Modelo para otorgar tasas y montos a clientes con tarjeta de crédito en una cooperativa de ahorro y crédito.

Sergio Armando Serrano Becerra, Luis Adolfo Caballero Villamizar

Proyecto de grado para optar al título de Especialista en Estadística

Director,

Carlos Mantilla

Magister en Estadística

Universidad Industrial de Santander

Facultad de Ciencias

Escuela de Matemáticas

Especialización en Estadística

Bucaramanga

2020

Agradecimientos

A Dios por la oportunidad poder vivir una nueva experiencia estudiantil y poder haber tenido la salud para finalizar la especialización.

A nuestras familias por todo el apoyo recibido durante los periodos cursados.

A nuestra empresa Financiera Comultrasan por la oportunidad de capacitar el talento humano de sus empleados.

A todos los profesores que desde su experiencia hicieron parte de este maravilloso ciclo.

Contenido

Pág.

Lista de Figuras

Pág.

Lista de Tablas

Pág.

Resumen

Título: Modelo para otorgar tasas y montos a clientes con tarjeta de crédito en una cooperativa de ahorro y crédito[*]

Autores: Sergio Armando Serrano Becerra y Luis Adolfo Caballero Villamizar[**]

Palabras claves: Tarjeta Crédito, Modelo Logit

Descripción:

Se realizó un estudio estadístico cuyo propósito era generar valor en un producto relativamente nuevo para una cooperativa de ahorro y crédito, como lo es la tarjeta crédito y el cual tienen un comportamiento diferente a un crédito tradicional, para lo cual se aplicó un análisis descriptivo de correspondencia para revisar cuales variables se relacionaban más con el incumplimiento del no pago de la tarjeta y de esta manera poder realizar un modelo de regresión logística que permitirá calcular la probabilidad que tiene un asociado con tarjeta habiente para incurrir en el no pago de sus obligaciones.

Estos modelos fueron desarrollados en el lenguaje de desarrollo estadístico R y como resultado aplicado no se obtuvo una clasificación alta, sin embargo, es importantes resaltar que un modelo de probabilidad para entrar en mora crediticia puede integrarse con gran acogida en diferentes campañas o asignaciones de tasas e incluso en segmentaciones que se hagan o se tenga establecidas en una entidad financiera.

Por lo tanto, se concluye que estos modelos pueden ser de gran ayuda, teniendo una mayor información histórica y más variables que se relacionen con la mora, para poder obtener un mejor entrenamiento del modelo y este permita una clasificación más alta.

[*] Trabajo de grado

[**] Facultad De Ciencias. Escuela De Matemáticas. Especialización en Estadistica Director: Carlos Alfonso Mantilla Duarte.

Summary

Title: Model for granting rates and amounts to credit card customers in a savings and credit cooperative[*]

Authors: Sergio Armando Serrano Becerra y Luis Adolfo Caballero Villamizar[**]

Key Words: Credit Card, Logit Model

Description:

A statistical study was carried out whose purpose was to generate value in a relatively new product for a savings and credit cooperative, such as the credit card and which has a different behavior from a traditional credit, for which a descriptive analysis was applied of correspondence to review which variables were more related to non-payment of the card and in this way to be able to carry out a logistic regression model that will be able to calculate the probability that an associate with a cardholder has to incur in non-payment of their obligations.

These models were developed in the statistical development language R and as an applied result a high classification was not obtained, however, it is important to highlight that a probability model to enter into credit default can be integrated with great acceptance in different campaigns or rate assignments and even in segmentations that are made or have established in a financial institution.

Therefore, it is concluded that these models can be of great help, having more historical information and more variables that are related to delinquency, in order to obtain a better training of the model and this allow a higher classification.

[*] Degree Work

[**] Facultad De Ciencias. Escuela De Matemáticas. Especialización en Estadistica Director: Carlos Alfonso Mantilla Duarte

Introducción

Las tarjetas de crédito hoy en día son un producto ofrecido por la gran mayoría de entidades bancarias y estas generan comportamientos diferentes a un crédito tradicional, dado que parten de un valor desembolsado en un plástico o tarjeta que le permite realizar transacciones en: cajeros, establecimientos, comercios electrónicos o cajeros automáticos y dada la facilidad de generar varias transacciones en diferentes lugares, su rentabilidad es mucho mejor que un crédito, pero también el riesgo de tener un cliente en mora es bastante alto.

Para las entidades financieras es de gran importancia conocer mejor a sus clientes y transformar la información de cada uno de ellos en propuesta de valor que les permitan brindar mejores tarifas, montos o cualquier beneficio; por tal razón en FINANCIERA COMULTRASAN una cooperativa de ahorro y crédito quien tiene en su portafolio el producto de tarjeta de crédito desde el año 2018, desea ofrecer mejores tasas y montos, basándose en un modelo estadístico que le permita conocer el comportamiento de sus asociados habientes y establecer campañas de marketing que permitan mejorar la estrategia obteniendo la probabilidad de riesgo en mora que podría tener cada de sus asociados.

1. Planteamiento del Problema

Las estrategias de mercadeo implementadas por la cooperativa de ahorro y crédito hacia sus clientes con tarjeta de crédito se realizan a través de análisis de datos muy superficiales, que arrojan mercados subjetivos, con poca intensión de compra y riesgo crediticio alto, convirtiendo estas estrategias de mercado en poco eficientes, aumentando la incertidumbre del nivel de riesgo crediticio y ofreciendo los mismos incentivos que la competencia otorga.

La finalidad de este trabajo es desarrollar un modelo de clasificación de clientes con tarjeta de crédito, según su hábito de pago, hábito de consumo y nivel de riesgo crediticio, de tal manera que permita generar estrategias de mercadeo que maximicen la rentabilidad del producto con el menor riesgo posible

2. Objetivos

2.1 Objetivo General

Diseñar un modelo estadístico de clasificación de los asociados con tarjeta crédito de FINANCIERA COMULTRASAN, que permita otorgar tasa y monto, de acuerdo con su hábito de pago, hábito de consumo y nivel de riesgo crediticio.

2.2 Objetivos Específicos

- Elegir una técnica estadística para diseñar el modelo planteado.
- Seleccionar las variables que utilizará al modelo estadístico.
- Normalizar el conjunto de datos para al modelo estadístico planteado.
- Codificar en lenguaje R el modelo estadístico planteado.

3. Marco histórico de la información

El horizonte de tiempo de la información seleccionada para este trabajo comprende el historial de los años 2018, 2019 y marzo de 2020 de las compras, pagos, información financiera y datos sociodemográficos, de los clientes con tarjeta de crédito de una cooperativa de ahorro y crédito

4. Selección de la muestra

La base de clientes con tarjeta de crédito de la cooperativa de ahorro y crédito está conformada por 19767 registros con información socioeconómica, comportamiento de compra, habito de pago y monto del cupo otorgado a su tarjeta de crédito.

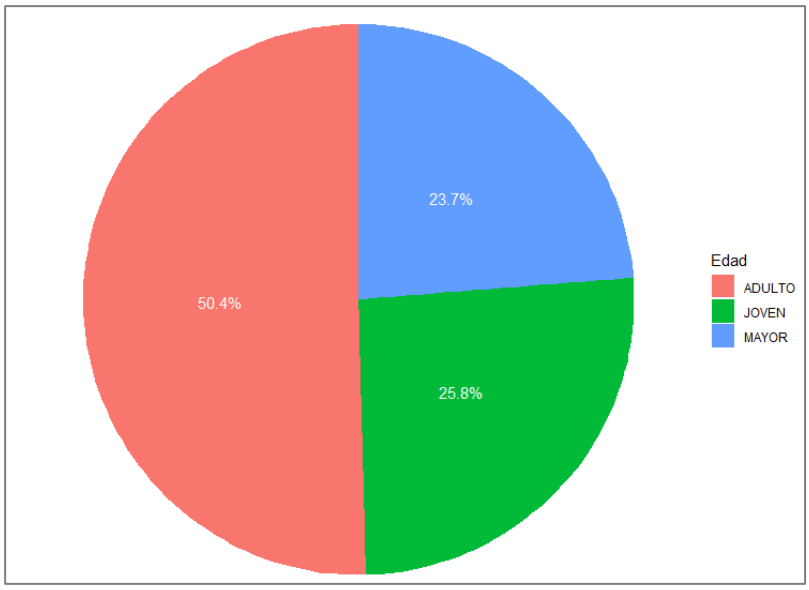
Luego de hacer una reclasificación de variables y posteriormente una limpieza de datos, se obtuvo una base final de 8630 registros útiles con seis variables categóricas de interés

5. Descripción de variables

La gráfica 1 muestra la distribución de los clientes según la edad. La variable se categorizó en 3 niveles, clientes entre los 36 y 55 años representan la población adulta, equivalente al 50.4% del total de los clientes, el nivel jóvenes está comprendido por los clientes entre los 18 y 34 años, con una representación del 25.8% del total de clientes y con un 23.7% de representación de los clientes, el nivel mayor agrupa a los clientes mayores de 55 años.

