

APLICACIÓN DE UNA REGRESIÓN BETA PARA LA ESTIMACIÓN DEL DESCUENTO POR OBLIGACIÓN DE CLIENTES CON CARTERA MOROSA.

APPLICATION OF BETA REGRESSION

Edinson Vicente Aguilar Rodriguez.^a
edinsonaguilar@usantotomas.edu.co

Director de trabajo: Hanwen Zang.^b

Resumen

Este trabajo de grado propone un modelo de regresión con distribución beta usando la metodología clásica y su implementación en R. El objetivo principal es evaluar la regresión beta para la estimación del descuento de las obligaciones en una empresa administradora de cartera vencida, con la información histórica de los descuentos que se han efectuado en un lapso de tiempo de un año. La idea es proporcionar el modelo cuyos resultados propongan el mejor ajuste, comprobando los supuestos de los residuales y llevando a cabo las comparaciones de los diferentes modelos que se prueben, con el fin de obtener el que mejor se ajuste a los datos proporcionados.

Palabras clave: . Regresión, descuento, ajuste, residuales.

Abstract

This paper proposes a grade regression model with beta distribution using the classical methodology and its implementation in R. The main objective is to evaluate the beta regression to estimate the discount obligations in a management company of nonperforming loans, with historical information discounts that have taken place in a time span of one year. The idea is to provide the model whose results propose the best fit, checking the assumptions of residuals and carrying out comparisons of different models that are tested, in order to obtain the best fit to the data provided.

Keywords: . Regression, discount, fit, residuals.

1. Introducción

La cartera vencida ha sido uno de los insumos de compañías que en Colombia se dedican a la recuperación de este activo que los bancos sacan de sus libros al cumplir con el 100 % de su provisión, es decir que son deudas incobrables “en mi opinión mal llamadas por el sector financiero”.

Uno de los interrogantes de las compañías que se dedican a cobrar estas obligaciones es; ¿qué descuento le podrán dar a cada obligación?. En este trabajo, se ilustra una de las metodologías estadísticas que podría ayudar resolver esta pregunta, se trata de la regresión beta que dadas sus propiedades nos permite evaluar tasas y porcentajes; razón por la cual puede ser un buen inicio para abordar la problemática.

^aEstudiante de estadística Universidad Santo Tomás Bogotá

^bPhd. en ciencias, docente de la universidad Santo Tomás

La metodología que analizaremos empezará desde la extracción de los datos ya que por lo general en este tipo de estudios se cuenta con la información, lo que aquí intentaremos, será observar también las dificultades a las que se enfrentan los estadísticos a la hora de construir las bases de datos. Finalizado el paso anterior, continuamos con la construcción del modelo, en donde revisaremos cada posible mejora a aplicar a nuestra estimación, y por ultimo haremos una comprobación de supuestos de los residuales que arroje el modelo.

Generalidades

Refinancia es una compañía que se dedica a la compra y administración de cartera morosa de clientes del sector financiero, especialmente en bancos y en todas sus modalidades de crédito. Esta cartera se clasifica en dos líneas del negocio; NPL's para la cartera comprada y Servicing para la cartera administrada. También ofrece avales mediante cheques y pagares, y posteriormente se dedica a la cobranza de los siniestros (incumplimientos) que se presentan en los avales otorgados.

NPL's es la principal línea de negocio de la compañía, la cual genera un 60 % aproximadamente de los ingresos. Todas las obligaciones que llegan a esta línea de negocio antes de ser compradas son sometidas a un proceso de valoración con el cual se determina si la compra es rentable o no. El proceso de valoración tiene como objeto determinar el comportamiento de pago de los clientes del portafolio en el que se va a invertir, adicionalmente busca establecer el precio a pagar por ese portafolio, velando porque la inversión genere rentabilidad para el inversionista y utilidad para Refinancia.

Por ultimo, aprobada y ejecutada la compra, las obligaciones pasan por un proceso revaloración del portafolio con el fin de establecer el flujo futuro y los niveles de descuento a lo largo del tiempo.

Posteriormente se migra la información de cada obligación al software utilizado por Refinancia y después se aplica una estrategia de operación del portafolio en donde se define el capacity (fuerza comercial) y los canales a los cuales serán asignados cada uno de los clientes.

Los canales a los que podrán ser asignados son los siguientes:

Judicialización: Se envían todas aquellas obligaciones que presenten procesos jurídicos en el momento de su compra.

Cierres: Aquí llegan todos los clientes que son sujetos a ser cobrados mediante el call center, distribuidos en los segmentos masivo (saldos bajos), medio, alto, top, ultra top.

Localización: Clientes de los cuales no se tiene certeza de si su información telefónica es efectiva.

Canales no tradicionales: Es el canal a donde ingresan los clientes que tienen información geográfica o email.

Tercerización: Es cuando la cartera es enviada a una compañía externa para que esta realice la labor de cobranza.

Diagnóstico de la situación actual:

El proceso de la fijación de precios en NPL's, en la actualidad se viene realizando de dos maneras diferentes; la primera aplica para portafolios adquiridos antes de septiembre de 2012 y la segunda mediante la metodología propuesta por el área de Valoración.

Para los portafolios de antes del 2012 se utiliza una regresión lineal la cual se basó en la información de pago del cliente (4 variables), y con los coeficientes de esta determinar el valor a cobrar según el tiempo en Refinancia.

Para las carteras posteriores a septiembre de 2012 se usa una metodología que básicamente consiste en calcular un descuento por cada segmento mediante los promedios obtenidos por variables como sal-

do a capital, edad de mora, etc., en portafolios de características similares trabajados previamente en Refinancia.

Presentación del problema:

En la sección anterior, podemos identificar algunas dificultades que se presentan actualmente en NPL's, entre ellas la fijación de un precio ajustado a la historia para el pago de las obligaciones de cada uno de los clientes. Esto se debe a que la metodología utilizada actualmente no es la más conveniente para la operación, dado que genera cargas operacionales y comités extraordinarios que resultan desgastantes y costosos. Esto se ve reflejado en que eventualmente una oferta de pago debe escalar mas de 4 instancias para poder ser aprobada, lo que hace que la operación sea demorada.

Un problema adicional, es que el modelo anterior a septiembre de 2012 presenta un R square muy bajo cercano al 30 % lo que sugiere que el modelo posiblemente no sea lo suficientemente ajustado para modelar las variables elegidas. También se identifica una falla adicional la cual es la falta de calibración del modelo, ya que desde que se empezó a usar no se realiza un testeo que permita determinar si las condiciones de las obligaciones han cambiado con el paso del tiempo.

2. Objetivos

2.1. Objetivo Principal

Evaluar la regresión Beta clásica para la construcción de un modelo estadístico, que basado en información histórica, sirva como herramienta para establecer el porcentaje de descuento de cada obligación de los clientes con cartera morosa, teniendo en cuenta características como antigüedad, mora, saldo, entre otras.

2.2. Objetivo Específicos

- Construir la base de datos que se usará para el desarrollo del modelo.
- Proporcionar el modelo beta con mejor ajuste para los datos obtenidos.
- Identificar las características a tener en cuenta a la hora de construir un modelo.

3. Metodología

3.1. Regresión Beta Clásica

La densidad de la distribución beta está dada por la correspondiente función:

$$f(y, \alpha, \beta) = \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)} y^{\alpha-1} (1-y)^{\beta-1}$$

Sin embargo, para un análisis de regresión es típicamente más útil modelar el promedio que la respuesta.

También es típico definir un modelo que contenga la precisión o el parámetro de dispersión.

Sea

$$\mu = \frac{\alpha}{\alpha+\beta} \quad \phi = \alpha + \beta$$

entonces si:

$$\alpha = \mu\phi \quad \beta = (1 - \mu)\phi$$

Se define el valor esperado como:

$$E(y) = \mu$$

La varianza de y es:

$$V(y) = \frac{V(\mu)}{1+\phi}$$

Esto es fácil de verificar porque

$$V(y) = \frac{\alpha\beta}{(\alpha+\beta)^2(\alpha+\beta+1)}$$

sustituyendo:

$$\alpha = \mu\phi \quad \beta = (1 - \mu)\phi \quad V(y) = \frac{\mu\phi(1-\mu)\phi}{\phi^2(\phi+1)} \text{ da como resultado: } V(y) = \frac{V(\mu)}{1+\phi}$$

como $V(\mu) = \mu(1 - \mu)$ entonces μ es la media de la variable respuesta y ϕ el parámetro de precisión y siguen una distribución

$$f(y, \mu\phi, (1 - \mu)\phi) = \frac{\Gamma(\mu\phi + (1-\mu)\phi)}{\Gamma(\mu\phi)\Gamma((1-\mu)\phi)} y^{\mu\phi-1} (1-y)^{(1-\mu)\phi-1}$$

Donde: $0 < \mu < 1 \quad \phi > 0$

Es importante resaltar que:

1. $f(y, \mu, \phi)$ es simétrico cuando $\mu = 1/2$
2. la función decrece cuando μ decrece y ϕ se incrementa
3. En betareg cuando la variable respuesta esta restringida al intervalo (a, b) resulta útil modelar $\frac{y-a}{b-a}$ en vez de la variable y directamente

Función de enlace

Sea $y_1, \dots, y_n$ la variable aleatoria independiente donde cada y_i sigue la densidad $f(y, \mu, \phi)$ con media μ_t y precisión desconocida ϕ el modelo es obtenido asumiendo que la media de y puede ser escrita como:

$$g(\mu_t) = \sum_{i=1}^n x_{ti} \beta_i$$

donde: β_i es el vector de parametros desconocidos x_{ti} son observaciones de k covariables

la función de enlace es estrictamente monótona , 2 veces diferenciable y toma valores ente cero y uno en los reales

Dentro de las posibles elecciones para la variable enlace están logit, probit, y el complemento loglog, sin embargo para efectos prácticos se implementaría la logit.

$$g(\mu_t) = \log\left(\frac{\mu}{1-\mu}\right) \quad g(\mu_t) = \log\left(\frac{e^{x_t^T \beta}}{1+e^{x_t^T \beta}}\right)$$

Es importante resaltar que el objetivo de este trabajo es modelar μ por medio de la función de enlace. Un pensamiento natural sería pensar en:

$$\mu = x'\beta$$

Pero esto sería un error ya que μ podría tomar valores menores que cero y mayores que uno. Una función de link particularmente útil es

$$\mu_t = \frac{\exp^{x_t^t \beta}}{1 + \exp^{x_t^t \beta}}$$

Debido a la siguiente propiedad, supongase que el i -ésimo regresor se incrementa en c unidades y las otras unidades permanecen sin cambios, se puede mostrar que:

$$e^{c\beta_i} = \frac{\frac{\mu^*}{(1-\mu^*)}}{\frac{\mu}{(1-\mu)}}$$

donde $e^{c\beta_i}$ corresponde al odds.

