Comparison of Cross Selling Models with Symmetrical and Asymmetrical Links with Classical and Bayesian Estimation for Predicting Client Propensity for Acquiring Credit Cards of Financial Bank Corp.

Alexis Coronado Ortiz, Milagros Luz Huanca Huamaní, Alex Pierre Huarcaya Sairitupac

Asignatura: Métodos Econométricos I

Escuela Profesional de Ingeniería Estadística

Facultad de Ingeniería Económica, Estadística y Ciencias Sociales

Universidad Nacional de Ingeniería

RESUMEN

En el presente trabajo se realizaron modelos con enlaces simétricos y asimétricos de estimación clásica y bayesiana para predecir la propensión de que los clientes del banco A.M.A. adquieran una tarjeta de crédito. Se decide realizar un modelo estadístico predictivo debido a la baja tasa de aceptación de las tarjetas de crédito (8.5%)

Se evidenció que el modelo asimétrico con enlace Power Logit fue el más adecuado porque se ajusta más a la realidad que atraviesa el banco.

Realizando un análisis a las probabilidades del modelo Power Logit se determinó que si nos enfocamos en el 20% de los clientes con mayor propensión a la aceptación de la tarjeta de crético, obtenemos una tasa de aceptación del 77%. Con esto queda evidenciado la eficiencia de realizar modelos predictivos.

INTRODUCCIÓN

En la continua competencia de las entidades financieras por ampliar su participación en el mercado de ventas de tarjetas de crédito, buscan desarrollar estrategias de marketing identificando diversos perfiles de clientes.

Dentro de este contexto, el banco A.M.A busca ofrecer tarjetas de crédito a los clientes que se encuentran dentro de su cartera. Por tal motivo, se propone la creación de modelos estadísticos que nos permitan conocer la propensión de los clientes del banco A.M.A a adquirir una tarjeta de crédito. De esta manera, se podrán focalizar las distintas estrategias para la captación de clientes.

PRESENTACIÓN DEL PROBLEMA

El banco “A.M.A” ha observado que hasta el año 2014, aproximadamente, el 8.5% de todos los clientes que cuentan con el producto Ahorro Sueldo tiene una tarjeta de crédito.Debido a que esta proporción es muy baja, se tiene la necesidad de aumentar la venta de tarjetas de créditos sin caer en el ofrecimiento masivo que involucre altos costos. Estos costos elevados son debido a la forma tradicional de captar clientes tales como envío de cartas, llamadas telefónicas y correos electrónicos sin considerar los comportamientos diferenciados de los clientes.

OBJETIVOS

- OBJETIVO GENERAL

Predecir la propensión de los clientes del banco A.M.A. para aceptar una tarjeta de crédito mediante la comparación de modelos predictivos de regresión de respuesta binaria con enlaces simétricos y asimétricos.

- OBJETIVOS ESPECÍFICOS

Determinar el modelo más adecuado de regresión binaria para conocer la propensión de los clientes del banco A.M.A a aceptar una tarjeta de crédito

Determinar cuál es el factor que más influyeen los clientes del banco A.M.A. para que acepten una tarjeta de crédito.

COMPRENSIÓN DE DATOS

El banco A.M.A. cuenta con una base de datos de 34461 registros y 80 variables. La característica de esta base de datos es que contiene a aquellos clientes que cuentan con el producto Ahorro sueldo que ofrece el banco A.M.A.De todos estos clientes, tan solo el 8.5% ha adquirido una tarjeta de crédito.

PREPARACIÓN DE DATOS

- **Limpieza de datos**

De las 80 variables contenidas en la base de datos, se eliminaron las variables que contenían información sobre el tipo de servicio BFP. Asi mismo se eliminaron variables que tienen que ver con Microfinanzas, y a aquellas como Saldo en los siguientes: Comex, Creditos por liquidar, Descuentos, Factoring, Lease Back, Mediano Plazo, Prestamos, Sobregiros, Enlatado, Cartas de crédito, Avales, Cartas Fianza. Esto debido a que los datos tienen un muy bajo porcentaje de aparición.

Las variables categóricas tales como Grupo generacional, Estado civil y Zona tuvieron un tratamiento especial: Cada categoría pasó a ser una variable para poder escoger a aquellas categorías más significativas.

Dado que se tenía la variable de Tiempo de Corte y la variable de Inicio de afiliación, se creó la variable Tiempo de Vida que determina el número de días que el cliente se encuentra en el banco A.M.A.

Las variables continuas fueron examinadas gráficamente para determinar el tipo de transformación que debían tener. Por tanto, se determinó trabajar con la raíz cuadrada de las variables Deuda_SBS y Tiempo_Vida.