Figura 1.

Distribución por edad

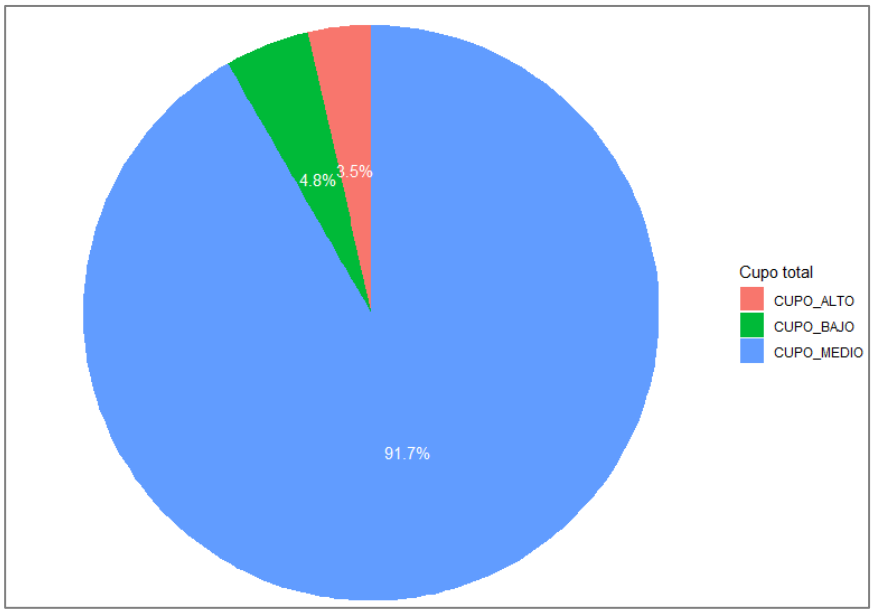


La gráfica 2 muestra el valor del cupo total asignado a la tarjeta de crédito de cada cliente, se transformó a una variable categórica de 3 niveles así: BAJO: cupo entre 1.2 y 2 SMMLV,

MEDIO: cupo entre 2 y 3 SMMLV y ALTO: entre 3 SMMLV y 10.000.000 millones de pesos. Luego de la reclasificación, la población es representada en un cupo medio con el 91.7%.

Figura 2.

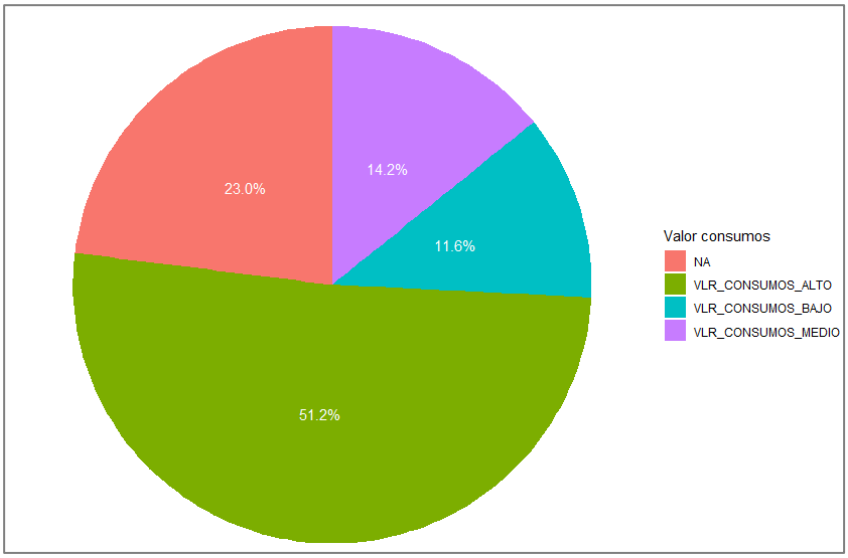
Distribución de cupo



La grafica 3 representa el valor de los consumos (compras y/o avances) realizados con la tarjeta de crédito y clasifica la población en 4 categorías así: NA: Corresponden a los clientes que no han realizado compras y los que han realizado compras superiores a 10.000.000 de pesos, BAJO: consumos (compras y/o avances) realizados por valores iguales o menores a 1.000.000 de pesos, MEDIO: consumos realizados por valores entre 1.000.001 y 2.000.000 de pesos y ALTO: los consumos realizados por valores entre 2.000.001 y 10.000.000 de pesos. Al realizar esta clasificación se reagrupan los clientes en un 51.2% con consumos altos, el 25.8% en clientes con consumos medios y altos y el 23.02% restante clasificados en NA.

Figura 3.

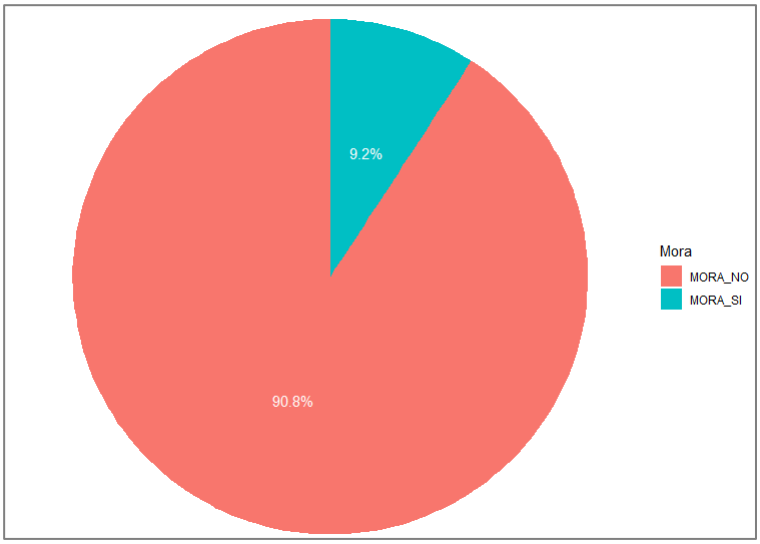
Valor consumos



La grafica 4 representa los clientes según los días de mora que han presentado en su tarjeta de crédito, esta viable se transformó a dicotómica, donde NO equivale a los clientes que tiene 0 días de mora y SI para los que tienen 1 o más días de mora, en el cual el 90.8% de los clientes no presentan mora y el 9.2% restante ha tenido algún día de mora.

Figura 4.

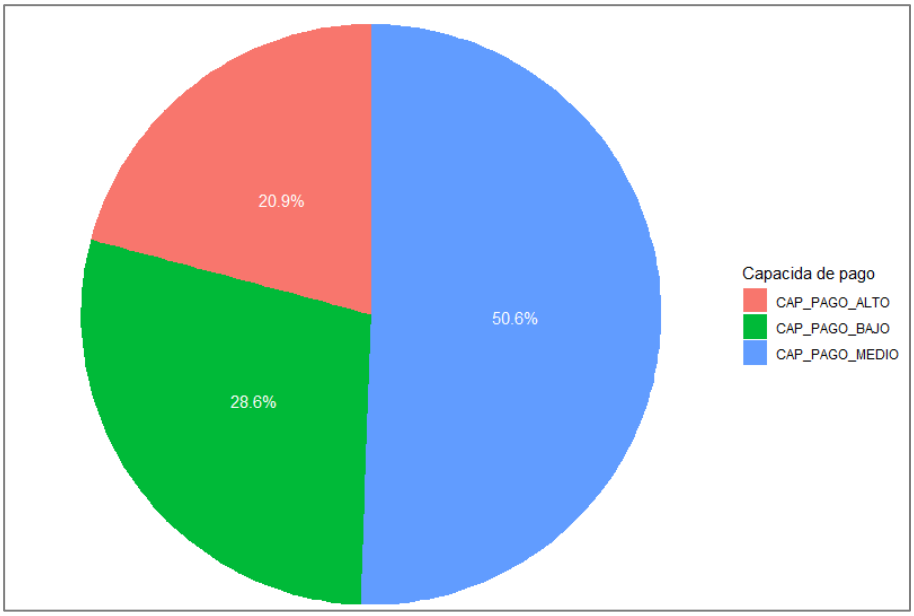
Mora de la tarjeta de crédito



La grafica 5 representa la capacidad de pago que tiene un cliente, valor que resulta luego de hacer una diferencia entre su valor de ingresos y el valor de sus egresos. Se convirtió en una variable politómica de 3 niveles, donde el 50.6% de los clientes tienen una capacidad de pago media, el 28.6% cuenta con una capacidad de pago baja y para el 20.9% restante de los clientes su capacidad de pago es alta.