La Función Score

la logverosimilitud basada en una muestra de n observaciones independientes es

$$l(\beta, \phi) = \sum_{i=1}^n l_t(\mu_t, \phi)$$

$$l(\beta, \phi) = \log(\Gamma(\phi)) - \log(\Gamma(\mu\phi)) - \log(\Gamma(1 - \mu_t)\phi) + (\mu\phi - 1)\log(y_t) + ((1 - \mu_t)\phi - 1)\log(1 - y_t)$$

sea

$$y_t^* = \log\left(\frac{y_t}{1-y_t}\right) \text{ y } \mu_t^* = \varphi(\mu_t * \phi) - \varphi((1 - \mu_t) * \phi)$$

$$\frac{\partial l(\beta, \phi)}{\partial \beta_i} = \sum_{i=1}^n \frac{\partial l_t(\mu_t, \phi)}{\partial \beta_{\mu_t}} \frac{\partial(\mu_t)}{\partial \eta_t} \frac{\partial(\eta_k)}{\partial \beta_i}$$

$$\frac{\partial l_t(\mu_t, \phi)}{\partial \beta_{\mu_t}} = \phi(y^* - \mu^*)$$

donde

$$\varphi(x) = \frac{\partial \log(\Gamma(z))}{\partial z}$$

para $z > 0$

$$\frac{\partial \beta \phi}{\partial \beta_i} = \frac{\phi(\sum_{i=1}^n (y_t^* - \mu_t^*))}{g'(\mu_t)}$$

entonces:

$$U_\beta(\beta, \phi) = \phi x^t T (y_t^* - \mu_t^*) \text{ donde } T = \text{diag}^{-1} \frac{1}{g(\mu_t)}$$

$$U_\phi(\beta, \phi) = \sum_{i=1}^n (\mu_t(y_t^* - \mu_t^*)) + \log(1 - y_t) - \varphi((1 - \mu_t)\phi + \varphi(\phi))$$

la función score esta dada por:

$$(U_\beta(\beta, \phi))^t, U_\phi(\beta, \phi))^t$$

La matriz de información de Fisher

sea W una matriz diagonal con:

$$w_t = \phi(\psi'(\mu_t\phi) + \psi'(1 - \mu_t\phi)) \frac{1}{g'(\mu_t)^2}$$

sea c

$$c = (c_1, \dots, c_n)^t \quad c_t = \phi(\psi'(\mu_t\phi)\mu_t + \psi'((1 - \mu_t)\phi))(1 - \mu_t)$$

sea dt

$$dt = \psi'(\mu_t\phi)\mu_t^2 + \psi'(1 - \mu_t\phi)(1 - \mu_t)^2 - \psi'(\phi)$$

Creando la matriz k : se sabe por la logverosimilitud que la función gamma se puede escribir como

$$l(\beta, \phi) = \sum_{i=1}^n l_t(\mu_t, \phi)$$

donde la derivada respecto a β_i de esta verosimilitud se puede expresar como

$$\frac{\partial(\beta, \phi)}{\partial \beta_i} = \sum_{t=1}^n \frac{\partial l_t(\mu_t, \phi)}{\partial \mu_t} \frac{\partial(\mu_t)}{\partial \eta_t} \frac{\partial(\eta_t)}{\partial \beta_i}$$

se sabe que

$$l(\beta, \phi) = \log(\Gamma(\phi)) - \log(\Gamma(\mu\phi)) - \log(\Gamma(1 - \mu_t\phi)) + (\mu\phi - 1)\log(y_t) + ((1 - \mu_t\phi) - 1)\log(1 - y_t)$$

usando esto se puede establecer que:

$$\frac{\partial l(\mu_t, \phi)}{\partial \mu_t} = \phi \left(\log \frac{y_t}{1 - y_t} - \psi(\mu_t\phi) - (\psi(1 - \mu_t\phi)) \right)$$

donde ψ es la función digamma por ejemplo, $\psi(z) = \frac{\partial \log \Gamma(z)}{\partial z}$ para $z > 0$. Por condiciones de regularidad se sabe que en la función de verosimilitud la derivada se igual a cero lo que genera que

$$\mu = E(y_t^*) \quad \frac{\partial(\beta, \phi)}{\partial \beta_i} = \phi \sum_{i=1}^n (y_t^* - \mu_t^*) \frac{1}{g'(\mu_t)} x_{ti}$$

Expresión de la cual deriva la función de score. Sin embargo como nuestro objetivo es la segunda derivada para construir la matriz de fisher se deriva por segunda vez:

■ cálculo de $K\beta\beta = E\left(\frac{\partial^2 l(\beta, \phi)}{\partial \beta \partial \beta}\right)$

$$\begin{aligned} \frac{\partial(\beta, \phi)}{\partial \beta \partial \beta} &= \sum_{t=1}^n \frac{\partial}{\partial \mu_t} \left(\frac{\partial l_t(\mu_t, \phi)}{\partial \mu_t} \frac{\partial(\mu_t)}{\partial \eta_t} \right) \frac{\partial \mu_t}{\partial \eta_t} \frac{\partial(\eta_t)}{\partial \beta_j} x_{ti} \\ \frac{\partial(\beta, \phi)}{\partial \beta \partial \beta} &= \sum_{t=1}^n \left(\frac{\partial^2 l_t(\mu_t, \phi)}{\partial \mu_t^2} + \frac{\partial l_t(\mu_t, \phi)}{\partial \mu_t} \frac{\partial}{\partial \mu_t} \frac{\partial \mu_t}{\partial \eta_t} \right) \frac{\partial(\mu_t)}{\partial \eta_t} x_{ti} x_{tj} \end{aligned}$$

como $E\left(\frac{\partial l(\mu_t, \phi)}{\partial \mu_t}\right) = 0$ se tiene que

$$E\left(\frac{\partial(\beta, \phi)}{\partial \beta \partial \beta}\right) = \sum_{t=1}^n E\left(\frac{\partial^2 l(\mu_t, \phi)}{\partial \mu_t^2}\right) \left(\frac{\partial \mu_t}{\partial \eta_t}\right)^2 x_{ti} x_{tj}$$

por

$$\frac{\partial l(\mu_t, \phi)}{\partial \mu_t} = \phi \left(\log \frac{y_t}{1-y_t} - \psi(\mu_t \phi) - (\psi(1 - \mu_t) \phi) \right)$$

se puede decir que

$$\frac{\partial^2 l(\mu_t, \phi)}{\partial \mu_t^2} = -\phi^2 (\psi'(\mu_t \phi) + (\psi'(1 - \mu_t) \phi))$$

y como resultado tomando en cuenta el valor de w_i definido arriba

$$E\left(\frac{\partial^2 l(\beta, \phi)}{\partial \beta^2}\right) = -\phi^2 \sum_{t=1}^n w_i x_{ti} x_{tj}$$

que matricialmente puede escribirse como

$$E\left(\frac{\partial^2 l(\beta, \phi)}{\partial \beta \partial \beta}\right) = -\phi X^t W X$$

- cálculo de $K\beta\phi = E\left(\frac{\partial^2 l(\beta, \phi)}{\partial \beta \partial \phi}\right)$

$$\frac{\partial(\beta, \phi)}{\partial \beta_i} = \phi \sum_{i=1}^n (y_t^* - \mu_t^*) \frac{1}{g'(\mu_t)} x_{ti}$$

Derivando por segunda vez, evaluando la esperanza y considerando el valor de c definido arriba se tiene

$$E\left(\frac{\partial^2 l(\beta, \phi)}{\partial \beta_i \partial \phi}\right) = -\sum_{i=1}^n c_t \frac{1}{g'(\mu_t)} x_{ti}$$

matricialmente

$$E\left(\frac{\partial^2 l(\beta, \phi)}{\partial \beta_i \partial \phi}\right) = X^t T C$$

- cálculo de $K\phi\phi = E\left(\frac{\partial^2 l(\beta, \phi)}{\partial \phi^2}\right)$ se toma esta expresión del score y se deriva contra ϕ

$$U_\phi(\beta, \phi) = \sum_{i=1}^n (\mu_t (y_t^* - \mu_t^*)) + \log(1 - y_t) - \phi((1 - \mu_t)\phi + \varphi(\phi))$$

derivando por segunda vez considerando el valor de d_t definido arriba y evaluando la esperanza

$$E\left(\frac{\partial U_\phi(\beta, \phi)}{\partial \phi}\right) = -\sum_{t=1}^n d_t$$

matricialmente

$$K\phi\phi = -\text{traza}(d_t)$$

Finalmente la matriz de fisher sería:

$$K = \begin{bmatrix} k\beta\beta & k\beta\phi k\phi\beta & k\phi\phi \end{bmatrix}$$

Inferencia

Bajo condiciones de regularidad el estimador de máxima verosimilitud está dado por:

$$[\hat{\beta}\hat{\phi}] \sim N([\beta\phi], K^{-1})$$

Los estimadores de máxima verosimilitud de β y ϕ se obtienen de las ecuaciones $U\beta(\beta, \phi) = 0$ y $U\phi(\beta, \phi) = 0$, pero no tienen forma reducida. Por lo tanto, tienen que ser obtenidos numéricamente maximizando la función de log-verosimilitud usando un algoritmo de optimización no lineal, tal como el algoritmo de Newton.

Los algoritmos de optimización requieren la especificación de los valores iniciales para ser utilizados en el esquema iterativo. Se sugiere utilizar como punto inicial para la estimación para β la estimación de mínimos cuadrados ordinarios para este vector de parámetros obtenidos de una regresión lineal de las respuestas transformadas $g(y_1), \dots, g(y_n)$ en x .

Un buen estimador de β podría ser $(X^T X)^{-1} X^T Z$ siendo $z = (g(y_1), \dots, g(y_n))^t$.

