -0.25951    0.01627 -15.946  < 2e-16 ***
d_inq_4         -0.43563    0.01847 -23.589  < 2e-16 ***
d_inc_1         -0.13810    0.01699  -8.127 4.40e-16 ***
d_inc_2         -0.50203    0.03411 -14.716  < 2e-16 ***
d_inc_3         -0.55952    0.03840 -14.572  < 2e-16 ***
d_inc_4         -0.26755    0.03192  -8.383  < 2e-16 ***
d_int_2         -0.51516    0.05275  -9.765  < 2e-16 ***
d_int_3         -0.81896    0.05966 -13.728  < 2e-16 ***
d_int_4         -1.45361    0.08885 -16.361  < 2e-16 ***
d_int_5         -1.60071    0.08832 -18.123  < 2e-16 ***
d_int_6         -1.67260    0.08746 -19.124  < 2e-16 ***
d_int_7         -1.90983    0.08802 -21.697  < 2e-16 ***
d_int_8         -2.06895    0.09019 -22.941  < 2e-16 ***
d_int_9         -2.13751    0.09131 -23.410  < 2e-16 ***
d_int_10        -2.34409    0.09181 -25.531  < 2e-16 ***
d_int_11        -2.47823    0.09254 -26.781  < 2e-16 ***
d_delinq_1       0.25138    0.05442   4.619 3.85e-06 ***
d_delinq_2       0.38373    0.11105   3.456 0.000549 ***
d_delinq_3       0.29201    0.09893   2.952 0.003160 **
d_delinq_4       0.22899    0.09391   2.439 0.014747 *
d_delinq_6       0.35758    0.05445   6.567 5.12e-11 ***
d_record_2       0.16029    0.01606   9.983  < 2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Figura 5: Estimaciones de los coeficientes del modelo

10.3. Comprobación del modelo

Curva ROC

La Curva ROC se presenta en la Figura 6, y se aprecia que no está tan ampliamente abierta. El AUC abarca el área que hay entre la curva ROC y la recta identidad más el área del triangulo

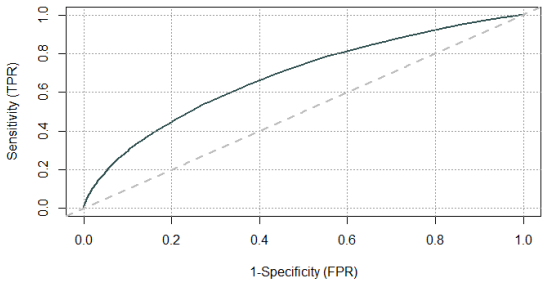

Figura 6: Curva ROC

inferior derecho cuya hipotenusa es la recta identidad .El coefficente de Gini se visualiza mediante el triángulo superior izquierdo; se define como el área que hay entre la curva ROC y la recta identidad divido entre el área de todo el triángulo, el cual es igual a 0.5

Específicamente, según el modelo ajustado, se logró un área bajo la curva de 0.68; esto sugiere que la capacidad del modelo para distinguir entre clases no es muy destacada, aunque tampoco es completamente aleatoria.

En relación con el Coeficiente de Gini, se alcanzó un valor de 0.36, lo que señala que el modelo presenta aproximadamente un 36 % de ventaja en comparación con una clasificación aleatoria para discernir correctamente entre un cliente bueno y uno malo.

Densidades de las clases

La Figura 7 presenta las probabilidades de cumplimiento tanto para los clientes buenos como para los malos. Se nota que, en ciertas áreas, las distribuciones se superponen, lo que complica la diferenciación entre ellas.

