Universidad de Valladolid
Facultad de Ciencias

TRABAJO FIN DE GRADO

Grado en Estadística

Métodos de predicción de fuga con grandes volúmenes de datos

Autor:
D. Raquel García Fernández

Tutor/es:
D. Eusebio Arenal Gutiérrez

Índice

1 Introducción

DATASEGUROS es una empresa del sector seguros líder en el mercado. Su producto estrella es el seguro “Premium-Hogar”, ya que es el producto con el que consiguen mayor rentabilidad. La empresa trabaja fundamentalmente con seguros (de hogar, vida, salud), y de forma adicional trabaja con otros productos bancarios (cuentas de ahorro, tarjetas de crédito y débito, inversión).

DATASEGUROS observa que la proporción de clientes que dan de baja el seguro “Premium Hogar” es muy alta entre los clientes que tienen contratados además productos de ahorro o inversión.

Para proponer solución a dicho problema, la empresa se plantea obtener un modelo de propensión a la fuga de dicho producto con objeto de hacer mayor esfuerzo comercial sobre aquellos clientes que tengan mayor probabilidad de darlo de baja, para intentar retenerlos.

Para ello se van a realizar varios modelos utilizando los siguientes métodos estadísticos:

- **Árboles de decisión y regresión logística:** con los que se estimará para cada cliente la probabilidad de fuga y su vida esperada.

- **Análisis de supervivencia:** con el que vamos a estimar el tiempo que tardan en dar de baja el producto desde que lo contrataron por primera vez.

Además, otro de los objetivos es comparar los resultados obtenidos con cada uno de los estos métodos dentro del ejemplo concreto que estamos abordando.

Se utilizará como herramienta analítica el software SAS, en concreto los módulos SAS Guide y SAS Enterprise Miner.

2 Metodología

En este capítulo se explicarán cada uno de los métodos estadísticos que se han utilizado a lo largo del documento.

2.1 Análisis descriptivo univariante

Es necesario realizar en primer lugar un análisis descriptivo de los datos, ya que es muy importante conocer la calidad de la información con la que se cuenta para explicar el problema.
Se representan los siguientes estadísticos, dependiendo de si se trata de una variable numérica o categórica:

- Para las variables **numéricas** se muestra:

  – Histograma
  – Nº de observaciones
  – Nº de valores perdidos
  – Nº de valores distintos que tiene la variable
  – Proporción de registros que toman el mismo valor
  – Nº de valores que toman valor infinito
  – Nº de registros con valores negativos
  – Nº de registros en los que la variable tome valor cero
  – Nº de registros con valores positivos
  – Se comprueba si la moda es 0
  – Se comprueba si hay valores atípicos por la izquierda. Se consideran valores atípicos por la izquierda a todos los valores que están por debajo del percentil veinticinco menos tres veces el rango intercuartílico:

$$X_i < PCT_{25} - 3 \cdot (PCT_{75} - PCT_{25})$$

  – Se comprueba si hay valores atípicos por la derecha. Se consideran valores atípicos por la derecha a todos los valores que están por encima del percentil setenta y cinco más tres veces el rango intercuartílico:

$$X_i > PCT_{75} + 3 \cdot (PCT_{75} - PCT_{25})$$

  – Mínimo
  – Máximo
  – Percentiles: Representados como PCTxx, siendo xx el porcentaje de las observaciones que deja por debajo
  – Media
  – Desviación típica

    Tanto para la media como para la desviación típica, en el caso de que la moda sea cero, se excluye el cero en el cálculo puesto que en este caso interesa

conocer los valores condicionados a que el cliente tenga el producto (valores distintos de cero).

- Para las variables **categóricas** se representa el gráfico de sectores para observar la distribución de cada una de las variables. En caso de haber valores perdidos se mostrará como otra categoría más.

El realizar un análisis descriptivo de las variables antes de comenzar a desarrollar un modelo es fundamental, no solamente para conocer la distribución de las mismas, sino además para determinar qué **variables no podemos utilizar** o para cuáles es necesario realizar algún **tratamiento especial**, como es el caso de:

- Variables recomendadas descartar por tomar el mismo valor en todos los casos o por tener algún valor/categoría predominante (es conveniente rechazarlas).
- Variables categóricas con muchas categorías distintas, ya que aumentarían la complejidad del modelo.
- Variables con valores perdidos: Si son pocos, es conveniente imputarlos pero si son muchos se recomienda descartar la variable.
- Si un mismo registro tiene muchas variables sin informar, se recomienda descartarlo siempre que sea posible.

2.2 *Árboles de decisión*

2.2.1 Descripción del modelo

Los árboles de decisión son modelos estadísticos en los que interesa explicar una variable dependiente cualitativa (en nuestro caso binaria) en función de varias variables explicativas.

El método consiste en dividir el tablón de datos en varios subconjuntos descendientes según la proporción de casos positivos de la variable objetivo (proporción de casos con Y=1) que se concentre en cada uno de ellos.

De esta forma, todas las observaciones se encuentran inicialmente en un mismo grupo, en el cual hay tanto casos positivos como negativos para la variable objetivo con una proporción de casos positivos igual a p (llamada prior o probabilidad inicial del evento). El método consiste en dividir este grupo en varios subconjuntos, de forma que cada uno de ellos tenga distinta proporción de casos positivos. Esta división se realiza a través de la variable independiente que sea más discriminante en cada caso y por el punto de corte más óptimo (el que mejor discrimina). Este proceso se realiza de forma recursiva hasta que se cumplan los criterios de parada que junto con los criterios de corte se comentarán en detalle más adelante.

La manera de representarlo es a través de un árbol: el grupo inicial se denomina **nodo raíz**, el cual se divide en distintos nodos conectados entre sí a través de ramas. Las distintas ramas del árbol terminarán en un nodo **hoja** cuando se cumpla el criterio de parada establecido. De esta forma, cada nodo hoja recogerá un segmento de la población que caracterizará un patrón de comportamiento respecto la variable objetivo. Cada nodo hoja junto con todos sus nodos predecesores hasta llegar al nodo raíz forma una **regla**, la cual tendrá asignado un valor de la variable respuesta.

Los criterios de corte y de parada más habituales cuando estamos ante una variable objetivo binaria (ya que es el caso que estamos abordando) son:

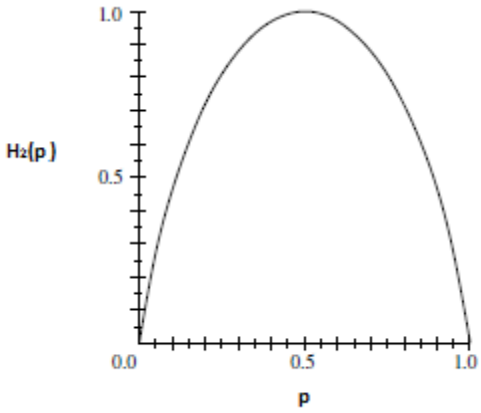
1. **Criterios de corte** ***(splitting)***

   - **Índice de entropía**:

     La entropía es una medida de la variabilidad de los datos que se obtiene de la siguiente forma:

     $$H_2(p) = -p \cdot \log_2(p) - (1-p) \cdot \log_2(1-p)$$

     siendo p la proporción de casos positivos. Toma valores entre 0 y 1:



Se calcula el índice de entropía para cada una de las variables: Si la variable es categórica, se obtiene sumando el índice de entropía de cada una de sus clases. En cambio, si es numérica, previamente se obtiene uno o varios puntos de corte por métodos iterativos.

Se elegirá aquella variable que tenga menor índice de entropía, que según puede observarse en la gráfica anterior, será aquella cuya proporción de casos positivos esté más alejada de 0,5.

- **Chi-cuadrado**:

  Mide el grado de asociación entre dos variables. Se calcula comparando la tabla de contingencia dada por el cruce de la variable objetivo versus cada una de las variables explicativas con una tabla donde no hubiera asociación.

  Se representa de la siguiente forma (veamos el caso en que la variable explicativa sea binaria):

	Y		
X	0	1	Total
a	n_{11}	n_{12}	$n_{1.}$
b	n_{21}	n_{22}	$n_{2.}$
Total	$n_{.1}$	$n_{.2}$	n

  Siendo:
  - n_{ij} cada uno de los valores observados para X=i y Y=j
  - $n_{i.}$ el número total de observaciones para los que X=i (también se denomina marginal fila)
  - $n_{.j}$ el número total de observaciones para los que Y=ji (también se denomina marginal columna)

– n número total de observaciones

El estadístico chi-cuadrado se calcula de la siguiente forma:

$$\chi^2_{(F-1)(C-1)} = \sum_{i=1}^{F}\sum_{j=1}^{C} \frac{\left(n_{ij} - e_{ij}\right)^2}{e_{ij}}$$

donde:

– F es el número de filas de la tabla de contingencia
– C el número de columnas
– e_{ij} la frecuencia esperada, la cual se obtiene multiplicando el total marginal columna por el total marginal fila dividido por el número total de observaciones

Se elegirá la variable explicativa que tenga mayor asociación con la variable objetivo, es decir, la variable con mayor valor del estadístico chi-cuadrado.

- **Índice de Gini**:

Mide la probabilidad de no sacar dos registros con el mismo valor para la variable objetivo dentro del mismo nodo. Se calcula de la siguiente forma:

$$G = 1 - \sum_{j=1}^{r} p_j^2$$

donde p_1,..,p_r son las probabilidades de que la variable objetivo sea igual a uno en cada uno de los grupos en los que se divide cada variable explicativa.

Cuando menor es el índice de Gini mayor es la pureza del corte. Por lo tanto, el corte propuesto en primer lugar será el de aquella variable que tenga menor valor del índice de Gini.

2. **Criterios de parada *(stopping)***

- **Máxima profundidad del árbol** (Maximum Depth): Se especifica el número de niveles que puede tener el árbol. Cuando se llega a dicho número de niveles no se realizan más particiones.

- **Máximo número de ramas por nodo** (Maximum Branch): Número de ramas máximo en que puede dividirse un nodo.
- **Número mínimo de observaciones por nodo final** (Leaf Size): Número mínimo de observaciones que tiene que tener un nodo final para que se construya la regla.
- **Número mínimo de observaciones para dividir un nodo** (Split Size): Número mínimo de observaciones que tiene que tener un nodo para que se pueda cortar por la variable seleccionada.
- **Variables discriminantes** (Discriminant Variables): No encontrar ninguna variable que sea lo suficientemente discriminante en el nodo es motivo de parada.

En último lugar es recomendable **podar** (pruning) el árbol. Consiste en reducirlo, haciéndolo más sencillo dejando sólo los nodos más importantes y a su vez eliminando los redundantes.

2.2.2 Pasos a seguir para construir un árbol de decisión

- ***Paso 1: Preparación del tablón de datos***

Para poder hacer modelos predictivos, es necesario dividir el conjunto de datos que disponemos de forma aleatoria en tres bloques:

- **Tablón de entrenamiento (training):** Conjunto de datos que se utilizará para entrenar el modelo. Con estos datos se construye el árbol.
- **Tablón de validación (validate):** Conjunto de datos que se utilizan para validar el modelo. Su función es detectar cuál es el modelo que comete menos error según el criterio que se haya definido. Estos datos no intervienen en la creación del árbol (es decir, en la selección de las variables y de los puntos de corte), solamente se aplica el modelo sobre ellos y se valida si se ajusta bien también a estos datos.
- **Tablón de prueba (test):** Conjunto de datos que no interviene ni en entrenamiento ni en validación. Una vez que se ha construido el modelo, es necesario utilizar los datos de prueba para evaluar la manera en que el modelo seleccionado se generaliza para datos que no jugaron ningún papel en la selección del mismo.

Antes de comenzar a construir el modelo es necesario comprobar cuál es la probabilidad a priori del evento que estamos analizando (prior). Si esta probabilidad es muy baja (menor del 5% aproximadamente pero depende la calidad de la información con la que se cuenta), no vamos a poder ser capaces de construir ningún modelo, va a

ser prácticamente imposible encontrar algún patrón en los datos que discrimine entre los casos positivos y los negativos de la variable dependiente. Para poder hacerlo es necesario balancear la muestra de datos.

Balancear es dar mayor peso al evento de nuestra variable objetivo, (es decir dar mayor peso a los casos positivos), con el fin de que los patrones que determinan que se cumpla el evento tengan peso suficiente en la construcción del modelo.

Balancear los datos demasiado también puede ser contraproducente, es decir, no podemos pasar de tener una penetración del 1% al 50% por ejemplo, ya que se está sesgando el patrón natural de los datos y podría darse el caso de que el modelo prediga bien solamente sobre la muestra balanceada, pero no sobre los datos originales (sin balancear). Por eso en la práctica es recomendable ir incrementándola poco a poco y parar en el momento en que veamos que el software que empleemos es capaz de construir un modelo.

- ***Paso 2: Selección del criterio de corte***

Se utilizan los tablones de entrenamiento y validación que se detalla en el punto anterior. A partir de estos tablones, se realizan varios árboles de forma automática variando el criterio de corte, con el objetivo de comprobar con cuál de ellos se obtienen mejores resultados.

Para evaluar cuál es el criterio elegido nos basaremos en los siguientes indicadores, los cuales deben comportarse de forma similar tanto para el tablón de entrenamiento como el de validación. En caso de no ser así, es síntoma de que el modelo obtenido no es un buen modelo, debido a que estaría muy ajustado a los datos con los que se ha entrenado:

- **Porcentaje de éxitos capturados (Cumulative % Capturated Response):**

Mide el porcentaje de casos positivos reales, fugas en nuestro caso, que el modelo es capaz de detectar (captar) en función del porcentaje de clientes que se seleccionen.

Si se extrae una muestra de n clientes de forma aleatoria el porcentaje de fugas captado en media (media de la variable número de fugas en la muestra) sería el mismo que el relativo al tamaño de la muestra respecto al número de datos. Si queremos obtener esta misma información teniendo en cuenta el modelo los clientes se ordenan de mayor a menor probabilidad de fuga y la muestra se forma con los n primeros clientes. La proporción de fugas captadas debería ser superior a la obtenida sin modelo.

Para poder comparar este indicador en muestras de distinto tamaño se obtiene el gráfico de “Porcentaje de éxitos capturados”, en el cual en el eje X se representa el porcentaje de clientes seleccionados y en el eje Y el de fugas captadas.

El mejor modelo será aquel para el que a menor porcentaje de clientes seleccionados se capte el mayor porcentaje posible de fugas.

- **Mejora acumulada (Cumulative Lift):**

Mide la ganancia acumulada del modelo en función del volumen de clientes seleccionados.
Se calcula dividiendo proporción de fugas capturadas por el modelo entre la proporción de fugas capturadas si no se tuviera modelo (selección aleatoria), ambas seleccionando la misma proporción de clientes.
El mejor modelo será aquel con el que se consiga mayor ganancia.

- **Curva ROC**

Mide el poder de clasificación del modelo. Los ejes del gráfico ROC representan lo siguiente:

– Sensitividad: proporción de verdaderos positivos (verdaderos positivos entre el total de positivos)

– 1-Especificidad: proporción de falsos positivos (falsos positivos entre el total de negativos)

Se basa por lo tanto en la matriz de clasificación:

	Real 0	Real 1
Estimado 0	VN	FN
Estimado 1	FP	VP

Donde:

– VN o especificidad es la proporción de verdaderos negativos: P(E=0/R=0)
– FP es la proporción de falsos positivos: P(E=1/R=0)
– FN es la proporción de falsos negativos: P(E=0/R=1)
– VP o sensitividad es la proporción de verdaderos positivos: P(E=1/R=1)

Los árboles de decisión asignan a cada cliente una probabilidad estimada de que cumpla el evento (Y=1). Un cliente se estima que va a ser un caso positivo, es decir, estimamos que va a tomar el valor 1, si la probabilidad que se le ha asignado con el modelo es mayor que la probabilidad a priori (tasa global de fuga sin modelo), o lo que es lo mismo, si la ganancia es >1.

El mejor modelo según este indicador será aquel que tenga mayor proporción de individuos clasificados correctamente:

$$\text{tasa global acierto} = \frac{\text{VP} + \text{VN}}{\text{n}}$$

siendo n el número total de individuos.

- **Otros estadísticos de ajuste:**

A modo resumen y con el fin de poder comparar fácilmente cada uno de los árboles obtenidos, se muestran los resultados de los siguientes indicadores para cada uno de ellos:

- **Probabilidad de clasificación errónea (Misclassification Rate):**
  Proporción de observaciones clasificadas en el grupo incorrecto (clasificado como 0 y en realidad es 1, o 1 y en realidad es 0). Cuanto menor sea este coeficiente mejor es la estimación del modelo

- **Suma de cuadrados del error (SSE):**
  Suma de cuadrados del error de predicción. Cuanto menor sea mejor es la estimación

- **Error medio (ASE o MSE):**
  Media de los cuadrados de los errores. Cuanto menor sea mejor es la estimación

- **Ganancia (Gain)**:
  Mide la ganancia que consigue el modelo en el primer decil. A mayor valor mejor es la estimación del modelo

- **Mejora (Lift)**:
  Mide la mejora que se consigue con el modelo versus una selección aletoria en el primer decil. A mayor valor mejor es la estimación del modelo

- **Índice ROC (ROC Index)**:
  Mide el área bajo la curva ROC, es decir, el poder de clasificación del modelo. El índice ROC toma valores de 0,5 (indica que el modelo clasifica igual que una selección aleatoria) a 1 (para un ajuste perfecto)

- **Coeficiente de Gini (Gini Coeficient):**
  Mide la probabilidad de no extraer dos registros de la misma clase dentro del mismo nodo para el modelo final obtenido. Cuando menor es el índice de Gini mayor es la pureza del corte.

Se seleccionará como mejor criterio, aquel con el que se hayan obtenido mejores resultados en cada uno de estos indicadores pero los más importantes en el caso que estamos abordando son:

- Porcentaje de éxitos capturados (Cumulative % Capturated Response)
- Ganancia acumulada (Cumulative Lift)
- Curva ROC
- Probabilidad de clasificación errónea

- ***Paso 3: Construcción y validación del árbol***

Con los árboles automáticos normalmente se pierde el sentido de negocio a la solución del problema que se quiere resolver: Así por ejemplo, puede darse el caso de que a la hora de hacer un corte, haya dos variables que explican prácticamente lo mismo sobre la variable objetivo. El árbol, elegirá una de las dos (en concreto la que tenga mayor valor según el criterio que se haya seleccionado, pero esta diferencia puede ser muy pequeña), pero es mucho más que probable que tenga mayor sentido de negocio una variable que la otra.

Una vez que hemos obtenido cuál es el criterio de corte con el que podemos conseguir mejores resultados, se entrenará un árbol de decisión de forma "manual" con dicho criterio. Para ello utilizamos también los tablones de entrenamiento y validación

- En primer lugar se tienen todas las observaciones del tablón de entrenamiento en un único nodo (nodo raíz). Vamos a "partir" este nodo por la variable más significativa que tenga sentido de negocio y que a su vez tenga mejor valor para el estadístico que se está evaluando dependiendo del criterio seleccionado. El software muestra en cada paso las variables ordenadas de mejor a peor según dicho estadístico, y de forma manual se elige la deseada.

- Para decidir cuál es el punto de corte, el software también da una recomendación basada en el criterio elegido, pero manualmente es posible modificarlo. Una vez aplicados los cambios se tendrán el árbol parcial que se ha obtenido.

- Para obtener un buen árbol es importante comprobar cada vez que se realiza un corte que:

  - La variable elegida discrimina correctamente la variable objetivo, es decir cada uno de los nodos terminales deben tener distinta proporción de casos positivos versus el nodo anterior.
  - Si se predicen las fugas con el árbol parcial sobre el conjunto de datos de validación se obtienen gráficos y valores de los estadísticos de ajuste similares a los obtenidos sobre el tablón de entrenamiento.

- Se continuarán realizando cortes de esta forma en cada uno de los nodos que se van formando hasta que se cumplen los criterios de parada.

- Una vez que se obtiene el modelo, es necesario comprobar si es un modelo válido (aplicar los mismos criterios que en los árboles automáticos). Especialmente nos fijaremos en:

  – Porcentaje de éxitos capturados (Cumulative % Capturated Response)
  – Ganancia acumulada (Cumulative Lift)
  – Curva ROC
  – Probabilidad de clasificación errónea

- Se mostrará a nivel técnico cuales son las reglas obtenidas

- Además se analizará la capacidad predictiva de cada una de las reglas.
  Para ello se construirá la siguiente tabla, la cual tiene una fila por cada una de las reglas obtenidas:

Orden	Prior	Ampliada	Reducida	Candidatos	Éxitos	Penetración	Ganancia	%Cand	%Éxitos

  – **Orden**: orden de cada una de las reglas.
  – **Prior**: Probabilidad de fuga (variable objetivo igual a 1) obtenida para cada regla en el entrenamiento del modelo
  – **Ampliada**: Indicador creado para identificar las reglas con ganancia superior a uno.
  – **Reducida**: Indicador creado para identificar las reglas de mayor ganancia con el fin de identificar las mejores reglas.
  – **Candidatos**: Número de clientes que se encuentran en cada una de las reglas.
  – **Éxitos**: Número de clientes que realmente se han fugado (variable objetivo=1) en cada una de las reglas.
  – **Penetración**: Proporción de clientes que han sido éxito en cada una de las reglas. Es el resultado de dividir el número de casos positivos entre el número total de individuos.
  – **Ganancia**: Ganancia proporcionada por cada una de las reglas. Se calcula dividiendo la penetración de cada regla entre la penetración del total.
  – **Proporción de candidatos**: es la proporción de individuos que caen en cada regla sobre el total.
  – **Proporción de éxitos**: es la proporción de casos positivos que hay en cada regla sobre el número de casos positivos que hay en total.

- A partir de la tabla anterior, se calculará de forma manual la matriz de clasificación junto con la tasa global de acierto.

- ***Paso 4: Test***

Consiste en medir los resultados del modelo obtenido en el Paso3 con un conjunto de datos que no han intervenido en el entrenamiento ni en la validación del mismo. De esta forma nos aseguramos que el modelo obtenido no se ajusta exclusivamente a los datos utilizados para construir el modelo.

Se contrastará la capacidad predictiva de cada una de las reglas del modelo y la matriz de clasificación con dichos datos de test, de forma que con este conjunto de datos se tiene que llegar a las mismas conclusiones que los obtenidos con los datos de entrenamiento y validación (resultados del paso3).

En caso de que se lleguen a las mismas conclusiones, ya se tendría el **modelo definitivo** construido a partir de los árboles de decisión, pero en caso de no ser así, habría que entrenar otro modelo interactivo (repetir desde el paso 3).

- ***Paso 5: Resultados obtenidos e interpretación***

En este paso se explicará:

- El **perfil** de los clientes de cada una de las reglas a partir de la combinación de variables y puntos de corte que las forman.
- Cuáles son las **variables** que más han aportado en la explicación de la variable objetivo ordenadas de mayor a menor importancia.
- Los **resultados** y **conclusiones** más importantes del modelo obtenido.
- Para cada cliente se estimará la **esperanza de vida**, la cual se calculará como el inverso de la probabilidad de fuga.

2.3 Regresión logística

2.3.1 Descripción del modelo

Los modelos de regresión logística son modelos estadísticos en los que se desea conocer la relación entre una variable dependiente binaria y una o más variables explicativas. La variable respuesta es la probabilidad de éxito (Y=1) para cada individuo, luego permitirá clasificarlos en una de las categorías de esta variable.

La ecuación de partida en los modelos de regresión logística es:

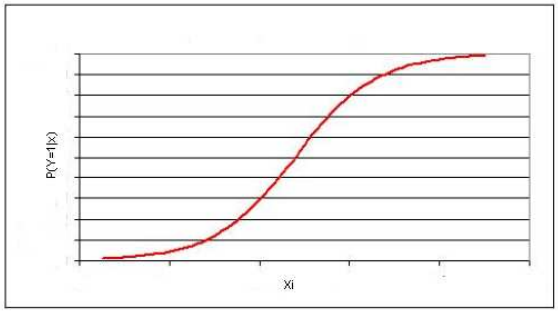
<u>1- Ecuación de partida:</u>
(ecuación logística)

$$P(y=1|x)=\frac{\exp(b_0+\sum_{i=1}^{n} b_i x_i)}{1+\exp(b_0+\sum_{i=1}^{n} b_i x_i)}$$

Donde

- P(y=1|x)=p probabilidad de que se cumpla el evento
- x_i: desde i=1 hasta n variables explicativas

Representación gráfica:
p vs X_i



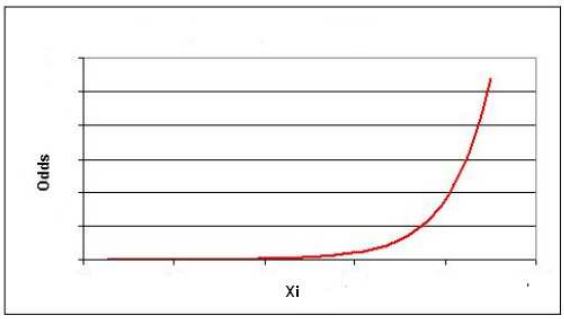
<u>2- Si se divide por su complementario:</u>

$$\frac{p}{1-p}=\exp\left(b_0+\sum_{i=1}^{n} b_i x_i\right)$$

Donde

- $\frac{p}{1-p}$ se denomina odds ratio

Representación gráfica:
Odds ratio vs X_i



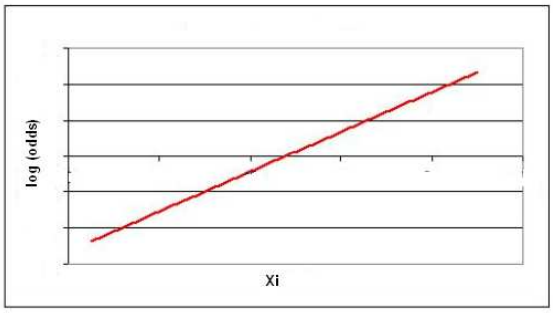
3- Transformación con logaritmo natural

$$\log\left(\frac{p}{1-p}\right) = b_0 + \sum_{i=1}^{n} b_i x_i$$

Donde

- $\log\left(\frac{p}{1-p}\right)$ = logit

Representación gráfica:
Log(odds) vs X_i



Se puede observar que el logaritmo de la odds ratio es una función lineal. Sin embargo, el modelo no puede tratarse como una regresión lineal puesto que los errores no siguen una distribución normal de media cero y varianza constante, sino una distribución binomial con media y varianza proporcionales al tamaño muestral y a la probabilidad de que se cumpla el evento en presencia de los regresores.

Sea i=k, se aprecia claramente que el estimador del parámetro β_k se podría interpretar en el caso de variables numéricas como la variación en el término logit (el logaritmo del cociente de probabilidades) causada por la variación unitaria en la variable X_k suponiendo constantes el resto de variables explicativas:

$$OR = e^{\beta_k}$$

O lo que es lo mismo:

$$\log(OR) = \beta_k$$

La interpretación de β_k es diferente en el caso de variables categóricas de dos o más categorías:

- Si X_k es dicotómica, indica lo que varía el logit entre los clientes con valor de X_k igual a uno versus los clientes con valor de X_k igual a cero.
- Si X_k es categórica en general, indica lo que varía el logit entre los clientes con valor de X_k igual a una de sus categorías versus los clientes con valor de X_k igual a la categoría que se haya definido como referencia, ya que para obtener la codificación de las variables categóricas se han utilizado los contrastes tratamiento.

Se va a aplicar el principio de parsimonia, es decir, el mejor modelo será aquel más reducido que mejor explique los datos pero que a la vez sea fácilmente interpretable.

2.3.2 Pasos a seguir para realizar regresión logística

- ***Paso 1: Búsqueda de variables correladas con la variable objetivo***

El tablón de datos cuenta con un elevado número de variables explicativas. Es conveniente conocer cómo se relaciona cada una de ellas con la variable objetivo para poder hacer una primera selección y quedarnos con aquellas que más aportan en la explicación de la misma.

Para ello, aplicamos métodos de selección de variables univariantes y multivariantes:

- **Selección de variables univariante:**

Son aquellas que miden la asociación que existe entre cada variable explicativa y la variable objetivo, sin tener en cuenta el resto de las variables explicativas. Hay varios tipos de análisis que lo miden:

- **Estadístico Chi-cuadrado:** Mide la asociación entre la variable objetivo y cada una de las variables categóricas.

- **Estadístico V de Cramer:** Al igual que el estadístico Chi-cuadrado, mide la asociación entre el variable objetivo y cada una de las variables categóricas. La diferencia principal consiste en que el estadístico V de Cramer toma valores entre 0 y 1, lo cual facilita comparar el grado de asociación entre distintas variables.

  Su fórmula es:

$$V = \sqrt{\frac{\chi^2}{N \cdot m}}$$

  Donde :
  - N: nº de observaciones de la tabla de contingencia donde se representa cada variable versus la variable objetivo
  - M: mínimo entre (nº de filas-1 y nº de columnas-1) de dicha tabla
  - X^2: estadístico Chi-cuadrado

  Las variables explicativas más correladas con la variable objetivo serán aquellas para los que el estadístico V de Cramer está más cerca de 1, ya que indican fuerte asociación.

- **Importancia:** Mide la capacidad discriminante entre la variable objetivo y cada una de las variables explicativas bien sean categóricas o numéricas utilizando una predicción basada en la construcción de árboles de decisión con el criterio Chi-cuadrado. A mayor importancia, más correlada estará dicha variable vs la variable objetivo.

- **Selección de variables multivariante**:
  Son aquellas que miden la asociación que existe entre cada variable independiente y la variable objetivo, teniendo en cuenta el resto de las variables. Hay varios tipos de análisis:

  - **Importancia relativa multivariante:** Análisis multivariante que mide la contribución global de las variables, utilizando una predicción basada en la construcción de árboles de decisión con el criterio Chi-cuadrado.

  - **Árbol extenso:** Análisis multivariante que consiste en la construcción de un árbol completo incorporando reglas subrogadas en cada nodo. (Reglas subrogadas son reglas "sustitutas" que se crean como segunda opción en caso de que en la primera opción las observaciones no puedan clasificarse por tener valores atípicos).

    Se ordenan las variables de mayor a menor importancia. Se destacan las variables cuya importancia relativa en test es al menos el 75% de la importancia relativa en entrenamiento.

➢ *Paso 2: Categorización de variables numéricas*

Vamos a categorizar las variables que han sido seleccionadas en el paso anterior, es decir, que tienen relación con la variable objetivo. Para determinar cuáles son los puntos de corte óptimos para cada una de ellas vamos a utilizar los árboles de decisión. De esta forma cada variable numérica se dividirá en "x" tramos, de forma que cada uno de ellos tenga diferente proporción de casos positivos para la variable objetivo, ya que es lo que interesa ir detectando.

Otra opción para convertir una variable en categórica es dividirla en "x" tramos, de forma que cada uno de estos tramos tenga la misma proporción de observaciones. No se recomienda esta opción, ya que es posible que se pierda parte de la asociación natural entre cada variable y la variable objetivo.

➢ *Paso 3: Preparación del tablón de datos*

A diferencia de en el caso de los árboles de decisión, para entrenar un modelo de regresión de logística, el número de registros que se dispone en el tablón de datos debe de ser moderado. El algoritmo es mucho más complejo y en el caso de tener muchas observaciones se requiere un tiempo de ejecución muy elevado e incluso es posible que una máquina convencional no disponga de recursos suficientes para poder realizar tales cálculos.

Es por ello por lo que es necesario extraer una muestra aleatoria simple del tablón de datos y entrenar el modelo con dicha muestra. El tamaño de la muestra depende del software empleado, del número de variables, de la capacidad del servidor,… no se

puede determinar por tanto un número exacto, lo que sí es imprescindible es que dicha muestra sea representativa del tablón de datos original.

En este caso no vamos a tener tablón de validación. Tan solo:

- Tablón de entrenamiento: muestra con la que se estimarán los parámetros del modelo.
- Tablón de test: Conjunto de datos que no interviene en la fase de entrenamiento del modelo y con los que se evaluará si el modelo seleccionado se puede generalizar para datos que no jugaron ningún papel en la selección del mismo. El modelo se aplicará sobre todos los registros, es decir, sin extraer ninguna muestra.

➢ *Paso 4: Entrenamiento del modelo inicial (sin interacciones)*

Se va a aplicar el **principio jerárquico**, que consiste en que si en el modelo de regresión logística se incluye un término cualquiera, todos sus términos de menor orden deben de permanecer en el modelo, y que si se elimina un término cualquiera, todos los términos de mayor orden en los que intervenga también deben sacarse del modelo.

Por ello se va a buscar un modelo inicial solamente con los efectos principales, es decir, sin interacciones, con las variables explicativas que han resultado significativas en el paso1 y a partir de él (en el paso 5) se analizará si las interacciones entre ellas (solamente entre las que forman parte del modelo obtenido en esta fase) son significativas.

Para determinar qué variables aportan más en la explicación de la variable objetivo y que a su vez no expliquen lo mismo entre ellas se utilizarán métodos de selección de variables. Hay varios métodos:

- Forward (hacia delante):
  - Se inicia con un modelo vacío (solo con el término independiente)
  - Se ajusta un modelo con el método de máxima verosimilitud y se calcula el estadístico chi-cuadrado junto con el p-valor de incluir cada variable por separado
  - Se selecciona el modelo con la variable más significativa
  - Se ajusta el modelo con la(s) variable(s) seleccionadas y se calcula el p-valor de añadir cada una de las variables no seleccionadas por separado
  - Se selecciona el modelo con la variable más significativa
  - Se repiten estos pasos hasta que no queden variables significativas por incluir

- Backward (hacia atrás):
  - Se inicia con un modelo con todas las variables candidatas
  - Se eliminan, una a una, cada variable y se calcula la pérdida de ajuste al eliminar

    - Se selecciona para eliminar la variable menos significativa
    - Se repiten estos pasos hasta que todas las variables incluidas sean significativas y no pueda eliminarse ninguna sin que se pierda ajuste

- Stepwise:
    - Se combinan los métodos forward y backward
    - Se puede comenzar o con el modelo vacío o con el modelo completo, pero en cada paso se exploran las variables incluidas por si deben salir y las no seleccionadas por si deben entrar
    - Se repiten estos pasos hasta que todas las variables incluidas sean significativas y no entre ni salga ninguna más

Vamos a utilizar en este caso el método stepwise, ya que no está influenciado por el orden en el que se vayan seleccionando las variables en el modelo. Se incluirán aquellas que tengan fuerte grado de relación (elegimos como nivel de significación de entrada para el estadístico chi-cuadrado: p-valor<0,05 y para permanecer en el modelo: p-valor<0,03).

La estimación de los parámetros β_0, $\beta_{1,..,}$ β_k asociados a cada una de las variables explicativas se realiza mediante la estimación de máxima verosimilitud. Para ello es necesario resolver un sistema de ecuaciones complejo que solamente puede calcularse a través de métodos iterativos. El método que utiliza el software que vamos a emplear es el método de Newton-Raphson.

A través del contraste de la regresión se comparará de forma global si los coeficientes estimados que forman parte del modelo versus el modelo formado solamente por una constante son iguales.

Con el test parcial de variables individuales se contrasta la hipótesis de que un coeficiente aislado es distinto de cero. Es decir, no se tiene en cuenta el resto de las variables que también forman parte del modelo.

Con el test secuencial se contrasta si cada coeficiente es igual a cero teniendo en cuenta el resto de coeficientes que forman parte también del modelo. Cada estadístico sigue una distribución chi-cuadrado con un grado de libertad. En caso de no ser significativo implica que el modelo sin la variable no empeora respecto el modelo completo (es decir, da igual que esté o que no esté en el modelo), con lo que siguiendo el principio de parsimonia (es mejor el modelo más reducido siempre y cuando explique lo mismo), no se incluirá dicha variable en el modelo, ya que no aporta nada al mismo.

La estimación de estos parámetros puede hacerse tanto con test de razón de verosimilitud como con el test de Wald. En cada caso se mostrará la salida proporcionada por el software.

Las variables que forman parte del modelo son las que han sido seleccionadas por el método stepwise.

A cada cliente se le asigna una probabilidad de fuga. Para poder visualizar los resultados obtenidos fácilmente, se van a crear grupos de clientes de similar probabilidad: se ordenan de mayor a menor probabilidad de evento y posteriormente los dividimos en grupos. Cada grupo es un intervalo de probabilidad de amplitud 0,05.

- **Medidas para comprobar la capacidad predictiva del modelo:**

Ver detalle sobre la parte teórica en el apartado de árboles de decisión

  - Porcentaje de éxitos capturados (Cumulative % Capturated Response):
  - Mejora acumulada (Cumulative Lift):
  - Curva ROC
  - Matriz de clasificación

- **Otros estadísticos de ajuste:**

  - **AIC de Akaike**
    Matemáticamente, la verosimilitud aumenta conforme aumenta el número de variables explicativas del modelo, lo cual es factible si se tiene suficiente número de observaciones. Sin embargo los modelos mejor interpretables son los más simples. Por ello, utilizamos el criterio de información de Akaike que permite comparar modelos penalizando aquellos con mayor número de variables. Se calcula de la siguiente forma:

$$\mathrm{AIC} = -2\,(\mathrm{lnL} - \mathrm{p})$$

    donde p es el número de parámetros.
    Son mejores los modelos con menor valor de AIC

  - **Suma de cuadrados del error (SSE)**
    Suma de cuadrados del error de predicción. Cuanto menor sea mejor es la estimación

  - **Error medio (ASE o MSE)**
    Media de los cuadrados de los errores. Cuanto menor sea mejor es la estimación

- ***Paso 5: Entrenamiento del modelo (con interacciones)***

Una vez que tenemos un modelo válido con los efectos principales de las variables, vamos a ver si hay interacciones entre ellas que resulten significativas y se mejore por lo tanto la explicación de la variable objetivo.

Para ello, con las variables que están en el modelo obtenido en el “Paso4”, más sus interacciones de orden 2, se va a entrenar un modelo siguiendo los pasos y mismos criterios descritos en el Paso 4.

Debido al elevado número de variables que tenemos en el tablón de datos, no vamos a considerar interacciones de orden 3 ni superiores con el fin de no aumentar la complejidad del modelo.

- ***Paso 6: Entrenamiento del modelo jerárquico (sin selección de variables)***

Ya sabemos cuáles son las variables que más aportan en la explicación de la variable dependiente. Ahora bien, en el modelo obtenido en el paso anterior, puede haber variables para los que es significativa su interacción con otra variable pero no lo es su efecto principal. Como se ha explicado con anterioridad, si es posible se va a aplicar el principio jerárquico, que indica que si una variable (interacción) está incluida en el modelo, todas las anteriores de menor orden también deberán estar incluidas. Por lo tanto, vamos a ajustar un modelo sin selección de variables, donde estimemos los coeficientes de las siguientes variables que nos interesa incluir en el modelo:

- Variables significativas obtenidas en el modelo anterior
- Además incluimos las variables de primer orden cuya interacción con otra variable ha resultado significativa.

- ***Paso 7: Elección del mejor modelo***

Se han estimado tres modelos: Los modelos que incluyen interacciones son más complejos, pero probablemente tenga mayor capacidad de predicción. Para poder cuantificarlo, vamos a comparar entre ellos los siguientes indicadores:

- Porcentaje de éxitos capturados
- Mejora acumulada
- Matriz de clasificación
- Otros estimadores: AIC, SSE, ASE

En caso de que los modelos tengan similar capacidad de predicción entre ellos, nos quedaremos con el modelo más sencillo.

- ***Paso 8: Test***

Consiste en medir los resultados del modelo obtenido en el Paso7 con un conjunto de datos que no han intervenido en el entrenamiento del mismo. De esta forma nos aseguramos que el modelo no se ajusta exclusivamente a los datos utilizados en su construcción.

Se contrastará la capacidad predictiva del mismo con dichos datos de test, de forma que con este conjunto de datos se tiene que llegar a las mismas conclusiones que las que obtuvieron con los datos de entrenamiento (resultados del paso7). En caso de ser

así, ya se tendría el **modelo definitivo** construido a partir de la regresión logística, pero en caso contrario, habría que entrenar otro (repetir desde el paso 4).

- ***Paso 9: Resultados obtenidos e interpretación***

Se mostrará:

- las variables que han resultado más significativas en el modelo
- las conclusiones más importantes del modelo obtenido
- la vida estimada para cada cliente

2.4 Análisis de supervivencia

2.4.1 Descripción del modelo

El análisis de supervivencia consiste en modelar no solamente la relación entre la tasa de supervivencia y el tiempo, sino también la posible relación con diferentes variables. Se trata por tanto de calcular la tasa de fuga como función del tiempo y de las variables explicativas (que en concreto, en análisis de supervivencia se denominan covariables).

La variable dependiente es el tiempo, que en nuestro caso es el número de meses que transcurren desde que dan de alta el contrato.

Existen varios modelos de análisis de supervivencia. En este caso, vamos a utilizar el modelo de riesgos proporcionales o modelo de Cox, el cual, es un modelo semiparamétrico que se usa mucho para explicar los efectos de las variables explicativas en los tiempos de supervivencia. El objetivo es determinar no sólo si un evento ha ocurrido, sino cuándo ocurrió.

La fórmula del modelo de riesgo proporcional de COX es:

$$H(t,X)=h_0(t)\bullet \exp(\sum \beta_i X_i)$$

donde:

- X_i son las variables explicativas y βi sus coeficientes estimados. Son independientes del tiempo. Al igual que en regresión logística, para estimar qué variables son significativas y cuáles son sus parámetros estimados se va a utilizar el método de selección de variables stepwise.

- $h_0(t)$ se denomina riesgo base. No depende de las variables X_i, solamente depende del tiempo.

Es un modelo semiparamétrico debido a que una parte de su fórmula depende del tiempo y otra parte no.

Se podría hacer una utilización mucho más amplia de este modelo, pero la parte que vamos a aplicar en este estudio, es ampliar la información obtenida con los árboles de decisión y la regresión logística, dando otra visión adicional, que consiste en comprobar si el hecho de que cancelen los clientes el producto depende además de otros factores del tiempo que llevan con dicho producto. Dicho de otra manera, vamos a obtener la función de distribución de supervivencia en función del tiempo que los clientes llevan con dicho producto contratado.

La función de supervivencia se calcula de la siguiente forma:

$$S(t,X) = [S_0(t)] \exp(\sum \beta_i X_i)$$

Es importante también conocer el concepto censura: las observaciones censuradas son aquellas para las que no se contempla su duración total durante el periodo de estudio. En nuestro caso, consideramos observaciones censuradas a aquellas que aún no se han fugado.

2.4.2 Pasos a seguir en el análisis de supervivencia

- ***Paso 1: Preparación del tablón de datos***

Vamos a estimar la función de supervivencia exclusivamente sobre el conjunto de datos de test (nivel de agregación cliente), ya que en este caso, solamente nos interesa complementar los resultados del modelo final obtenido con árboles o con regresión logística aplicado sobre dichos datos. No tenemos un objetivo predictivo, por lo que no es problema el que la curva pueda ajustarse exclusivamente a estos datos.

- ***Paso 2: Construcción del modelo***

Para estimar el tiempo de supervivencia con la ayuda de variables explicativas, se utilizará el método de selección de variables stepwise. La estimación de los parámetros se realiza a través del método de máxima verosimilitud.

Con el test del ratio de verosimilitud, Score y Wald se contrasta si el modelo con todas las variables es mejor que no tener modelo.

Para validar el grado de ajuste se utilizan los siguientes estadísticos: -2logverosimilud, AIC, SBC. Recordar que valores más bajos indican mejor ajuste.

- ***Paso 3: Resultados obtenidos e interpretación***

Se mostrará la curva de supervivencia obtenida.

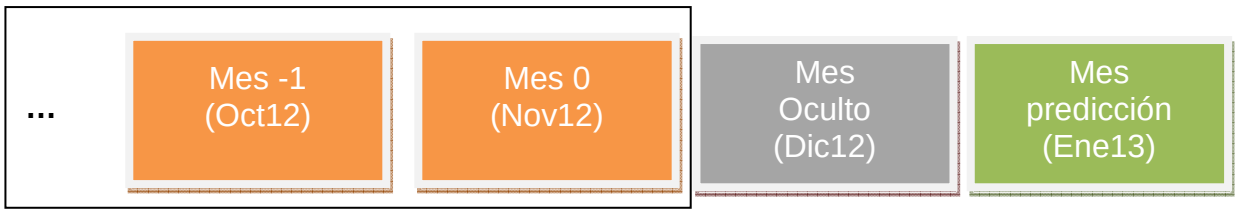
3 Datos

En este capítulo vamos a comprobar de qué datos se disponen y se realizará un análisis descriptivo con el objetivo de analizar la calidad de los mismos.

3.1 Tablón de investigación

El universo de clientes que vamos a analizar son aquellos que tienen contratados al menos un producto de ahorro o inversión y que además han contratado también un seguro “Premium-Hogar”. Adicionalmente pueden tener contratados otros productos.

El nivel de agregación de los datos es cliente-mes y se construye de la siguiente forma (vamos a explicar cómo se ha construido con una de las particiones):



- En Nov12 se selecciona el universo de clientes con los que se va a entrenar el modelo
- Para estos clientes se va a predecir si se fugan o no en Ene13
- Las variables explicativas se construyen con los datos de Nov12 y de los meses previos (Oct12, Sep12...) (ver recuadro). Por ejemplo la variable cuota con la tarjeta de crédito en los últimos tres meses, se calcula como la media de la cuota que tiene el cliente en los meses de Nov12, Oct12 y Sep12.
- Esto mismo se ejecuta para más meses. Cada ejecución es una partición que se unen en un único tablón formando el tablón de investigación. Al construir el tablón de esta forma, predecimos no solamente el perfil de los clientes que se fugan, sino también si lo han a hacer en el mes de predicción. Además podremos calcular la esperanza de vida, es decir, cuantos meses faltan para que el cliente de baja el producto.

Los meses disponibles para la construcción del modelo son de Noviembre12 – Marzo13 (ambos incluidos)

Mes predicción	Nº de registros
Noviembre12	94.006
Diciembre12	91.042
Enero13	91.412
Febrero13	97.429
Marzo13	99.386
Total	473.275

Respecto las variables que disponemos, muchas de ellas tienen nombre largo. Para facilitar su legibilidad, se ha asignado a cada una de ellas una etiqueta (el nombre abreviado que figura en las tablas inferiores). A lo largo de este documento utilizaremos una u otra notación, siguiendo el siguiente criterio:

- Se utilizará la **notación ampliada** siempre que sea posible con el fin de entender de forma más ágil el significado de los resultados que se estén analizando.
- Se utilizará la **notación reducida** (etiqueta) solamente en los casos en los que el nombre de la variable dificulte el entendimiento de los resultados, como es el caso por ejemplo de las interacciones entre variables en regresión logística.