También es necesario un valor inicial para ϕ como anteriormente se demostró

$$var(y_t) = \frac{\mu_t(1-\mu_t)}{(1+\phi)}$$

lo cual implica que

$$\phi = \frac{\mu_t(1-\mu_t)}{var(y_t)} - 1$$

es decir,

$$var(y_t) \approx var(g(\mu_t)) + (y_t - \mu_t)g'(\mu_t) = var(y_t)(g'(\mu_t))^2$$

notese que:

$$var(y_t) \approx var(g(y_t)(g'(\mu_t)))^{-2}$$

por lo que el valor inicial de ϕ que se sugiere es:

$$\hat{\phi} = \frac{1}{n} \sum_{i=1}^n \frac{\check{\mu}(1-\check{\mu})}{\check{\sigma}^2} - 1$$

dónde $\check{\mu}$ se obtiene aplicando $g^{-1}()$ al i -ésimo valor ajustado de la regresión lineal $g(y_1), \dots, g(y_n)$ en X

$$\check{\mu} = g^{-1}(x_t(X^T X)^{-1}(X^T z))$$

$$\check{\sigma} = \frac{(\check{e})^t \check{e}}{(n-k)(g'(\check{\mu}_t))^2} \quad \check{e} = z - X(X^T X)^{-1} X^T Z.$$

Estadísticos para realizar inferencia

A continuación, se muestra como realizar inferencia en una muestra en el modelo de regresión beta. Considere, por ejemplo, la prueba de la hipótesis nula

$$H_0 : \beta_1 = \beta_1^{(0)} \quad H_a : \beta_1 \neq \beta_1^{(0)}$$

Donde β_1 corresponde al vector de los betas del modelo

$$\beta_1 = (\beta_1, \dots, \beta_m)^t \text{ y } \beta_1^{(0)} = (\beta_1^{(0)}, \dots, \beta_m^{(0)})^t \text{ con } m < k$$

- La estadística de razón de verosimilitudes esta dada por

$$w = 2(l(\hat{\beta}, \hat{\phi}) - l(\tilde{\beta}, \tilde{\phi}))$$

donde: $l(\hat{\beta}, \hat{\phi})$ corresponde al modelo reducido bajo la hipótesis nula $l(\tilde{\beta}, \tilde{\phi})$ corresponde a la estimación del modelo estimado por la máxima verosimilitud cumpliéndose la hipótesis nula. Bajo condiciones de regularidad el estadístico w se distribuye ji cuadrado con m grados de libertad por lo que este puede ser empleado para calcular las regiones críticas.

- La estadística de razón de Wald esta dada por

$$w = (\hat{\beta}_1 - \beta_1^{(0)})^t (k_1^{-1})^{-1} (\hat{\beta}_1 - \beta_1^{(0)})$$

donde $k_{11}^{\beta\beta}$ corresponde a $K_{\beta\beta}$ evaluado en el beta de interés. este estadístico presenta también ji cuadrado con m grados de libertad

- La estadística de razón de odds el intervalo para un beta dado por la razón de odds este intervalo corresponde al inverso de la función de enlace dado por

$$\exp^{c(\hat{\beta}_1 \pm \phi^{-1}(1-\alpha/2)se(\hat{\beta}_1))}$$

Residuales

Después de ajustar el modelo es importante realizar análisis diagnósticos con el fin de verificar la bondad de ajuste del modelo estimado. Se busca introducir medidas de variabilidad explicada y herramientas gráficas con el fin de encontrar fallas en el ajuste o medidas de influencia.

- El pseudo R^2 definido como el coeficiente de correlación muestral entre $\hat{\eta}$ y $g(y)$ siendo $R^2 = 1$ cuando existe asociación perfecta entre μ y y .
- La discrepancia de bondad de ajuste: puede ser medida como 2 veces la diferencia entre el modelo de máxima verosimilitud conseguible (modelo saturado) y el obtenido bajo el modelo planteado. sea

$$D(y; \mu; \phi) = \sum_{t=1}^n 2(l_t(\tilde{\mu}_t, \phi) - l_t(\mu_t, \phi))$$

dónde $\tilde{\mu}_t$ es el valor que resuelve $\frac{\partial l_t}{\partial \mu_t} = 0$ lo que implica que $\phi(y_t^* - \mu_t^*) = 0$ cuando ϕ toma valores muy grandes $\mu_t^* \approx \frac{\mu_t}{1-\mu_t}$ lo que implica que $\tilde{\mu}_t \approx y_t$. cuando ϕ es conocido la medida de discrepancia equivale a $D(y, \tilde{\mu}, \phi)$ donde $\tilde{\mu}$ es el modelo de máxima verosimilitud del modelo. Cuando ϕ es desconocido una aproximación a esta cantidad corresponde a la devianza del modelo donde la devianza es:

$$D(y; \hat{\mu}, \hat{\phi}) = \sum_{t=1}^n (r_t^d)^2 = \text{signo}(y_t - \hat{\mu}_t) 2(l_t(\tilde{\mu}_t, \hat{\phi}) - l_t(\hat{\mu}_t, \hat{\phi}))$$

Es importante destacar que los residuales r_t^d contribuyen un $(r_t^d)^2$ a la devianza por lo que un residual con un valor muy alto se puede establecer como discrepante.

- residuales estandarizado es posible definir los residuales estandarizados como

$$\frac{y_t - \mu_t}{\text{var}(y_i)^{1/2}}$$

donde

$$\text{var}(\hat{y}_i) = \frac{\mu_i(1-\mu_i)}{(1+\phi_i)} \hat{\mu}_i = g^{-1}(x_i' \hat{\beta})$$

Es importante destacar que si el gráfico de los residuales contra $\hat{\eta}$ presenta tendencia podría presentarse un problema de mala especificación de la función de enlace.

3.2. Multicolinealidad

Como la multicolinealidad es un problema muestral, ya que va asociada a la configuración concreta de la matriz X , no existen contrastes estadísticos, propiamente dichos, que sean aplicables para su detección. En cambio, se han desarrollado numerosas reglas prácticas que tratan de determinar en qué medida la multicolinealidad afecta gravemente a la estimación y contraste de un modelo. Estas reglas no son siempre fiables, siendo en algunos casos muy discutibles. A continuación se van a exponer dos procedimientos (el factor de agrandamiento de la varianza y el número de condición) que son los que gozan de mayor soporte - especialmente el segundo - en la literatura econométrica actual.

Factor de inflación de la varianza (FIV)

En un modelo de regresión múltiple, si el regresor j -ésimo fuera ortogonal con respecto a los demás regresores (es decir, si la correlación con el resto de los regresores fuera nula), la fórmula para la varianza quedaría reducida a:

$$\text{var}(\hat{y}_i) = \frac{\sigma^2}{TS_j^2}$$

El cociente entre (1) y (2) es precisamente el factor de agrandamiento de la varianza (FIV), cuya expresión será

$$FIV(\hat{\beta}) = \frac{1}{1-R_j^2}$$

Así, pues, el $FIV(\hat{\beta})$ es la razón entre la varianza observada y la que habría sido en caso de que X_j estuviera incorrelacionada con el resto de regresores del modelo. Dicho de otra forma, el FIV muestra en qué medida se «agrandan» la varianza del estimador como consecuencia de la no ortogonalidad de los regresores. Algunos autores consideran que existe un problema grave de multicolinealidad cuando el FIV de algún coeficiente es mayor de 10, es decir, cuando el $R_j^2 > 0,90$. En algunos programas de ordenador (el SPSS, por ejemplo) se define el término de tolerancia como la diferencia entre 1 y el R_j^2 . Análogamente con el criterio aplicado a las otras medidas, se puede decir que existe un problema de multicolinealidad cuando la tolerancia $< 0,10$.

El problema que tiene el FIV (o el R_j^2 o la tolerancia) es que no suministra ninguna información que pueda utilizarse para corregir el problema.

Número de condición

Este procedimiento de detección de la multicolinealidad es el más adecuado entre los actualmente disponibles, según afirman Judge et al. (1985) III-4. Fue planteado inicialmente por Rachudel (1971) y desarrollado posteriormente por Belsley et al. (1980), y Belsley (1982).

El número de condición, $k(X)$, es igual a la raíz cuadrada de la razón entre la raíz característica más grande ($\lambda_{(max)}$) y la raíz característica más pequeña ($\lambda_{(min)}$) de la matriz $X'X$, es decir,

$$k(X) = \sqrt{\frac{\lambda_{(max)}}{\lambda_{(min)}}}$$

Como la matriz $X'X$ es de dimensión $k * K$ se obtienen k raíces características, pudiéndose calcular para cada una de ellas un índice de condición definido de la siguiente forma:

$$ic(\lambda_i) = \sqrt{\frac{\lambda_{(max)}}{\lambda_i}}$$

El número de condición mide la sensibilidad de las estimaciones mínimocuadráticas ante pequeños cambios en los datos. De acuerdo con los estudios realizados por Belsley y otros, tanto con datos observados como con datos simulados, el problema de la multicolinealidad es grave cuando el número de condición toma un valor entre 20 y 30. Naturalmente, si este indicador superase el valor de 30, el problema sería ya manifiestamente grave. Estos valores vienen generalmente referidos a regresores medidos con escala de longitud unidad (es decir, con los regresores divididos por la raíz cuadrada de la suma de los valores de las observaciones), pero no centrados. Parece que no es conveniente centrar los datos (es decir, restarles sus correspondientes medias), ya que esta operación oscurece cualquier dependencia lineal que implique al término independiente.

4. Construcción Del Modelo

Para la construcción del modelo se siguieron una metodología que consta de los siguientes pasos:

4.1. Construcción De La Base Le Datos

Para trabajar en SQL SERVER se tuvo en cuenta una serie de pasos básicos:

Primer paso: Definir las variables.

La definición de las variables se realizó en la reunión de expertos, en donde se discutió cuales variables deberían incluirse al modelo y también las que eran necesarias para la extracción de toda la información, es decir, existen variables que a pesar de que no van en el modelo, deben ser tenidas en cuenta ya que de ellas depende el cálculo de otras.

Base	Variable	Observaciones
Acuerdos	ultacuerdo	Ultima fecha de acuerdo
Acuerdos	idacuerdo	Identificación interna del acuerdo
Acuerdos	tipoacuerdo	Tipo de acuerdo (Al instante, a plazos)
Acuerdos	fechaacuerdo	Fecha de constitución del acuerdo
Acuerdos	tlcuotasacuerdo	Cuenta de cuotas del acuerdo
Acuerdos	valoracuerdo	Sumatoria del total en valor del idacuerdo
Clientes	idcliente	Identificación interna del cliente
Clientes	Multiportafolio	1 Si cumple 0 si no.
Negocios	maxdiasmora	Máximo de días mora que presentó la obligación en el banco
Negocios	diasmorarefi	Días mora que presentó la obligación en la compañía de cobro
Negocios	Promponpagorigi	Promedio de pago por portafolio que presentó el cliente

Negocios	sumsaldocapitalt	Sumatoria del saldo de las obligaciones en el portafolio
Negocios	interesescorrientest	Total intereses corrientes de las obligaciones en el portafolio
Negocios	interesemorat	Total intereses moratorios de las obligaciones en el portafolio
Negocios	interesestrost	Total intereses causados por otros conceptos en la obligación
Negocios	interesersfnt	Sumatoria del saldo de los intereses en la compañía
Negocios	deudatotal	Sumatoria de todos los intereses y el capital de la obligación
Negocios	tarjetasport	Cuenta de tarjetas de crédito en cada uno de los portafolios
Negocios	depto	Departamento de origen de la obligación
Negocios	Segmento	Saldo a capital categorico (masivo, medio, alto, top y ultratop)
Pagos	maxfechacancelacion	Fecha de finalización de la obligación
Pagos	valorpagado	Valor total del pago por obligación
Pagos	cosecha	Mes y fecha en cual se realizó el pago
Portafolios	idportafolio	Identificación interna del portafolio
Portafolios	fechacompra	Fecha en la que el banco vendió la obligación
Portafolios	tiempogestion	Tiempo que lleva gestionada la obligación en la compañía
Portafolios	originador	Banco del cual proviene la obligación
Calculadas	acuerdoalinstante	1 Si paga de contado.
Calculadas	PortAntiguo	1 Si el portafolio e antiguo.
Calculadas	obligportafolio	Cuenta obligaciones del cliente en cada uno de los portafolios
Calculadas	noportafolios	Cuenta de los portafolios en los que se encuentra cada cliente
Calculadas	descuentocap	Capital total menos el valor pagado
Calculadas	tiempogestioncat	Categorización del tiempo de gestionada la obligación
Calculadas	idclienteportafolio	Unión del idcliente con el idportafolio
Calculadas	Negociaciones	Negociaciones que tuvo cada obligación antes de ser cancelada
Campañas	campanas	Cuenta de campañas en las que estuvo involucrada la obligación
Campañas	sede	Ciudad donde se aplica la campaña

Tabla 1: variables elegidas para la construcción de la base de datos

Segundo paso: Identificar las bases a las cuales se les realizará la extracción de la información. A continuación describimos cada uno de los repositorios de los que se realizó la extracción:

- Base de pagos: Contiene la información de todos los pagos realizados en la empresa desde su inicio, bien sea abono o cancelación total.
- Base de negocios: Contiene la información de cada una de las obligaciones en la empresa.
- Base de clientes: Contiene la información personal de cada uno de los clientes que ha ingresado con al menos una obligación a la compañía.
- Base de negociaciones: Contiene las negociaciones realizadas por cada cliente.

