Ciencia Latina
Internacional

Análisis Predictivo En La Banca:
Regresión Logística Para La Adquisición De Tarjetas De Crédito

CID editorial
ALIANZA RED
Red de Redes
INDEXA
Crossref

Mayo 2025 – CID - Centro de Investigación y Desarrollo
Copyright © CID - Centro de Investigación y Desarrollo
Copyright del texto © 2025 de Autores
libros.ciencialatina.org
editorial@ciencialatina.org
Atención por WhatsApp al +52 22 26903834

Datos Técnicos de Publicación Internacional
Título: "Análisis predictivo en la Banca Regresión logística para la adquisición de tarjetas de crédito"
Autores: Solange Lesly Basualdo Najera, Richard Fernando Fernández Vásquez
Editor: CID - Centro de Investigación y Desarrollo
Diseño de tapa: CID - Centro de Investigación y Desarrollo
Corrección de Estilo: CID - Centro de Investigación y Desarrollo
Formato: PDF
Páginas: 43 pág.
Tamaño: Sobre C5 162 x 229 mm
Requisitos de sistema: Adobe Acrobat Reader
Modo de acceso: World Wide Web
ISBN: 978-1-967344-80-2
DOI: https://doi.org/10.37811/cli_w1207

1ª. Edición. Año 2025. Editorial CID - Centro de Investigación y Desarrollo.

El contenido del libro y sus datos en su forma, corrección y fiabilidad son responsabilidad exclusiva de los autores. Permite la descarga de la obra y compartir siempre que los créditos se atribuyan a los autores, pero sin la posibilidad de cambiarlo de cualquier forma o utilizarlo con fines comerciales

Prohibida su reproducción por cualquier medio.

Distribución gratuita

Dedicamos este trabajo a todos

los apasionados de la

Ciencia de Datos

ÍNDICE

PRÓLOGO **7**
CAPITULO I **8**
Conceptos clave 8
1.1 La tarjeta de crédito en una Entidad Bancaria 8
1.2 Tipos de tarjeta de crédito ofrecida 8
1.3 El proceso de captación de tarjeta de crédito 9
1.4 Resolución S.B.S. N° 11356 - 2008 9
CAPITULO II **11**
Pruebas estadísticas 11
2.1 Prueba de Levene para igualdad de varianzas 11
2.2 Prueba T para muestras independientes 11
2.3 Prueba Chi-Cuadrado de Pearson 11
CAPITULO III **13**
Modelo de regresión logística 13
3.1 Definición 13
3.2 Modelo dicotómico 13
3.3 Hipótesis de Regresión Logística 14
3.4 Razón de apuestas (Odd ratio) 15
3.5 Estimación 15
3.6 Contrastes de significación 19
CAPITULO IV **20**
Medidas de bondad de ajuste 20
4.1 Coeficiente Pseudo R^2 20
4.2 Contraste de bondad de ajuste de Hosmer y Lemeshow 20
4.3 Curva ROC 21
4.4 Tablas de clasificación 21
CAPITULO V **23**
Caso de aplicación para la adquisición de tarjetas de crédito 23
5.1 Población 23
5.2 Muestra 23
5.3 Unidad de análisis 23
5.4 Fuentes de información 23
5.5 Definición de variables 23
5.6 Muestra de entrenamiento y validación 26
5.7 Análisis de significancia entre la variable dependiente e independientes 26
5.8 Análisis de regresión logística 30
5.9 Análisis de bondad de ajuste 34
CAPITULO VI **38**
Conclusiones 38
REFERENCIAS BIBLIOGRAFICAS **40**

ÍNDICE DE TABLAS

Tabla 1. Prueba de medias para la variable share of wallet deudas retail 27
Tabla 2. Prueba de medias para la variable saturación total de tarjeta de crédito 28
Tabla 3. Prueba de medias para la variable comparación entre líneas de tarjeta de crédito 29
Tabla 4. Prueba Chi-Cuadrado de Pearson para variables independientes cualitativas 29
Tabla 5. Modelo de Regresión Logística 30
Tabla 6. Prueba Ómnibus 34
Tabla 7. Resumen del modelo 34
Tabla 8. Prueba de Hosmer y Lemeshow 34
Tabla 9. Valor del área bajo la curva ROC 35
Tabla 10. Tabla de clasificación para la muestra de entrenamiento 36
Tabla 11. Tabla de clasificación para la muestra de validación 36
Tabla 12. Clasificación de clientes según probabilidad de adquisición para la muestra de entrenamiento 37
Tabla 13. Clasificación de clientes según probabilidad de adquisición para la muestra de validación 37

ÍNDICE DE FIGURA

Figura. 1 Curva ROC 35

ANÁLISIS PREDICTIVO EN LA BANCA: REGRESIÓN LOGÍSTICA PARA LA ADQUISICIÓN DE TARJETAS DE CRÉDITO

- **Título:** Análisis predictivo en la Banca: Regresión logística para la adquisición de tarjetas de crédito
- **Autor:** Solange Lesly Basualdo Najera, Richard Fernando Fernández Vásquez
- **Ciudad:** Lima - Perú
- **Fecha:** Abril de 2025

Mag. Solange Lesly Basualdo Najera[1]

Dr. Richard Fernando Fernández Vásquez[2]

[1] Universidad Nacional de Ingeniería, Docente Universitario, correo electrónico: lbasualdo@uni.edu.pe. ORCID: https://orcid.org/0009-0001-0567-689X

[2] Universidad Nacional de Ingeniería, Docente Universitario, correo electrónico: rfernandezv@uni.edu.pe. ORCID: https://orcid.org/0000-0003-1721-8527

PRÓLOGO

En un entorno bancario cada vez más competitivo, comprender el comportamiento de los clientes se ha vuelto esencial para el diseño de estrategias comerciales efectivas. Las entidades financieras no solo deben captar nuevos usuarios, sino también optimizar sus esfuerzos al dirigir sus productos a quienes realmente muestran mayor disposición a adoptarlos.

Este trabajo nace precisamente de esa necesidad: identificar, con base en evidencia estadística, a los clientes bancarizados con mayor propensión a adquirir una tarjeta de crédito. Para ello, se utilizó un modelo de regresión logística, una herramienta potente en la estadística aplicada, capaz de clasificar y predecir comportamientos a partir de variables significativas.

El caso de aplicación considera el análisis de factores como el nivel educacional, el número de tarjetas utilizadas, la línea de crédito ofrecida, entre otros, y permite identificar con claridad patrones que no siempre son visibles a simple vista. Uno de los grandes aportes de este caso de aplicación es, justamente, poner en valor el uso de datos internos de la banca para afinar procesos comerciales, generar eficiencia y, sobre todo, tomar decisiones más informadas.

Con una tasa de clasificación global del 84,1%, el modelo estadístico de regresión logística no solo demuestra una alta capacidad predictiva, sino que también resalta el potencial de la analítica avanzada como aliada estratégica para la industria financiera. Este libro está dirigido a quienes buscan entender cómo la estadística aplicada puede contribuir de forma concreta en la toma de decisiones dentro del sector bancario, desde ejecutivos hasta analistas, pasando por estudiantes y profesionales interesados en el cruce entre economía, datos y comportamiento del consumidor.

Más allá de los números, esta investigación busca también inspirar nuevas preguntas y líneas de estudio. Porque en la banca del futuro, los datos no serán una herramienta más, serán el centro de toda estrategia.

CAPITULO I

Conceptos clave

1.1 La tarjeta de crédito en una Entidad Bancaria

a) Tarjeta de crédito

Mediante el contrato de tarjeta de crédito una Entidad Bancaria concede una línea de crédito al titular por un plazo determinado y otorga la correspondiente tarjeta, con la finalidad de que el usuario de dicha tarjeta adquiera bienes o servicios en los establecimientos afiliados que los proveen, pague obligaciones o, de así permitirlo la empresa emisora y no mediar renuncia expresa por parte del titular, hacer uso del servicio de disposición de efectivo u otros servicios conexos, dentro de los límites y condiciones pactados, obligándose a su vez, a pagar a la empresa que expide la correspondiente tarjeta, el importe de los bienes y servicios que haya adquirido, obligaciones pagadas, y demás cargos, conforme a lo establecido en el respectivo contrato. (Monografias.com, s.f.)

b) Cuenta de tarjeta de crédito

Es un archivo magnético donde se almacenan la línea de crédito de tarjeta asignada al cliente, saldos de la cuenta, indicadores de control de la cuenta e información demográfica del cliente. (Monografias.com, s.f.)

c) Línea de tarjeta de crédito

Es la línea que se le otorga al cliente en la tarjeta de crédito para que pueda ser utilizada según disponga el cliente y se renueva según el cliente pague sus consumos. (Monografias.com, s.f.)

1.2 Tipos de tarjeta de crédito ofrecida

Las tarjetas de crédito en una Entidad Bancaria ofrecidas a sus clientes potenciales son tarjeta de crédito clásica y tarjeta de crédito visa.

El tipo de tarjeta de crédito que se les otorgará a los clientes dependerá de los resultados de la evaluación crediticia realizada por el funcionario de tarjetas de crédito y la gerencia de riesgos en la Entidad Bancaria. Universidad Dr. José Matías Delgado. (s.f.)

1.3 El proceso de captación de tarjeta de crédito

El proceso abarca desde el primer contacto con el potencial cliente y la solicitud de requisitos hasta el envío del expediente del cliente al área de Back Office para su respectiva evaluación, luego se realizan los siguientes pasos:

- Evaluación crediticia del cliente

- Gestión de desembolsos

- Gestión de garantías y seguros

- Gestión de entregas de tarjetas de crédito

La persona debe cumplir con requisitos mínimos para calificar a una tarjeta de crédito:

- Edad mínima 22 años

- Ingreso mínimo 750 nuevos soles.