Hay dos tipos de variables:

- **Variable objetivo (Y):** Variable que queremos predecir. Dependiendo del tipo de modelo que vayamos a realizar deberemos de elegir una entre:

Variable	Etiqueta	Descripción
TARGET	T1	Variable objetivo en los modelos de regresión logística y árboles de decisión. Toma valor 1 si el cliente cancela el producto "Premium Hogar". 0 en caso contrario
TIEMPO	T2	Variable objetivo en el caso de análisis de supervivencia Número de meses que transcurren desde que contratan el seguro

- **Variables explicativas (X_i)**: Variables que vamos a utilizar para explicar la variable objetivo:

Variable	Etiqueta	Descripción
ANTIGUEDAD_CLIENTE	V1	Antigüedad del cliente en la compañía
COC_CUOTA_TJCRED_ULT1	V2	Cuota en la tarjeta de crédito del cliente entre la cuota media en el último mes
COC_IMP_FINAN_CRED_ULT1	V3	Importe financiado a crédito entre la media en el último mes
COC_IMP_OPER_TJCRED_ULT1	V4	Importe de operaciones realizadas con la tarjeta de crédito entre la media en el último mes
COC_LIM_TJCRED_ULT1	V5	Límite de la tarjeta de crédito entre la media en el último mes
COC_NUM_OPER_TJCRED_ULT1	V6	Número de operaciones realizadas con la tarjeta de crédito entre la media en el último mes
COC_TJCRED_ULT1	V7	Número de tarjetas de crédito que tiene el cliente entre la media en el último mes
CRED_DISPONIBLE_TJCRED_ULT1	V8	Crédito disponible con su tarjeta de crédito en el último mes
CUOTA_TJCRED_ULT1	V9	Cuota con la tarjeta de crédito en el último mes
CUOTA_TJCRED_ULT3	V10	Cuota con la tarjeta de crédito en los últimos tres meses
EDAD	V11	Edad del cliente
ESTADO_CIVIL_CAT	V12	Estado civil
IMP_FINANCIADO_CRED_ULT1	V13	Importe financiado a crédito en el último mes

IMP_FINANCIADO_CRED_ULT3	V14	Importe financiado a crédito en los últimos tres meses
IMP_OPER_CRED_ULT1	V15	Importe de operaciones a crédito en el último mes
IMP_OPER_CRED_ULT3	V16	Importe de operaciones a crédito en los últimos tres meses
IMP_PROD_INVAH_HACE3	V17	Importe en productos de inversión y ahorro hace tres meses
IMP_PROD_INVAH_ULT1	V18	Importe en productos de inversión y ahorro en el último mes
IMP_TRANSF_PROD_AH_ULT1	V19	Importe de transferencias realizadas a los productos de ahorro en el último mes
IMP_TRANSF_PROD_AH_ULT3	V20	Importe de transferencias realizadas a los productos de ahorro durante los últimos 3 meses
IMP_TRANSF_PROD_INV_ULT1	V21	Importe de transferencias realizadas a los productos de inversión en el último mes
IMP_TRANSF_PROD_INV_ULT3	V22	Importe de transferencias realizadas a los productos de inversión durante los últimos 3 meses
IMP_USO_TJDEB_CAJERO_ULT1	V23	Importe gastado con la tarjeta de débito en cajeros en el último mes
IMP_USO_TJDEB_CAJERO_ULT3	V24	Importe gastado con la tarjeta de débito en cajeros en los últimos tres meses
IMP_USO_TJDEB_CAJERO_ULT6	V25	Importe gastado con la tarjeta de débito en cajeros en los últimos seis meses
IMP_USO_TJDEB_COMPRAS_ULT1	V26	Importe gastado con la tarjeta de débito en compras (comercios) en el último mes
IMP_USO_TJDEB_COMPRAS_ULT3	V27	Importe gastado con la tarjeta de débito en compras (comercios) en los últimos tres meses
IMP_USO_TJDEB_COMPRAS_ULT6	V28	Importe gastado con la tarjeta de débito en compras (comercios) en los últimos seis meses
IMP_USO_TJDEB_ULT1	V29	Importe gastado con la tarjeta de débito en el último mes
IMP_USO_TJDEB_ULT3	V30	Importe gastado con la tarjeta de débito en los últimos tres meses
IMP_USO_TJDEB_ULT6	V31	Importe gastado con la tarjeta de débito en los últimos seis
IMPORTE_ULTIMO_COBRO	V32	Importe del último recibo
IND_EMPLEADO	V33	Indicador de Empleado (SI/NO)
IND_NACIONAL_CAT	V34	Indicador de nacionalidad española (SI/NO)
IND_PARTES_HOGAR	V35	Indicador de haber dado algún parte con su seguro de hogar (SI/NO)
IND_RECLAMACIONES_ULT3_CAT	V36	Indicador de haber realizado alguna reclamación en los últimos tres meses (SI/NO)
IND_RESIDENTE_CAT	V37	Indicador de residente en España (SI/NO)
IND_TJDEB_ULT1	V38	Indicador de tenencia de tarjeta de débito activa en el último mes (SI/NO)
IND_TJDEB_ULT3	V39	Indicador de tenencia de tarjeta de débito activa en los últimos tres meses (SI/NO)
IND_TJDEB_ULT6	V40	Indicador de tenencia de tarjeta de débito activa en los últimos seis meses (SI/NO)
INGRESOS_ANUALES_ESTIMADOS	V41	Ingresos anuales estimados del cliente
LIMITE_TJCRED_ULT1	V42	Límite con su tarjeta de crédito en el último mes
NIVEL_ESTUDIOS_CAT	V43	Nivel de estudios del cliente
NIVEL_SATISFACCION_CAT	V44	Nivel de satisfacción del cliente
NUM_CAMPANIAS_ULT3	V45	Número de campañas que ha recibido el cliente durante los últimos tres meses
NUM_DIAS_CAT	V46	Número de días transcurridos desde el pago del último recibo: 1= recientemente / 0= tienen que pagar dentro de poco
NUM_DIAS_DESDE_COBRO	V47	Número de días transcurridos desde el cobro del último recibo
NUM_DIAS_ULT_USO_TJDEB	V48	Número de días transcurridos desde que usó por última vez la tjd e débito. Si vale -1 es que nunca la ha usado
NUM_OPER_CRED_ULT1	V49	Número de operaciones con la tarjeta de crédito en el último mes

NUM_OPER_CRED_ULT3	V50	Número de operaciones con la tarjeta de crédito en los últimos 3 meses
NUM_PARTES_VIDA_SALUD_ULT2	V51	Número de partes de vida o salud durante los últimos 2 meses
NUM_PARTES_VIDA_SALUD_ULT4	V52	Número de partes de vida o salud durante los últimos 4 meses
NUM_PROD_AH_ULT1	V53	Número de productos de ahorro en el último mes
NUM_PROD_AHINV_ULT1	V54	Número de productos de ahorro e inversión en el último mes
NUM_PROD_INV_ULT1	V55	Número de productos de inversión en el último mes
NUM_SALUD_HACE3	V56	Número de seguros de salud hace 3 meses
NUM_SALUD_ULT1	V57	Número de seguros de salud durante el último mes
NUM_TJDEB_HACE1	V58	Número de tarjetas de débito en el último mes
NUM_TJDEB_HACE3	V59	Número de tarjetas de débito hace3 meses
NUM_TJDEB_HACE6	V60	Número de tarjetas de débito hace 6 meses
NUM_TRANSF_PROD_AH_ULT1	V61	Número de transferencias realizadas a los productos de ahorro en el último mes
NUM_TRANSF_PROD_AH_ULT3	V62	Número de transferencias realizadas a los productos de ahorro durante los últimos 3 meses
NUM_TRANSF_PROD_INV_ULT1	V63	Número de transferencias realizadas a los productos de inversión en el último mes
NUM_TRANSF_PROD_INV_ULT3	V64	Número de transferencias realizadas a los productos de inversión durante los últimos 3 meses
NUM_VIDA_HACE3	V65	Número de seguros de vida hace 3 meses
NUM_VIDA_ULT1	V66	Número de seguros de vida en el último mes
RATIO_NUM_PROD_AHINV_MED_HACE1	V67	Número de productos de ahorro entre la media de todos los clientes
RATIO_NUM_PROD_AH_MEDIA_ULT1	V68	Número de productos ahorro e inversión entre la media de todos los clientes
RATIO_NUM_PROD_INV_MEDIA_ULT1	V69	Número de productos de inversión entre la media de todos los clientes
RECIB_PROMOCION	V70	Toma el valor 1 si al cliente le ofrecieron un descuento especial en el último recibo. 0 en caso de precio normal
SEXO_CAT	V71	Sexo
VALOR_VIVIENDA	V72	Valor de la vivienda estimada
VINCULACION_CAT	V73	Vinculación

3.2 Análisis descriptivo univariante

En los anexos se encuentra el análisis descriptivo de cada variable.

El objetivo de este análisis no es solamente conocer la distribución de cada variable, sino que además es posible detectar problemas en las mismas derivadas de la calidad de la información de la que se dispone.

Los problemas que nos hemos encontrado y las soluciones que vamos a aplicar son:

- Se detectan variables que están mal informadas, y por lo tanto vamos a descartar de los modelos:

Variable	Motivo
VALOR VIVIENDA	Para el 95% de los registros es cero
INGRESOS ANUALES ESTIMADOS	Para el 99,9% de los registros es cero
IMP_TRANSF_PROD_AH_ULT1	Para el 99,9% de los registros es cero
IMP_TRANSF_PROD_AH_ULT3	Para el 99,7% de los registros es cero
IMP_TRANSF_PROD_INV_ULT1	Para el 99,8% de los registros es cero
IMP_TRANSF_PROD_INV_ULT3	Para el 99,9% de los registros es cero
IMPORTE_ULTIMO_COBRO	Tiene muchos valores negativos y no es posible. Además tiene 90.000 valores perdidos. Está mal informada

Aunque no es realmente el objetivo de este trabajo, se ha investigado el motivo de porqué estas cuatro variables de transferencias son cero en casi todos los casos, y se debe a que realmente son muy pocos los clientes que realizan transferencias.

- Se detectan variables con algunos valores ausentes:

Variable con valores perdidos	Nº de valores perdidos
COC_CUOTA_TJCRED_ULT1	13
COC_IMP_FINAN_CRED_ULT1	13
COC_IMP_OPER_TJCRED_ULT1	13
COC_LIM_TJCRED_ULT1	13
COC_NUM_OPER_TJCRED_ULT1	13
CRED_DISPONIBLE_TJCRED_ULT1	13
CUOTA_TJCRED_ULT1	13
CUOTA_TJCRED_ULT3	13
IMP_FINANCIADO_CRED_ULT1	13
IMP_FINANCIADO_CRED_ULT3	13
IMP_OPER_CRED_ULT1	13
IMP_OPER_CRED_ULT3	13
LIMITE_TJCRED_ULT1	13
NUM_OPER_CRED_ULT1	13
NUM_OPER_CRED_ULT3	13
NUM_PROD_AHINV_ULT1	13
NUM_PROD_AH_ULT1	13
NUM_PROD_INV_ULT1	13
NUM_TRANSF_PROD_AH_ULT1	13
NUM_TRANSF_PROD_AH_ULT3	13
NUM_TRANSF_PROD_INV_ULT1	13
NUM_TRANSF_PROD_INV_ULT3	13
RATIO_NUM_PROD_AHINV_MED_HACE1	13
RATIO_NUM_PROD_AH_MEDIA_ULT1	13
RATIO_NUM_PROD_INV_MEDIA_ULT1	13

Estos 13 valores ausentes que tiene cada una de estas variables, corresponden a los mismos clientes. Puesto que contamos con un número más que suficiente de observaciones, descartamos estos 13 registros de los análisis.

4 Resultados

4.1 *Árboles de decisión*

➢ *Paso 1: Preparación del tablón de datos*

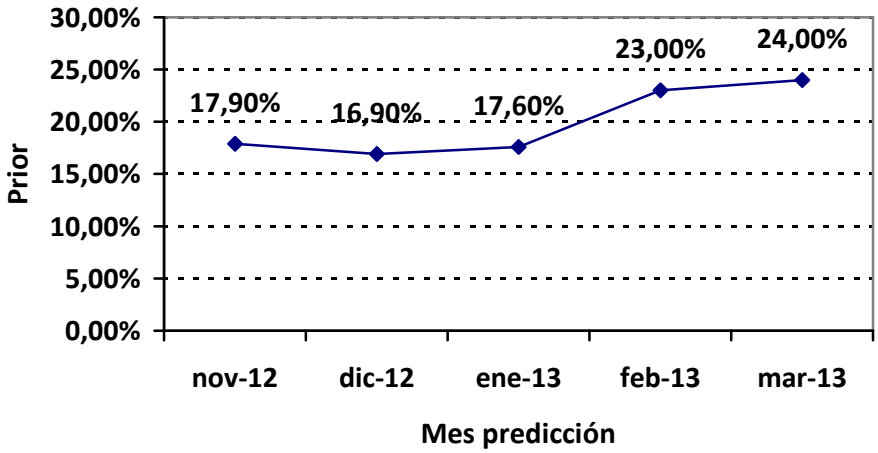
Dividimos el tablón de investigación en entrenamiento, validación y test:

- Tablón de entrenamiento →Se construye con el 80% de los datos de los meses de Noviembre 12 a Febrero 13
- Tablón de validación → Se construye con el 20% de los datos de los meses de Noviembre 12 a Febrero 13
- Tablón de test → Se construye con el 100% de los datos de los meses de Marzo13

El prior (o probabilidad de fuga) en cada uno de los meses es:

Mes predicción	Nº clientes con Y=1	P(Y=1)
Noviembre12	16.896	17,9%
Diciembre12	15.369	16,9%
Enero13	16.089	17,6%
Febrero13	22.423	23,0%
Marzo13	23.878	24,0%
Total	94.655	20,1%

En nuestro caso no es conveniente balancear los datos, ya que se tienen casos positivos suficientes para entrenar el modelo.



➢ ***Paso 2: Selección del criterio de corte***

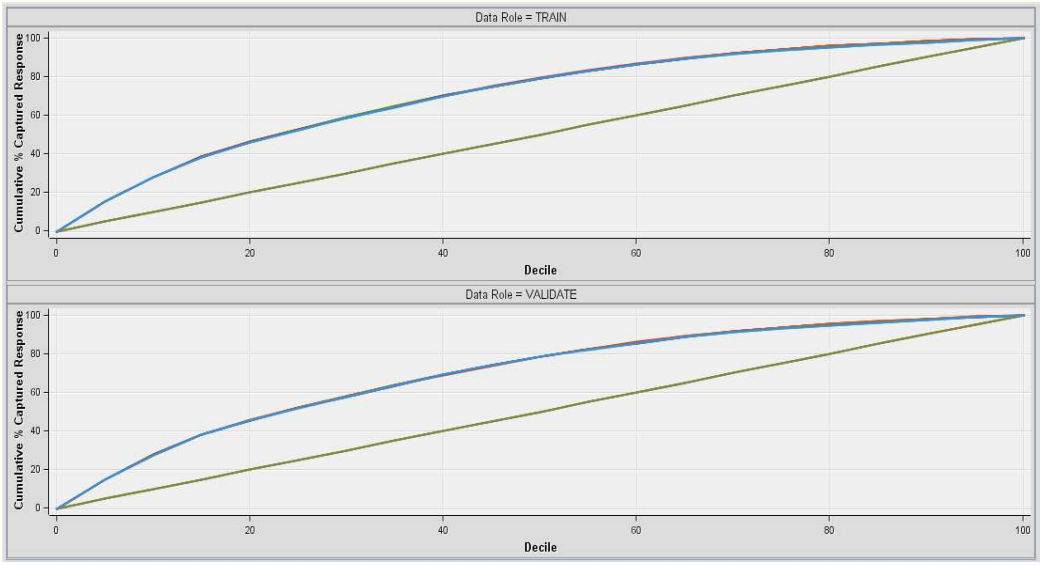
Se realizan varios árboles de forma automática con los siguientes criterios de corte:

Criterios	Árbol 1	Árbol 2	Árbol 3	Árbol 4	Árbol 5	Árbol 6	Árbol 7	Árbol 8	Árbol 9
Criterio corte	Entropy	ProbChisq	Gini	Entropy	ProbChisq	Gini	Entropy	ProbChisq	Gini
Tamaño del árbol	Largest	Largest	Largest	N	N	N	Assesment	Assesment	Assesment

Los resultados obtenidos son:

- **Porcentaje de éxitos capturados:**

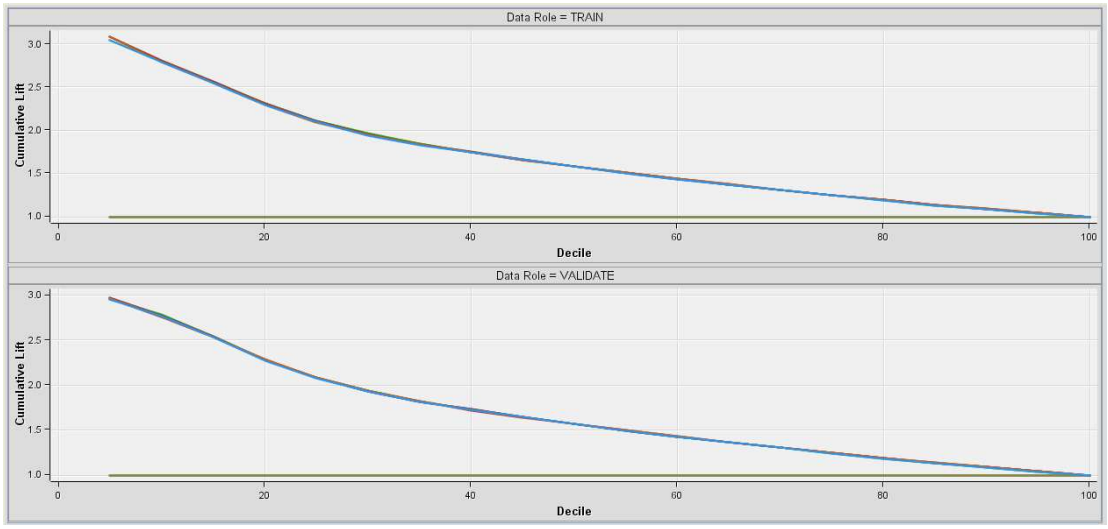
  En la parte superior se muestran los resultados para el tablón de entrenamiento y en la parte inferior para el tablón de validación. Ambas gráficas son similares, lo cual es importante, ya que en caso contrario sería indicativo de que el modelo se estaría ajustando exclusivamente a los datos con los que se ha entrenado.



La línea verde representa los resultados en caso de no tener modelo y la línea azul los resultados obtenidos con uno de los árboles. No se aprecia pero hay una línea por cada uno de los modelos. No se visualizan porque están superpuestas (debajo de la línea azul), lo cual indica que para todos los árboles se obtienen similares resultados.

Cualquiera de los 9 árboles sería válido según este indicador

- **Mejora acumulada:**



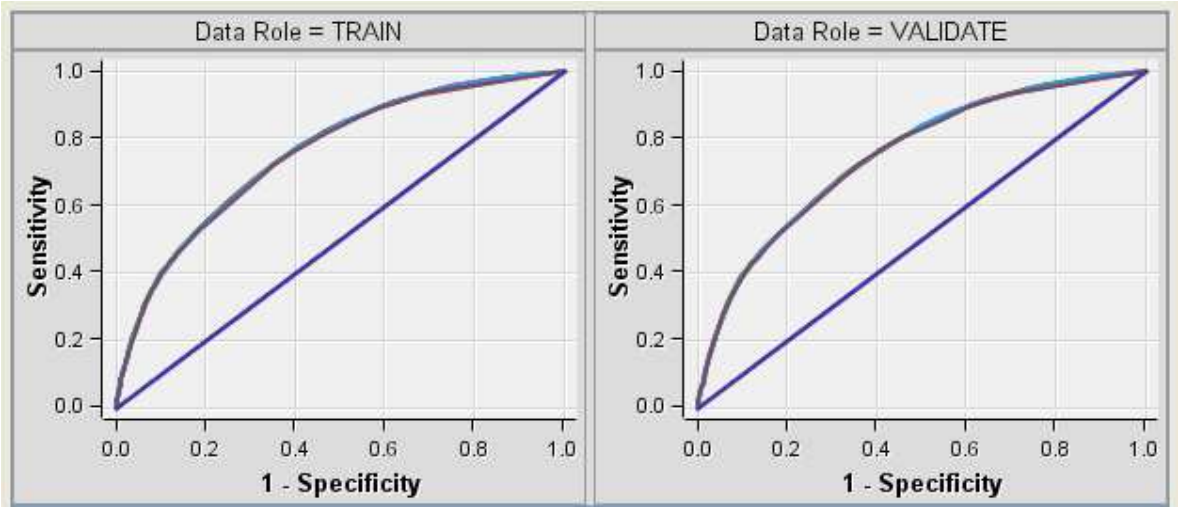
La línea verde representa los resultados en caso de no tener modelo. Además hay una línea por cada uno de los modelos, pero al igual que en gráfico anterior, no se aprecian porque están superpuestas.

Vemos que la ganancia máxima que vamos a poder conseguir es de x3 en alguna de las reglas. Es decir, de 3 veces más al usar el modelo frente a no utilizarlo.

Cualquiera de los 9 árboles sería válido según este indicador

- **Curva ROC:**

En la parte izquierda se muestran los resultados para el tablón de entrenamiento y en la derecha para el tablón de validación.



La línea azul representa los resultados sin modelo. Además al igual que en los gráficos anteriores, hay una línea por cada uno de los modelos, pero no se aprecian porque están superpuestas.

Cualquiera de los 9 árboles sería válido según este indicador, las diferencias son mínimas (ver valor del estadístico índice de ROC en el apartado “estadísticos de ajuste”. Este índice representa el área comprendida entre la curva ROC y la selección aleatoria (línea azul).

- **Otros estadísticos de ajuste:**

  – **Probabilidad de clasificación errónea** (Misclassification Rate):

Id Árbol	Entrenamiento	Validación
Árbol 1	0.1808	0.1830
Árbol 2	**0.1807**	**0.1829**
Árbol 3	0.1814	0.1831
Árbol 4	0.1893	0.1893
Árbol 5	0.1893	0.1893
Árbol 6	0.1893	0.1893
Árbol 7	0.1809	0.1832
Árbol 8	0.1816	0.1834
Árbol 9	0.1809	0.1831

Se obtienen mejores resultados para el árbol 2, aunque las diferencias son mínimas.

  – **Suma de cuadrados del error** (SSE):

Id Árbol	Entrenamiento	Validación
Árbol 1	68844.9359	29790.9861
Árbol 2	**68837.0686**	**29790.1318**
Árbol 3	68967.6716	29808.7703
Árbol 4	80329.2458	34428.6530
Árbol 5	80329.2458	34428.6530
Árbol 6	80329.2458	34428.6530
Árbol 7	69005.1389	29796.8848
Árbol 8	69135.7046	29837.6434
Árbol 9	68945.2248	29811.0134

Se obtienen mejores resultados para el árbol 2, aunque las diferencias son mínimas, excepto en los árboles de 4 al 6 inclusive que tienen mayor error.

– **Error medio** (ASE o MSE):

Id Árbol	Entrenamiento	Validación
Árbol 1	0.1316	0.1329
Árbol 2	**0.1315**	**0.1328**
Árbol 3	0.1318	0.1329
Árbol 4	0.1535	0.1535
Árbol 5	0.1535	0.1535
Árbol 6	0.1535	0.1535
Árbol 7	0.1318	0.1328
Árbol 8	0.1321	0.1330
Árbol 9	0.1317	0.1329

Se obtienen mejores resultados para el árbol 2, aunque las diferencias son mínimas, excepto en los árboles de 4 al 6 inclusive que tienen mayor error.

– **Ganancia** (Gain):

Id Árbol	Entrenamiento	Validación
Árbol 1	208.6439	196.3351
Árbol 2	**208.6823**	**196.5119**
Árbol 3	204.5759	195.0545
Árbol 4	0.0000	0.0000
Árbol 5	0.0000	0.0000
Árbol 6	0.0000	0.0000
Árbol 7	207.7544	195.7294
Árbol 8	207.7928	196.3351
Árbol 9	207.7928	195.7294

Se obtienen mejores resultados para el árbol 2, aunque las diferencias son mínimas respecto los árboles 1 y 9. El 4, 5 y 6 quedan descartados ya que no tienen ninguna ganancia.

– **Mejora (Lift):**

Id Árbol	Entrenamiento	Validación
Árbol 1	3.0864	2.9634
Árbol 2	**3.0868**	**2.9651**
Árbol 3	3.0458	2.9505
Árbol 4	1.0000	1.0000
Árbol 5	1.0000	1.0000
Árbol 6	1.0000	1.0000
Árbol 7	3.0775	2.9555
Árbol 8	3.0451	2.9496
Árbol 9	3.0779	2.9573

Se obtienen mejores resultados para el árbol 2, aunque las diferencias son mínimas excepto en los árboles 4,5 y 6 ya que no tienen ninguna mejora.

– **Índice ROC** (ROC Index):

Id Árbol	Entrenamiento	Validación
Árbol 1	0.7555	0.7484
Árbol 2	**0.7547**	**0.7479**
Árbol 3	0.7528	0.7464
Árbol 4	0.5000	0.5000
Árbol 5	0.5000	0.5000
Árbol 6	0.5000	0.5000
Árbol 7	0.7538	0.7488
Árbol 8	0.7505	0.7453
Árbol 9	0.7537	0.7478

Vemos que el mejor árbol es el 2, aunque las diferencias son mínimas excepto en los árboles 4,5 y 6 que como veíamos en los gráficos anteriores son similares que en una selección aleatoria.

– **Coeficiente de Gini (Gini Coeficient):**

Id Árbol	Entrenamiento	Validación
Árbol 1	0.5111	0.4967
Árbol 2	0.5095	0.4958
Árbol 3	0.5056	0.4928
Árbol 4	0.0000	0.0000
Árbol 5	0.0000	0.0000
Árbol 6	0.0000	0.0000
Árbol 7	0.5076	0.4976
Árbol 8	**0.5010**	**0.4907**
Árbol 9	0.5073	0.4956

Según este indicador el árbol 8 sería el mejor aunque la diferencia es mínima sobre todo con los árboles 2, 3 y 9.

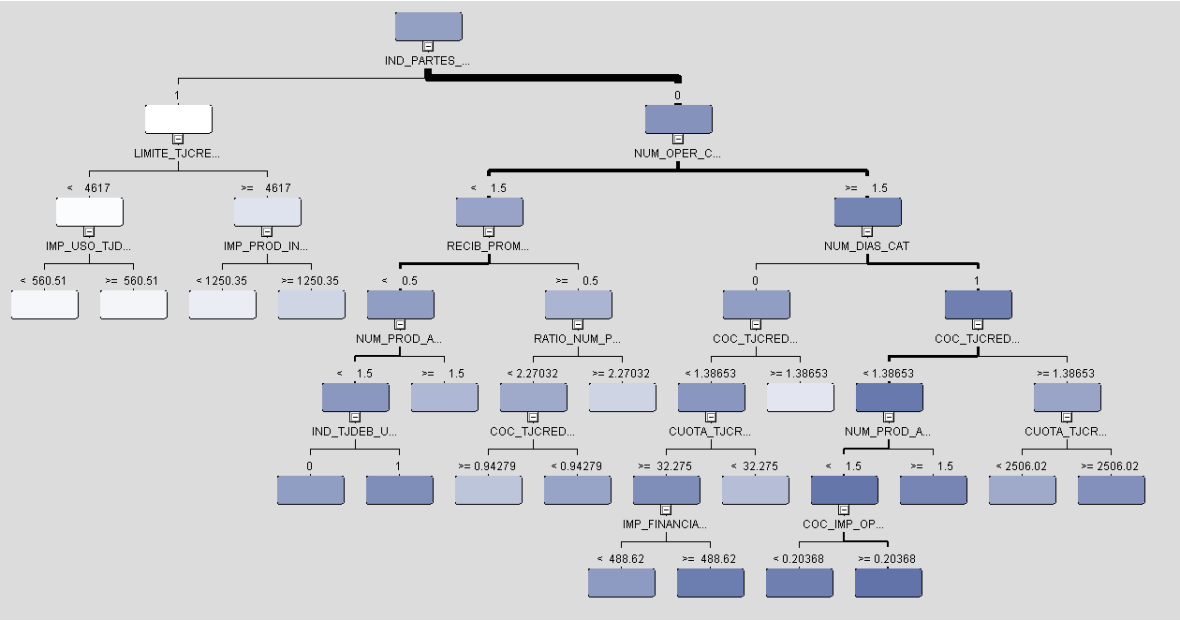
– **Conclusiones**

El mejor árbol es el árbol 2, con lo cual el criterio de corte elegido es el de Chi-cuadrado.

➢ ***Paso 3: Construcción y validación del árbol***

Una vez que hemos elegido el criterio de corte, se ha obtenido un árbol de decisión de forma “manual” con el criterio Chi-cuadrado.

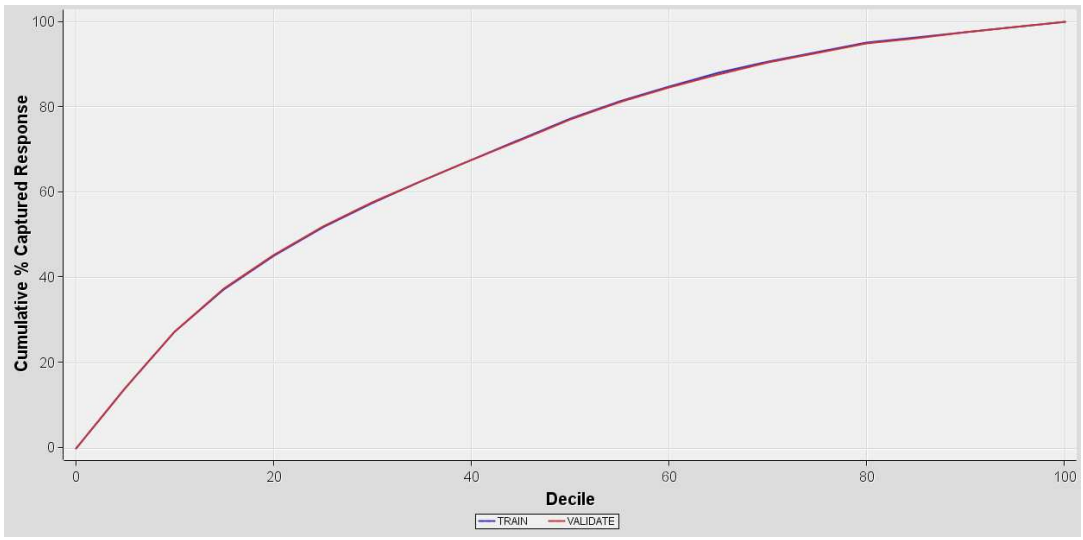
El árbol obtenido es:



Más adelante se explicará el detalle de cada una de las reglas que lo forman.

Antes veamos si es un modelo válido. Veremos el resultado de cada uno de los indicadores para los datos de entrenamiento y validación:

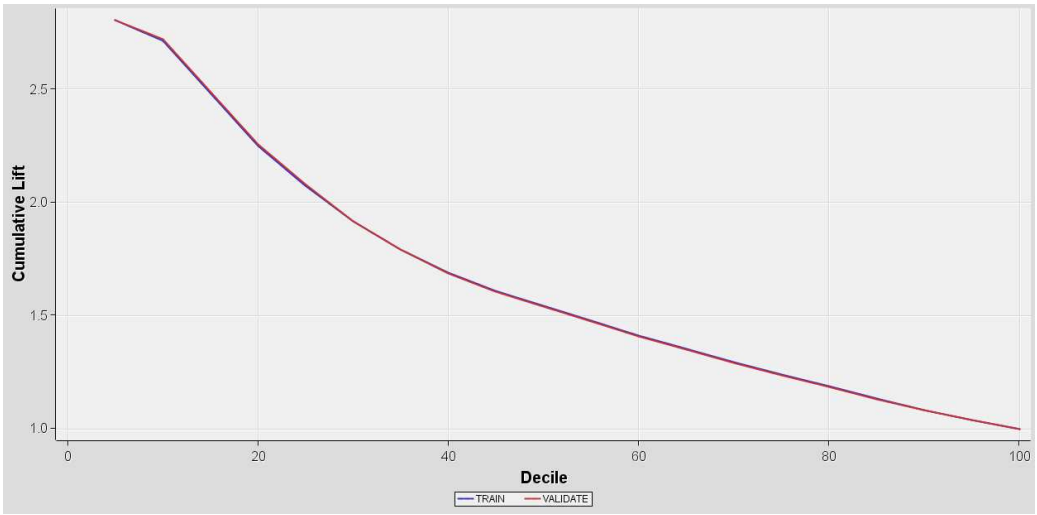
- **Porcentaje de éxitos capturados:**



  - Con un 12% de los clientes predecimos el 31% de las fugas
  - Con un 28% de los clientes predecimos el 56% de las fugas
  - Con un 40% de los clientes predecimos el 65% de las fugas

  Se obtienen buenos resultados, los cuales se mantienen tanto en entrenamiento como en validación (ambas líneas están superpuestas)

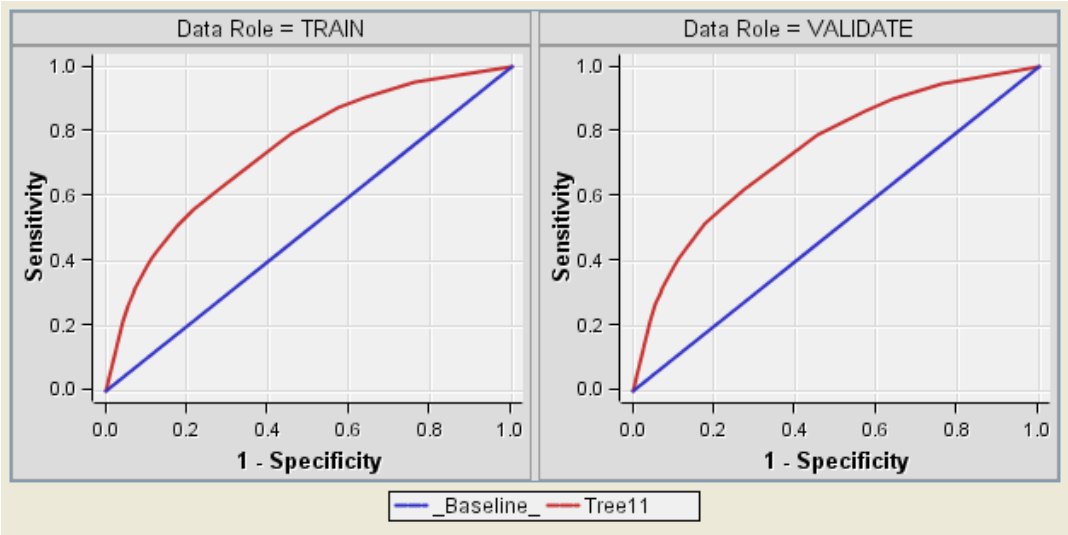
- **Mejora acumulada:**



– Con un 12% de los clientes se obtiene una ganancia de x2,6
– Con un 28% de los clientes se obtiene una ganancia de x2,0
– Con un 40% de los clientes se obtiene una ganancia de x1,6

Se obtienen buenos, los cuales se mantienen tanto en entrenamiento como en validación (ambas líneas están superpuestas)

- **Curva ROC:**



– Para conseguir un 29,5% de verdaderos positivos se tendrá un 6,7% de falsos positivos
– Para conseguir un 42,7% de verdaderos positivos se tendrá un 12,7% de falsos positivos

Los resultados se mantienen en entrenamiento y validación

- **Otros estadísticos de ajuste:**

– **Probabilidad de clasificación erronea (Misclassification Rate):**

Entrenamiento	Validación
0.1845	0.1844

Es estable entre entrenamiento y validación

– **Suma de cuadrados del error (SSE):**

Entrenamiento	Validación
70141.84	30067.49

No son comparables entre entrenamiento y validación, ya que influye el número de observaciones de cada uno

– **Error medio (ASE o MSE):**

Entrenamiento	Validación
0.1340	0.1340

Es estable entre entrenamiento y validación

– **Ganancia (Gain):**

Entrenamiento	Validación
180.6824	180.6664

Es estable entre entrenamiento y validación

– **Mejora (Lift):**

Entrenamiento	Validación
2.8068	2.8067

Es estable entre entrenamiento y validación

– **Índice ROC (ROC Index):**

Entrenamiento	Validación
0.7396	0.7386

Es estable entre entrenamiento y validación

– **Coeficiente de Gini (Gini Coeficient):**

Entrenamiento	Validación
0.4792	0.4772

Es estable entre entrenamiento y validación

- **Descripción de las reglas obtenidas**

Las reglas obtenidas son:

ORDEN	REGLA	NODO ARBOL	PRIOR	REGLA
1	1	7	53.1	IF IMP_USO_TJDEB_ULT1 < 560.51 AND LIMITE_TJCRED_ULT1 < 4617 AND IND_PARTES_HOGAR = 1
2	2	8	46.2	560.51 <= IMP_USO_TJDEB_ULT1 AND LIMITE_TJCRED_ULT1 < 4617 AND IND_PARTES_HOGAR = 1
3	3	21	43.5	IMP_PROD_INVAH_ULT1 < 1250.345 AND 4617 <= LIMITE_TJCRED_ULT1 AND IND_PARTES_HOGAR = 1
4	10	72	41.4	1.3865259066 <= COC_TJCRED_ULT1 AND NUM_DIAS_CAT = 0 AND 1.5 <= NUM_OPER_CRED_ULT3 AND IND_PARTES_HOGAR = 0
5	4	22	35.9	1250.345 <= IMP_PROD_INVAH_ULT1 AND 4617 <= LIMITE_TJCRED_ULT1 AND IND_PARTES_HOGAR = 1
6	14	95	35.8	2.2703216975 <= RATIO_NUM_PROD_AH_MEDIA_ULT1 AND 0.5 <= RECIB_PROMOCION AND NUM_OPER_CRED_ULT3 < 1.5 AND IND_PARTES_HOGAR = 0
7	16	106	30.8	0.9427850916 <= COC_TJCRED_ULT1 AND RATIO_NUM_PROD_AH_MEDIA_ULT1 < 2.2703216975 AND 0.5 <= RECIB_PROMOCION AND NUM_OPER_CRED_ULT3 < 1.5 AND IND_PARTES_HOGAR = 0
8	11	76	28.8	CUOTA_TJCRED_ULT1 < 32.275 AND COC_TJCRED_ULT1 < 1.3865259066 AND NUM_DIAS_CAT = 0 AND 1.5 <= NUM_OPER_CRED_ULT3 AND IND_PARTES_HOGAR = 0

9	17	117	26.5	1.5 <= NUM_PROD_AHINV_ULT1 AND RECIB_PROMOCION < 0.5 AND NUM_OPER_CRED_ULT3 < 1.5 AND IND_PARTES_HOGAR = 0
10	5	56	22.1	CUOTA_TJCRED_ULT3 < 2506.015 AND 1.3865259066 <= COC_TJCRED_ULT1 AND NUM_DIAS_CAT = 1 AND 1.5 <= NUM_OPER_CRED_ULT3 AND IND_PARTES_HOGAR = 0
11	15	105	20.1	COC_TJCRED_ULT1 < 0.9427850916 AND RATIO_NUM_PROD_AH_MEDIA_ULT1 < 2.2703216975 AND 0.5 <= RECIB_PROMOCION AND NUM_OPER_CRED_ULT3 < 1.5 AND IND_PARTES_HOGAR = 0
12	18	118	18.4	IND_TJDEB_ULT1 = 0 AND NUM_PROD_AHINV_ULT1 < 1.5 AND RECIB_PROMOCION < 0.5 AND NUM_OPER_CRED_ULT3 < 1.5 AND IND_PARTES_HOGAR = 0
13	12	81	17.2	IMP_FINANCIADO_CRED_ULT3 < 488.62 AND 32.275 <= CUOTA_TJCRED_ULT1 AND COC_TJCRED_ULT1 < 1.3865259066 AND NUM_DIAS_CAT = 0 AND 1.5 <= NUM_OPER_CRED_ULT3 AND IND_PARTES_HOGAR = 0
14	6	57	13.9	2506.015 <= CUOTA_TJCRED_ULT3 AND 1.3865259066 <= COC_TJCRED_ULT1 AND NUM_DIAS_CAT = 1 AND 1.5 <= NUM_OPER_CRED_ULT3 AND IND_PARTES_HOGAR = 0
15	19	119	13.3	IND_TJDEB_ULT1 = 1 AND NUM_PROD_AHINV_ULT1 < 1.5 AND RECIB_PROMOCION < 0.5 AND NUM_OPER_CRED_ULT3 < 1.5 AND IND_PARTES_HOGAR = 0
16	7	62	10.6	1.5 <= NUM_PROD_AHINV_ULT1 AND COC_TJCRED_ULT1 < 1.3865259066 AND NUM_DIAS_CAT = 1 AND 1.5 <= NUM_OPER_CRED_ULT3 AND IND_PARTES_HOGAR = 0
17	8	63	8.5	COC_IMP_OPER_TJCRED_ULT1 < 0.2036808347 AND NUM_PROD_AHINV_ULT1 < 1.5 AND COC_TJCRED_ULT1 < 1.3865259066 AND NUM_DIAS_CAT = 1 AND 1.5 <= NUM_OPER_CRED_ULT3 AND IND_PARTES_HOGAR = 0
18	13	82	7.9	488.62 <= IMP_FINANCIADO_CRED_ULT3 AND 32.275 <= CUOTA_TJCRED_ULT1 AND COC_TJCRED_ULT1 < 1.3865259066 AND NUM_DIAS_CAT = 0 AND 1.5 <= NUM_OPER_CRED_ULT3 AND IND_PARTES_HOGAR = 0

19	9	64	4.7	0.2036808347 <= COC_IMP_OPER_TJCRED_ULT1 AND NUM_PROD_AHINV_ULT1 < 1.5 AND COC_TJCRED_ULT1 < 1.3865259066 AND NUM_DIAS_CAT = 1 AND 1.5 <= NUM_OPER_CRED_ULT3 AND IND_PARTES_HOGAR = 0

- **Capacidad predictiva por regla:**
  Datos de entrenamiento y validación:

Orden	Prior	Ampliada	Reducida	Candidatos	Éxitos	Penetración	Ganancia	%Cand	%Éxitos
1	53.1	**1**	**1**	28,527	5,157	53.13%	2.81	7.6%	21.4%
2	46.2	**1**	**1**	7,425	3,446	46.41%	2.45	2.0%	4.9%
3	43.5	**1**	**1**	5,363	2,300	42.89%	2.27	1.4%	3.2%
4	41.4	**1**	**1**	2,833	1,173	41.40%	2.19	0.8%	1.7%
5	35.9	**1**	0	3,809	1,385	36.36%	1.92	1.0%	2.0%
6	35.8	**1**	0	9,953	3,585	36.02%	1.90	2.7%	5.1%
7	30.8	**1**	0	5,833	1,806	30.96%	1.64	1.6%	2.6%
8	28.8	**1**	0	4,973	1,417	28.49%	1.51	1.3%	2.0%
9	26.5	**1**	0	21,769	5,782	26.56%	1.40	5.8%	8.2%
10	22.1	**1**	0	14,692	3,218	21.90%	1.16	3.9%	4.5%
11	20.1	0	0	22,958	4,604	20.05%	1.06	6.1%	6.5%
12	18.4	0	0	54,585	9,988	18.30%	0.97	14.6%	14.1%
13	17.2	0	0	12,067	2,088	17.30%	0.91	3.2%	3.0%
14	13.9	0	0	3,782	549	14.52%	0.77	1.0%	0.8%
15	13.3	0	0	36,872	4,868	13.20%	0.70	9.9%	6.9%
16	10.6	0	0	22,931	2,451	10.69%	0.56	6.1%	3.5%
17	8.5	0	0	30,118	2,550	8.47%	0.45	8.1%	3.6%
18	7.9	0	0	9,722	800	8.23%	0.43	2.6%	1.1%
19	4.7	0	0	75,677	3,610	4.77%	0.25	20.2%	5.1%
		Total		373,889	70,777	18,93%	1.00	100.0%	100%

Selección ampliada	1.97	**28.1%**	**55.5%**
No seleccionados	0.62	71.9%	44.5%
Total	1.00	100.0%	100.0%

Selección reducida	2.64	**11.8%**	**31.2%**
No seleccionados	0.78	88.2%	68.8%
Total	1.00	100.0%	100.0%

Selección reducida:
Las reglas de mayor propensión a la fuga son las reglas 1,2,3 y4:

– Con dichas reglas se obtiene una ganancia global de X2,64
– Con un 11,8% de los datos se predice el 31,2% de las fugas

Selección ampliada:
Si seleccionamos todas las reglas con algo de ganancia (de la regla 1-10):

– Con dichas reglas se obtiene una ganancia global de X1,97
– Con un 28,1% de los datos se predice el 55,5% de las fugas

- **Matriz de clasificación:**

	Real 0	Real 1
Estimado 0	237.204 (76%)	31.508 (52%)
Estimado 1	75.908 (24%)	29.269 (48%)
Total	268.712 (100%)	105.177 (100%)

Es decir:

Verdaderos positivos:	48%
Falsos positivos:	24%
Verdaderos negativos:	76%
Falsos negativos:	52%
Estimado correctamente:	71%

Estos datos los compararemos con los datos de test para verificar si finalmente el árbol conseguido es un árbol válido.

- ***Paso 4: Test***

Tras aplicar el modelo a los datos del conjunto de test, se han obtenido los siguientes resultados:

- **Capacidad predictiva por regla:**

Orden	Prior	Ampliada	Reducida	Candidatos	Éxitos	Penetración	Ganancia	%Cand	%Éxitos
1	53.1	**1**	**1**	9,691	5,557	57.34%	2.39	9.8%	23.3%
2	46.2	**1**	**1**	2,290	1,167	50.96%	2.12	2.3%	4.9%
3	43.5	**1**	**1**	1,710	889	51.99%	2.16	1.7%	3.7%
4	41.4	**1**	**1**	1,397	725	51.90%	2.16	1.4%	3.0%
5	35.9	**1**	0	1,316	597	45.36%	1.89	1.3%	2.5%
6	35.8	**1**	0	4,527	1,719	37.97%	1.58	4.6%	7.2%
7	30.8	**1**	0	80	28	35.00%	1.46	0.1%	0.1%
8	28.8	**1**	0	1,726	646	37.43%	1.56	1.7%	2.7%
9	26.5	**1**	0	5,463	1,758	32.18%	1.34	5.5%	7.4%
10	22.1	**1**	0	3,807	1,044	27.42%	1.14	3.8%	4.4%
11	20.1	0	0	6,741	1,989	29.51%	1.23	6.8%	8.3%
12	18.4	0	0	12,515	2,485	19.86%	0.83	12.6%	10.4%
13	17.2	0	0	3,893	1,063	27.31%	1.14	3.9%	4.5%
14	13.9	0	0	933	204	21.86%	0.91	0.9%	0.9%
15	13.3	0	0	9,121	1,375	15.08%	0.63	9.2%	5.8%
16	10.6	0	0	5,603	684	12.21%	0.51	5.6%	2.9%
17	8.5	0	0	7,101	668	9.41%	0.39	7.1%	2.8%
18	7.9	0	0	2,881	296	10.27%	0.43	2.9%	1.2%
19	4.7	0	0	18,591	984	5.29%	0.22	18.7%	4.1%
		Total		99,386	23,878	24,03%	1.00	100.0%	100%

Selección ampliada	1.84	**32.2%**	**59.2%**
No seleccionados	0.60	67.8%	40.8%
Total	1.00	100.0%	100.0%

Selección reducida	2.30	**15.2%**	**34.9%**
No seleccionados	0.77	84.8%	65.1%
Total	1.00	100.0%	100.0%

En términos generales se obtiene similares resultados a los obtenidos con los datos de entrenamiento y validación:

<u>Selección reducida:</u>
Las reglas de mayor propensión a la fuga son las reglas 1,2,3 y4:

– Se obtiene una ganancia global de X2,30.
– Con un 15,2% de los datos se predice el 34,9% de las fugas.

<u>Selección ampliada:</u>
Si seleccionamos todas las reglas con las que se obtiene algo de ganancia (de la regla 1-10):

– Se obtiene una ganancia global de X1,84.
– Con un 32,2% de los datos se predice el 59,2% de las fugas.

Si analizamos regla por regla, vemos que cada una de ellas tiene ganancias similares tanto en los datos de entrenamiento y validación como con los de test, luego el modelo obtenido es un modelo válido.

Orden	Prior	Ampliada	Reducida	Ganancia Entrenam	Ganancia Test
1	53.1	**1**	**1**	2.81	2.39
2	46.2	**1**	**1**	2.45	2.12
3	43.5	**1**	**1**	2.27	2.16
4	41.4	**1**	**1**	2.19	2.16
5	35.9	**1**	0	1.92	1.89
6	35.8	**1**	0	1.90	1.58
7	30.8	**1**	0	1.64	1.46
8	28.8	**1**	0	1.51	1.56
9	26.5	**1**	0	1.40	1.34
10	22.1	**1**	0	1.16	1.14
11	20.1	0	0	1.06	1.23
12	18.4	0	0	0.97	0.83
13	17.2	0	0	0.91	1.14
14	13.9	0	0	0.77	0.91
15	13.3	0	0	0.70	0.63
16	10.6	0	0	0.56	0.51
17	8.5	0	0	0.45	0.39
18	7.9	0	0	0.43	0.43
19	4.7	0	0	0.25	0.22

- **Matriz de clasificación:**

  Se obtienen los siguientes resultados:

	Real 0	Real 1
Estimado 0	57.631 (76%)	9.748 (41%)
Estimado 1	17.877 (24%)	14.130 (59%)
Total	75.508 (100%)	32.007 (100%)

Verdaderos positivos:	59%
Falsos positivos:	24%
Verdaderos negativos:	76%
Falsos negativos:	41%
Estimado correctamente:	72%

Que son muy similares a los datos obtenidos con los datos de entrenamiento y validación.

Por lo tanto, se puede concluir que el modelo es válido.