Luego de haber hecho una limpieza de datos exhaustiva, la base paso a tener 23 variables.

- **Selección de variables**

A continuación, se detallan los resultados de las pruebas Chi-Cuadrado para las variables categóricas.

TABLA N° 01:*Prueba de asociación Chicuadrado con las variables independientes categóricas*

Variable	Valor	gl	Sig
Deuda Linea Tarjeta de Crédito SBS	3450155	1	0.0
Saldo en CTS (Dolares)	1545108	1	0.0
Saldo en Ahorro Efectivo (Dolares)	558343	1	0.0
Saldo en Deposito a plazo (Dolares)	293622	1	0.0
Saldo en Vista (Dolares)	151884	1	0.0
El cliente es conviviente?	143506	1	0.0
Pertenece a la zona de Lima_Moderna?	85212	1	0.0
Sexo	25115	1	0.0
Saldo en Credito Carsa (Dolares)	4212	1	0.0
Saldo en Credito por Convenios (Dolares)	3154	1	0.0

A continuación, se detallan los resultados de las pruebas de independencia de medias para las variables continuas. Esto se realizó teniendo en cuenta la prueba de Levene.

TABLA N°02:*Pruebas de Levene y de igualdad de medias con las variables independientes continuas.*

	Prueba de Levene de calidad de varianzas		Prueba t para la igualdad de medias	
	F	Sig.	t	Sig.
Saldo en Deposito a plazo (Dolares)	1152.10	0.00	8.83	0.00
Raíz cuadrada de la Deuda Total en la SBS	1140.48	0.00	29.74	0.00
Raíz cuadrada del tiempo de vida del cliente en dias	1.99	0.16	28.30	0.00

Por lo tanto, las variables con las que trabajaremos en la construcción de los modelos son las siguientes:

TABLA N° 03: *Variables independientes con las que se construye el modelo.*

Variable	Descripción
CrossSell	Número de productos del cliente en el Banco A.M.A.
Lin_TCre_SBS	Determina si el cliente tiene Deuda en Tarjetas de Crédito según SBS
Conviviente	Determina si el cliente es conviviente
Casado	Determina si el cliente es casado
Tiempo de Vida	El tiempo de vida del cliente en el banco A.M.A en días
Deuda_SBS	Cantidad de Deuda Total según SBS
Sexo	Determina el sexo del cliente
Ahorro Efectivo	Determina si cliente tiene el producto Ahorro Efectivo
Lima_Moderna	Determina si el cliente vive en Lima Moderna
Vista	Determina si cliente tiene el producto Vista
Cts	Determina si cliente tiene el producto Cts
Plazo	Determina el saldo en el producto depósito a Plazo
CredCarsa	Determina si cliente tiene el producto CredCarsa
Convenios	Determina si cliente tiene el producto Crédito por convenios

- **Muestras de partición**

Para la creación de los modelos, se extrajo aleatoriamente una muestra de construcción que representa el 70% de los registros. El otro 30% de registros fue utilizado para validar el modelo.

MODELADO

Modelos Simétricos

Para la creación de los modelos simétricos se debe tener en cuenta que la distribución de los niveles del target (Línea de Crédito) es de aproximadamente 50%-50%. Por tal motivo, se balanceo la muestra de construcción.

- **Modelo Logit clásico**

Estimación de los parámetros

TABLA N° 04:*Estimación de parámetros de las variables independientes usando regresión binaria con enlace Logit clásico.*

Variable	B	Error estándar	Exp(B)
Conviviente	1.897	0.494	6.664
Sexo	-0.256	0.167	0.774
Lima_Moderna	-1.080	0.225	0.340
AhEfec	-0.406	0.248	0.666
Plazo	-4.433	0.579	0.012
Cts	-2.891	0.204	0.056
Vista	-1.280	0.389	0.278
CredCarsa	-6.063	1.131	0.002
Convenios	-8.923	0.607	0.000
CrossSell	5.139	0.195	170.622
Lin_TCre_SBS	4.014	0.255	55.353
Deuda_SBS_Raiz	-.003	0.001	0.997
Vida_R	.008	0.006	1.008
Constante	-11.205	0.523	0.000

Tabla de clasificación

TABLA N° 05:*Tabla de clasificación usando regresión binaria con enlace Logit clásico.*

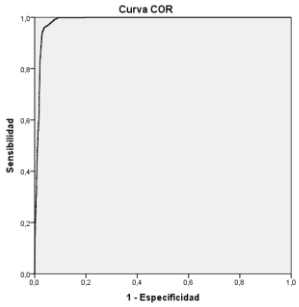
		Pronosticado		
		0	1	%
Observado	0	1937	486	**95.7**
	1	91	1932	**95.5**
	%			**95.6**

En la presente tabla de clasificación, de todos los clientes que si aceptan una línea de crédito, el modelo puede predecir al 95.5% de ellos (**Sensibilidad**). El modelo puede predecir en general al 95.6% de clientes que aceptarán o no una línea de tarjeta de crédito (**Porcentaje de acierto**).