- Persona bancarizada: Calificación sbs normal en doce meses de historia.

- Sin información negativa en las centrales de riesgos

- Antigüedad laboral mínima de 3 años

- Línea ofrecida mínima de 300 nuevos soles

- Línea ofrecida máxima de 40000 nuevos soles

- Residencia particular mínima de un año

1.4 Resolución S.B.S. N° 11356 - 2008

Definiciones

a) Créditos: se refiere a la suma de los créditos directos más indirectos.

b) Créditos directos: representa los financiamientos que, bajo cualquier modalidad, las empresas del sistema financiero otorguen a sus clientes, originando a cargo de éstos la obligación de entregar una suma de dinero determinada, en uno o varios actos.

c) Créditos indirectos: representan los avales, las cartas fianza, las aceptaciones bancarias, las cartas de crédito, los créditos aprobados no desembolsados y las líneas de crédito no utilizadas, otorgados por las empresas del sistema financiero.

d) Créditos revolventes: son aquellos créditos en los que se permite que el saldo fluctúe en función de las decisiones del deudor, por ejemplo, las tarjetas de crédito. Asimismo, se consideran dentro de este tipo de crédito los productos que permiten reutilizaciones parciales, es decir, que tienen un componente revolvente y otro no revolvente.

e) Créditos no revolventes: son aquellos créditos reembolsables por cuotas, siempre que los montos pagados no puedan ser reutilizables por el deudor. En este tipo de crédito no se permite que los saldos pendientes fluctúen en función de las propias decisiones del deudor.

f) Deudor minorista: persona natural o jurídica que cuenta con créditos directos o indirectos clasificados como de consumo (revolventes y no revolventes), a microempresas, a pequeñas empresas o hipotecarios para vivienda.

(Superintendencia de Banca, Seguros y AFP, 2008)

Tipos de crédito

- Créditos de consumo revolvente: son aquellos créditos revolventes otorgados a personas naturales, con la finalidad de atender el pago de bienes, servicios o gastos no relacionados con la actividad empresarial.

- Créditos de consumo no revolvente: son aquellos créditos no revolventes otorgados a personas naturales, con la finalidad de atender el pago de bienes, servicios o gastos no relacionados con la actividad empresarial.

(Superintendencia de Banca, Seguros y AFP, 2008)

Clasificación del deudor

La clasificación del deudor está determinada principalmente por su capacidad de pago, definida por el flujo de fondos y el grado del cumplimiento de sus obligaciones. Si un deudor es responsable de varios tipos de créditos con una misma entidad, la calificación de basará en la categoría del crédito con mayor riesgo. Las categorías con las que se reportan a los clientes son: Normal (0), Con Problemas Potenciales – CPP (1), Deficiente (2), Dudoso (3) y Pérdida (4). (Superintendencia de Banca, Seguros y AFP, 2008)

CAPITULO II

Pruebas estadísticas

2.1 Prueba de Levene para igualdad de varianzas

La prueba de Levene, es un contraste sobre homogeneidad o igualdad de varianzas, el resultado de este contraste nos permite decidir si podemos o no suponer que las varianzas poblacionales son iguales: si la probabilidad asociada al estadístico de Levene es mayor que 0.05, podemos suponer que las varianzas poblaciones son iguales; si la probabilidad asociada del estadístico de Levene es menor que 0.05, rechazaremos la hipótesis de igualdad de varianzas y supondremos que son distintas. (Basualdo, 2014)

$H_o: \sigma_1^2 = \sigma_2^2 = \ldots = \sigma_k^2$

$H_a: \sigma_i^2 = \sigma_j^2$ para algún $i \neq j$

2.2 Prueba T para muestras independientes

La prueba T para muestras independientes utiliza una distribución t de student, permite contrastar la hipótesis de igualdad de medias, es decir la tipificación de la diferencia entre las dos medias muestrales que se obtiene restando a esa diferencia su valor esperado en la población y dividiendo el resultado por el error típico de la diferencia. (Basualdo, 2014)

2.3 Prueba Chi-Cuadrado de Pearson

La prueba Chi-Cuadrado de Pearson, mide la discrepancia entre una distribución observada y otra teórica (bondad de ajuste), indicando en qué medida las diferencias existentes entre ambas, de haberlas, se deben al azar en el contraste de hipótesis. También contrasta la hipótesis de que las variables categóricas son independientes, frente a la hipótesis alternativa de que una variable se distribuye de modo diferente para diversos niveles de la otra, mediante la presentación de los datos en tablas de contingencia. (Basualdo, 2014)

H_o: Las dos variables son independientes

H_a: Las dos variables no son independientes

La fórmula del estadístico es la siguiente:

$$\chi^2 = \sum_i \frac{(observada_i - teorica_i)^2}{teorica_i}$$

Cuanto mayor sea el valor de χ^2, menos verosímil es que la hipótesis sea correcta. De la misma forma, cuanto más se aproxima a cero el valor de Chi-Cuadrado, más ajustadas están ambas distribuciones. Los grados de libertad (gl) vienen dados por:

gl= (r-1)(k-1). Donde r es el número de filas y k el de columnas.

No se rechaza H_0 cuando $\chi^2 < \chi_t^2(r-1)(k-1)$. En caso contrario sí se rechaza.

"t" representa el valor proporcionado por las tablas, según el nivel de significación estadística elegido.

CAPITULO III

Modelo de regresión logística

3.1 Definición

La regresión logística es un modelo de elección discreta en el que la variable dependiente es cualitativa. Este requiere realizar menos supuestos, lo que permite obtener resultados más robustos y es flexible en cuanto a la naturaleza de las variables explicativas, pues éstas pueden ser de escala y categóricas. Permite estudiar el impacto que tiene cada una de las variables independientes en la probabilidad de que ocurra el suceso de estudio. (Gujarati, 2004)

3.2 Modelo dicotómico

Para este modelo se considera que la variable respuesta, es una variable dicotómica que tomas dos valores.

Para estos modelos dicotómicos, las dos categorías deben de ser mutuamente excluyentes. La variable respuesta se puede expresar de la siguiente forma:

$$Y_i = \begin{cases} 1, \Pr ob(Y_i = 1) = P_i \\ 0, \Pr ob(Y_i = 0) = 1 - P_i \end{cases}$$

Tenemos que: $E(Y_i) = 1xP_i + 0x(1 - P_i) = P_i$

Se puede apreciar que la media teórica es igual a que Y_i tome el valor de 1.

Se va a considerar que Y_i es explicado por las variables independientes X_{2i}, X_{3i}, …, X_{ki}, designamos Z_i como:

$$Z_i = \beta_1 + \beta_2 X_{2i} + \ldots + \beta_k X_{ki} = \begin{bmatrix} 1 & X_{2i} & \ldots & X_{ki} \end{bmatrix} \begin{bmatrix} \beta_1 \\ \beta_2 \\ \vdots \\ \beta_k \end{bmatrix}$$

Entonces la esperanza de Y_i dado las variables independientes, es:

$$E(Y_i / X_{2i}, \cdots, X_{ki}) = F(\beta_1 + \beta_2 X_{2i} + \cdots + \beta_k X_{ki}) = F(Z_i)$$

El modelo se expresa como:

$$Y_i = E(Y_i / X_{2i}, \cdots, X_{ki}) + U_i = F(Z_i) + U_i$$

U_i es una perturbación aleatoria

Según lo cual, la función F puede generar distintos modelos dicotómicos, como el modelo de probabilidad lineal, el modelo probit y el modelo logit. Para el caso, se usó el modelo logit.

Con el objeto de que las predicciones del modelo dicotómico sean coherentes con la teoría, se debe cumplir que:

$$\lim_{Z_i \to +00} \Pr ob(Y_i = 1) = 1$$

$$\lim_{Z_i \to -00} \Pr ob(Y_i = 0) = 0$$

Esto se cumple si la función F es una función de distribución aleatoria. Para nuestro caso tomamos como función F a la función logística, obteniéndose el modelo logit:

$$P_i = E(Y_i / X_{2i}, \cdots, X_{ki}) + U_i = F(Z_i) = \frac{1}{1 + e^{-Z_i}} = \Lambda(Z_i)$$

(Menard, 2002)

3.3 Hipótesis de Regresión Logística

En los modelos de regresión logística, se asume que las perturbaciones son homocedásticas y no autocorrelacionadas, no se adopta el supuesto de linealidad entre la variable dependiente y las variables independientes, ya que su relación es de naturaleza no lineal.

Las hipótesis iniciales del modelo de regresión logística son las siguientes:

- Las observaciones $Y_1, \dots, Y_n$ son independientes.

- Cada Y_i tiene una distribución Bernoulli

- La probabilidad de que Y_i sea igual a 1 (probabilidad de que el individuo pertenezca al grupo A) depende de los valores de $X_1, \dots, X_k$ a través del siguiente modelo:

$$P_i = \frac{1}{1 + e^{-Z_i}}$$

(Basualdo, 2014)

3.4 Razón de apuestas (Odd ratio)

La razón de apuestas es una relación entre dos probabilidades, es la razón entre la probabilidad de que se produzca un suceso y la probabilidad de que no se produzca ese suceso.

