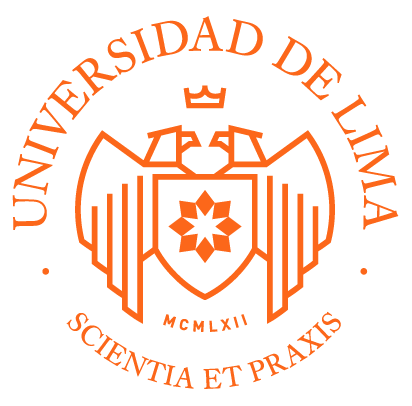
Universidad de Lima

Facultad de Ingeniería

Estadística Aplicada





**TRABAJO DE APLICACIÓN: CRÉDITO BANCARIO**

Informe – Sección 523

**Sebastián Alonso Álvarez Domínguez**

**Código 20220109**

[Hector Felipe Gutierrez Ascasibar](mailto:20221178@aloe.ulima.edu.pe)

**20221178**

[Ralf Fernando Callalli Villalta](mailto:20220449@aloe.ulima.edu.pe)

**20220449**

**Sáez Olin Gabriel Enrique**

**20215303**

**Jara Espinoza Rodrigo**

**20224280**

Aldo Richard Meza Rodriguez

Lima – Perú

2024

**TABLA DE CONTENIDO**

[**INTRODUCCIÓN**](#_heading=h.rtq5g7h9cqw5) **1**

[**CAPÍTULO I: PLANTEAMIENTO DE OBJETIVOS**](#_heading=h.pkrl1hvf752v) **2**

[**CAPÍTULO II: ANÁLISIS E INTERPRETACIÓN DE RESULTADOS**](#_heading=h.pe9eiz7138me) **3**

[2.1 Análisis exploratorio y/o descriptivo](#_heading=h.1ehg45p66puq) 3

[2.2 Modelación](#_heading=h.cczrcfuewbkn) 13

[2.2.1 Modelo de Regresión Lineal](#_heading=h.vm1rhvgys17d) 13

[2.2.2 Modelo de Regresión Logística](#_heading=h.i9vv3rpghpcw) 20

[**CONCLUSIONES Y RECOMENDACIONES**](#_heading=h.5etd8mqbodo2) **27**

[**ANEXOS**](#_heading=h.klj0fvpj06w0) **28**

**ÍNDICE DE FIGURAS**

**Figura 2.1 *Gráfico de pie de distribución por sexo*****4**

**Figura 2.2 *Gráfico de barras de distribución por Estado Civil*****4**

**Figura 2.3 *Gráfico de barras por distribución nivel de educación*****5**

**Figura 2.4 *Gráfico de pie por distribución por prioridad*****6**

**Figura 2.5 *Gráfico de barras por concesión de crédito*****6**

**Figura 2.6 *Histograma de distribución de edad*****7**

**Figura 2.7 *Boxplot de distribución de cantidad de tarjetas*****8**

**Figura 2.8 *Boxplot de distribución de saldo*****8**

**Figura 2.9 *Histograma de distribución de saldo*****9**

**Figura 2.10 *Histograma de distribución de puntaje de crédito*****10**

**Figura 2.11 *Histograma de distribución de años de empleo*****10**

**Figura 2.12 *Histograma de distribución de ingresos*****11**

**Figura 2.13 *Histograma de distribución de monto de crédito*****11**

**Figura 2.2.1 *Modelo de regresión lineal*** **12**

**Figura 2.2.2 *Modelo de regresión lineal*****14**

**Figura 2.2.3 *Modelo de regresión lineal ajustado*****15**

**Figura 2.2.4 *Prueba de Anderson-Darling*****16**

**Figura 2.2.5 *Supuesto de no multicolinealidad*****17**

**Figura 2.2.6 *Modelo de regresión logística*****18**

**Figura 2.2.7 *Estadístico chi2*****19**

**Figura 2.2.8 *Grados de libertad*****19**

**Figura 2.2.9 *P-value*****19**

**Figura 2.2.10 *Modelo de regresión logística*****20**

**Figura 2.2.11 *Modelo de Regresión Logística Ajustado*****21**

**Figura 2.2.12 *Odds ratio*****21**

**Figura 2.2.13 *Predicción 1*****22**

**Figura 2.2.14 *Predicción 2*****23**

**Figura 2.2.15 *Predicción 3*****23**

**Figura 2.2.16 *Predicción 4*****24**

# INTRODUCCIÓN

En el ámbito bancario, la concesión de créditos es una de las actividades más fundamentales y complejas. Evaluar adecuadamente el riesgo crediticio asociado con cada solicitante es crucial para tomar decisiones informadas y mitigar los riesgos financieros. Para ello, los bancos recopilan una gran cantidad de datos sobre los clientes potenciales, que incluyen información financiera, historial de crédito, ingresos, deudas y otros factores relevantes que aportan a su evaluación. Estos datos son utilizados para desarrollar modelos estadísticos, como la regresión, que permiten predecir la probabilidad de incumplimiento de los préstamos. Estos modelos son esenciales para minimizar las pérdidas financieras y optimizar la gestión de la cartera de préstamos.

A pesar de los avances significativos en la modelización de riesgos crediticios, los bancos todavía enfrentan varios desafíos importantes. Uno de estos desafíos radica en la identificación de las variables más influyentes en el cálculo del riesgo crediticio. Determinar qué factores tienen mayor impacto puede ser complejo en la mayoría de los casos debido a la interrelación de múltiples variables. Además, es importante encontrar el modelo adecuado para obtener resultados correctos. Por otro lado, la interpretación también representa un obstáculo, ya que los modelos de regresión son difíciles de comprender, lo que limita la capacidad de los bancos para tomar decisiones informadas.

Por lo tanto, es evidente la necesidad de abordar estos desafíos para mejorar la efectividad de los modelos de regresión en la evaluación del riesgo crediticio. Una comprensión más profunda de las variables relevantes, la selección adecuada del modelo y una interpretación clara de los resultados son elementos clave para optimizar la toma de decisiones en el ámbito bancario. En este estudio, se busca explorar estos aspectos con el fin de ofrecer recomendaciones prácticas que mejoren la gestión del riesgo crediticio y fortalezcan la salud financiera de las instituciones bancarias.

# CAPÍTULO I: PLANTEAMIENTO DE OBJETIVOS

**1.1 Objetivo general:**

Mejorar la eficacia de los modelos de regresión utilizados para evaluar el riesgo crediticio, permitiendo a los bancos tomar decisiones más informadas sobre la concesión de créditos y, de esta manera gestionar mejor su cartera de préstamos y minimizar las pérdidas financieras.

(Determinar los factores determinantes para ver si se le debe un crédito bancario a un cliente)

**1.2 Objetivo específicos**

* Identificar las variables más relevantes: Utilizar técnicas de análisis exploratorio de datos y selección de características para identificar las variables que tienen el mayor impacto en el riesgo crediticio. Esto incluye el uso de técnicas sobre el manejo de datos y análisis estadístico.

(Determinar qué variables son influyentes)

(Determinar cuál es el mejor modelo)

(Determinar cuántas variables se deben retirar del modelo)

(Identificar/dar solución al primer problema de la banca)

* Desarrollar modelos de regresión robustos: Explorar diferentes algoritmos de regresión y regresión logística. Esto también implica validar los modelos utilizando técnicas de validación cruzada y conjuntos de datos de prueba independientes.
* Interpretar el modelo: Emplear métodos de interpretación de modelo para comprender mejor cómo cada variable influye en el riesgo crediticio. Esto ayuda a desglosar el impacto de cada variable y proporciona una interpretación más clara y accesible de los resultados del modelo.
* Evaluar el rendimiento del modelo: Utilizar métricas de evaluación del rendimiento, como la precisión, la sensibilidad, la especificidad, el valor predictivo positivo, el valor predictivo negativo, para evaluar y comparar la eficacia de los modelos desarrollados. Estas métricas permitirán determinar qué modelos son los más efectivos para predecir el riesgo de incumplimiento y ayudarán a los bancos a tomar decisiones más informadas sobre la concesión de créditos.