Curva Cor y área bajo la curva

El área bajo la curva mostrada es de 98.59%. Este porcentaje es el valor del COR. Mediante una transformación antes explicada, obtenemos el valor del **Gini** que es **97.19%.**

GRÁFICO N° 01:*Curva COR del modelo de regresión binaria con enlace Logit clásico.*



Modelos Asimétricos

Para la creación de los modelos asimétricos no se maneja ningún supuesto de porcentaje distribucional de la variable dependiente. Por tal motivo, no es necesario balancear la muestra de construcción. Existen muchos tipos de enlaces asimétricos. Uno de ellos es el enlace cloglog.

- **Modelo ClogLog clásico**

Estimación de los parámetros

TABLA N° 06:*Estimación de parámetros de las variables independientes usando regresión binaria con enlace Cloglog clásico.*

	Estimado	Std.Error	Pr(>|z|)
Intercepto	-7.38	0.15	0.00
Conviviente	0.65	0.12	0.00
Sexo_Masculino	-0.20	0.05	0.00
Lima_Moderna	-0.37	0.07	0.00
AhEfec	-0.48	0.08	0.00
Plazo	-1.57	0.15	0.00
Cts	-0.84	0.07	0.00
Vista	-0.70	0.12	0.00
CredCarsa	-1.91	0.41	0.00
Convenios	-3.69	0.24	0.00
CrossSell	1.56	0.04	0.00
Lin_Tcre_SBS	3.38	0.13	0.00
Deuda_SBS_Raiz	0.00	0.00	0.00
Vida_R	0.01	0.00	0.00

Tabla de clasificación

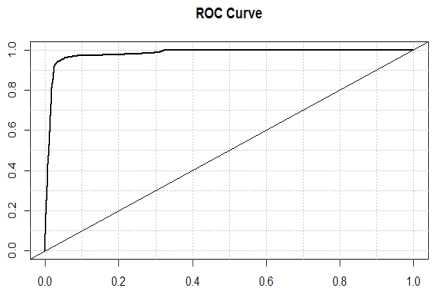
TABLA N° 07:*Tabla de clasificación usando regresión binaria con enlace Logit clásico.*

		Pronosticado		
		0	1	%
Observado	0	9368	88	99.07%
	1	503	379	42.97%
	%			94.28%

En la presente tabla de clasificación, de todos los clientes que si aceptan una línea de crédito, el modelo puede predecir al 42.97% de ellos (**Sensibilidad**). El modelo puede predecir en general al 94.28% de clientes que aceptarán o no una línea de tarjeta de crédito (**Porcentaje de acierto**).

Curva Cor y área bajo la curva

GRÁFICO N° 02:*Curva COR del modelo de regresión binaria con enlace Logit clásico.*



El área bajo la curva mostrada es de 98.01%. Este porcentaje es el valor del COR. Mediante una transformación antes explicada, obtenemos el valor del **Gini** que es **96.03%.**

- **Modelo ClogLog bayesiano**

Tabla de Clasificación

TABLA N° 08:*Tabla de clasificación usando regresión binaria con enlace ClogLog bayesiano.*

		Pronosticado		
		0	1	%
Observado	0	9385	71	99.25%
	1	514	368	41.72%
	%			94.34%

En la presente tabla de clasificación, de todos los clientes que si aceptan una línea de crédito, el modelo puede predecir al 41.72% de ellos (**Sensibilidad**). Finalmente, el modelo puede predecir en general al 94.34% de clientes que aceptarán o no una línea de tarjeta de crédito (**Porcentaje de acierto**).

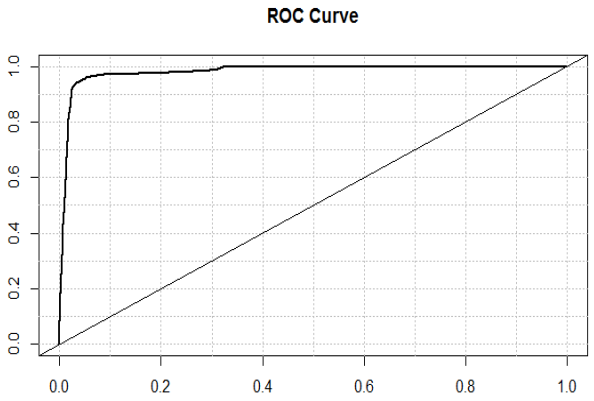
Estimación de los parámetros

TABLA N° 09:*Estimación de parámetros de las variables independientes usando regresión binaria con enlace Cloglog Bayesiano.*