Sabemos que la probabilidad de que el suceso no tome el valor de 1 viene dada por:

$$1 - P_i = 1 - \frac{1}{1 + e^{-Z_i}} = \frac{e^{-Z_i}}{1 + e^{-Z_i}}$$

Adicionalmente, sabemos que la probabilidad de que el suceso tome el valor de 1 viene dada por:

$$P_i = \frac{1}{1 + e^{-Z_i}}$$

Por lo tanto, la razón de apuestas para Y_i sería:

$$\frac{P_i}{1 - P_i} = e^{Z_i}$$

Para nuestro caso, se utiliza el logaritmo neperiano:

$$\ln\left(\frac{P_i}{1 - P_i}\right) = Z_i = X_i'\beta = \beta_1 + \sum_{j=2}^{k} \beta_j X_j$$

En conclusión, se puede decir que mientras la probabilidad Pi es una función no lineal de los parámetros β, la razón de apuestas si es una función lineal de los parámetros β. (Langrand, 2009).

3.5 Estimación

El método de estimación que se suele utilizar para estimar sus parámetros, es el de máxima verosimilitud, mediante procedimientos iterativos.

Supondremos una muestra aleatoria de datos (x_i, y_i), i=1,…, n, la función de probabilidades para una respuesta y_i cualquiera es:

$$P(y_i) = p_i^{y_i}(1 - p_i)^{1 - y_i} \; y_i = 0, 1$$

Y para la muestra:

$$P(y_1, \dots, y_n) = \prod_{i=1}^{n} p_i^{y_i} (1 - p_i)^{1-y_i}$$

Tomando logaritmos:

$logP(y) = \sum_{i=1}^{n} y_i \log\left(\frac{p_i}{1-p_i}\right) + \sum log\,(1 - p_i)$ (a)

La función soporte (de verosimilitud en logaritmos) puede escribirse como

$logP(\beta) = \sum_{i=1}^{n} (y_i \log p_i + (1 - y_i) \log\,(1 - p_i))$ (b)

Donde $\beta' = (\beta_0, \beta_1, \dots, \beta_p)$ es un vector de p+1 componentes, incluyendo la constante β_0 que determina las probabilidades p_i. Maximizar la verosimilitud puede expresarse como minimizar una función que mide la desviación entre los datos y el modelo. Entonces la desviación del modelo será:

$D(\beta) = -2 \sum_{i=1}^{n} (y_i \log p_i + (1 - y_i) \log\,(1 - p_i))$ (c)

Y hablaremos indistintamente de maximizar el soporte o minimizar la desviación del modelo. Se define la desviación (deviance) de cada dato por:

$d_i = -2(y_i \log\, p_i + (1 - y_i) \log\,(1 - p_i))$ (d)

Y mide el ajuste del modelo al dato (y_i, x_i). En efecto, observemos, en primer lugar, que como los $\widehat{p_i}$ son menores que uno, sus logaritmos son negativos, por lo que la desviación es siempre positiva. Además, en el cálculo de la desviación solo interviene uno de sus dos términos, ya que y_i solo puede valer cero o uno. Entonces:

- Si y_i=1 y la observación pertenece a la segunda población, el segundo término de la desviación es nulo y $d_i = -2logp_i$. La observación tendrá una desviación grande si la probabilidad estimada de pertenecer a la segunda población p_i, es pequeña, es decir, cuando esta observación está mal explicada por el modelo.

- Si y_i=0, y la observación pertenece a la primera población, solo interviene el segundo término de la desviación $d_i = -2log(1 - p_i)$. La desviación será grande si p_i es grande, lo que indica la probabilidad de pertenecer a la verdadera población es pequeña y el modelo ajusta mal dicho dato.

Para maximizar la verosimilitud, expresando p_i en función de los parámetros de interés, β, obtenemos la función soporte:

$$L(\beta) = \sum_{i=1}^{n} y_i \, x_i'\beta - \sum_{i=1}^{n} log\,(1 + e^{x_i'\beta}) \qquad \text{(e)}$$

Que derivaremos para obtener los estimadores MV. Escribiendo los resultados como vector columna:

$$\frac{\partial L(\beta)}{\partial \beta} = \sum_{i=1}^{r} y_i \, x_i - \sum_{i=1}^{n} x_i \left(\frac{e^{x_i'\beta}}{1+e^{x_i'\beta}}\right) \qquad \text{(f)}$$

E igualando este vector a cero y llamando $\hat{\beta}$ a los parámetros que satisfacen el sistema de ecuaciones:

$$\sum_{i=1}^{n} y_i \, x_i = \sum_{i=1}^{n} x_i \left(\frac{1}{1+e^{-x_i'\hat{\beta}}}\right) = \sum_{i=1}^{n} \hat{y}_i \, x_i \qquad \text{(g)}$$

Estas ecuaciones establecen que los productos de los valores observados por las variables explicativas deben ser igual al de los valores previstos. También, que los residuos del modelo,$e_i = y_i - \hat{y}_i$, deben ser ortogonales a las variables x. Esta condición es análoga a la obtenida en el modelo de regresión estándar, pero ahora el sistema resultante es no lineal en los parámetros $\hat{\beta}$. Para obtener el valor $\hat{\beta}_{MV}$ que maximiza la verosimilitud, acudiremos a un algoritmo tipo Newton-Raphson.

Desarrollando el vector $\left(\partial L(\beta) \big/ \partial \beta\right)$ alrededor de un punto β_a, se tiene:

$$\frac{\partial L(\beta)}{\partial \beta} = \frac{\partial L(\beta_a)}{\partial \beta} + \frac{\partial^2 L(\beta_a)}{\partial \beta \partial \beta'}(\beta - \beta_a):$$

Para que en el punto β_a corresponda el máximo de la verosimilitud su primera derivada debe anularse. Imponiendo la condición $\left(\partial L(\beta_a) \big/ \partial \beta\right) = 0$, se obtiene:

$$\beta_a = \hat{\beta} + \left(-\frac{\partial^2 L(\beta_a)}{\partial \beta \partial \beta'}\right)^{-1} \left(\frac{\partial L(\beta)}{\partial \beta}\right) \qquad \text{(h)}$$

Que expresa como obtener el punto máximo β_a, a partir de un punto próximo cualquiera β. La ecuación depende de la matriz de segundas derivadas, que, en el óptimo, es la inversa de la matriz de varianzas y covarianzas asintótica de los estimadores MV. Para obtener su expresión, derivando por segunda vez se obtiene:

$$\widehat{M}^{-1} = \left(-\frac{\partial^2 L(\beta)}{\partial \beta \partial \beta'}\right) = \sum_{i=1}^{n} x_i \, x'_i w_i \qquad \text{(i)}$$

Donde los coeficientes w_i estan dados por:

$$w_i = \frac{e^{x_i'\beta}}{(1+e^{x_i'\beta})^2} = p_i(1-p_i) \qquad \text{(j)}$$

Sustituyendo en (h) las expresiones (i) y (f) y evaluando las derivadas en un estimador inicial $\hat{\beta}$, se obtiene el siguiente método para obtener un nuevo valor del estimador, β_a, a partir del $\hat{\beta}$:

$$\beta_a = \hat{\beta} + (\sum_{i=1}^{n} x_i\, x_i'\hat{w}_i)^{-1}(\sum x_i(y_i - \hat{p}_i)) \qquad \text{(k)}$$

Donde $\hat{p}_i$ y $\hat{w}_i$ se calculan con el valor $\hat{\beta}$. El algoritmo puede escribirse como:

$$\beta_a = \hat{\beta} + \left(X'\widehat{W}X\right)^{-1}X'(Y - \hat{Y}) \qquad \text{(l)}$$

Donde $\widehat{W}$ es una matriz diagonal con términos $\hat{p}_i(1-\hat{p}_i)$ y $\hat{Y}$ es el vector de valores esperados de Y. La matriz de varianzas y covarianzas de los estimadores asi obtenidos es aproximadamente, según (l), $\left(X'\widehat{W}X\right)^{-1}$. Observemos que la ecuación (l) indica que debemos modificar el estimador si los residuos nos son ortogonales a las variables explicativas, es decir, si $X'(Y-\hat{Y}) \neq 0$. La modificación del estimador depende de esta diferencia y se reparte entre los componentes de $\hat{\beta}$ en función de su matriz de varianzas y covarianzas estimada.

La forma habitual de implementar este método es el siguiente algoritmo iterativo que proporciona, en convergencia, el estimador MV de β:

1. Fijar un valor arbitrario inicial, $\hat{\beta}_1$, para los parámetros y obtener el vector $\hat{Y}_1$ para dicho valor en el modelo logit.

2. Definir una variable auxiliar z_i de residuos estandarizados por:

$$z_i = \frac{y_i - \hat{y}_i}{\sqrt{\hat{y}_i(1-\hat{y}_i)}} = \frac{y_i - \hat{p}_i}{\sqrt{\hat{p}_i(1-\hat{p}_i)}}$$

O vectorialmente:

$$Z = \widehat{W}^{-\frac{1}{2}}(Y - \hat{Y})$$

Donde $\widehat{W}$ es una matriz diagonal con términos $\hat{y}_i(1-\hat{y}_i)$.

3. Estimar por mínimos cuadrados una regresión con variable dependiente Z y matriz de regresores $T = \widehat{W}^{\frac{1}{2}}X$. Los parámetros estimados con esta regresión, $\hat{b}_1$, vendrán dados por:

$$\hat{b}_1 = (T'T)^{-1}T'Z = \left(X'\widehat{W}X\right)^{-1}X'(Y - \hat{Y})$$

Y, comparando con (I), vemos que $\hat{b}_1$ estima el incremento $\beta_a - \hat{\beta}_1$ de los parámetros que nos acerca al máximo.