# CAPÍTULO II: ANÁLISIS E INTERPRETACIÓN DE RESULTADOS

## 2.1 Análisis exploratorio y/o descriptivo

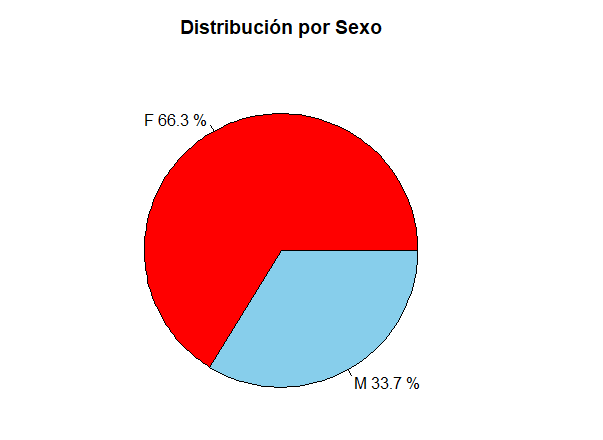
El conjunto de datos utilizado en este estudio Crédito de Banco, consta de 600 registros y abarca un total de 13 variables relevantes.

* Monto (Y1): Monto de crédito
* Crédito (Y2): Si se le debe dar el crédito al cliente (Si, No)
* Sexo: Sexo del cliente (Masculino, femenino)
* E.Civil: Estado civil del cliente
* Educación: Grado de instrucción
* Prioridad: Si el cliente es prioritario (Si, No)
* Edad: Edad del cliente
* Tarjetas: Cantidad de tarjetas de crédito del cliente
* Deuda: Deuda del cliente
* Saldo: Saldo del cliente
* CrediScore: Puntaje de crédito del cliente
* Años\_empleo: Años de empleo del cliente
* Ingresos: Ingresos del cliente

En primer lugar, el monto del crédito solicitado (Y1) tiene un impacto directo en la evaluación del riesgo crediticio. Cuanto mayor sea el monto, mayor podría ser el riesgo para el banco. Comprender cómo esta variable se relaciona con otras características del cliente es fundamental para determinar la viabilidad del crédito. En segundo lugar, la variable Y2 (Crédito) indica si el crédito debe ser otorgado o no. Esta decisión es fundamental para los bancos, ya que determina la concesión final del crédito. Modelar y predecir esta variable es esencial para tomar decisiones informadas y gestionar de manera efectiva el riesgo crediticio.

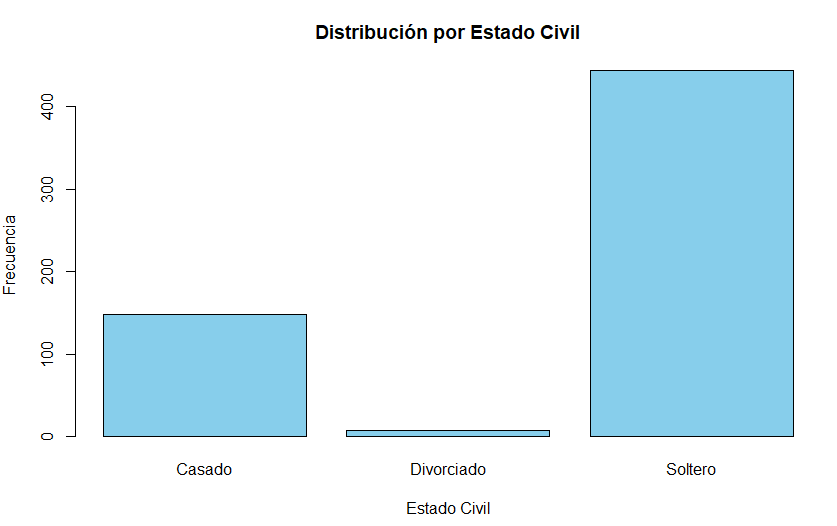
Variables cualitativas (Sexo, Estado civil, Educación, prioridad, Crédito)  
En el análisis de variables cualitativas, utilizamos gráfico de barras(barplot) y pie para poder visualizar la distribución y comparaciones de diferentes categorías.

**Figura 2.1***Gráfico de pie de distribución por sexo*



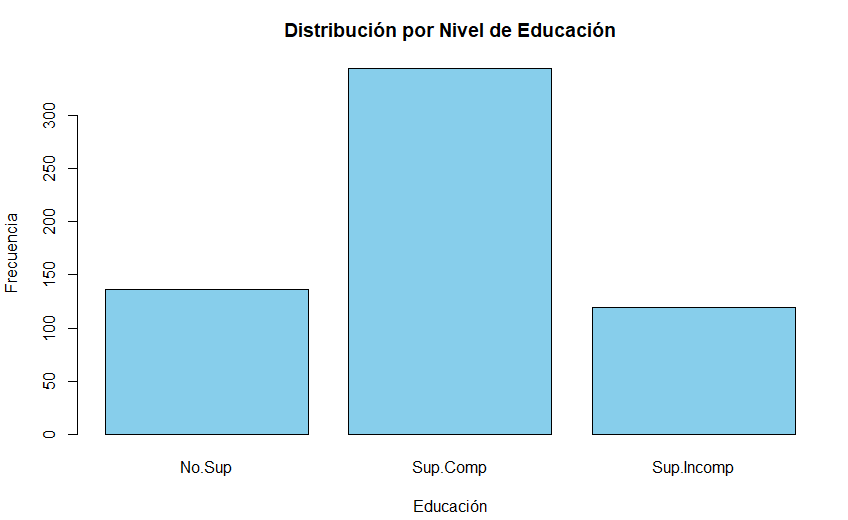
En el siguiente gráfico, se observa dentro de nuestra muestra la categoría “Sexo”, nos indica la superioridad del genero femenino con un 66.3% en comparación del género masculino con un 33.7%. En otras palabras, de cada 3 personas de la muestra, 2 son mujeres aproximadamente.

**Figura 2.2***Gráfico de barras de distribución por Estado Civil*

**

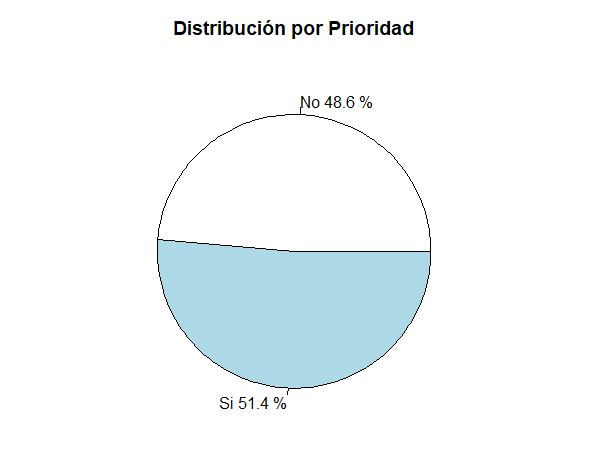
En la figura 2.2 se observa la distribución de la muestra de estudio según el "Estado Civil". Se distinguen tres categorías: Casados, Divorciados y Solteros. Predominan los "Solteros" con una frecuencia superior a 400 personas, lo cual indica que la mayoría de los participantes en la muestra no están casados. En contraste, los "Divorciados" son la categoría menos representada, con una frecuencia muy baja, lo que sugiere que hay pocos participantes en esta situación. Los "Casados" presentan una frecuencia intermedia, con una cantidad significativa de personas, aunque menor en comparación con los "Solteros".