- ***Paso 5: Resultados obtenidos e interpretación***

- **Descripción de las reglas obtenidas**
  El perfil de los clientes de cada una de las reglas es:

Orden	Prior	Ganancia	Descripción	Regla
1	53.1	2.39	Clientes que han dado partes en su seguro de hogar, tienen menos de 4.600€ como límite en su tarjeta de crédito y se gastan menos de 560€ con su tarjeta de débito	IF IMP_USO_TJDEB_ULT1 < 560.51 AND LIMITE_TJCRED_ULT1 < 4617 AND IND_PARTES_HOGAR = 1
2	46.2	2.12	Clientes que han dado partes en su seguro de hogar, tienen menos de 4.600€ como límite en su tarjeta de crédito y se gastan más de 560€ con su tarjeta de débito	560.51 <= IMP_USO_TJDEB_ULT1 AND LIMITE_TJCRED_ULT1 < 4617 AND IND_PARTES_HOGAR = 1
3	43.5	2.16	Clientes que han dado partes en su seguro de hogar, tienen más de 4.600€ como límite en su tarjeta de crédito y tienen menos de 1200€ en productos de inversión y ahorro	IMP_PROD_INVAH_ULT1 < 1250.345 AND 4617 <= LIMITE_TJCRED_ULT1 AND IND_PARTES_HOGAR = 1
4	41.4	2.16	No han dado partes con su seguro de hogar, hacen más de una operación con su tarjeta de crédito durante los últimos 3 meses, tienen que pagar dentro de poco el próximo recibo y tiene más tarjetas de crédito que la media	1.3865259066 <= COC_TJCRED_ULT1 AND NUM_DIAS_CAT = 0 AND 1.5 <= NUM_OPER_CRED_ULT3 AND IND_PARTES_HOGAR = 0
5	35.9	1.89	Han dado partes con su seguro de hogar, tienen más de 4.600 de límite en su tarjeta de crédito y tienen más de 1.200€ en productos de inversión y ahorro	1250.345 <= IMP_PROD_INVAH_ULT1 AND 4617 <= LIMITE_TJCRED_ULT1 AND IND_PARTES_HOGAR = 1
6	35.8	1.58	No han dado partes con su seguro de hogar, hace menos de 2 operaciones con su tarjeta de crédito en los últimos 3 meses, en el pago del anterior recibo le hicieron oferta especial y tienen el doble de productos de ahorro que la media	2.2703216975 <= RATIO_NUM_PROD_AH_MEDIA_ULT1 AND 0.5 <= RECIB_PROMOCION AND NUM_OPER_CRED_ULT3 < 1.5 AND IND_PARTES_HOGAR = 0
7	30.8	1.46	No han dado partes con su seguro de hogar, hace menos de 2 operaciones con su tarjeta de crédito en los últimos 3 meses, en el pago del anterior recibo le hicieron oferta especial, no tienen el doble de productos de ahorro que la media y tienen el mismo número de tarjetas de crédito o más que la media	0.9427850916 <= COC_TJCRED_ULT1 AND RATIO_NUM_PROD_AH_MEDIA_ULT1 < 2.2703216975 AND 0.5 <= RECIB_PROMOCION AND NUM_OPER_CRED_ULT3 < 1.5 AND IND_PARTES_HOGAR = 0
8	28.8	1.56	No han dado partes de hogar, realizan más de una operación con su tarjeta de crédito en los últimos 3 meses, tienen que pagar dentro de poco el próximo recibo, el número de tarjetas de crédito que tienen en relación a la media es menor de 1.38 y la cuota con su tarjeta de crédito es menor de 32€	CUOTA_TJCRED_ULT1 < 32.275 AND COC_TJCRED_ULT1 < 1.3865259066 AND NUM_DIAS_CAT = 0 AND 1.5 <= NUM_OPER_CRED_ULT3 AND IND_PARTES_HOGAR = 0
9	26.5	1.34	No han dado partes de hogar, realizan menos de 2 operaciones con su tarjeta de crédito durente los últimos 3 meses, en el último recibo no tuvieron oferta especial y tienen 2 productos o más de ahorro e inversión	1.5 <= NUM_PROD_AHINV_ULT1 AND RECIB_PROMOCION < 0.5 AND NUM_OPER_CRED_ULT3 < 1.5 AND IND_PARTES_HOGAR = 0
10	22.1	1.14	No han dado partes de hogar, realizan 2 operaciones o más con su tarjeta de crédito , han pagado el último recibo recientemente, el número de tarjetas de crédito sobre la media es mayor de 1.38 y la cuota con su tarjeta de crédito en los últimos 3 meses es menor de 2.500€	CUOTA_TJCRED_ULT3 < 2506.015 AND 1.3865259066 <= COC_TJCRED_ULT1 AND NUM_DIAS_CAT = 1 AND 1.5 <= NUM_OPER_CRED_ULT3 AND IND_PARTES_HOGAR = 0
11	20.1	1.23	No han dado partes con su seguro de hogar, realizan menos de 2 operaciones con su tarjeta de crédito en los últimos 3 meses, han recibido oferta especial cuando pagaron el último recibo, el ratio de productos de ahorro sobre la media es menor	COC_TJCRED_ULT1 < 0.9427850916 AND RATIO_NUM_PROD_AH_MEDIA_ULT1 < 2.2703216975 AND 0.5 <= RECIB_PROMOCION AND NUM_OPER_CRED_ULT3 < 1.5

			de 2,27 y tienen menos tarjetas de crédito que la media	AND IND_PARTES_HOGAR = 0
12	18.4	0.83	No han dado partes de hogar, realizan menos de 1 operación con su tarjeta de crédito durante los últimos 3 meses, no han recibido promoción en el pago del último recibo, tienen menos de 2 productos de ahorro e inversión y no tienen tarjeta de débito	IND_TJDEB_ULT1 = 0 AND NUM_PROD_AHINV_ULT1 < 1.5 AND RECIB_PROMOCION < 0.5 AND NUM_OPER_CRED_ULT3 < 1.5 AND IND_PARTES_HOGAR = 0
13	17.2	1.14	No han dado partes de hogar, realizan más de 1 operación con su tarjeta de crédito durante los últimos 3 meses, tienen que pagar dentro de poco el próximo recibo, el cociente de tarjetas de crédito sobre la media es menor de 1.38, la cuota con su tarjeta de crédito es mayor de 32€ y tienen menos de 488€ financiado a crédito en los últimos 3 meses	IMP_FINANCIADO_CRED_ULT3 < 488.62 AND 32.275 <= CUOTA_TJCRED_ULT1 AND COC_TJCRED_ULT1 < 1.3865259066 AND NUM_DIAS_CAT = 0 AND 1.5 <= NUM_OPER_CRED_ULT3 AND IND_PARTES_HOGAR = 0
14	13.9	0.91	No han dado partes de hogar y el número de operaciones con su tarjeta de crédito en los últimos 3 meses es mayor de una, acaban de pagar el recibo, el número tarjetas de crédito sobre la media es mayor que 1.38 y la cuota con su tarjeta de crédito en los últimos 3 meses es mayor de 2.506€	2506.015 <= CUOTA_TJCRED_ULT3 AND 1.3865259066 <= COC_TJCRED_ULT1 AND NUM_DIAS_CAT = 1 AND 1.5 <= NUM_OPER_CRED_ULT3 AND IND_PARTES_HOGAR = 0
15	13.3	0.63	No han dado partes de hogar y realizan menos de 2 operaciones con la tarjeta de crédito en los últimos 3 meses y no recibieron promoción en el pago del anterior recibo y tienen 1 producto o menos de ahorro e inversión en el último mes y tiene tarjeta de débito activa	IND_TJDEB_ULT1 = 1 AND NUM_PROD_AHINV_ULT1 < 1.5 AND RECIB_PROMOCION < 0.5 AND NUM_OPER_CRED_ULT3 < 1.5 AND IND_PARTES_HOGAR = 0
16	10.6	0.51	No han dado partes de hogar y realizan 2 operaciones o más con la tarjeta de crédito en los últimos 3 meses y acaban de pagar el recibo de este año y el número de tarjetas de crédito en relación a la media es menor de 1.38 y tienen 2 productos o más de ahorro e inversión en el último mes	1.5 <= NUM_PROD_AHINV_ULT1 AND COC_TJCRED_ULT1 < 1.3865259066 AND NUM_DIAS_CAT = 1 AND 1.5 <= NUM_OPER_CRED_ULT3 AND IND_PARTES_HOGAR = 0
17	8.5	0.39	No han dado partes de hogar y el número de operaciones con su tarjeta de crédito en los últimos 3 meses es mayor o igual de 2 y acaban de pagar el recibo de este año y el número de tarjetas de crédito en relación a la media es menor de 1.38 y tienen menos de 2 productos de ahorro e inversión en el último mes y el importe realizado con las tarjetas de crédito en el último mes respecto la media esmenor de 0.20	COC_IMP_OPER_TJCRED_ULT1 < 0.2036808347 AND NUM_PROD_AHINV_ULT1 < 1.5 AND COC_TJCRED_ULT1 < 1.3865259066 AND NUM_DIAS_CAT = 1 AND 1.5 <= NUM_OPER_CRED_ULT3 AND IND_PARTES_HOGAR = 0
18	7.9	0.43	No han dado partes de hogar y el número de operaciones con su tarjeta de crédito en los últimos 3 meses es mayor o igual de 2 y les quedan pocos días para pagar el recibo y el número de tarjetas de crédito en relación a la media es menor de 1.38 y la cuota con su tarjeta de crédito en el último mes es mayor de 32€ y importe financiado a crédito en los últimos 3 meses es mayor de 488€	488.62 <= IMP_FINANCIADO_CRED_ULT3 AND 32.275 <= CUOTA_TJCRED_ULT1 AND COC_TJCRED_ULT1 < 1.3865259066 AND NUM_DIAS_CAT = 0 AND 1.5 <= NUM_OPER_CRED_ULT3 AND IND_PARTES_HOGAR = 0
19	4.7	0.22	No han dado partes de hogar y el número de operaciones con su tarjeta de crédito en los últimos 3 meses es mayor o igual de 2 y acaban de pagar el recibo de este año y el número de tarjetas de crédito en relación a la media es menor de 1.38 y el número de productos de ahorro e inversión es menor o igual que uno y el importe de operaciones con su tarjeta de crédito relativizado entre la media es mayor o igual 0.20	0.2036808347 <= COC_IMP_OPER_TJCRED_ULT1 AND NUM_PROD_AHINV_ULT1 < 1.5 AND COC_TJCRED_ULT1 < 1.3865259066 AND NUM_DIAS_CAT = 1 AND 1.5 <= NUM_OPER_CRED_ULT3 AND IND_PARTES_HOGAR = 0

- **Importancia de las variables que forman parte del árbol**

  Las variables más han aportado en la explicación del variable objetivo (ordenadas de mayor a menor importancia son:

obs	Name	nrules	importance	vimportance	ratio
1	IND_PARTES_HOGAR	1	1.00000	1.00000	1.00000
2	NUM_OPER_CRED_ULT3	1	0.41889	0.41122	0.98169
3	COC_TJCRED_ULT1	3	0.33365	0.32587	0.97668
4	NUM_DIAS_CAT	1	0.22780	0.23419	1.02804
5	NUM_PROD_AHINV_ULT1	2	0.21650	0.22972	1.06104
6	RECIB_PROMOCION	1	0.18415	0.19129	1.03880
7	RATIO_NUM_PROD_AH_MEDIA_ULT1	1	0.16949	0.17582	1.03735
8	CUOTA_TJCRED_ULT1	1	0.14475	0.12995	0.89775
9	LIMITE_TJCRED_ULT1	1	0.14080	0.14884	1.05708
10	IND_TJDEB_ULT1	1	0.11064	0.10688	0.96596
11	IMP_FINANCIADO_CRED_ULT3	1	0.09876	0.09084	0.91988
12	COC_IMP_OPER_TJCRED_ULT1	1	0.07973	0.07579	0.95058
13	IMP_USO_TJDEB_ULT1	1	0.07739	0.06815	0.88058
14	CUOTA_TJCRED_ULT3	1	0.06540	0.03717	0.56843
15	IMP_PROD_INVAH_ULT1	1	0.05213	0.01087	0.20861

- **Capacidad predictiva del modelo**

  En términos generales se han obtenido los siguientes resultados:

  <u>Selección reducida:</u>

  Las reglas de mayor propensión a la fuga son las reglas 1,2,3 y4:

  – Se obtiene una ganancia global de X2,30
  – Con un 15,2% de los datos se predice el 34,9% de las fugas

  <u>Selección ampliada:</u>

  Si seleccionamos todas las reglas con algo de ganancia (de la regla 1-10):

  – Se obtiene una ganancia global de X1,84
  – Con un 32,2% de los datos se predice el 59,2% de las fugas

- **Matriz de clasificación:**

  El modelo consigue clasificar correctamente al 72% de los clientes:

Verdaderos positivos:	44%
Falsos positivos:	14%
Verdaderos negativos:	86%
Falsos negativos:	56%

- **Cálculo de la esperanza de vida:**

  El número de meses estimados que transcurrirán hasta que el cliente de baja el producto es:

Regla	P(fuga)	Nº Meses hasta la baja
1	0,531	1,88
2	0,462	2,16
3	0,435	2,30
4	0,414	2,42
5	0,359	2,79
6	0,358	2,79
7	0,308	3,25
8	0,288	3,47
9	0,265	3,77
10	0,221	4,52
11	0,201	4,98
12	0,184	5,43
13	0,172	5,81
14	0,139	7,19
15	0,133	7,52
16	0,106	9,43
17	0,085	11,76
18	0,079	12,66
19	0,047	21,28

  Es decir, los clientes con mayor propensión a dar de baja el producto (los clientes de la regla 1) darán de baja el producto dentro de casi 2 meses (1,88 meses)

4.2 *Regresión logística*

- ***Paso 1: Búsqueda de variables correladas con la variable objetivo***

- **Métodos univariantes:**

VARIABLE	CRAMERS'V	CHI-CUADRADO	PROB	GRADOS DE LIBERTAD	ORDEN
NUM_DIAS_CAT	0,233	20214,25	0,0000	1	1
IND_TJDEB_ULT1	0,076	2180,81	0,0000	1	2
IND_TJDEB_ULT3	0,071	1861,27	0,0000	1	3
IND_TJDEB_ULT6	0,063	1507,49	0,0000	1	4
NIVEL_SATISFACCION_CAT	0,039	579,38	0,0000	1	5
IND_NACIONAL_CAT	0,037	524,58	0,0000	1	6
RECIB_PROMOCION	0,035	458,75	0,0000	1	7
VINCULACION_CAT	0,023	191,02	0,0000	2	8
NIVEL_ESTUDIOS_CAT	0,021	166,48	0,0000	2	9
ESTADO_CIVIL_CAT	0,010	35,09	0,0000	3	10
SEXO_CAT	0,002	2,23	0,1350	1	11

En la tabla se muestra las variables categóricas con mayor asociación respecto el variable objetivo, ordenadas de mayor a menor fuerza de asociación.

Destaca la variable NUM_DIAS_CAT sobre el resto, aunque su relación con la variable objetivo es moderada (valor V-cramer=0,233)

VARIABLE	IMPORTANCIA	VALOR
NUM_DIAS_CAT	1	0,019
NUM_OPER_CRED_ULT3	2	0,013
IMP_OPER_CRED_ULT3	3	0,013
CUOTA_TJCRED_ULT3	4	0,012
CUOTA_TJCRED_ULT1	5	0,012
COC_CUOTA_TJCRED_ULT1	6	0,011
COC_NUM_OPER_TJCRED_ULT1	7	0,009
NUM_OPER_CRED_ULT1	8	0,009
COC_IMP_OPER_TJCRED_ULT1	9	0,009
IMP_OPER_CRED_ULT1	10	0,009
COC_IMP_FINAN_CRED_ULT1	11	0,006
IMP_FINANCIADO_CRED_ULT3	12	0,005
COC_TJCRED_ULT1	13	0,003
RATIO_NUM_PROD_AHINV_MED_HACE1	14	0,003
IMP_USO_TJDEB_ULT1	15	0,002
IMP_USO_TJDEB_CAJERO_ULT1	16	0,002
LIMITE_TJCRED_ULT1	17	0,002
COC_LIM_TJCRED_ULT1	18	0,002

IND_TJDEB_ULT1	19	0,002
IMP_USO_TJDEB_ULT3	20	0,002
IMP_USO_TJDEB_CAJERO_ULT3	21	0,002
IMP_FINANCIADO_CRED_ULT1	22	0,002
IND_TJDEB_ULT3	23	0,001
NUM_DIAS_ULT_USO_TJDEB	24	0,001
IMP_USO_TJDEB_ULT6	25	0,001
NUM_PROD_AHINV_ULT1	26	0,001
IMP_USO_TJDEB_CAJERO_ULT6	27	0,001
IMP_USO_TJDEB_COMPRAS_ULT1	28	0,001
RATIO_NUM_PROD_AH_MEDIA_ULT1	29	0,001
IND_TJDEB_ULT6	30	0,001
IMP_USO_TJDEB_COMPRAS_ULT3	31	0,001
NUM_TJDEB_HACE1	32	0,001
N_IMPORTE_ULTIMO_COBRO	33	0,001
NUM_TJDEB_HACE3	34	0,001
IMP_USO_TJDEB_COMPRAS_ULT6	35	0,001
RECIB_PROMOCION	36	0,001
IND_NACIONAL_CAT	37	0,001
EDAD	38	0,001
NUM_CAMPANIAS_ULT3	39	0,001
CRED_DISPONIBLE_TJCRED_ULT1	40	0,000
RATIO_NUM_PROD_INV_MEDIA_ULT1	41	0,000
NUM_TJDEB_HACE6	42	0,000
NUM_CAMPANIAS_ULT3_2	43	0,000
NIVEL_SATISFACCION_CAT	44	0,000
NUM_PROD_AH_ULT1	45	0,000
NUM_PROD_INV_ULT1	46	0,000
NIVEL_ESTUDIOS_CAT	47	0,000
VINCULACION_CAT	48	0,000
ANTIGUEDAD_CLIENTE	49	0,000
NUM_PARTES_VIDA_SALUD_ULT4	50	0,000
IMP_PROD_INVAH_ULT1	51	0,000
IMP_PROD_INVAH_HACE3	52	0,000
NUM_SALUD_HACE3	53	0,000
NUM_SALUD_ULT1	54	0,000
NUM_PERSONAS_DEPENDIENTES	55	0,000
ESTADO_CIVIL_CAT	56	0,000
NUM_VIDA_ULT1	57	0,000
NUM_VIDA_HACE3	58	0,000
SEXO_CAT	59	0,000

- **Métodos multivariantes**

VARIABLE	NÚMERO DE REGLAS EN ÁRBOL	IMPORTANCIA RELATIVA
NUM_DIAS_CAT	1	1,000
COC_CUOTA_TJCRED_ULT1	3	0,894
COC_TJCRED_ULT1	4	0,375
NUM_PROD_AHINV_ULT1	4	0,293
IMP_OPER_CRED_ULT1	3	0,267
RECIB_PROMOCION	2	0,227
COC_LIM_TJCRED_ULT1	1	0,210
NUM_PROD_AH_ULT1	3	0,182
COC_IMP_FINAN_CRED_ULT1	1	0,134
CUOTA_TJCRED_ULT3	2	0,119
NUM_OPER_CRED_ULT3	3	0,111
IND_RESIDENTE_CAT	1	0,101
IMP_FINANCIADO_CRED_ULT3	2	0,094
NUM_DIAS_ULT_USO_TJDEB	1	0,032

Obs	VARIABLE	NÚMERO DE REGLAS	NÚMERO DE REGLAS SUBROGADAS	IMPORTANCIA ENTRENAMIENTO	IMPORTANCIA TEST	RATIO
1	COC_LIM_TJCRED_ULT1	4	28	1	1	**1**
2	LIMITE_TJCRED_ULT1	3	23	0,9889	0,98878	**0,99987**
3	NUM_DIAS_CAT	1	0	0,96281	0,9772	**1,01494**
4	CUOTA_TJCRED_ULT3	6	15	0,95592	0,95416	**0,99815**
5	NUM_OPER_CRED_ULT3	2	25	0,94688	0,94232	**0,99519**
6	IMP_OPER_CRED_ULT3	5	13	0,92912	0,92914	**1,00002**
7	CUOTA_TJCRED_ULT1	2	24	0,92516	0,92013	**0,99455**
8	COC_CUOTA_TJCRED_ULT1	2	24	0,92414	0,91974	**0,99524**
9	COC_IMP_OPER_TJCRED_ULT1	3	21	0,89912	0,89094	**0,99091**
11	IMP_USO_TJDEB_CAJERO_ULT6	1	14	0,8933	0,89089	**0,99731**
13	NUM_CAMPANIAS_ULT3	4	7	0,88225	0,88397	**1,00195**
14	IMP_FINANCIADO_CRED_ULT3	1	16	0,85279	0,86058	**1,00913**
21	COC_IMP_FINAN_CRED_ULT1	6	11	0,85174	0,85757	**1,00684**
22	COC_TJCRED_ULT1	4	14	0,54199	0,53869	**0,9939**
23	RATIO_NUM_PROD_AHINV_MED_HACE1	4	16	0,52631	0,52438	**0,99633**
24	RATIO_NUM_PROD_INV_MEDIA_ULT1	1	20	0,50662	0,50008	**0,9871**
26	NUM_PROD_AHINV_ULT1	5	12	0,45524	0,44547	**0,97855**
28	NUM_PROD_AH_ULT1	1	16	0,41071	0,41324	**1,00617**
30	IMP_PROD_INVAH_ULT1	2	11	0,33592	0,30223	**0,8997**
31	IMP_USO_TJDEB_CAJERO_ULT1	2	13	0,27957	0,2112	**0,75543**
33	IMP_USO_TJDEB_COMPRAS_ULT1	1	17	0,27309	0,24051	**0,88072**
34	EDAD	2	12	0,26325	0,24785	**0,94151**
37	N_IMPORTE_ULTIMO_COBRO	1	2	0,23428	0,22733	**0,97033**
40	RECIB_PROMOCION	1	0	0,20261	0,20304	**1,00213**
41	IMP_USO_TJDEB_ULT1	4	9	0,19884	0,1559	**0,78404**

42	NUM_DIAS_ULT_USO_TJDEB	1	4	0,1799	0,1588	**0,88272**
43	IND_TJDEB_ULT1	1	7	0,17396	0,14523	**0,8348**
44	IND_NACIONAL_CAT	1	3	0,15762	0,16412	**1,04124**
47	IMP_USO_TJDEB_COMPRAS_ULT6	2	10	0,15445	0,11807	**0,76443**
48	VINCULACION_CAT	1	2	0,11936	0,10039	**0,8411**

En resumen, las variables más significativas que son con las que vamos a entrenar el modelo de regresión logística son:

Variables numércias
COC_CUOTA_TJCRED_ULT1
COC_IMP_FINAN_CRED_ULT1
COC_IMP_OPER_TJCRED_ULT1
COC_LIM_TJCRED_ULT1
COC_NUM_OPER_TJCRED_ULT1
COC_TJCRED_ULT1
CUOTA_TJCRED_ULT1
CUOTA_TJCRED_ULT3
EDAD
IMP_FINANCIADO_CRED_ULT1
IMP_FINANCIADO_CRED_ULT3
IMP_OPER_CRED_ULT1
IMP_OPER_CRED_ULT3
IMP_PROD_INVAH_ULT1
IMP_USO_TJDEB_CAJERO_ULT1
IMP_USO_TJDEB_CAJERO_ULT3
IMP_USO_TJDEB_CAJERO_ULT6
IMP_USO_TJDEB_COMPRAS_ULT1
IMP_USO_TJDEB_ULT1
IMP_USO_TJDEB_ULT3
IMP_USO_TJDEB_ULT6
LIMITE_TJCRED_ULT1
NUM_CAMPANIAS_ULT3
NUM_DIAS_ULT_USO_TJDEB
NUM_OPER_CRED_ULT1
NUM_OPER_CRED_ULT3
NUM_PROD_AHINV_ULT1
NUM_PROD_AH_ULT1
NUM_PROD_INV_ULT1
NUM_TJDEB_HACE1
NUM_TJDEB_HACE3

Variables categóricas
IND_NACIONAL_CAT
IND_PARTES_HOGAR
IND_RECLAMACIONES_ULT3_CAT
IND_RESIDENTE_CAT
IND_TJDEB_ULT1
IND_TJDEB_ULT3
IND_TJDEB_ULT6
NIVEL_ESTUDIOS_CAT
NIVEL_SATISFACCION_CAT
NUM_DIAS_CAT
RECIB_PROMOCION
SEXO_CAT
VINCULACION_CAT

NUM_TJDEB_HACE6
NUM_VIDA_HACE3
NUM_VIDA_ULT1
RATIO_NUM_PROD_AHINV
RATIO_NUM_PROD_AH_MEDIA_ULT1
RATIO_NUM_PROD_AHINV_MED_HACE1

- ***Paso 2: Categorización de variables numéricas***

Se categorizan aquellas variables numéricas que tienen relación con la variable dependiente, (que son las obtenidas en el paso anterior). Se crearán por lo tanto las siguientes variables que sustituyen a las anteriores:

NUEVA VARIABLE	VALORES POSIBLES
CAT_COC_CUOTA_TJCRED_ULT1	1: si COC_CUOTA_TJCRED_ULT1 < 0.66 2:en caso contrario
CAT_COC_IMP_FINAN_CRED_ULT1	1: si COC_IMP_FINAN_CRED_ULT1< 0.4 2:en caso contrario
CAT_COC_IMP_OPER_TJCRED_ULT1	1: si COC_IMP_OPER_TJCRED_ULT1 < 0.17 2:en caso contrario
CAT_COC_LIM_TJCRED_ULT1	1: si COC_LIM_TJCRED_ULT1 < 0.18 2:en caso contrario
CAT_COC_NUM_OPER_TJCRED_ULT1	1: si COC_NUM_OPER_TJCRED_ULT1 < 0.14 2:en caso contrario
CAT_COC_TJCRED_ULT1	1: si COC_TJCRED_ULT1 < 0.46 2: si 0.46<= COC_TJCRED_ULT1 < 0.94 3: si 0.94<= COC_TJCRED_ULT1 < 1.83 4: si COC_TJCRED_ULT1 >= 1.83
CAT_CUOTA_TJCRED_ULT1	1: si CUOTA_TJCRED_ULT1< 31.705 2:en caso contrario
CAT_CUOTA_TJCRED_ULT3	1: si CUOTA_TJCRED_ULT3< 38.775 2:en caso contrario
CAT_EDAD	1: si EDAD< 498 2: si 498<= EDAD < 638 3: si EDAD >= 638
CAT_IMP_FINANCIADO_CRED_ULT1	1: si IMP_FINANCIADO_CRED_ULT1< 113.13 2:en caso contrario
CAT_IMP_FINANCIADO_CRED_ULT3	1: si IMP_FINANCIADO_CRED_ULT3< 290.15 2:en caso contrario
CAT_IMP_OPER_CRED_ULT1	1: si IMP_OPER_CRED_ULT1< 282.12 2:en caso contrario
CAT_IMP_OPER_CRED_ULT3	1: si IMP_OPER_CRED_ULT< 422.57 2:en caso contrario
CAT_IMP_PROD_INVAH_ULT1	1: si IMP_PROD_INVAH_ULT1< 466.19 2:en caso contrario
CAT_IMP_USO_TJDEB_CAJERO_ULT1	1: si IMP_USO_TJDEB_CAJERO_ULT1< 13.5 2:en caso contrario

CAT_IMP_USO_TJDEB_CAJERO_ULT3	1: si IMP_USO_TJDEB_CAJERO_ULT3< 310.5 2:en caso contrario
CAT_IMP_USO_TJDEB_CAJERO_ULT6	1: si IMP_USO_TJDEB_CAJERO_ULT6< 41 2:en caso contrario
CAT_IMP_USO_TJDEB_COMPRAS_ULT1	1: si IMP_USO_TJDEB_COMPRAS_ULT1 < 75.25 2:en caso contrario
CAT_IMP_USO_TJDEB_ULT1	1: si IMP_USO_TJDEB_ULT1< 97.8 2:en caso contrario
CAT_IMP_USO_TJDEB_ULT3	1: si IMP_USO_TJDEB_ULT3< 0.5 2:en caso contrario
CAT_IMP_USO_TJDEB_ULT6	1: si IMP_USO_TJDEB_ULT6< 97.7 2:en caso contrario
CAT_LIMITE_TJCRED_ULT1	1: si LIMITE_TJCRED_ULT1< 2944 2:en caso contrario
CAT_NUM_CAMPANIAS_ULT3	1: si NUM_CAMPANIAS_ULT3< 9.5 2:en caso contrario
CAT_NUM_DIAS_ULT_USO_TJDEB	1: si NUM_DIAS_ULT_USO_TJDEB< 35.5 2:en caso contrario
CAT_NUM_OPER_CRED_ULT1	1: si NUM_OPER_CRED_ULT1< 0.5 2:en caso contrario
CAT_NUM_OPER_CRED_ULT3	1: si NUM_OPER_CRED_ULT3< 0.5 2:en caso contrario
CAT_NUM_PROD_AH_ULT1	1: si NUM_PROD_AH_ULT1< 0.5 2: si 0.5<= NUM_PROD_AH_ULT1< 1.5 3: si NUM_PROD_AH_ULT1>= 1.5
CAT_NUM_PROD_AHINV_ULT1	1: si NUM_PROD_AHINV_ULT1< 1.5 2:en caso contrario
CAT_NUM_PROD_INV_ULT1	1: si NUM_PROD_INV_ULT1< 1.5 2:en caso contrario
CAT_NUM_TJDEB_HACE1	1: si NUM_TJDEB_HACE1< 1 2:en caso contrario
CAT_NUM_TJDEB_HACE3	1: si NUM_TJDEB_HACE3< 1 2:en caso contrario
CAT_NUM_TJDEB_HACE6	1: si NUM_TJDEB_HACE6< 1 2:en caso contrario
CAT_NUM_VIDA_HACE3	1: si NUM_VIDA_HACE3< 0.5 2:en caso contrario
CAT_NUM_VIDA_ULT1	1: si NUM_VIDA_ULT1< 0.5 2:en caso contrario
CAT_RATIO_NUM_PROD_AH_MEDIA_ULT1	1: si RATIO_NUM_PROD_AH_MEDIA_ULT1< 2.27 2:en caso contrario
CAT_RPROD_AHINV_MED_HACE1	1: si RATIO_NUM_PROD_AHINV_MED_HACE1< 0.91 2: si 0.91<= RATIO_NUM_PROD_AHINV_MED_HACE1 < 1.4 3: si 1.4<= RATIO_NUM_PROD_AHINV_MED_HACE1 < 1.83 4: si 1.83<=RATIO_NUM_PROD_AHINV_MED_HACE1 < 2.01 5: resto

Los nombres de las variables son largos, y cuando más adelante se muestren los resultados de las interacciones entre ellas va a ser prácticamente ilegible. Por ello, a partir de ahora nos referiremos a dichas variables con su notación corta.

Variable	Etiqueta
CAT_COC_CUOTA_TJCRED_ULT1	C2
CAT_COC_IMP_FINAN_CRED_ULT1	C3
CAT_COC_IMP_OPER_TJCRED_ULT1	C4
CAT_COC_LIM_TJCRED_ULT1	C5
CAT_COC_NUM_OPER_TJCRED_ULT1	C6
CAT_COC_TJCRED_ULT1	C7
CAT_CUOTA_TJCRED_ULT1	C9
CAT_CUOTA_TJCRED_ULT3	C10
CAT_EDAD	C11
CAT_IMP_FINANCIADO_CRED_ULT1	C13
CAT_IMP_FINANCIADO_CRED_ULT3	C14
CAT_IMP_OPER_CRED_ULT1	C15
CAT_IMP_OPER_CRED_ULT3	C16
CAT_IMP_PROD_INVAH_ULT1	C18
CAT_IMP_USO_TJDEB_CAJERO_ULT1	C23
CAT_IMP_USO_TJDEB_CAJERO_ULT3	C24
CAT_IMP_USO_TJDEB_CAJERO_ULT6	C25
CAT_IMP_USO_TJDEB_COMPRAS_ULT1	C26
CAT_IMP_USO_TJDEB_ULT1	C29
CAT_IMP_USO_TJDEB_ULT3	C30
CAT_IMP_USO_TJDEB_ULT6	C31
CAT_LIMITE_TJCRED_ULT1	C42
CAT_NUM_CAMPANIAS_ULT3	C45
CAT_NUM_DIAS_ULT_USO_TJDEB	C48
CAT_NUM_OPER_CRED_ULT1	C49
CAT_NUM_OPER_CRED_ULT3	C50
CAT_NUM_PROD_AHINV_ULT1	C54
CAT_NUM_PROD_AH_ULT1	C53
CAT_NUM_PROD_INV_ULT1	C55
CAT_NUM_TJDEB_HACE1	C58
CAT_NUM_TJDEB_HACE3	C59
CAT_NUM_TJDEB_HACE6	C60
CAT_NUM_VIDA_HACE3	C65
CAT_NUM_VIDA_ULT1	C66
CAT_RATIO_NUM_PROD_AHINV	C68
CAT_RATIO_NUM_PROD_AH_MEDIA_ULT1	C67
CAT_RPROD_AHINV_MED_HACE1	C69
IND_EMPLEADO	C33
IND_NACIONAL_CAT	C34
IND_PARTES_HOGAR	C35
IND_RECLAMACIONES_ULT3_CAT	C36
IND_RESIDENTE_CAT	C37

IND_TJDEB_ULT1	C38
IND_TJDEB_ULT3	C39
IND_TJDEB_ULT6	C40
NIVEL_ESTUDIOS_CAT	C43
NIVEL_SATISFACCION_CAT	C44
NUM_DIAS_CAT	C46
RECIB_PROMOCION	C70
SEXO_CAT	C71
VINCULACION_CAT	C75

- ***Paso 3: Preparación del tablón de datos***

- **Tablón de entrenamiento**

El tablón de entrenamiento contiene 373.889 registros. Como se ha comentado en la parte teórica, para entrenar un modelo de regresión logística son demasiados, por lo que es necesario realizar una muestra aleatoria simple de 50.000 registros (13%) y entrenar el modelo con dicha muestra.

Al tratarse de una muestra aleatoria simple, la proporción de fugas de dicha muestra se debe mantener vs la proporción de fugas del tablón de entrenamiento original. No obstante vamos a comprobarlo:

Datos	N	P(Y=1)
Tablón de entrenamiento sin muestrear	373.889	19%
Muestra	50.000	19%

En ambos la proporción de fugas es del 19%, por lo que se ha extraído la muestra correctamente.

- **Tablón de test:**

Está formado con los datos de todos los clientes de Marzo13, es decir, sin extraer sobre ellos ninguna muestra.

- ***Paso 4: Entrenamiento del modelo inicial (sin interacciones)***

Los resultados obtenidos son:

Paso	Variable que entra	Variable que sale	Grados de libertad	Orden de entrada	Score Chi-cuadrado	Pr > Chi
1	C35		1	1	3898.6403	<.0001
2	C16		1	2	785.5852	<.0001
3	C69		4	3	687.2316	<.0001
4	C46		1	4	382.2001	<.0001
5	C3		1	5	339.3434	<.0001
6	C4		1	6	141.9545	<.0001
7	C44		1	7	77.6544	<.0001
8	C7		3	8	68.9151	<.0001
9	C70		1	9	65.8672	<.0001
10	C53		2	10	63.8738	<.0001
11	C29		1	11	61.6669	<.0001
12	C11		2	12	55.4631	<.0001
13	C42		1	13	32.8404	<.0001
14	C37		1	14	27.0944	<.0001
15	C2		1	15	20.3852	<.0001
16	C71		1	16	19.6328	<.0001
17	C75		2	17	22.4991	<.0001
18	C18		1	18	14.5125	0.0001
19	C34		1	19	13.4126	0.0002
20	C36		1	20	12.243	0.0005
21	C40		1	21	12.081	0.0005
22	C48		1	22	16.7131	<.0001
23	C49		1	23	10.2868	0.0013
24	C60		1	24	7.8026	0.0052
25	C14		1	25	5.7284	0.0167
26	C58		1	26	5.1273	0.0236
27	C23		1	27	4.7866	0.0287

Esta tabla muestra el resumen de la selección de variables con el método stepwise. Se puede observar que una vez que han entrado en el modelo, ninguna de ellas ha salido por la inclusión de otra variable en el modelo.

Todas ellas tienen un p-valor menor o igual a 0,03 que es el requisito que habíamos marcado para que cada una de las variables permanezcan en el modelo.

En total hay 27 variables significativas en el modelo inicial sin interacciones

Test Ratio de verosimilitud para la hipótesis nula global: Beta=0
(Likelihood Ratio Test for Global Null Hypothesis: BETA=0)

-2 Log Verosimilitud				
Término independiente solamente	Término independiente y variables	Ratio de verosilimilitud chi-cuadrado	Grados de libertad	Pr > Chi
48369.009	42320.112	6048.8963	35	<.0001

En esta tabla vemos que el p-valor es significativo, lo cual quiere decir, que el modelo obtenido tiene mejor ajuste que el modelo que está formado solamente por el término independiente.

Tipo 3 Análisis de los efectos
(TipeIII Analysis effects)

Effect	DF	Wald Chi-Square	Pr > ChiSq
C2	1	24.2837	<.0001
C14	1	5.6819	0.0171
C16	1	87.1743	<.0001
C18	1	14.2199	0.0002
C23	1	4.7847	0.0287
C29	1	15.9409	<.0001
C3	1	106.5379	<.0001
C42	1	25.3174	<.0001
C48	1	20.9057	<.0001
C49	1	9.9513	0.0016
C53	2	40.9363	<.0001
C4	1	44.3122	<.0001
C58	1	5.0495	0.0246
C60	1	10.5273	0.0012
C69	4	231.6957	<.0001
C34	1	14.7698	0.0001
C35	1	74.0826	<.0001
C36	1	12.8573	0.0003
C37	1	14.9149	0.0001
C40	1	10.8941	0.001
C44	1	16.6954	<.0001
C46	1	509.628	<.0001
C70	1	44.0777	<.0001
C71	1	20.084	<.0001
C75	2	19.9491	<.0001
C7	3	118.2245	<.0001
C11	2	42.3607	<.0001

En esta tabla se puede observar que todas las variables por si solas son significativas respecto la variable objetivo.

Análisis de estimación de maxima verosimilitud
(Analysis of Maximum Likelihood Estimates)

Parámetro	Categoría	Grados de libertad	Valor estimado	Error estandar	Chi-cuadrado Wald	Pr > Chi	Exp (Valor estimado)
Intercept		1	-1.5699	0.1624	93.48	<.0001	0.208
C2	1	1	0.1241	0.0252	24.28	<.0001	1.132
C14	1	1	-0.1019	0.0428	5.68	0.0171	0.903
C16	1	1	0.2229	0.0239	87.17	<.0001	1.25
C18	1	1	0.0612	0.0162	14.22	0.0002	1.063
C23	1	1	0.0548	0.0251	4.78	0.0287	1.056
C29	1	1	0.1109	0.0278	15.94	<.0001	1.117
C3	1	1	0.4575	0.0443	106.54	<.0001	1.58
C42	1	1	0.0683	0.0136	25.32	<.0001	1.071
C48	1	1	0.0704	0.0154	20.91	<.0001	1.073
C49	1	1	-0.1613	0.0511	9.95	0.0016	0.851
C53	1	1	-0.1513	0.0248	37.32	<.0001	0.86
C53	2	1	0.00921	0.0238	0.15	0.6981	1.009
C4	1	1	0.3349	0.0503	44.31	<.0001	1.398
C58	1	1	0.0926	0.0412	5.05	0.0246	1.097
C60	1	1	-0.1331	0.041	10.53	0.0012	0.875
C69	1	1	-0.2889	0.0488	34.99	<.0001	0.749
C69	2	1	-0.4659	0.0342	185.68	<.0001	0.628
C69	3	1	0.1446	0.051	8.04	0.0046	1.156
C69	4	1	0.0386	0.0376	1.06	0.3043	1.039
C34	NO	1	-0.1076	0.028	14.77	0.0001	0.898
C35	0	1	-0.2327	0.027	74.08	<.0001	0.792
C36	NO	1	-0.3475	0.0969	12.86	0.0003	0.706
C37	NO	1	-0.4076	0.1055	14.91	0.0001	0.665
C40	0	1	-0.0694	0.021	10.89	0.001	0.933
C44	ALTA	1	-0.0936	0.0229	16.7	<.0001	0.911
C46	0	1	0.5183	0.023	509.63	<.0001	1.679
C70	0	1	-0.1149	0.0173	44.08	<.0001	0.891
C71	HOMBRE	1	-0.0563	0.0126	20.08	<.0001	0.945
C75	ALTA	1	-0.4403	0.1478	8.87	0.0029	0.644
C75	BAJA	1	0.2695	0.0757	12.69	0.0004	1.309
C7	1	1	-0.2455	0.0293	70.05	<.0001	0.782
C7	2	1	-0.2365	0.0469	25.45	<.0001	0.789
C7	3	1	0.052	0.0273	3.64	0.0565	1.053

C11	1	1	0.098	0.019	26.46	<.0001	1.103
C11	2	1	-0.1165	0.0189	37.84	<.0001	0.89

En esta tabla se muestra el resultado de la estimación de cada uno de los parámetros y su significatividad teniendo en cuenta el resto de variables que forman parte del modelo.

Estimación de odds ratio
(Odds Ratio Estimates)

Effect		Point estimate
C2	1 vs 2	1.282
C14	1 vs 2	0.816
C16	1 vs 2	1.562
C18	1 vs 2	1.13
C23	1 vs 2	1.116
C29	1 vs 2	1.248
C3	1 vs 2	2.497
C42	1 vs 2	1.146
C48	1 vs 2	1.151
C49	1 vs 2	0.724
C53	1 vs 3	0.746
C53	2 vs 3	0.876
C4	1 vs 2	1.954
C58	1 vs 2	1.203
C60	1 vs 2	0.766
C69	1 vs 5	0.423
C69	2 vs 5	0.354
C69	3 vs 5	0.653
C69	4 vs 5	0.587
C34	NO vs SI	0.806
C35	0 vs 1	0.628
C36	NO vs SI	0.499
C37	NO vs SI	0.443
C40	0 vs 1	0.87
C44	ALTA vs BAJA	0.829
C46	0 vs 1	2.819
C70	0 vs 1	0.795
C71	HOMBRE vs MUJER	0.894
C75	ALTA vs MEDIA	0.543
C75	BAJA vs MEDIA	1.104
C7	1 vs 4	0.509

C7	2 vs 4	0.514
C7	3 vs 4	0.685
C11	1 vs 3	1.083
C11	2 vs 3	0.874

Esta tabla muestra la odds ratio entre cada una de las categorías de las variables versus la categoría de referencia.

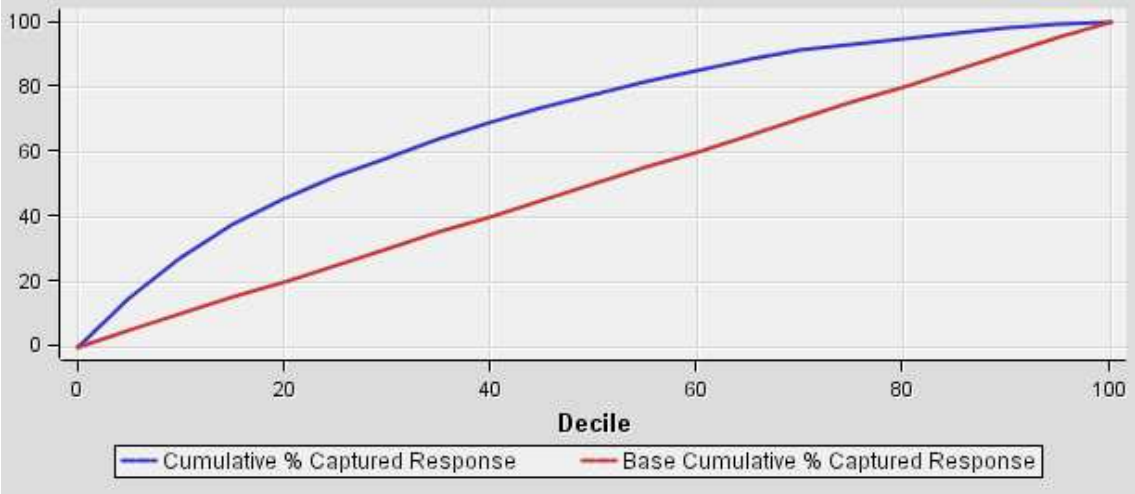
En la siguiente tabla se muestra la probabilidad estimada tramificada en intervalos de 0.05 de amplitud y se observan el número de aciertos y no aciertos que realmente se han obtenido:

Distribución del score
(Assessment Score Distribution)

Rango de probabilidad estimada	Nº eventos	Nº no eventos	Probabilidad a posteriori media	Porcentaje
0.95 - 1.00	0	0	.	0
0.90 - 0.95	0	0	.	0
0.85 - 0.90	1	0	0.85539	0.002
0.80 - 0.85	18	7	0.82252	0.05
0.75 - 0.80	28	13	0.77159	0.082
0.70 - 0.75	57	50	0.71939	0.214
0.65 - 0.70	82	69	0.67388	0.302
0.60 - 0.65	113	94	0.62164	0.414
0.55 - 0.60	395	309	0.57088	1.408
0.50 - 0.55	746	680	0.5223	2.852
0.45 - 0.50	824	859	0.47489	3.366
0.40 - 0.45	656	824	0.42586	2.96
0.35 - 0.40	421	743	0.37529	2.328
0.30 - 0.35	442	914	0.32387	2.712
0.25 - 0.30	706	1697	0.27205	4.806
0.20 - 0.25	1264	4315	0.2218	11.158
0.15 - 0.20	1652	7745	0.17435	18.794
0.10 - 0.15	1097	7578	0.12577	17.35
0.05 - 0.10	740	10220	0.0737	21.92
0.00 - 0.05	171	4470	0.03705	9.282

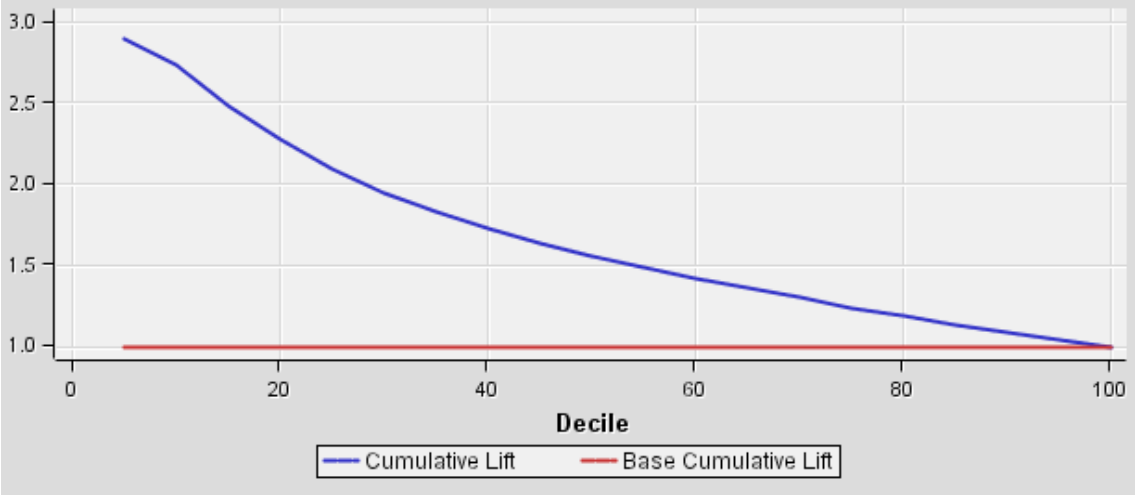
Con esta tabla se calculará la matriz de clasificación

- **Porcentaje de éxitos capturados:**



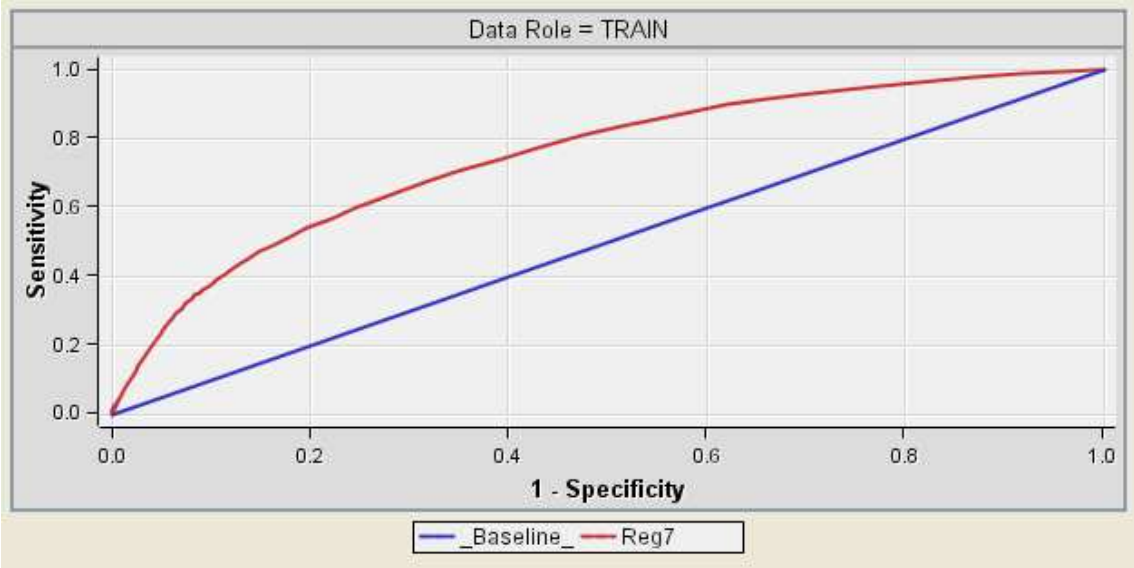
  – Con un 20% de los clientes predecimos el 45% de las fugas
  – Con un 40% de los clientes predecimos el 69% de las fugas

- **Mejora acumulada:**



  – Con un 20% de los clientes se obtiene una ganancia de x2,3
  – Con un 40% de los clientes se obtiene una ganancia de x1,7

- **Curva ROC:**



  - Para conseguir un 21% de verdaderos positivos se tendrá un 5,5% de falsos positivos
  - Para conseguir un 40% de verdaderos positivos se tendrá un 11,6% de falsos positivos

- **Matriz de clasificación:**
  La probabilidad estimada se resume en la siguiente tabla:

Rango de probabilidad estimada	Nº eventos	Nº no eventos	Probabilidad a posteriori media	Porcentaje
0.95 - 1.00	0	0	.	0
0.90 - 0.95	0	0	.	0
0.85 - 0.90	1	0	0.85539	0.002
0.80 - 0.85	18	7	0.82252	0.05
0.75 - 0.80	28	13	0.77159	0.082
0.70 - 0.75	57	50	0.71939	0.214
0.65 - 0.70	82	69	0.67388	0.302
0.60 - 0.65	113	94	0.62164	0.414
0.55 - 0.60	395	309	0.57088	1.408
0.50 - 0.55	746	680	0.5223	2.852
0.45 - 0.50	824	859	0.47489	3.366
0.40 - 0.45	656	824	0.42586	2.96
0.35 - 0.40	421	743	0.37529	2.328
0.30 - 0.35	442	914	0.32387	2.712

0.25 - 0.30	706	1697	0.27205	4.806
0.20 - 0.25	1264	4315	0.2218	11.158
0.15 - 0.20	1652	7745	0.17435	18.794
0.10 - 0.15	1097	7578	0.12577	17.35
0.05 - 0.10	740	10220	0.0737	21.92
0.00 - 0.05	171	4470	0.03705	9.282

El resultado de la regresión logística es una probabilidad. Si dicha probabilidad es mayor que la tasa a priori (es decir, mayor que 0,19), se dirá que el cliente se fuga. (Y=1)

Con esta premisa, se marcan en verde los clientes bien clasificados y en rojo los mal clasificados. La matriz de clasificación será por tanto:

	Real 0	Real 1
Estimado 0	30.013 (74%)	3.660 (39%)
Estimado 1	10.574 (26%)	5.753 (61%)
Total	40.587 (100%)	9.413 (100%)

Es decir:

Verdaderos positivos:	61%
Falsos positivos:	26%
Verdaderos negativos:	74%
Falsos negativos:	39%
Estimado correctamente:	72%

- **Otros estadísticos de ajuste:**

  Los siguientes estadísticos por sí solos no nos dicen nada. Solamente si los comparamos con el resto de modelos que se estimarán más adelante:

  – AIC

Entrenamiento
42392.11

– Suma de cuadrados del error (SSE):

Entrenamiento
13.323

– Error medio (ASE):

Entrenamiento
0.1332

➢ ***Paso 5: Entrenamiento del modelo (con interacciones)***

Una vez que tenemos los efectos principales que han resultado significativos, el siguiente paso es incluir como variables candidatas las interacciones de nivel 2 entre ellas (no vamos a considerar interacciones de mayor grado para no incluir demasiada complejidad en el modelo, debido al elevado número de variables con el que se cuenta).
No se van a incluir todas las variables en el modelo sin más, sino que se va a aplicar selección de variables con el método stepwise (con nivel de significación de entrada igual a 0,05 y de permanencia igual a 0,03).