Figura 5.

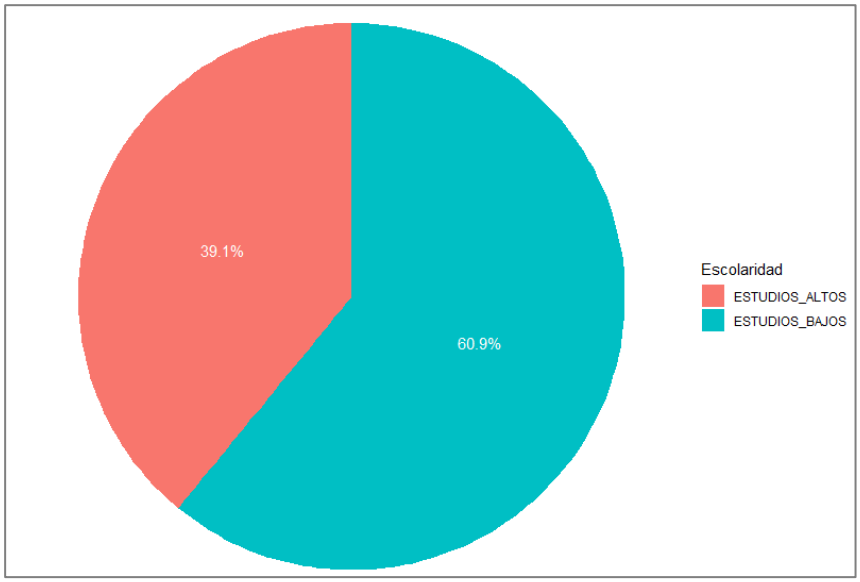
Capacidad de pago



La figura 6 representa el nivel de estudios de los clientes, variable que fue reclasificada en 2 niveles: clientes con estudios bajos que equivale el 60.9% y el 39.1% restante corresponden a clientes con estudio altos.

Figura 6.

Escolaridad



La tabla 1 muestra un resumen general de las variables objeto de estudio.

Tabla 1.

Descripción de variables

VARIABLE	VARIABLE EN MODELO	TIPO	DESCRIPCIÓN GENERAL	DESCRIPCION NIVELES
edad	EDAD	factor	Edad del cliente	joven: 18 - 35 años adulto: 36 - 55 años mayor: >55 años
cupo total	CUPO_TOTAL	factor	Valor de cupo a utilizar en la tarjeta de crédito	bajo: 1 - 1.2 SMMLV medio: 2 - 3 SMMLV alto: 3 SMMLV y $10.000.000 millones
valor consumos	VALOR_CONSUMOS	factor	valor de las compras o avances de la tarjeta de credito	bajo: <=1.000.000 millones medio: 1.000.001 - 2.000.0000 millones alto: 2.000.001 - 10.000.0000 millones na: 0 pesos ó >10.000.000 millones
dias mora	DIASMORA	factor	número de días de mora que presenta el cliente	no: 0 días de mora si: >=1 día de mora
capacidad pago	CAPACIDAD_PAGO	factor	Valor que resulta de la diferencia del valor de los ingresos y los egresos del cliente	bajo: 1.000.000 - 1.500.001 millones medio: 1.500.001 - 3.000.000 millones alto: 3.000.001 - 5.000.000 millones
nivel de estudios	ESCOLARIDAD	factor	Nivel de estudios del cliente	bajos: Primaria, Secundario, Tecnico, Tecnólogo altos: Universidad, Posgrado

6. Marco Teórico

6.1 Análisis de correspondencia

Uno de los objetivos del análisis de correspondencias es describir las relaciones existentes entre dos variables nominales, recogidas en una tabla de correspondencias, sobre un espacio de pocas dimensiones, mientras que al mismo tiempo se describen las relaciones entre las categorías de cada variable. Para cada variable, las distancias sobre un gráfico entre los puntos de categorías reflejan las relaciones entre las categorías, con las categorías similares representadas próximas unas a otras. La proyección de los puntos de una variable sobre el vector desde el origen hasta un punto de categoría de la otra variable describe la relación entre ambas variables.

El análisis de las tablas de contingencia a menudo incluye examinar los perfiles de fila y de columna, así como contrastar la independencia a través del estadístico de chi-cuadrado. Sin embargo, el número de perfiles puede ser bastante grande y la prueba de chi-cuadrado no revelará la estructura de la dependencia. El procedimiento Tablas cruzadas ofrece varias medidas y pruebas de asociación, pero no puede representar gráficamente ninguna relación entre las variables.

El análisis factorial es una técnica estándar para describir las relaciones existentes entre variables en un espacio de pocas dimensiones. Sin embargo, el análisis factorial requiere datos de intervalo y el número de observaciones debe ser cinco veces el número de variables. Por su parte, el análisis de correspondencias asume que las variables son nominales y permite describir las relaciones entre las categorías de cada variable, así como la relación entre las variables. Además,

el análisis de correspondencias se puede utilizar para analizar cualquier tabla de medidas de correspondencia que sean positivas.

Dependencia e independencia de una tabla de correspondencia

La existencia o no de algún tipo de relación entre las variables X e Y se analiza mediante contrastes de hipótesis sobre la independencia de dichas variables. El test de hipótesis habitualmente utilizado es el de la Chi-cuadrado de Pearson. Se contrasta la hipótesis nula que presupone la independencia entre ambas variables, mediante el estadístico $\chi 2$ de Pearson.

H0: ambas variables son independientes

H1: existe una relación de dependencia

El test se basa en comparar los perfiles fila y columna con los perfiles marginales correspondientes, considerando que si H0 es cierta todos los perfiles fila (respecto columna) son iguales entre sí e iguales al perfil marginal de X (respecto de Y).

Compara los valores del estadístico que define los términos observados n_{ij} con los valores esperados (i,j):

$$\sum_{i=1}^{k}\sum_{j=1}^{m}\frac{(n_{ij}-e_{ij})^2}{e_{ij}}=\chi^2_{(k-1).(m-1)}$$

Siendo:

$$e_{ij}=E\left[n_{ij}/H_0 \text{ es cierta}\right]=\frac{N_{i\bullet}.N_{\bullet j}}{N_{\bullet\bullet}}$$

el estadístico observado se puede expresar también:

$$\sum_{i=1}^{k}\sum_{j=1}^{m}\frac{(n_{ij}-e_{ij})^2}{e_{ij}}=\sum_{i=1}^{k}\sum_{j=1}^{m}\frac{N_{i\bullet}\left[\frac{n_{ij}}{N_{i\bullet}}-\frac{N_{\bullet j}}{N_{\bullet\bullet}}\right]^2}{\frac{N_{\bullet j}}{N_{\bullet\bullet}}}=\sum_{i=1}^{k}\sum_{j=1}^{m}\frac{N_{\bullet j}\left[\frac{n_{ij}}{N_{\bullet j}}-\frac{N_{i\bullet}}{N_{\bullet\bullet}}\right]^2}{\frac{N_{i\bullet}}{N_{\bullet\bullet}}}$$

La región crítica para el contraste de independencia se determina:

$$P\left[\chi^2_{(k-1).(m-1)}\geq k/H_0\right]=\alpha$$

Así, pues, para un nivel de significación α:

$$\begin{cases}\chi^2_{(k-1).(m-1)}<\chi^2_{\alpha;(k-1).(m-1)} & \Rightarrow \text{ X e Y son independientes al nivel } \alpha\\ \chi^2_{(k-1).(m-1)}\geq\chi^2_{\alpha;(k-1).(m-1)} & \Rightarrow \text{ X e Y no son independientes al nivel } \alpha\end{cases}$$

$$\begin{cases}\text{Sig.asintótica}(p\_\text{value})\leq 0{,}05 & \Rightarrow \text{ Se rechaza } H_0\\ \text{Sig.asintótica}(p\_\text{value})>0{,}05 & \Rightarrow \text{ Se acepta } H_0\end{cases}$$

Si la hipótesis nula se rechaza, las variables X e Y son dependientes. En este caso conviene analizar los perfiles condicionales fila y columna así como los residuos del modelo para estudiar qué tipo de dependencia existe entre ellas. (IBM, 2020)

7. Desarrollo del análisis de correspondencia múltiple

Es una técnica de análisis multivariante que tiene como objetivo principal obtener una representación gráfica que permita visualizar las relaciones existentes entre las variables de estudio y sus diversas categorías.