4. Obtener un nuevo estimador de los parámetros $\hat{\beta}_2$ del modelo logístico mediante

$$\hat{\beta}_2 = \hat{\beta}_1 - \hat{b}_1$$

5. Tomar el valor estimado resultante de la etapa anterior, que en general llamaremos $\hat{\beta}_h$, y sustituirlo en la ecuación del modelo logístico para obtener el vector de estimadores $Y(\hat{\beta}_h)$= $\hat{Y}_h$. Utilizando este vector $\hat{Y}_h$ construir la matriz $\widehat{W}_h$ y las nuevas variables Z_h y T_h

$Z_h = \widehat{W}_h^{-\frac{1}{2}}(Y - Y_h)$,

$$T_h = \widehat{W}_h^{-\frac{1}{2}}X$$

Y volver a la etapa 2. El proceso se repite hasta obtener la convergencia ($\hat{\beta}_{h+1} \cong \hat{\beta}_h$). (Basualdo, 2014)

3.6 Contrastes de significación

Para contrastar la significatividad global en los modelos de regresión logística, se utiliza el estadístico de razón de verosimilitud (RV). Normalmente se usa la prueba ómnibus. (Veliz, 2016).

Las hipótesis nula y alternante son:

$H_o : \beta_2 = \beta_3 = \cdots \beta_k = 0$
$H_a : NoH_o$

CAPITULO IV

Medidas de bondad de ajuste

4.1 Coeficiente Pseudo R^2

En vez de utilizar el R^2 de la regresión lineal, en los modelos logit se utilizan los Pseudo R^2, el cual es un estadístico que mide cuánto de la falta inicial de ajuste chi cuadrado es explicada por las variables en el modelo. (Gujarati, 2004)

El primero que apareció fue el de Mac Fadden:

$$PR^2_{MacFadden} = 1 - \frac{\ln L}{LnL_0}$$

Luego el de Cox y Snell:

$$PR^2_{Cox-Snell} = 1 - e^{-\frac{2}{N}(LnL_0 - LnL)} = 1 - e^{-\frac{RV_0}{N}}$$

Finalmente, el de Nagelberke:

$$PR^2_{Nagelberke} = \frac{PR^2_{Cox-Snell}}{PR^2_{MaxCox-Snell}} = \frac{PR^2_{Cox-Snell}}{1 - e^{-\frac{2}{N}LnL_0}}$$

Cabe resaltar que, "L" es el estimador de máxima verosimilitud del modelo con todas las variables explicativas y "L_0" es el estimador de máxima verosimilitud del modelo sin variables explicativas.

4.2 Contraste de bondad de ajuste de Hosmer y Lemeshow

Se dividen todos los casos en deciles basados en las probabilidades predichas, el primer decil se cuentan los casos con las probabilidades más altas, siendo el estadístico:

$$HL = \sum_{i=1}^{10} \frac{\left[O_i - N_i \bar{\pi}_i \right]^2}{N_i \bar{\pi}_i (1 - \bar{\pi}_i)}$$

Donde:

O_i : número de unos en el decil i-ésimo

$\bar{\pi}_i$: media de probabilidades en el decil i-ésimo

N_i : número de observaciones en el decil i-ésimo

Las hipótesis nula y alternante son:

H_0 : no existen diferencias entre los valores observados y predichos

H_a : existen diferencias entre los valores observados y predichos

"HL" se distribuye como una Chi-cuadrado con 8 grados de libertad. Si rechazamos la "H_0", implica que el modelo ajustado no es el adecuado. (Hosmer, 2000)

4.3 Curva ROC

La curva ROC (del inglés *Receiver Operating Characteristic curves*) indica que cuanto más alejada este de la diagonal principal mejor es el método de diagnóstico, ya que la curva ROC ideal sería la que con una especificidad de 1 tuviera una sensibilidad de 1, y cuanto más cercana esté a dicha diagonal peor será el método de diagnóstico. Cabe recordar que la diagonal principal es la que corresponde al peor test de diagnóstico y que tiene un área bajo de ella de 0.5. Esto se puede afirmar con el área bajo la curva ROC, que se utiliza como medida de discriminación y representa para todos los pares posibles de individuos formados por un individuo en el que ocurrió el evento y otro en el que no, la proporción de los que el modelo predice una mayor probabilidad para el que tuvo el evento. (Alonzo, 2002)

Las hipótesis nula y alternante son:

H_0 : el área bajo la curva ROC es igual a 0.5

H_a : el área bajo la curva ROC no es igual a 0.5

Si rechazamos la "H_0" asociado a un "p-value", implica que el modelo ajustado es el adecuado.

A partir de un área de 0,7 la discriminación del modelo se considera aceptable.

4.4 Tablas de clasificación

La tabla de clasificación muestra la distribución de valores observados y estimados. Los valores estimados se obtienen a partir del modelo. (Veliz, 2016)

La regla de clasificación predeterminada para un caso es que si la probabilidad estimada de pertenencia en el grupo de respuesta con el valor más alto es mayor o que igual a 0.5, entonces

predice la pertenencia a ese grupo. De lo contrario, predice la pertenencia al grupo con el valor de respuesta más bajo. (Veliz, 2016)

Por otro lado, la capacidad de que nuestro modelo estime el suceso de interés de cuyo valor es 1, se denomina sensibilidad. Por el contrario, la capacidad de que nuestro modelo no estime el suceso de interés cuyo valor es 0, se denomina especificidad. (Veliz, 2016)

CAPITULO V

Caso de aplicación para la adquisición de tarjetas de crédito

5.1 Población

La población a estudiar fueron las personas naturales bancarizadas, que hayan sido elegidas por una Entidad Bancaria después de pasar por los filtros de riesgos y ser aprobadas para ofrecerles una tarjeta de crédito en sus campañas de captación de nuevos clientes.

5.2 Muestra

La muestra a estudiar fueron las personas naturales bancarizadas, que hayan sido elegidas por la Entidad Bancaria después de pasar por los filtros de riesgos y ser aprobadas para ofrecerles una tarjeta de crédito en sus campañas de captación de nuevos clientes, la cual asciende a 2,559 personas naturales bancarizadas.

5.3 Unidad de análisis

La unidad de análisis fue la persona natural bancarizada en la Entidad Bancaria.

5.4 Fuentes de información

a) Data Warehouse de la Entidad Bancaria

Es el repositorio central de información de la Entidad Bancaria, que tiene información de las campañas de captación masivas de nuevos clientes en un periodo de tiempo determinado.

b) Reporte Crediticio Consolidado

El Reporte Crediticio Consolidado (RCC), es el reporte que emite la Superintendencia de Banca, Seguros y AFP (SBS) de manera mensual de todos los deudores de las Entidades Financieras que reportaron información sobre las deudas, líneas y otras variables relativas a la morosidad de sus clientes.

5.5 Definición de variables

a) Variable dependiente

La variable dependiente es denominada Adquisición, que es cualitativa nominal, la cual toma los valores de 1 si a la persona natural bancarizada adquirió una tarjeta de crédito, definida por aquellas personas naturales bancarizadas que adquirieron una tarjeta de crédito en la

campaña de captación de nuevos clientes; y 0 si la persona natural bancarizada no adquirió la tarjeta de crédito, definida por aquellas personas naturales bancarizadas que no adquirieron una tarjeta de crédito en la campaña de captación de nuevos clientes.

b) Variables independientes

Las variables independientes son las siguientes:

- Share of wallet deudas retail: la variable Share Of Wallet Deudas Retail (SOW_DRETAIL), es cuantitativa continua y recoge el ratio entre la deuda total de sus tarjetas de crédito en las entidades bancarias retail y la deuda total de sus tarjetas de crédito en el sistema financiero de la persona natural bancarizada.

- Saturación total de tarjetas de crédito: la variable Saturación de la Tarjeta de Crédito (SATURACIONTC), es cuantitativa continua y recoge el ratio entre la deuda total de sus tarjetas de crédito en el sistema financiero y la línea total de sus tarjetas de crédito en el sistema financiero de la persona natural bancarizada.

- Comparación entre líneas de tarjeta de crédito: la variable ratio de comparación entre líneas de tarjeta de crédito (COMPARA_LINEATC), es cuantitativa continua y recoge el ratio entre la línea de tarjeta de crédito ofrecida por una Entidad Bancaria y la línea máxima de sus tarjetas de crédito en el sistema financiero de la persona natural bancarizada.

- Invitaciones para obtener una tarjeta de crédito: la variable invitaciones para obtener una tarjeta de crédito, es cualitativa ordinal y muestra el número de veces que la persona natural bancarizada ha recibido invitaciones por teléfono de una Entidad Bancaria para obtener una tarjeta de crédito en seis meses de campaña de captación, codificado con 0, 1, 2, 3, 4 y 5 si no ha recibido ninguna invitación por teléfono en seis meses de campaña de captación, si ha recibido una invitación en seis meses de campaña de captación, si ha recibido dos invitaciones en seis meses de campaña de captación, si ha recibido tres invitaciones en seis meses de campaña de captación, si ha recibido cuatro invitaciones en seis meses de campaña de captación, si ha recibido cinco invitaciones en seis meses de campaña de captación, respectivamente.

- Tipo de tarjetas de crédito ofrecidas: la variable tipo de tarjeta de crédito ofrecidas por una Entidad Bancaria (TIPO_TC), es cualitativa nominal y muestra si a la persona natural

bancarizada se le ofrece la tarjeta de crédito clásica, codificado con 0 y 1 si a la persona natural bancarizada se le ofrece la tarjeta de crédito visa, respectivamente.