Los resultados obtenidos son:

Paso	Variable que entra	Variable que sale	Grados de libertad	Orden de entrada	Score Chi-cuadrado	Wald Chi-cuadrado	Pr > ChiSq
1	C35		1	1	3898.6403		<.0001
2	C46*C10		1	2	837.8436		<.0001
3	C68*C7		9	3	783.6931		<.0001
4	C4*C37		1	4	310.6891		<.0001
5	C54*C69		4	5	180.4408		<.0001
6	C50*C70		1	6	150.8192		<.0001
7	C3*C44		1	7	143.6792		<.0001
8	C53*C46		2	8	127.2166		<.0001
9	C3*C75		2	9	92.6159		<.0001
10	C29*C37		1	10	71.2551		<.0001
11	C16*C75		2	11	74.757		<.0001
12	C36*C46		1	12	120.5101		<.0001
13	C54*C10		1	13	65.6212		<.0001
14	C53*C5		2	14	60.3631		<.0001
15	C11		2	15	48.8083		<.0001
16		C46*C10	1	14		3.5454	0.0597
17	C3*C35		1	15	28.6576		<.0001
18	C18		1	16	28.2983		<.0001

19	C48*C60		1	17	23.5282		<.0001
20	C71		1	18	24.078		<.0001
21	C53*C7		6	19	35.0459		<.0001
22	C53*C10		2	20	23.0112		<.0001
23	C45*C70		1	21	17.8428		<.0001
24	C34*C44		1	22	15.8614		<.0001
25		C3*C44	1	21		0.3314	0.5649
26	C14*C46		1	22	15.5546		<.0001
27	C2*C37		1	23	16.1937		<.0001
28	C23*C54		1	24	12.8422		0.0003
29	C45*C40		1	25	12.9885		0.0003
30	C38*C46		1	26	12.433		0.0004
31	C42*C60		1	27	12.4058		0.0004
32	C67*C35		1	28	12.393		0.0004
33		C53*C46	2	27		5.4608	0.0652
34	C26*C10		1	28	12.6881		0.0004
35	C50*C5		1	29	11.6668		0.0006
36	C42*C46		1	30	10.1829		0.0014
37	C42*C36		1	31	12.2652		0.0005
38	C18*C67		1	32	9.6795		0.0019
39	C13*C55		1	33	10.5837		0.0011
40	C7*C11		6	34	19.7588		0.0031
41	C38*C11		2	35	11.1882		0.0037
42	C24*C70		1	36	8.2281		0.0041
43	C70*C7		3	37	13.6727		0.0034
44	C49*C46		1	38	8.4221		0.0037
45	C18*C36		1	39	7.6353		0.0057
46	C53*C70		2	40	10.0852		0.0065
47	C14*C53		2	41	10.5408		0.0051
48	C42*C45		1	42	7.7531		0.0054
49	C58*C59		1	43	7.5972		0.0058
50	C29*C45		1	44	7.5173		0.0061
51	C66*C38		1	45	6.7267		0.0095
52		C29*C37	1	44		1.4428	0.2297
53	C66*C70		1	45	7.7536		0.0054
54	C36*C37		1	46	7.1464		0.0075
55	C42*C37		1	47	8.1188		0.0044
56	C42		1	48	8.1006		0.0044
57		C42*C36	1	47		1.934	0.1643
58	C13*C15		1	48	7.1929		0.0073
59	C10*C11		2	49	9.1156		0.0105
60	C5*C7		3	50	10.4946		0.0148
61	C18*C42		1	51	5.7004		0.017
62	C50*C44		1	52	6.4963		0.0108

63	C55*C58		1	53	5.7282		0.0167
64	C2*C60		1	54	8.8083		0.003
65	C48*C46		1	55	5.7891		0.0161
66	C48*C34		1	56	6.6231		0.0101
67		C48*C60	1	55		3.4634	0.0627
68		C55*C58	1	54		3.0836	0.0791
69	C2*C58		1	55	7.4357		0.0064
70	C30*C48		1	56	7.2298		0.0072
71	C37*C11		2	57	7.9262		0.019
72	C25*C38		1	58	5.2093		0.0225
73	C16*C60		1	59	4.9826		0.0256
74	C37*C46		1	60	4.7381		0.0295
75		C36*C37	1	59		3.4673	0.0626
76	C37		1	60	5.3865		0.0203
77	C38*C71		1	61	4.5469		0.033
78	C65*C11		2	62	6.5827		0.0372
79	C14*C5		1	63	4.1955		0.0405
80	C18*C26		1	64	4.1776		0.041
81	C66*C44		1	65	4.6702		0.0307
82		C66*C70	1	64		3.1011	0.0782
83	C25*C35		1	65	4.0304		0.0447
84	C30*C35		1	66	9.5342		0.002
85	C58*C71		1	67	3.9852		0.0459
86	C24*C54		1	68	3.9949		0.0456
87	C23*C24		1	69	4.591		0.0321
88	C55*C60		1	70	3.9729		0.0462
89		C16*C60	1	69		3.5783	0.0585
90	C24*C10		1	70	4.8493		0.0277

Esta tabla muestra el resumen de cómo las variables han ido entrando y saliendo del modelo. Así por ejemplo, en el paso2 se puede observar que la variable C46*C10 entra en el modelo, pero en el paso 16 sale debido a que ha dejado de ser significativa al entrar en el modelo las variables C68*C7, C4*C37, C54*C69, C50*C70, C3*C44, C53*C46, C3*C75, C29*C37, C16*C75, C36*C46 , C54*C10, C53*C5 y C11.

Test Ratio de verosimilitud para la hipótesis nula global: Beta=0
(Likelihood Ratio Test for Global Null Hypothesis: BETA=0)

-2 Log Verosimilitud				
Término independiente solamente	Término independiente y variables	Ratio de verosilimilitud chi-cuadrado	Grados de libertad	Pr > Chi
48369.009	41547.689	6821.3195	106	<.0001

Vemos que el p-valor es significativo, lo cual quiere decir, que el modelo obtenido tiene mejor ajuste que el modelo que está formado solamente por el término independiente.

Tipo 3 Análisis de los efectos
(TipeIII Analysis effects)

Effect	DF	Wald Chi-Square	Pr > ChiSq
C18	1	14.7826	0.0001
C42	1	18.2342	<.0001
C35	1	31.4069	<.0001
C37	1	4.5376	0.0332
C71	1	6.8259	0.009
C11	2	7.7835	0.0204
C2*C58	1	9.203	0.0024
C2*C60	1	19.744	<.0001
C2*C37	1	7.6615	0.0056
C13*C15	1	8.2198	0.0041
C13*C55	1	15.5665	<.0001
C14*C53	2	11.9437	0.0025
C14*C5	1	4.5929	0.0321
C14*C46	1	56.7943	<.0001
C16*C75	2	55.6333	<.0001
C18*C26	1	5.2492	0.022
C18*C42	1	5.9296	0.0149
C18*C67	1	11.5465	0.0007
C18*C36	1	10.5901	0.0011
C23*C24	1	4.8982	0.0269
C23*C54	1	15.6446	<.0001
C24*C54	1	8.9992	0.0027
C24*C70	1	6.6131	0.0101
C24*C10	1	4.8489	0.0277
C25*C35	1	10.4936	0.0012
C25*C38	1	8.5019	0.0035
C26*C10	1	5.4585	0.0195
C29*C45	1	12.6225	0.0004
C3*C35	1	32.74	<.0001

C3*C75	2	23.9556	<.0001
C30*C48	1	9.0928	0.0026
C30*C35	1	10.0968	0.0015
C42*C45	1	9.041	0.0026
C42*C60	1	6.404	0.0114
C42*C37	1	13.013	0.0003
C42*C46	1	15.8428	<.0001
C45*C40	1	18.6345	<.0001
C45*C70	1	21.6664	<.0001
C48*C34	1	20.3475	<.0001
C48*C46	1	7.5895	0.0059
C49*C46	1	9.6801	0.0019
C50*C5	1	16.364	<.0001
C50*C44	1	6.8087	0.0091
C50*C70	1	82.4794	<.0001
C54*C69	4	124.2175	<.0001
C54*C10	1	52.0153	<.0001
C53*C5	2	8.3738	0.0152
C53*C70	2	15.9413	0.0003
C53*C7	6	40.1541	<.0001
C53*C10	2	19.4748	<.0001
C55*C60	1	6.1028	0.0135
C4*C37	1	82.4986	<.0001
C58*C59	1	7.443	0.0064
C58*C71	1	4.1671	0.0412
C65*C11	2	8.3207	0.0156
C66*C38	1	21.544	<.0001
C66*C44	1	7.7434	0.0054
C68*C7	9	72.1326	<.0001
C67*C35	1	22.8447	<.0001
C34*C44	1	31.8356	<.0001
C5*C7	3	9.4305	0.0241
C36*C46	1	6.756	0.0093
C37*C46	1	9.209	0.0024
C37*C11	2	6.7005	0.0351
C38*C46	1	9.779	0.0018

C38*C71	1	8.3584	0.0038
C38*C11	2	11.484	0.0032
C70*C7	3	16.4133	0.0009
C7*C11	6	20.0042	0.0028
C10*C11	2	12.6302	0.0018

Se puede observar que todas las variables por si solas son significativas respecto la variable objetivo, es decir, sin tener en cuenta el resto de variables que han entrado en el modelo

Análisis de estimación de máxima verosimilitud
(Analysis of Maximum Likelihood Estimates)

Parameter	Categoría variable1	Categoría variable2	DF	Estimate	Standard Error	Wald Chi-Square	Pr > ChiSq	Exp(Est)
Intercept			1	1.0413	0.6998	2.21	0.1367	2.833
C18	1		1	0.409	0.1064	14.78	0.0001	1.505
C42	1		1	0.6909	0.1618	18.23	<.0001	1.996
C35	0		1	-0.5045	0.09	31.41	<.0001	0.604
C37	NO		1	-0.3757	0.1764	4.54	0.0332	0.687
C71	HOMBRE		1	-0.0388	0.0148	6.83	0.009	0.962
C11	1		1	0.4843	0.1869	6.72	0.0096	1.623
C11	2		1	-0.221	0.2023	1.19	0.2747	0.802
C2*C58	1	1	1	0.1233	0.0406	9.2	0.0024	1.131
C2*C60	1	1	1	-0.1831	0.0412	19.74	<.0001	0.833
C2*C37	1	NO	1	-0.0795	0.0287	7.66	0.0056	0.924
C13*C15	1	1	1	-0.0704	0.0246	8.22	0.0041	0.932
C13*C55	1	1	1	0.0918	0.0233	15.57	<.0001	1.096
C14*C53	1	1	1	-0.0641	0.0399	2.58	0.1086	0.938
C14*C53	1	2	1	0.118	0.0379	9.7	0.0018	1.125
C14*C5	1	1	1	0.055	0.0257	4.59	0.0321	1.057
C14*C46	1	0	1	0.1851	0.0246	56.79	<.0001	1.203
C16*C75	1	ALTA	1	-0.3888	0.0669	33.78	<.0001	0.678
C16*C75	1	BAJA	1	0.1361	0.0395	11.88	0.0006	1.146
C18*C26	1	1	1	-0.0365	0.016	5.25	0.022	0.964
C18*C42	1	1	1	-0.0353	0.0145	5.93	0.0149	0.965
C18*C67	1	1	1	0.0529	0.0156	11.55	0.0007	1.054
C18*C36	1	NO	1	-0.3413	0.1049	10.59	0.0011	0.711
C23*C24	1	1	1	-0.0427	0.0193	4.9	0.0269	0.958
C23*C54	1	1	1	0.0732	0.0185	15.64	<.0001	1.076
C24*C54	1	1	1	-0.0602	0.0201	9	0.0027	0.942
C24*C70	1	0	1	0.041	0.0159	6.61	0.0101	1.042

C24*C10	1	1	1	0.0382	0.0174	4.85	0.0277	1.039
C25*C35	1	0	1	-0.0651	0.0201	10.49	0.0012	0.937
C25*C38	1	0	1	-0.0556	0.0191	8.5	0.0035	0.946
C26*C10	1	1	1	0.0389	0.0167	5.46	0.0195	1.04
C29*C45	1	1	1	0.0783	0.022	12.62	0.0004	1.081
C3*C35	1	0	1	0.4907	0.0858	32.74	<.0001	1.633
C3*C75	1	ALTA	1	-0.0526	0.1573	0.11	0.7379	0.949
C3*C75	1	BAJA	1	0.1055	0.0808	1.7	0.1918	1.111
C30*C48	1	1	1	-0.044	0.0146	9.09	0.0026	0.957
C30*C35	1	0	1	0.0685	0.0216	10.1	0.0015	1.071
C42*C45	1	1	1	-0.0626	0.0208	9.04	0.0026	0.939
C42*C60	1	1	1	-0.0356	0.0141	6.4	0.0114	0.965
C42*C37	1	NO	1	0.5775	0.1601	13.01	0.0003	1.782
C42*C46	1	0	1	0.0546	0.0137	15.84	<.0001	1.056
C45*C40	1	0	1	-0.0909	0.0211	18.63	<.0001	0.913
C45*C70	1	0	1	-0.0907	0.0195	21.67	<.0001	0.913
C48*C34	1	NO	1	-0.0683	0.0152	20.35	<.0001	0.934
C48*C46	1	0	1	0.04	0.0145	7.59	0.0059	1.041
C49*C46	1	0	1	0.0647	0.0208	9.68	0.0019	1.067
C50*C5	1	1	1	-0.0723	0.0179	16.36	<.0001	0.93
C50*C44	1	ALTA	1	-0.054	0.0207	6.81	0.0091	0.947
C50*C70	1	0	1	-0.1799	0.0198	82.48	<.0001	0.835
C54*C69	1	1	1	-2.8871	0.9775	8.72	0.0031	0.056
C54*C69	1	2	1	-2.9115	0.74	15.48	<.0001	0.054
C54*C69	1	3	1	2.3985	0.4142	33.53	<.0001	11.007
C54*C69	1	4	1	1.9504	0.6533	8.91	0.0028	7.031
C54*C10	1	1	1	0.134	0.0186	52.02	<.0001	1.143
C53*C5	1	1	1	0.0694	0.0254	7.49	0.0062	1.072
C53*C5	2	1	1	-0.0111	0.0236	0.22	0.6378	0.989
C53*C70	1	0	1	0.0915	0.0246	13.79	0.0002	1.096
C53*C70	2	0	1	-0.035	0.0262	1.78	0.1824	0.966
C53*C7	1	1	1	-0.0084	0.0533	0.02	0.8748	0.992
C53*C7	1	2	1	-0.0191	0.0518	0.14	0.7121	0.981
C53*C7	1	3	1	-0.1524	0.0346	19.35	<.0001	0.859
C53*C7	2	1	1	0.0189	0.0579	0.11	0.7437	1.019
C53*C7	2	2	1	0.0583	0.0572	1.04	0.3083	1.06
C53*C7	2	3	1	0.0929	0.0357	6.76	0.0093	1.097
C53*C10	1	1	1	0.0273	0.0284	0.93	0.3359	1.028
C53*C10	2	1	1	-0.1518	0.0372	16.63	<.0001	0.859
C55*C60	1	1	1	0.0449	0.0182	6.1	0.0135	1.046
C4*C37	1	NO	1	-0.2783	0.0306	82.5	<.0001	0.757
C58*C59	1	1	1	-0.1701	0.0624	7.44	0.0064	0.844
C58*C71	1	HOMBRE	1	0.0338	0.0165	4.17	0.0412	1.034
C65*C11	1	1	1	0.087	0.0303	8.27	0.004	1.091

C65*C11	1	2	1	-0.0546	0.0288	3.6	0.0578	0.947
C66*C38	1	0	1	0.0814	0.0175	21.54	<.0001	1.085
C66*C44	1	ALTA	1	0.0568	0.0204	7.74	0.0054	1.058
C68*C7	1	1	1	0.1569	0.3349	0.22	0.6395	1.17
C68*C7	1	2	1	0.1802	0.3425	0.28	0.5987	1.197
C68*C7	1	3	1	-1.3186	0.3598	13.43	0.0002	0.268
C68*C7	2	1	1	0.0576	0.1654	0.12	0.7277	1.059
C68*C7	2	2	1	-0.1746	0.3107	0.32	0.5742	0.84
C68*C7	2	3	1	0.4215	0.1691	6.21	0.0127	1.524
C68*C7	3	1	1	-0.1876	0.3348	0.31	0.5754	0.829
C68*C7	3	2	1	0.2339	0.3242	0.52	0.4706	1.264
C68*C7	3	3	1	0.8979	0.3247	7.65	0.0057	2.455
C67*C35	1	0	1	-0.0808	0.0169	22.84	<.0001	0.922
C34*C44	NO	ALTA	1	0.1075	0.0191	31.84	<.0001	1.114
C5*C7	1	1	1	0.0307	0.0333	0.85	0.3559	1.031
C5*C7	1	2	1	0.0767	0.0355	4.67	0.0308	1.08
C5*C7	1	3	1	-0.0283	0.0269	1.11	0.2931	0.972
C36*C46	NO	0	1	0.2134	0.0821	6.76	0.0093	1.238
C37*C46	NO	0	1	-0.2482	0.0818	9.21	0.0024	0.78
C37*C11	NO	1	1	0.4171	0.1852	5.07	0.0243	1.518
C37*C11	NO	2	1	-0.1371	0.2007	0.47	0.4945	0.872
C38*C46	0	0	1	-0.0572	0.0183	9.78	0.0018	0.944
C38*C71	0	HOMBRE	1	-0.045	0.0156	8.36	0.0038	0.956
C38*C11	0	1	1	0.0439	0.019	5.31	0.0212	1.045
C38*C11	0	2	1	0.0166	0.0195	0.72	0.3962	1.017
C70*C7	0	1	1	0.0849	0.0436	3.79	0.0517	1.089
C70*C7	0	2	1	0.0228	0.0403	0.32	0.5711	1.023
C70*C7	0	3	1	-0.086	0.0289	8.82	0.003	0.918
C7*C11	1	1	1	-0.011	0.0366	0.09	0.7644	0.989
C7*C11	1	2	1	0.1041	0.0386	7.26	0.0071	1.11
C7*C11	2	1	1	0.0626	0.039	2.58	0.1081	1.065
C7*C11	2	2	1	-0.0759	0.0413	3.37	0.0663	0.927
C7*C11	3	1	1	-0.0625	0.0285	4.79	0.0286	0.939
C7*C11	3	2	1	-0.019	0.0294	0.42	0.5192	0.981
C10*C11	1	1	1	-0.0709	0.0204	12.03	0.0005	0.932
C10*C11	1	2	1	0.0264	0.0209	1.6	0.2053	1.027

Se muestra el resultado de la estimación de cada uno de los parámetros y su significatividad teniendo en cuenta el resto de variables que forman parte del modelo.

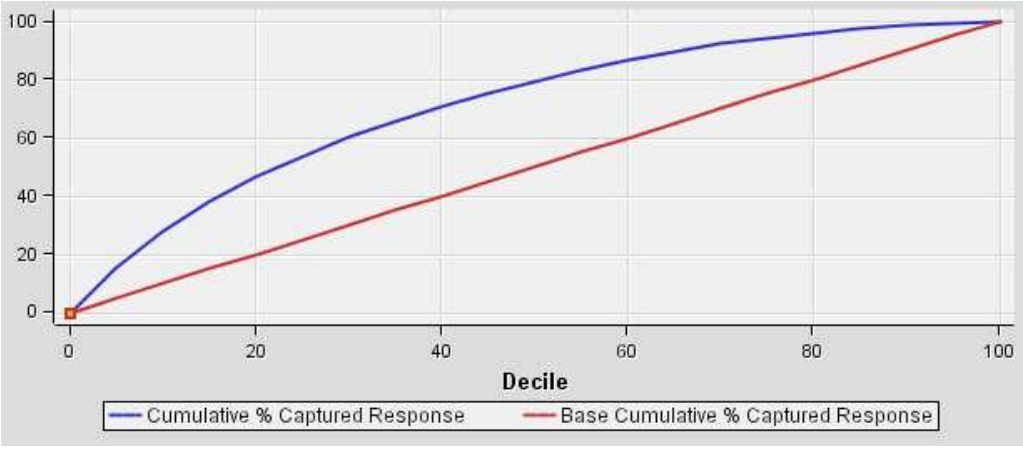
Si tramificamos la probabilidad estimada en intervalos de 0.05 de amplitud y se observan el número de aciertos y no aciertos que realmente se han obtenido:

Distribución del score
(Assessment Score Distribution)

Rango de probabilidad estimada	Nº eventos	Nº no eventos	Probabilidad a posteriori media	Porcentaje
0.95 - 1.00	0	0	.	0
0.90 - 0.95	0	0	.	0
0.85 - 0.90	1	0	0.88347	0.002
0.80 - 0.85	3	2	0.81015	0.01
0.75 - 0.80	14	10	0.77291	0.048
0.70 - 0.75	55	27	0.72142	0.164
0.65 - 0.70	107	49	0.67115	0.312
0.60 - 0.65	230	153	0.62179	0.766
0.55 - 0.60	492	407	0.57039	1.798
0.50 - 0.55	765	681	0.52354	2.892
0.45 - 0.50	791	816	0.47465	3.214
0.40 - 0.45	613	844	0.42515	2.914
0.35 - 0.40	519	930	0.37491	2.898
0.30 - 0.35	614	1179	0.32352	3.586
0.25 - 0.30	765	2094	0.27235	5.718
0.20 - 0.25	1111	3702	0.22258	9.626
0.15 - 0.20	1385	6759	0.17379	16.288
0.10 - 0.15	1012	6836	0.12574	15.696
0.05 - 0.10	672	8960	0.07202	19.264
0.00 - 0.05	264	7138	0.036	14.804

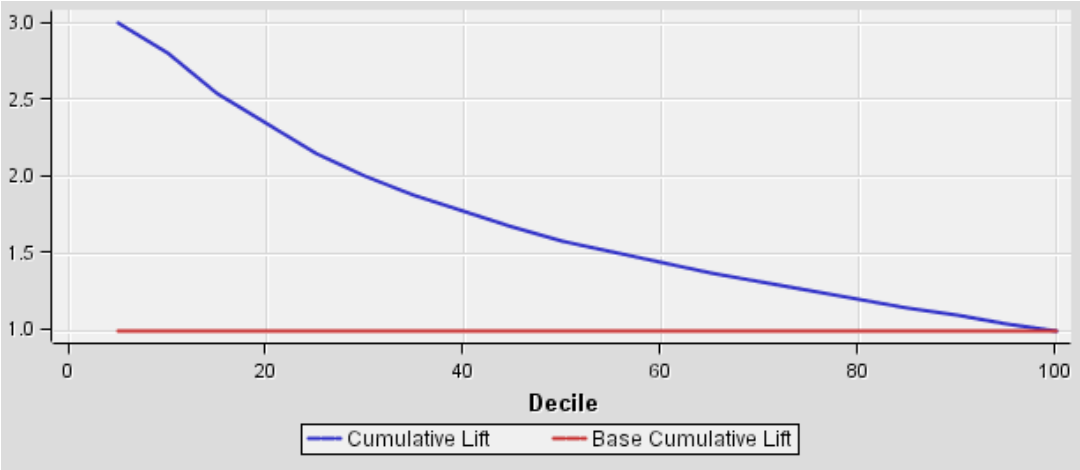
Con esta tabla se calculará la matriz de clasificación

- **Porcentaje de éxitos capturados:**



– Con un 20% de los clientes predecimos el 47% de las fugas
– Con un 40% de los clientes predecimos el 71% de las fugas

- **Mejora acumulada:**



– Con un 20% de los clientes se obtiene una ganancia de x2,3
– Con un 40% de los clientes se obtiene una ganancia de x1,8

- **Curva ROC:**

– Para conseguir un 21% de verdaderos positivos se tendrá un 4,0% de falsos positivos
– Para conseguir un 40% de verdaderos positivos se tendrá un 10,7% de falsos positivos

- **Matriz de clasificación:**

  La probabilidad estimada se resume en la siguiente tabla:

Rango de probabilidad estimada	Nº eventos	Nº no eventos	Probabilidad a posteriori media	Porcentaje
0.95 - 1.00	0	0	.	0
0.90 - 0.95	0	0	.	0
0.85 - 0.90	1	0	0.85539	0.002
0.80 - 0.85	3	2	0.82252	0.05
0.75 - 0.80	14	10	0.77159	0.082
0.70 - 0.75	55	27	0.71939	0.214
0.65 - 0.70	107	49	0.67388	0.302
0.60 - 0.65	230	153	0.62164	0.414
0.55 - 0.60	492	407	0.57088	1.408
0.50 - 0.55	765	681	0.5223	2.852
0.45 - 0.50	791	816	0.47489	3.366
0.40 - 0.45	613	844	0.42586	2.96
0.35 - 0.40	519	930	0.37529	2.328
0.30 - 0.35	614	1179	0.32387	2.712
0.25 - 0.30	765	2094	0.27205	4.806
0.20 - 0.25	1111	3702	0.2218	11.158
0.15 - 0.20	1385	6759	0.17435	18.794
0.10 - 0.15	1012	6836	0.12577	17.35
0.05 - 0.10	672	8960	0.0737	21.92
0.00 - 0.05	264	7138	0.03705	9.282

  La matriz de clasificación será por tanto:

	Real 0	**Real 1**
Estimado 0	29.693 (73%)	3.333 (35%)
Estimado 1	10.894 (27%)	6.080 (65%)
Total	40.587 (100%)	9.413 (100%)

  Es decir:

Verdaderos positivos:	65%
Falsos positivos:	27%
Verdaderos negativos:	73%
Falsos negativos:	35%
Estimado correctamente:	72%

- **Otros estadísticos de ajuste:**

  – AIC

Entrenamiento
41761.69

  – Suma de cuadrados del error (SSE):

Entrenamiento
13089.31

  – Error medio (ASE):

Entrenamiento
0.1300

➢ ***Paso 6: Entrenamiento del modelo jerárquico (sin selección de variables)***

Test Ratio de verosimilitud para la hipótesis nula global: Beta=0
(Likelihood Ratio Test for Global Null Hypothesis: BETA=0)

-2 Log Verosimilitud				
Sin variables	Con variables	Ratio de verosilimilitud chi-cuadrado	Grados de libertad	Pr > Chi
69314.718	41677.880	27636.8384	136	<.0001

En esta tabla vemos que el p-valor es significativo, lo cual quiere decir, que el modelo obtenido tiene mejor ajuste que el modelo que no hacer modelo (en este caso no se compara con el término independiente ya que vimos que no resultaba significativo)

Tipo 3 Análisis de los efectos
(TipeIII Analysis effects)

Effect	DF	Wald Chi-Square	Pr > ChiSq
C2	1	3.8922	0.0485
C13	1	0.0566	0.8119
C14	1	1.8864	0.1696
C15	1	0.2229	0.6368
C16	1	0.0044	0.9469
C18	1	8.7319	0.0031
C23	1	2.0026	0.157
C24	1	3.0708	0.0797
C25	1	1.9264	0.1652
C26	1	0.0861	0.7693
C29	1	0.0225	0.8808
C3	1	0.8168	0.3661
C30	1	2.9204	0.0875
C42	1	13.3528	0.0003
C45	1	0.1287	0.7198
C48	1	0.6339	0.4259
C49	1	0.1986	0.6559
C50	1	0.0211	0.8845
C53	2	6.4063	0.0406
C55	1	6.9618	0.0083
C4	1	2.9038	0.0884
C58	1	1.2906	0.2559
C59	1	0.4172	0.5183
C60	1	0.6101	0.4347
C65	1	0.401	0.5266
C66	1	0.1106	0.7395
C68	3	4.1785	0.2428
C67	1	5.0865	0.0241
C34	1	1.1966	0.274
C5	1	1.9631	0.1612
C36	1	0.0037	0.9518
C37	1	1.975	0.1599
C38	1	0.1035	0.7477
C40	1	0.0762	0.7825
C44	1	0.4636	0.496
C46	1	2.7177	0.0992
C70	1	0.3605	0.5482
C71	1	5.552	0.0185
C75	2	0.4889	0.7831
C7	3	6.5015	0.0896
C10	1	0.0137	0.9067
C11	2	7.5743	0.0227
C2*C58	1	4.6107	0.0318
C2*C60	1	11.4254	0.0007
C2*C37	1	4.6364	0.0313

C13*C15	1	6.3361	0.0118
C13*C55	1	9.7705	0.0018
C14*C53	2	6.9198	0.0314
C14*C5	1	5.1398	0.0234
C14*C46	1	80.9483	<.0001
C16*C75	2	13.1263	0.0014
C18*C26	1	4.9845	0.0256
C18*C42	1	6.0682	0.0138
C18*C67	1	12.1219	0.0005
C18*C36	1	6.3218	0.0119
C23*C24	1	5.0213	0.025
C24*C70	1	7.5183	0.0061
C24*C10	1	4.3781	0.0364
C25*C38	1	7.2408	0.0071
C26*C10	1	4.5919	0.0321
C29*C45	1	6.3264	0.0119
C3*C75	2	6.8809	0.032
C30*C48	1	6.247	0.0124
C42*C45	1	6.6364	0.01
C42*C60	1	5.5472	0.0185
C42*C37	1	10.036	0.0015
C42*C46	1	15.2844	<.0001
C45*C40	1	9.69	0.0019
C45*C70	1	10.2247	0.0014
C48*C34	1	3.007	0.0829
C48*C46	1	7.6581	0.0057
C49*C46	1	29.456	<.0001
C50*C5	1	8.3039	0.004
C50*C44	1	8.3592	0.0038
C50*C70	1	83.7869	<.0001
C53*C5	2	13.5062	0.0012
C53*C70	2	20.3417	<.0001
C53*C7	6	31.0563	<.0001
C53*C10	2	41.5748	<.0001
C55*C60	1	8.204	0.0042
C4*C37	1	0.7476	0.3872
C58*C59	1	5.9711	0.0145
C58*C71	1	4.2234	0.0399
C65*C11	2	6.9005	0.0317
C66*C38	1	16.309	<.0001
C66*C44	1	3.0196	0.0823
C68*C7	9	38.1231	<.0001
C34*C44	1	0.5357	0.4642
C5*C7	3	13.7779	0.0032
C36*C46	1	2.1737	0.1404
C37*C11	2	6.7221	0.0347
C38*C46	1	18.5218	<.0001
C38*C71	1	8.6464	0.0033
C38*C11	2	12.3966	0.002
C70*C7	3	16.6547	0.0008
C7*C11	6	19.1263	0.004
C10*C11	2	9.2978	0.0096

Se puede observar que no todas las variables por si solas son significativas respecto la variable objetivo

Análisis de estimación de máxima verosimilitud
(Analysis of Maximum Likelihood Estimates)

Parameter	Categoría variable1	Categoría variable2	DF	Estimate	Standard Error	Wald Chi-Square	Pr > ChiSq	Exp(Est)
C2	1		1	-0.6277	0.3181	3.89	0.0485	0.534
C13	1		1	-0.00862	0.0362	0.06	0.8119	0.991
C14	1		1	0.0869	0.0633	1.89	0.1696	1.091
C15	1		1	0.0666	0.141	0.22	0.6368	1.069
C16	1		1	0.0147	0.2206	0	0.9469	1.015
C18	1		1	0.3964	0.1341	8.73	0.0031	1.486
C23	1		1	0.0444	0.0314	2	0.157	1.045
C24	1		1	-0.0478	0.0273	3.07	0.0797	0.953
C25	1		1	-0.0405	0.0292	1.93	0.1652	0.96
C26	1		1	0.00745	0.0254	0.09	0.7693	1.007
C29	1		1	0.0082	0.0546	0.02	0.8808	1.008
C3	1		1	0.9262	1.0248	0.82	0.3661	2.525
C30	1		1	0.0474	0.0278	2.92	0.0875	1.049
C42	1		1	0.7169	0.1962	13.35	0.0003	2.048
C45	1		1	0.0101	0.0281	0.13	0.7198	1.01
C48	1		1	0.0231	0.029	0.63	0.4259	1.023
C49	1		1	-0.066	0.1482	0.2	0.6559	0.936
C50	1		1	0.0169	0.1166	0.02	0.8845	1.017
C53	1		1	-0.1147	0.1006	1.3	0.2545	0.892
C53	2		1	-0.0884	0.0613	2.08	0.1492	0.915
C55	1		1	-0.1384	0.0524	6.96	0.0083	0.871
C4	1		1	0.5672	0.3329	2.9	0.0884	1.763
C58	1		1	0.0785	0.0691	1.29	0.2559	1.082
C59	1		1	-0.0513	0.0795	0.42	0.5183	0.95
C60	1		1	-0.0461	0.0591	0.61	0.4347	0.955
C65	1		1	-0.0433	0.0684	0.4	0.5266	0.958
C66	1		1	0.0235	0.0706	0.11	0.7395	1.024
C68	1		1	-3.5529	3.5591	1	0.3181	0.029
C68	2		1	8.3116	4.7888	3.01	0.0826	999
C68	3		1	-3.0284	1.5901	3.63	0.0568	0.048
C67	1		1	-0.059	0.0262	5.09	0.0241	0.943
C34	NO		1	-0.071	0.0649	1.2	0.274	0.931
C5	1		1	-0.0604	0.0431	1.96	0.1612	0.941
C36	NO		1	0.0108	0.1791	0	0.9518	1.011
C37	NO		1	-0.274	0.195	1.97	0.1599	0.76
C38	0		1	-0.0164	0.051	0.1	0.7477	0.984
C40	0		1	0.011	0.0398	0.08	0.7825	1.011
C44	ALTA		1	-0.0452	0.0663	0.46	0.496	0.956
C46	0		1	0.2457	0.149	2.72	0.0992	1.278

C70	0		1	-0.0268	0.0447	0.36	0.5482	0.974
C71	HOMBRE		1	-0.0351	0.0149	5.55	0.0185	0.966
C75	ALTA		1	-0.9473	2.0221	0.22	0.6395	0.388
C75	BAJA		1	0.4867	1.0115	0.23	0.6304	1.627
C7	1		1	-2.3505	1.0529	4.98	0.0256	0.095
C7	2		1	8.7043	4.7922	3.3	0.0693	999
C7	3		1	-2.0319	1.0867	3.5	0.0615	0.131
C10	1		1	-0.00497	0.0424	0.01	0.9067	0.995
C11	1		1	0.4836	0.1968	6.04	0.014	1.622
C11	2		1	-0.1798	0.2101	0.73	0.3922	0.835
C2*C58	1	1	1	0.1017	0.0474	4.61	0.0318	1.107
C2*C60	1	1	1	-0.1601	0.0474	11.43	0.0007	0.852
C2*C37	1	NO	1	-0.6825	0.317	4.64	0.0313	0.505
C13*C15	1	1	1	-0.0695	0.0276	6.34	0.0118	0.933
C13*C55	1	1	1	0.1134	0.0363	9.77	0.0018	1.12
C14*C53	1	1	1	-0.1017	0.0933	1.19	0.2758	0.903
C14*C53	1	2	1	0.1553	0.0601	6.66	0.0098	1.168
C14*C5	1	1	1	0.0689	0.0304	5.14	0.0234	1.071
C14*C46	1	0	1	0.2113	0.0235	80.95	<.0001	1.235
C16*C75	1	ALTA	1	-0.3425	0.3781	0.82	0.365	0.71
C16*C75	1	BAJA	1	0.1152	0.1899	0.37	0.5442	1.122
C18*C26	1	1	1	-0.0374	0.0167	4.98	0.0256	0.963
C18*C42	1	1	1	-0.0357	0.0145	6.07	0.0138	0.965
C18*C67	1	1	1	0.0583	0.0168	12.12	0.0005	1.06
C18*C36	1	NO	1	-0.3342	0.1329	6.32	0.0119	0.716
C23*C24	1	1	1	-0.052	0.0232	5.02	0.025	0.949
C24*C70	1	0	1	0.0517	0.0189	7.52	0.0061	1.053
C24*C10	1	1	1	0.0362	0.0173	4.38	0.0364	1.037
C25*C38	1	0	1	-0.0591	0.022	7.24	0.0071	0.943
C26*C10	1	1	1	0.036	0.0168	4.59	0.0321	1.037
C29*C45	1	1	1	0.0786	0.0312	6.33	0.0119	1.082
C3*C75	1	ALTA	1	0.8555	2.0477	0.17	0.6761	2.353
C3*C75	1	BAJA	1	-0.3615	1.0242	0.12	0.7241	0.697
C30*C48	1	1	1	-0.0394	0.0157	6.25	0.0124	0.961
C42*C45	1	1	1	-0.0539	0.0209	6.64	0.01	0.948
C42*C60	1	1	1	-0.0334	0.0142	5.55	0.0185	0.967
C42*C37	1	NO	1	0.6169	0.1947	10.04	0.0015	1.853
C42*C46	1	0	1	0.0533	0.0136	15.28	<.0001	1.055
C45*C40	1	0	1	-0.0975	0.0313	9.69	0.0019	0.907
C45*C70	1	0	1	-0.0901	0.0282	10.22	0.0014	0.914
C48*C34	1	NO	1	-0.0471	0.0272	3.01	0.0829	0.954
C48*C46	1	0	1	0.0408	0.0148	7.66	0.0057	1.042
C49*C46	1	0	1	0.0928	0.0171	29.46	<.0001	1.097
C50*C5	1	1	1	-0.0513	0.0178	8.3	0.004	0.95
C50*C44	1	ALTA	1	-0.0612	0.0212	8.36	0.0038	0.941
C50*C70	1	0	1	-0.183	0.02	83.79	<.0001	0.833

C53*C5	1	1	1	0.0912	0.027	11.42	0.0007	1.096
C53*C5	2	1	1	-0.0109	0.0271	0.16	0.6858	0.989
C53*C70	1	0	1	0.1088	0.0269	16.41	<.0001	1.115
C53*C70	2	0	1	-0.0393	0.0289	1.84	0.1746	0.962
C53*C7	1	1	1	-0.0302	0.0538	0.31	0.5751	0.97
C53*C7	1	2	1	-0.00106	0.0529	0	0.984	0.999
C53*C7	1	3	1	-0.1236	0.0364	11.52	0.0007	0.884
C53*C7	2	1	1	-0.0643	0.0592	1.18	0.2775	0.938
C53*C7	2	2	1	0.0764	0.0597	1.64	0.2009	1.079
C53*C7	2	3	1	0.1133	0.0397	8.16	0.0043	1.12
C53*C10	1	1	1	0.1554	0.0353	19.41	<.0001	1.168
C53*C10	2	1	1	-0.1132	0.0399	8.06	0.0045	0.893
C55*C60	1	1	1	0.0616	0.0215	8.2	0.0042	1.064
C4*C37	1	NO	1	0.2847	0.3293	0.75	0.3872	1.329
C58*C59	1	1	1	-0.1573	0.0644	5.97	0.0145	0.854
C58*C71	1	HOMBRE	1	0.0343	0.0167	4.22	0.0399	1.035
C65*C11	1	1	1	0.0812	0.0313	6.72	0.0095	1.085
C65*C11	1	2	1	-0.0588	0.0287	4.2	0.0404	0.943
C66*C38	1	0	1	0.0791	0.0196	16.31	<.0001	1.082
C66*C44	1	ALTA	1	0.0442	0.0255	3.02	0.0823	1.045
C68*C7	1	1	1	3.5786	3.5592	1.01	0.3147	35.825
C68*C7	1	2	1	-7.5255	3.3772	4.97	0.0259	0.001
C68*C7	1	3	1	2.8134	3.5899	0.61	0.4332	16.667
C68*C7	2	1	1	-8.4254	4.7893	3.09	0.0785	0
C68*C7	2	2	1	23.9254	11.9053	4.04	0.0445	999
C68*C7	2	3	1	-8.4567	4.7815	3.13	0.077	0
C68*C7	3	1	1	2.7378	1.5931	2.95	0.0857	15.454
C68*C7	3	2	1	-7.7483	4.6405	2.79	0.095	0
C68*C7	3	3	1	3.6717	1.6141	5.17	0.0229	39.318
C34*C44	NO	ALTA	1	0.0473	0.0646	0.54	0.4642	1.048
C5*C7	1	1	1	0.0695	0.0388	3.21	0.0732	1.072
C5*C7	1	2	1	0.1076	0.0387	7.75	0.0054	1.114
C5*C7	1	3	1	-0.00297	0.0335	0.01	0.9293	0.997
C36*C46	NO	0	1	0.2186	0.1483	2.17	0.1404	1.244
C37*C11	NO	1	1	0.4199	0.1951	4.63	0.0314	1.522
C37*C11	NO	2	1	-0.1038	0.2086	0.25	0.6187	0.901
C38*C46	0	0	1	-0.0644	0.015	18.52	<.0001	0.938
C38*C71	0	HOMBRE	1	-0.0461	0.0157	8.65	0.0033	0.955
C38*C11	0	1	1	0.0471	0.019	6.13	0.0133	1.048
C38*C11	0	2	1	0.0158	0.0195	0.65	0.4201	1.016
C70*C7	0	1	1	0.1033	0.0464	4.97	0.0259	1.109
C70*C7	0	2	1	0.0339	0.0437	0.6	0.4377	1.035
C70*C7	0	3	1	-0.0757	0.037	4.18	0.0408	0.927
C7*C11	1	1	1	-0.0038	0.0367	0.01	0.9175	0.996
C7*C11	1	2	1	0.0992	0.0386	6.59	0.0102	1.104
C7*C11	2	1	1	0.0596	0.0388	2.36	0.1246	1.061

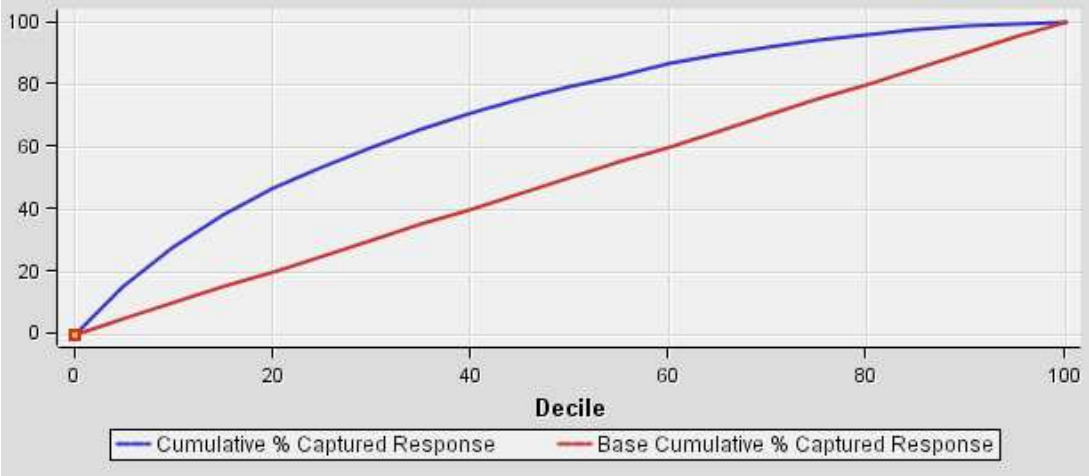
C7*C11	2	2	1	-0.0732	0.0412	3.16	0.0753	0.929
C7*C11	3	1	1	-0.0568	0.0285	3.97	0.0463	0.945
C7*C11	3	2	1	-0.0232	0.0294	0.62	0.43	0.977
C10*C11	1	1	1	-0.06	0.0204	8.67	0.0032	0.942
C10*C11	1	2	1	0.0199	0.0208	0.92	0.337	1.02

Distribución del score
(Assessment Score Distribution)

Rango de probabilidad estimada	Nº eventos	Nº no eventos	Probabilidad a posteriori media	Porcentaje
0.95 - 1.00	0	0	.	0
0.90 - 0.95	0	0	.	0
0.85 - 0.90	1	0	0.85475	0.002
0.80 - 0.85	2	1	0.8086	0.006
0.75 - 0.80	30	16	0.77303	0.092
0.70 - 0.75	78	40	0.72262	0.236
0.65 - 0.70	114	88	0.6736	0.404
0.60 - 0.65	182	129	0.62136	0.622
0.55 - 0.60	433	344	0.57026	1.554
0.50 - 0.55	744	656	0.52324	2.8
0.45 - 0.50	744	787	0.47512	3.062
0.40 - 0.45	691	925	0.42525	3.232
0.35 - 0.40	591	958	0.37504	3.098
0.30 - 0.35	575	1219	0.32442	3.588
0.25 - 0.30	711	1960	0.27281	5.342
0.20 - 0.25	1111	3526	0.22286	9.274
0.15 - 0.20	1418	6775	0.17385	16.386
0.10 - 0.15	1046	7095	0.12568	16.282
0.05 - 0.10	733	9774	0.07205	21.014
0.00 - 0.05	209	6294	0.03709	13.006

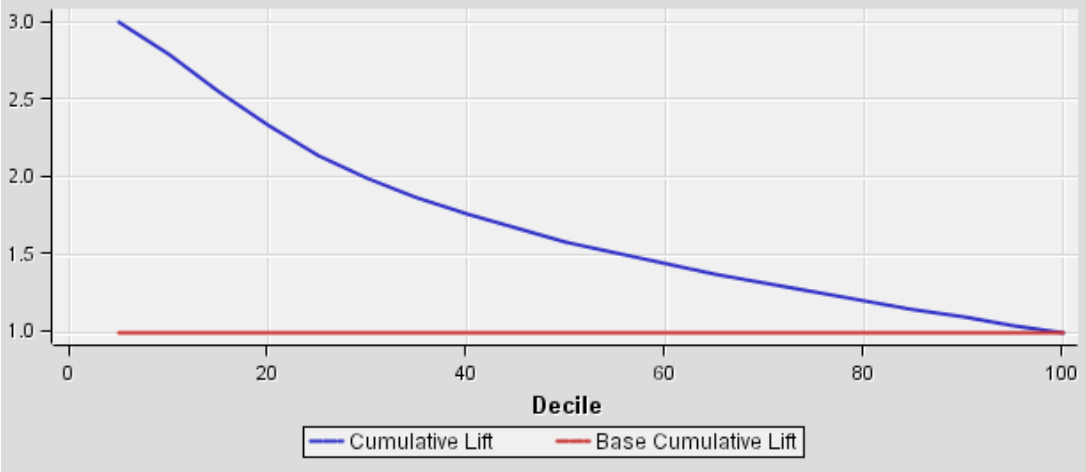
Con esta tabla se calculará la matriz de clasificación

- **Porcentaje de éxitos capturados:**



  – Con un 20% de los clientes predecimos el 47% de las fugas
  – Con un 40% de los clientes predecimos el 71% de las fugas

- **Mejora acumulada:**



  – Con un 20% de los clientes se obtiene una ganancia de x2,3
  – Con un 40% de los clientes se obtiene una ganancia de x1,8

- **Curva ROC:**

  – Para conseguir un 21% de verdaderos positivos se tendrá un 3,8% de falsos positivos
  – Para conseguir un 40% de verdaderos positivos se tendrá un 10,7% de falsos positivos

- **Matriz de clasificación:**
  La probabilidad estimada se resume en la siguiente tabla:

Rango de probabilidad estimada	Nº eventos	Nº no eventos	Probabilidad a posteriori media	Porcentaje
0.95 - 1.00	0	0	.	0
0.90 - 0.95	0	0	.	0
0.85 - 0.90	1	0	0.85539	0.002
0.80 - 0.85	2	1	0.82252	0.05
0.75 - 0.80	30	16	0.77159	0.082
0.70 - 0.75	78	40	0.71939	0.214
0.65 - 0.70	114	88	0.67388	0.302
0.60 - 0.65	182	129	0.62164	0.414
0.55 - 0.60	433	344	0.57088	1.408
0.50 - 0.55	744	656	0.5223	2.852
0.45 - 0.50	744	787	0.47489	3.366
0.40 - 0.45	691	925	0.42586	2.96
0.35 - 0.40	591	958	0.37529	2.328
0.30 - 0.35	575	1219	0.32387	2.712
0.25 - 0.30	711	1960	0.27205	4.806
0.20 - 0.25	1111	3526	0.2218	11.158
0.15 - 0.20	1418	6775	0.17435	18.794

0.10 - 0.15	1046	7095	0.12577	17.35
0.05 - 0.10	733	9774	0.0737	21.92
0.00 - 0.05	209	6294	0.03705	9.282

La matriz de clasificación será por tanto:

	Real 0	Real 1
Estimado 0	29.938 (74%)	3.406 (36%)
Estimado 1	10.649 (26%)	6.007 (64%)
Total	40.587 (100%)	9.413 (100%)

Es decir:

Verdaderos positivos:	64%
Falsos positivos:	26%
Verdaderos negativos:	74%
Falsos negativos:	36%
Estimado correctamente:	72%

- **Otros estadísticos de ajuste:**
  - AIC

Entrenamiento
41949.88

  - Suma de cuadrados del error (SSE):

Entrenamiento
13139.31

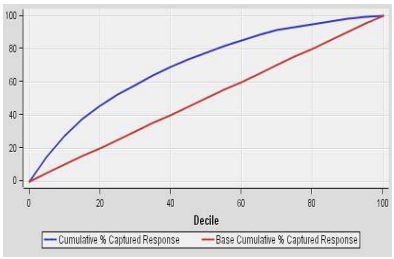
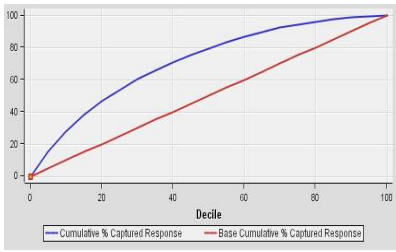
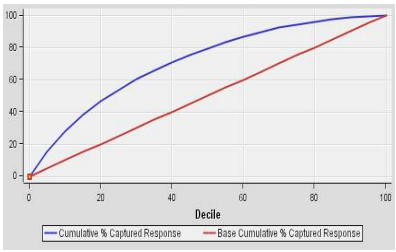
  - Error medio (ASE):

Entrenamiento
0.1313

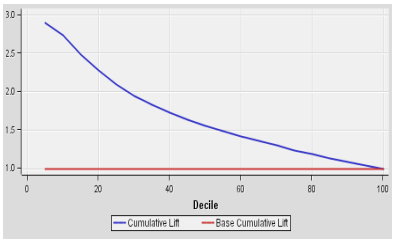
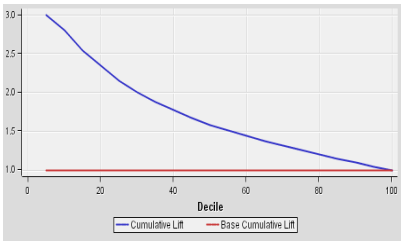
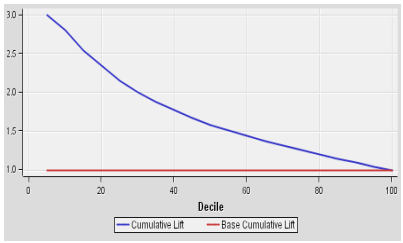
➢ ***Paso 7: Elección del mejor modelo***

Vamos a comparar cada uno de los modelos entre sí para ver con cuál de ellos nos quedamos:

- **Porcentaje de éxitos capturados:**

Modelo1: solo con efectos pples	Modelo2: incluyendo interacciones	Modelo3: principio jerarquico
		
- Con un 20% de los clientes predecimos el 45% de las fugas - Con un 40% de los clientes predecimos el 69% de las fugas	- Con un 20% de los clientes predecimos el 47% de las fugas - Con un 40% de los clientes predecimos el 71% de las fugas	- Con un 20% de los clientes predecimos el 47% de las fugas - Con un 40% de los clientes predecimos el 71% de las fugas

- **Mejora acumulada**

Modelo1: solo con efectos pples	Modelo2: incluyendo interacciones	Modelo3: principio jerarquico
		
- Con un 5% de los clientes se obtiene una ganancia de x2,9 - Con un 20% de los clientes se obtiene una ganancia de x2,3 - Con un 40% de los clientes se obtiene una ganancia de x1,7	- Con un 5% de los clientes se obtiene una ganancia de x3,0 - Con un 20% de los clientes se obtiene una ganancia de x2,3 - Con un 40% de los clientes se obtiene una ganancia de x1,8	- Con un 5% de los clientes se obtiene una ganancia de x3,0 - Con un 20% de los clientes se obtiene una ganancia de x2,3 - Con un 40% de los clientes se obtiene una ganancia de x1,8

- **Matriz de clasificación**

Modelo1: solo con efectos pples		Modelo2: incluyendo interacciones		Modelo3: principio jerarquico	
Verdaderos positivos:	61%	Verdaderos positivos:	65%	Verdaderos positivos:	64%
Falsos positivos:	26%	Falsos positivos:	27%	Falsos positivos:	26%
Verdaderos negativos:	74%	Verdaderos negativos:	73%	Verdaderos negativos:	74%
Falsos negativos:	39%	Falsos negativos:	35%	Falsos negativos:	36%
Estimado correctamente:	72%	Estimado correctamente:	72%	Estimado correctamente:	72%

- **Otros indicadores:**

Modelo1: solo con efectos pples	Modelo2: incluyendo interacciones	Modelo3: principio jerarquico
AIC 42392.11 SSE 13.323 ASE 0.1332	AIC 41761.69 SSE 13089.31 ASE 0.1300	AIC 41949.88 SSE 13139.31 ASE 0.1313

- **Conclusiones:**

  – El Modelo1, que es el que contiene solamente los efectos principales queda descartado, ya que el incluir las interacciones mejora la predicción de la variable dependiente
  – Entre el Modelo2 y el Modelo3, vamos a descartar el Modelo3, ya que el hecho de que se cumpla el principio jerárquico hace que aumente la complejidad del modelo, pero no mejora el poder de predicción.
  – Por lo tanto, el modelo final estimado con regresión logística es el Modelo2

➢ ***Paso 8: Test***

- **Matriz de clasificación:**

Rango de probabilidad estimada	Nº eventos	Nº no eventos	Probabilidad a posteriori media	Porcentaje
0.95 - 1.00			.	.
0.90 - 0.95			.	.
0.85 - 0.90	0	1	0.8452	0.9000
0.80 - 0.85	13	1	0.8048	0.8451
0.75 - 0.80	39	28	0.7512	0.7989
0.70 - 0.75	130	92	0.7000	0.7499
0.65 - 0.70	315	209	0.6501	0.6997
0.60 - 0.65	705	418	0.6000	0.6500
0.55 - 0.60	1136	784	0.5500	0.6000
0.50 - 0.55	1577	1217	0.5000	0.5500
0.45 - 0.50	1904	1604	0.4500	0.5000
0.40 - 0.45	1952	1931	0.4000	0.4500

0.35 - 0.40	1471	1741	0.3500	0.4000
0.30 - 0.35	1365	1895	0.3000	0.3500
0.25 - 0.30	1894	3340	0.2500	0.3000
0.20 - 0.25	2873	6859	0.2000	0.2500
0.15 - 0.20	12144	3540	0.1500	0.2000
0.10 - 0.15	12538	2629	0.1000	0.1500
0.05 - 0.10	15981	1599	0.0500	0.1000
0.00 - 0.05	14726	736	0.0034	0.0500

La matriz de clasificación será por tanto:

	Real 0	Real 1
Estimado 0	55.389 (73%)	8.504 (36%)
Estimado 1	20.120 (27%)	15.374 (64%)
Total	75.509 (100%)	23.878 (100%)

Es decir si comparamos la matriz de clasificación que obtuvimos con los datos de entrenamiento versus los obtenidos con los datos de test se tiene que:

Matriz clasificación datos entrenamiento		Matriz clasificación datos test	
Verdaderos positivos:	65%	Verdaderos positivos:	64%
Falsos positivos:	27%	Falsos positivos:	27%
Verdaderos negativos:	73%	Verdaderos negativos:	73%
Falsos negativos:	35%	Falsos negativos:	36%
Estimado correctamente:	72%	Estimado correctamente:	72%

Los resultados son similares, por lo que el modelo predice bien para los datos de test.