También se considera como una extensión del análisis de correspondencia simple aplicado a más de dos variables categóricas, puede verse como una generalización del análisis de componentes principales cuando las variables a analizar son categóricas en lugar de cuantitativas (Abdi y Williams 2010).

7.1 Test de independencia(X2)

El test $\chi 2$ de independencia, también conocido como $\chi 2$ de Pearson se emplea para estudiar si existe asociación entre dos variables categóricas, es decir si las proporciones de una variable son diferentes dependiendo del valor que adquiera la otra variable, cuando los datos son independientes. Se trata por lo tanto de una expansión del Z-test para dos proporciones cuando una de las variables estudiadas tiene dos o más niveles. Cuando ambas variables tienen dos niveles (tabla 2x2) ambos test $\chi 2$ goodness of fit y Z-test para dos proporción son equivalentes.

Es importante tener en cuenta que cuando el número de observaciones esperadas para alguno de los niveles es igual o menor a 5 la aproximación por el test $\chi 2$ no es buena, se debe usar una prueba que no incluya aproximaciones, como la prueba exacta de Fisher.

El test de independencia cuantifica y sumariza que tan distinto es el número de eventos observados en cada nivel con respecto al número esperado acorde con Ho. Esto permite contrastar si las diferencias observadas entre los grupos son atribuibles al azar.

7.2 Estadistico X2

$$\chi^2 = \sum_{i,j} \frac{(\text{observado}_{ij} - \text{esperado}_{ij})^2}{\text{esperado}_{ij}}$$

El valor esperado de cada grupo se obtiene multiplicando las frecuencias marginales de la fila y columna en la que se encuentra la celda y dividiendo por el total de observaciones. Se suman las diferencias de todos los niveles. Elevar al cuadrado las diferencias permite hacerlas todas positivas y además magnificar aquellas más grandes. (Joaquin, 2017)

7.3 Hipótesis

H0 : Las variables son independientes por lo que una variable no varía entre los distintos niveles de la otra variable.

Ha: Las variables son dependientes, una variable varía entre los distintos niveles de la otra variable.

La prueba de hipótesis permitirá identificar si la variable DIASMORA varía entre los niveles de las variables EDAD_ASOCIADO, CUPO_TOTAL, VALOR_CONSUMOS, ESCOLARIDAD y CAPACIDAD_PAGO, por lo cual se deberá rechazar o aceptar las siguientes hipótesis:

$H0_1$: Clientes en mora y la edad son independientes, el % de clientes en mora no varía entre los diferentes niveles de la variable EDAD_ASOCIADO.

Ha_1: Clientes en mora y la edad son dependientes, el % de clientes en mora varía entre los diferentes niveles de la variable EDAD_ASOCIADO.

$H0_2$: Clientes en mora y el valor del cupo de tarjeta de crédito son independientes, el % de clientes en mora no varía entre los diferentes niveles de la variable CUPO_TOTAL.

Ha_2: Clientes en mora y el valor del cupo de tarjeta de crédito son dependientes, el % de clientes en mora varía entre los diferentes niveles de la variable CUPO_TOTAL.

$H0_3$: Clientes en mora y el valor de los consumos de tarjeta de crédito son independientes, el % de clientes en mora no varía entre los diferentes niveles de la variable VALOR_CONSUMOS.

Ha_3: Clientes en mora y el valor de los consumos de tarjeta de crédito son dependientes, el % de clientes en mora varía entre los diferentes niveles de la variable VALOR_CONSUMOS.

$H0_4$: Clientes en mora y el nivel de escolaridad son independientes, el % de clientes en mora no varía entre los diferentes niveles de la variable ESCOLARIDAD.

Ha_4: Clientes en mora y el nivel de escolaridad son dependientes, el % de clientes en mora varía entre los diferentes niveles de la variable ESCOLARIDAD.

$H0_5$: Clientes en mora y la capacidad de pago del cliente son independientes, el % de clientes en mora no varía entre los diferentes niveles de la variable CAPACIDAD_PAGO.

Ha_5: Clientes en mora y la capacidad de pago del cliente son dependientes, el % de clientes en mora varía entre los diferentes niveles de la variable CAPACIDAD_PAGO.

En la tabla 2 se encuentran almacenados los resultados de las pruebas X2, en donde las variables ESCOLARIDAD y CAPACIDAD_PAGO con un p-valor de p-value(2.841635e-01) y p-value(1.127698e-01) respectivamente son significativas, lo cual indica que el estado de la mora de los clientes es independiente de la variable escolaridad y la capacidad de pago. Sin embargo estas se incluirán en el MCA debido a que se consideran importantes y en conjunto contribuirían a explicar la variabilidad total de los datos.

Tabla 2.

Pruebas Chi Cuadrado

	VARIABLES	X-squared	df	p-value
1	DIASMORA EDAD_ASOCIADO	32.111228	2	1.064475e-07
2	DIASMORA CUPO_TOTAL	5742.943221	1	0.000000e+00
3	DIASMORA VALOR_CONSUMOS	15.237287	2	4.912076e-04
4	DIASMORA ESCOLARIDAD	1.147069	1	2.841635e-01
5	DIASMORA CAPACIDAD_PAGO	4.364813	2	1.127698e-01

7.4 Visualización e interpretación del modelo

El conjunto de datos analizados contiene características de los clientes con tarjeta de crédito de una cooperativa de ahorro y crédito, representadas en 6 variables (edad, valor del cupo, valor de consumos, días de mora, escolaridad y capacidad de pago).

La variabilidad total de los datos estudiados está representada en 8 dimensiones y cada una de estas almacena un porcentaje de dicha varianza. Las dimensiones 1, 2 y 3 suponen el 42.61% de la inercia total de los datos, los cuales permitirá estudiar a través de un plano de 2 dimensiones, las asociaciones entre las categorías de cada variable y principalmente con la variable de interés días mora.

La figura 7 muestra la varianza total de los datos extraída en 8 dimensiones, ordenadas de forma descendente de acuerdo al porcentaje de representación de la varianza total.

Figura 7.

Varianza explicada por las dimensiones

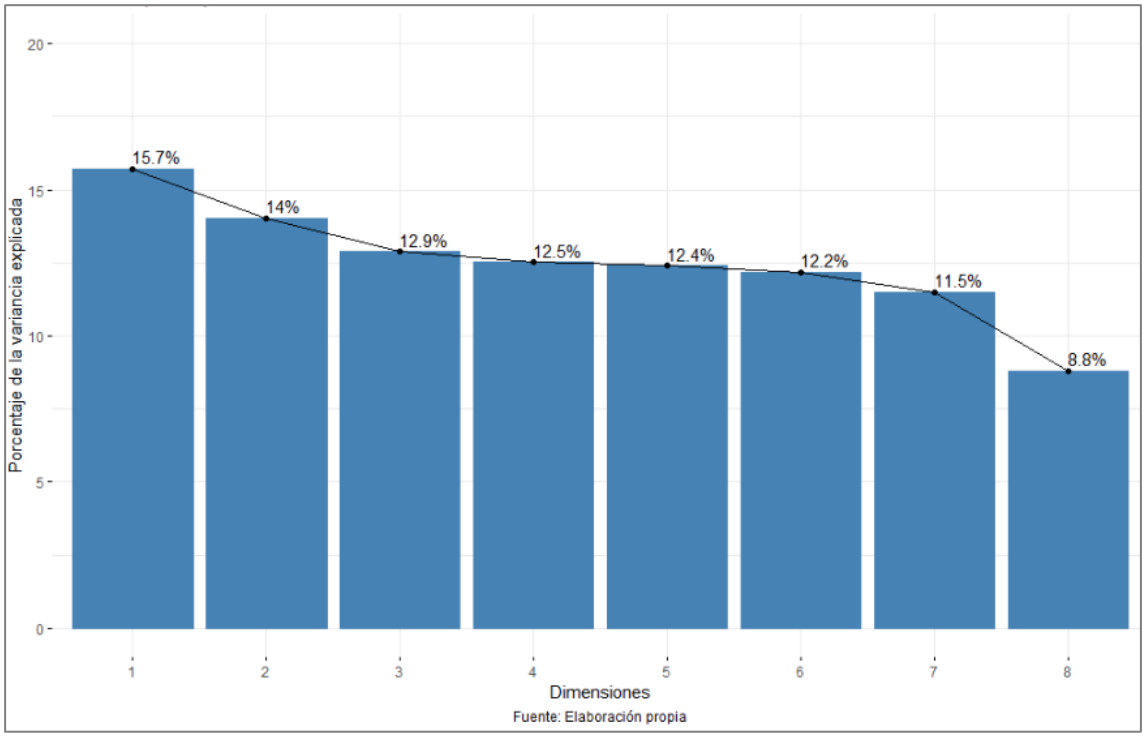


Tabla 3.