**Figura 2.3***Gráfico de barras por distribución nivel de educación*

**

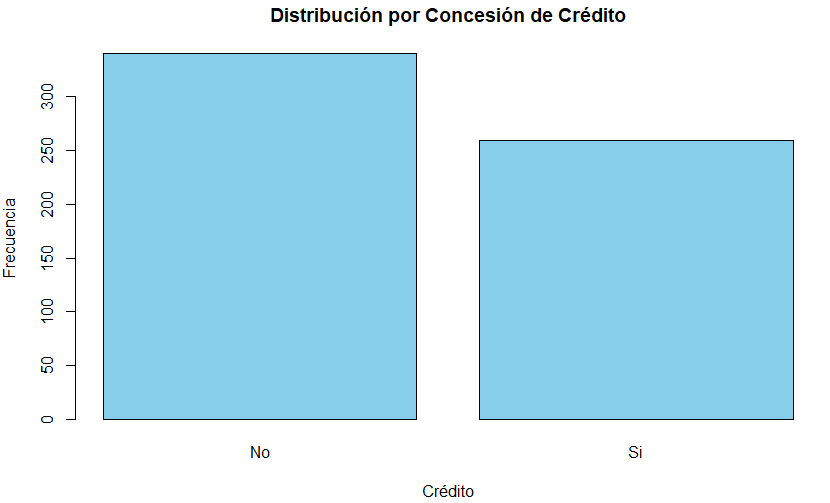
En la figura 2.3 se observa la distribución de la muestra de estudio según el "Nivel de Educación". Se distinguen tres categorías: No.Superior (No.Sup), Superior Completo (Sup.Comp) y Superior Incompleto (Sup.Incomp). Predomina el nivel "Superior Completo" con una frecuencia superior a 300 personas, indicando que la mayoría de los participantes han finalizado sus estudios superiores. En contraste, el nivel "No.Superior" presenta una frecuencia significativamente menor, alrededor de 100 personas, lo que sugiere que una minoría de los participantes no ha accedido a la educación superior. Este patrón puede reflejar diversas tendencias educativas y socioeconómicas de la población estudiada, como el acceso a la educación superior, la tasa de finalización de estudios y las diferencias en la formación académica.

**Figura 2.4***Gráfico de pie por distribución por prioridad*



En la figura 2.4 se observa en este gráfico circular que se tiene dos grupos, uno que pertenecen al “Si” con el 51.4%, que si poseen prioridad al recibir el crédito, y por el otro existen un 48.6% que no son acreedores del Crédito.

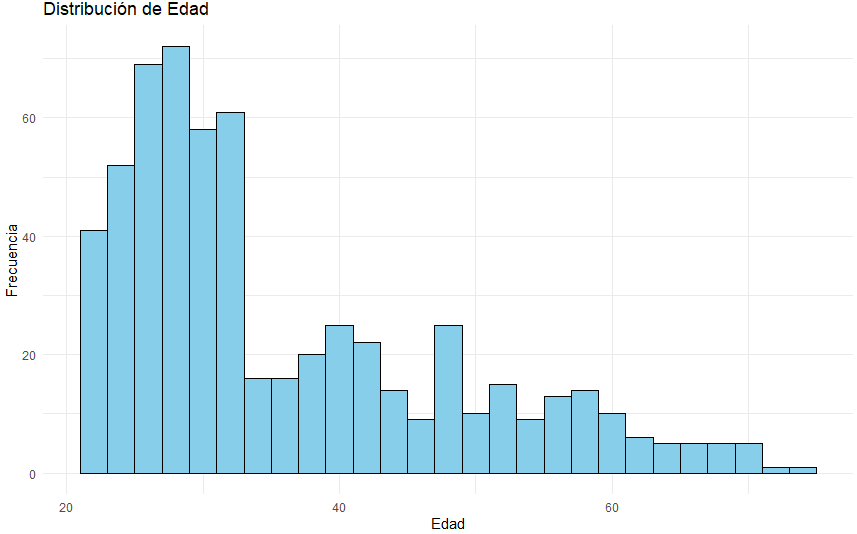
**Figura 2.5***Gráfico de barras de barras por concesión de crédito*

**

En la figura 2.5 se observa lo visto en la anterior figura, pero en forma de frecuencia permite ver cómo se encuentran las cantidades de personas a las que se les dió el crédito y a quienes no se les dio el crédito.

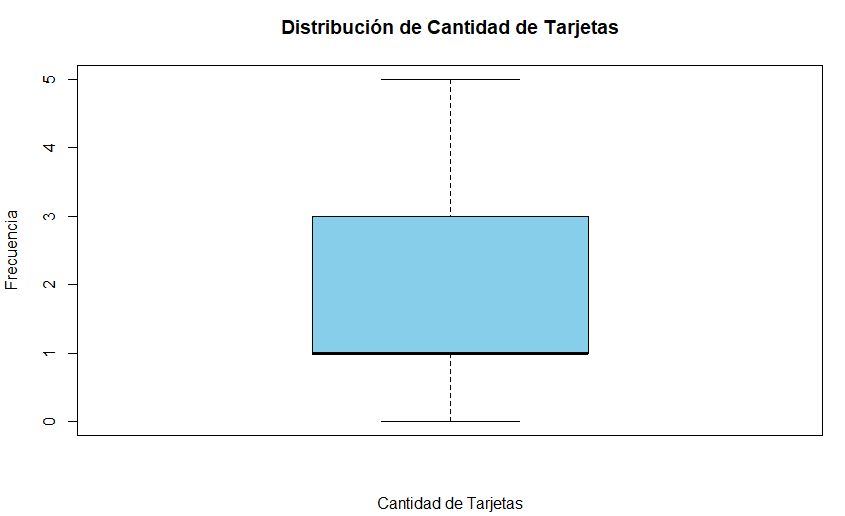
Variables cuantitativas(Sexo, Estado civil, Educación, prioridad, Crédito)  
Para las variables cuantitativas, utilizamos histogramas y boxplot para visualizar cómo los datos se distribuyen a lo largo de un rango continuo. Estos permiten observar la forma de la distribución de los datos, detectar asimetrías, picos y dispersión

**Figura 2.6***Histograma de distribución de edad*



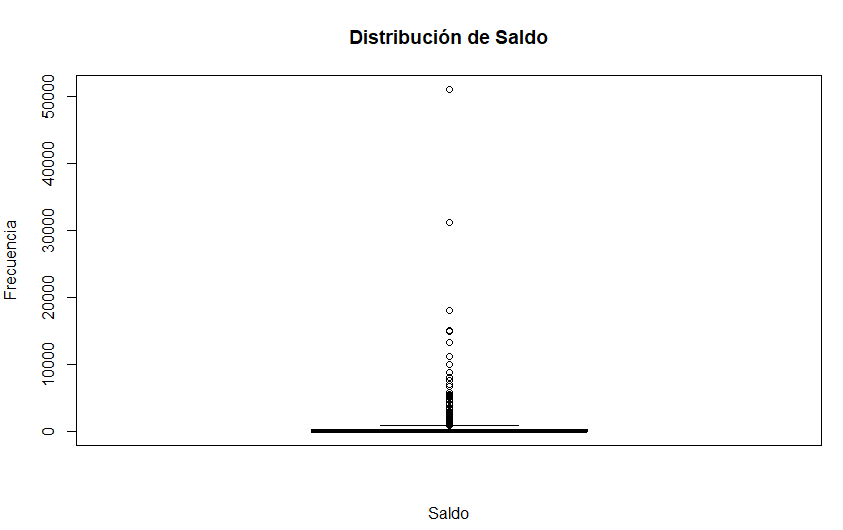
En la figura 2.6 se observa las personas que buscan el crédito, el histograma de distribución de edad muestra una alta concentración de individuos entre 20 y 40 años, con un pico alrededor de los 25-30 años. La frecuencia disminuye gradualmente con la edad, extendiéndose hasta los 60-65 años. Esta distribución sugiere una población mayormente joven, posiblemente de estudiantes universitarios o trabajadores jóvenes.