El Modelo 2 por lo tanto, es el modelo final obtenido con regresión logística.

- ***Paso 9: Resultados obtenidos e interpretación***

- **Descripción del modelo obtenido**

La probabilidad de fuga (probabilidad de dar de baja el seguro Premium Hogar) se estima de la siguiente forma:

$$P(Y=1) = 1/ (1+ e^{(-\beta 0- \beta 1X1 - \beta 2X2 -\ldots- \beta kXk)})$$

Siendo cada uno los parámetros β_i los siguientes:

Parameter	Categoría variable1	Categoría variable2	Estimate
Intercept			0
C18	1		0.409
C42	1		0.6909
C35	0		-0.5045
C37	NO		-0.3757
C71	HOMBRE		-0.0388
C11	1		0.4843
C11	2		0
C2*C58	1	1	0.1233
C2*C60	1	1	-0.1831
C2*C37	1	NO	-0.0795
C13*C15	1	1	-0.0704
C13*C55	1	1	0.0918
C14*C53	1	1	0
C14*C53	1	2	0.118
C14*C5	1	1	0.055
C14*C46	1	0	0.1851
C16*C75	1	ALTA	-0.3888
C16*C75	1	BAJA	0.1361
C18*C26	1	1	-0.0365
C18*C42	1	1	-0.0353
C18*C67	1	1	0.0529
C18*C36	1	NO	-0.3413
C23*C24	1	1	-0.0427
C23*C54	1	1	0.0732
C24*C54	1	1	-0.0602
C24*C70	1	0	0.041
C24*C10	1	1	0.0382
C25*C35	1	0	-0.0651
C25*C38	1	0	-0.0556

C26*C10	1	1	0.0389
C29*C45	1	1	0.0783
C3*C35	1	0	0.4907
C3*C75	1	ALTA	0
C3*C75	1	BAJA	0
C30*C48	1	1	-0.044
C30*C35	1	0	0.0685
C42*C45	1	1	-0.0626
C42*C60	1	1	-0.0356
C42*C37	1	NO	0.5775
C42*C46	1	0	0.0546
C45*C40	1	0	-0.0909
C45*C70	1	0	-0.0907
C48*C34	1	NO	-0.0683
C48*C46	1	0	0.04
C49*C46	1	0	0.0647
C50*C5	1	1	-0.0723
C50*C44	1	ALTA	-0.054
C50*C70	1	0	-0.1799
C54*C69	1	1	-2.8871
C54*C69	1	2	-2.9115
C54*C69	1	3	2.3985
C54*C69	1	4	1.9504
C54*C10	1	1	0.134
C53*C5	1	1	0.0694
C53*C5	2	1	0
C53*C70	1	0	0.0915
C53*C70	2	0	0
C53*C7	1	1	0
C53*C7	1	2	-0.0191
C53*C7	1	3	-0.1524
C53*C7	2	1	0
C53*C7	2	2	0
C53*C7	2	3	0.0929
C53*C10	1	1	0
C53*C10	2	1	-0.1518
C55*C60	1	1	0.0449
C4*C37	1	NO	-0.2783
C58*C59	1	1	-0.1701
C58*C71	1	HOMBRE	0.0338
C65*C11	1	1	0.087
C65*C11	1	2	0
C66*C38	1	0	0.0814
C66*C44	1	ALTA	0.0568

C68*C7	1	1	0
C68*C7	1	2	0
C68*C7	1	3	-1.3186
C68*C7	2	1	0
C68*C7	2	2	0
C68*C7	2	3	0.4215
C68*C7	3	1	0
C68*C7	3	2	0
C68*C7	3	3	0.8979
C67*C35	1	0	-0.0808
C34*C44	NO	ALTA	0.1075
C5*C7	1	1	0
C5*C7	1	2	0.0767
C5*C7	1	3	0
C36*C46	NO	0	0.2134
C37*C46	NO	0	-0.2482
C37*C11	NO	1	0.4171
C37*C11	NO	2	0
C38*C46	0	0	-0.0572
C38*C71	0	HOMBRE	-0.045
C38*C11	0	1	0.0439
C38*C11	0	2	0
C70*C7	0	1	0
C70*C7	0	2	0
C70*C7	0	3	-0.086
C7*C11	1	1	0
C7*C11	1	2	0.1041
C7*C11	2	1	0
C7*C11	2	2	0
C7*C11	3	1	-0.0625
C7*C11	3	2	0
C10*C11	1	1	-0.0709
C10*C11	1	2	0

Es decir, la ecuación será la siguiente para cada cliente:

$$P(Y=1) = 1/ (1+ e^{(-0.409 \cdot C18\_1 - 0.6909 \cdot C42\_1 + 0.5045 \cdot C35\_0 - \ldots.. + 0.0709 \cdot C10\_1V9\_2)})$$

Siendo:

- C18_1=1 si C18=1
  0 en caso contrario
  (es decir, si CAT_IMP_PROD_INVAH_ULT1=1 entonces C18_1=1)
- C42_1=1 si C42=1

0 en caso contrario
(es decir, si CAT_LIMITE_TJCRED_ULT1=1 entonces C42_1=1)

– C35_0=1 si C35=0
0 en caso contrario
(es decir, si IND_PARTES_HOGAR =0 entonces C35_0=1)

– C10_1V9_2=1 si C10=1 y C11=2
0 en caso contrario
(es decir, si CAT_CUOTA_TJCRED_ULT3=1 y CAT_EDAD=2 entonces C10_1V9_2=1)

- **Capacidad predictiva del modelo**

score	total acum	exitos acum	%acum total	%acum exitos
0.95 - 1.00	0	0	0%	0%
0.90 - 0.95	0	0	0%	0%
0.85 - 0.90	1	0	0%	0%
0.80 - 0.85	15	13	0%	0%
0.75 - 0.80	82	52	0%	0%
0.70 - 0.75	304	182	0%	1%
0.65 - 0.70	828	497	1%	2%
0.60 - 0.65	1951	1202	2%	5%
0.55 - 0.60	3871	2338	4%	10%
0.50 - 0.55	6665	3915	7%	16%
0.45 - 0.50	10173	5819	10%	24%
0.40 - 0.45	14056	7771	14%	33%
0.35 - 0.40	17268	9242	17%	39%
0.30 - 0.35	20528	10607	21%	44%
0.25 - 0.30	25762	12501	26%	52%
0.20 - 0.25	35494	15374	36%	64%
0.15 - 0.20	51178	18914	51%	79%
0.10 - 0.15	66345	21543	67%	90%
0.05 - 0.10	83925	23142	84%	97%
0.00 - 0.05	99387	23878	100%	100%

Los clientes con mayor propensión a la fuga son los que tienen un score>0,19 (prior con el que se ha entrenado el modelo)

Selección reducida:

(Seleccionamos aproximadamente el mismo volumen que con los árboles para que sea comparable)

– Con un 14% de los datos se predice el 33% de las fugas
– Supone una ganancia de x2,36

Selección ampliada:

– Con un 26% de los datos se predice el 52% de las fugas
– Supone una ganancia de x2,00

- **Matriz de clasificación:**

  El modelo consigue clasificar correctamente al 72% de los clientes:

Verdaderos positivos:	64%
Falsos positivos:	27%
Verdaderos negativos:	73%
Falsos negativos:	36%

- **Cálculo de la esperanza de vida:**

  Por lo tanto, el número de meses que transcurrirán hasta que el cliente de baja el producto es (se muestran los datos para las siguientes puntuaciones):

Orden	Score	Nº meses hasta la baja
1	0.975	1.03
2	0.925	1.08
3	0.875	1.14
4	0.825	1.21
5	0.775	1.29
6	0.725	1.38
7	0.675	1.48
8	0.625	1.60
9	0.575	1.74
10	0.525	1.90
11	0.475	2.11
12	0.425	2.35
13	0.375	2.67
14	0.325	3.08
15	0.275	3.64
16	0.225	4.44
17	0.175	5.71
18	0.125	8.00
19	0.075	13.33
20	0.025	40.00

Es decir, los clientes con mayor propensión a dar de baja el producto (los clientes con score menor o igual a 0.825) darán de baja el producto dentro de algo más de 1 mes (1,21 meses)

4.3 *Análisis de supervivencia*

➢ *Paso 1: Preparación del tablón de datos*

Realizamos este análisis sobre los datos de test: Marzo13.

➢ *Paso 2: Construcción del modelo*

Model Fit Statistics		
Criterion	Without Covariates	With Covariates
-2 LOG L	404941.83	397539.02
AIC	404941.83	397589.02
SBC	404941.83	397789.47

Valores más bajos indican mejor ajuste, lo cual es indicativo que si tenemos en cuenta las covariables el ajuste es mejor

Testing Global Null Hypothesis: BETA=0			
Test	Chi-Square	DF	Pr > ChiSq
Likelihood Ratio	7402.8047	25	<.0001
Score	7663.0589	25	<.0001
Wald	7294.1526	25	<.0001

El p-valor es significativo en los tres test, por lo que el modelo con todas las variables claramente es mejor.

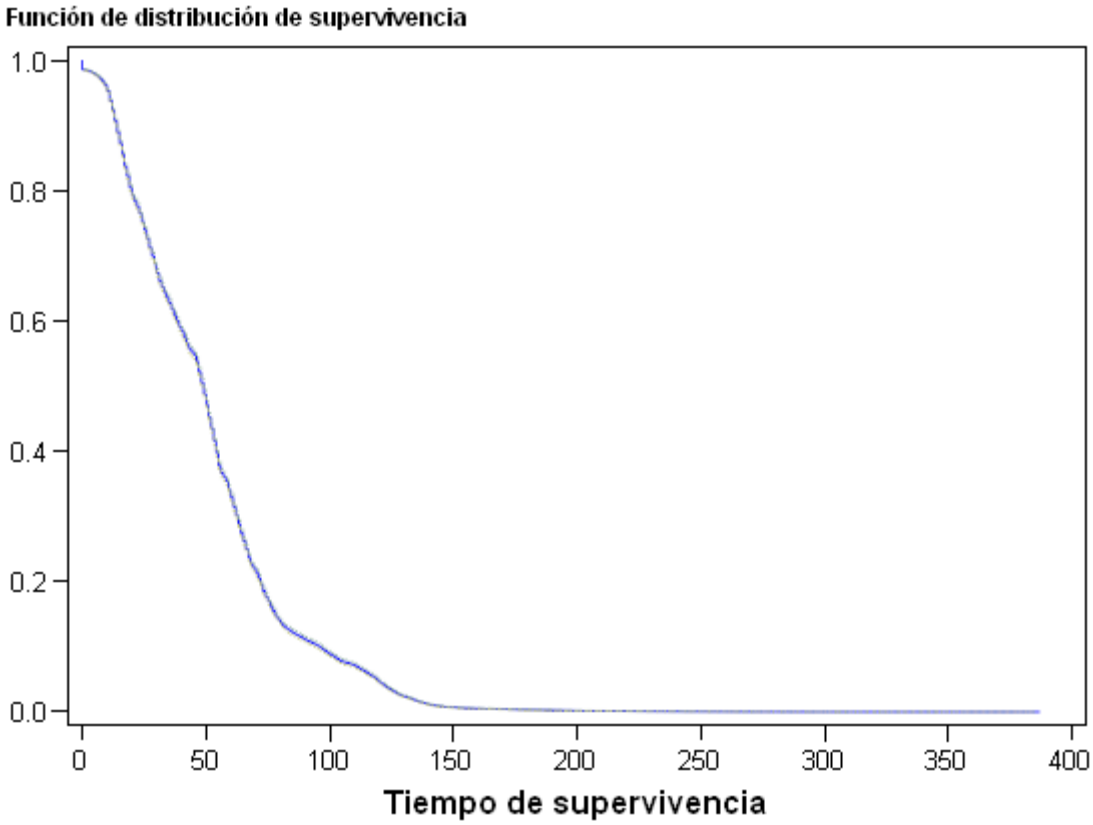
Analysis of Maximum Likelihood Estimates								
Variable	DF	Parameter Estimate	Standard Error	Chi-Square	Pr > Chi Sq	Hazard Ratio	95% Hazard Ratio Confidence Limits	
NUM_PROD_AHINV_ULT1	1	-3.11035	0.90501	11.8117	0.0006	0.045	0.008	0.263
RATIO_NUM_PROD_AHINV_MED_HACE1	1	3.15526	0.98498	10.2616	0.0014	23.459	3.403	161.705
RATIO_NUM_PROD_AH_MEDIA_ULT1	1	0.03588	0.00535	44.9035	<.0001	1.037	1.026	1.047
NUM_OPER_CRED_ULT1	1	-0.01696	0.00426	15.8546	<.0001	0.983	0.975	0.991
IMP_OPER_CRED_ULT3	1	0.0000167	3.25E-06	26.3847	<.0001	1	1	1
CUOTA_TJCRED_ULT1	1	0.0004891	0.0002209	4.9037	0.0268	1	1	1.001
ANTIGUEDAD_CLIENTE	1	-0.00158	0.0000799	392.5175	<.0001	0.998	0.998	0.999

EDAD	1	-0.0005717	0.0000471	147.3994	<.0001	0.999	0.999	1
NUM_PERSONAS_DEPENDIENTES	1	-0.06958	0.00518	180.5403	<.0001	0.933	0.923	0.942
NUM_PARTES_VIDA_SALUD_ULT4	1	-0.00964	0.0018	28.6256	<.0001	0.99	0.987	0.994
NUM_VIDA_ULT1	1	0.1038	0.00931	124.2929	<.0001	1.109	1.089	1.13
NUM_TJDEB_HACE1	1	0.20901	0.01948	115.0762	<.0001	1.232	1.186	1.28
NUM_TJDEB_HACE6	1	-0.08886	0.01884	22.2439	<.0001	0.915	0.882	0.949
NUM_DIAS_ULT_USO_TJDEB	1	-0.00447	0.0001884	562.7901	<.0001	0.996	0.995	0.996
IND_TJDEB_ULT1	1	0.08604	0.02575	11.1654	0.0008	1.09	1.036	1.146
NUM_CAMPANIAS_ULT3	1	0.02423	0.00194	156.1838	<.0001	1.025	1.021	1.028
COC_CUOTA_TJCRED_ULT1	1	-0.12128	0.04453	7.4192	0.0065	0.886	0.812	0.967
COC_LIM_TJCRED_ULT1	1	-0.19886	0.00779	651.9139	<.0001	0.82	0.807	0.832
COC_NUM_OPER_TJCRED_ULT1	1	0.14214	0.02623	29.3679	<.0001	1.153	1.095	1.214
COC_TJCRED_ULT1	1	0.12912	0.01514	72.7218	<.0001	1.138	1.105	1.172
NUM_DIAS_DESDE_COBRO	1	-0.0001219	0.0000333	13.4309	0.0002	1	1	1
IMPORTE_ULTIMO_COBRO	1	-0.00182	0.0002337	60.68	<.0001	0.998	0.998	0.999
IND_TJDEB_ULT3	1	0.05988	0.02615	5.2444	0.022	1.062	1.009	1.118
NUM_DIAS_CAT	1	-0.16522	0.01603	106.2893	<.0001	0.848	0.821	0.875
RECIB_PROMOCION	1	1.03142	0.02157	2285.5121	<.0001	2.805	2.689	2.926

Con estos parámetros se estima para cada cliente su tasa de supervivencia.

- ***Paso 3: Resultados obtenidos e interpretación***

La función de supervivencia obtenida es:



- Un cliente que lleva 10 meses con su contrato, tiene una tasa de supervivencia del 95%
- Un cliente que lleva 50 meses con su contrato, tiene una tasa de supervivencia del 46%
- Un cliente que lleva 82 meses con su contrato, tiene una tasa de supervivencia del 12%

5 Conclusiones

5.1 Modelo final

El problema que se quiere resolver es intentar evitar que los clientes den de baja el seguro Premium Hogar. Para ello tenemos dos modelos uno con árboles de decisión y el otro con regresión logística. Ahora bien, vamos a comprobar con qué modelo se consiguen mejores predicciones.

A continuación resumimos las similitudes y diferencias entre uno y otro:

- Según la capacidad predictiva de los modelos

Criterio	Árboles de decisión	Regresión logística
Matriz de clasificación	Verdaderos positivos: 44% Falsos positivos: 14% Verdaderos negativos: 86% Falsos negativos: 56% Estimado correctamente: 72%	Verdaderos positivos: 64% Falsos positivos: 27% Verdaderos negativos: 73% Falsos negativos: 36% Estimado correctamente: 72%
Porcentaje de éxitos capturados	· Con un 15,2% de los datos se predice el 34,9% de las fugas (X2,30) · Con un 32,2% de los datos se predice el 59,2% de las fugas (X1,84)	· Con un 14% de los datos se predice el 33% de las fugas (x2,36) · Con un 26% de los datos se predice el 52% de las fugas (x2,00)

- Según su funcionalidad

Criterio	Árboles de decisión	Regresión logística
Puntuación asignada a los clientes	Se asigna por bloques, es decir, los clientes que pertenezcan a la misma regla tienen la misma probabilidad de fuga asignada	Cada cliente tiene su propia probabilidad de fuga. Es una ventaja, ya que no se tienen prácticamente “empates entre ellos”
Sencillez	Los árboles de decisión son más sencillos de interpretar por personas con conocimientos básicos de estadística	

En el caso que estamos abordando el objetivo es intentar retener a los clientes que se fugan, por lo que lo más costoso sería clasificar a un individuo que realmente se fuga dentro del grupo “no fuga”. Es decir, nos interesa que la proporción de falsos negativos sea la menor posible. Con el árbol esta proporción es de un 56% mientras que con la regresión logística es de un 36%. Esta diferencia es lo suficientemente grande como para elegir el modelo de regresión logística aunque los árboles de decisión sean más sencillos de interpretar.

5.2 Aplicabilidad

Con los resultados del modelo vamos a intentar reducir la tasa de fuga, es decir, vamos a intentar reducir el número de clientes que den de baja el seguro Premium Hogar. Para ello, vamos a lanzar una campaña por Outbound (contratamos a un Contact Center para que llame a los clientes) sobre 2.000 clientes, ya que no se dispone de presupuesto para impactar a más. Les seleccionaremos con el modelo de regresión logística, es decir, llamaremos a los 2.000 clientes que tengan mayor probabilidad de fuga estimada.

Rango de probabilidad estimada	Nº clientes
0.95 - 1.00	0
0.90 - 0.95	0
0.85 - 0.90	1
0.80 - 0.85	14
0.75 - 0.80	67
0.70 - 0.75	222
0.65 - 0.70	524
0.60 - 0.65	1,123
0.55 - 0.60	1,920
0.50 - 0.55	2,794
0.45 - 0.50	3,508
0.40 - 0.45	3,883
0.35 - 0.40	3,212
0.30 - 0.35	3,260
0.25 - 0.30	5,234
0.20 - 0.25	9,732
0.15 - 0.20	15,684
0.10 - 0.15	15,167
0.05 - 0.10	17,580
0.00 - 0.05	15,462

6 Bibliografía

- Allison, P.D. (2001). Survival Analysis Using the SAS System. A Practival Guide. SAS Institute. Books by Users Press.
- Barry de Ville (2006). Decision Trees for Business Intelligence and Data Mining using SAS Enterprise Miner
- Dobson, A.J. and Barnett, A.G. (2008). An Introduction to Generalized Linear Models, third edition. Chapman & Hall/CRC.
- Fahrmeir, L. and Tutz, G. (2001). Multivariate Statistical Modelling Based on Generalized Linear Models, second edition. Springer.
- Hilbe, Joseph M. (2009). Logistic Regression Models. Chapman & Hall/CRC.
- Hosmer, D.W. and Lemeshow,S. (2008). Applied Survival Analysis. Regression Modeling of Time to Event data. Wiley
- Hougaard, P (2000). Analysis of Multivariate Survival Data. Springer.
- Klein, P.J. And Moechberger, M.L. (2003). Survival Analysis. Springer.
- Kleinbaum, D.G. and Klein, M. (2005) . Survival Analysis. A self learning text. Springer
- Myers, R.H., Montgomery, D.C., Vining, G.G. and Robinson, T.J. (2010) Generalized Linear Models with Applications in Engineering and the Sciences, second edition. Wiley.
- SAS Institute (2008). Advanced Predictive Modeling Using SAS Enterprise Miner. Course Notes
- SAS Institute (2006). Decision Tree Modeling. Course Notes
- SAS Institute (2005). Predictive Modeling Using Logistic Regression. Course Notes

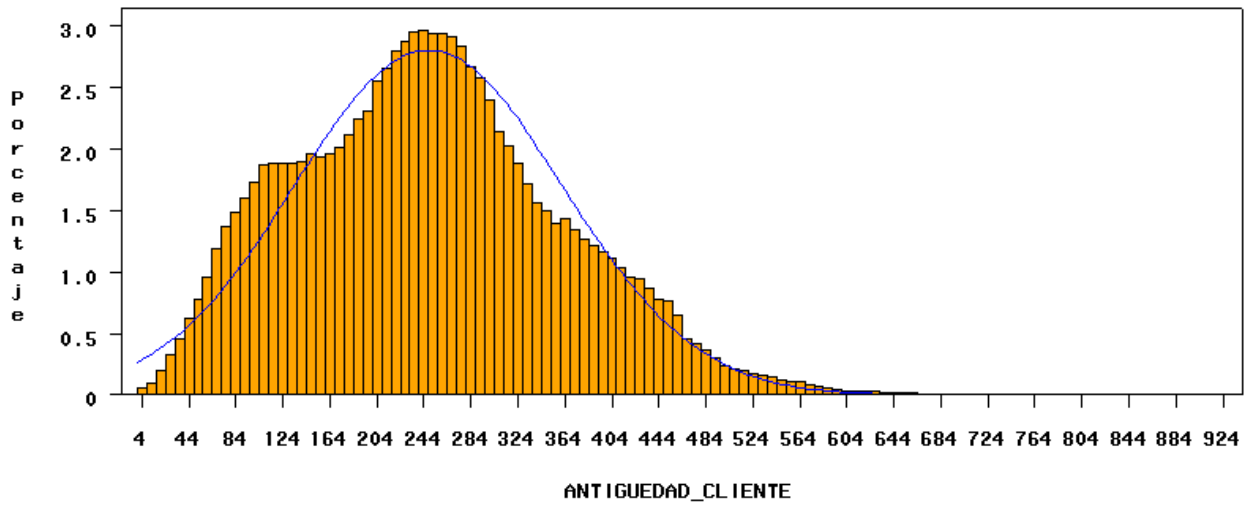
7 Anexos

7.1 Análisis descriptivo univariante

En este punto se muestra el análisis descriptivo univariante realizado sobre cada una de las variables con el fin de conocer cómo es la distribución de las mismas.

7.1.1 Variables numéricas

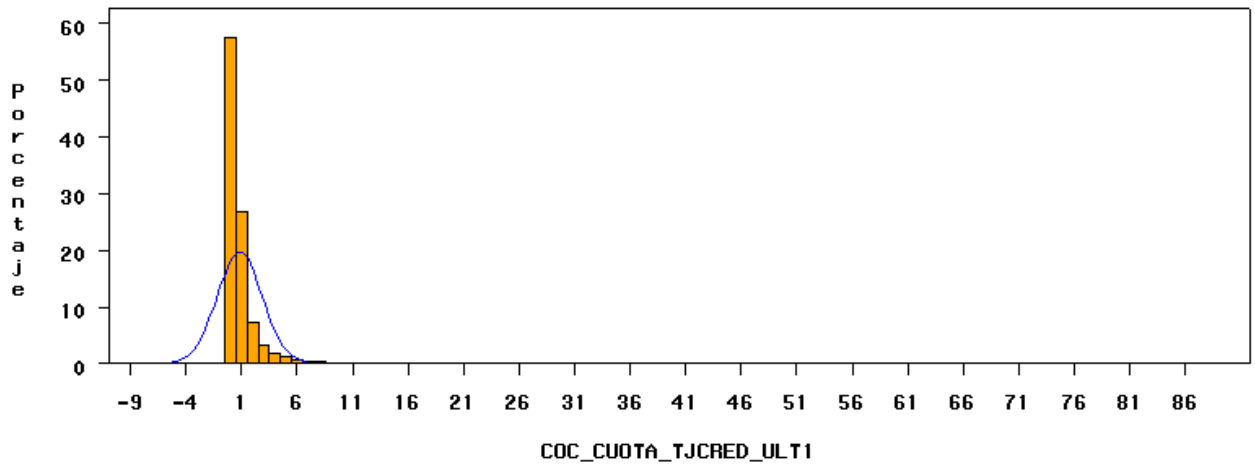
ANTIGUEDAD_CLIENTE: Antigüedad del cliente



N	Perdidos	Distintos	% Iguales	Infinitos	Negativos	Ceros	Positivos	¿Tiene Moda 0?	¿Tiene Valores atípicos por la izda?	¿Tiene Valores atípicos por la dcha?	Media (si moda es 0 se excluye del cálculo)	Desv.Tipica (si moda es 0 se excluye del cálculo)
473.275	0	3.650	99%	0	0	144	473.131	NO	NO	SI	247,91	114

Mínimo	Pct01	Pct05	Pct10	Pct25	Pct50	Pct75	Pct90	Pct95	Pct99	Max
0,00	37,20	73,20	99,80	163,20	243,40	319,60	403,80	446,80	535,60	899,80

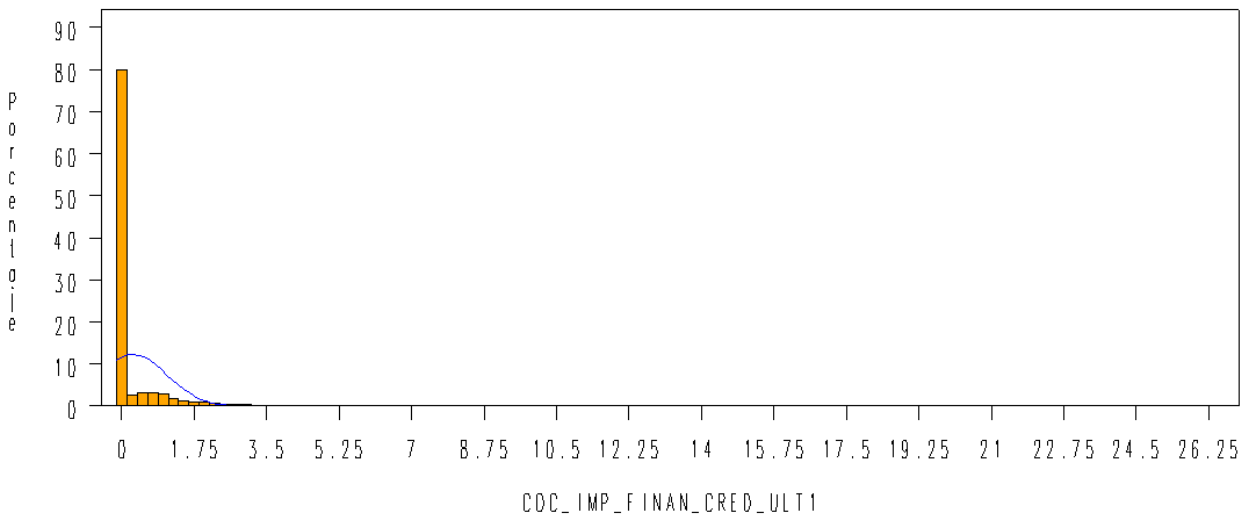
COC_CUOTA_TJCRED_ULT1: Cociente cuota con la tarjeta de crédito



N	Perdidos	Distintos	% Iguales	Infinitos	Negativos	Ceros	Positivos	¿Tiene Moda 0?	¿Tiene Valores atípicos por la izda?	¿Tiene Valores atípicos por la dcha?	Media (si moda es 0 se excluye del cálculo)	Desv.Tipica (si moda es 0 se excluye del cálculo)
473.275	13	112.596	76%	0	139	271.585	201.538	SI	SI	SI	1,98	3

Mínimo	Pct01	Pct05	Pct10	Pct25	Pct50	Pct75	Pct90	Pct95	Pct99	Max
-7,19	0,55	0,63	0,69	0,82	1,15	2,08	3,99	5,81	12,70	76,83

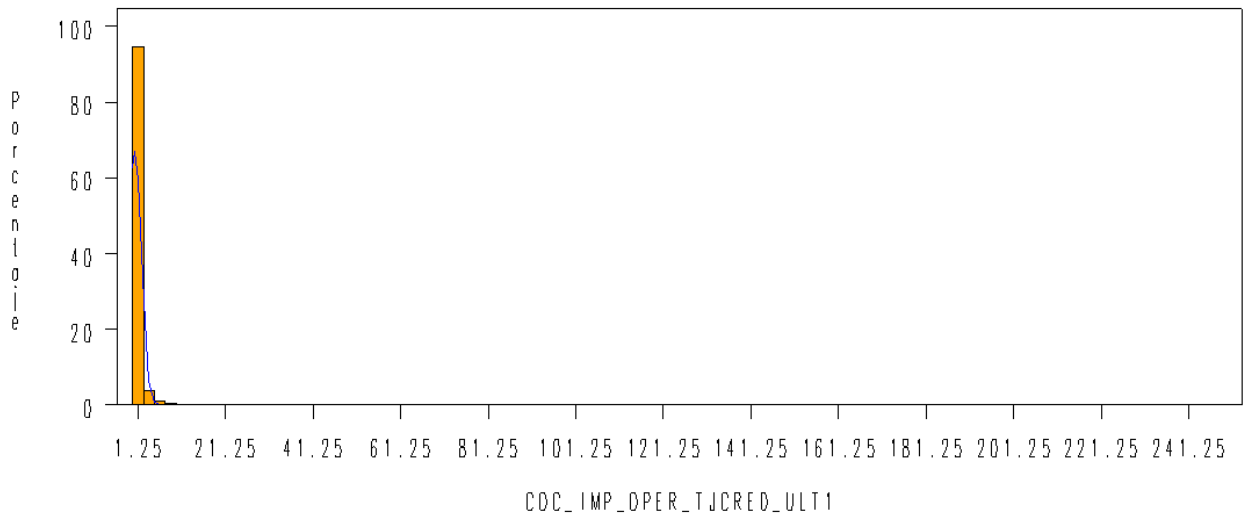
COC_IMP_FINAN_CRED_ULT1: Cociente importe financiado a crédito



N	Perdidos	Distintos	% Iguales	Infinitos	Negativos	Ceros	Positivos	¿Tiene Moda 0?	¿Tiene Valores atípicos por la izda?	¿Tiene Valores atípicos por la dcha?	Media (si moda es 0 se excluye del cálculo)	Desv.Tipica (si moda es 0 se excluye del cálculo)
473.275	13	91.129	81%	0	0	378.676	94.586	SI	NO	SI	1,36	1

Mínimo	Pct01	Pct05	Pct10	Pct25	Pct50	Pct75	Pct90	Pct95	Pct99	Max
0,22	0,22	0,27	0,33	0,55	0,95	1,65	2,75	3,76	6,86	21,82

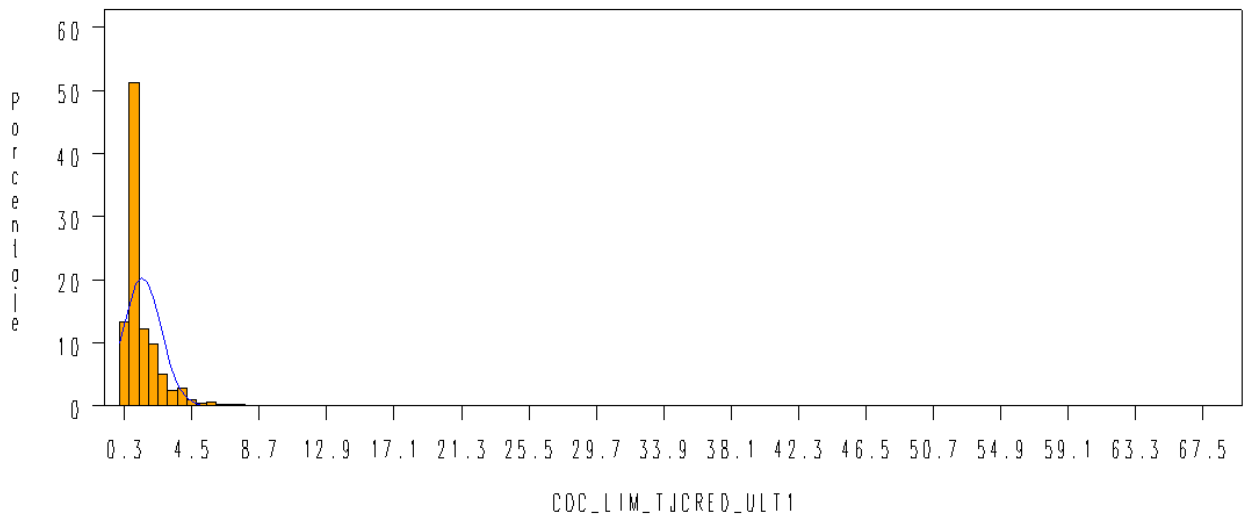
COC_IMP_OPER_TJCRED_ULT1: Cociente importe de las operaciones con la tarjeta de crédito



N	Perdidos	Distintos	% Iguales	Infinitos	Negativos	Ceros	Positivos	¿Tiene Moda 0?	¿Tiene Valores atípicos por la izda?	¿Tiene Valores atípicos por la dcha?	Media (si moda es 0 se excluye del cálculo)	Desv.Tipica (si moda es 0 se excluye del cálculo)
473.275	13	118.022	75%	0	0	296.271	176.991	SI	NO	SI	1,42	2

Mínimo	Pct01	Pct05	Pct10	Pct25	Pct50	Pct75	Pct90	Pct95	Pct99	Max
0,15	0,16	0,18	0,24	0,38	0,78	1,67	3,15	4,56	9,49	97,29

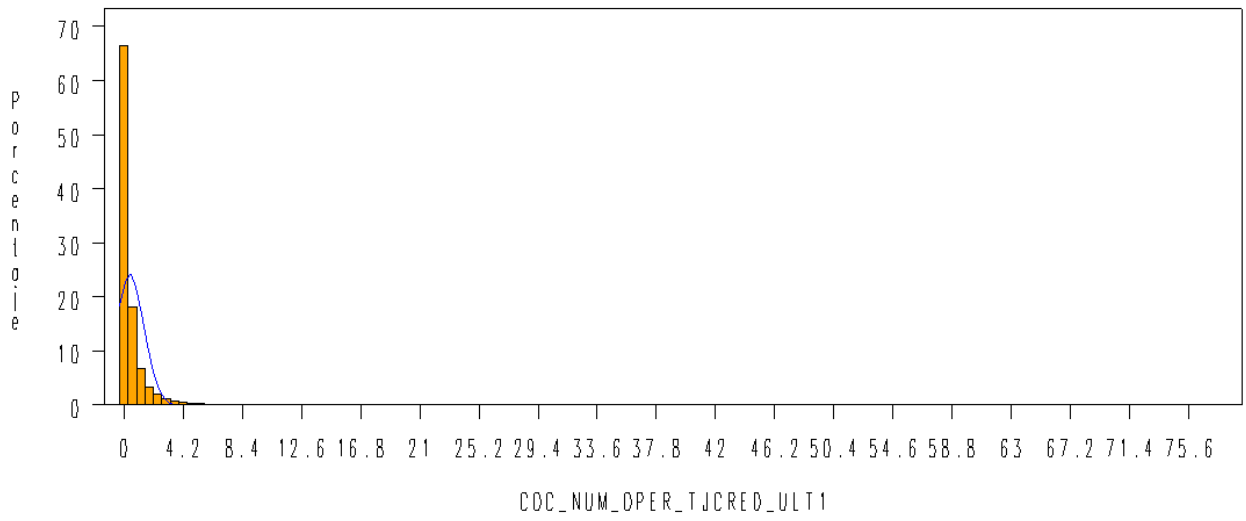
COC_LIM_TJCRED_ULT1: Cociente límite con la tarjeta de crédito



N	Perdidos	Distintos	% Iguales	Infinitos	Negativos	Ceros	Positivos	¿Tiene Moda 0?	¿Tiene Valores atípicos por la izda?	¿Tiene Valores atípicos por la dcha?	Media (si moda es 0 se excluye del cálculo)	Desv.Tipica (si moda es 0 se excluye del cálculo)
473.275	13	3.877	99%	0	0	375	472.887	NO	NO	SI	1,41	1

Mínimo	Pct01	Pct05	Pct10	Pct25	Pct50	Pct75	Pct90	Pct95	Pct99	Max
0,00	0,42	0,42	0,50	0,67	1,05	1,76	2,91	3,77	5,92	39,88

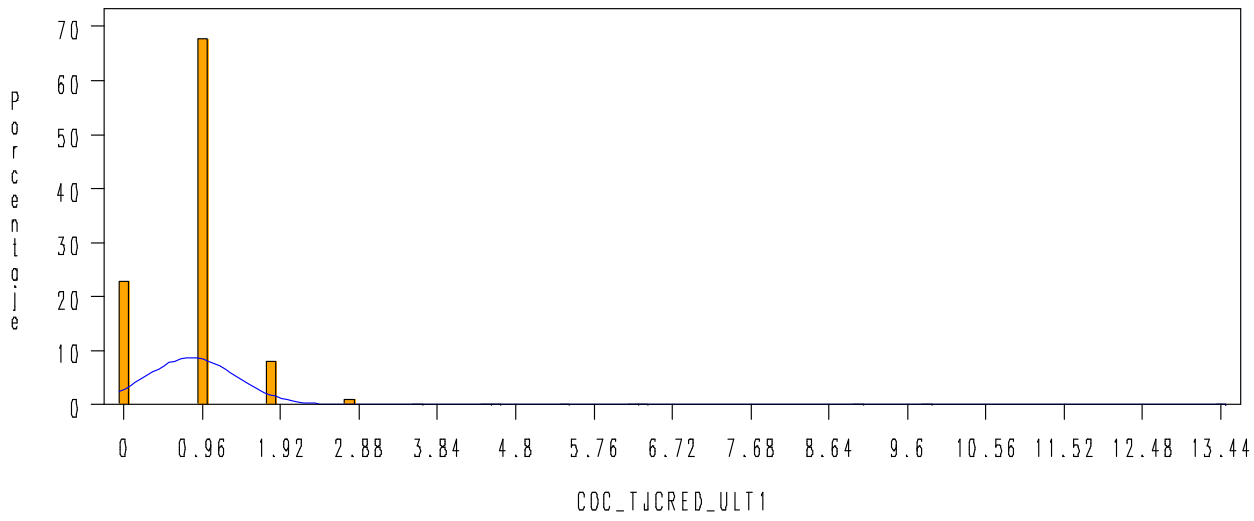
COC_NUM_OPER_TJCRED_ULT1: Cociente número de operaciones con la tarjeta de crédito



N	Perdidos	Distintos	% Iguales	Infinitos	Negativos	Ceros	Positivos	¿Tiene Moda 0?	¿Tiene Valores atípicos por la izda?	¿Tiene Valores atípicos por la dcha?	Media (si moda es 0 se excluye del cálculo)	Desv.Tipica (si moda es 0 se excluye del cálculo)
473.275	13	403	100%	0	0	296.271	176.991	SI	NO	SI	1,18	1

Mínimo	Pct01	Pct05	Pct10	Pct25	Pct50	Pct75	Pct90	Pct95	Pct99	Max
0,30	0,30	0,30	0,30	0,30	0,69	1,45	2,64	3,63	6,40	40,50

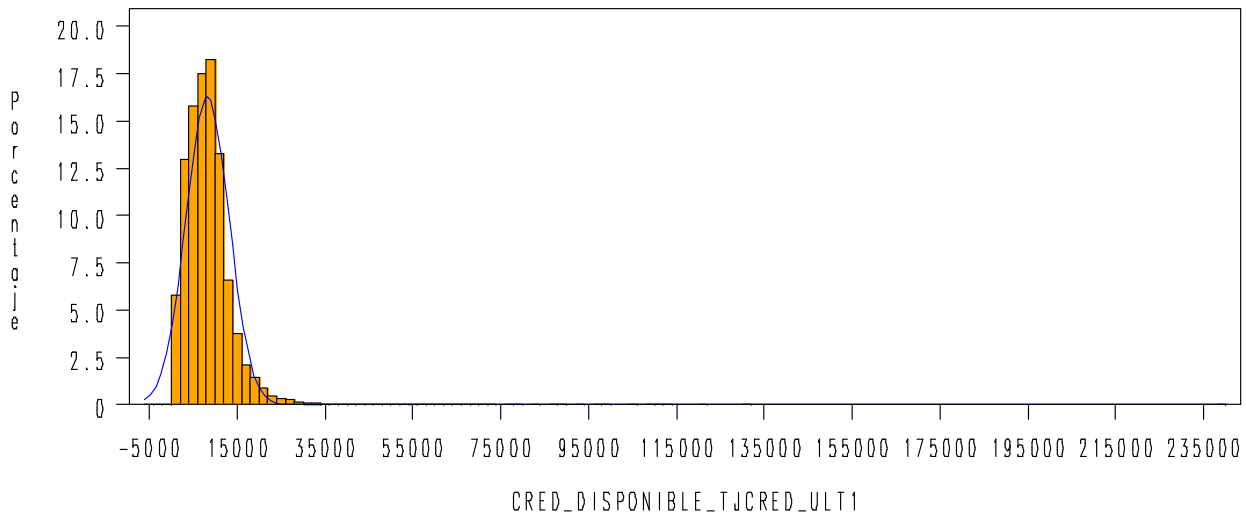
COC_TJCRED_ULT1: Cociente tarjeta de crédito



N	Perdidos	Distintos	% Iguales	Infinitos	Negativos	Ceros	Positivos	¿Tiene Moda 0?	¿Tiene Valores atípicos por la izda?	¿Tiene Valores atípicos por la dcha?	Media (si moda es 0 se excluye del cálculo)	Desv.Tipica (si moda es 0 se excluye del cálculo)
473.275	0	44	100%	0	0	108.373	364.902	NO	SI	SI	0,82	1

Mínimo	Pct01	Pct05	Pct10	Pct25	Pct50	Pct75	Pct90	Pct95	Pct99	Max
0,00	0,00	0,00	0,00	0,94	0,94	0,94	1,12	1,83	2,72	8,96

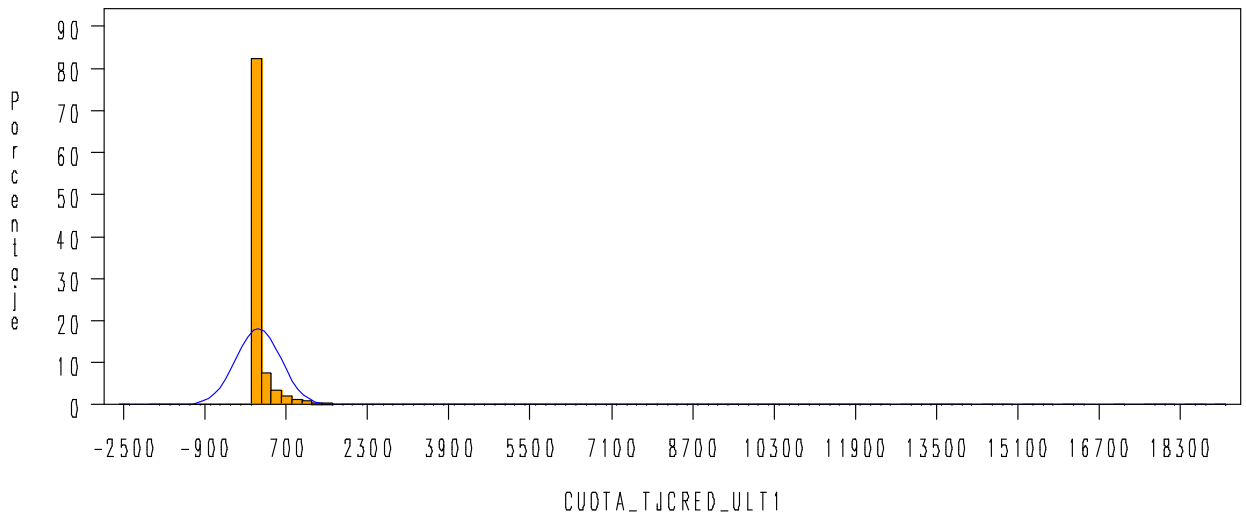
CRED_DISPONIBLE_TJCRED_ULT1: Crédito disponible en la tarjeta de crédito



N	Perdidos	Distintos	% Iguales	Infinitos	Negativos	Ceros	Positivos	¿Tiene Moda 0?	¿Tiene Valores atípicos por la izda?	¿Tiene Valores atípicos por la dcha?	Media (si moda es 0 se excluye del cálculo)	Desv.Tipica (si moda es 0 se excluye del cálculo)
473.275	13	264.190	44%	0	24	739	472.499	SI	NO	SI	8.254,82	4.900

Mínimo	Pct01	Pct05	Pct10	Pct25	Pct50	Pct75	Pct90	Pct95	Pct99	Max
-3.284,40	753,55	1.869,22	2.782,65	4.833,52	7.772,84	10.559,07	13.922,57	16.883,52	24.308,94	140.830,52

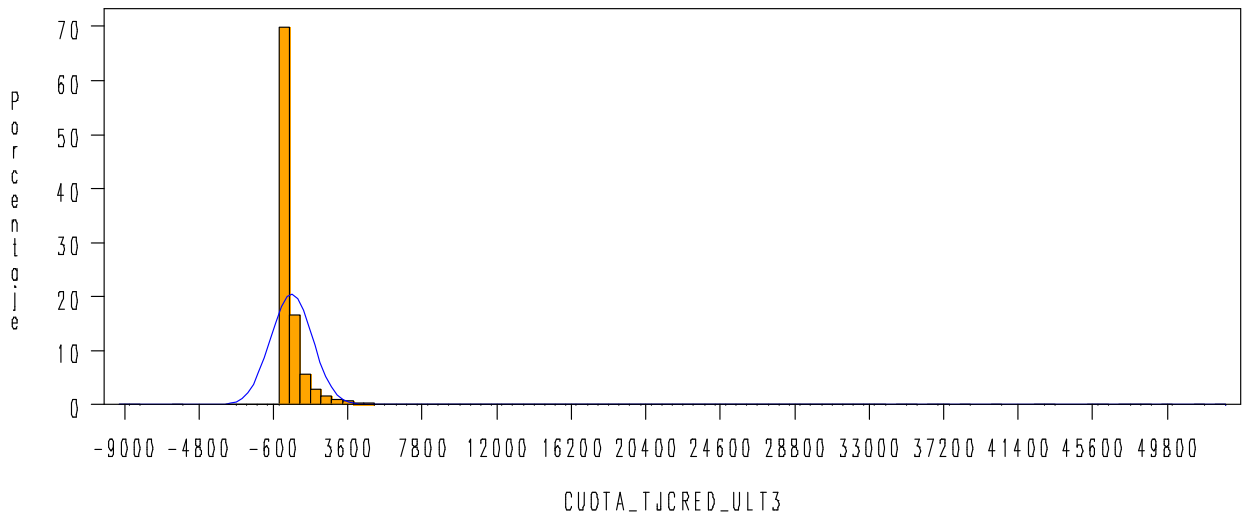
CUOTA_TJCRED_ULT1: Cuota con la tarjeta de crédito en el último mes