Porcentaje varianza explicada por cada dimensión

	eigenvalue	variance.percent	cumulative.variance.percent
Dim.1	0.2096544	15.724078	15.72408
Dim.2	0.1868016	14.010118	29.73420
Dim.3	0.1717720	12.882900	42.61710
Dim.4	0.1669381	12.520355	55.13745
Dim.5	0.1652609	12.394569	67.53202
Dim.6	0.1623708	12.177811	79.70983
Dim.7	0.1530395	11.477966	91.18780
Dim.8	0.1174960	8.812204	100.00000

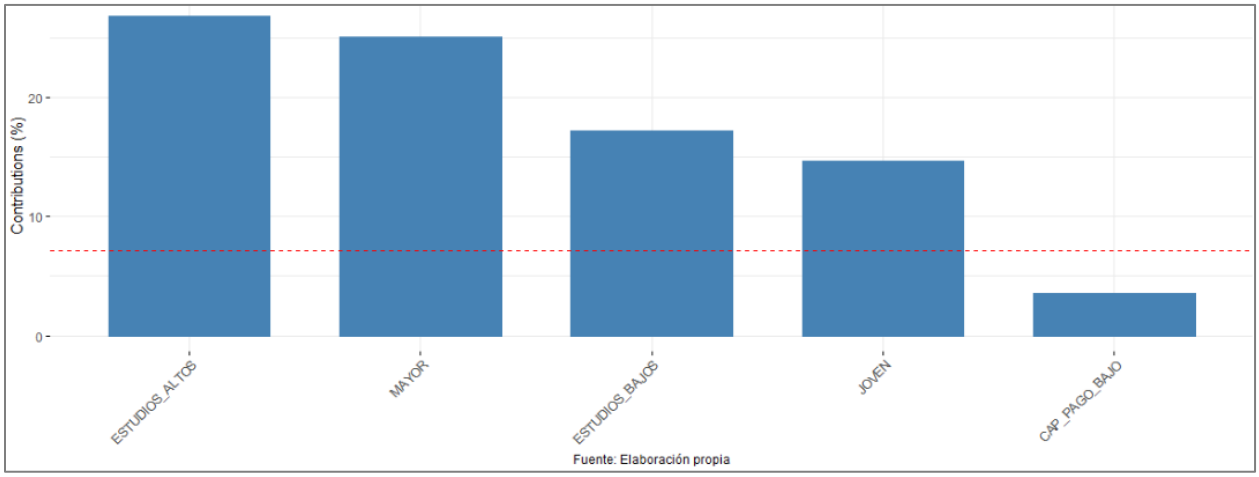
Las asociaciones que se puedan encontrar en al análisis de correspondencia múltiple dependen del grado de contribución y calidad de representación que una categoría de variable represente en la dimensión. Por lo cual se describen a continuación las categorías que son mejor

representadas en las dimensiones 1,2 y 3 las cuales explican el 42.61% de la inercia total de los datos.

La grafica 8 muestra la categoría estudios altos con un valor de 26.81 y los clientes mayores con un valor de 25.08, convirtiéndolas en las variables que más contribuyen a la inercia total de la dimensión 1 y a la explicación de la misma. La línea punteada indica el valor promedio esperado si las contribuciones fueran uniformes y marca un punto de división entre las variables más representativas y las menos representativas de la dimensión 1.

Figura 8.

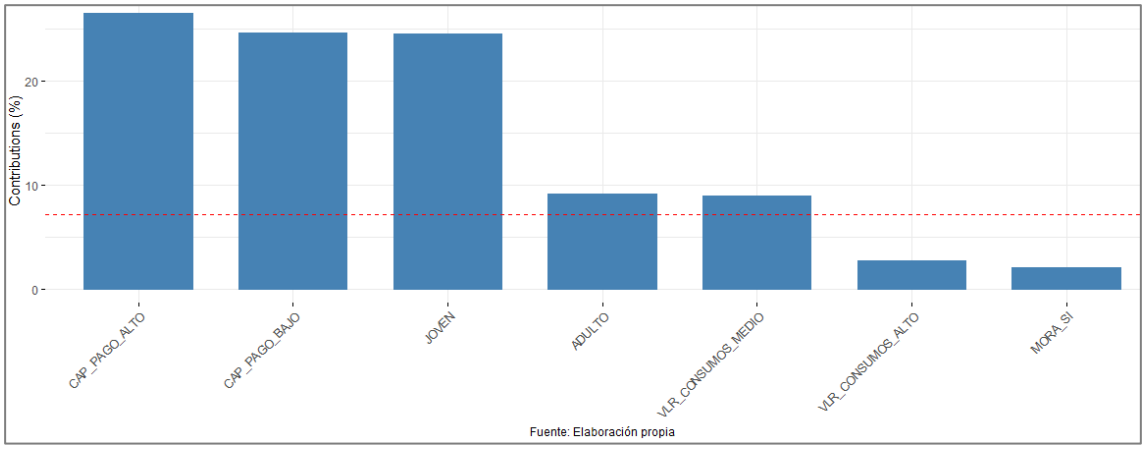
Contribución de variables a la dimensión 1



En la gráfica 9 las variables capacidad de pago alto, capacidad de pago bajo y los jóvenes son los que aportan el mayor porcentaje a la explicación de la dimensión 2, permitiendo concluir que esta representa la capacidad de pago de los clientes jóvenes. La línea punteada indica el valor promedio esperado si las contribuciones fueran uniformes y marca un punto de división entre las variables más representativas y las menos representativas de la dimensión 2

Figura 9.

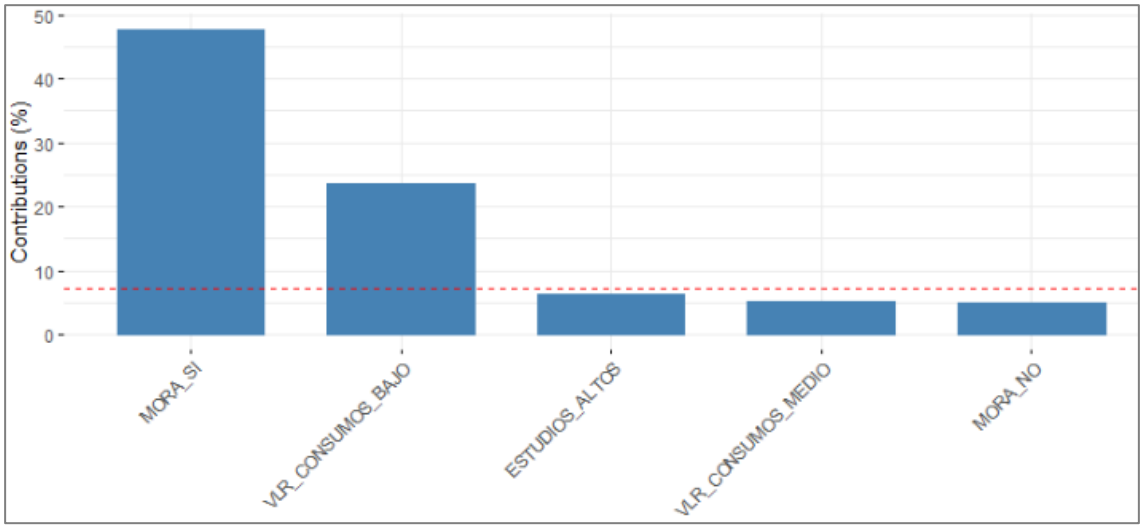
Contribución de variables a la dimensión 2



En la gráfica 10 las variables de interés mora y el valor de consumos bajo, son las que aportan el mayor porcentaje a la explicación de la dimensión 3, permitiendo concluir que esta representa la mora y los consumos de la tarjeta de crédito. La línea punteada indica el valor promedio esperado si las contribuciones fueran uniformes y marca un punto de división entre las variables más representativas y las menos representativas de la dimensión 3

Figura 10.