	mean	sd	MC_error
Constante	-4.628	0.8288	0.1421
Conviviente	0.6525	0.1526	0.01953
Sexo	-0.6066	0.1384	0.02438
Lima_mod	-0.3621	0.1091	0.0172
AhEfec	-0.1381	0.1908	0.03182
Plazo	-1.19	0.3494	0.05972
Cts	-0.5778	0.3393	0.05622
Vista	-0.357	0.2044	0.03089
CredCarsa	-2.077	0.5707	0.07401
Convenios	-2.729	0.6733	0.1155
CrossSell	1.362	0.241	0.04064
Lin_TCre	2.188	0.4377	0.07719
Deuda_SBS	2.75E-04	0.001405	1.57E-04
Vida_R	-0.01089	0.01162	0.001974

Curva Cor y área bajo la Curva

GRÁFICO N° 03:*Curva COR del modelo de regresión binaria con enlace Logit clásico.*



El área bajo la curva mostrada es de 96.26%. Este porcentaje es el valor del COR. Mediante una transformación antes explicada, obtenemos el valor del **Gini** que es **96.26%.**

- **Modelo Scobit bayesiano**

Estimación de los parámetros

TABLA N°10:*Estimación de parámetros de las variables independientes usando regresión binaria con enlace Scobit Bayesiano.*

	mean	sd	MC_error
Constante	-7.482	0.3293	0.07017
Conviviente	1.319	0.1791	0.01543
Sexo	-0.895	0.08318	0.01736
Lima_Moderna	-0.7228	0.1208	0.01618
AhEfec	-0.2103	0.1133	0.01537
Plazo	-3.231	0.3369	0.05996
Cts	-2.269	0.1577	0.03242
Vista	-0.8162	0.1883	0.02277
CredCarsa	-4.33	0.6167	0.07346
Convenios	-6.316	0.4785	0.07802
CrossSell	3.469	0.1408	0.03029
Lin_Tcre_SBS	3.028	0.1533	0.03203
Deuda_SBS_raiz	-1.74E-04	0.001079	2.21E-04
Vida_R	5.76E-04	0.002009	3.86E-04

Tabla de clasificación

TABLA N°11:*Tabla de clasificación usando regresión binaria con enlace Scobit bayesiano.*

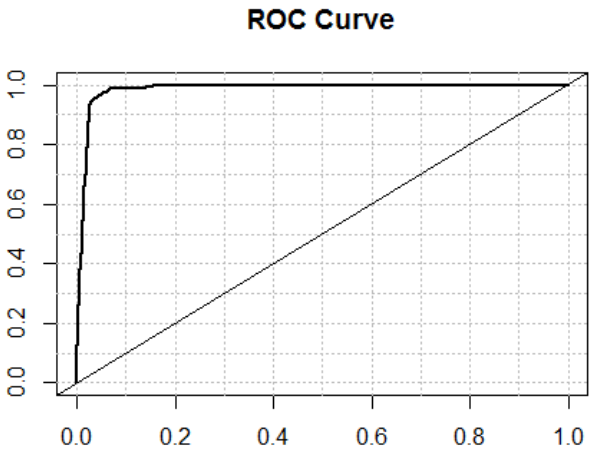
		Pronosticado		
		0	1	%
Observado	0	9282	174	98.16%
	1	267	615	69.73%
	%			95.73%

En la presente tabla de clasificación, de todos los clientes que si aceptan una línea de crédito, el modelo puede predecir al 69.73% de ellos (**Sensibilidad**). Finalmente, el modelo puede predecir en general al 95.73% de clientes que aceptarán o no una línea de tarjeta de crédito (**Porcentaje de acierto**).

Curva Cor y área bajo la Curva

El área bajo la curva mostrada es de 98.494%. Este porcentaje es el valor del COR. Mediante una transformación antes explicada, obtenemos el valor del **Gini** que es **96.811%.**

GRÁFICO N° 04:*Curva COR del modelo de regresión binaria con enlace Scobit Bayesiano.*



- **Modelo Power_Logit bayesiano**

Estimación de los parámetros

TABLA N°12:*Estimación de parámetros de las variables independientes usando regresión binaria con enlace Power Logit Bayesiano.*

	mean	sd	MC_error
Constante	-4.628	0.8288	0.1421
Conviviente	0.6525	0.1526	0.01953
Sexo	-0.6066	0.1384	0.02438
Lima_mod	-0.3621	0.1091	0.0172
AhEfec	-0.1381	0.1908	0.03182
Plazo	-1.19	0.3494	0.05972
Cts	-0.5778	0.3393	0.05622
Vista	-0.357	0.2044	0.03089
CredCarsa	-2.077	0.5707	0.07401
Convenios	-2.729	0.6733	0.1155
CrossSell	1.362	0.241	0.04064
Lin_TCre	2.188	0.4377	0.07719
Deuda_SBS	2.74E-04	0.001405	1.57E-04
Vida_R	-0.01089	0.01162	0.001974