N	Perdidos	Distintos	% Iguales	Infinitos	Negativos	Ceros	Positivos	¿Tiene Moda 0?	¿Tiene Valores atípicos por la izda?	¿Tiene Valores atípicos por la dcha?	Media (si moda es 0 se excluye del cálculo)	Desv.Tipica (si moda es 0 se excluye del cálculo)
473.275	13	115.892	76%	0	932	271.588	200.742	SI	SI	SI	335,76	623

Mínimo	Pct01	Pct05	Pct10	Pct25	Pct50	Pct75	Pct90	Pct95	Pct99	Max
-1.791,97	3,74	22,44	37,12	67,85	143,58	357,77	797,94	1.219,29	2.804,70	17.566,98

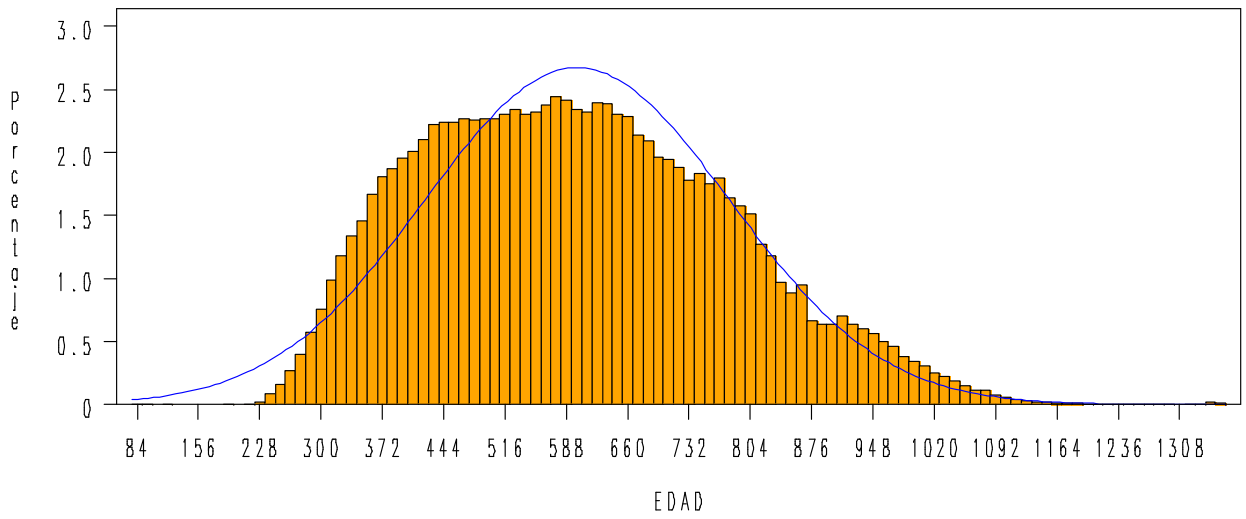
CUOTA_TJCRED_ULT3: Cuota con la tarjeta de crédito en los últimos 3 meses



N	Perdidos	Distintos	% Iguales	Infinitos	Negativos	Ceros	Positivos	¿Tiene Moda 0?	¿Tiene Valores atípicos por la izda?	¿Tiene Valores atípicos por la dcha?	Media (si moda es 0 se excluye del cálculo)	Desv.Tipica (si moda es 0 se excluye del cálculo)
473.275	13	162.528	66%	0	717	241.172	231.373	SI	SI	SI	877,82	1.543

Mínimo	Pct01	Pct05	Pct10	Pct25	Pct50	Pct75	Pct90	Pct95	Pct99	Max
-5.855,21	7,55	40,44	82,42	171,24	392,64	977,93	2.096,98	3.141,47	6.802,30	49.339,28

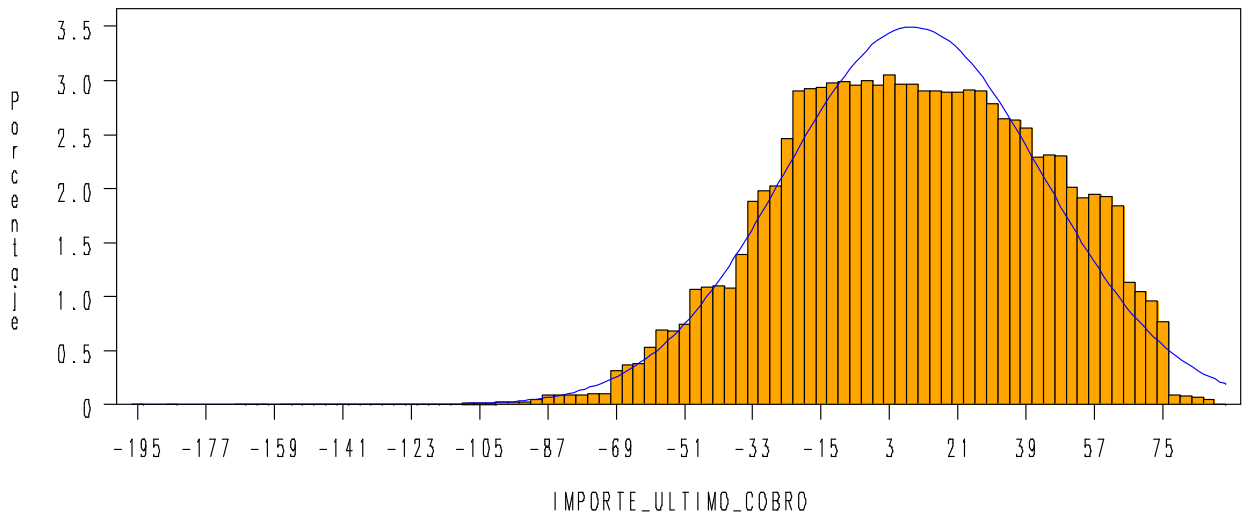
EDAD: Edad (en meses)



N	Perdidos	Distintos	% Iguales	Infinitos	Negativos	Ceros	Positivos	¿Tiene Moda 0?	¿Tiene Valores atípicos por la izda?	¿Tiene Valores atípicos por la dcha?	Media (si moda es 0 se excluye del cálculo)	Desv.Tipica (si moda es 0 se excluye del cálculo)
473.275	0	4.939	99%	0	0	0	473.275	NO	NO	NO	600,62	179

Mínimo	Pct01	Pct05	Pct10	Pct25	Pct50	Pct75	Pct90	Pct95	Pct99	Max
165,00	283,20	334,60	372,80	459,40	588,60	725,40	841,40	921,60	1.033,40	1.353,80

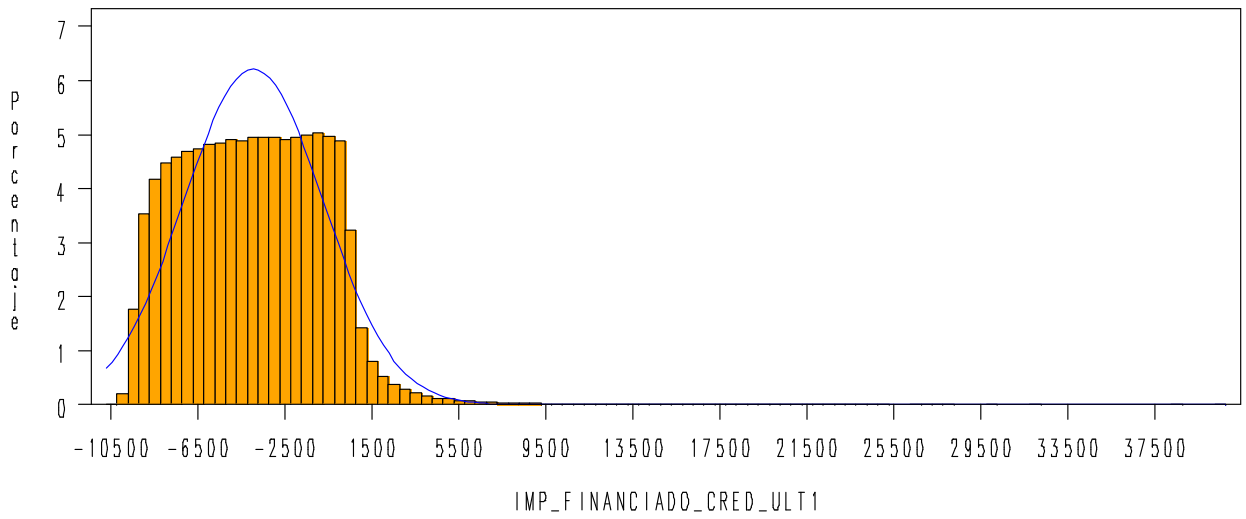
IMPORTE_ULTIMO_COBRO: Importe último cobro



N	Perdidos	Distintos	% Iguales	Infinitos	Negativos	Ceros	Positivos	¿Tiene Moda 0?	¿Tiene Valores atípicos por la izda?	¿Tiene Valores atípicos por la dcha?	Media (si moda es 0 se excluye del cálculo)	Desv.Tipica (si moda es 0 se excluye del cálculo)
473.275	96.237	1.703	100%	0	151.668	3.750	221.620	NO	SI	NO	9,19	34

Mínimo	Pct01	Pct05	Pct10	Pct25	Pct50	Pct75	Pct90	Pct95	Pct99	Max
-175,00	-67,80	-48,20	-35,00	-16,00	9,40	35,40	55,60	63,20	73,60	88,80

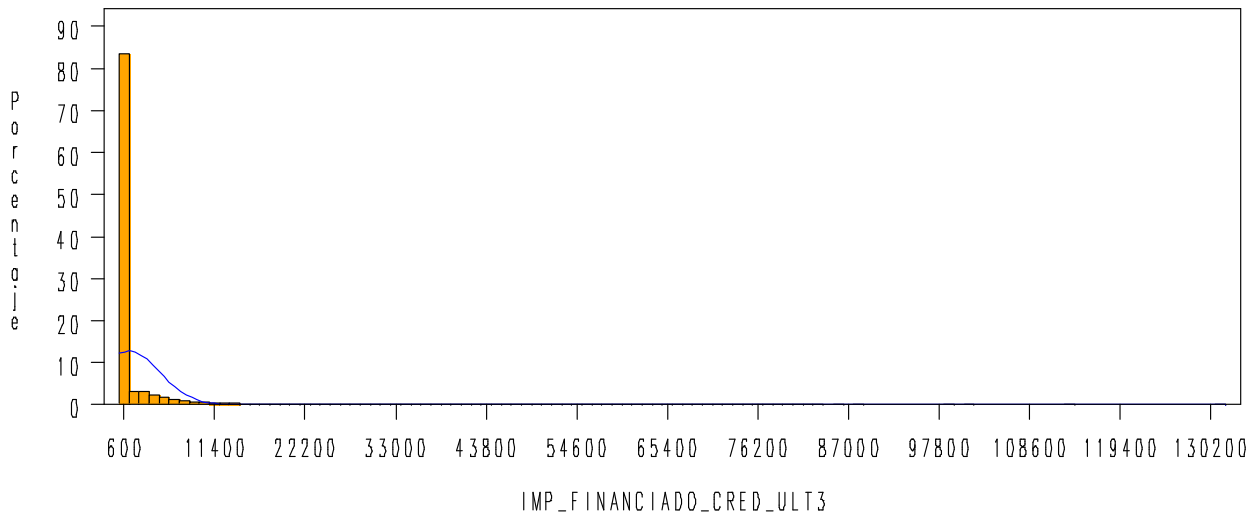
IMP_FINANCIADO_CRED_ULT1: Importe financiado a crédito en el último mes



N	Perdidos	Distintos	% Iguales	Infinitos	Negativos	Ceros	Positivos	¿Tiene Moda 0?	¿Tiene Valores atípicos por la izda?	¿Tiene Valores atípicos por la dcha?	Media (si moda es 0 se excluye del cálculo)	Desv.Tipica (si moda es 0 se excluye del cálculo)
473.275	13	234.036	51%	0	424.978	381	47.903	SI	NO	SI	-3.969,43	3.211

Mínimo	Pct01	Pct05	Pct10	Pct25	Pct50	Pct75	Pct90	Pct95	Pct99	Max
-10.149,33	-9.454,00	-8.814,00	-8.214,21	-6.586,86	-4.010,14	-1.485,84	14,19	664,62	3.600,28	33.946,19

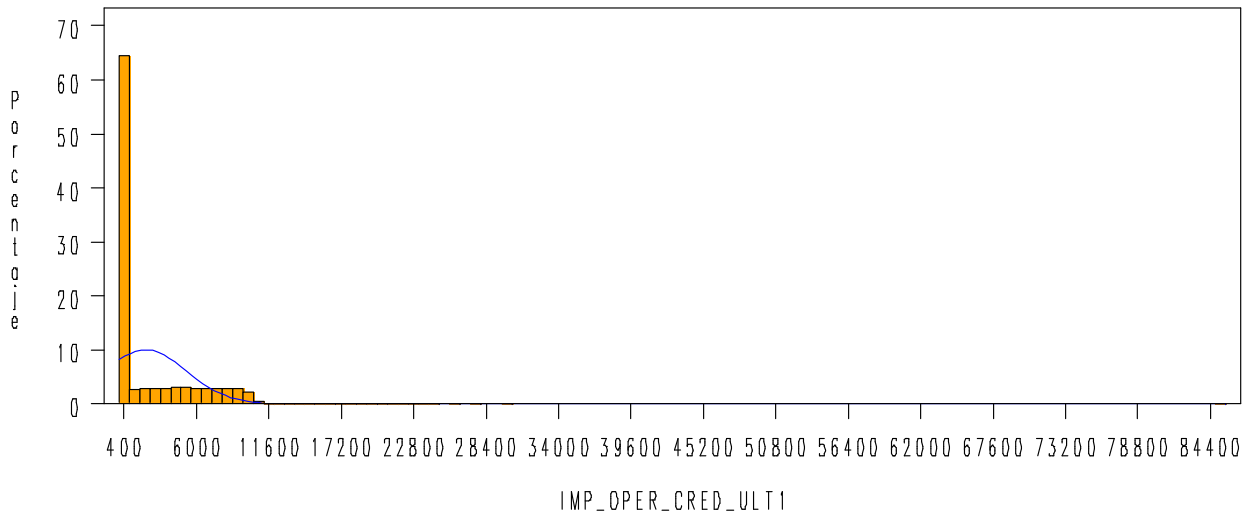
IMP_FINANCIADO_CRED_ULT3: Importe financiado a crédito en los últimos 3 meses



N	Perdidos	Distintos	% Iguales	Infinitos	Negativos	Ceros	Positivos	¿Tiene Moda 0?	¿Tiene Valores atípicos por la izda?	¿Tiene Valores atípicos por la dcha?	Media (si moda es 0 se excluye del cálculo)	Desv.Tipica (si moda es 0 se excluye del cálculo)
473.275	13	99.129	79%	0	0	371.961	101.301	SI	NO	SI	5.298,50	6.595

Mínimo	Pct01	Pct05	Pct10	Pct25	Pct50	Pct75	Pct90	Pct95	Pct99	Max
0,16	9,78	120,47	351,96	1.349,13	3.365,48	6.775,64	11.985,90	17.137,65	32.321,02	105.697,74

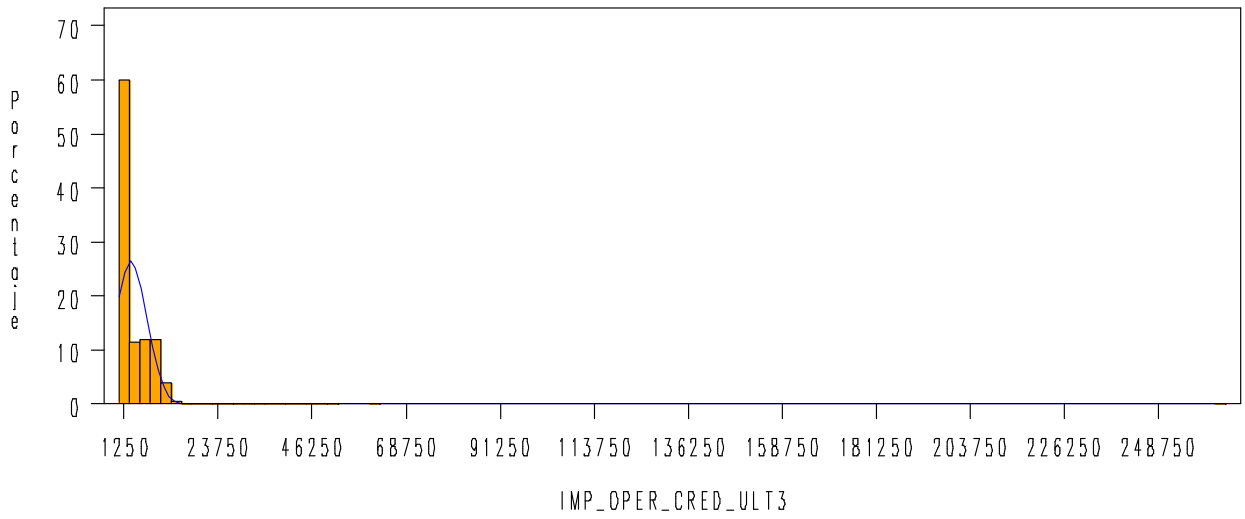
IMP_OPER_CRED_ULT1: Importe de las operaciones con la tarjeta de crédito en el último mes



N	Perdidos	Distintos	% Iguales	Infinitos	Negativos	Ceros	Positivos	¿Tiene Moda 0?	¿Tiene Valores atípicos por la izda?	¿Tiene Valores atípicos por la dcha?	Media (si moda es 0 se excluye del cálculo)	Desv.Tipica (si moda es 0 se excluye del cálculo)
473.275	13	165.898	65%	0	0	296.271	176.991	SI	NO	SI	5.409,03	2.964

Mínimo	Pct01	Pct05	Pct10	Pct25	Pct50	Pct75	Pct90	Pct95	Pct99	Max
8,69	277,44	811,16	1.373,35	2.894,55	5.393,78	7.901,08	9.405,57	9.912,61	10.850,81	38.585,09

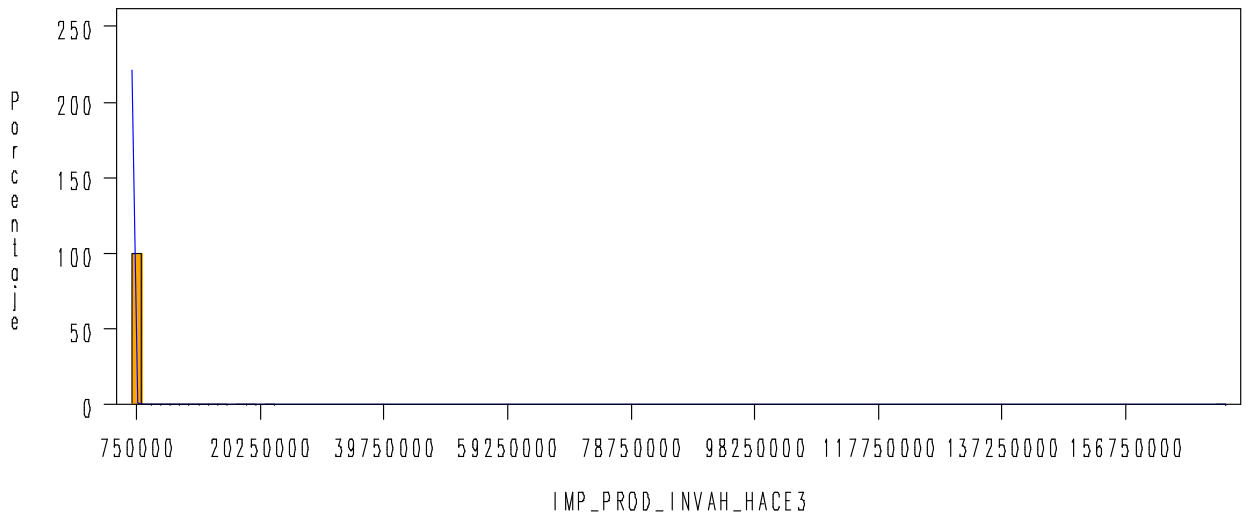
IMP_OPER_CRED_ULT3: Importe de las operaciones con la tarjeta de crédito en los últimos 3 meses



N	Perdidos	Distintos	% Iguales	Infinitos	Negativos	Ceros	Positivos	¿Tiene Moda 0?	¿Tiene Valores atípicos por la izda?	¿Tiene Valores atípicos por la dcha?	Media (si moda es 0 se excluye del cálculo)	Desv.Tipica (si moda es 0 se excluye del cálculo)
473.275	13	215.979	54%	0	0	246.187	227.075	SI	NO	SI	5.991,78	3.308

Mínimo	Pct01	Pct05	Pct10	Pct25	Pct50	Pct75	Pct90	Pct95	Pct99	Max
14,38	401,80	1.081,67	1.727,50	3.373,68	5.933,45	8.456,77	9.975,51	10.772,90	13.602,23	94.455,49

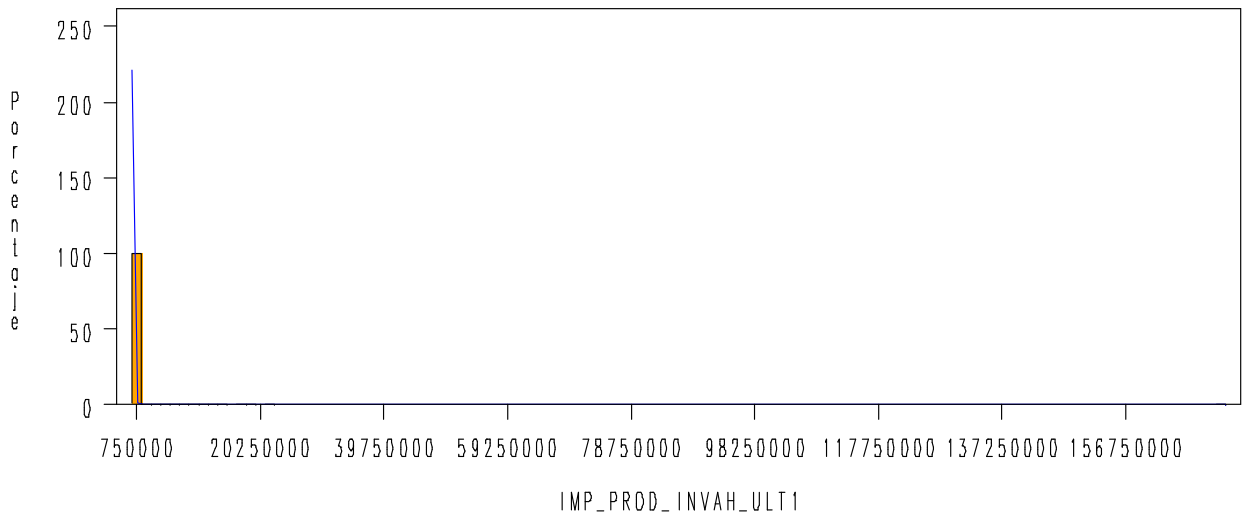
IMP_PROD_INVAH_HACE3: Importe productos de inversión y ahorro hace 3 meses



N	Perdidos	Distintos	% Iguales	Infinitos	Negativos	Ceros	Positivos	¿Tiene Moda 0?	¿Tiene Valores atípicos por la izda?	¿Tiene Valores atípicos por la dcha?	Media (si moda es 0 se excluye del cálculo)	Desv.Tipica (si moda es 0 se excluye del cálculo)
473.275	0	143.188	70%	0	0	299.168	174.107	SI	NO	SI	35.696,35	322.614

Mínimo	Pct01	Pct05	Pct10	Pct25	Pct50	Pct75	Pct90	Pct95	Pct99	Max
0,01	0,56	5,11	44,19	750,70	6.556,90	27.801,59	75.850,57	132.108,70	401.347,77	47.173.072,95

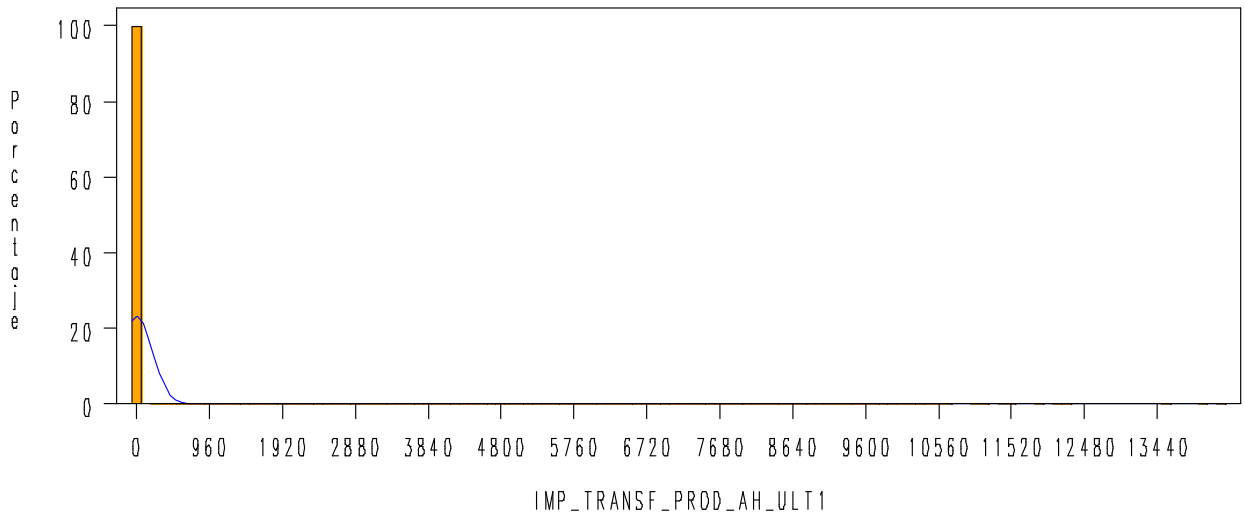
IMP_PROD_INVAH_ULT1: Importe productos de inversión y ahorro en el último mes



N	Perdidos	Distintos	% Iguales	Infinitos	Negativos	Ceros	Positivos	¿Tiene Moda 0?	¿Tiene Valores atípicos por la izda?	¿Tiene Valores atípicos por la dcha?	Media (si moda es 0 se excluye del cálculo)	Desv.Tipica (si moda es 0 se excluye del cálculo)
473.275	0	141.853	70%	0	0	299.805	173.470	SI	NO	SI	35.520,33	322.447

Mínimo	Pct01	Pct05	Pct10	Pct25	Pct50	Pct75	Pct90	Pct95	Pct99	Max
0,01	0,53	4,77	41,77	726,60	6.537,91	27.664,46	75.871,34	131.835,00	400.833,61	47.387.010,07

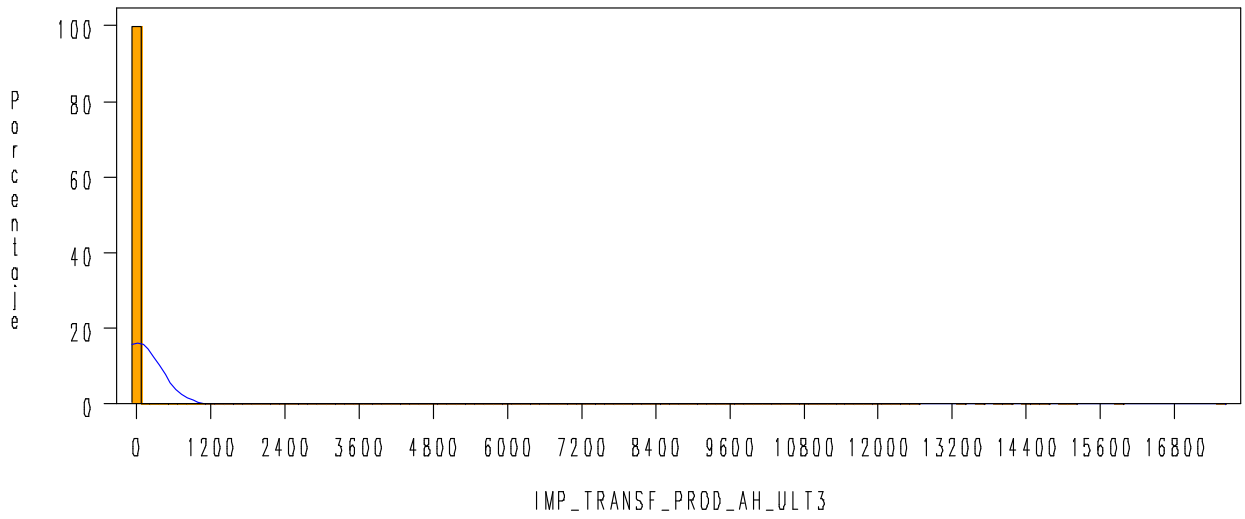
IMP_TRANSF_PROD_AH_ULT1: Importe transferido a productos de ahorro en el último mes



N	Perdidos	Distintos	% Iguales	Infinitos	Negativos	Ceros	Positivos	¿Tiene Moda 0?	¿Tiene Valores atípicos por la izda?	¿Tiene Valores atípicos por la dcha?	Media (si moda es 0 se excluye del cálculo)	Desv.Tipica (si moda es 0 se excluye del cálculo)
473.275	13	505	100%	0	0	472.762	500	SI	NO	SI	5.645,73	2.939

Mínimo	Pct01	Pct05	Pct10	Pct25	Pct50	Pct75	Pct90	Pct95	Pct99	Max
437,96	531,57	1.184,43	1.679,77	3.306,74	5.573,48	7.942,34	9.730,49	10.248,55	11.905,34	12.881,10

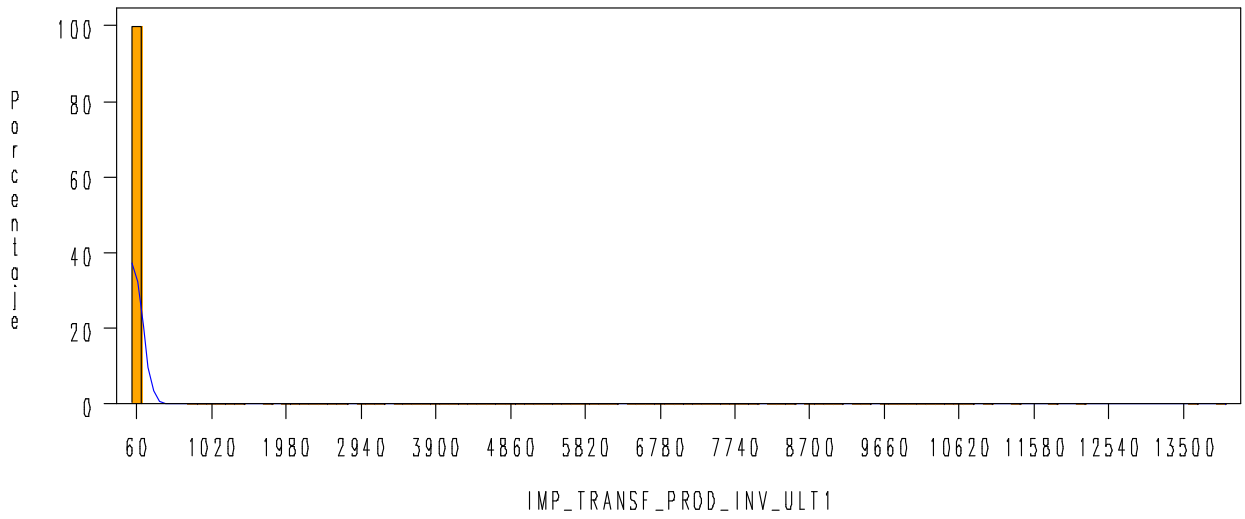
IMP_TRANSF_PROD_AH_ULT3: Importe transferido a productos de ahorro en los últimos 3 meses



N	Perdidos	Distintos	% Iguales	Infinitos	Negativos	Ceros	Positivos	¿Tiene Moda 0?	¿Tiene Valores atípicos por la izda?	¿Tiene Valores atípicos por la dcha?	Media (si moda es 0 se excluye del cálculo)	Desv.Tipica (si moda es 0 se excluye del cálculo)
473.275	13	1.477	100%	0	0	471.786	1.476	SI	NO	SI	5.916,89	3.053

Mínimo	Pct01	Pct05	Pct10	Pct25	Pct50	Pct75	Pct90	Pct95	Pct99	Max
300,88	528,70	1.143,11	1.829,56	3.437,59	5.941,88	8.361,00	9.934,29	10.456,91	12.951,43	15.284,19

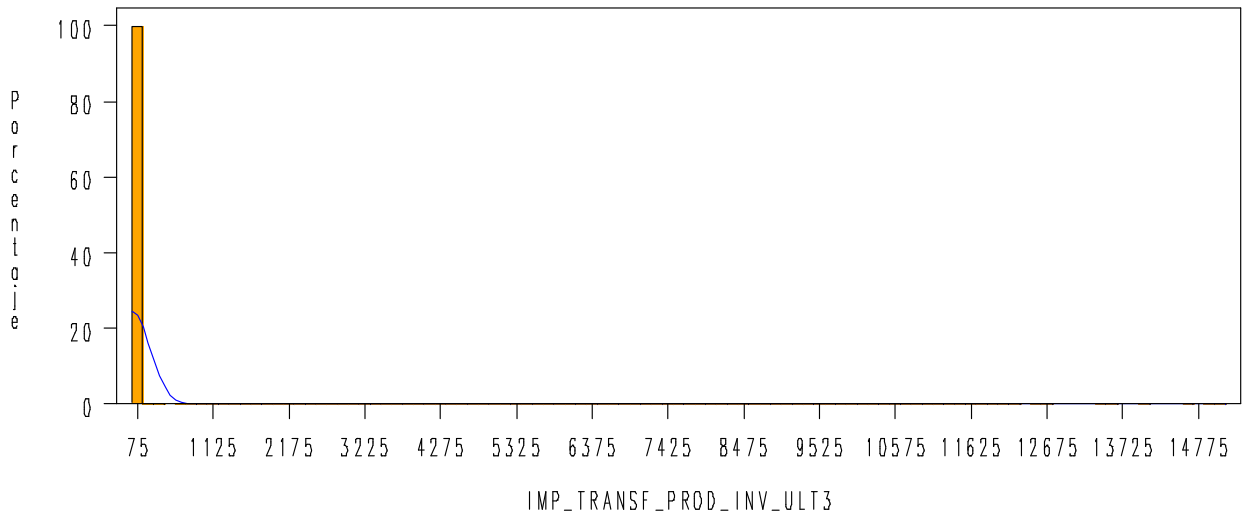
IMP_TRANSF_PROD_INV_ULT1: Importe transferido a productos de inversión en el último mes



N	Perdidos	Distintos	% Iguales	Infinitos	Negativos	Ceros	Positivos	¿Tiene Moda 0?	¿Tiene Valores atípicos por la izda?	¿Tiene Valores atípicos por la dcha?	Media (si moda es 0 se excluye del cálculo)	Desv.Tipica (si moda es 0 se excluye del cálculo)
473.275	13	180	100%	0	0	473.087	175	SI	NO	SI	5.823,74	2.993

Mínimo	Pct01	Pct05	Pct10	Pct25	Pct50	Pct75	Pct90	Pct95	Pct99	Max
1.034,30	1.034,30	1.288,61	2.082,85	3.408,00	5.828,03	7.949,80	10.108,80	10.689,80	12.057,70	12.057,70

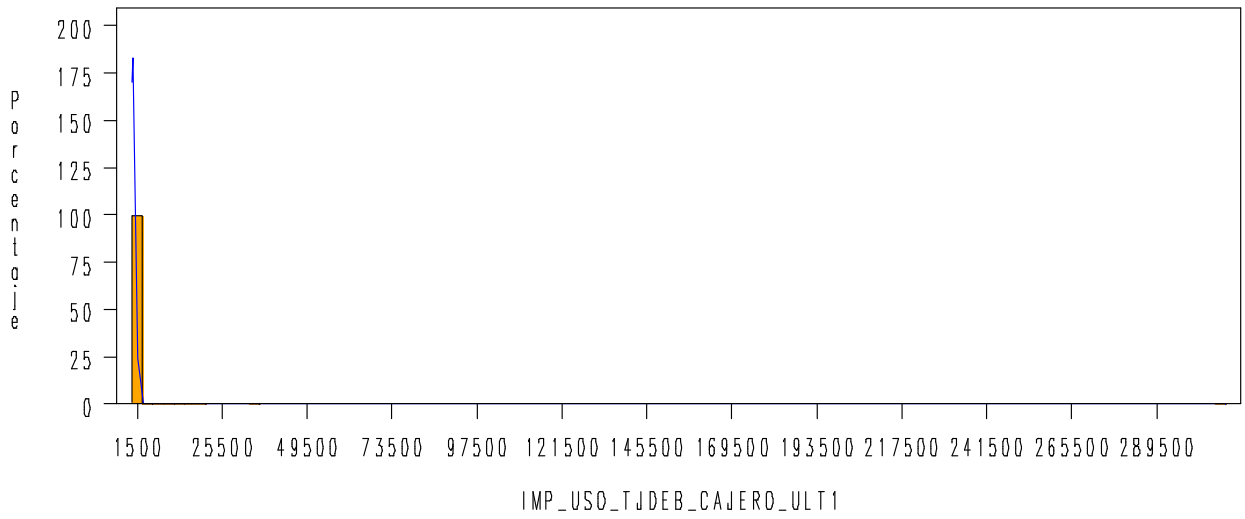
IMP_TRANSF_PROD_INV_ULT3: Importe transferido a productos de inversión en los últimos 3 meses



N	Perdidos	Distintos	% Iguales	Infinitos	Negativos	Ceros	Positivos	¿Tiene Moda 0?	¿Tiene Valores atípicos por la izda?	¿Tiene Valores atípicos por la dcha?	Media (si moda es 0 se excluye del cálculo)	Desv.Tipica (si moda es 0 se excluye del cálculo)
473.275	13	588	100%	0	0	472.679	583	SI	NO	SI	6.329,96	3.106

Mínimo	Pct01	Pct05	Pct10	Pct25	Pct50	Pct75	Pct90	Pct95	Pct99	Max
750,60	837,60	1.519,22	2.222,80	3.743,80	6.386,59	8.727,98	10.295,00	11.053,60	13.543,40	13.912,20

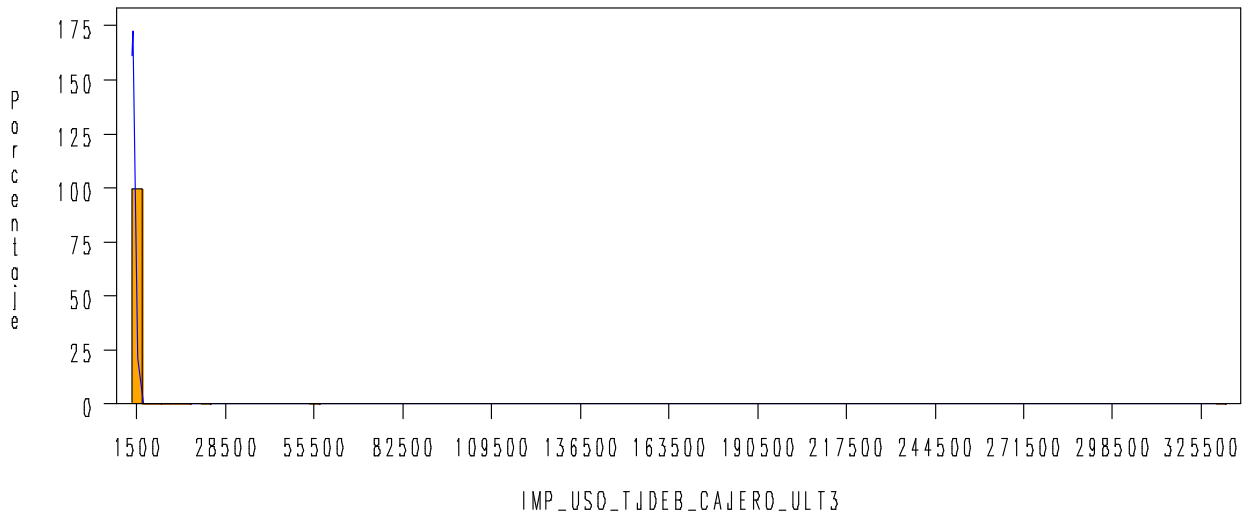
IMP_USO_TJDEB_CAJERO_ULT1: Importe con la tarjeta de débito en cajeros en el último mes



N	Perdidos	Distintos	% Iguales	Infinitos	Negativos	Ceros	Positivos	¿Tiene Moda 0?	¿Tiene Valores atípicos por la izda?	¿Tiene Valores atípicos por la dcha?	Media (si moda es 0 se excluye del cálculo)	Desv.Tipica (si moda es 0 se excluye del cálculo)
473.275	0	15.358	97%	0	0	283.558	189.717	SI	NO	SI	614,57	811

Mínimo	Pct01	Pct05	Pct10	Pct25	Pct50	Pct75	Pct90	Pct95	Pct99	Max
18,23	49,00	95,00	130,60	235,60	455,00	802,20	1.267,00	1.640,80	2.674,78	77.394,60

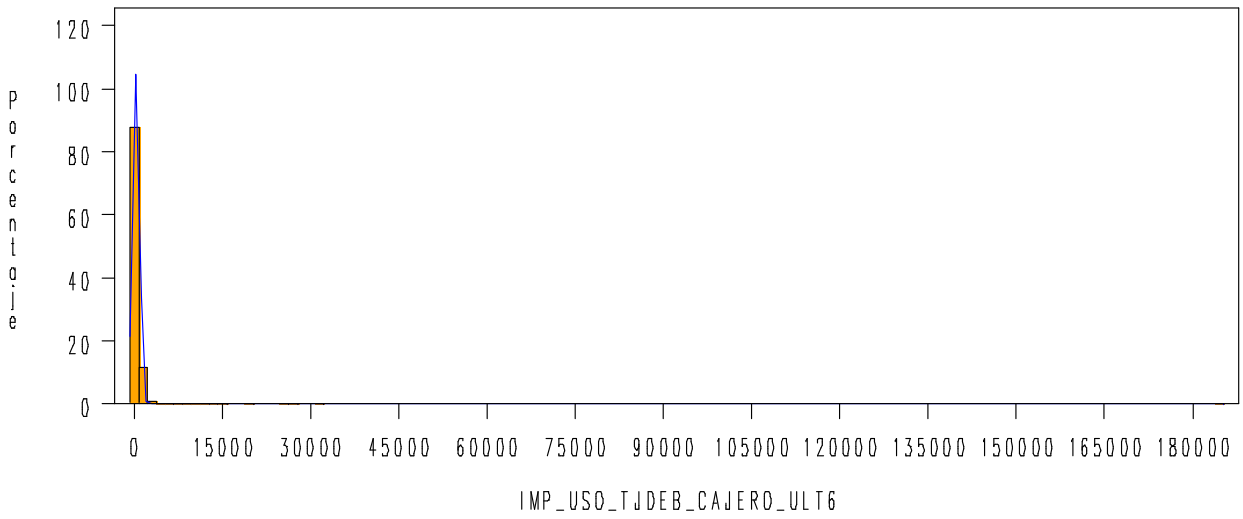
IMP_USO_TJDEB_CAJERO_ULT3: Importe con la tarjeta de débito en cajeros en los últimos 3 meses



N	Perdidos	Distintos	% Iguales	Infinitos	Negativos	Ceros	Positivos	¿Tiene Moda 0?	¿Tiene Valores atípicos por la izda?	¿Tiene Valores atípicos por la dcha?	Media (si moda es 0 se excluye del cálculo)	Desv.Tipica (si moda es 0 se excluye del cálculo)
473.275	0	15.938	97%	0	0	279.961	193.314	SI	NO	SI	630,58	833

Mínimo	Pct01	Pct05	Pct10	Pct25	Pct50	Pct75	Pct90	Pct95	Pct99	Max
19,20	51,40	96,60	133,00	240,20	468,30	823,00	1.298,78	1.684,10	2.747,60	88.235,60

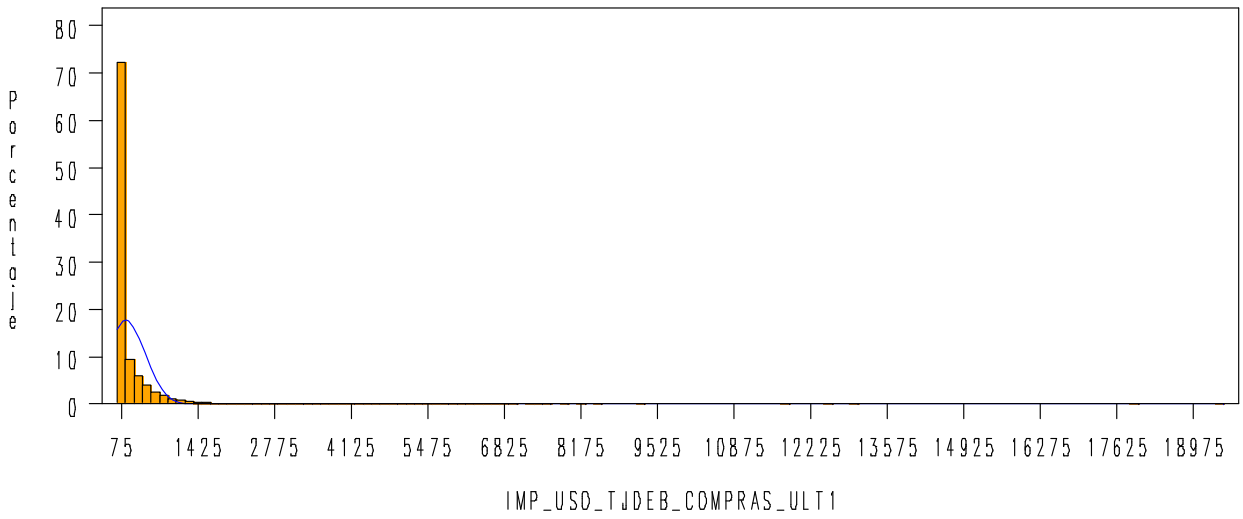
IMP_USO_TJDEB_CAJERO_ULT6: Importe con la tarjeta de débito en cajeros en los últimos 6 meses



N	Perdidos	Distintos	% Iguales	Infinitos	Negativos	Ceros	Positivos	¿Tiene Moda 0?	¿Tiene Valores atípicos por la izda?	¿Tiene Valores atípicos por la dcha?	Media (si moda es 0 se excluye del cálculo)	Desv.Tipica (si moda es 0 se excluye del cálculo)
473.275	0	16.287	97%	0	0	278.358	194.917	SI	NO	SI	648,53	713

Mínimo	Pct01	Pct05	Pct10	Pct25	Pct50	Pct75	Pct90	Pct95	Pct99	Max
15,96	52,60	99,20	136,40	249,40	486,50	850,20	1.330,60	1.736,71	2.807,80	56.708,40

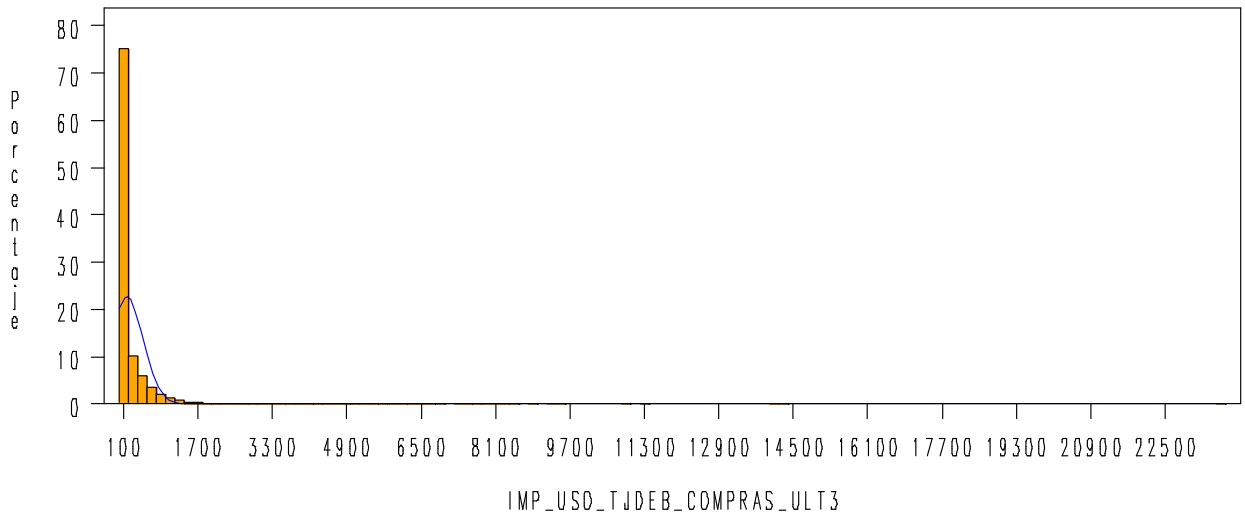
IMP_USO_TJDEB_COMPRAS_ULT1: Importe con la tarjeta de débito en compras en el último mes



N	Perdidos	Distintos	% Iguales	Infinitos	Negativos	Ceros	Positivos	¿Tiene Moda 0?	¿Tiene Valores atípicos por la izda?	¿Tiene Valores atípicos por la dcha?	Media (si moda es 0 se excluye del cálculo)	Desv.Tipica (si moda es 0 se excluye del cálculo)
473.275	0	137.365	71%	0	0	295.949	177.326	SI	NO	SI	422,19	435

Mínimo	Pct01	Pct05	Pct10	Pct25	Pct50	Pct75	Pct90	Pct95	Pct99	Max
1,82	27,25	59,51	84,10	146,19	293,73	550,19	905,05	1.197,44	2.027,95	14.264,99