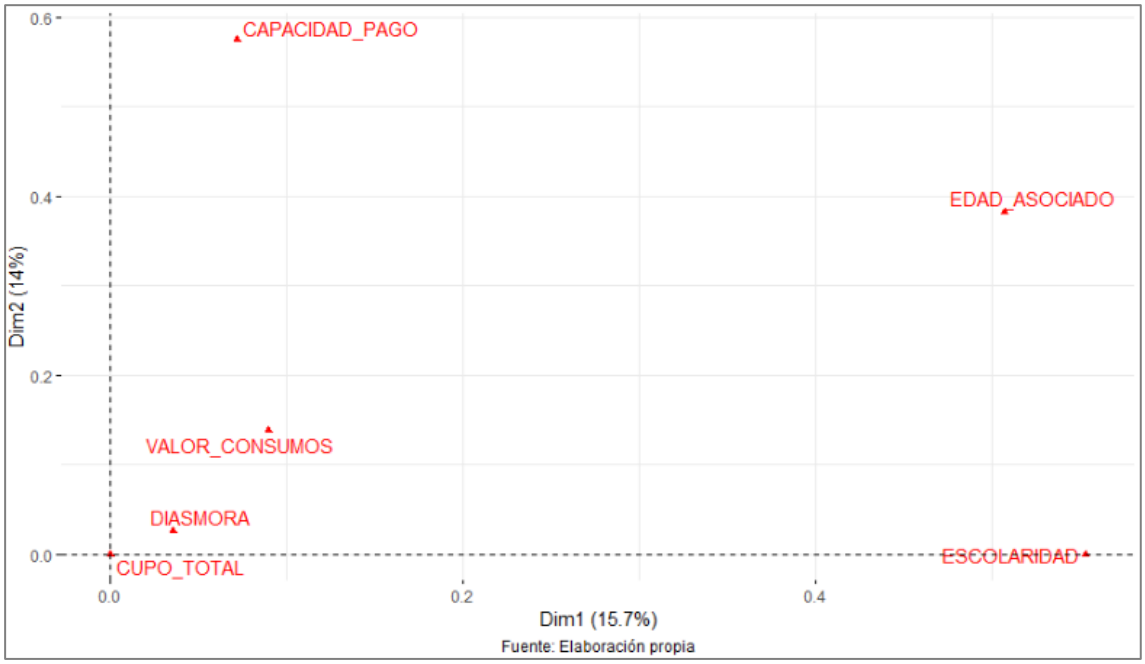
Contribución de variables a la dimensión 3.



La grafica 11 muestra en un plano de dos dimensiones, las variables mejor explicadas por la dimensión 1 y la dimensión 2, en donde la CAPACIDAD_PAGO Y ESCOLARIDAD son las mejor explicadas por la dimensión 1 y la dimensión 2 respectivamente debido a que son las más cercanas al eje de cada dimensión y las más alejadas del centro de gravedad de la nube de datos, la variable EDAD_ASOCIADO es la mejor representada por las 2 dimensiones debido a que tiene los valores de coordenadas más altos en X y Y.

Figura 11.

Representación de variables en las dimensiones 1 y 2



La grafica 12, 13 y 14 muestran en un plano de dos dimensiones las asociaciones y el nivel de explicación de las variables objeto de estudio, en el cual se pueden identificar asociaciones importantes donde se destacan las siguientes:

- El valor de los consumos medios está fuertemente relacionado con los clientes de capacidad de pago baja, ver gráfica 12. Asociación que se vuelve más fuerte en la gráfica 14 en donde los clientes jóvenes complementan la relación. Este comportamiento puede ser explicado por los jóvenes que tienden a consumir más a través del crédito.

- Los clientes mayores con estudios bajos tienden a realizar consumos bajos (menores a 1.000.000 de pesos), podría indicar que gracias a su edad prefieren usar el efectivo y su nivel de endeudamiento en el sector real es más concienzudo. Ver figura 12.

- La mora que es nuestra variable de interés pareciera tener un poco de relación con los clientes con estudios altos o con las personas adultas.
- Los clientes jóvenes tienden a tener estudios altos, comportamiento interesante debido a que el nicho de mercado de la cooperativa está compuesto principalmente de clientes de estratos bajos que generalmente su nivel de educación es bajo. Ver figura 12.

La tabla 4 resume las asociaciones más fuertes con la variable de interés (mora_si), luego de realizar una segmentación por departamentos con mayor representación de clientes con tarjeta de crédito, como lo son Santander, Cundinamarca y norte de Santander, se destacan las siguientes relaciones:

- En los departamentos de Santander y Cundinamarca existe una clara relación entre la mora y los clientes con estudios altos, sin embargo, los adultos santandereanos tienden a no cancelar sus obligaciones a tiempo, comportamiento que también se refleja en los jóvenes cundinamarqueses.
- La mora del norte santandereano se encuentra asociada a los adultos con consumos altos de su tarjeta de crédito.
- La grafica 13 y 14 permite identificar que la categoría de interés: MORA_SI, se encuentra alejada de las demás categorías, con lo cual se podría inferir que las asociaciones encontradas anteriormente no son muy fuertes y no pueden ser generalizadas sobre todo el conjunto de datos estudiado.

Figura 12.

Contribuciones dimensión 1 y dimensión 2

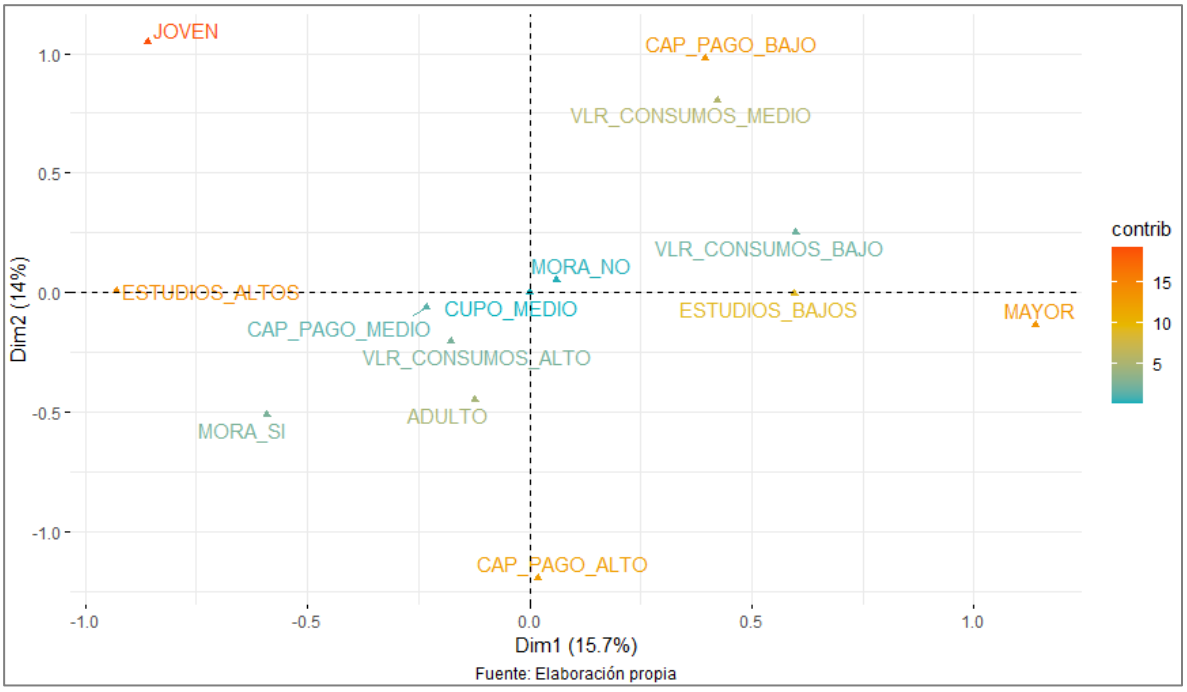


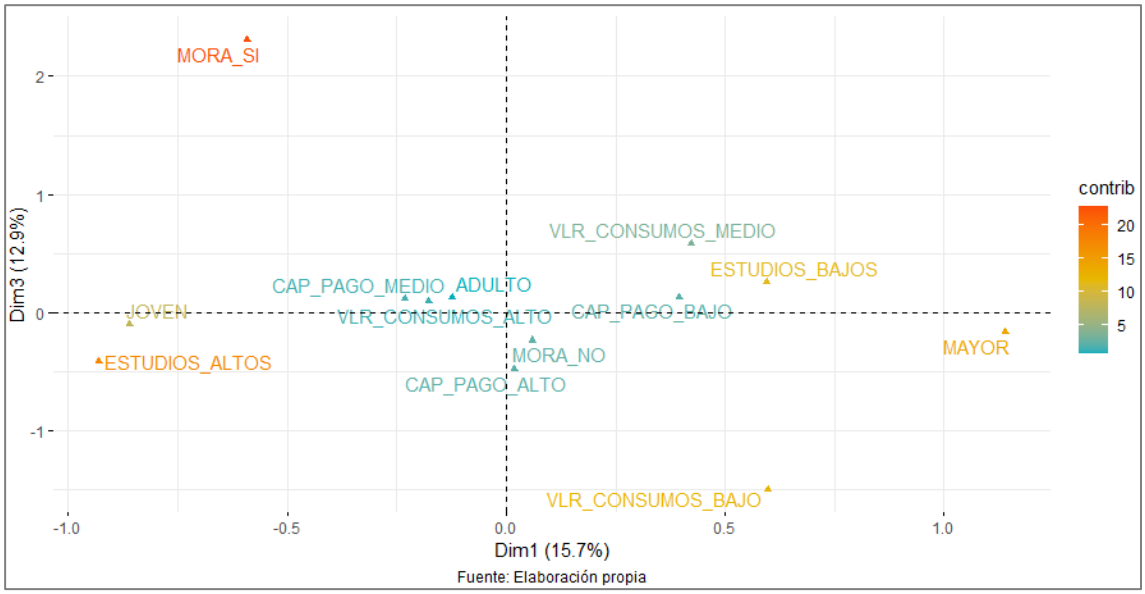
Tabla 4.