Una negociación se constituye cuando el cliente realiza una promesa de pago bien sea telefónica o personal, pero sin ningún abono.

- Base de acuerdos: Esta contiene la información de cada acuerdo que ha realizado un cliente históricamente en la empresa.

Un acuerdo se constituye cuando el cliente realiza un abono o en su defecto el pago de la obligación.

- Base de campañas: Contiene la identificación de las campañas en las que han estado cada una de las

obligaciones de los clientes.

- Base de portafolios: Es la base que contiene la información de cada uno de los portafolios.

Tercer paso: Realizar query¹ teniendo en cuenta la lógica del cálculo de las variables. Es importante tener en cuenta que a la hora de hacer la consulta que nos trae las variables, esta sea construida por un experto que conozca los repositorios, y sobretodo el funcionamiento del negocio, empresa, etc.

Resultado

Al ejecutar el query, obtenemos una base de datos con 8220 registros únicos y 12 variables. Este será nuestro insumo para la construcción del modelo.

4.2. Variables Elegidas Para La Construcción Del Modelo

Las variables para un modelo se eligen de acuerdo con la opinión de los dueños de la información, es decir, se debe acatar la sugerencia que hagan los expertos de acuerdo a las variables a incluir en el modelo, ya que sin ello podemos incurrir en errores de interpretación y de construcción del modelo.

Las variables elegidas son:

Variable	Observaciones
maxdiasmora	Máximo de días mora que presentó la obligación en el banco
obligportafolio	Cuenta obligaciones del cliente en cada uno de los portafolios
diasmorarefi	Días mora que presentó la obligación en la compañía de cobro
Promponpagorigi	Promedio de pago por portafolio que presentó el cliente
sumsaldocapitalt	Sumatoria del saldo de las obligaciones en el portafolio
tarjetasport	Cuenta de tarjetas de crédito en cada uno de los portafolios
Segmento	Saldo a capital categorico (masivo, medio, alto, top y ultratop)
PortAntiguo	1 Si cumple 0 si no.
descuentocap	Capital total menos el valor pagado
Negociaciones	Negociaciones que tuvo cada obligación antes de ser cancelada
campañas	Cuenta de campañas en las que estuvo involucrada la obligación

Tabla 2: variables elegidas para la construcción del modelo

5. Análisis Descriptivo

A continuación se presentan los gráficos de distribución individuales de las variables elegidas para el modelo:

¹Se anexa a este documento

5.1. Gráficos de participación De Las Variables Categóricas

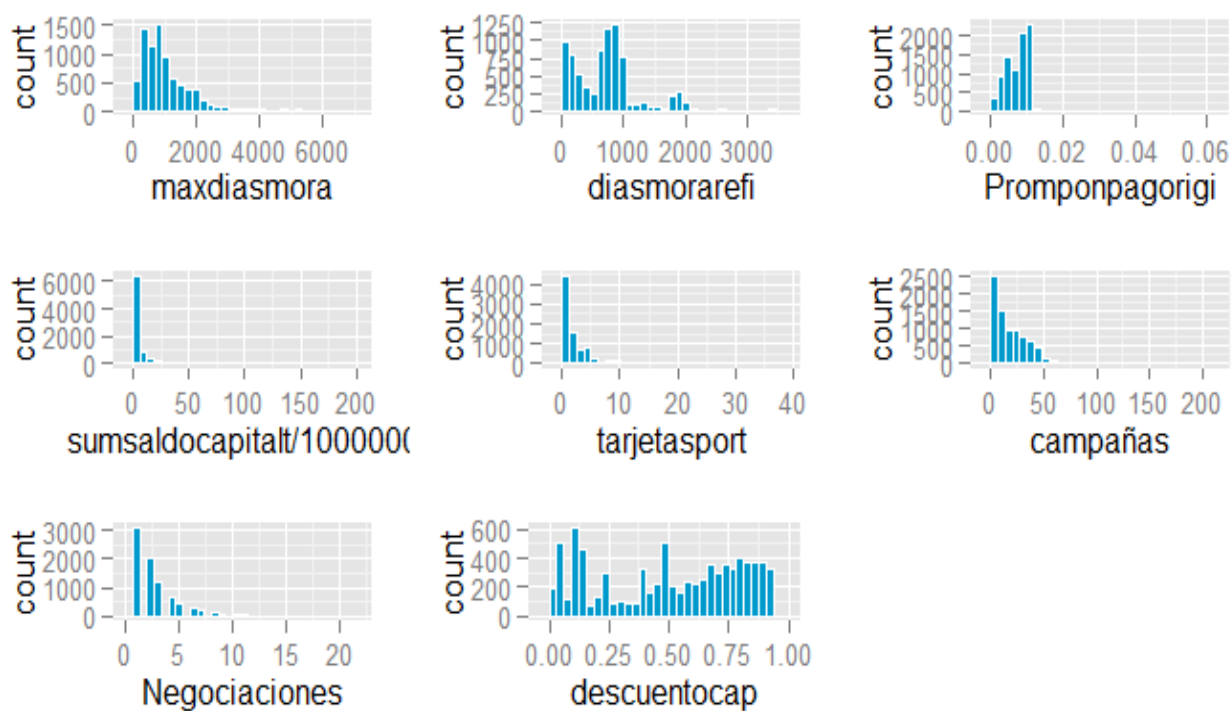
Distribuciones por categorías



H

5.2. Gráficos de Distribución De Las Variables Continuas

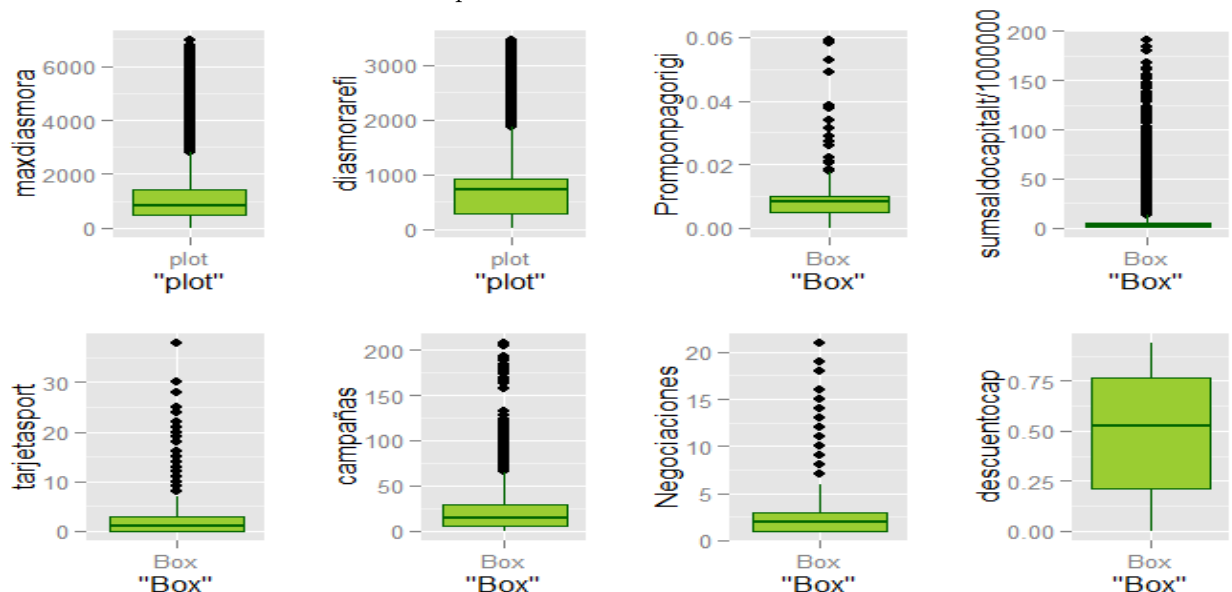
Distribuciones de las variables continuas



H

5.3. Gráficos De Box Plot

Box plot de variables continuas



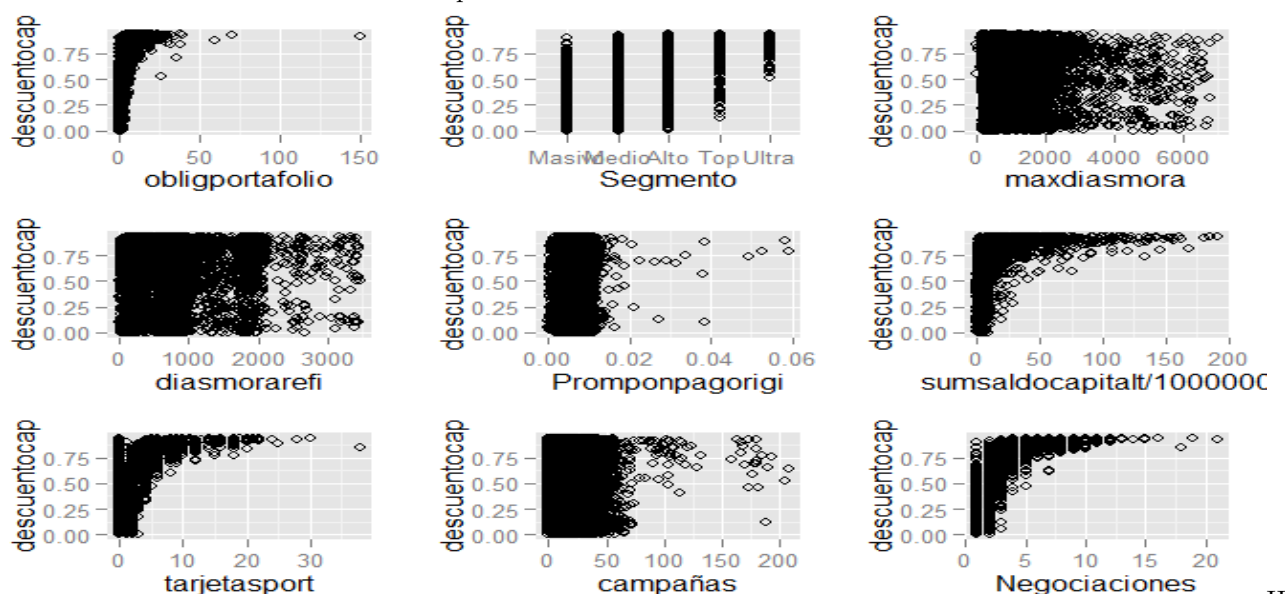
H

- En los gráficos de distribuciones de las variables continuas además de observar cada una de las variables, podemos identificar que la variable respuesta de nuestro modelo (descuentocap) no presenta una forma trivial de identificar y es una razón adicional, por la cual debemos usar la regresión beta.