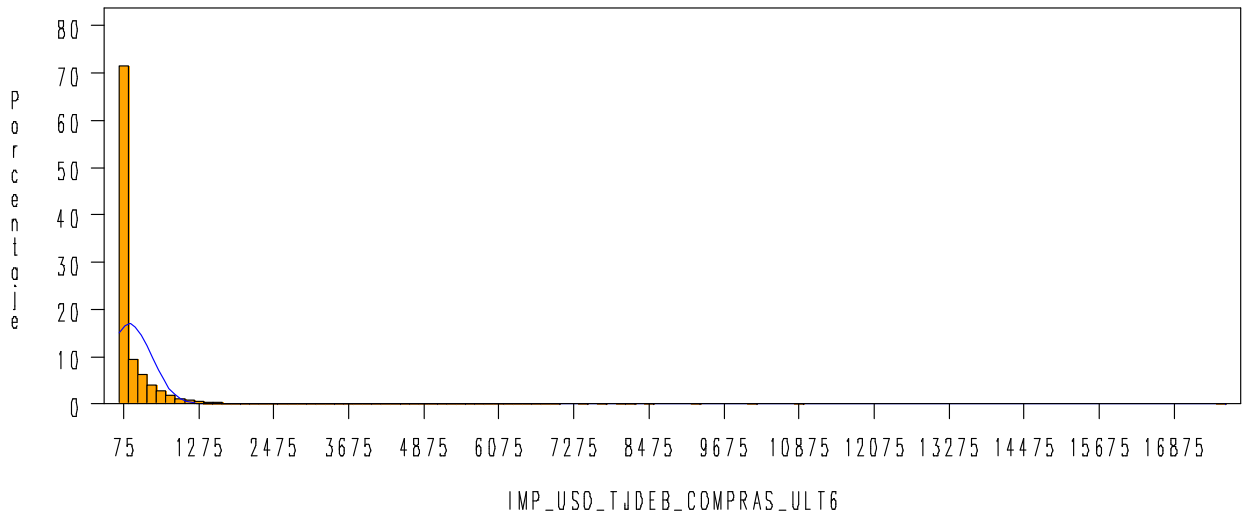
IMP_USO_TJDEB_COMPRAS_ULT3: Importe con la tarjeta de débito en compras en los últimos 3 meses



N	Perdidos	Distintos	% Iguales	Infinitos	Negativos	Ceros	Positivos	¿Tiene Moda 0?	¿Tiene Valores atípicos por la izda?	¿Tiene Valores atípicos por la dcha?	Media (si moda es 0 se excluye del cálculo)	Desv.Tipica (si moda es 0 se excluye del cálculo)
473.275	0	140.491	70%	0	0	292.995	180.280	SI	NO	SI	434,73	452

Mínimo	Pct01	Pct05	Pct10	Pct25	Pct50	Pct75	Pct90	Pct95	Pct99	Max
2,39	27,77	60,68	86,24	149,68	300,37	566,36	935,64	1.238,83	2.101,76	14.446,64

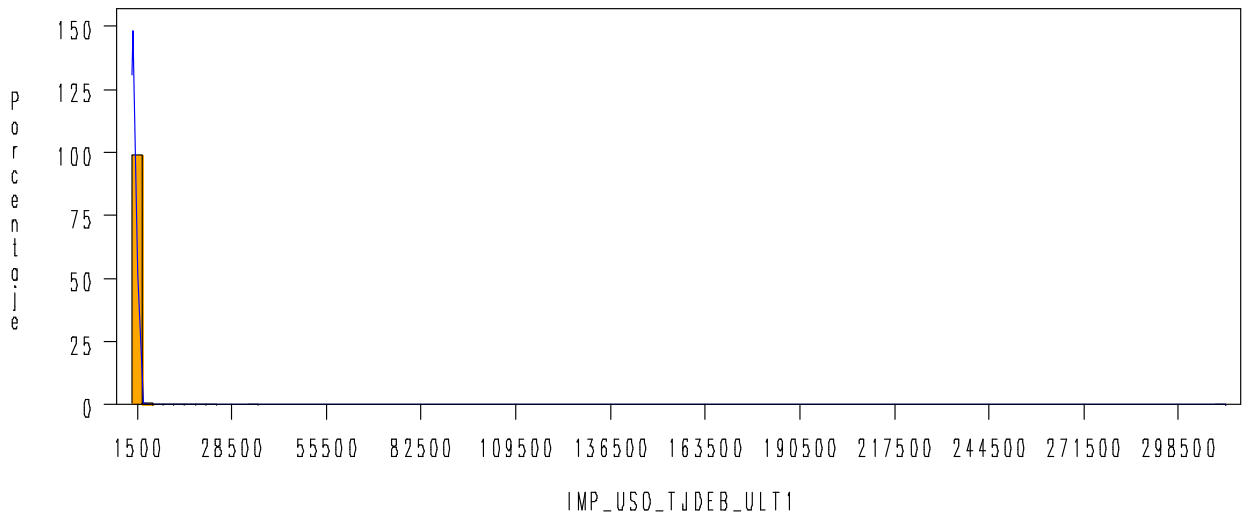
IMP_USO_TJDEB_COMPRAS_ULT6: Importe con la tarjeta de débito en compras en los últimos 6 meses



N	Perdidos	Distintos	% Iguales	Infinitos	Negativos	Ceros	Positivos	¿Tiene Moda 0?	¿Tiene Valores atípicos por la izda?	¿Tiene Valores atípicos por la dcha?	Media (si moda es 0 se excluye del cálculo)	Desv.Tipica (si moda es 0 se excluye del cálculo)
473.275	0	140.839	70%	0	0	292.083	181.192	SI	NO	SI	435,82	454

Mínimo	Pct01	Pct05	Pct10	Pct25	Pct50	Pct75	Pct90	Pct95	Pct99	Max
1,95	27,97	60,81	86,26	148,51	297,66	564,38	942,65	1.260,22	2.148,37	10.849,20

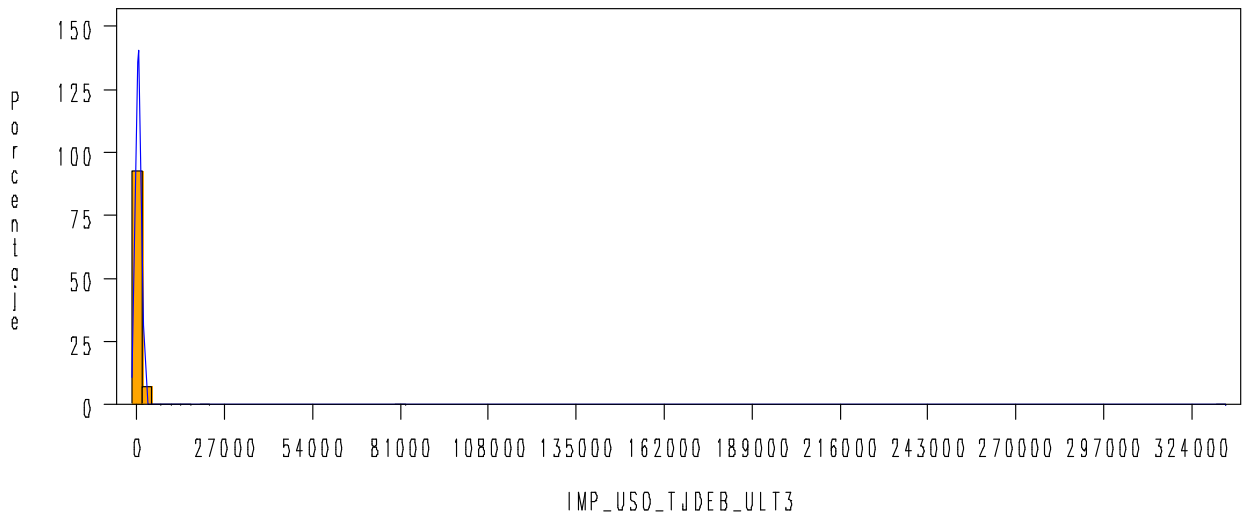
IMP_USO_TJDEB_ULT1: Importe con la tarjeta de débito en el último mes



N	Perdidos	Distintos	% Iguales	Infinitos	Negativos	Ceros	Positivos	¿Tiene Moda 0?	¿Tiene Valores atípicos por la izda?	¿Tiene Valores atípicos por la dcha?	Media (si moda es 0 se excluye del cálculo)	Desv.Tipica (si moda es 0 se excluye del cálculo)
473.275	0	161.211	66%	0	0	245.282	227.993	SI	NO	SI	840,39	929

Mínimo	Pct01	Pct05	Pct10	Pct25	Pct50	Pct75	Pct90	Pct95	Pct99	Max
2,71	45,41	99,26	151,47	332,64	664,70	1.125,27	1.703,25	2.156,77	3.422,63	80.224,64

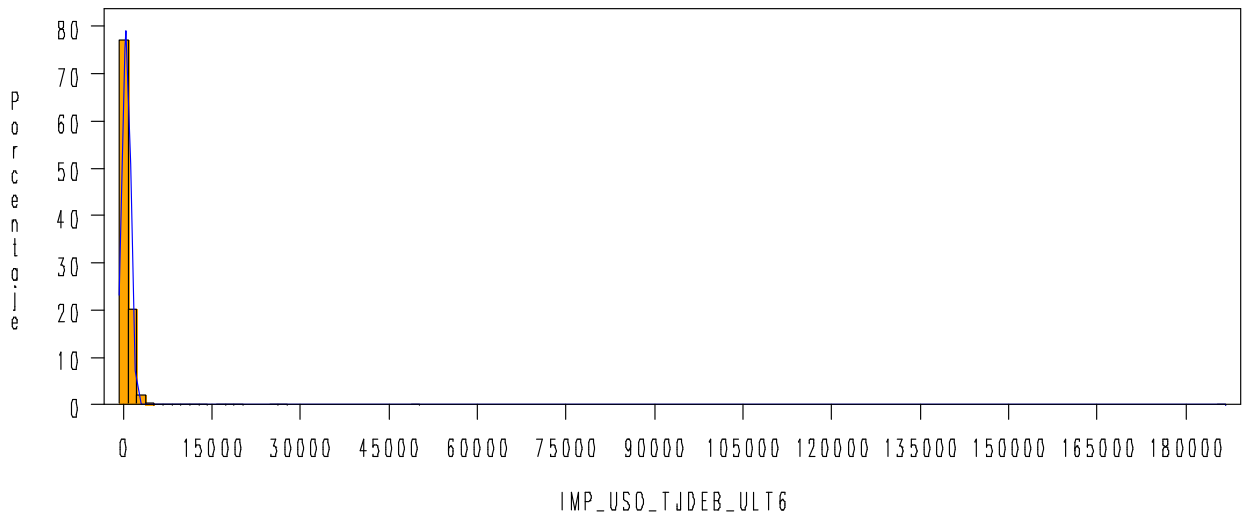
IMP_USO_TJDEB_ULT3: Importe con la tarjeta de débito en los últimos 3 meses



N	Perdidos	Distintos	% Iguales	Infinitos	Negativos	Ceros	Positivos	¿Tiene Moda 0?	¿Tiene Valores atípicos por la izda?	¿Tiene Valores atípicos por la dcha?	Media (si moda es 0 se excluye del cálculo)	Desv.Tipica (si moda es 0 se excluye del cálculo)
473.275	0	164.469	65%	0	0	242.417	230.858	SI	NO	SI	867,86	961

Mínimo	Pct01	Pct05	Pct10	Pct25	Pct50	Pct75	Pct90	Pct95	Pct99	Max
2,94	46,03	102,31	156,05	342,96	685,56	1.162,61	1.758,39	2.237,26	3.531,58	93.447,26

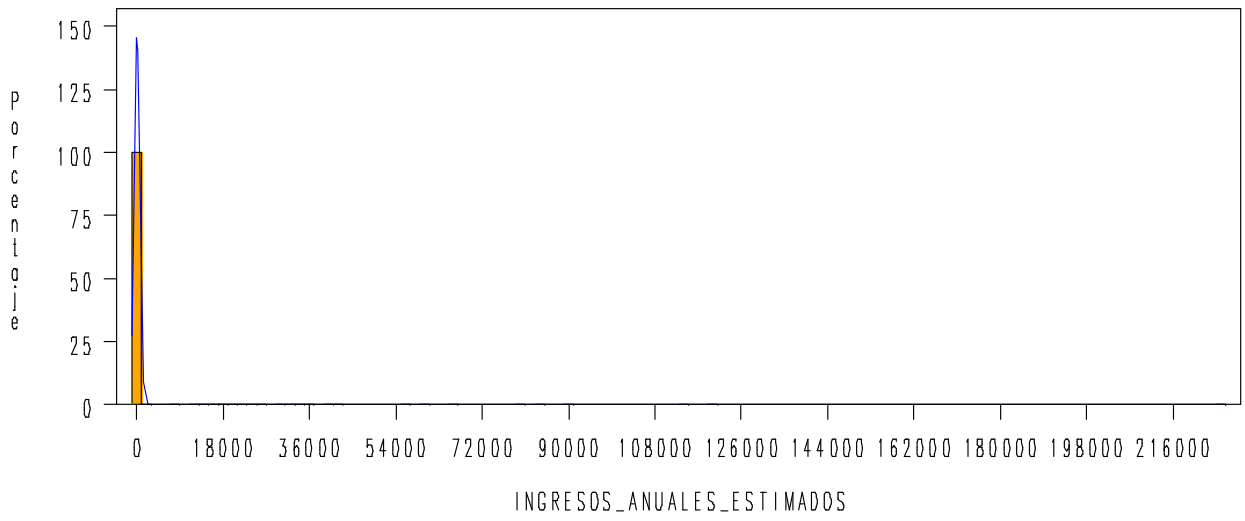
IMP_USO_TJDEB_ULT6: Importe con la tarjeta de débito en los últimos 6 meses



N	Perdidos	Distintos	% Iguales	Infinitos	Negativos	Ceros	Positivos	¿Tiene Moda 0?	¿Tiene Valores atípicos por la izda?	¿Tiene Valores atípicos por la dcha?	Media (si moda es 0 se excluye del cálculo)	Desv.Tipica (si moda es 0 se excluye del cálculo)
473.275	0	165.965	65%	0	0	241.450	231.825	SI	NO	SI	886,08	865

Mínimo	Pct01	Pct05	Pct10	Pct25	Pct50	Pct75	Pct90	Pct95	Pct99	Max
2,95	47,32	103,50	158,15	349,23	697,53	1.187,94	1.802,12	2.293,72	3.658,81	61.290,18

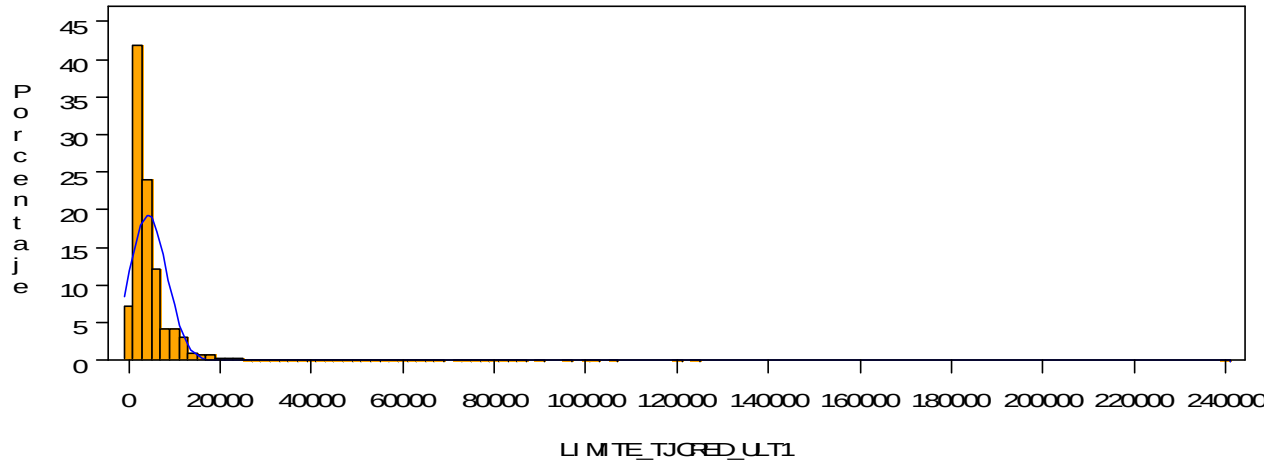
INGRESOS_ANUALES_ESTIMADOS: Ingresos anuales estimados



N	Perdidos	Distintos	% Iguales	Infinitos	Negativos	Ceros	Positivos	¿Tiene Moda 0?	¿Tiene Valores atípicos por la izda?	¿Tiene Valores atípicos por la dcha?	Media (si moda es 0 se excluye del cálculo)	Desv.Tipica (si moda es 0 se excluye del cálculo)
473.275	0	49	100%	0	0	473.231	44	SI	NO	SI	38.094,16	33.064

Mínimo	Pct01	Pct05	Pct10	Pct25	Pct50	Pct75	Pct90	Pct95	Pct99	Max
9.015,61	9.015,61	9.015,61	10.748,11	17.815,11	28.936,10	40.779,37	88.870,52	110.666,08	110.666,08	110.666,08

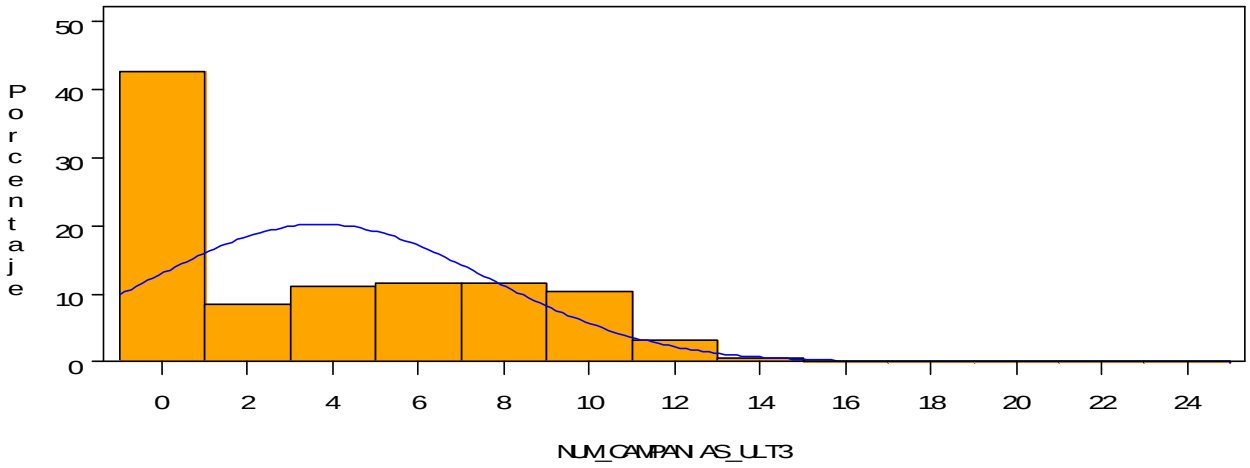
LIMITE_TJCRED_ULT1: Límite con la tarjeta de crédito en el último mes



N	Perdidos	Distintos	% Iguales	Infinitos	Negativos	Ceros	Positivos	¿Tiene Moda 0?	¿Tiene Valores atípicos por la izda?	¿Tiene Valores atípicos por la dcha?	Media (si moda es 0 se excluye del cálculo)	Desv.Tipica (si moda es 0 se excluye del cálculo)
473.275	13	72.169	85%	0	0	375	472.887	SI	NO	SI	4.279,09	4.128

Mínimo	Pct01	Pct05	Pct10	Pct25	Pct50	Pct75	Pct90	Pct95	Pct99	Max
104,00	439,40	834,80	1.149,40	1.769,80	3.050,20	5.411,00	9.335,60	12.385,80	19.977,40	138.265,20

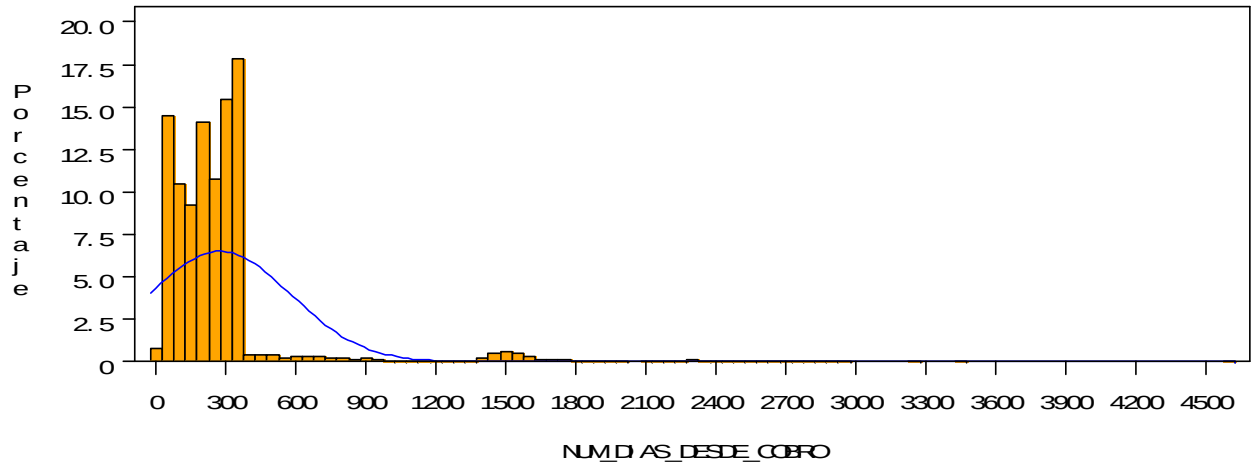
NUM_CAMPANIAS_ULT3: Número de campañas en los últimos 3 meses



N	Perdidos	Distintos	% Iguales	Infinitos	Negativos	Ceros	Positivos	¿Tiene Moda 0?	¿Tiene Valores atípicos por la izda?	¿Tiene Valores atípicos por la dcha?	Media (si moda es 0 se excluye del cálculo)	Desv.Tipica (si moda es 0 se excluye del cálculo)
473.275	0	439	100%	0	0	195.337	277.938	SI	NO	NO	6,28	3

Mínimo	Pct01	Pct05	Pct10	Pct25	Pct50	Pct75	Pct90	Pct95	Pct99	Max
0,33	0,33	1,33	2,13	3,73	6,33	8,80	10,40	11,20	13,07	20,47

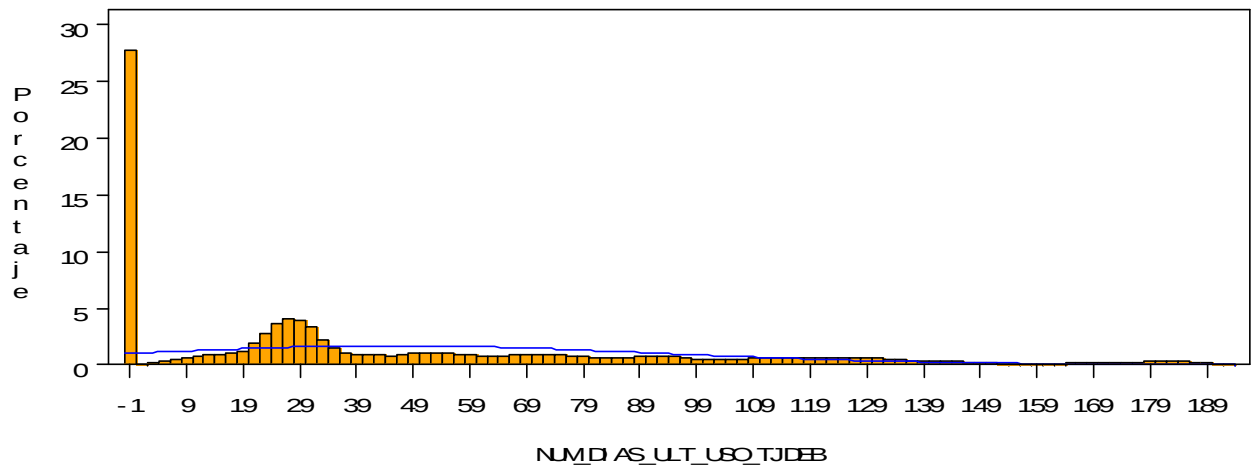
NUM_DIAS_DESDE_COBRO: Número de días desde el cobro



N	Perdidos	Distintos	% Iguales	Infinitos	Negativos	Ceros	Positivos	¿Tiene Moda 0?	¿Tiene Valores atípicos por la izda?	¿Tiene Valores atípicos por la dcha?	Media (si moda es 0 se excluye del cálculo)	Desv.Tipica (si moda es 0 se excluye del cálculo)
473.275	0	5.992	99%	0	0	225	473.050	NO	NO	SI	298,68	278

Mínimo	Pct01	Pct05	Pct10	Pct25	Pct50	Pct75	Pct90	Pct95	Pct99	Max
1,00	26,00	34,40	63,40	153,80	281,60	394,40	400,00	431,00	1.584,20	3.647,20

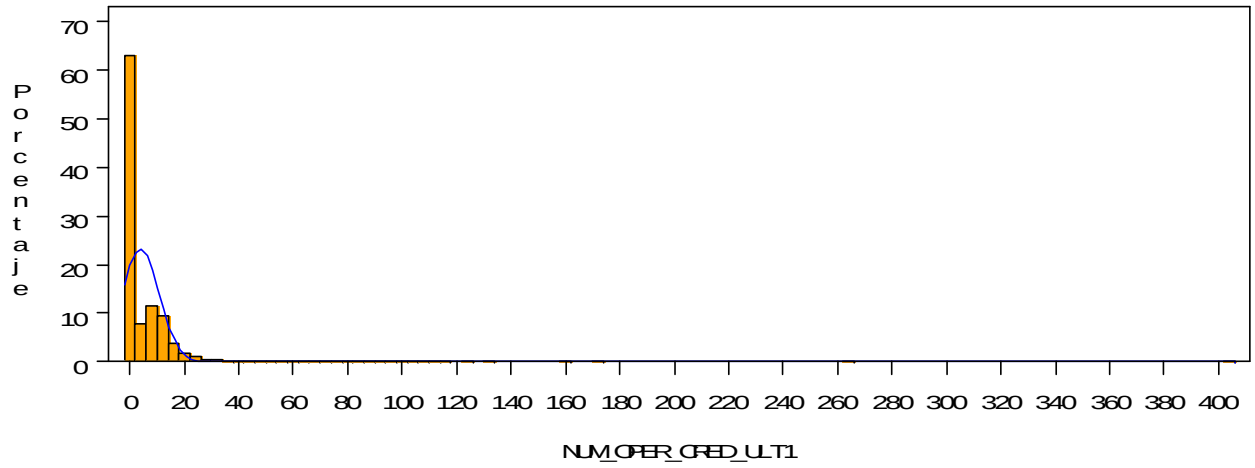
NUM_DIAS_ULT_USO_TJDEB: Número de días desde el último uso de la tarjeta de débito



N	Perdidos	Distintos	% Iguales	Infinitos	Negativos	Ceros	Positivos	¿Tiene Moda 0?	¿Tiene Valores atípicos por la izda?	¿Tiene Valores atípicos por la dcha?	Media (si moda es 0 se excluye del cálculo)	Desv.Tipica (si moda es 0 se excluye del cálculo)
473.275	0	965	100%	0	131.535	50	341.690	NO	NO	NO	45,54	48

Mínimo	Pct01	Pct05	Pct10	Pct25	Pct50	Pct75	Pct90	Pct95	Pct99	Max
-1,00	-1,00	-1,00	-1,00	-1,00	29,00	74,00	120,80	139,80	182,00	191,00

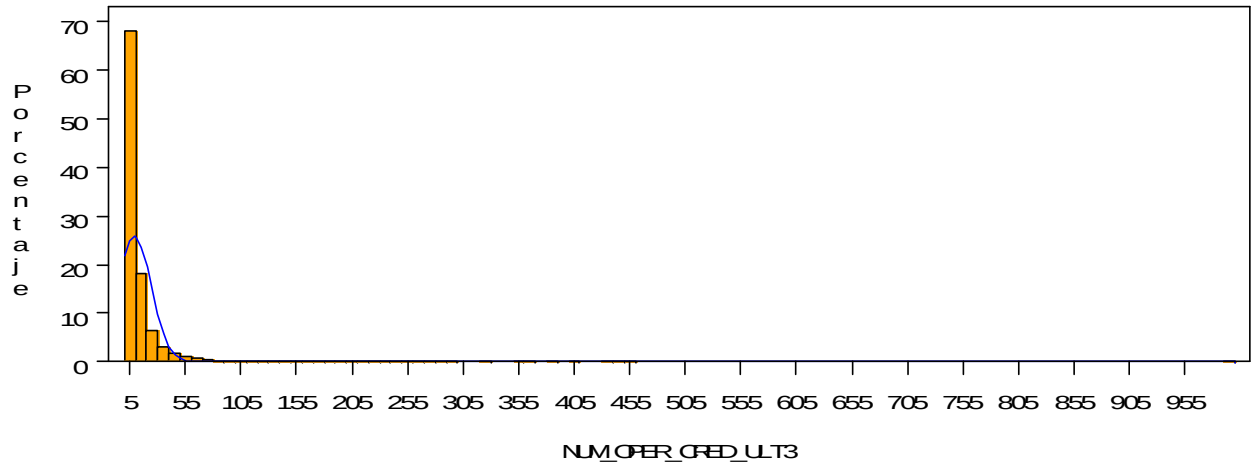
NUM_OPER_CRED_ULT1: Número de operaciones a crédito en el último mes



N	Perdidos	Distintos	% Iguales	Infinitos	Negativos	Ceros	Positivos	¿Tiene Moda 0?	¿Tiene Valores atípicos por la izda?	¿Tiene Valores atípicos por la dcha?	Media (si moda es 0 se excluye del cálculo)	Desv.Tipica (si moda es 0 se excluye del cálculo)
473.275	13	418	100%	0	0	296.271	176.991	SI	NO	SI	10,56	7

Mínimo	Pct01	Pct05	Pct10	Pct25	Pct50	Pct75	Pct90	Pct95	Pct99	Max
1,00	1,00	2,80	3,80	6,00	9,00	12,80	18,80	24,00	38,00	213,00

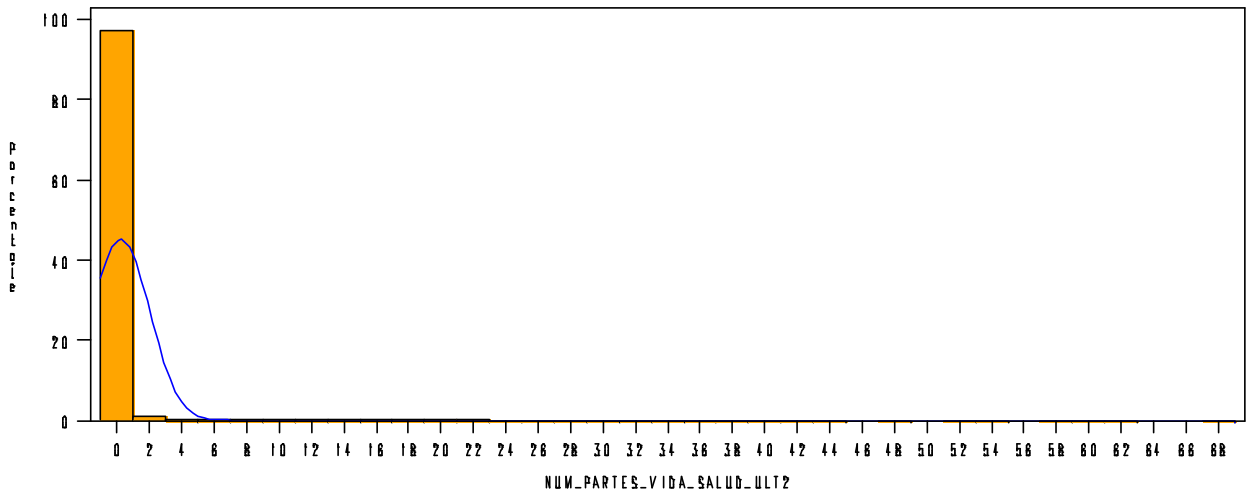
NUM_OPER_CRED_ULT3: Número de operaciones a crédito en los últimos 3 meses



N	Perdidos	Distintos	% Iguales	Infinitos	Negativos	Ceros	Positivos	¿Tiene Moda 0?	¿Tiene Valores atípicos por la izda?	¿Tiene Valores atípicos por la dcha?	Media (si moda es 0 se excluye del cálculo)	Desv.Tipica (si moda es 0 se excluye del cálculo)
473.275	13	952	100%	0	0	246.187	227.075	SI	NO	SI	18,17	18

Mínimo	Pct01	Pct05	Pct10	Pct25	Pct50	Pct75	Pct90	Pct95	Pct99	Max
1,00	2,00	3,40	5,00	8,00	12,40	21,60	37,80	52,00	90,00	537,60

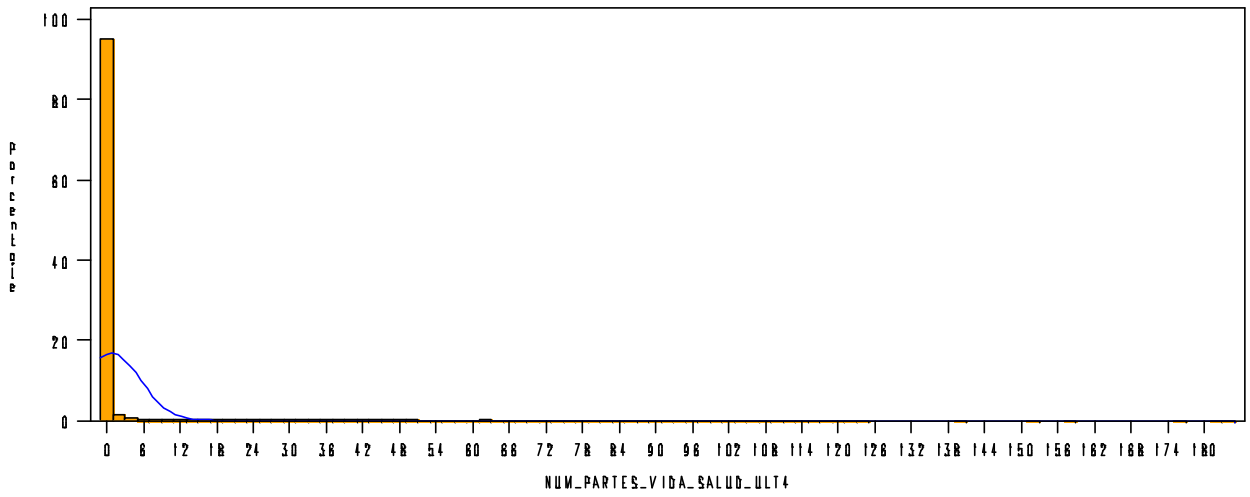
NUM_PARTES_VIDA_SALUD_ULT2: Número de partes de vida y salud en los últimos 2 meses



N	Perdidos	Distintos	% Iguales	Infinitos	Negativos	Ceros	Positivos	¿Tiene Moda 0?	¿Tiene Valores atípicos por la izda?	¿Tiene Valores atípicos por la dcha?	Media (si moda es 0 se excluye del cálculo)	Desv.Tipica (si moda es 0 se excluye del cálculo)
473.275	0	189	100%	0	0	459.505	13.770	SI	NO	SI	7,70	7

Mínimo	Pct01	Pct05	Pct10	Pct25	Pct50	Pct75	Pct90	Pct95	Pct99	Max
1,00	1,00	1,00	1,00	1,80	5,00	12,80	18,80	20,60	23,30	51,60

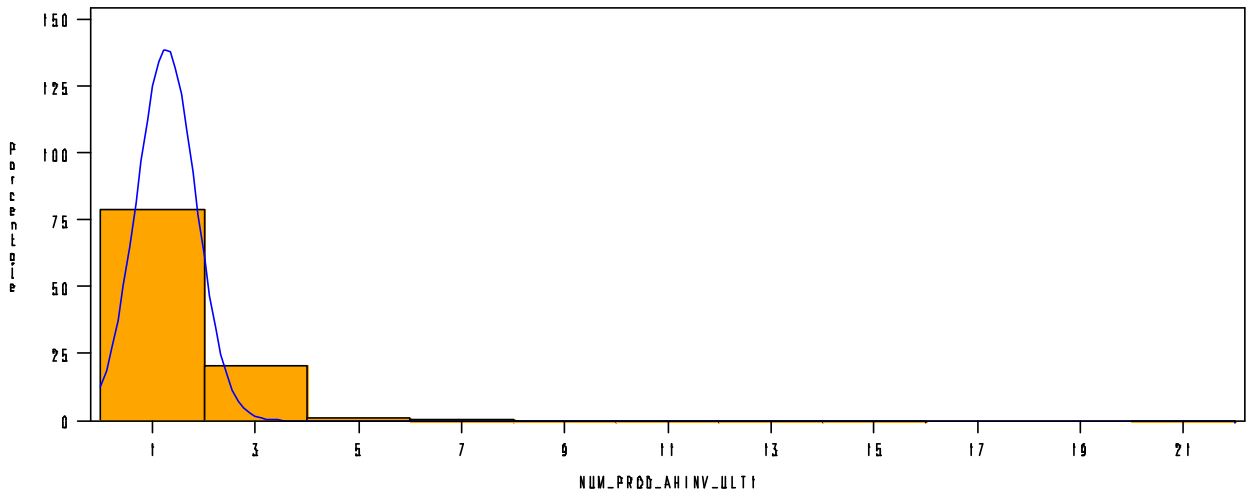
NUM_PARTES_VIDA_SALUD_ULT4: Número de partes de vida y salud en los últimos 4 meses



N	Perdidos	Distintos	% Iguales	Infinitos	Negativos	Ceros	Positivos	¿Tiene Moda 0?	¿Tiene Valores atípicos por la izda?	¿Tiene Valores atípicos por la dcha?	Media (si moda es 0 se excluye del cálculo)	Desv.Tipica (si moda es 0 se excluye del cálculo)
473.275	0	426	100%	0	0	449.789	23.486	SI	NO	SI	13,48	17

Mínimo	Pct01	Pct05	Pct10	Pct25	Pct50	Pct75	Pct90	Pct95	Pct99	Max
1,00	1,00	1,00	1,00	1,80	6,00	18,60	39,70	51,70	62,40	149,00

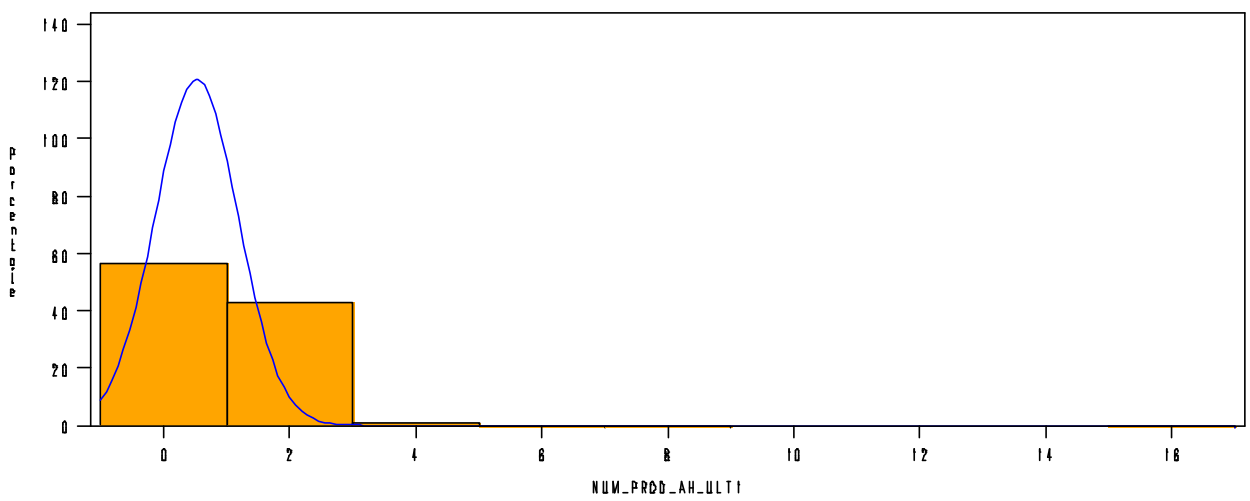
NUM_PROD_AHINV_ULT1: Número de productos de ahorro e inversión en el último mes



N	Perdidos	Distintos	% Iguales	Infinitos	Negativos	Ceros	Positivos	¿Tiene Moda 0?	¿Tiene Valores atípicos por la izda?	¿Tiene Valores atípicos por la dcha?	Media (si moda es 0 se excluye del cálculo)	Desv.Tipica (si moda es 0 se excluye del cálculo)
473.275	13	54	100%	0	0	0	473.262	NO	NO	SI	1,26	1

Mínimo	Pct01	Pct05	Pct10	Pct25	Pct50	Pct75	Pct90	Pct95	Pct99	Max
1,00	1,00	1,00	1,00	1,00	1,00	1,00	2,00	2,00	3,40	13,80

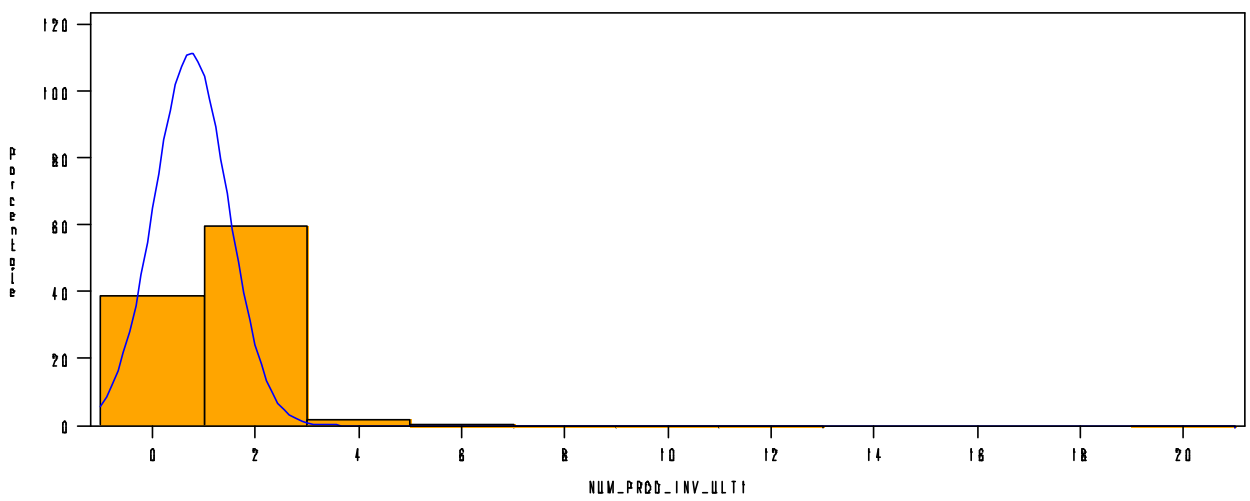
NUM_PROD_AH_ULT1: Número de productos de ahorro



N	Perdidos	Distintos	% Iguales	Infinitos	Negativos	Ceros	Positivos	¿Tiene Moda 0?	¿Tiene Valores atípicos por la izda?	¿Tiene Valores atípicos por la dcha?	Media (si moda es 0 se excluye del cálculo)	Desv.Tipica (si moda es 0 se excluye del cálculo)
473.275	13	41	100%	0	0	266.052	207.210	SI	NO	SI	1,18	0

Mínimo	Pct01	Pct05	Pct10	Pct25	Pct50	Pct75	Pct90	Pct95	Pct99	Max
1,00	1,00	1,00	1,00	1,00	1,00	1,00	2,00	2,00	3,00	8,80

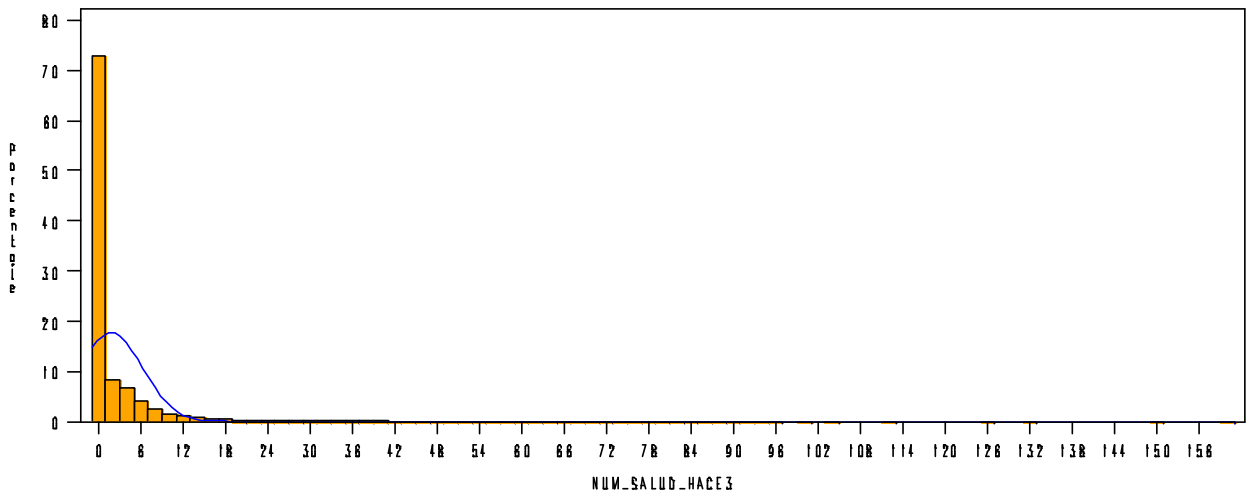
NUM_PROD_INV_ULT1: Número de productos de inversión



N	Perdidos	Distintos	% Iguales	Infinitos	Negativos	Ceros	Positivos	¿Tiene Moda 0?	¿Tiene Valores atípicos por la izda?	¿Tiene Valores atípicos por la dcha?	Media (si moda es 0 se excluye del cálculo)	Desv.Tipica (si moda es 0 se excluye del cálculo)
473.275	13	52	100%	0	0	183.855	289.407	NO	NO	SI	0,74	1

Mínimo	Pct01	Pct05	Pct10	Pct25	Pct50	Pct75	Pct90	Pct95	Pct99	Max
0,00	0,00	0,00	0,00	0,00	1,00	1,00	2,00	2,00	3,00	12,20

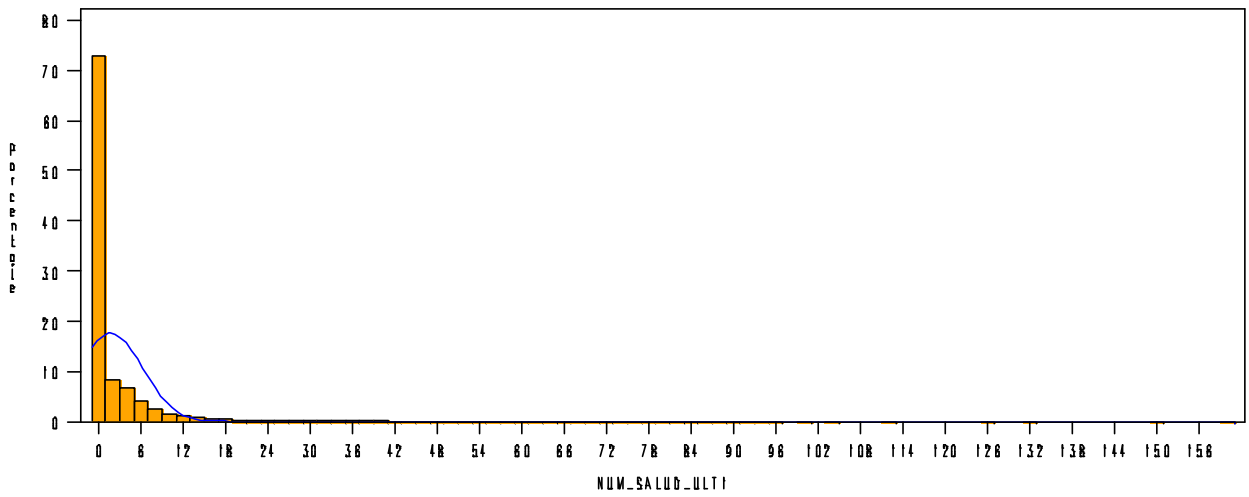
NUM_SALUD_HACE3: Número de productos de salud hace 3 meses



N	Perdidos	Distintos	% Iguales	Infinitos	Negativos	Ceros	Positivos	¿Tiene Moda 0?	¿Tiene Valores atípicos por la izda?	¿Tiene Valores atípicos por la dcha?	Media (si moda es 0 se excluye del cálculo)	Desv.Tipica (si moda es 0 se excluye del cálculo)
473.275	0	391	100%	0	0	344.666	128.609	SI	NO	SI	6,35	7

Mínimo	Pct01	Pct05	Pct10	Pct25	Pct50	Pct75	Pct90	Pct95	Pct99	Max
1,00	2,00	2,00	2,00	2,00	4,00	8,00	13,40	18,60	34,00	127,00

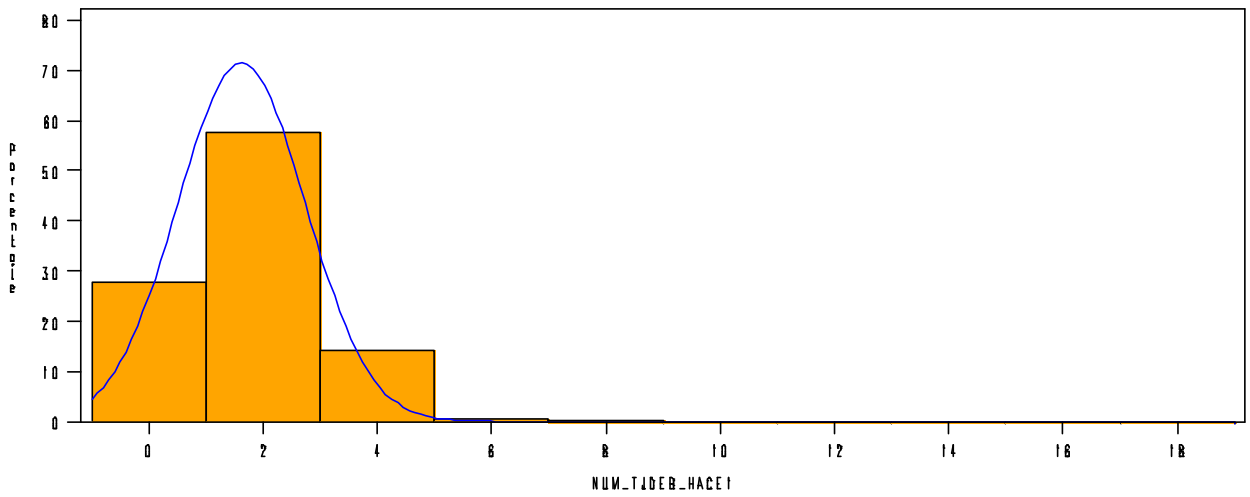
NUM_SALUD_ULT1: Número de productos de salud en el último mes



N	Perdidos	Distintos	% Iguales	Infinitos	Negativos	Ceros	Positivos	¿Tiene Moda 0?	¿Tiene Valores atípicos por la izda?	¿Tiene Valores atípicos por la dcha?	Media (si moda es 0 se excluye del cálculo)	Desv.Tipica (si moda es 0 se excluye del cálculo)
473.275	0	390	100%	0	0	344.528	128.747	SI	NO	SI	6,36	7