**Figura 2.7***Boxplot de distribución de cantidad de tarjetas*

**

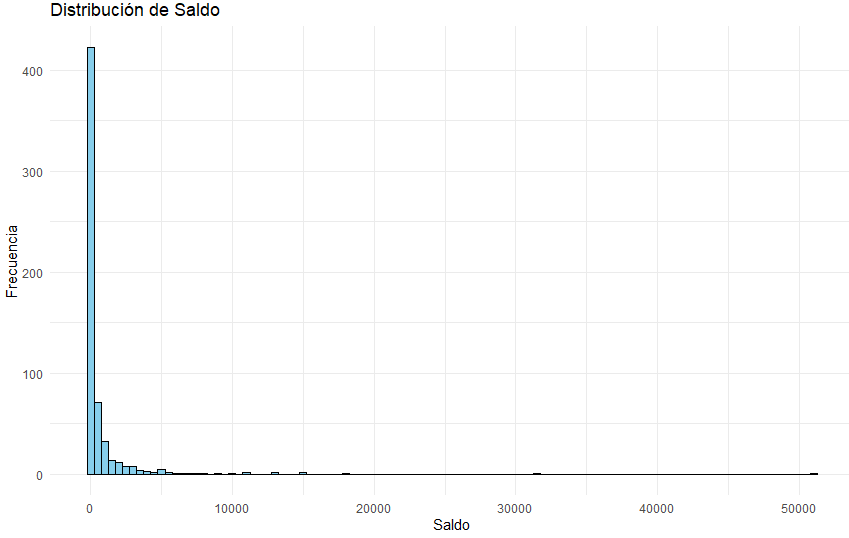
En la figura 2.8, se observa, el diagrama de cajas muestra la distribución de la cantidad de tarjetas. El rango intercuartil se encuentra entre aproximadamente 2 y 4 tarjetas, con una mediana en 3 tarjetas. No se observan valores atípicos significativos y la distribución parece ser simétrica.

**Figura 2.8***Boxplot de distribución de saldo*

**

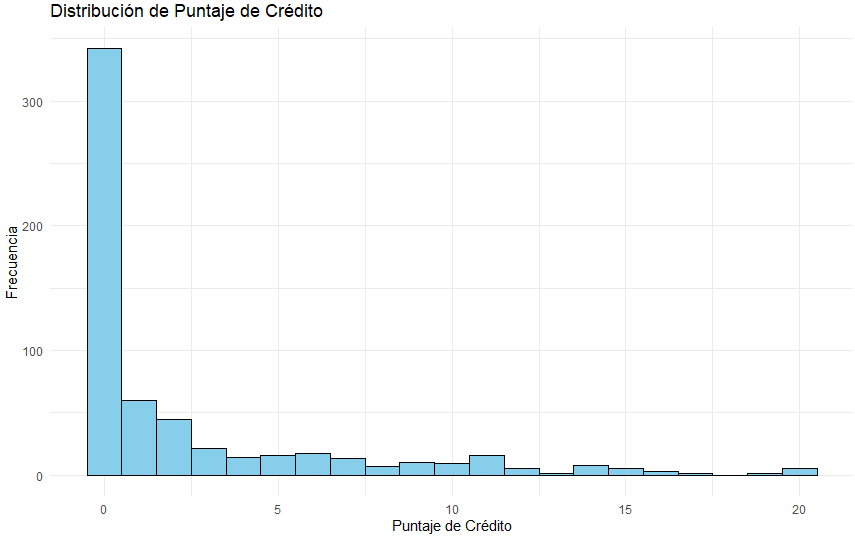
En la figura 2.8 se observa que tanto el diagrama de cajas muestra la distribución del saldo, destacando una gran cantidad de valores atípicos que se extienden por encima del límite superior. La mayoría de los datos se concentran cerca del límite inferior, sugiriendo una distribución asimétrica con una larga cola hacia la derecha. La mediana del saldo es baja en comparación con los valores atípicos, indicando que la mayoría de los saldos son pequeños.

**Figura 2.9***Histograma de distribución de saldo*

**

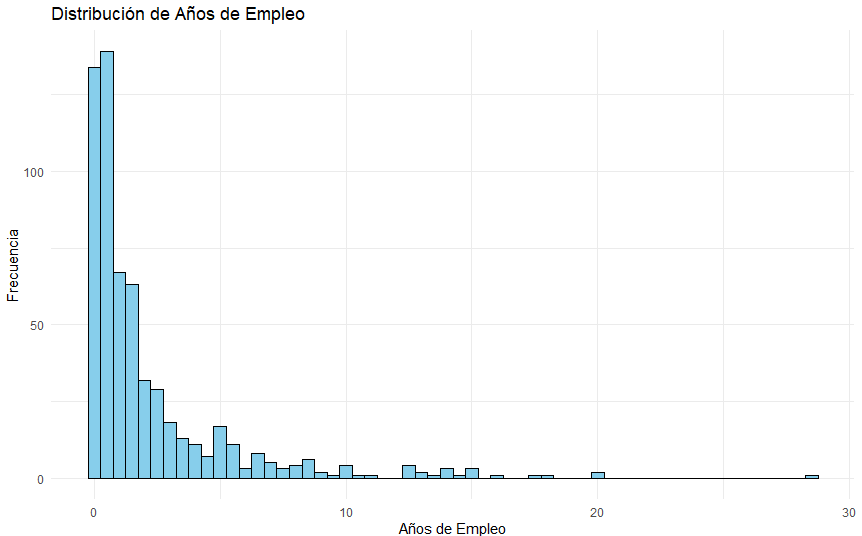
En la figura 2.9 se observa al histograma que muestra una alta concentración de saldos bajos, con la mayoría de los datos cerca del valor mínimo. A medida que el saldo aumenta, la frecuencia disminuye rápidamente, indicando que pocos casos tienen saldos elevados. La distribución es claramente sesgada a la derecha.

**Figura 2.10***Histograma de distribución de puntaje de crédito*

**

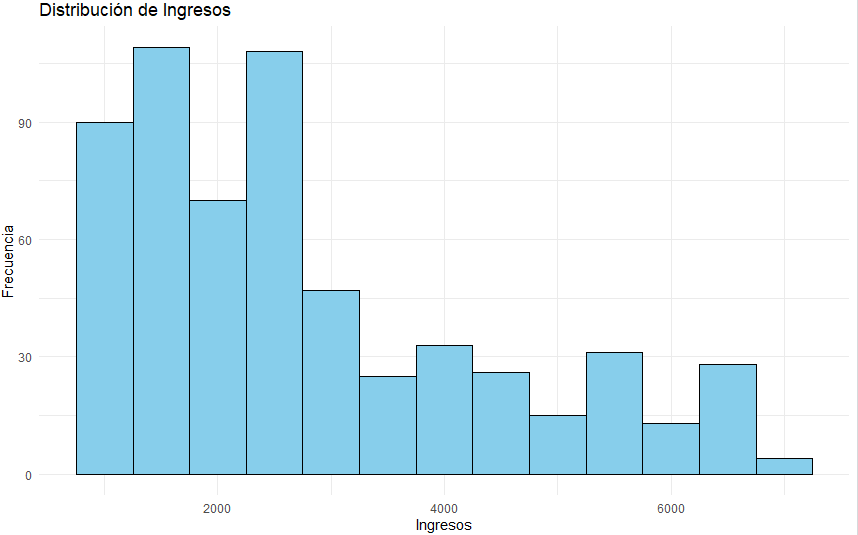
En la figura 2.10 el gráfico muestra un histograma de la distribución del puntaje de crédito. La mayoría de los puntajes están concentrados en valores bajos, con una alta frecuencia en el primer rango. A medida que el puntaje aumenta, la frecuencia disminuye rápidamente, indicando que pocos casos tienen puntajes de crédito elevados. La distribución es sesgada a la derecha.