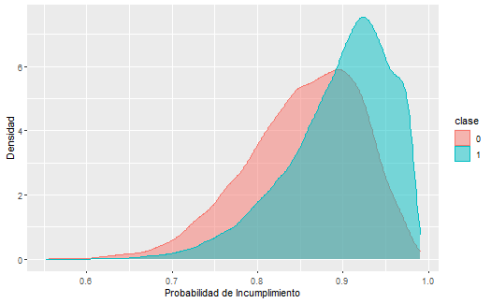

Figura 7: Probabilidad de cumpliento por clase

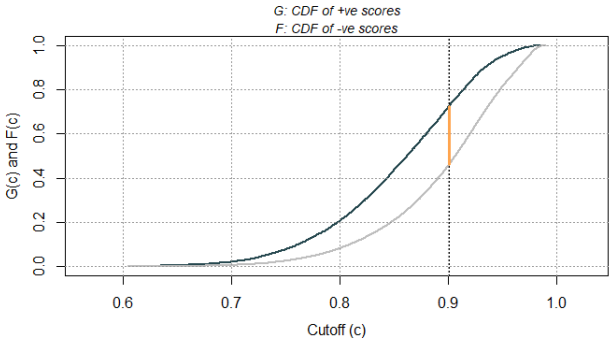

Figura 8: Gráfica de Kolmogorov-Smirnov

De manera formal, mediante el test de Kolmogorov-Smirnov, se probó la hipotesis nula de que, bajo la predicción del modelo, la clase positiva (clientes buenos) y la clase negativa (clientes malos) tienen la misma distribución. Se obtuvo que el mayor grado de separación entre las funciones de distribución empírica para cada clase es de 0.26, esto es posible de comparar visualmente por medio de la Figura 8, en donde la curva menos desplazada hacia la derecha representa la función de distribución acumulada de la clase negativa, mienta que la curva restante representa la función de distribución acumulada de la clase positiva. Dado que la línea de color naranja, que representa la distancia vertical máxima entre las dos funciones, no se encuentra en un solo punto, surge la sospecha de que el modelo exhibe cierto desempeño en la clasificación. Si la distancia vertical máxima fuera igual a cero, implicaría que el modelo estaría realizando predicciones de manera aleatoria.

Bajo esta prueba, se calculó la zona de rechazo, la cual está dada por $\left\{T > 1.36\sqrt{\frac{m+n}{nm}}\right\}$ para un nivel de significancia del 5 %, donde $m = 82954$ y $n = 10303$ representan el tamaño de la muestra de la clase positiva (clientes buenos) y la clase negativa (clientes malos), respectivamente. Por lo tanto, $\{T > 0.0142\}$ y puesto que, la distancia máxima entre las funciones es mayor a 0.01423, hay suficiente evidencia para rechazar la hipotesis nula y decidirse por la hipotesis alternativa, la cual indica que las funciones de distribución para cada muestra son diferentes.

Curva CAP

La Figura 9 presenta la curva de Aprovechamiento Cumulativo (CAP) en color rosa, generada por el modelo ajustado. En este contexto, la recta identidad representa la curva CAP correspondiente al modelo sin capacidad discriminatoria, mientras que la curva morada representa la curva CAP de un modelo perfecto. Se observa que la forma de la curva CAP del modelo desarrollado no se aproxima significativamente a la forma de la curva CAP de un modelo perfecto. Esto sugiere que la capacidad de discriminación del modelo ajustado no es muy robusta. No obstante, también se nota que el modelo ajustado supera a un modelo aleatorio, indicando un desempeño moderado en términos de capacidad discriminatoria.

En particular, se puede observar que, por ejemplo, si se deseara rechazar aproximadamente el 25 % de los malos, se tendría que rechazar aproximadamente el 12.5 % del total de solicitantes de crédito.

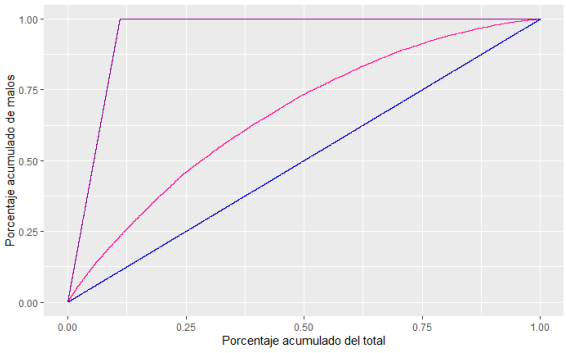

Figura 9: Enter Caption

10.4. Umbral de discriminación óptimo

El índice de Youden refleja la capacidad general del modelo de no cometer el error tipo 1 y tipo 2, pues se calcula de la siguiente manera:

$$\text{YOUDEN} \;=\; \text{Sensibilidad} \;+\; \text{Especifidad} \;-1$$

donde la especifidad es la tasa de verdaderos negativos (TNR, por sus siglas en inglés).