Influencia de variables en departamentos

DEPARTAMENTO	MORA	ESTUDIOS ALTOS	JOVEN	ADULTO	VALOR CONSUMOS ALTOS
Santander	x	x		x	
Cundinamarca	x	x	x		
Norte de Santander	x			x	x

Figura 13.

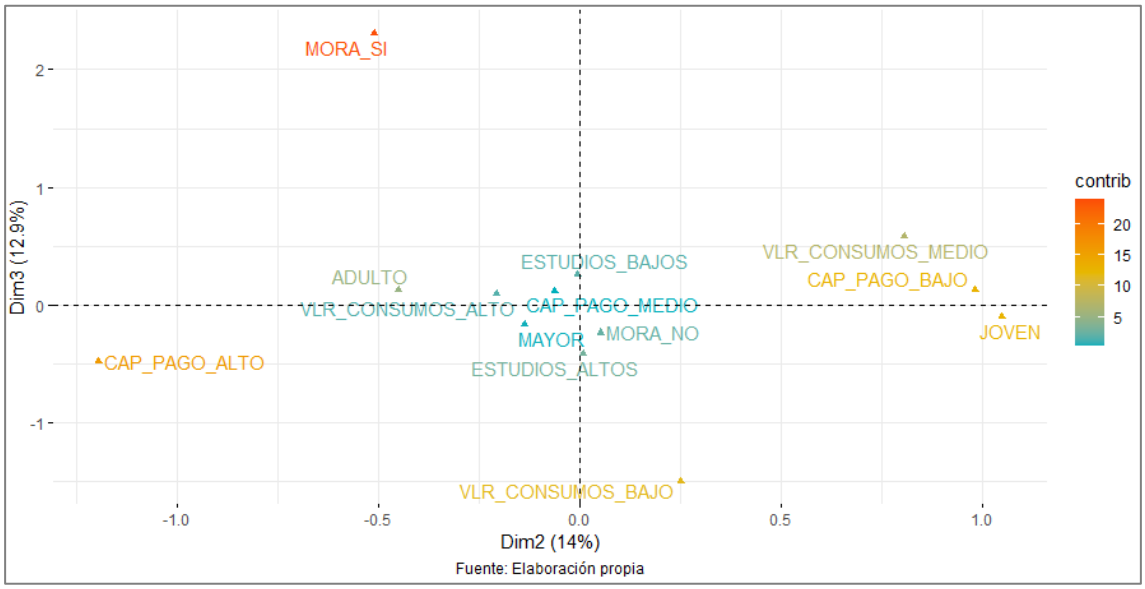
Contribuciones dimensión 1 y dimensión 3



Fuente: Elaboración propia

Figura 14.

Contribuciones dimensión 2 y dimensión 3



Fuente: Elaboración propia

8. Desarrollo modelo Logit

La regresión logística es una de las técnicas más conocidas y utilizadas para modelar una variable de respuesta categórica en función de variables predictoras continuas o categóricas. Forma parte de los modelos lineales generalizados, introducidos por (McCullagh and Nelder, 1989) y se aplica en campos tan distintos como la epidemiología, ecología, sociología o en los sectores bancario y asegurador. En este caso, los clientes con tarjeta de crédito de una cooperativa de ahorro y crédito, la regresión logística permite estudiar en términos de probabilidad si la mora de los clientes es explicada por las variables predictoras objeto de estudio. (Cañadas Reche, 2013)

8.1 Selección de variables

Para crear el modelo logit, se utiliza como variable respuesta DIAS_MORA, la cual esta codificada en 1 para la categoría de interés (cliente en mora) y 0 para los clientes sin mora y como variables explicativas numéricas CUPO_TOTAL, VALOR_CONSUMOS, CAPACIDAD_PAGO y EDAD_ASOCIADO.

8.2 Selección de la muestra

La muestra está conformada por 340 observaciones seleccionadas a través de un M.A.S(Muestreo Aleatorio Simple) con un intervalo de confianza del 95%, de una base de 3180 observaciones conformado por 1590 pares de individuos con características similares a sí:

DIAS_MORA: Variable codificada en 0(cliente sin mora) y 1(cliente en mora), en el cual el par de esta variable tendrá valores opuestos.

CUPO_TOTAL: Variables continua, cada valor del par varía entre más o menos 25% de su opuesto.

VALOR_CONSUMOS: Variables continua, cada valor del par varía entre más o menos 25% de su opuesto.

CAPACIDAD_PAGO: Variables continua, cada valor del par varía entre más o menos 25% de su opuesto.

EDAD_ASOCIADO: Variables ordinal, cada valor del par varía entre más o 3 unidades de su opuesto.

8.3 Selección del modelo Logit

La selección del modelo se realiza a través del software R con la librería MAS que incorpora el método stepIAC e implementa la selección automática de variables a través del criterio de información de Akaike.

Se crean dos modelos, el modelo.full que con tiene la interacciones de orden 4 entre las variables cupo_total, valor_consumos, capacidad_pago y edad_asociado así como todas las

interacciones de orden inferior y los efectos principales. Como modelo inicial tomamos el que sólo tiene el término constante, como se observa en la tabla 5.

Tabla 5.

Parámetros para elección del mejor modelo

```
modelo.full <- glm(DIASMORA ~ CUPO_TOTAL * VALOR_CONSUMOS * CAPACIDAD_PAGO * EDAD_ASOCIADO,
                   data = muestra2, family = binomial)

modelo.inicial <- glm(DIASMORA ~ 1, data = muestra2, family = binomial)
```

La tabla 6 muestra el modelo que mejor explica el estado de mora de los clientes con tarjeta de crédito, es el que incluye como predictores al valor de consumos y la capacidad de pago, con un AIC de 450.62 es el valor más bajo entre todos los modelos que resultan de las combinaciones posibles entre los predictores y la variable explicada.

Tabla 6.

Elección del mejor modelo según criterio AIC

```
Start:  AIC=467.1
DIASMORA ~ 1

                 Df Deviance    AIC
+ VALOR_CONSUMOS  1   451.71 455.71
+ CAPACIDAD_PAGO  1   454.17 458.17
+ CUPO_TOTAL      1   460.97 464.97
<none>                465.10 467.10
+ EDAD_ASOCIADO   1   465.02 469.02

Step:  AIC=455.71
DIASMORA ~ VALOR_CONSUMOS

                 Df Deviance    AIC
+ CAPACIDAD_PAGO  1   444.62 450.62
<none>                451.71 455.71
+ CUPO_TOTAL      1   451.20 457.20
+ EDAD_ASOCIADO   1   451.70 457.70
- VALOR_CONSUMOS  1   465.10 467.10

Step:  AIC=450.62
DIASMORA ~ VALOR_CONSUMOS + CAPACIDAD_PAGO

                                Df Deviance    AIC
<none>                               444.62 450.62
+ CUPO_TOTAL                     1   442.72 450.72
+ VALOR_CONSUMOS:CAPACIDAD_PAGO  1   444.27 452.27
+ EDAD_ASOCIADO                  1   444.52 452.52
- CAPACIDAD_PAGO                 1   451.71 455.71
- VALOR_CONSUMOS                 1   454.17 458.17
```

8.4 Validación del modelo

Para determinar si los predictores incluidos en el modelo elegido anteriormente, contribuyen de forma significativa a explicar la variable respuesta, se realiza mediante el test *Wald chi-test.* La tabla 7 contiene el resultado del test *Wald chi-test*, e indica que los predictores VALOR_CONSUMOS y CAPACIDAD_PAGO con un p-valor de (p-value=0.0002531) y (p-value=0.0077596) respectivamente, contribuyen de forma significativa al modelo.