**Figura 2.11***Histograma de distribución de años de empleo*

**

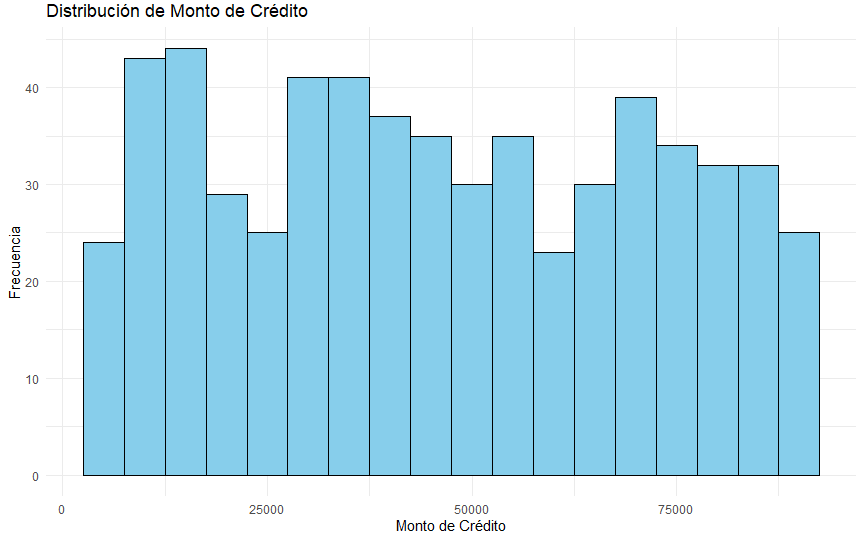
En la figura 2.11 el histograma muestra la distribución de años de empleo. La mayoría de los datos se concentran en los primeros años, con una alta frecuencia al inicio. A medida que los años de empleo aumentan, la frecuencia disminuye rápidamente. La distribución está sesgada a la derecha, indicando que pocos casos tienen muchos años de empleo.

**Figura 2.12***Histograma de distribución de ingresos*

**

En la figura 2.12 el histograma consiste en barras verticales que representan la frecuencia de ingresos en diferentes rangos. Se observa que las frecuencias más altas se encuentran en los rangos cercanos al valor de 2000 en el eje x, disminuyendo a medida que los ingresos aumentan. Esto sugiere una distribución donde la mayoría de los ingresos se concentran alrededor de un valor central, con menos individuos ganando cantidades significativamente más altas o más bajas.

**Figura 2.13***Histograma de distribución de monto de crédito*

**

En la figura 2.13 se puede observar que la imagen es un histograma que representa la distribución de montos de crédito, con el título "Distribución de Monto de Crédito". En el eje x se encuentra "Monto de Crédito" con valores de 0 a 75,000, y en el eje y está "Frecuencia" con valores de 0 a más de 40. Se observa una distribución desigual con picos alrededor de 25,000, 50,000 y 75,000, indicando diferentes montos de crédito con frecuencias variables. Las barras muestran las frecuencias de cada rango de montos de crédito, sin un modo claro, lo que sugiere una distribución variada de los montos de crédito otorgados.

## 2.2 Modelación

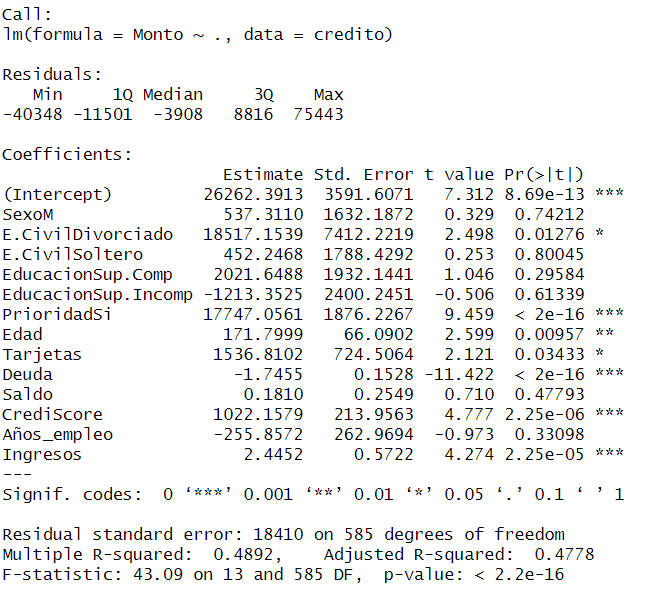
## 2.2.1 Modelo de Regresión Lineal

En este modelo de Regresión Lineal se toma la variable Monto cómo variable dependiente (Y) y las variables Sexo, E.Civil, Educacion, Prioridad, Edad, Tarjetas, Deuda, Saldo, CrediScore, Años\_empleo e Ingresos como variables independientes (X).

Para el uso de este modelo, se debe excluir la variable dependiente Credito (Y).

**Figura 2.2.1**

*Modelo de regresión lineal*



La variable Monto depende las variables SexoM, E.CivilDivorciado, E.CivilSoltero,

EducacionSup.Comp, EducacionSup.Incomp, PrioridadSi, Edad, Tarjetas, Deuda, Saldo, CrediScore, Años\_empleo e Ingresos de la siguiente manera:

Monto = 26262.391 + 537.311\*SexoM + 18517.154\*E.CivilDivorciado + 452.247\*E.CivilSoltero + 2021.649\*EducacionSup.Comp - 1213.352\*EducacionSup.Incomp + 17747.056\*PrioridadSi + 171.800\*Edad + 1536.810\*Tarjetas - 1.745\*Deuda + 0.181\*Saldo + 1022.158\*CrediScore - 255.857\*Años\_empleo + 2.445\*Ingresos

B1 = 537.311: Si el sexo es masculino, el monto de crédito aumenta en 537.311 soles.

B2 = 18517.154: Si el estado civil del cliente es divorciado, el monto de crédito aumenta en 18517.154 soles.

B3 = 452.247: Si el estado civil del cliente es soltero, el monto de crédito aumenta en 452.247 soles.

B4 = 2021.649: Si el grado de educación es de superior completo, el monto de crédito aumenta en 2021.649 soles.

B5 = 1213.352: Si el grado de educación es de superior incompleto, el monto de crédito disminuye en 1213.352 soles.

B6 = 17747.056: Si el cliente es prioritario, el monto de crédito aumenta en 17747.056 soles.

B7 = 171.800: Por cada año adicional en la edad del cliente, el monto de crédito aumenta en 171.800 soles.

B8 = 1536.810: Por cada tarjeta de crédito adicional que posea el cliente, el monto de crédito aumenta en 1536.810 soles.

B9 = 1.745: Por cada sol que tenga de deuda el cliente, el monto de crédito disminuye en 1.745 soles.

B10 = 0.181: Por cada sol de saldo que posea el cliente, el monto de crédito aumenta en 0.181 soles.

B11 = 1022.158: Por cada punto de crédito que tenga el cliente, el monto de crédito aumenta en 1022.158 soles.

B12 = 255.857: Por cada año de empleo adicional del cliente, el monto de crédito disminuye en 255.857 soles.

B13 = 2.445: Por cada sol de ingreso extra del cliente, el monto de crédito aumenta en 2.445 soles.

**Prueba Global**

Procedemos a evaluar si el modelo lineal múltiple es adecuado para predecir el monto de crédito. Usando α = 0.05.

H0: B1 = B2 = B3 = B4 = B5 = B6 = B7 = B8 = B9 = B10 = B11 = B12 = B13 = 0

H1: Al menos un Bi es diferente de 0.

F-statistic: 43.09 on 13 and 585 DF, p-value: < 2.2e-16

Conclusión: como p-value < α, entonces se rechaza H0. Por lo tanto, al menos un Bi es diferente a 0, a un nivel se significancia del 0.05. Esto quiere decir que el modelo es válido y adecuado para calcular el monto de crédito.

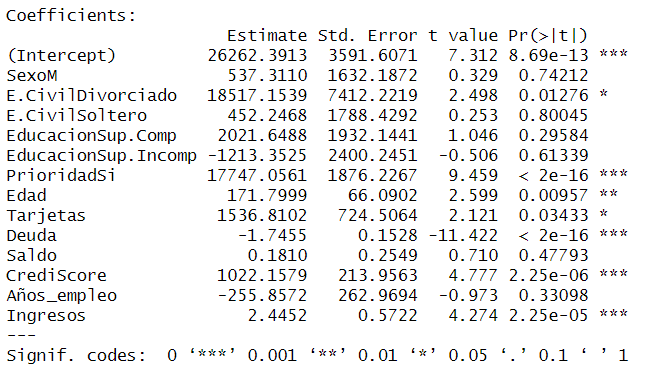
También podemos observar que Adjusted R-squared = 0.4778. Esto quiere decir que la influencia de las variables mencionadas es de un 47.78% sobre el monto de crédito.