$$\text{Especifidad} \;= \frac{\text{Verdaderos negativos}}{\text{Verdaderos negativos} \;+\; \text{falsos positivos}} = 1 - FPR$$

En consecuencia, considerando que la muestra de prueba es representativa de la clientela poblacional, el punto de corte óptimo (threshold) se determina como aquel que maximiza el índice de Youden, calculado a partir de la matriz de confusión obtenida de la muestra de prueba. Este threshold también optimiza la sensibilidad al tiempo que minimiza la tasa de falsos positivos, colocándolo en el punto más cercano a (0,1) en la curva ROC.

Según el modelo ajustado, se determinó que el umbral de discriminación óptimo, redondeado, es de 0.90.

10.5. Scoring

Una interpretación del puntaje asignado por el Score es:

Score	Tipo de Cliente
↑	Bueno
↓	Malo

Para el cálculo del score correspondiente a cada categoría en el modelo final, incluyendo las categorías de referencia en la creación del modelo final, se utilizó como base un score máximo de 850 y mínimo de 300 . Siendo la fórmula para calcularlos la siguiente:

$$SCORE = \begin{cases} \left(\frac{\beta_i - min\_suma}{max\_suma - min\_suma}\right)(max\_score - min\_score) + min\_score & i = 0 \\ \left(\frac{\beta_i - min\_suma}{max\_suma - min\_suma}\right) & de\ otra\ forma \end{cases}$$

donde max_suma y min_suma se obtienen de sumar los valores máximo y mínimo para cada base definido anteriormente.

Los scores obtenidos se detallan en el anexo.

El cuadro **??** muestra el máximo y mínimo por variable. En particular, para el modelo ajustado, se obtuvo que la suma de los mínimos coeficientes (min_suma) es -0.31, mientras que la suma de los máximos (max_suma) es 5.47

Variable	Min_coeficiente	Max_coeficiente
addr_state	-0.13	0.00
mths_since_last_delinq	0.00	0.38
emp_length	-0.13	0.00
grade	0.00	0.64
home_ownership	-0.15	0.00
annual_inc	-0.56	0.00
inq_last_6mths	-0.44	0.00
int_rate	-2.48	0.00
Intercepto	3.81	3.81
initial_list_status	0.00	0.29
purpose	-0.24	0.00
mths_since_last_record	0.00	0.16
term	0.00	0.07
verification_status	0.00	0.11

Cuadro 1: Mínimos y máximos coeficientes

10.6. Probabilidades como scores

Se puede reemplazar una probabilidad de cumplimiento por el score correspondiente mediante la siguiente fórmula:

$$Probabilidad\,como\,score = \left(ln\left(\frac{prob\_bueno}{1 - prob\_bueno}\right) - min\_suma \right) \frac{max\_score - min\_score}{max\_suma - min\_suma} + min\_sum$$

La siguiente tabla presenta la proporción de individuos aprobados y rechazados en relación con un score considerado como umbral, junto con la probabilidad de incumplimiento (PD) correspondiente a dicho umbral. A partir de la tabla, se evidencia que a medida que la probabilidad de incumplimiento aumenta, el puntaje asociado tiende a disminuir.

PD	Score	Tasa_Aprob	Tasa_Rechazo	Tasa_Aprob1000	Tasa_Rechazo1000
1 %	732.00	0.00	1.00	0.05	999.95
2 %	626.00	0.03	0.97	26.85	973.15
3 %	567.00	0.10	0.90	95.63	904.37
4 %	525.00	0.17	0.83	171.57	828.43
5 %	492.00	0.25	0.75	247.97	752.03
6 %	464.00	0.34	0.66	338.67	661.33
7 %	441.00	0.44	0.56	436.20	563.80
8 %	421.00	0.54	0.46	541.04	458.96
9 %	402.00	0.64	0.36	642.70	357.30
10 %	386.00	0.74	0.26	738.91	261.09

La tasa de aprobación para cada puntaje se calculó a partir de la muestra de prueba, dividiendo

el número de registros con un puntaje mayor al establecido como umbral, entre el total de registros en la muestra de prueba. La tasa de rechazo es simplemente el complemento de la tasa de aprobación.