- Mediante los boxplot podemos identificar los comportamientos atípicos individuales de las variables, y podrían servir de referencia para excluir una variable según sea el caso.

5.4. Gráficos De Dispersión

Dispersiones de las variables



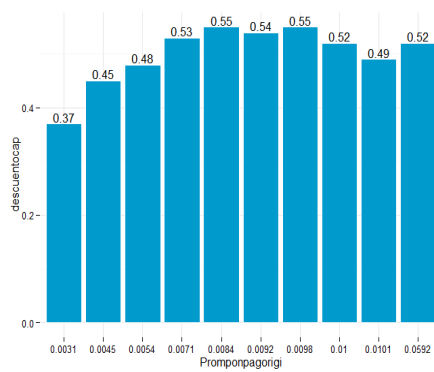
H

Los gráficos de dispersión se construyen con la variable respuesta y cada una de las variables a incluir en el modelo. En ellos podemos observar comportamientos de los datos, los cuales nos pueden sugerir algunas transformaciones en caso de necesitar un mejor ajuste del modelo.

5.5. Gráficos De Distribución De la Variable Respuesta Vs Variables Explicativas

Variable Respuesta Vs Variable Explicativa

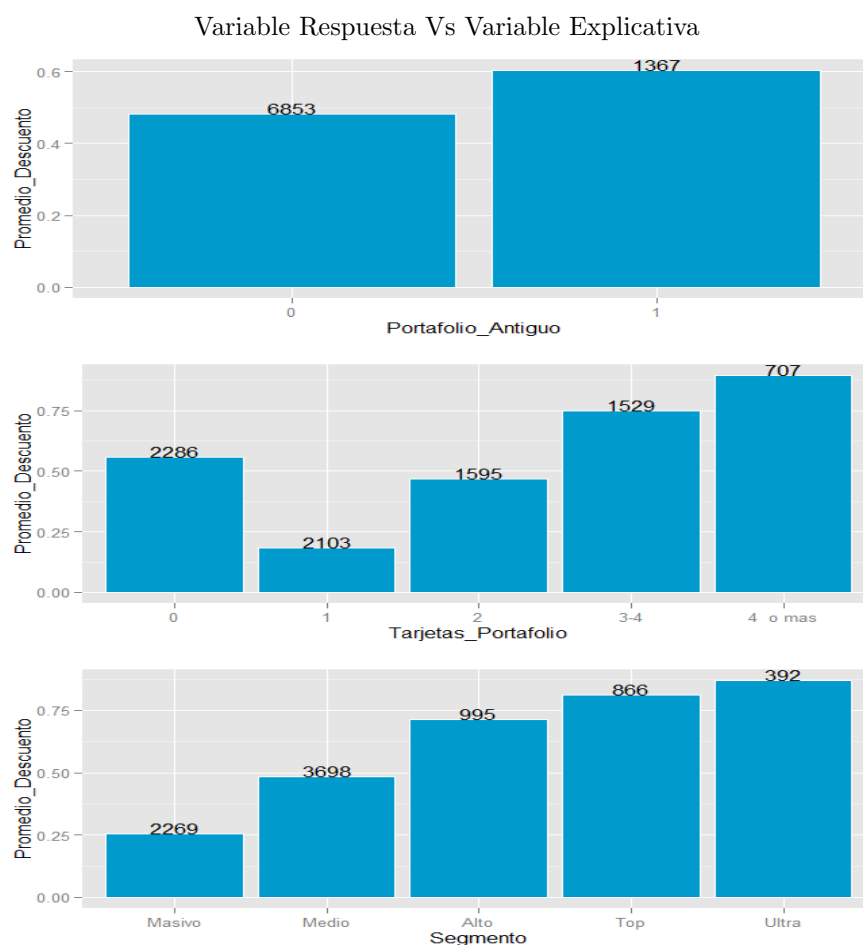




Los gráficos de distribución de las variables explicativas contra la variable respuesta, son de utilidad para identificar comportamientos que en los descriptivos no se logran observar. Esto sirve como herramienta para establecer si hay necesidad de aplicar transformaciones y/o categorizaciones a las variables explicativas para que sean capturadas adecuadamente por el modelo.

En la siguiente sección observaremos los gráficos de distribución de variables categoricas vs la variable respuesta.

5.6. Gráficos de distribución de variables categoricas vs la variable respuesta



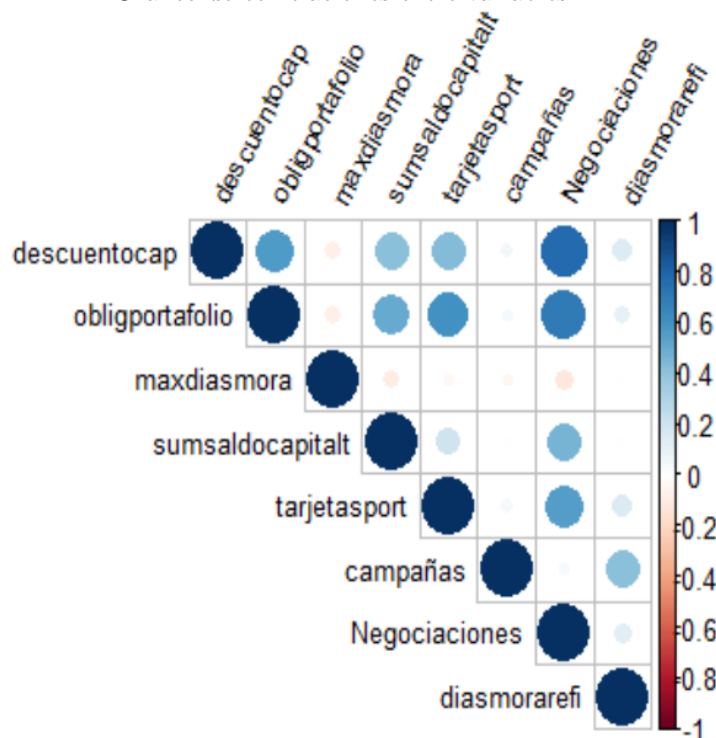
Los tres gráficos anteriores se deben interpretar cautelosamente. La altura de la barra significa el porcentaje promedio de descuento, la etiqueta en el número de clientes en esa categoría y el eje X es la categoría. De esta manera se puede observar la distribución por categoría tanto del número de clientes como del descuento.

6. Construcción Del Modelo

Analisis de correlación y multicolinealidad

Iniciamos la construcción del modelo evaluando la correlación que existe entre las variables elegidas para la construcción del modelo. Aquí visualmente podemos darnos una idea de si existe o no multicolinealidad entre las variables.

Gráfico de correlaciones entre variables



Como el gráfico no es suficiente para determinar si existe o no multicolinealidad, usaremos el FIV para confirmar que entre las variables elegidas para el modelo no existe multicolinealidad significativa:

```
at1=lm(descontocap PortAntiguo+obligportafolio+maxdiasmora+sumsaldocapital+tarjetasport+campañas+
Negociaciones+diasmorarefi+Segmento+Promponpagorigi ,data=basetes1)
vif(at1)
```

Resultado:

VARIABLE	VIF
PortAntiguo	3.11
obligportafolio	2.63
maxdiasmora	1.14
sumsaldocapitalt	4.30
tarjetasport	1.80
campanas	1.32
Negociaciones	2.55
diasmorarefi	3.37
Segmento	6.90
Promponpagorigi	1.27

Usando el criterio que algunos autores proponen de usar un FIV menor a 10 puede determinar que los datos no presentan multicolinealidad; descartamos la multicolinealidad entre las variables.

6.1. Aplicación de la regresión beta utilizando el paquete betareg de R.

El primer paso que realizamos antes de aplicar la función betareg de R, es tomar una muestra aleatoria del 70 % de los datos para entrenar el modelo y con el 30 % restante evaluar las predicciones.

Posteriormente evaluamos la función betareg del paquete betareg², la cual tiene inicialmente dos entradas; el modelo y la data

```
betareg(formula = modelo, data )
```

iniciamos la ejecución de modelos buscando el de mejor ajuste realizando los siguientes pasos:

Paso 1

Al realizar el primer modelo("primod") después de probar 25 transformaciones e interacciones de las variables para conseguir un mejor ajuste, no encontramos un cambio significativo en el Pseudo R-squared siendo este nuestro criterio de comparación de los modelos, por lo tanto relacionamos el resultado del modelo inicial.

Paso 2

Aplicamos el cambio en la función link a la opción "loglog", ya que por defecto el paquete betareg utiliza "log". En seguida También se probaron modelos independientes para cada uno de los segmentos, sin conseguir un mejor resultado ya que el Pseudo R-squared se encontraba por debajo de 0.50 en los 5 modelos

Paso 3

Probamos en total 31 transformaciones obteniendo la del paso 3 como la de mejor ajuste. Esta consiste en tomar la raíz cuadrada de las variables "sumsaldocapitalt", "tarjetasportz "negociaciones"

²otras entradas se pueden apreciar en la ayuda del paquete de R

Paso 4

Eliminamos la variable portantiguo ya que en el modelo primodm2 es no significativa, la variable "dias-morarefi" que también es no significativa no se quita teniendo en cuenta que en la reunión de expertos se manifestó que era importante para la compañía al igual que "sumsaldocapitalt".

A continuación presentamos el resumen de los resultados:

Resumen de resultados				
Paso	Paso 1	Paso 2	Paso 3	Paso 4
Cambio	-	Función Link	Transformacion de variables	-
Modelo	primod	primodm1	primodm2	modelf
Pseudo R-squared	0,5843	0,7287	0,8129	0,8129
AIC	-6291,4	-7292,321	-8750,912	-8752,789
Promedio Observado	0,4974869	0,4974869	0,4974869	0,4974869
Predicción Promedio	0,4788721	0,4776936	0,4803174	0,4803239
Promedio Absoluto del Error	0,1182161	0,1048378	0,08895171	0,08891908
Promedio Error	0,01861481	0,01979335	0,01716949	0,01716296

Las transformaciones realizadas se adoptaron después de verificar mediante ensayo y error las formas que presentan las variables observadas en los gráficos descriptivos y exploratorios.