**Selección de variables y modelo óptimo**

Ahora vamos a determinar las variables predictoras que son significativas en este modelo lineal múltiple. Con α = 0.05. Para esto nos fijamos en el valor del p-value de cada variable. Si, en cada variable, el p-value < α, entonces la variable es significativa para el modelo.

**Figura 2.2.2**

*Modelo de regresión lineal*

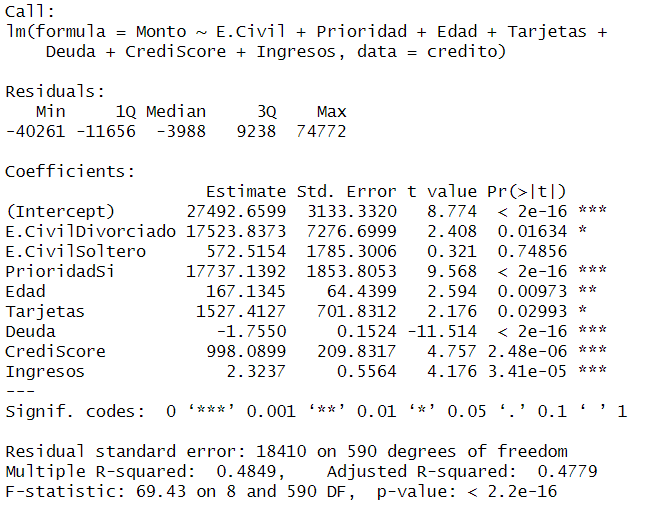


| B1: Sexo  H0: B1 = 0  H1: B1 ≠ 0  T = 0.329,  Pval = 0.74212  No se rechaza H0 | B2: E.CivilDivorciado  H0: B2 = 0  H1: B2 ≠ 0  T = 2.498,  Pval = 0.01276  Rechaza H0 | B3: E.CivilSoltero  H0: B3 = 0  H1: B3 ≠ 0  T = 0.253,  Pval = 0.80045  No se rechaza H0 | B4: EduacionSup.Comp  H0: B4 = 0  H1: B4 ≠ 0  T = 1.046,  Pval = 0.29584  No se rechaza H0 |
| --- | --- | --- | --- |
| B5: EduacacionSup.Incomp  H0: B5 = 0  H1: B5 ≠ 0  T = -0.506,  Pval = 0.61339  No se rechaza H0 | B6: PrioridadSi  H0: B6 = 0  H1: B6 ≠ 0  T = 9.459,  Pval = <2e-16  Rechaza H0 | B7: Edad  H0: B7 = 0  H1: B7 ≠ 0  T = 2.599,  Pval = 0.00957  Rechaza H0 | B8: Tarjetas  H0: B8 = 0  H1: B8 ≠ 0  T = 2.121,  Pval = 0.03433  Rechaza H0 |
| B9: Deuda  H0: B9 = 0  H1: B9 ≠ 0  T = -11.422,  Pval = <2e-16  Rechaza H0 | B10: Saldo  H0: B10 = 0  H1: B10 ≠ 0  T = 0.710,  Pval = 0.47793  No se rechaza H0 | B11: CrediScore  H0: B11 = 0  H1: B11 ≠ 0  T = 4.777,  Pval = 2.25e-06  Rechaza H0 | B12: Años\_empleo  H0: B12 = 0  H1: B12 ≠ 0  T = -0.973,  Pval = 0.33098  Rechaza H0 |
| B13: Ingresos  H0: B13 = 0  H1: B13 ≠ 0  T = 4.274,  Pval = 2.25e-05  Rechaza H0 |  |  |  |

Ahora vamos a estimar el modelo solo con las variables predictoras significativas.

**Figura 2.2.3**

*Modelo de regresión lineal ajustado*



Monto = 27492.6599 + 17523.8373\*E.CivilDivorciado + 17737.1392\*PrioridadSi + 167.1345\*Edad + 1527.4127\*Tarjetas - 1.755\*Deuda + 998.0899\*CrediScore + 2.3237\*Ingresos

En este modelo podemos observar algunos datos, los cuáles serán interpretados a continuación.

En primer lugar, se puede observar el coeficiente de determinación “Adjusted R-squared: 0.4779”. Esto significa que la variabilidad del monto de crédito es explicado por las variables en estudio en un 47.79% a través del modelo de regresión.

En segundo lugar, se puede observar que todos los p-value menores a 0.05 tienen un \*. Esto significa que son variables significativas y que influyen en el monto que se va a prestar por parte del banco.

En tercer lugar, observamos la columna “Estimate”. Esta columna nos indica los coeficientes de cada variable que conforma el modelo.

**Prueba Global**

Procedemos a evaluar si el nuevo modelo lineal múltiple es adecuado para predecir el monto de crédito. Usando α = 0.05.

H0: B1 = B2 = B3 = B4 = B5 = B6 = B7 = 0

H1: Al menos un Bi es diferente de 0.

F-statistic: 69.43 on 8 and 590 DF, p-value: < 2.2e-16

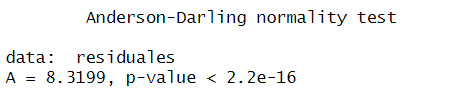
Conclusión: como p-value < α, entonces se rechaza H0. Por lo tanto, al menos un Bi es diferente a 0, a un nivel se significancia del 0.05. Esto quiere decir que el modelo es válido y adecuado para calcular el monto de crédito.

**Supuesto de normalidad de los errores**

Empezamos con el supuesto de normalidad de los errores. Para este supuesto se realiza la Prueba de Anderson-Darling, ya que hay más de 50 datos. Con un α = 0.05.

**Figura 2.2.4**

*Prueba de Anderson-Darling*

****

La hipótesis planteada es la siguiente:

H0: Los errores se distribuyen normalmente

H1: Los errores no se distribuyen normalmente

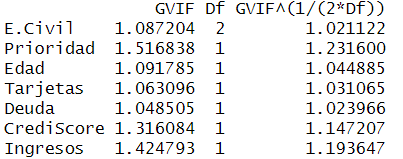
Conclusión: Como el p-value < alfa, entonces se rechaza H0. Es decir, los errores no siguen una distribución normal, a un nivel de significancia del 0.05.

**Supuesto de no multicolinealidad**

Continuamos con el supuesto de no multicolinealidad. Para este supuesto se utiliza el VIF Ajustado (los valores de los factores de inflación de varianza ). Con un α = 0.05.

**Figura 2.2.5**

*Supuesto de no multicolinealidad*

****

Hipótesis: Si VIF < 5, no existe multicolinealidad

Conclusión: Como los VIF < 5, entonces no existe multicolinealidad entre las variables regresoras.

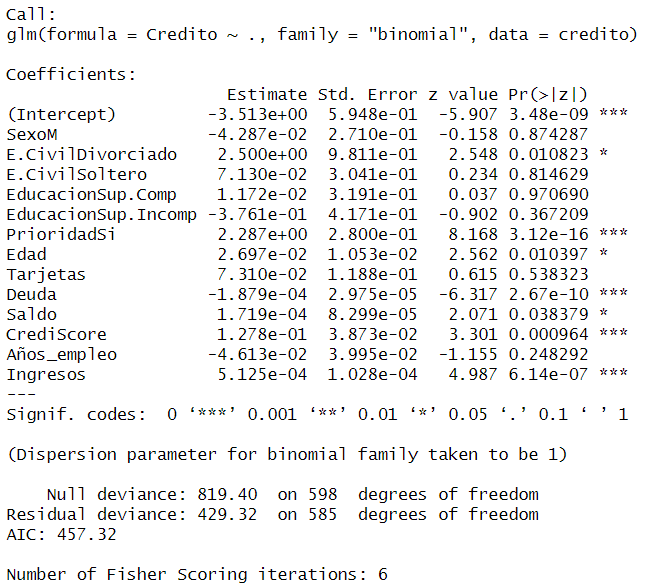
## 2.2.2 Modelo de Regresión Logística

En este modelo de Regresión Lineal se toma la variable Credito cómo variable dependiente (Y) y las variables Sexo, E.Civil, Educacion, Prioridad, Edad, Tarjetas, Deuda, Saldo, CrediScore, Años\_empleo e Ingresos como variables independientes (X).