Según el modelo ajustado, el puntaje óptimo considerado como umbral de discriminación es 386. Con este umbral, de cada 1000 solicitantes de crédito, se aprobarían 738 y se rechazarían 261 individuos, lo que representa poco menos de la mitad del total de solicitantes.

11. Conclusión

El modelo resultante de la regresión logística no alcanza la categoría de excelente, pero exhibe un poder de discriminación aceptable. Por ende, los puntajes derivados de las estimaciones de los coeficientes para cada categoría pueden servir como referencia para calcular la puntuación total de un solicitante de crédito. Esta puntuación se obtiene sumando los puntajes asociados a las características cumplidas por el cliente. Si la puntuación resultante es inferior a 386, se rechazaría la solicitud, ya que el cliente sería considerado de alto riesgo.

12. Anexo

Categorías del modelo final

Categoria	WOE	Total_buenos	Total_malos	Total	Total_porcentaje
ANY, NONE, OTHER, RENT	0.16	18994	132046	151040	0.40
OWN	-0.01	3601	29770	33371	0.09
MORTGAGE	-0.14	18070	170543	188613	0.51

Cuadro 2: Propiedad de vivenda ($Home\_ownership$)

Categoría	WOE	Total Buenos	Total Malos	T
NE, IA, ID, NV, HI, FL, AL, LA, NY, MD, NC, MO, NM, OK, NJ, VA	0.09	15,394	114,525	129,
Complemento	-0.05	25,271	217,834	243,

Cuadro 3: Estado ($Addr\_state$)

Categoria	WOE	Total_buenos	Total_malos	Total	Total_porcentaje
Verified	0.18	17173	117473	134646	0.36
Source Verified	-0.06	12389	107412	119801	0.32
Not Verified	-0.17	11103	107474	118577	0.32

Cuadro 4: Verificación de estatus ($Verification\_status$)

Categoria	WOE	Total_buenos	T
small business,educational,moving,renewable energy,other,house,medical,vacation,wedding	0.36	5510	
debt consolidation	0.04	24696	
home improvement,major purchase	-0.12	2864	
credit_card_car	-0.26	7595	

Cuadro 5: Propósito ($Purpose$)

Categoria	WOE	Total_buenos	Total_malos	Total	Total_porcentaje
f	-0.11	29106	213479	242585	0.65
w	0.23	11559	118880	130439	0.35

Cuadro 6: Estado inicial de la lista ($Initial\_list\_status$)

Categoria	WOE	Total_buenos	Total_malos	Total	Total_porcentaje
A	-1.11	2323	57583	59906	0.16
B	-0.36	8624	101046	109670	0.29
C	0.06	11518	88749	100267	0.27
D	0.39	9411	52065	61476	0.16
E_F_G	0.78	8789	32916	41705	0.11

Cuadro 7: Grado (*Grade*)

Categoria	WOE	Total_buenos	Total_malos	Total	Total_porcentaje
36 months	0.13	26256	244294	270550	0.73
60 months	-0.29	14409	88065	102474	0.27

Cuadro 8: Plazo (*Term*)

Categoria	WOE	Total_buenos	Total_malos	Total	Total_porcentaje
< 1 year	0.14	5663	40218	45881	0.12
1 year - 4 years	0.01	11944	96662	108606	0.29
5 years - 6 years	0.07	5267	40211	45478	0.12
7 year - 9 years	0.01	5853	47165	53018	0.14
10+ years	-0.10	11938	108103	120041	0.32

Cuadro 9: Duración de empleo (*Emp_length*)

Categoria	WOE	Total_buenos	Total_malos	Total	Total_porcentaje
0	-0.20	17579	175537	193116	0.52
1	0.06	11963	92251	104214	0.28
2	0.25	6300	39977	46277	0.12
3>=	0.47	4823	24594	29417	0.08