En este cuadro, las variables que presentan un efecto positivo en la predicción de obtener una tarjeta de crédito son: CrossSelling (Cantidad de productos que tiene el cliente en el banco A.M.A.) y Linea de tarjeta de crédito en otros bancos. Las demás variables tienen un efecto negativo para la obtención de una tarjeta de crédito.

Tabla de clasificación

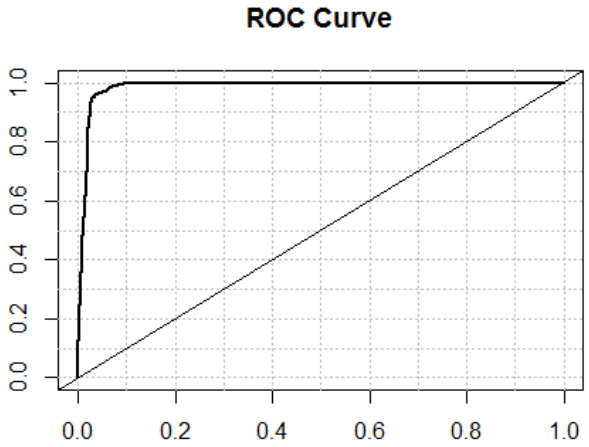
TABLA N°13:*Tabla de clasificación usando regresión binaria con enlace Power Logit bayesiano.*

		Pronosticado		
		0	1	%
Observado	0	9270	186	98.03%
	1	248	634	71.88%
	%			95.80%

En la presente tabla de clasificación, de todos los clientes que si aceptan una línea de crédito, el modelo puede predecir al 71.88% de ellos (**Sensibilidad**). Finalmente, el modelo puede predecir en general al 95.80% de clientes que aceptarán o no una línea de tarjeta de crédito (**Porcentaje de acierto**).

Curva Cor y área bajo la Curva

GRÁFICO N° 05:*Curva COR del modelo de regresión binaria con enlace Power Logit Bayesiano.*



El área bajo la curva mostrada es de 98.572%. Este porcentaje es el valor del COR. Mediante una transformación antes explicada, obtenemos el valor del **Gini** que es **97.14%.**

EVALUACIÓN

A continuación compararemos los modelos de regresión binaria con los distintos enlaces simétricos y asimétricos para determinar el modelo más adecuado.

Esta evaluación se realizará tomando en cuenta 5 importantes indicadores tales como el Gini, Sensibilidad, DIC, EAIC y EBIC.

TABLA N°14:*Tabla de evaluación de modelos.*

	Logit Clásico	Cloglog Clásico	Cloglog Bayesiano	Power Logit Bayesiano	Scobit Bayesiano
Gini	97.2%	96.0%	96.3%	97.1%	97.0%
Sensibilidad	95.7%	99.1%	99.2%	98.0%	98.2%
DIC	4389	5697	7001	4045	4299
EAIC	4417	5725	6598	4060	4306
EBIC	4505	5813	6711	4173	4419

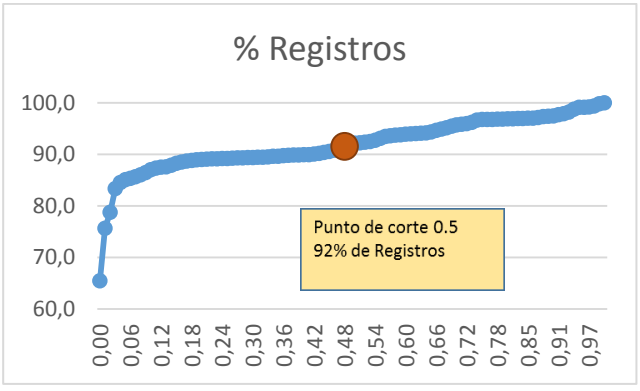
El primer indicador que observamos es el Gini. Este indicador es muy parecido en los 5 modelos porque se encuentra entre 96% y 97%.

El segundo indicador es la sensibilidad, este porcentaje representa el porcentaje de clientes que el modelo predice que aceptan una tarjeta de crédito con respecto a la cantidad total de clientes que aceptan la tarjeta. Buscamos que este porcentaje sea lo más alto posible. Es por esta razón que los modelos que cumplen esa condición son el Cloglog clásico, el Power Logit y el Scobit Bayesiano.