- Sexo: la variable sexo, es cualitativa nominal y muestra el sexo de la persona natural bancarizada, codificado con 0 y 1 si es de sexo femenino o masculino, respectivamente.

- Edad: la variable edad está dividida en rangos (R_EDAD), es cualitativa ordinal y recoge la edad de la persona natural bancarizada en años, codificado con 1, 2, 3, 4 y 5 si tiene una edad menor o igual a 36 años, si tiene una edad entre 36 y 40 años, si tiene una edad entre 40 y 50 años, si tiene una edad entre 50 y 60 años; y si tiene una edad mayor a 60 años, respectivamente.

- Línea de Tarjeta de Crédito Ofrecida: la variable Línea de Tarjeta de Crédito Ofrecida está dividida en rangos (R_LINEATC), es cualitativa ordinal y recoge la línea de tarjeta ofrecida por la Entidad Bancaria a la persona natural bancarizada en Soles, codificado con 1, 2, 3, 4 y 5 si tiene un línea de tarjeta de crédito ofrecida menor o igual a 1,000, si tiene un línea de tarjeta de crédito ofrecida entre 1,000 y 4,000, si tiene un línea de tarjeta de crédito ofrecida entre 4,000 y 8,000, si tiene un línea de tarjeta de crédito ofrecida entre 8,000 y 15,000 y si tiene un línea de tarjeta de crédito ofrecida mayor o igual a 15,000, respectivamente.

- Tenencia de crédito hipotecario: la variable Tenencia de crédito hipotecario (FLAG_HIPOTECARIO), es cualitativa nominal y muestra si la persona natural bancarizada tiene crédito hipotecario en el Sistema Financiero, codificado con 0 y 1 si no tiene crédito hipotecario o si tiene crédito hipotecario, respectivamente.

- Apalancamiento de la persona natural bancarizada: la variable Rango de Apalancamiento (R_APALANCAMIENTO), es cualitativa ordinal y muestra el nivel de apalancamiento de la persona natural bancarizada por rangos; codificado con 1, 2, 3 y 4 si está apalancado menos 0.25, si está apalancado más o igual que 0.25 y menos que 2.25 y si está apalancado más de 2.25. Donde:

Apalancamiento = Ratio entre la deuda consumo total en el sistema financiero sin incluir deuda automotriz ni hipotecario y el ingreso bruto de la persona natural bancarizada.

- Nivel educacional: la variable Nivel educacional (R_NE), es cualitativa nominal y muestra si la persona natural bancarizada tiene educación superior o no, codificado con 0 y 1 si no educación superior o si tiene educación superior, respectivamente.

- Número de tarjetas de crédito de consumo usadas: la variable Número de tarjetas de crédito de consumo usadas por la persona natural bancarizada (NROTC_USADAS), es cualitativa ordinal y muestra el número de tarjetas de crédito para consumo que tienen deudas consumo asignadas en el sistema financiero, codificado con 1, 2, 3, 4 y 5 si tiene una tarjeta de crédito para consumo con deuda asignada, si tiene dos tarjetas de crédito para consumo con deudas asignadas, si tiene tres tarjetas de crédito para consumo con deudas asignadas, si tiene cuatro tarjetas de crédito para consumo con deudas asignadas o si tiene cinco tarjetas de crédito para consumo con deudas asignadas, respectivamente.

5.6 Muestra de entrenamiento y validación

Se realizó la aplicación mediante el análisis de 2,559 personas naturales bancarizadas, correspondiente a la base de datos de una Entidad Bancaria. El 80% de la muestra (2,047 personas naturales bancarizadas) se usó como la muestra para la construcción del modelo estadístico de regresión logística, mientras que el 20% de la muestra (512 personas naturales bancarizadas) se usó como muestra para la validación del modelo estadístico.

5.7 Análisis de significancia entre la variable dependiente e independientes

a) Variable dependiente versus variables independientes cuantitativas

Para realizar el análisis de significancia entre las variables independientes share of wallet deudas retail, saturación total de tarjeta de crédito, comparación de líneas de tarjeta de crédito y la variable dependiente adquisición, se usó el contraste de Levene, seguido de la prueba T para muestras independientes.

La tabla 1, recoge en primer lugar, el contraste de Levene (F). Las columnas siguientes contienen el estadístico t, sus grados de libertad (gl), el nivel de significación (bilateral), la diferencia entre el share of wallet deudas retail promedio de cada grupo, el error típico de esa diferencia, y los límites inferior y superior del intervalo de confianza al 95%. Toda esta información está calculada tanto para el caso de asumir varianzas iguales y de no asumir varianzas iguales.

De este modo, se puede apreciar que la probabilidad asociada al estadístico de Levene es 0.00, el cual es menor a un nivel de significancia de 0.05, por lo que debemos rechazar la hipótesis de igualdad de varianzas, por lo tanto, se usó la información para el caso que no se asume varianzas iguales. Al usar esta información, se tiene que el estadístico t toma el valor de 127.85

y tiene asociado un nivel de significancia de 0.00, el cual es menor a un nivel de significancia de 0.05, por lo que se rechaza la hipótesis de igualdad de medias y, por ende, se concluye al 95% de confianza que el share of wallet deudas retail de las personas naturales bancarizadas que adquirieron la tarjeta de crédito y los que no la adquirieron no es el mismo.

El intervalo de confianza permite estimar que la verdadera diferencia entre el share of wallet deudas retail promedio de la población de personas naturales bancarizadas que adquirieron la tarjeta de crédito y no la adquirieron se encuentra entre 0.43 y 0.49. El hecho de que el intervalo obtenido no incluya el valor cero, también permite rechazar la hipótesis de igualdad de medias.

Tabla 1.*Prueba de medias para la variable share of wallet deudas retail*

Share of walet deudas retail	Prueba de Levene		Prueba T para la igualdad de medias						
	F	Sig.	t	gl	Sig. (bilateral)	Diferencia de medias	Error típ. de la	95% de Intervalo de Inferior	Superior
Se han asumido varianzas iguales	127.85	0.00	29.58	2,557.00	0.00	0.46	0.02	0.43	0.49
No se han asumido varianzas iguales			28.59	1,993.22	0.00	0.46	0.02	0.43	0.49

La tabla 2, recoge en primer lugar, el contraste de Levene (F). Las columnas siguientes contienen el estadístico t, sus grados de libertad (gl), el nivel de significación (bilateral), la diferencia entre la saturación de tarjeta de crédito promedio de cada grupo, el error típico de esa diferencia, y los límites inferior y superior del intervalo de confianza al 95%. Toda esta información está calculada tanto para el caso de asumir varianzas iguales y de no asumir varianzas iguales.

De este modo, se puede apreciar que la probabilidad asociada al estadístico de Levene es 0.00, el cual es menor a un nivel de significancia de 0.05, por lo que se rechaza la hipótesis de igualdad de varianzas, por lo tanto, se usó la información para el caso que no se ha asumido varianzas iguales. Al usar esta información, se tiene que el estadístico t toma el valor de -3.169 y tiene asociado un nivel de significancia de 0.00, el cual es menor a un nivel de significancia de 0.05, por lo que se rechaza la hipótesis de igualdad de medias y, por ende, se concluye que al 95% de confianza, la saturación total de tarjeta de crédito de las personas naturales que adquirieron la tarjeta de crédito y los que no la adquirieron no es la misma.

El intervalo de confianza permite estimar que la verdadera diferencia entre la saturación total de tarjeta de crédito promedio de la población de personas naturales bancarizadas que

adquirieron la tarjeta de crédito y no la adquirieron se encuentra entre -0.05 y -0.01. El hecho de que el intervalo obtenido no incluya el valor cero, también permite rechazar la hipótesis de igualdad de medias.

Tabla 2. *Prueba de medias para la variable saturación total de tarjeta de crédito*

Saturación total de tarjetas de crédito	Prueba de Levene		Prueba T para la igualdad de medias						
	F	Sig.	t	gl	Sig. (bilateral)	Diferencia de medias	Error típ. de la diferencia	95% de Intervalo de Confianza	
								Inferior	Superior
Se han asumido varianzas iguales	3.00	0.00	-3.18	2,557.00	0.00	-0.03	0.01	-0.05	-0.01
No se han asumido varianzas iguales			-3.17	2,268.23	0.00	-0.03	0.01	-0.05	-0.01

La table 3, recoge en primer lugar, el contraste de Levene (F). Las columnas siguientes contienen el estadístico t, sus grados de libertad (gl), el nivel de significación (bilateral), la diferencia entre la comparación de líneas de tarjeta de crédito promedio de cada grupo, el error típico de esa diferencia, y los límites inferior y superior del intervalo de confianza al 95%. Toda esta información está calculada tanto para el caso de asumir varianzas iguales y de no asumir varianzas iguales.

De este modo, se puede apreciar que la probabilidad asociada al estadístico de Levene es 0.00, el cual es menor a un nivel de significancia de 0.05, por lo que se rechaza la hipótesis de igualdad de varianzas, por lo tanto, se usó la información para el caso que no se ha asumido varianzas iguales. Al usar esta información, se tiene que el estadístico t toma el valor de 35.04 y tiene asociado un nivel de significancia de 0.00, el cual es menor a un nivel de significancia de 0.05, por lo que se rechaza la hipótesis de igualdad de medias y, por ende, se concluye que al 95% de confianza, la comparación de líneas de tarjeta de crédito de las personas naturales bancarizadas que adquirieron la tarjeta de crédito y los que no la adquirieron no es el mismo.