Cuadro 10: Consultas en los últimos 6 meses (*Inq_last_6mths*)

Categoria	WOE	Total_buenos	Total_malos	Total	Total_porcentaje
<4.93e+04	0.25	15259	97345	112604	0.30
(4.93e+04,5.82e+04]	0.08	5823	43873	49696	0.13
(5.82e+04,6.11e+04]	0.12	2301	16709	19010	0.05
(6.11e+04„1.5e+05]	-0.21	17282	174432	191714	0.51
1.5e+05>	-	-	-	-	0.05

Cuadro 11: Ingreso anual (*annual_inc*)

Categoria	WOE	Total_buenos	Total_malos	Total	Total_porcentaje
<7.48	-1.56	553	21572	22125	0.06
(7.48,8.52]	-1.03	1141	26021	27162	0.07
(8.52,10.6]	-0.71	2153	35921	38074	0.10
(10.6,11.6]	-0.39	2078	24982	27060	0.07
(11.6,12.6]	-0.33	2904	33061	35965	0.10
(12.6,13.7]	-0.13	3842	35901	39743	0.11
(13.7,14.7]	0.02	4224	33812	38036	0.10
(14.7,15.7]	0.09	3751	28151	31902	0.09
(15.7,16.8]	0.35	3378	19514	22892	0.06
(16.8,17.8]	0.40	3916	21502	25418	0.07
17.8>	0.69	12725	51922	64647	0.17

Cuadro 12: Tasa de interés (Int_rate)

Categoria	WOE	Total_buenos	Total_malos	Total	Total_porcentaje
1	0.06	2684	20644	23328	0.06
2	-0.01	3119	25873	28992	0.08
3	-0.05	2639	22625	25264	0.07
4	-0.08	2316	20456	22772	0.06
5>	-0.01	7820	64566	72386	0.19
NA	0.01	22087	178195	200282	0.54

Cuadro 13: Meses desde el último incumplimiento de pago ($Mths\_since\_last\_delinq$)

Categoria	WOE	Total_buenos	Total_malos	Total	Total_porcentaje
1>=	-0.04	5293	44860	50153	0.13
NA	0.01	35372	287499	322871	0.87

Cuadro 14: Meses desde el último registro ($Mths\_since\_last\_record$)

Scores

Variable	Coeficientes	Score
(Intercept)	3.81	614.49
d_grade_2	0.53	74.91
d_grade_3	0.63	89.18
d_grade_4	0.64	91.23
d_grade_5	0.35	49.99
d_home_1	-0.15	-21.53
d_home_2	-0.01	-1.22
d_addr_state_1	-0.13	-18.42
d_ver_1	0.11	15.55
d_ver_2	0.11	15.64
d_pur_1	-0.24	-33.46
d_pur_2	-0.11	-15.85
d_pur_3	-0.08	-11.77
d_list_2	0.29	41.45
d_term_1	0.07	10.07
d_emp_1	-0.13	-18.69
d_emp_2	-0.04	-6.27
d_emp_3	-0.10	-14.53
d_emp_4	-0.07	-10.06
d_inq_2	-0.13	-17.77
d_inq_3	-0.26	-36.79
d_inq_4	-0.44	-61.76
d_inc_1	-0.14	-19.58
d_inc_2	-0.50	-71.17
d_inc_3	-0.56	-79.32
d_inc_4	-0.27	-37.93
d_int_2	-0.52	-73.03
d_int_3	-0.82	-116.10
d_int_4	-1.45	-206.07
d_int_5	-1.60	-226.92
d_int_6	-1.67	-237.11
d_int_7	-1.91	-270.74
d_int_8	-2.07	-293.30
d_int_9	-2.14	-303.02
d_int_10	-2.34	-332.30
d_int_11	-2.48	-351.32
d_delinq_1	0.25	35.64
d_delinq_2	0.38	54.40
d_delinq_3	0.29	41.40
d_delinq_4	0.23	32.46
d_delinq_6	0.36	50.69
d_record_2	0.16	22.72

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]