Los indicadores de DIC, EAIC y EBIC deben ser los menores posibles para poder elegir al modelo más adecuado. De esto, decimos que el modelo con menores valores en estos 3 indicadores es el Modelo Power Logit.

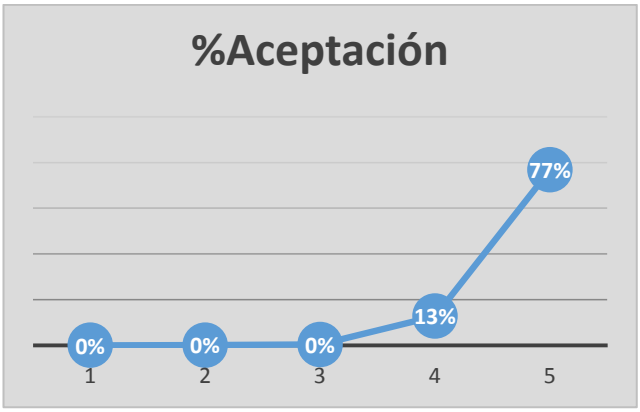
En conclusión, el modelo más adecuado bajo estos 5 criterios es el Modelo de regresión binaria con enlace asimétrico Power Logit Bayesiano.

- **Porcentaje de registros alcanzados de acuerdo al punto de Corte.**



Podemos observar que con un punto de corte del 0.50 a las probabilidades, obtenemos el 92% de los datos. Esto quiere decir que en la muestra de validación tendremos que el 8% clientes aceptan la tarjeta de crédito. Esto se ajusta claramente a la realidad.

- **Tasa de Aceptación de acuerdo probabilidades**



A continuación se muestra la tasa de aceptación de tarjetas de crédito por quintil de probabilidad. Esto quiere decir que si nos enfocamos en el 20% de clientes con propensión más alta a aceptar una tarjeta de crédito, obtenemos una tasa de aceptación del 77%.

- **Estrategias Direccionadas**

Quintil 1, 2, 3
Correos Electronicos masivos

Quintil 4
Correos electronicos personalizados

Quintil 5
A clientes con Cross Selling de 1 a 3 productos se les llamara ofreciendo recompensas como tasas de interes 0 para los primeros 6 meses.

Quintil 5
A clientes con Cross Selling mayor o igual a 4 productos se les llamara a ofrecerles altas lineas de credito y recompensas como millas aereas.

RESULTADOS

- **Orden de importancia de las variables independientes.**

TABLA N° 14: *Tabla de importancia de variables independientes.*

VARIABLE	ODDS RATIO
CrossSell	6.02
Lin_Tcre_SBS	4.40
Conviviente	1.85
Vida_R	1.00
Deuda_SBS_raiz	1.00
AhEfec	0.88
Sexo	0.80
Lima_Moderna	0.71
Vista	0.67
Cts	0.32
Plazo	0.19
CredCarsa	0.13
Convenios	0.04

En la tabla N° 14 tenemos los valores de Odds Ratio para cada una de las variables.

$$Odds\ Ratio\ (Xi) = \frac{P(Y = 1/Xi)}{1 - P(Y = 1/Xi)}$$

El Odds ratio nos ayuda a determinar un orden de importancia de las variables. Dado estos valores, decimos que la variable Cross Selling y la tenencia de Tarjeta de crédito en la SBS son las 2 variables que más influyen en el hecho de que un cliente del banco A.M.A. acepte una tarjeta de crédito.

CONCLUSIONES

El modelo más adecuado de regresión binaria para conocer la propensión de los clientes del banco A.M.A a aceptar una tarjeta de crédito es el enlace asimétrico Power Logit bayesiano.

El factor que más influye en los clientes del banco A.M.A. para que acepten una tarjeta de crédito es Cross Sell.

BIBLIOGRAFÍA

[2]Jorge Bazán, Cristian Bayes. (2010). *Inferencia bayesiana en modelos de regresión binaria usando BRMUW*

[3]Jorge Luis Bazán. *Manual de Uso de BRMUW.* Departamento de ciencias. Pontificia Universidad Católica del Perú

[4]*Wikipedia.* Función Gumbel

[5]*Wikipedia.* Función Burr

ANEXOS

I. JUSTIFICACIÓN EN ENLACES ASIMÉTRICOS EN LA REGRESIÓN BINARIA

Con la regresión binaria, se modela las probabilidades de una variable de respuesta que toma dos valores en función de otras variables explicativas considerando una función de enlace o *link*. Este enlace trata los datos dicotómicos como una variable de respuesta y explora su relación con otras variables explicativas, que son combinadas como un predictor lineal. Es decir, los modelos de regresión binaria estiman la probabilidad de éxito de uno de los valores de la variable respuesta como función de un conjunto de predictores o regresores considerando un enlace entre estas variables.