Tabla 7.

Test Wald chi-test

```
Analysis of Deviance Table

Model: binomial, link: logit

Response: DIASMORA

Terms added sequentially (first to last)


               Df Deviance Resid. Df Resid. Dev  Pr(>Chi)
NULL                            339     465.10
VALOR_CONSUMOS  1  13.3890       338     451.71 0.0002531 ***
CAPACIDAD_PAGO  1   7.0881       337     444.62 0.0077596 **
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

8.5 Interpretación del modelo

La tabla 8 muestra que el coeficiente estimado para la intersección es el valor esperado del logaritmo de *odds* de que un cliente entre en mora, teniendo un 0 de valor de consumos de la tarjeta crédito y 0 pesos en su capacidad de pago. Los odds(1.019e+00) del intercepto arrojan que la probabilidad de que un cliente entre en mora, con los predictores del modelo en 0 es del exp(1.019e+00)/(1+exp(1.019e+00)) ≈ 73.47%.

El segundo coeficiente llamado VALOR_CONSUMOS es el logaritmo de la razón de probabilidad de un cliente en mora. La pendiente negativa de β_1 -1.026e-07, indica que cuando el valor de consumos aumenta en 1 unidad y los demás predictores se mantienen constantes, la probabilidad de mora de un cliente baja en un 1%.

El coeficiente llamado CAPACIDAD _PAGO tiene un comportamiento similar al coeficiente VALOR_CONSUMOS. Con una pendiente negativa de β_2 -1.415e-07, in

dica que si la capacidad de pago aumenta en 1 unidad y los demás predictores se mantienen constantes, la probabilidad de un cliente en mora baja en 1%.

Tabla 8.

Modelo logit final

```
Call:
glm(formula = DIASMORA ~ VALOR_CONSUMOS + CAPACIDAD_PAGO, family = "binomial",
    data = muestra2)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-1.5503  -1.2598   0.8777   0.9999   2.0771

Coefficients:
                  Estimate Std. Error z value Pr(>|z|)
(Intercept)     1.019e+00  2.127e-01   4.789 1.67e-06 ***
VALOR_CONSUMOS -1.026e-07  3.577e-08  -2.870  0.00411 **
CAPACIDAD_PAGO -1.415e-07  5.776e-08  -2.450  0.01427 *
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 465.10  on 339  degrees of freedom
Residual deviance: 444.62  on 337  degrees of freedom
AIC: 450.62

Number of Fisher Scoring iterations: 4
```

El modelo logit creado para predecir la probabilidad que un cliente entre en mora a partir de los predictores valor de consumos y capacidad de pago, está dado por la siguiente formula:

$$logic(mora) = 1.019\text{e} + 00 - 1.026\text{e} - 07 \ * \ \text{valorConsumos} - 1.415\text{e} - 07 \ * \ \text{capacidadPago}$$

$$P(mora) = \frac{e\text{^}(1.019e + 00 - 1.026e - 07 \ * \ valorConsumos - 1.415e - 07 * \ capacidadPago)}{1 + e\text{^}(1.019e + 00 - 1.026e - 07 \ * \ valorConsumos - 1.415e - 07 * \ capacidadPago)}$$

8.6 Poder de clasificación del modelo

En el estudio se utilizó un *threshold* de 0.5. Si la probabilidad de que la variable DIASMORA supere un 0.5 se asigna al nivel 1(cliente en mora) en caso contrario se asigna al nivel 0(cliente sin mora).

La tabla 9 muestra la capacidad de clasificación del modelo logit creado. Con una tasa del $\frac{63+83}{63+83+84+110} \approx 42.09\%$ de clasificación correcta, el modelo no es bueno para predecir. Por otro lado la sensibilidad del modelo es del $\frac{63}{63+84} \approx 42.85\%$, es decir la capacidad para detectar los clientes que entran en mora y la especificidad es decir la capacidad para detectar los clientes que no entraron en mora es del $\frac{83}{83+110} \approx 45.35\%$.

Tabla 9.

Matriz de confusión

	clasificado	
observado	0	1
1	84	63
0	83	110

9. Conclusiones

El análisis de correspondencia múltiple permitió encontrar relaciones importantes entre las variables objeto de estudio, las cuales ayudan a entender mejor el comportamiento de consumo y comportamiento de pago de los clientes con tarjeta de crédito.

Estas relaciones encontradas permitieron seleccionar los predictores que mejor explican la variable de interés, las cuales fueron utilizadas en la construcción del modelo logit.

El modelo estadístico obtenido para predecir si un cliente puede entrar en mora o no, tiene un poder de clasificación del 42.09%, valor que mide la calidad del modelo y resulta del porcentaje de aciertos del mismo.

Tomando como base las conclusiones del análisis de correspondencia múltiple, el valor obtenido de la calidad del modelo posiblemente obedece a las asociaciones débiles encontradas entre las variables objeto de estudio y la variable explicada.

El modelo estadístico final tiene una calidad que no puede ser alta, pero sirve como herramienta, para que en conjunto con otros factores como la experticia y capacidad estratégica a nivel comercial, permita llegar a generar estrategias de marketing sobre segmentos de mercado que maximicen la rentabilidad del producto tarjeta de crédito con el menor riesgo posible.

Se recomienda realizar análisis con nuevos predictores que permitan mejorar la calidad del modelo, y así obtener estadísticamente una herramienta más sólida para la toma de decisiones de marketing sobre el producto tarjeta de crédito.

Referencias Bibliográficas

Bambino, C. C. (27 de 01 de 2020). *flacsoandes.* Obtenido de www.flacsoandes.edu.ec: https://biblio.flacsoandes.edu.ec/catalog/resGet.php?resId=18022.

Blanco, V. (01 de 01 de 2015). *www.rdu.unc.edu.ar.* Obtenido de https://rdu.unc.edu.ar/bitstream/handle/11086/4707/Blanco%2C%20Victoria.%20Segmentacion%20de%20clientes%20en%20una%20entidad%20financiera..pdf?sequence=1&isAllowed=y

Bliss, C. (1934). The Method of Probits. *Science*, 38-39.

Campoverde, V. F. (27 de 01 de 2020). *zonaeconomica*. Obtenido de www.zonaeconomica.com: https://www.zonaeconomica.com/riesgo-crediticio?__cf_chl_jschl_tk__=c123d250e81cfed342a4583fb8877397c960fabe-1579571852-0-AavfX3WEpNsu2s4xb1Us3Fcdudk0NGv3o0pK3BXqrXdJSZ7kFl29A5SWXN0LTmvm2z8l6VITDtvkUR--YcIZH793vsdIdXk2ED_dpB6QEjORTP1sqcYoFkKkc4bKGYoijAM26

De La Fuente, F. S. (27 de 01 de 2020). *estadistica.* Obtenido de www.estadistica.net: http://www.estadistica.net/Master-Econometria/Analisis_Cluster.pdf

De La Fuente, F. S. (27 de 01 de 2020). *fuenterrebollo.* Obtenido de http://www.fuenterrebollo.com/: http://www.fuenterrebollo.com/Economicas/ECONOMETRIA/SEGMENTACION/DISCRIMINANTE/analisis-discriminante.pdf

Granda, O., & Niño, J. M. (01 de 05 de 2016). *http://tangara.uis.edu.co.* Obtenido de http://tangara.uis.edu.co/biblioweb/tesis/2016/163514.pdf

IBM. (27 de 01 de 2020). *IBM*. Obtenido de www.ibm.com: https://www.ibm.com/support/knowledgecenter/es/SSLVMB_sub/statistics_mainhelp_ddita/spss/categories/idh_cors.html

Marín, J. M. (27 de 01 de 2020). *halweb.uc3m.* Obtenido de halweb.uc3m.es: http://halweb.uc3m.es/esp/Personal/personas/jmmarin/esp/AMult/tema6am.pdf

Ochoa, P. J., Galeano, M. W., & Agudelo, V. L. (2010). Construcción de un modelo de scoring para el otorgamiento de credito en una institución financiera. *Perfil de Coyuntura Económica No. 16,* 191-222.

Seh-Lelha. (27 de 01 de 2020). *seh-lelha*. Obtenido de www.seh-lelha.org: https://www.seh-lelha.org/metodos-estadisticos-clasificacion/

uv. (27 de 01 de 2020). *uv*. Obtenido de www.uv.es: https://www.uv.es/ceaces/multivari/cluster/CLUSTER2.htm