El resultado del modelo lo podemos verificar en la siguiente salida:

```
betareg(formula = descuentocap ~ obligportafolio + maxdiasmora + campañas
+sqrt(sumsaldocapitalt) + sqrt(tarjetasport) + sqrt(Negociaciones) +
+diasmorarefi + Segmento + Promponpagorigi, data = ebaseto, link = "loglog")
```

```
Standardized weighted residuals 2:
      Min      1Q   Median      3Q      Max
-10.8071  -0.5006   0.0536   0.6058   4.5553
```

```
Coefficients (mean model with loglog link):
              Estimate Std. Error z value Pr(>|z|)
(Intercept) -2.308e+00  2.624e-02 -87.956 < 2e-16 ***
obligportafolio -1.330e-02  2.634e-03  -5.047 4.49e-07 ***
maxdiasmora      3.596e-05  5.878e-06   6.118 9.50e-10 ***
campañas        -4.676e-04  3.557e-04  -1.315 0.188561
sqrt(sumsaldocapitalt) 1.152e-04  1.501e-05   7.675 1.65e-14 ***
sqrt(tarjetasport)  4.163e-02  8.777e-03   4.742 2.11e-06 ***
sqrt(Negociaciones)  1.644e+00  2.032e-02  80.922 < 2e-16 ***
diasmorarefi      7.505e-06  1.108e-05   0.677 0.498334
SegmentoMedio     1.110e-01  1.719e-02   6.457 1.07e-10 ***
SegmentoAlto      2.424e-01  3.574e-02   6.782 1.19e-11 ***
SegmentoTop       2.047e-01  5.455e-02   3.752 0.000175 ***
SegmentoUltra    -1.487e-01  1.046e-01  -1.422 0.155136
Promponpagorigi  -1.245e+01  1.849e+00  -6.735 1.63e-11 ***
```

```
Phi coefficients (precision model with identity link):
```

```
              Estimate Std. Error z value Pr(>|z|)
(phi)      11.123      0.204    54.52 <2e-16 ***
```

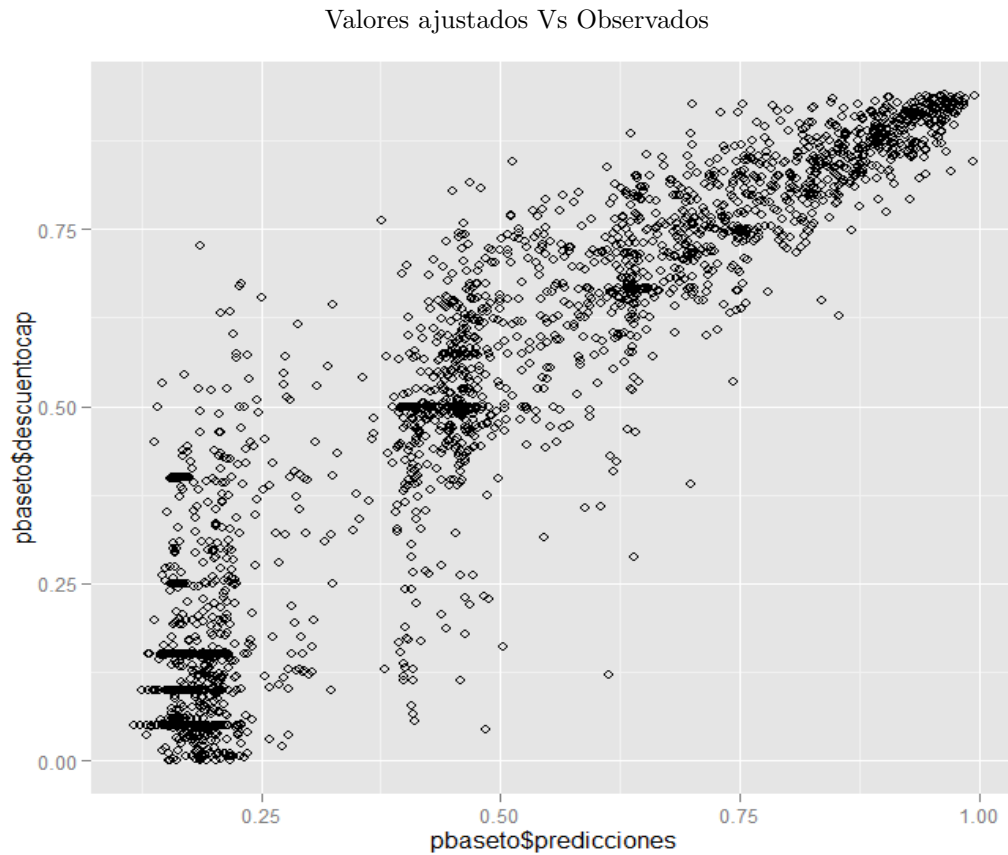
```
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
Type of estimator: ML (maximum likelihood)
Log-likelihood: 4390 on 14 Df
Pseudo R-squared: 0.8129
Number of iterations: 25 (BFGS) + 3 (Fisher scoring)
```

H

6.2. Valores observados contra los ajustados del modelo elegido:

El gráfico a continuación muestra los valores ajustados contra los observados, se observa en este que los pronósticos tienen una alta dispersión en los descuentos menores a 30 %. Sin embargo los valores ajustados por encima de ese porcentaje, presentan pendiente aparentemente de 45 grados que es lo que buscamos observar con esta gráfica.



6.3. Expresión del modelo de regresión beta:

Como la función link o de enlace que usamos es la "loglog"

$$g(\mu) = -\log(-\log(\mu))$$

al despejar μ de la función, el modelo de regresión beta quedaria expresado

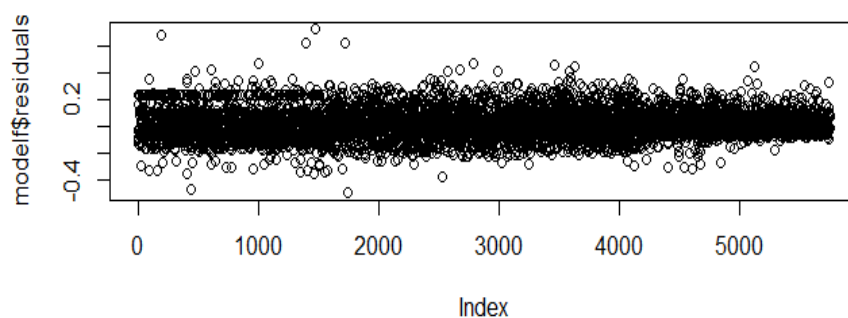
$$\hat{\mu} = e^{-e^{-x'\beta}}$$

donde x y β son los vectores de variables y coeficientes del modelo respectivamente

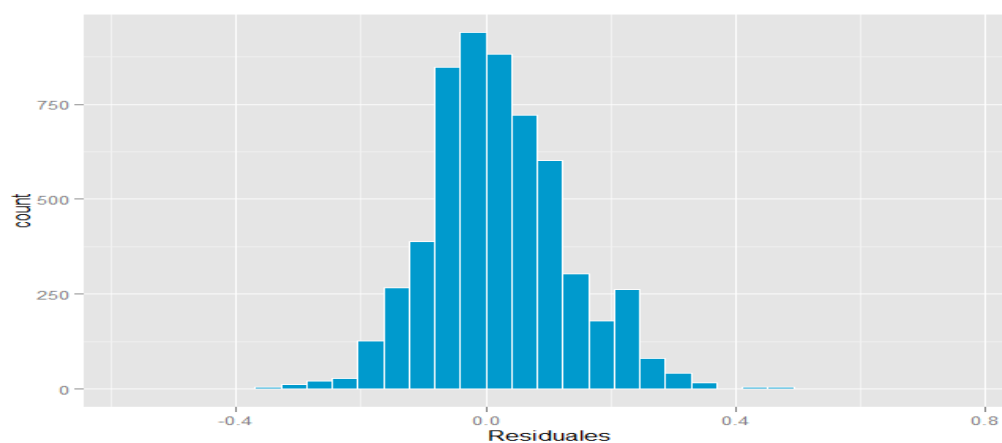
x	β
Intercepto	-2.308
obligportafolio	-0,0133
maxdiasmora	0,00003596
campañas	-0,0004676
$\sqrt{\text{sumsaldocapitalt}}$	0,0001152
$\sqrt{\text{tarjetasport}}$	0,04163
$\sqrt{\text{Negociaciones}}$	1,644
diasmorarefi	0,000007505
SegmentoMedio	0,111
SegmentoAlto	0,2424
SegmentoTop	0,2047
SegmentoUltra	-0,1487
Promponpagorigi	-12,45

6.4. Normalidad de los residuales:

Residuales del modelo



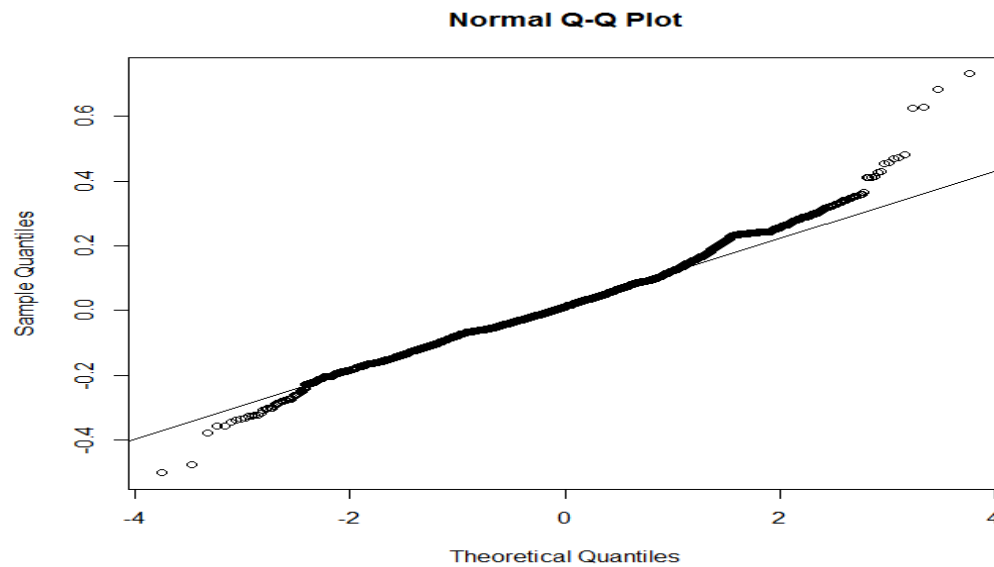
Distribución de los residuales



Visualmente en la gráfica los datos parecieran distribuirse normales.