En la regresión binaria, los enlaces usados comúnmente son los enlaces probit y logit, que originan la regresión probit y la regresión logística, respectivamente. En ambos modelos (probit y logit), esta probabilidad tiene una forma simétrica de alrededor de 0.5.

Sin embargo, cuando hay probabilidades extremas, es decir, cuando hay presencia predominante de uno de los valores de la variable respuesta, los enlaces simétricos son inadecuados[2].

Aun cuando la regresión binaria se ha discutido en la literatura en los últimos cincuenta años bajo una visión frecuentista clásica, el enfoque bayesiano ha sido tratado solo recientemente.

Se ha demostrado que la metodología bayesiana es útil, especialmente bajo los métodos *Markov Chain Monte Carlo Methods* (MCMC)[2]. La aproximación bayesiana nos sirve para conocer la distribución a posteriori de los parámetros de los modelos y también para observar medidas alternativas de bondad de ajuste. Esto es útil para comparaciones entre modelos alternativos.

II. CLASIFICACION DE LOS ENLACES EN REGRESION BINARIA

Sea Y el vector de variables respuesta nx1.

$\beta = (\beta_1, \beta_2, \dots, \beta_P)$ el vector de parámetros. La matriz de datos Xnxp con filas x_i'donde $x'_i = (x_{i1}, x_{i2}, \dots, x_{ip})$es el vector de parámetros. Considere que y_i=1 es la probabilidad de éxito con probabilidad p_i y y_i=0 es la probabilidad de fracaso con probabilidad de (1- p_i)

En los modelos de datos binarios tenemos F(x) como la función de distribución acumulada (fda). Entonces:

$$Y_i \sim Bernoulli(p_i)$$

$$p_i = F(x_i'\beta)$$

2.1. MODELO BINARIO SIMETRICO

- **FUNCION DE ENLACE LOGIT**

Si F es simétrico entonces el enlace resultante es simétrico y pi tiene una forma simétrica alrededor de pi=0.5. En el caso de F sea la fda de una distribución logística obtenemos el enlace Logit: [2]

$$F(t) = \frac{e^t}{1 + e^t}$$

2.2. MODELO BINARIO ASIMETRICO

Ahora si la probabilidad de la respuesta binaria se aproxima a 0 en una tasa diferente que cuando se aproxima a 1, los enlaces simétricos para el ajuste de datos pueden ser inadecuados. En este caso, hay que considerar los enlaces asimétricos. Un ejemplo muy popular es el enlace log-log complementario o cloglog, donde la fda usada en el enlace corresponde a la distribución de Gumbel.

- **FUNCION DE ENLACE LOG-LOG COMPLEMENTARIO O CLOGLOG**

Distribución de Gumbel [4]

Es utilizada para modelar la distribución del máximo (o el mínimo), por lo que se usa para calcular valores extremos. La aplicabilidad potencial de la distribución de Gumbel para representar los máximos se debe a la teoría de valores extremos que indica que es probable que sea útil si la muestra de datos tiene una distribución normal o exponencial.

La función de distribución acumulada de Gumbel: [2]

$$F(x, \mu, \beta) = e^{-e^{-(x-\mu)/\beta}}$$

- **FUNCION DE ENLACE SCOBIT [5]**

Los modelos logit scobit están basados en la distribución Burr. Esta distribución también llamada distribución Singh-Maddala es una distribución continua para variables aleatorias no negativas. Es una de las tantas distribuciones logística generalizada y su mayor utilidad es para modelar los ingresos del hogar.

La distribución acumulada es la siguiente:

$$F(x_i'\beta) = 1 - (1 + e^{x_i'\beta})^{-\lambda}$$

➢ **FUNCION DE ENLACE POWER LOGIT[3]**

Los modelos con función de enlace Power Logit son utilizados para la creación de modelos con target de distribución asimétrica.

La distribución acumulada es la siguiente:

$$F(x_i'\beta) = (1 + e^{-x_i'\beta})^{-\lambda}$$

III. COMPARACIÓN DE MODELOS

Como hemos visto podemos considerar diferentes modelos para un conjunto de datos binarios con solo cambiar la función de enlace, en esta sección revisaremos diferentes criterios para la comparación de modelos que nos ayudaran a decidir qué modelo es más apropiado.

Existen una seria de metodologías para comparar modelos alternativos, entre los principales criterios para comparación de modelos en la inferencia bayesiana tenemos: (deviance information criterion) (DIC), el esperado del criterio de información de Akaike (EAIC) y el esperado del criterio de información de Schwarz o Bayesiano (EBIC). Además de esto, añadiremos el índice de GINI y la sensibiilidad.