Para el uso de este modelo, se debe excluir la variable dependiente Monto (Y).

**Figura 2.2.6**

*Modelo de regresión logística*

**

La variable Credito depende de las variables SexoM, E.CivilDivorciado, E.CivilSoltero, EducacionSup.Comp, EducacionSup.Incomp, PrioridadSi, Edad, Tarjetas, Deuda, Saldo, CrediScore, Años\_empleo e Ingresos de la siguiente manera:

logit(p) = -3.5134696 - 0.0428694\*SexoM + 2.5001011\*E.CivilDivorciado + 0.0712996\*E.CivilSoltero + 0.0117230\*EducacionSup.Comp - 0.3761023\*EducacionSup.Incomp + 2.2873763\*PrioridadSi + 0.0269723\*Edad + 0.0731001\*Tarjetas - 0.0001879\*Deuda + 0.0001719\*Saldo + 0.1278364\*CrediScore - 0.0461274\*Años\_empleo + 0.0005125\*Ingresos

**Prueba Global**

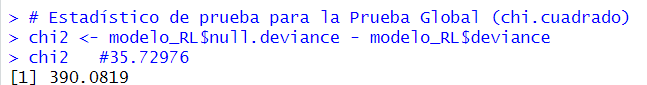
Procedemos a evaluar si el modelo de regresión logística es adecuado para predecir el monto de crédito. Usando α = 0.05.

H0: B1 = B2 = B3 = B4 = B5 = B6 = B7 = B8 = B9 = B10 = B11 = B12 = B13 = 0

H1: Al menos un Bi es diferente de 0.

**Figura 2.2.7**

*Estadístico chi2*



Se calcula restando el la desviación nula menos la desviación. Chi2 = 390.0819

**Figura 2.2.8**

*Grados de libertad*



**Figura 2.2.9**

*P-value*



El p-value se calcula de la manera mostrada en la imagen. En este caso el p-value es un valor muy cercano a 0.

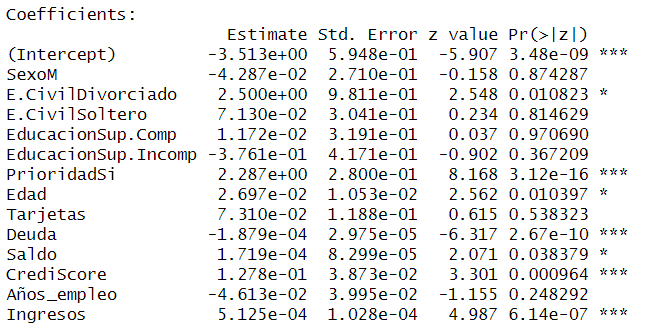
Conclusión: como p-value < α, entonces se rechaza H0. Por lo tanto, al menos un Bi es diferente a 0, a un nivel se significancia del 0.05. Esto quiere decir que el modelo es válido y adecuado para calcular el si es que se realiza el préstamo.

**Selección de variables y modelo óptimo**

Ahora vamos a determinar las variables predictoras que son significativas en este modelo de regresión logística. Con α = 0.05, nos fijamos en el valor de p-value de cada variable. Si en cada variable, el p-value < α, entonces la variable es significativa para el modelo.

**Figura 2.2.10**

*Modelo de regresión logística*

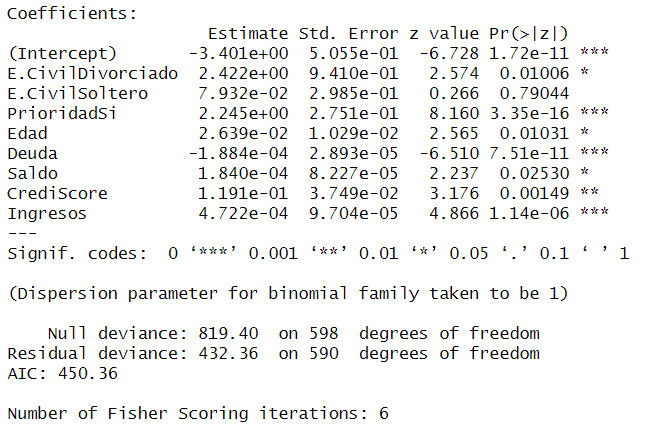
******

| B1: SexoM  H0: B1 = 0  H1: B1 ≠ 0  T = -0.158,  Pval = 0.874287  No se rechaza H0 | B2: E.CivilD  H0: B2 = 0  H1: B2 ≠ 0  T = 2.548,  Pval = 0.010823  Rechaza H0 | B3: E.CivilSol  H0: B3 = 0  H1: B3 ≠ 0  T = 0.234,  Pval = 0.814629  No se Rechaza H0 | B4: EducacionSup.comp  H0: B4 = 0  H1: B4 ≠ 0  T = 0.037 ,  Pval = 0.970690  No se rechaza H0 |
| --- | --- | --- | --- |
| B5: EducacionSup.incom  H0: B5 = 0  H1: B5 ≠ 0  T = -0.902,  Pval = 0.367209  No se rechaza H0 | B6: PrioridadSi  H0: B6 = 0  H1: B6 ≠ 0  T = 8.168  Pval = 3.12e-16  Rechaza H0 | B7: Edad  H0: B7 = 0  H1: B7 ≠ 0  T = 2.562,  Pval = 0.010397  Rechaza H0 | B8: Tarjetas  H0: B8 = 0  H1: B8 ≠ 0  T = 0.615  Pval = 0.538323  No se Rechaza H0 |
| B9: Deuda  H0: B9 = 0  H1: B9 ≠ 0  T = -6.317,  Pval = 2.67e-10  Rechaza H0 | B10: Saldo  H0: B10 = 0  H1: B10 ≠ 0  T = 2.071,  Pval = 0.038379  Rechaza H0 | B11: CrediScore  H0: B11 = 0  H1: B11 ≠ 0  T = 3.301  Pval = 0.000964  Rechaza H0 | B12: Años\_empleo  H0: B12 = 0  H1: B12 ≠ 0  T = -1.155,  Pval = 0.248292  No se rechaza H0 |
| B13: Ingresos  H0: B13 = 0  H1: B13 ≠ 0  T = 4.987,  Pval = 6.14e-07  Rechaza H0 |  |  |  |

Ahora vamos a estimar el modelo solo con las variables predictoras significativas.

**Figura 2.2.11**

*Modelo de Regresión Logística Ajustado*



Credito = - 3.4007459 + 2.4217069\*E.CivilDivorciado + 2.2447500\*PrioridadSi + 0.0263860\*Edad - 0.0001884\*Deuda + 0.0001840\*Saldo + 0.1190778\*CrediScore + 0.0004722\*Ingresos

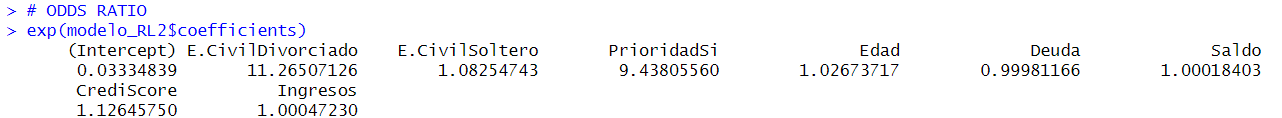
En este modelo podemos observar los datos que van a ser objeto de esta investigación.

**Odds Ratio**

A continuación vamos a realizar la interpretación de los coeficientes, para esto tenemos que sacarle el exponencial a la función mencionada para conseguir los odds ratio (OR).