La hipótesis de normalidad se suele comprobar mediante un QQ plot de los residuos:

QQ plot de los residuales



Dado que los puntos están bastante alineados, la normalidad también parece aceptable.

Continuamos con una prueba de Kolmogorov - Smirnov para verificar la normalidad de los residuales del modelo(modelf)

```
data: modelf$residuals
D = 0.043231, p-value = 9.201e-10
alternative hypothesis: two-sided
```

Con un $\alpha = 0,05$ se rechaza la hipótesis nula, es decir que los residuales no se ajustan a una distribución normal.

7. Conclusiones

- La construcción de la base de datos es una tarea que requiere experiencia y habilidades con el manejo de datos. La mala construcción de la base de trabajo, puede llevar a cometer errores de pronóstico, que para las compañías se traduce en pérdidas y en muchas ocasiones podría comprometer la continuidad del negocio.
- Realizar reuniones periódicas con los expertos del área a tratar, evita desvíos en el trabajo investigativo. Adicionalmente se descubren errores de mal manejo de la información o mala interpretación a tiempo.
- El análisis exploratorio de los datos antes de la construcción de un modelo es de gran ayuda al momento de realizar transformaciones en las variables, debido a que se pueden identificar comportamientos que en algunas ocasiones no logran ser percibidos por las técnicas estadísticas, pero que pueden influir considerablemente en el resultado final.
- La evaluación previa de la multicolinealidad permite un desarrollo adecuado del modelo, ya que no identificarla antes de empezar la construcción del modelo, puede generar retrasos, errores y dudas, en ocasiones difíciles de identificar cuando los modelos se encuentran contruidos, generando retrasos considerables.
- Las transformaciones de las variables logran mejorar en muchas ocasiones los modelos. No obstante se debe tener prudencia con la transformación de la variable respuesta, ya que esto podría generar errores de interpretación y en muchas ocasiones no se puede devolver a la variable original.
- El resultado de la regresión beta en la estimación de descuentos por obligación en los clientes con cartera vencida, presenta un promedio de estimación cercano al promedio observado y un error promedio relativamente bajo.
- Los residuales según la prueba de Kolmogorov Smirnov no se ajustan a una distribución normal, sin embargo en los gráficos de histograma y QQ plot aparentemente estos si son normales. En la practica muchos de los supuestos son violados debido a que puede ser muy complejo cumplir con todos.
- Para mejorar el ajuste, los residuales se podría realizar un modelamiento de la varianza, también se podría intentar ajustar una regresión beta bayesiana.
- Una sugerencia para mejorar el modelo sería quitar todos aquellos clientes que cancelaron influenciados por campañas especiales, dado que estas tienen descuentos muy agresivos que pueden hacer que el modelo no se ajuste a la realidad

Referencias

- [1] REVISTA SALUD PÚBLICA MÉX VOL.48 N.5 , *Modelos de regresión para variables expresadas como una proporción continua* Cuernavaca Sep./Oct. 2006.
- [2] EDILBERTO CEPEDA CUERVO, *Bayesian beta regression models joint mean and precision modelig*,
- [3] SILVIA L.P. FERRARI, *Beta regression for modelling rates and proportions*,
- [4] [HTTP://WWW.RESEARCHGATE.NET/PUBLICATION/31553228](http://www.researchgate.net/publication/31553228),
- [5] [HTTP://WWW.UV.ES/URIEL/MATERIAL/MULTICOLINEALIDAD3.PDF](http://www.uv.es/uriel/material/multicolinealidad3.pdf),
- [6] [HTTPS://AZURE.MICROSOFT.COM/ES-ES/DOCUMENTATION/ARTICLES/SQL-DATABASE-GET-STARTED/](https://azure.microsoft.com/es-es/documentation/articles/sql-database-get-started/),
- [7] [HTTPS://WWW.1KEYDATA.COM/ES/SQL/](https://www.1keydata.com/es/sql/),
- [8] [HTTPS://MSDN.MICROSOFT.COM/ES-ES/LIBRARY/DD264901.ASPX](https://msdn.microsoft.com/es-es/library/dd264901.aspx)
- [9] FRANCISCO CRIBARI-NIETO Y ACHIM ZEILEIS, *Beta Regression in R*, Journal of Statistical Software. 2010.

ANEXOS

Código SQL Server

WITH z

AS (

/***** Script para el comando SelectTopNRows de SSMS *****/

```

SELECT idclienteportafolio,
bbb.[idcliente],
bbb.[idportafolio],
bbb.obligportafolio,
bbb.fechacompra,
bbb.maxdiasmora,
bbb.maxfechacancelacion,
bbb.minfechadesembolso,
bbb.totaldeudaadquiridat,
( bbb.pagorixdeuda / bbb.totaldeudaadquiridat ) AS
promponpagorigi,
bbb.sumsaldocapitalt,
bbb.interesescorrientest,
bbb.interesesmorat,
bbb.interesesotrost,
bbb.interesesrftnt,
bbb.tarjetasport,
Max(d.[idacuerdo])          ultacuerdo
FROM   (SELECT idclienteportafolio,
[idcliente],
[idportafolio],
Count(idclienteportafolio) obligportafolio,
fechacompra,
Max([diasmora])              MaxDiasMora,
Max([fechacancelacion])     MaxFechaCancelacion,
Min([fechadesembolso])      minFechaDesembolso,
Sum([totaldeudaadquirida])   TotalDeudaAdquiridaT,
Sum(pagorixdeuda)            pagoriXdeuda,
Sum([saldocapital])          SumSaldoCapitalT,
Sum([interesescorrientes])   InteresesCorrientesT,
Sum([interesesmora])         InteresesMoraT,
Sum([interesesotros])        InteresesOtrosT,
Sum(interesesrftnr)          InteresesRFNT,
Sum(tipocredito)             TarjetasPort
FROM   (SELECT ( Cast(a.idcliente AS VARCHAR) + '-' + Cast(a.idportafolio AS VARCHAR) ) AS idcliente,
a.[idcliente],
a.[idportafolio],
b.fechacompra,
a.[diasmora],
a.[fechacancelacion],
a.[fechadesembolso],
a.[totaldeudaadquirida]
,(
a.[porcentajepagooriginador] * a.[totaldeudaadquirida] ) AS pagoriXdeuda,
a.[saldocapital],

```

```

a.[interesescorrientes],
a.[interesesmora],
a.[interesesotros],
c.[interesesrfn] AS InteresesRFNR,
CASE WHEN a.[tipocredito] LIKE '%tarjeta%' THEN 1
ELSE 0 END AS TipoCredito
FROM [SAB2NewExports].[dbo].[negocios] a
LEFT JOIN [SAB2NewExports].[dbo].[portafolios] b
ON a.idportafolio = b.idportafolio
LEFT JOIN
[SAB2NewExports].[dbo].[negociaciones_negocios] c
ON a.[idrefinancia] = c.[idrefinancia]
WHERE a.estadonegocio IN ( 'cancelado' )
AND a.idportafolio IN ( 1, 2, 3, 4,
5, 6, 7, 8,
9, 10, 12, 13,
14, 15, 16, 17,
18, 19, 20, 21,
22, 23, 24, 25,
26, 27, 28, 29,
30, 31, 32, 33,
34, 35, 36, 37,
38, 39, 40, 41,
42, 43, 44, 45,
47, 48, 49, 50,
51, 52, 53, 54,
55, 56, 59, 60,
61, 71, 72, 74,
77, 86, 89, 90,
100, 101, 105, 106,
113, 114, 115, 117,
120, 125, 127, 131,
132, 135, 136, 137,
140, 141, 151, 155,
156, 162, 163, 164,
165, 166, 167, 168, 169
)) AS aaa
GROUP BY [idcliente],
[idportafolio],
fechacompra,
idclienteportafolio) AS bbb
LEFT JOIN sab2newexports.[dbo].[acuerdos] d
ON bbb.idclienteportafolio = (
Cast(d.idcliente AS VARCHAR) + '-' + Cast(d.idportafolio AS VARCHAR)
)
WHERE totaldeudaadquiridat >= 1
GROUP BY idclienteportafolio,
bbb.[idcliente],
bbb.[idportafolio],
bbb.obligportafolio,
bbb.fechacompra,
bbb.maxdiasmora,
bbb.maxfechacancelacion,

```

```

bbb.minfechadesembolso,
bbb.totaldeudaadquiridat,
bbb.sumsaldocapitalt,
bbb.interesescorrientest,
bbb.interesesmorat,
bbb.interesesotrost,
bbb.interesesrfnt,
bbb.tarjetasport,
bbb.pagorixdeuda
),
y AS (SELECT z.idclienteportafolio,
z.[idcliente],
z.[idportafolio],
z.obligportafolio,
z.fechacompra,
z.maxdiasmora,
z.maxfechacancelacion,
z.minfechadesembolso,
z.totaldeudaadquiridat,
z.promponpagorigi,
z.sumsaldocapitalt,
z.interesescorrientest,
z.interesesmorat,
z.interesesotrost,
z.interesesrfnt,
z.tarjetasport,
z.ultacuerdo,
Acup.idacuerdo,
Ac.tipoacuerdo,
Max(totalcuotaspagadas) TLCuotasAcuerdo,
ac.fechaacuerdo,
Max(valoracuerdo)          ValorAcuerdo
FROM      z
LEFT JOIN sab2newexports.dbo.acuerdos_planpagos Acup
ON z.ultacuerdo = Acup.idacuerdo
LEFT JOIN sab2newexports.dbo.acuerdos Ac
ON Acup.idacuerdo = Ac.idacuerdo
GROUP BY z.idclienteportafolio,
z.[idcliente],
z.[idportafolio],
z.obligportafolio,
z.fechacompra,
z.maxdiasmora,
z.maxfechacancelacion,
z.minfechadesembolso,
z.totaldeudaadquiridat,
z.promponpagorigi,
z.sumsaldocapitalt,
z.interesescorrientest,
z.interesesmorat,
z.interesesotrost,
z.interesesrfnt,
z.tarjetasport,