- Los indicadores bayesianos son basados en media a posteriori del *desvío*$E[D(a,b,\lambda,\theta)]$, donde $D(a,b,\lambda,\theta)$es una medida de ajuste que puede ser aproximada utilizando la salida de la simulación MCMC de la distribución a posteriori.

$$E[D(a,b,\lambda,\theta)] = -2\ln(p(y|a,b,\lambda,\theta)$$
$$= -2\sum_{i=1}^{n} \ln P(Y_{ij} = y_{ij}|a,b,\lambda,\theta)$$

La aproximación de $E[D(a,b,\lambda,\theta)]$es dada por:

$$D_{bar} = \frac{1}{G}\sum_{i=1}^{G} D(a^g, b^g, \lambda^g, \theta^g)$$

Donde el índice g indica el g −ésimo valor simulado de un total de G simulaciones.

El EAIC, EBIC y DIC pueden ser estimados de la siguiente manera

$$\widehat{EAIC} = Dbar + 2p$$
$$\widehat{EBIC} = Dbar + plogN$$

y

$$\widehat{DIC} = Dbar + \widehat{\rho_D} = 2\,Dbar - Dhat$$

Respectivamente donde *p* es el número de parámetros en el modelo, *N* es el total de observaciones y ρ_D es el *número efectivo de parámetros* y es definido como

$\rho_D = E[D(a,b,\lambda,\theta)] - D[E(a),E(b),E(\lambda),E(\theta)]$

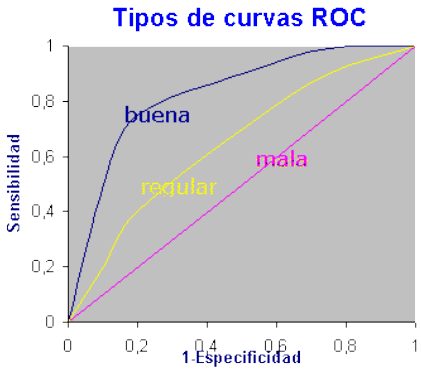
Donde $D[E(a),E(b),E(\lambda),E(\theta)]$ es el *desvio de la media a posteriori* obtenido cuando evaluamos la función desvío en la media a posteriori de los parámetros, el cual es estimado por

$$D_{hat} = D(\frac{1}{G}\sum_{i=1}^{G} a^g, \frac{1}{G}\sum_{i=1}^{G} b^g, \frac{1}{G}\sum_{i=1}^{G} \lambda^g, \frac{1}{G}\sum_{i=1}^{G} \theta^g)$$

Para comparar dos o más modelos alternativos, el modelo que presente mejor ajuste al conjunto de datos será el modelo que presente el menor valor de DIC, EAIC y EBIC.

- **Gini y sensibilidad**

Una curva ROC (Receiver Operating Characteristic, o Característica Operativa del Receptor) es una representación gráfica de la sensibilidad en función de los falsos positivos (complementario de la especificidad) para distintos puntos de corte. Un parámetro para evaluar la bondad de la prueba es el área bajo la curva que tomará valores entre 1 (prueba perfecta) y 0,5 (prueba inútil).



El índice de Gini es el indicador discriminatorio usado en el presente trabajo. Este valor es una transformación del valor de ROC, pues se cumple la siguiente relación: $\text{Gini} = 2 * (\text{ROC} - 0.5)$. De acuerdo a la clasificación del valor obtenido en el Gini podemos saber qué tan adecuado son los modelos construidos:

Valor Gini	Clasificación
0,00 - 0,25	Bajo
0,25 - 0,45	Aceptable
0,45 - 0,60	Bueno
0,60 - 1,00	Muy bueno

Otro indicador que nos ayuda a tener una visión más clara de la precisión del modelo es la sensibilidad. Esta se puede apreciar en una tabla de clasificación en donde se contrastan las cantidades de buenos y malos de acuerdo al modelo y la data real tomando en cuenta una data de validación.

Tabla de Clasificación			
Valid.		Real	
		1	0
Modelo	1	VP	FP
	0	FN	VN

$$Sensibilidad = \frac{VP}{VP + FN}$$

$$Especificidad = \frac{VN}{VN + FP}$$

AGRADECIMIENTOS ESPECIALES

Damos gracias de manera especial a aquellas personas que nos ayudaron a llevar este trabajo al éxito.

- Jorge Luis Bazán, jefe del área académica de la PUCP
- Omar Chíncharo, profesor de la Universidad Nacional de Ingeniería
- Richard Fernández, profesor de la Universidad Nacional de Ingeniería