El intervalo de confianza nos permite estimar que la verdadera diferencia entre la comparación de líneas de tarjeta de crédito promedio de la población de personas naturales bancarizadas que adquirieron la tarjeta de crédito y no la adquirieron se encuentra entre -0.10 y -0.03. El hecho de que el intervalo obtenido no incluya el valor cero, también permite rechazar la hipótesis de igualdad de medias.

Tabla 3. *Prueba de medias para la variable comparación entre líneas de tarjeta de crédito*

Comparación entre lineas de tarjetas de crédito	Prueba de Levene		Prueba T para la igualdad de medias						
	F	Sig.	t	gl	Sig. (bilateral)	Diferencia de medias	Error típ. de la	95% de Intervalo de Inferior	Superior
Se han asumido varianzas iguales	35.04	0.00	-3.85	2,557.00	0.00	-0.07	0.02	-0.10	-0.03
No se han asumido varianzas iguales			-3.75	2,048.96	0.00	-0.07	0.02	-0.10	-0.03

b) Variable dependiente versus variables independientes cualitativas

Para realizar el análisis de significancia entre las variables independientes Número de Entidades financieras, Tenencia de crédito hipotecario, Apalancamiento, Número de tarjetas de crédito de consumo usadas, Invitaciones para obtener una tarjeta de crédito, Tipo de tarjeta de crédito ofrecida, Línea de Tarjeta de Crédito Ofrecida, Sexo, Edad y Nivel educacional; con respecto a la variable dependiente adquisición, se usó la prueba Chi-Cuadrado de Pearson.

La tabla 4, recoge la significancia entre la variable adquisición (variable dependiente) y las variables independientes mediante el uso de la prueba Chi-Cuadrado de Pearson. En esta tabla, se puede apreciar que todas las variables presentadas, tienen un nivel de significación menor a 0.05, por lo que se rechaza la hipótesis Ho de independencia con respecto a la adquisición. Por lo tanto, se puede concluir que al 95% de confianza, la adquisición si depende de tales variables.

Tabla 4. *Prueba Chi-Cuadrado de Pearson para variables independientes cualitativas*

Variables vs Adquisición	Valor	gl	Sig. Asintótica (bilateral)
FLAG_HIPOTECARIO	6.16	1	0.01
INVITACIONES	18.38	5	0.00
R_EDAD	52.51	4	0.00
TIPO_TC	4.43	1	0.04
SEXO	32.11	1	0.00
NROTC_USADAS	180.74	4	0.00
R_NE	20.40	1	0.00
R_LINEATC	14.12	4	0.01
R_APALANCAMIENTO	18.67	2	0.00

5.8 Análisis de regresión logística

a) Interpretación de coeficientes y Odds ratios

En la tabla 5, se recoge el resultado del modelo estadístico de regresión logística, incluyendo aquellas variables que resultaron significativas.

Tabla 5. *Modelo de Regresión Logística*

Variables de la ecuación	Adquisición					
	B	E.T.	Wald	gl	Sig.	Exp(B)
SOW_DRETAIL	-4.896	0.249	386.107	1	0.000	0.007
COMPARA_LINEATC	1.089	0.174	39.030	1	0.000	2.971
INVITACIONES			11.981	5	0.035	
INVITACIONES(1)	0.121	0.173	0.488	1	0.485	1.129
INVITACIONES(2)	-0.089	0.182	0.239	1	0.625	0.915
INVITACIONES(3)	-0.514	0.179	8.244	1	0.004	0.598
INVITACIONES(4)	0.045	0.431	0.011	1	0.918	1.046
INVITACIONES(5)	0.361	0.680	0.282	1	0.595	1.434
R_LINEATC			40.740	4	0.000	
R_LINEATC(1)	-0.482	0.138	12.186	1	0.000	0.617
R_LINEATC(2)	-0.823	0.313	6.894	1	0.009	0.439
R_LINEATC(3)	-1.746	0.366	22.784	1	0.000	0.175
R_LINEATC(4)	-1.976	0.385	26.374	1	0.000	0.139
R_NE(1)	-0.293	0.138	4.529	1	0.033	0.746
NROTC_USADAS			198.361	4	0.000	
NROTC_USADAS(1)	-2.766	0.207	178.395	1	0.000	0.063
NROTC_USADAS(2)	-3.105	0.303	104.753	1	0.000	0.045
NROTC_USADAS(3)	-3.215	0.521	38.044	1	0.000	0.040
NROTC_USADAS(4)	-2.230	1.183	3.552	1	0.059	0.107
Constante	3.281	0.260	159.457	1	0.000	26.608

Los coeficientes B son los coeficientes reales del modelo estadístico de regresión logística, pero cabe recordar que tiene una relación no lineal con la probabilidad de que una persona natural bancarizada . una tarjeta de crédito. La segunda columna E.T. contiene los errores estándar para los coeficientes B. La tercera columna, muestra el estadístico de Wald y se utiliza para probar si el predictor se relaciona de manera significativa con la variable respuesta ajustando para las otras variables en la ecuación. La última columna presenta los coeficientes B exponenciales usando la función exponencial también llamados odds-ratio.

A continuación, se muestra la interpretación para cada una de las variables del modelo estadístico de regresión logística:

- Para la variable Comparación entre la línea ofrecida por una Entidad Bancaria y la máxima línea de tarjeta de crédito en el sistema financiero, se aprecia que tiene un odd-ratio con valor de 2.971, el cual está por encima de 1, lo que significa que ante un aumento en el Ratio de

Comparación entre la línea ofrecida y la máxima línea de tarjeta de crédito en el sistema financiero de la persona natural bancarizada, la probabilidad de que adquiera una tarjeta de crédito aumenta con un factor de 2.669.

- Para la variable Share of wallet de las deudas retail, se aprecia que tiene un odd-ratio con valor de 0.007, lo que significa que ante una disminución en el Share of wallet de las deudas retail de la persona natural bancarizada, la probabilidad de que adquiera una tarjeta de crédito disminuye con un factor de 0.007.

- Para la variable Invitaciones para obtener una tarjeta de crédito en seis meses de campaña de captación de nuevos clientes de una Entidad Bancaria, se aprecia que INVITACIONES(1) tiene un odd-ratio con valor de 1.129, lo que indica que, si los otros factores se mantienen constantes, pasar de que la persona natural bancarizada no haya recibido ninguna invitación a tener una invitación aumenta la probabilidad de que adquiera la tarjeta de crédito por alrededor de 1.129. Para INVITACIONES(2) tiene un odd-ratio con valor de 0.915, lo que indica que, si los otros factores se mantienen constantes, pasar de que la persona natural bancarizada no haya recibido ninguna invitación a tener dos invitaciones disminuye la probabilidad de que adquiera la tarjeta de crédito por alrededor de 0.915. Para INVITACIONES(3) tiene un odd-ratio con valor de 0.598, lo que indica que, si los otros factores se mantienen constantes, pasar de que la persona natural bancarizada no haya recibido ninguna invitación a tener tres invitaciones disminuye la probabilidad de que adquiera la tarjeta de crédito por alrededor de 0.598. Para INVITACIONES(4) tiene un odd-ratio con valor de 1.046, lo que indica que, si los otros factores se mantienen constantes, pasar de que la persona natural bancarizada no haya recibido ninguna invitación a tener cuatro invitaciones aumenta la probabilidad de que adquiera la tarjeta de crédito por alrededor de 1.046. Para INVITACIONES(5) tiene un odd-ratio con valor de 1.434, lo que indica que, si los otros factores se mantienen constantes, pasar de que la persona natural bancarizada no haya recibido ninguna invitación a tener cinco invitaciones aumenta la probabilidad de que adquiera la tarjeta de crédito por alrededor de 1.434.

- Para la variable Línea de tarjeta de crédito ofrecida por una Entidad Bancaria se aprecia que R_LINEATC(1) tiene un odd-ratio con valor de 0.615, lo que indica que, si los otros factores se mantienen constantes, pasar de que se le ofrezca a una persona natural bancarizada una línea de tarjeta de crédito menor o igual a 1,000 soles a que le ofrezcan de 1000 a 4000 soles

disminuye la probabilidad de que adquiera la tarjeta de crédito por alrededor de 0.615. Para el R_LINEATC(2) tiene un odd-ratio con valor de 0.439, lo que indica que, si los otros factores se mantienen constantes, pasar de que se le ofrezca a una persona natural bancarizada una línea de tarjeta de crédito menor o igual a 1,000 soles a que le ofrezcan de 4,000 a 8,000 soles disminuye la probabilidad de que adquiera la tarjeta de crédito por alrededor de 0.439. Para el R_LINEATC(3) tiene un odd-ratio con valor de 0.175, lo que indica que, si los otros factores se mantienen constantes, pasar de que se le ofrezca a una persona natural bancarizada una línea de tarjeta de crédito menor o igual a 1,000 soles a que le ofrezcan de 8,000 a 15,000 soles disminuye la probabilidad de que adquiera la tarjeta de crédito por alrededor de 0.175. Para el R_LINEATC(4) tiene un odd-ratio con valor de 0.139, lo que indica que, si los otros factores se mantienen constantes, pasar de que se le ofrezca a una persona natural bancarizada una línea de tarjeta de crédito menor o igual a 1,000 soles a que le ofrezcan más de 15,000 soles disminuye la probabilidad de que adquiera la tarjeta de crédito por alrededor de 0.139.

- Para la variable Nivel educacional de la persona natural bancarizada, se aprecia que R_NE(1) tiene un odd-ratio con valor de 0.746, lo que significa que pasar de que una persona natural bancarizada que no tenga grado superior a que tenga disminuye la probabilidad de que adquiera una tarjeta de crédito por alrededor de 0.746.