```

```

z.ultacuerdo,
Acup.idacuerdo,
Ac.tipoacuerdo,
ac.fechaacuerdo),
base
AS (SELECT * FROM (SELECT y.idclienteportafolio,
y.idcliente,
y.idportafolio,
y.obligportafolio,
y.fechacompra,
y.maxdiasmora,
Cast(( y.maxfechacancelacion - fechacompra ) AS INT)
Diasmorarefi,
y.maxfechacancelacion,
y.minfechadesembolso,
y.totaldeudaadquiridat,
CASE
WHEN y.promponpagorigi < 0 THEN 0
ELSE y.promponpagorigi
END
promponpagorigi
,
y.sumsaldocapitalt,
y.interesescorrientest,
y.interesesmorat,
y.interesesotrost,
y.interesesrft,
( y.sumsaldocapitalt + y.interesescorrientest
+ y.interesesmorat + y.interesesotrost
+ y.interesesrft )
Totaldeuda
,
y.tarjetasport,
y.ultacuerdo,
y.idacuerdo,
y.tipoacuerdo,
y.fechaacuerdo,
y.tlcuotasacuerdo,
CASE
WHEN y.tlcuotasacuerdo = 1 THEN 1
ELSE 0
END
Acuerdoalinstante,
y.valoracuerdo,
Sum (pagos.valor)
ValorPagado
-- ,( 1 - ( y.valoracuerdo / (
-- y.sumsaldocapitalt + y.interesescorrientest
-- + y.interesesmorat + y.interesesotrost) ) ) descuento
FROM y
LEFT JOIN sab2newexports.dbo.pagos
ON pagos.idacuerdo = y.idacuerdo
WHERE y.idacuerdo IS NOT NULL
AS
AS

```

```

AND y.valoracuerdo > 20000 and pagos.TipoOrigenPago in ('Consignacion','FlexiAbono',
'Otros Ingresos por Negociacion','Recaudo')
GROUP BY y.idclienteportafolio,
y.idcliente,
y.idportafolio,
y.obligportafolio,
y.fechacompra,
y.maxdiasmora,
y.maxfechacancelacion,
y.minfechadesembolso,
y.totaldeudaadquiridat,
y.promponpagorigi,
y.sumsaldocapitalt,
y.interesescorrientest,
y.interesesmorat,
y.interesesotrost,
y.interesesrnt,
y.tarjetasport,
y.ultacuerdo,
y.idacuerdo,
y.tipoacuerdo,
y.fechaacuerdo,
y.tlccuotasacuerdo,
y.valoracuerdo) AS qqqq
WHERE totaldeuda > 50000
AND ( maxfechacancelacion BETWEEN
CONVERT(DATETIME, '2010-01-01 00:00:00', 102)
AND
CONVERT(DATETIME, '2015-08-31 23:59:59',
102
) )
AND minfechadesembolso IS NOT NULL),
mult
AS (SELECT idcliente,
Count(idportafolio) Noportafolios
FROM base
WHERE fechacompra < maxfechacancelacion
GROUP BY idcliente),
n as (SELECT idclienteportafolio,
base.idcliente,
base.idportafolio,
obligportafolio,
fechacompra,
maxdiasmora,
diasmorarefi,
cast(maxfechacancelacion-minfechadesembolso as int) as Antiguedadoblig,
maxfechacancelacion,
minfechadesembolso,
totaldeudaadquiridat,
promponpagorigi,
sumsaldocapitalt,
interesescorrientest,
interesesmorat,

```

```

interesesotrost,
interesesrfnt,
totaldeuda,
tarjetasport,
ultacuerdo,
idacuerdo,
base.tipoacuerdo,
fechaacuerdo,
tlcuotasacuerdo,
acuerdoalinstante,
valoracuerdo,
valorpagado,
-- descuento,
noportafolios,
CASE
WHEN noportafolios = 1 THEN 1
ELSE 0
END AS Multiportafolio,
a.depto,
count(f.idcampana) campañas
FROM   base
LEFT JOIN mult
ON base.idcliente = mult.idcliente
LEFT JOIN SAB2NEWEXPORTS.DBO.CLIENTES a
ON base.idcliente=a.idcliente
LEFT JOIN sab2newexports.DBO.[CampanasComercialesXCliente] f
ON idclienteportafolio=(Cast(f.idcliente AS VARCHAR) +
'-'+ Cast(f.idportafolio AS VARCHAR)
)
WHERE noportafolios IS NOT NULL
AND diasmorarefi > 0
GROUP BY idclienteportafolio,
base.idcliente,
base.idportafolio,
obligportafolio,
maxdiasmora,
diasmorarefi,
maxfechacancelacion,
minfechadesembolso,
totaldeudaadquiridat,
promponpagorigi,
sumsaldocapitalt,
interesescorrientest,
interesesmorat,
interesesotrost,
interesesrfnt,
totaldeuda,
tarjetasport,
ultacuerdo,
idacuerdo,
BASE.tipoacuerdo,
fechaacuerdo,
tlcuotasacuerdo,

```

```

acuerdoalinstante,
valoracuerdo,
valorpagado,
noportafolios,
depto,
fechacomptra,
maxfechacancelacion,
minfechadesembolso,
ultacuerdo,
idacuerdo,
base.tipoacuerdo,
fechaacuerdo,
valoracuerdo,
noportafolios
--ORDER BY idclienteportafolio
)

select * into working.dbo.modelopreciosIni from n

--where (descuento between 0 and 1)
-----
with a as (select * from working.dbo.modelopreciosIni
),

base as (select a.[idclienteportafolio]
,a.[idcliente]
,a.[idportafolio]
,case when a.[idportafolio]<100 then 1 else 0 end as PortAntiguo
,a.[obligportafolio]
,a.[fechacomptra]
,a.[maxdiasmora]
,a.[diasmorarefi]
,a.[Antiguedadoblig]
,a.[maxfechacancelacion]
,a.[minfechadesembolso]
,a.[totaldeudaadquiridat]
,a.[promponpagorigi]/100 as Promponpagorigi
,a.[sumsaldocapitalt]
,a.[interesescorrientest]
,a.[interesesmorat]
,a.[interesesotrost]
,a.[interesesrftnt]
,a.[sumsaldocapitalt]+a.[interesescorrientest]+a.[interesesmorat]+
a.[interesesotrost] as deudatotal
,a.[tarjetasport]
,a.[ultacuerdo]
,a.[idacuerdo]
,a.[tipoacuerdo]
,a.[fechaacuerdo]
,a.[tlcuotasacuerdo]
,a.[acuerdoalinstante]
,a.[valoracuerdo]
,a.[valorpagado]

```

```
, (1-(a.[valoracuerdo]/(a.[sumsaldocapitalt]+a.[interesescorrientest]+
a.[interesesmorat]+a.[interesesotrost]))) as descuentot
,a.[noportafolios]
,a.[Multiportafolio]
,a.[depto]
,a.[campañas]
,count(distinct(b.idnegociacion)) as Negociaciones
from a
left join sab2newexports.dbo.negociaciones b
```

```
on a.idclienteportafolio=(cast(b.idcliente AS VARCHAR) +
'-'+ Cast(b.idportafolio AS VARCHAR))
```

```
where [FechaNegociacion]<[maxfechacancelacion]
group by a.[idclienteportafolio]
```

```
,a.[idcliente]
,a.[idportafolio]
,a.[obligportafolio]
,a.[fechacompra]
,a.[maxdiasmora]
,a.[diasmorarefi]
,a.[Antiguedadoblig]
,a.[maxfechacancelacion]
,a.[minfechadesembolso]
,a.[totaldeudaadquiridat]
,a.[promponpagorigi]
,a.[sumsaldocapitalt]
,a.[interesescorrientest]
,a.[interesesmorat]
,a.[interesesotrost]
,a.[interesesrfnt]
,a.[totaldeuda]
,a.[tarjetasport]
,a.[ultacuerdo]
,a.[idacuerdo]
,a.[tipoacuerdo]
,a.[fechaacuerdo]
,a.[tlcuotasacuerdo]
,a.[acuerdoalinstante]
,a.[valoracuerdo]
,a.[valorpagado]
,a.[noportafolios]
,a.[Multiportafolio]
,a.[depto]
,a.[campañas])
```

```
select *
into working.dbo.modelopreciosdos
from base
```

```
select * from working.dbo.modelopreciosdos
where (interesescorrientest+interesesmorat+interesesotrost) =0
```

```

-----
-----
drop table working.dbo.modeloprecios3;

with a as (Select *, (valorpagado-sumsaldocapitalt) as factor from modelopreciosdos
where sumsaldocapitalt>0 and valorpagado>0)

,b as (

select *, CASE
WHEN factor < 0 THEN (valorpagado/sumsaldocapitalt)
ELSE 1
END AS ppagocapital from a)

,c as (
select *, CASE
WHEN ppagocapital=1 and (interesescorrientest+interesesmorat+interesesotrost)=0 THEN 1
ELSE 0
END AS filtro  from b)

, d as (
select *,CASE
WHEN ppagocapital < 1 THEN 0
ELSE factor/((interesescorrientest+interesesmorat+interesesotrost)+0.0001)
END AS ppagointereses from c where filtro=0
),

f as (select distinct depto, sede from sab2newexports.dbo.clientes where pais in
('Colombia')
and (depto+'-'+sede) not in
('SANTANDER-Bogotá','EXTERIOR-Bogotá','CAUCA-Bogotá','ANTIOQUIA-Bogotá',
'CORDOBA-Bogotá','LA GUAJIRA-Bogotá')),

e as (
select idclienteportafolio,
idcliente,
idportafolio,
PortAntiguo,
obligportafolio,
fechacompra,
maxdiasmora,
diasmorarefi,
Antiguedadoblig,
maxfechacancelacion,
minfechadesembolso,
totaldeudaadquiridat,
Promponpagorigi,
sumsaldocapitalt,
interesescorrientest,
interesesmorat,
interesesotrost,
interesesrfnt,
deudatotal,

```

```

tarjetasport,
ultacuerdo,
idacuerdo,
tipoacuerdo,
fechaacuerdo,
tlcuotasacuerdo,
acuerdoalinstante,
valoracuerdo,
valorpagado,
noportafolios,
Multiportafolio,
d.depto,
campañas,
Negociaciones,
(1-ppagocapital) descuentocap,
ppagointereses
,sede
from d

left join f as sed on d.depto=sed.depto
-- filtro para que el descuento sea mayor o igual a 6%
where (ppagocapital)>=0.06
)
,j as (

select *,case when campañas <=25 then 1
when campañas between 25 and 101 then 2 else 3 end as campañascat
from e)

,l as (
select *
from j
where maxdiasmora<Antiguedadoblig and Antiguedadoblig<7301
and totaldeudaadquiridat<200000000
and Promponpagorigi<1 and sumsaldocapitalt<200000000
and interesescorrientest<200000000
and interesemorat<200000000 and ppagointereses<1000 and maxdiasmora<10000)
-----
select * into working.dbo.modeloprecios3 from l
-----
select * from working.dbo.modeloprecios3 where maxdiasmora<Antiguedadoblig and
Antiguedadoblig<7301
and totaldeudaadquiridat<200000000 and Promponpagorigi<1
and sumsaldocapitalt<200000000 and interesescorrientest<200000000
and interesemorat<200000000 and ppagointereses<1000
-----
select * from working.dbo.modeloprecios3
where sede is not null

```

Acuerdo de Confidencialidad

Agradecimientos

A mi madre y mis hermanas por su apoyo incondicional. A los docentes de la universidad Santo Tomas que con nobleza y entusiasmo compartieron sus conocimientos conmigo, a las directivas de Refinancia S.A. y a mi directora de tesis quienes me guiaron en la realización de este trabajo.