**Figura 2.2.12**

*Odds ratio*



OR\_E.CivilDivorciado = 11.26507126: si el estado civil del cliente es divorciado, hay 11.26 más posibilidades de que se le dé el crédito al cliente.

OR\_E.CivilSoltero = 1.08254743: si el estado civil del cliente es soltero, hay 1.08 más de posibilidad de que se le dé el crédito al cliente.

OR\_PrioridadSi = 9.43805560: Si la prioridad es si, el cliente tiene 9.43 más probabilidades de que se le dé el crédito.

OR\_Edad = 1.02673717. Por cada año adicional del cliente, la posibilidad de que se le dé el crédito aumenta en 1.026 veces.

OR\_Deuda = 0.99981166. Por cada sol adicional de deuda que tenga el cliente, la posibilidad de que se le dé el crédito disminuye en menos de 1 (0.99).

OR\_Saldo = 1.00018403: Por cada sol adicional de saldo que tenga el cliente, hay 1.00018403 veces más de probabilidad de que se le dé el crédito.

OR\_CrediScore: 1.12645750: Por cada punto de crédito adicional que tenga el cliente, hay 1.12645750 veces más de probabilidad de que se le dé el crédito.

OR\_Ingresos:1.00047230: Por cada sol adicional de sueldo que tenga el cliente, hay 1.00047230 veces más de probabilidad de que se le dé el crédito.

**2.2.3 Predicciones**

**Modelo de regresión lineal múltiple**

**Predicción 1**

1. Estimar el monto de crédito para un cliente si su estado civil es divorciado, que tenga prioridad, con 45 años, 2 tarjetas, una deuda de 1500 soles, con 13 puntos de crédito e ingresos de 6670 soles.

**Figura 2.2.13**

*Predicción 1*





Cuando se cumplen con las características mencionadas anteriormente, el monto de crédito que le corresponde al cliente es de S/.99171.52.

**Predicción 2**

1. Estimar el monto de crédito para un cliente si su estado civil es soltero, que no tenga prioridad, con 20 años, 1 tarjeta, una deuda de 500 soles, con 10 puntos de crédito e ingresos de 3000 soles.

**Figura 2.2.14**

*Predicción 2*





Cuando se cumplen con las características mencionadas anteriormente, el monto de crédito que le corresponde al cliente es de S/.49009.89.

**Modelo de regresión logística**

**Predicción 3**

1. Estimar la probabilidad de recibir el crédito de un cliente casado, con prioridad, con 55 años de edad, un saldo de 1300 soles, una deuda de 8540 soles, con 5 puntos de crédito e ingresos de 4000 soles.

**Figura 2.2.15**

*Predicción 3*





Si el cliente cuenta con las características mencionadas anteriormente, entonces la probabilidad de obtener el crédito es de 80.37%.

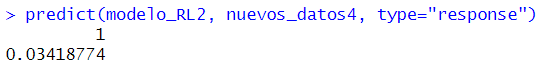
**Predicción 4**

1. Estimar la probabilidad de recibir el crédito de un cliente soltero, sin prioridad, con 25 años de edad, un saldo de 500 soles, una deuda de 11000 soles, con 2 puntos de crédito e ingresos de 2250 soles.

**Figura 2.2.16**

*Predicción 4*





Si el cliente cuenta con las características mencionadas anteriormente, entonces la probabilidad de obtener el crédito es de 3.42%.

# CONCLUSIONES Y RECOMENDACIONES

* De las 11 variables no dependientes proporcionadas para la construcción de los modelos predictivos, solo 7 de estas resultaron ser significativas para poder predecir de manera adecuada y confiable el comportamiento de las concesiones de crédito, así como el monto de dichas concesiones. Por esto, se recomienda el futuro descarte de las 4 variables restantes de las bases de datos proporcionadas y utilizadas para la creación, elección e interpretación de modelos que busquen analizar el estado de las concesiones y montos.
* Se logró desarrollar un robusto modelo de regresión para los casos dados, logrando aplicar diferentes algoritmos de regresión y técnicas de regularización para garantizar una correcta evaluación de la información recibida. Además, la relativa simplicidad del modelo permite su adaptación a diferentes aspectos dentro y fuera del área financiera, por esto se recomienda su adecuada modificación y prueba en otros campos.
* El análisis de la contribución e influencia de las variables relevantes en el modelo regresivo permitió una alta precisión en las predicciones realizadas. De similar manera, el análisis exploratorio de las variables permitió observar de manera simple y accesible su distribución en la base de datos dada y su posible influencia en el riesgo crediticio.
* Se utilizaron de manera exitosa métricas de evaluación, como la precisión y especificidad, para evaluar el desempeño mostrado por los múltiples modelos regresivos desarrollados y su eficacia al momento de predecir las variables requeridas, asimismo se pudo comparar este desempeño y refinar el modelo hasta encontrar uno que logró cumplir con los requerimientos impuestos tras un ajuste. Se recomienda someter el modelo a otros requerimientos como la predicción de otras variables distintas al crédito o monto.

# ANEXOS

**Código Rstudio para los gráficos**

library(readxl)

pie(table(data\_bancaria$Sexo),

main="Distribución por Sexo",

col=c("red", "skyblue"),

labels = paste(names(table(data\_bancaria$Sexo)),

round(100 \* prop.table(table(data\_bancaria$Sexo)), 1), "%"))

barplot(table(data\_bancaria$E.Civil), main="Distribución por Estado Civil", xlab="Estado Civil", ylab="Frecuencia", col="skyblue",)

barplot(table(data\_bancaria$Educacion), main="Distribución por Nivel de Educación", xlab="Educación", ylab="Frecuencia", col="skyblue")

priority\_labels <- paste(names(table(data\_bancaria$Prioridad)),

round(100 \* prop.table(table(data\_bancaria$Prioridad)), 1), "%")

barplot(table(data\_bancaria$Credito), main="Distribución por Concesión de Crédito", xlab="Crédito", ylab="Frecuencia", col="skyblue")

install.packages("ggplot2")

library(ggplot2)

ggplot(data\_bancaria, aes(x=Edad)) +

geom\_histogram(binwidth=2, fill="skyblue", color="black") +

labs(title="Distribución de Edad", x="Edad", y="Frecuencia") +

theme\_minimal()

boxplot(data\_bancaria$Tarjetas,

main="Distribución de Cantidad de Tarjetas",

xlab="Cantidad de Tarjetas",

ylab="Frecuencia",

col="skyblue")

boxplot(data\_bancaria$Saldo,

main="Distribución de Saldo",

xlab="Saldo",

ylab="Frecuencia",

col="skyblue")

ggplot(data\_bancaria, aes(x=Saldo)) +

geom\_histogram(binwidth=500, fill="skyblue", color="black") +

labs(title="Distribución de Saldo", x="Saldo", y="Frecuencia") +

theme\_minimal()

ggplot(data\_bancaria, aes(x=CrediScore)) +

geom\_histogram(binwidth=1, fill="skyblue", color="black") +

labs(title="Distribución de Puntaje de Crédito", x="Puntaje de Crédito", y="Frecuencia") +

theme\_minimal()

ggplot(data\_bancaria, aes(x=Años\_empleo)) +

geom\_histogram(binwidth=0.5, fill="skyblue", color="black") +

labs(title="Distribución de Años de Empleo", x="Años de Empleo", y="Frecuencia") +

theme\_minimal()

ggplot(data\_bancaria, aes(x=Ingresos)) +

geom\_histogram(binwidth=500, fill="skyblue", color="black") +

labs(title="Distribución de Ingresos", x="Ingresos", y="Frecuencia") +

theme\_minimal()

boxplot(data\_bancaria$Monto,

main="Distribución de Monto de Crédito",

xlab="Monto de Crédito",

ylab="Frecuencia",

col="skyblue")

**Código utilizado en R para cada regresión**

[Estadística Aplicada - Archivo en RStudio](https://drive.google.com/drive/folders/152W3gacmFupBvLufIt0xQuy3a_tgYnmD?usp=sharing)