- Para la variable Número de tarjetas de crédito de consumo usadas, se aprecia que NROTC_USADAS(1) tiene un odd-ratio con valor de 0.063, lo que indica que, si los otros factores se mantienen constantes, pasar de que una persona natural bancarizada use una tarjeta de crédito consumo a usar dos disminuye la probabilidad de que adquiera la tarjeta de crédito por alrededor de 0.063. Para la variable NROTC_USADAS(2) tiene un odd-ratio con valor de 0.045, lo que indica que, si los otros factores se mantienen constantes, pasar de que una persona natural bancarizada use una tarjeta de crédito a usar tres disminuye la probabilidad de que adquiera la tarjeta de crédito por alrededor de 0.045. Para la variable NROTC_USADAS(3) tiene un odd-ratio con valor de 0.040, lo que indica que, si los otros factores se mantienen constantes, pasar de que una persona natural bancarizada use una tarjeta de crédito a usar cuatro disminuye la probabilidad de que adquiera la tarjeta de crédito por alrededor de 0.040. Para la variable NROTC_USADAS(4) tiene un odd-ratio con valor de 0.107, lo que indica que, si los otros factores se mantienen constantes, pasar de que una

persona natural bancarizada use una tarjeta de crédito a usar cinco disminuye la probabilidad de que adquiera la tarjeta de crédito por alrededor de 0.107.

De este modo, el modelo más adecuado para clasificar a las personas naturales bancarizadas que adquieren la tarjeta de crédito de una Entidad Bancaria se obtuvo con 6 variables, las cuales son: Sow deudas retail (X_1), Comparación entre líneas de tarjeta de crédito (X_2), Invitaciones para obtener una tarjeta de crédito (X_3), Tipo de tarjeta de crédito ofrecida (X_4), Línea de tarjeta de crédito ofrecida (X_5) y Número de tarjetas de crédito de consumo usadas (X_6).

La ecuación del modelo estadístico de regresión logística más adecuado para clasificar a las personas naturales bancarizadas de las campañas de captación de la Entidad Bancaria que van a adquirir la tarjeta de crédito es:

$$P(Y=1/X_1,\cdots,X_k)=\frac{1}{1+e^{-Z_i}}$$

Donde Z_i es:

$$Z_i=\beta_0+\beta_1 X_{1i}+\ldots+\beta_k X_{ki}=\ln\left(\frac{p}{1-p}\right)=\log it\left[P(Y=1)\right]$$ = Logit(Adquisición)

Por lo tanto, usando los coeficientes B de la tabla 5, el modelo se representa de la siguiente manera:

Logit(Adquisición) = 26.608

+ 0.007X_1

+ 2.971X_2

+ 1.129X_{31} - 0.915X_{32} - 0.598X_{33} + 1.046X_{34} + 1.434X_{35}

- 0.617X_{41} - 0.439X_{42} - 0.175X_{43} - 0.139X_{44}

- 0.746X_5

- 0.063X_{61} – 0.045X_{62} - 0.040X_{63} - 0.107X_{64}

b) Prueba ómnibus

La tabla 6, recoge los valores de la prueba ómnibus, en la que se muestra el valor de chi-cuadrado de 1,138.50 con un p-value de 0.000<0.05, lo que indica que al 95% de confianza, hay una relación significativa entre las variables independientes, entonces el modelo es significativo.

Tabla 6. *Prueba Ómnibus*

Modelo	Chi Cuadrado	gl	Sig.
	1,125.57	16	0.000

5.9 Análisis de bondad de ajuste

a) Pseudo R-cuadrado

En la tabla 7, se recogen los valores de los diferentes pseudo R-cuadrados, los que indican que el modelo sólo explica una cantidad aceptable de chi-cuadrado inicial sin explicación.

Tabla 7. *Resumen del modelo*

-2 log de la verosimilitud	R cuadrado de Cox y Snell	R cuadrado de Nagelkerke
1,656.436	46.3%	58.92%

b) Prueba de Hosmer y Lemeshow

En tabla 8, se recogen los valores de la prueba de Hosmer y Lemeshow, en la que se muestra un valor de chi-cuadrado de 13.02 con un p-value de 0.111>0.05 (nivel de significancia), lo que indica que al 955 de confianza, no existen diferencias entre los valores observados y valores estimados, por la tanto se puede concluir que el modelo ajustado es significativo.

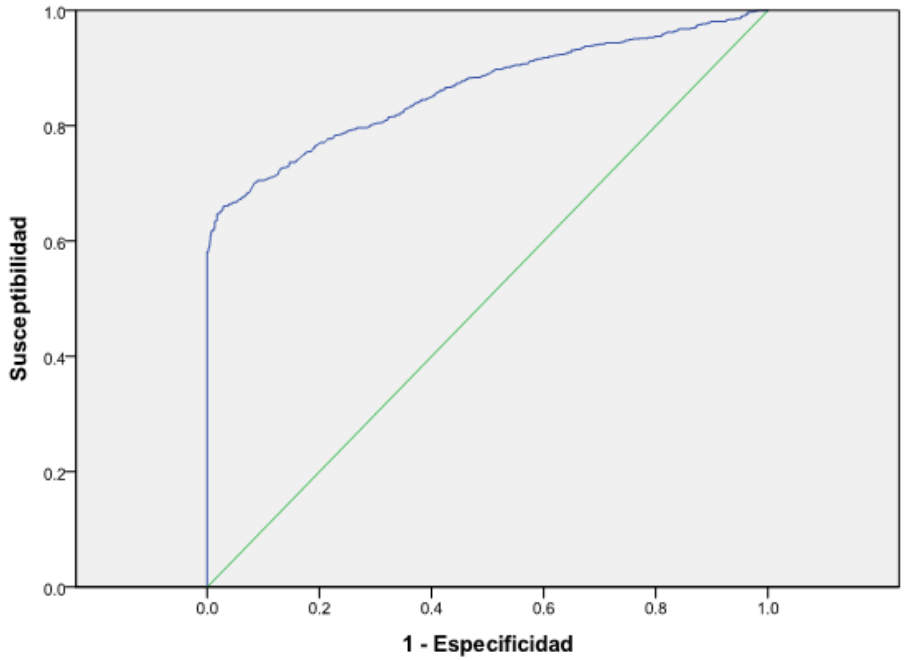
Tabla 8. *Prueba de Hosmer y Lemeshow*

Chi cuadrado	gl	Sig.
13.02	8	0.111

c) Área bajo la curva ROC

En la figura 1, se muestra que la curva ROC.

Figura. 1 *Curva ROC*



Se puede visualizar que la curva ROC está alejada de la diagonal principal y tiene una curvatura pronunciada, por lo tanto, se puede concluir que el modelo ajustado es el adecuado. Esto se puede afirmar con el valor del área bajo la curva ROC.

En la tabla 9, se muestra el valor del área bajo la curva ROC, en la que se muestra un área de 0.863>0.7 con un p-value de 0.000<0.05, lo cual indica una eficacia o capacidad predictiva del modelo muy bueno. Por Io tanto, se puede concluir que el modelo ajustado es el adecuado.

Tabla 9. *Valor del área bajo la curva ROC*

Área	Error típ.a	Sig. Asintóticab
0.863	0.009	0.000

d) Tabla de clasificación para la muestra de entrenamiento

En la tabla 10, se recogen los resultados obtenidos mediante la tabla de clasificación, lo que indica que aquellas personas naturales bancarizadas que adquirieron una tarjeta de crédito en la Entidad Bancaria se clasificaron con una precisión de 66.0% (sensibilidad) y aquellas personas naturales bancarizadas que no adquirieron una tarjeta de crédito, se clasificaron con

una precisión de 97.1% (especificidad). Para el conjunto de la muestra de entrenamiento (2,047 personas naturales bancarizadas) el porcentaje de aciertos del modelo ha sido del 84.1%.

Tabla 10. *Tabla de clasificación para la muestra de entrenamiento*

Observado		Pronosticado		
		Adquisición		Porcentaje correcto
		No Adquirió	Adquirió	
Adquisición	No Adquirió	1,157	35	97.1
	Adquirió	291	564	66.0
Porcentaje Global				84.1

e) Tabla de validación para la muestra de validación

En la tabla 11, se recogen los resultados de la validación del modelo estadístico de regresión logística, lo que indica que aquellas personas naturales bancarizadas que adquirieron una tarjeta de crédito en la Entidad Bancaria se clasificaron con una precisión de 65.4% y aquellas personas naturales bancarizadas que no adquirieron una tarjeta de crédito del banco se clasificaron con una precisión de 93.6%. Para el conjunto de la muestra de la validación (512 personas naturales bancarizadas) el porcentaje de aciertos del modelo ha sido del 81.8%.

Tabla 11. *Tabla de clasificación para la muestra de validación*

Observado		Pronosticado		
		Adquisición		Porcentaje correcto
		No Adquirió	Adquirió	
Adquisición	No Adquirió	279	19	93.6
	Adquirió	74	140	65.4
Porcentaje Global				81.8

f) Clasificación de clientes según probabilidades de adquisición

Al tomar los resultados de las probabilidades obtenidas mediante la aplicación del modelo estadístico de regresión logística para las personas naturales bancarizadas pertenecientes a las campañas masivas de captación de clientes de la Entidad Bancaria, se puede clasificar en tres categorías, las cuales se detallan a continuación:

- Clase con alta propensión a adquirir la tarjeta de crédito, la cual corresponde a las personas naturales bancarizadas con probabilidades de adquirir la tarjeta de crédito, mayores o iguales a 0.70.

- Clase con mediana propensión a adquirir la tarjeta de crédito, la cual corresponde a las personas naturales bancarizadas con probabilidades de adquirir la tarjeta de crédito, mayores o iguales a 0.50 y menores o iguales a 0.70.

- Clase con baja propensión a adquirir la tarjeta de crédito, la cual corresponde a las personas naturales bancarizadas con probabilidades de adquirir la tarjeta de crédito, menores o iguales a 0.50.

La tabla 12, recoge la clasificación de clientes según probabilidad de adquisición para la muestra de entrenamiento. De la tabla, se puede apreciar que el 24.7% (505 personas naturales bancarizadas) están en la categoría de alta propensión; un 4.6% (94 personas naturales bancarizadas) están en la categoría de mediana propensión; y el 70.7% (1448 personas naturales bancarizadas) se encuentran en la categoría de baja propensión.

Tabla 12. *Clasificación de clientes según probabilidad de adquisición para la muestra de entrenamiento*

Clase de Adquisición	Nro. de Personas	Adquisición	% Adquisición
Alta Propensión: ProbAdquisición [0.7 - 1]	505	502	99.4%
Mediana Propensión: ProbAdquisición [0.5 - 0.7>	94	62	66.0%
Baja Propensión: ProbAdquisición [0 - 0.5>	1,448	291	20.1%
Total	2,047	855	41.8%

Finalmente, en la tabla 13, se recoge la clasificación de personas naturales bancarizadas según probabilidad de adquisición para la muestra de validación. De la tabla, se puede apreciar que el 26.4% (135 personas) están en la categoría de baja propensión; sólo un 5.5% (28 personas) están en la categoría de mediana propensión; y el 68.2% (349 personas) se encuentran en la categoría de alta propensión.

Tabla 13. *Clasificación de clientes según probabilidad de adquisición para la muestra de validación*

Clase de Adquisición	Nro. de Personas	Adquisición	% Adquisición
Alta Propensión: ProbAdquisición [0.7 - 1]	135	131	97.0%
Mediana Propensión: ProbAdquisición [0.5 - 0.7>	28	17	60.7%
Baja Propensión: ProbAdquisición [0 - 0.5>	349	66	18.9%
Total	512	214	41.8%

CAPITULO VI

Conclusiones

En esta parte se exponen las conclusiones obtenidas como producto del desarrollo del caso de aplicación, las cuales son las siguientes:

1. Mediante la aplicación del modelo estadístico de regresión logística se pudo identificar a las personas naturales bancarizadas con propensión a adquirir una tarjeta de crédito de una Entidad Bancaria, obteniéndose una tasa de acierto de 66.0%. Finalmente, en la muestra de validación se obtuvo una tasa de acierto de 65.4%.

2. A través del modelo estadístico de regresión logística se encontró que las variables de mayor importancia que ayudan en la clasificación de las personas naturales bancarizadas con propensión a adquirir una tarjeta de crédito son: Share of wallet deudas retail, Comparación entre líneas de tarjeta de crédito, Número de tarjetas de crédito de consumo usadas, Invitaciones para obtener una tarjeta de crédito, Línea de Tarjeta de Crédito Ofrecida y Nivel educacional. Las cuales explican el 46.3% (R cuadrado de Cox y Snell), de varianza de la variable dependiente adquisición. Mientras que el R cuadrado de Nagelkerke, muestra que explican el 58.9% de la varianza de la variable dependiente.

3. Se evaluó el desempeño del modelo estadístico de regresión logística, obteniendo:

- Se tiene que pvalue=0.111 el cual es mayor a 0.05. Entonces no se rechaza la hipótesis nula que plantea que no existen diferencias entre los valores observados y predichos. Por lo que podemos decir que al 95% de confianza, el modelo se ajusta adecuadamente a los datos.

- En la tabla de clasificación se obtuvo una tasa de clasificación global de 84.1%, mientras que la tasa de clasificación de las personas que adquirieron la tarjeta de crédito fue de 66% y de las que no adquirieron de 97.1%. Esto se corrobora en la tabla de clasificación de la muestra de validación, ya que la clasificación global es de 81.8%, mientras que la tasa de clasificación de las personas que adquirieron la tarjeta de crédito fue de 65.4% y de las que no adquirieron 93.6%.

- Finalmente, se usó el área bajo la curva ROC para medir la bondad de ajuste del modelo. Se muestra que el área bajo la curva es igual a 0.863, la cual es mayor a 0.5, que es área bajo la curva de la diagonal principal que corresponde al peor test de diagnóstico. Esta área está

asociado a un p-value de 0.000, el cual es menor a 0.05, por lo que se rechaza la hipótesis nula que plantea que el área bajo la curva es igual a 0.5. De lo que podemos concluir que la discriminación del modelo es buena.

REFERENCIAS BIBLIOGRAFICAS

Alonzo T. A., & Pepe, M. S. (2002). *Distribution-free ROC analysis using binary regression techniques*. Biostatistics, 3(3), 421–432. https://doi.org/10.1093/biostatistics/3.3.421

Basualdo Najera, S.L. (2014). *Aplicación de la regresión logística para la clasificación de personas naturales bancarizadas con propensión a adquirir una tarjeta de crédito de una entidad bancaria*. Tesis, Universidad Nacional de Ingeniería. http://hdl.handle.net/20.500.14076/5999

Gujarati, D. N. (2004). *Econometría* (4ta ed.). McGraw-Hill Interamericana.

Hosmer, D. W., & Lemeshow, S. (2000). *Applied logistic regression* (2nd ed.). Wiley.

Langrand, C., & Pinzón, L. M. (2009). *Análisis de datos: Métodos y ejemplos*. Editorial Escuela Colombiana de Ingeniería.

Menard, S. (2002). *Applied logistic regression analysis* (2nd ed.). SAGE Publications.

Monografias.com. (s.f.). *Tarjetas de crédito*. Recuperado el 15 de diciembre de 2024 de https://acortar.link/9CWVUH

Superintendencia de Banca, Seguros y AFP. (2008). *Resolución S.B.S. N° 11356 - 2008*. https://acortar.link/UA9in3

Universidad Dr. José Matías Delgado. (s.f.). *Origen y evolución de las tarjetas de crédito*. https://acortar.link/z45Ohs

Veliz Capuñay, C. (2016). *Análisis Multivariante. Métodos estadísticos multivariantes para la investigación*. Cengage.

MAG. SOLANGE LESLY BASUALDO NAJERA.

AUTOR PRINCIPAL

ORCID: https://orcid.org/0009-0001-0567-689X

Peruana, se desempeña como docente de Pregrado y Posgrado en diferentes universidades en el Perú, destacando la Universidad Nacional de Ingeniería.

Magister en Marketing y Gestión Comercial por la Universidad del Pacífico. Ingeniera Estadística por la Universidad Nacional de Ingeniería. Con Especialización en Data Storytelling por INCAE Business School, Especialización en Data Science and Machine Learning por Massachusetts Institute Technology.

Con más de 14 años de experiencia liderando proyectos de Ciencia de Datos. Se desempeña como Lead Data Analytics en Zubale México, ha sido Head de Business Analytics en Bankea, Head de Data Analytics en Ideas Diferentes, Analista de Modelos Analíticos en Scotiabank, Analista de Inteligencia de Negocios en Banco Falabella y Analista de Riesgos en MiBanco.

DR. RICHARD FERNANDO FERNÁNDEZ VÁSQUEZ.

AUTOR PRINCIPAL

ORCID: https://orcid.org/0000-0003-1721-8527

Peruano, actualmente se desempeña como docente de Pregrado y Posgrado en diferentes universidades en el Perú, entre las cuales destaca la Universidad Nacional de Ingeniería. Asesor y revisor de Tesis en Pregrado y Posgrado.

Doctor en Estadística y Matemática en temas de Deep Learning e IA. Magister en Marketing y Gestión Comercial por la Universidad Peruana de Ciencias Aplicadas y la EOI de España. Magister Scientiae en Estadística Aplicada por la Universidad Nacional Agraria La Molina. Ingeniero Estadístico por la Universidad Nacional de Ingeniería. Con Especialización en Deep Learning por la Secretaria de Trabajo de México, Especialización en Gestión Integral de Riesgos por Equilibrium Financiero, Especialización en Ciencia de Datos por la Universidad Nacional Autónoma de México, Especialización en Innovación y Tecnología por San Francisco State University.

Con más de 12 años de experiencia liderando proyectos de Ciencia de Datos. Ha sido SubDirector de Inteligencia y Prospectiva Turística en PROMPERÚ, Head de Modelos & Data Analytics en Agrobanco, Data Science Head en el Grupo El Comercio, Data Science Lead en SUNAT, Subgerente Adjunto de Gestión de Portafolio de Productos en el Banco de Crédito del Perú, Analista de Business Analytics en Banco Financiero y Banco Falabella. Ganador del premio ORO en la categoría de Big Data, Analítica e Inteligencia Artificial de los Premios País a los Innovadores Financieros en las Américas 2023 por Fintech Americas. Reconocido por San Francisco State University en los Estados Unidos por contribuir de manera valiosa en el Programa de Innovación y Tecnología en San Francisco - Silicon Valley 2019 para los estudiantes del MBA de la Universidad de Lima.

Enero 2025 - CID - Centro de Investigación y Desarrollo
Copyright © - CID - Centro de Investigación y Desarrollo
Copyright del texto © 2024 de Autores
Formato: PDF
Tamaño: A4 210 x 297 mm
Requisitos de sistema: Adobe Acrobat Reader
Modo de acceso: World Wide Web
biblioteca.ciencialatina.org
editorial@ciencialatina.org
Atención por WhatsApp al +52 22 2690 3834