Mínimo	Pct01	Pct05	Pct10	Pct25	Pct50	Pct75	Pct90	Pct95	Pct99	Max
1,00	2,00	2,00	2,00	2,00	4,00	8,00	13,40	18,80	34,20	127,00

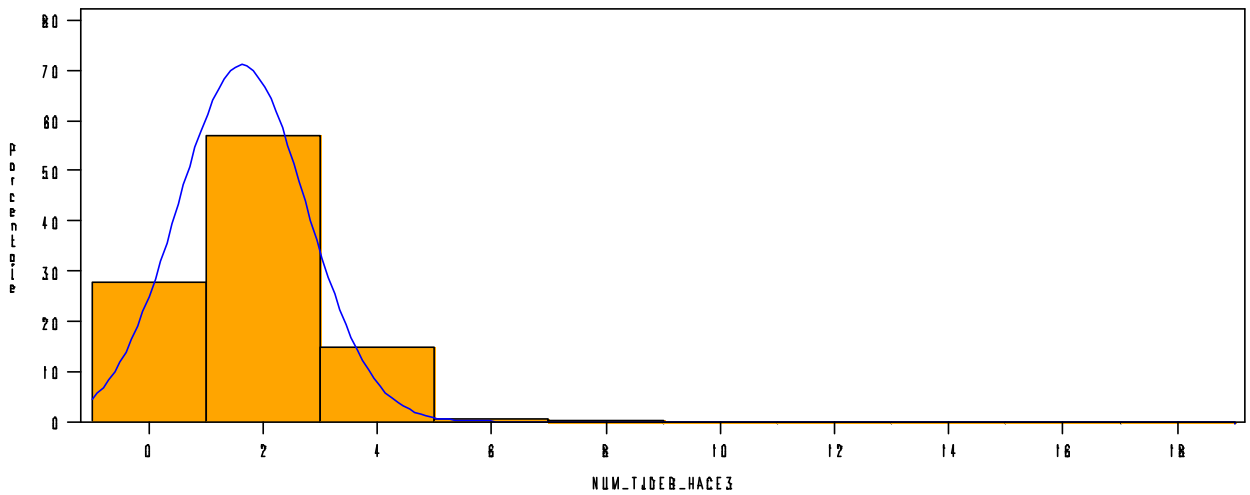
NUM_TJDEB_HACE1: Número de tarjetas de débito en el último mes



N	Perdidos	Distintos	% Iguales	Infinitos	Negativos	Ceros	Positivos	¿Tiene Moda 0?	¿Tiene Valores atípicos por la izda?	¿Tiene Valores atípicos por la dcha?	Media (si moda es 0 se excluye del cálculo)	Desv.Tipica (si moda es 0 se excluye del cálculo)
473.275	0	70	100%	0	0	131.535	341.740	NO	NO	SI	1,62	1

Mínimo	Pct01	Pct05	Pct10	Pct25	Pct50	Pct75	Pct90	Pct95	Pct99	Max
0,00	0,00	0,00	0,00	0,00	2,00	2,00	3,00	3,00	4,00	15,00

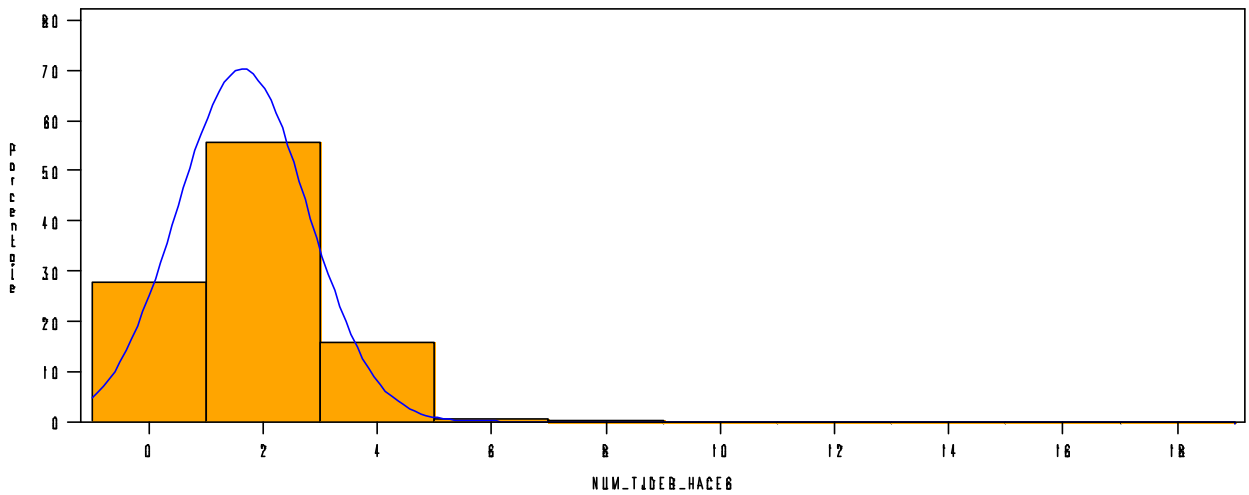
NUM_TJDEB_HACE3: Número de tarjetas de débito hace 3 meses



N	Perdidos	Distintos	% Iguales	Infinitos	Negativos	Ceros	Positivos	¿Tiene Moda 0?	¿Tiene Valores atípicos por la izda?	¿Tiene Valores atípicos por la dcha?	Media (si moda es 0 se excluye del cálculo)	Desv.Tipica (si moda es 0 se excluye del cálculo)
473.275	0	66	100%	0	0	131.200	342.075	NO	NO	SI	1,63	1

Mínimo	Pct01	Pct05	Pct10	Pct25	Pct50	Pct75	Pct90	Pct95	Pct99	Max
0,00	0,00	0,00	0,00	0,00	2,00	2,00	3,00	3,00	4,00	14,40

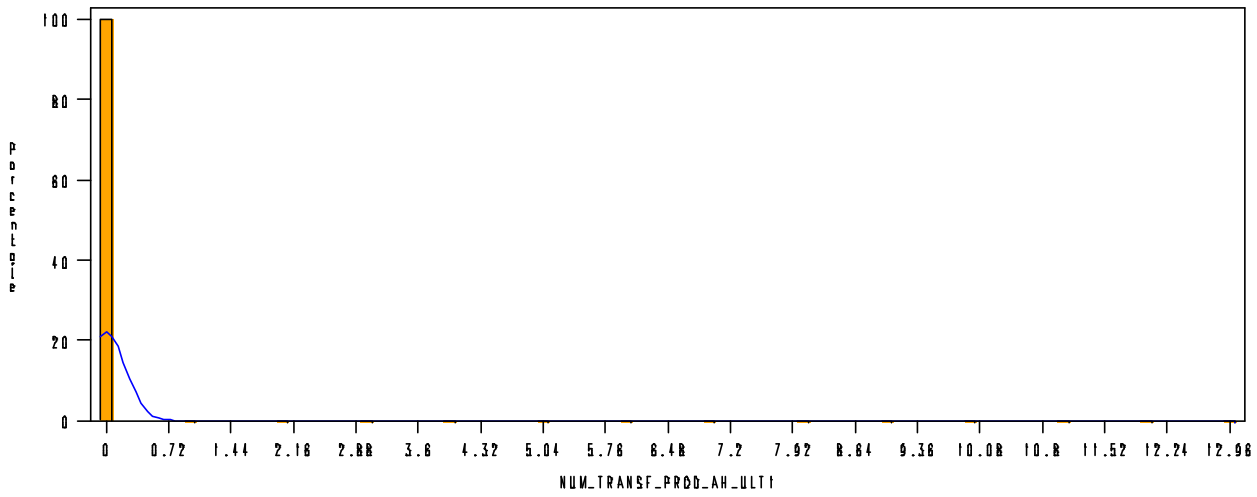
NUM_TJDEB_HACE6: Número de tarjetas de débito hace 6 meses



N	Perdidos	Distintos	% Iguales	Infinitos	Negativos	Ceros	Positivos	¿Tiene Moda 0?	¿Tiene Valores atípicos por la izda?	¿Tiene Valores atípicos por la dcha?	Media (si moda es 0 se excluye del cálculo)	Desv.Tipica (si moda es 0 se excluye del cálculo)
473.275	0	63	100%	0	0	132.025	341.250	NO	NO	SI	1,64	1

Mínimo	Pct01	Pct05	Pct10	Pct25	Pct50	Pct75	Pct90	Pct95	Pct99	Max
0,00	0,00	0,00	0,00	0,00	2,00	2,00	3,00	3,00	4,00	14,00

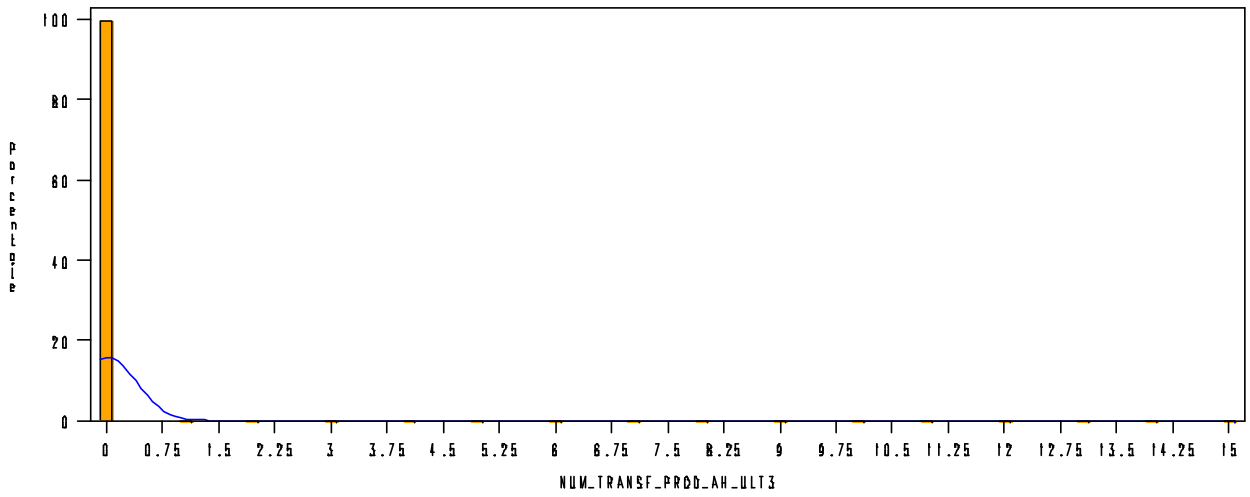
NUM_TRANSF_PROD_AH_ULT1: Número de transferencias en productos de ahorro en el último mes



N	Perdidos	Distintos	% Iguales	Infinitos	Negativos	Ceros	Positivos	¿Tiene Moda 0?	¿Tiene Valores atípicos por la izda?	¿Tiene Valores atípicos por la dcha?	Media (si moda es 0 se excluye del cálculo)	Desv.Tipica (si moda es 0 se excluye del cálculo)
473.275	13	65	100%	0	0	472.762	500	SI	NO	SI	6,05	3

Mínimo	Pct01	Pct05	Pct10	Pct25	Pct50	Pct75	Pct90	Pct95	Pct99	Max
1,00	1,00	1,50	2,10	3,80	6,10	8,30	10,00	10,60	11,70	12,20

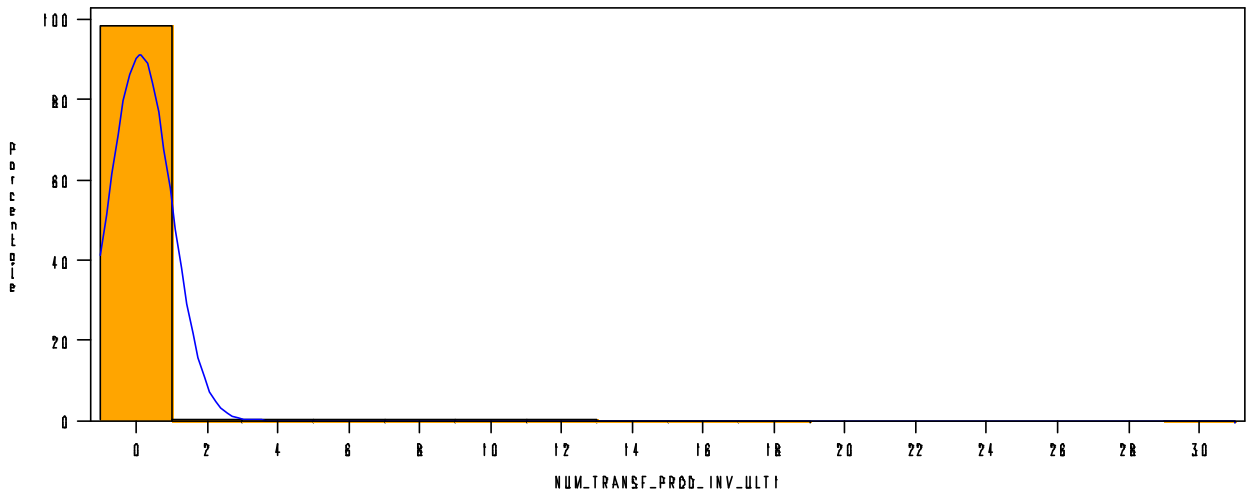
NUM_TRANSF_PROD_AH_ULT3: Número de transferencias en productos de ahorro en los últimos 3 meses



N	Perdidos	Distintos	% Iguales	Infinitos	Negativos	Ceros	Positivos	¿Tiene Moda 0?	¿Tiene Valores atípicos por la izda?	¿Tiene Valores atípicos por la dcha?	Media (si moda es 0 se excluye del cálculo)	Desv.Tipica (si moda es 0 se excluye del cálculo)
473.275	13	73	100%	0	0	471.786	1.476	SI	NO	SI	6,17	3

Mínimo	Pct01	Pct05	Pct10	Pct25	Pct50	Pct75	Pct90	Pct95	Pct99	Max
1,00	1,00	1,60	2,00	3,70	6,20	8,70	10,00	11,00	12,20	13,60

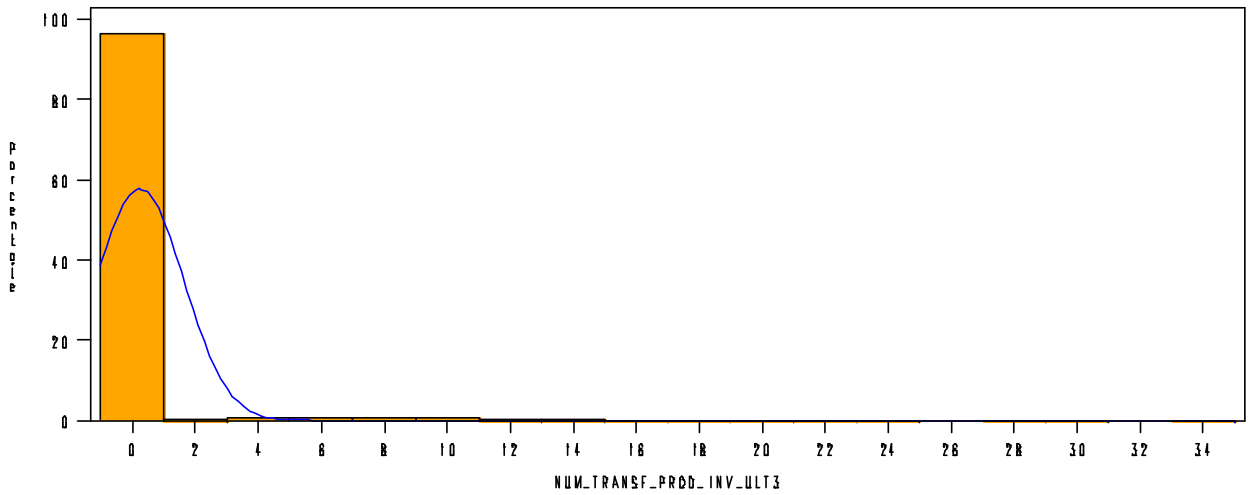
NUM_TRANSF_PROD_INV_ULT1: Número de transferencias en productos de inversión en el último mes



N	Perdidos	Distintos	% Iguales	Infinitos	Negativos	Ceros	Positivos	¿Tiene Moda 0?	¿Tiene Valores atípicos por la izda?	¿Tiene Valores atípicos por la dcha?	Media (si moda es 0 se excluye del cálculo)	Desv.Tipica (si moda es 0 se excluye del cálculo)
473.275	13	78	100%	0	0	465.616	7.646	SI	NO	SI	6,25	3

Mínimo	Pct01	Pct05	Pct10	Pct25	Pct50	Pct75	Pct90	Pct95	Pct99	Max
1,00	1,00	2,00	2,00	4,00	6,00	9,00	10,00	11,00	12,00	17,60

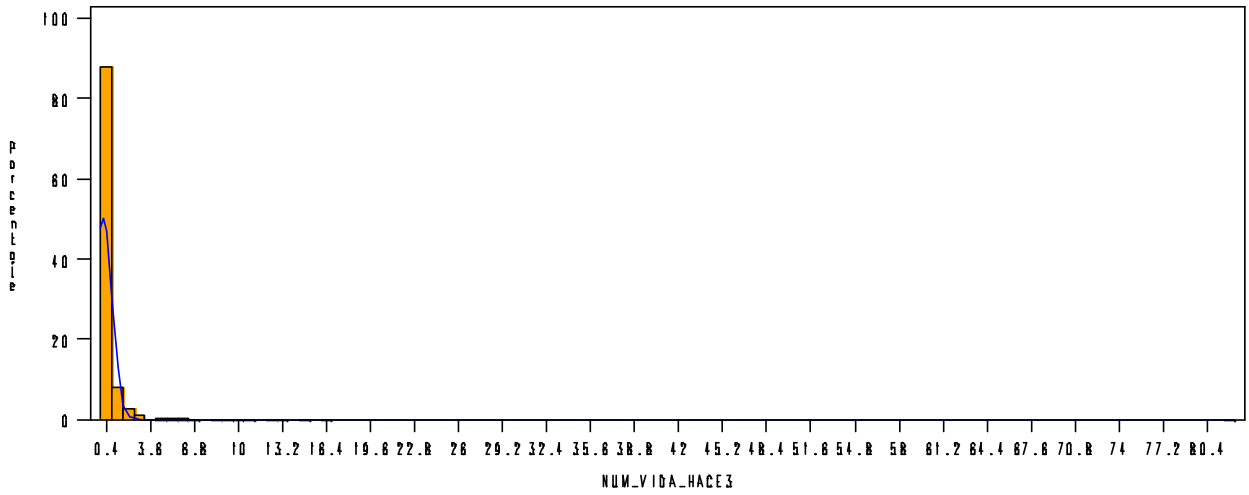
NUM_TRANSF_PROD_INV_ULT3: Número de transferencias en productos de inversión en últimos 3 meses



N	Perdidos	Distintos	% Iguales	Infinitos	Negativos	Ceros	Positivos	¿Tiene Moda 0?	¿Tiene Valores atípicos por la izda?	¿Tiene Valores atípicos por la dcha?	Media (si moda es 0 se excluye del cálculo)	Desv.Tipica (si moda es 0 se excluye del cálculo)
473.275	13	112	100%	0	0	456.355	16.907	SI	NO	SI	6,71	3

Mínimo	Pct01	Pct05	Pct10	Pct25	Pct50	Pct75	Pct90	Pct95	Pct99	Max
1,00	1,00	2,00	2,20	4,00	7,00	9,00	11,00	11,80	14,40	26,00

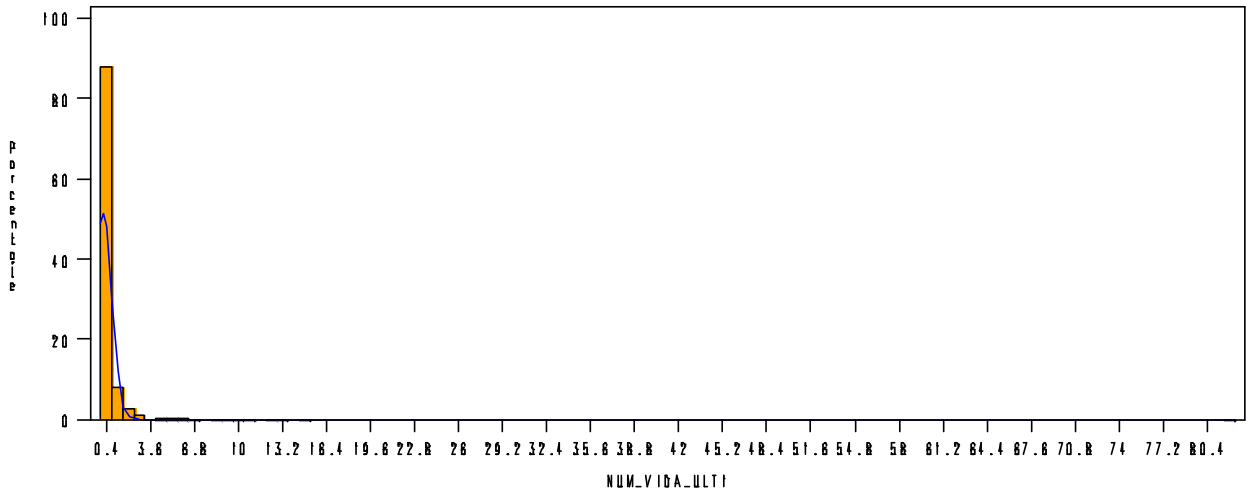
NUM_VIDA_HACE3: Número de productos de vida hace 3 meses



N	Perdidos	Distintos	% Iguales	Infinitos	Negativos	Ceros	Positivos	¿Tiene Moda 0?	¿Tiene Valores atípicos por la izda?	¿Tiene Valores atípicos por la dcha?	Media (si moda es 0 se excluye del cálculo)	Desv.Tipica (si moda es 0 se excluye del cálculo)
473.275	0	67	100%	0	0	402.285	70.990	SI	NO	SI	1,47	1

Mínimo	Pct01	Pct05	Pct10	Pct25	Pct50	Pct75	Pct90	Pct95	Pct99	Max
1,00	1,00	1,00	1,00	1,00	1,00	2,00	3,00	3,00	5,00	27,60

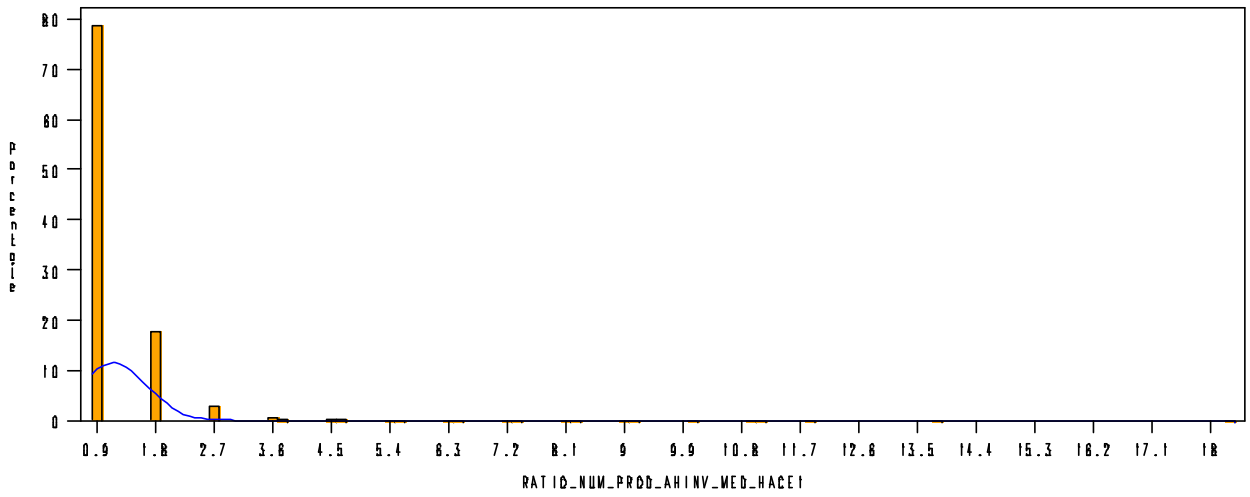
NUM_VIDA_ULT1: Número de productos de vida en el último mes



N	Perdidos	Distintos	% Iguales	Infinitos	Negativos	Ceros	Positivos	¿Tiene Moda 0?	¿Tiene Valores atípicos por la izda?	¿Tiene Valores atípicos por la dcha?	Media (si moda es 0 se excluye del cálculo)	Desv.Tipica (si moda es 0 se excluye del cálculo)
473.275	0	68	100%	0	0	402.800	70.475	SI	NO	SI	1,45	1

Mínimo	Pct01	Pct05	Pct10	Pct25	Pct50	Pct75	Pct90	Pct95	Pct99	Max
1,00	1,00	1,00	1,00	1,00	1,00	2,00	2,60	3,00	5,00	26,60

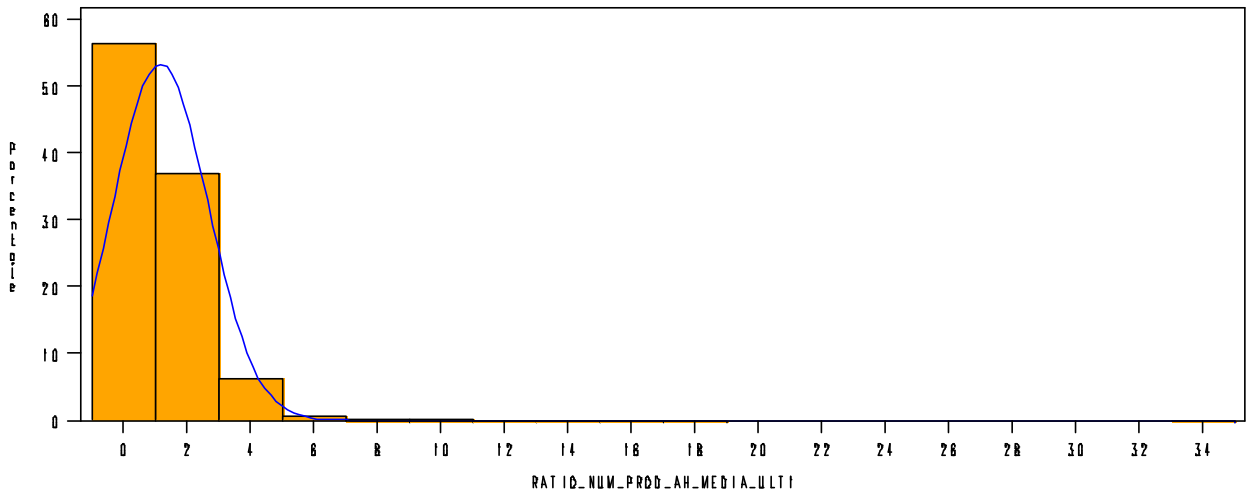
RATIO_NUM_PROD_AHINV_MED_HACE1: Ratio del número de productos de ahorro e inversión sobre la media



N	Perdidos	Distintos	% Iguales	Infinitos	Negativos	Ceros	Positivos	¿Tiene Moda 0?	¿Tiene Valores atípicos por la izda?	¿Tiene Valores atípicos por la dcha?	Media (si moda es 0 se excluye del cálculo)	Desv.Tipica (si moda es 0 se excluye del cálculo)
473.275	13	54	100%	0	0	0	473.262	NO	NO	SI	1,15	1

Mínimo	Pct01	Pct05	Pct10	Pct25	Pct50	Pct75	Pct90	Pct95	Pct99	Max
0,91	0,91	0,91	0,91	0,91	0,91	0,91	1,82	1,82	3,10	12,61

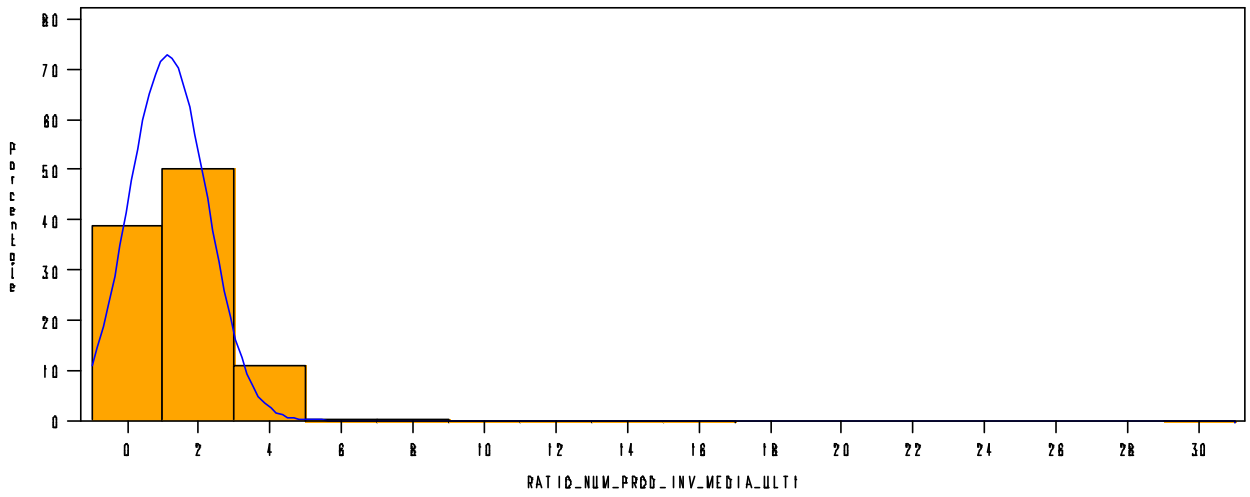
RATIO_NUM_PROD_AH_MEDIA_ULT1: Ratio del número de productos de ahorro sobre la media



N	Perdidos	Distintos	% Iguales	Infinitos	Negativos	Ceros	Positivos	¿Tiene Moda 0?	¿Tiene Valores atípicos por la izda?	¿Tiene Valores atípicos por la dcha?	Media (si moda es 0 se excluye del cálculo)	Desv.Tipica (si moda es 0 se excluye del cálculo)
473.275	13	41	100%	0	0	266.052	207.210	SI	NO	SI	2,68	1

Mínimo	Pct01	Pct05	Pct10	Pct25	Pct50	Pct75	Pct90	Pct95	Pct99	Max
2,26	2,26	2,26	2,26	2,26	2,26	2,26	4,53	4,53	6,79	19,97

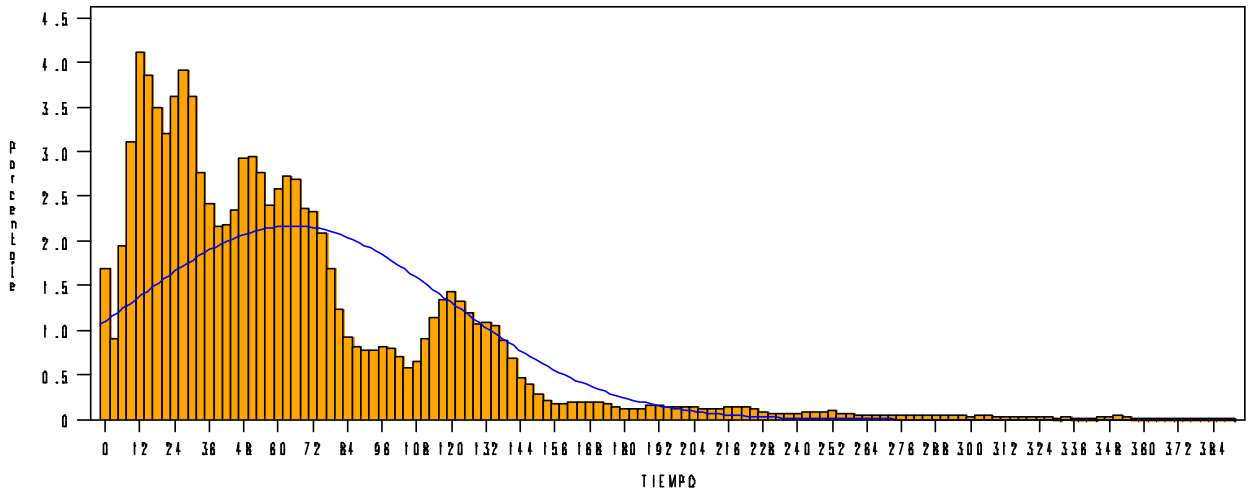
RATIO_NUM_PROD_INV_MEDIA_ULT1: Ratio del número de productos de inversión sobre la media



N	Perdidos	Distintos	% Iguales	Infinitos	Negativos	Ceros	Positivos	¿Tiene Moda 0?	¿Tiene Valores atípicos por la izda?	¿Tiene Valores atípicos por la dcha?	Media (si moda es 0 se excluye del cálculo)	Desv.Tipica (si moda es 0 se excluye del cálculo)
473.275	13	52	100%	0	0	183.855	289.407	NO	NO	SI	1,14	1

Mínimo	Pct01	Pct05	Pct10	Pct25	Pct50	Pct75	Pct90	Pct95	Pct99	Max
0,00	0,00	0,00	0,00	0,00	1,53	1,53	3,06	3,06	4,58	18,66

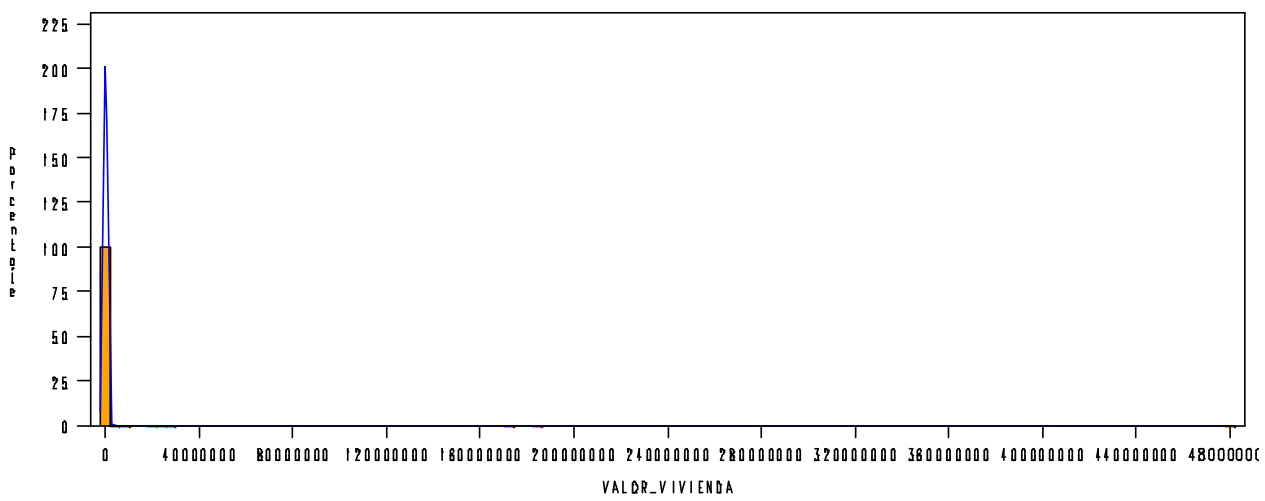
TIEMPO: Tiempo que los clientes tardan en fugarse



N	Perdidos	Distintos	% Iguales	Infinitos	Negativos	Ceros	Positivos	¿Tiene Moda 0?	¿Tiene Valores atípicos por la izda?	¿Tiene Valores atípicos por la dcha?	Media (si moda es 0 se excluye del cálculo)	Desv.Tipica (si moda es 0 se excluye del cálculo)
473.275	0	1.702	100%	0	378.620	1.728	92.927	NO	NO	SI	4,20	35

Mínimo	Pct01	Pct05	Pct10	Pct25	Pct50	Pct75	Pct90	Pct95	Pct99	Max
-9,00	-9,00	-9,00	-9,00	-9,00	-9,00	-9,00	41,80	73,40	142,40	386,40

VALOR_VIVIENDA: Valor de la vivienda

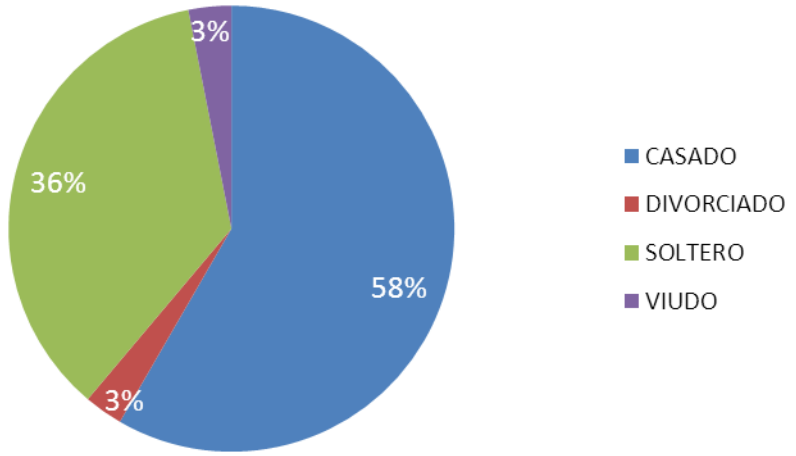


N	Perdidos	Distintos	% Iguales	Infinitos	Negativos	Ceros	Positivos	¿Tiene Moda 0?	¿Tiene Valores atípicos por la izda?	¿Tiene Valores atípicos por la dcha?	Media (si moda es 0 se excluye del cálculo)	Desv.Tipica (si moda es 0 se excluye del cálculo)
473.275	0	22.648	95%	0	0	450.223	23.052	SI	NO	SI	141.961,29	2.388.711

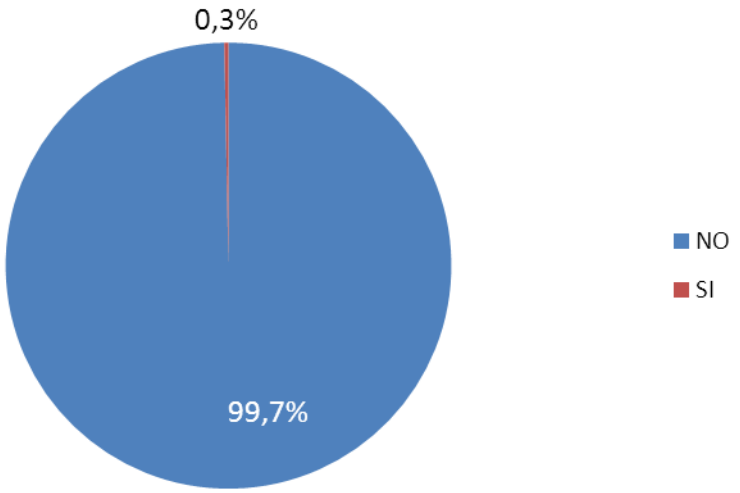
Mínimo	Pct01	Pct05	Pct10	Pct25	Pct50	Pct75	Pct90	Pct95	Pct99	Max
9,80	649,00	2.991,40	6.069,00	44.380,20	78.345,40	123.957,20	188.577,80	256.886,00	490.051,20	142.717.278,40

7.1.2 Variables categóricas

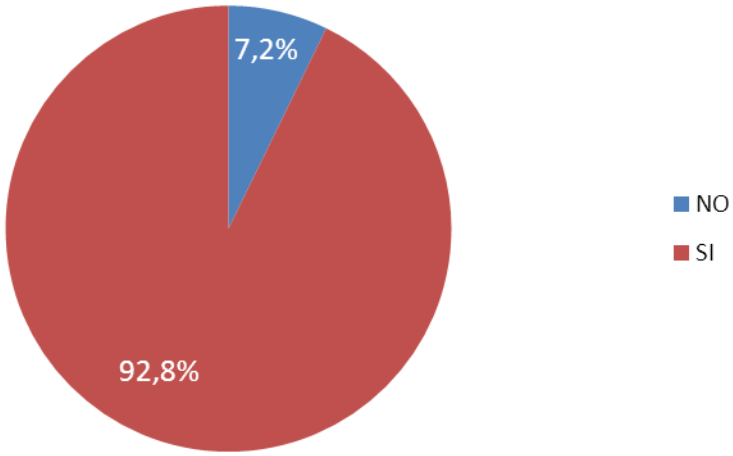
ESTADO_CIVIL: estado civil



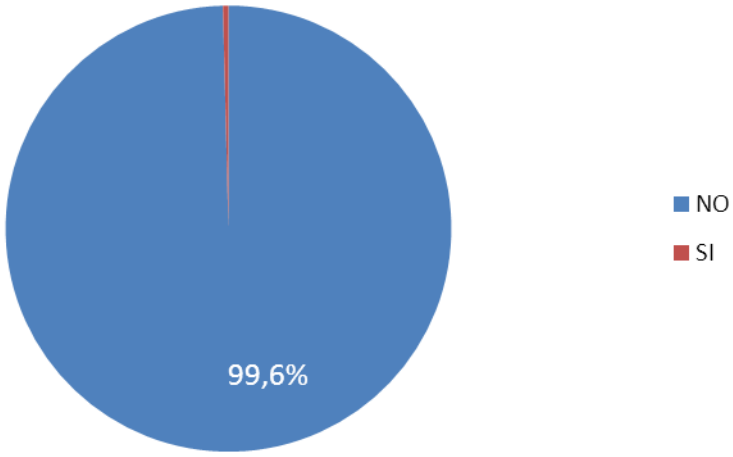
IND_EMPLEADO: Indicador de empleado



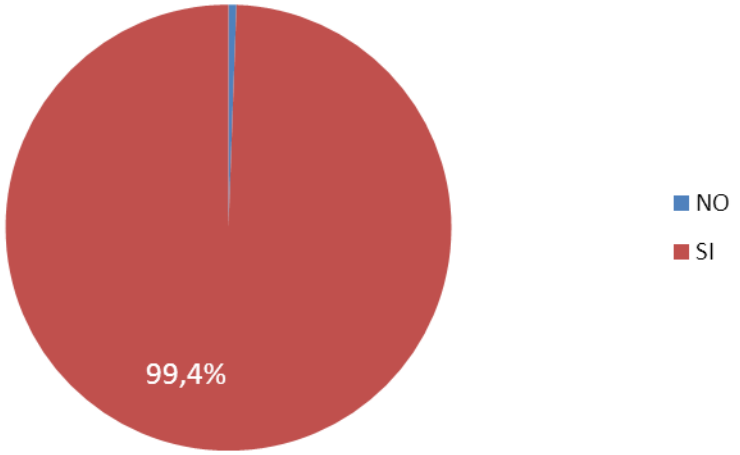
IND_NACIONAL_CAT: Indicador de nacionalidad española



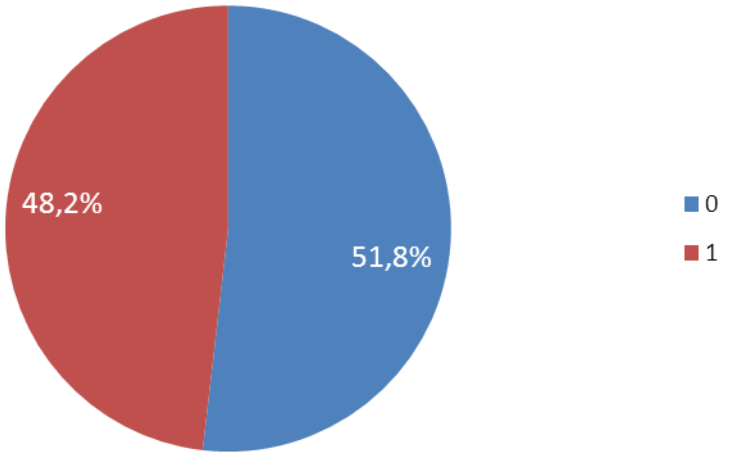
IND_RECLAMACIONES_ULT3_CAT: Indicador de nacionalidad española



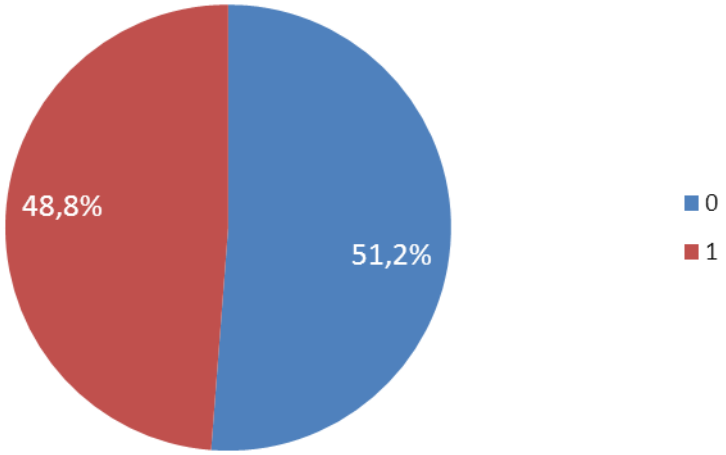
IND_RESIDENTE_CAT: Indicador de residente



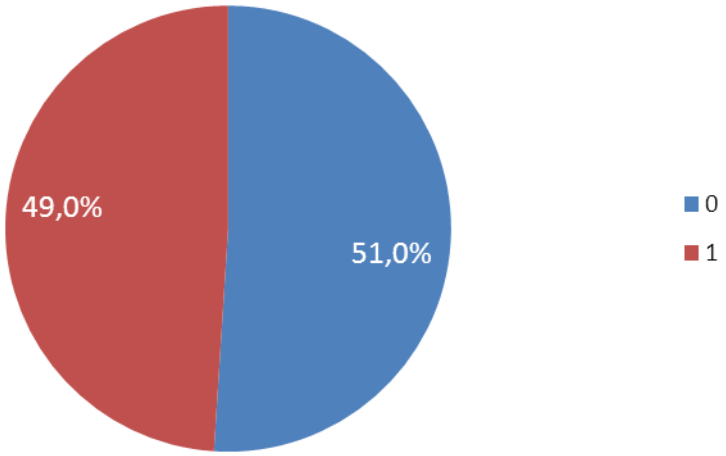
IND_TJDEB_ULT1: Indicador de tenencia de tarjeta de débito en el último mes



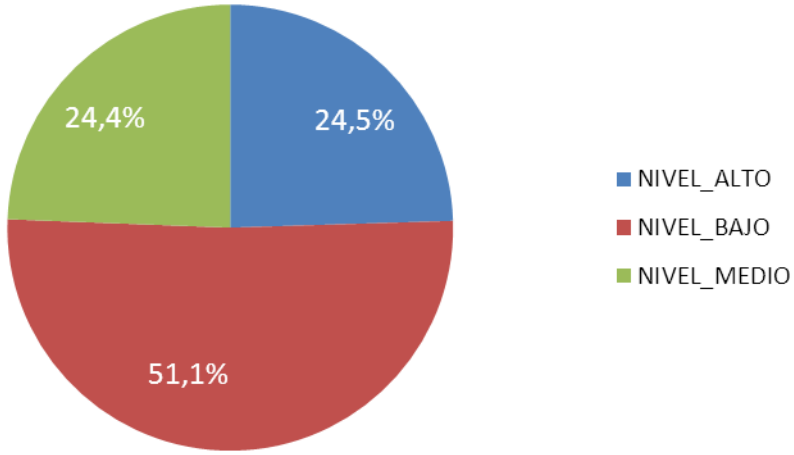
IND_TJDEB_ULT3: Indicador de tenencia de tarjeta de débito en los últimos 3 meses



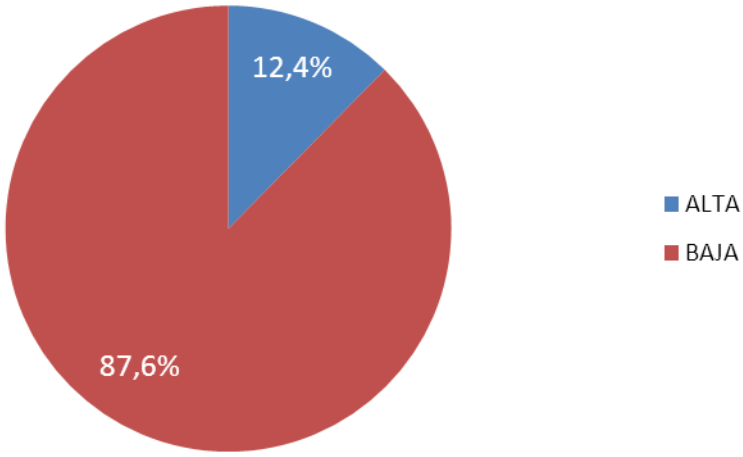
IND_TJDEB_ULT6: Indicador de tenencia de tarjeta de débito en los últimos 6 meses



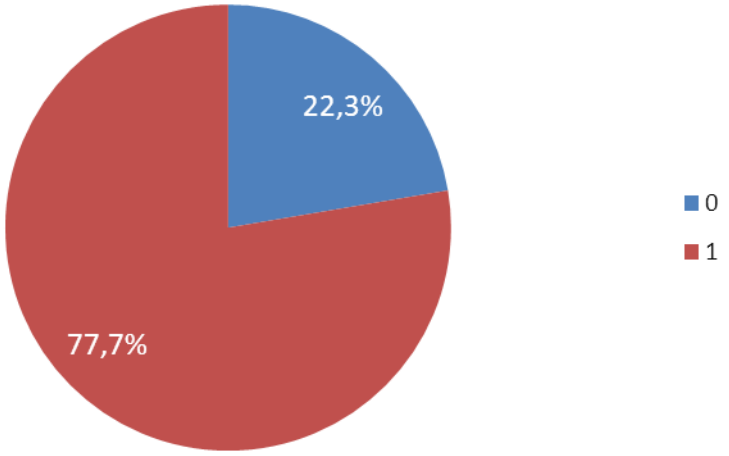
NIVEL_ESTUDIOS_CAT: Nivel de estudios



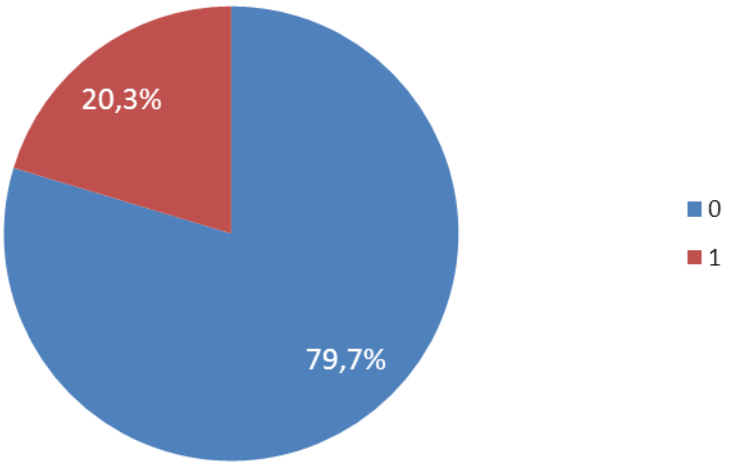
NIVEL_SATISFACCION_CAT: Nivel de satisfacción



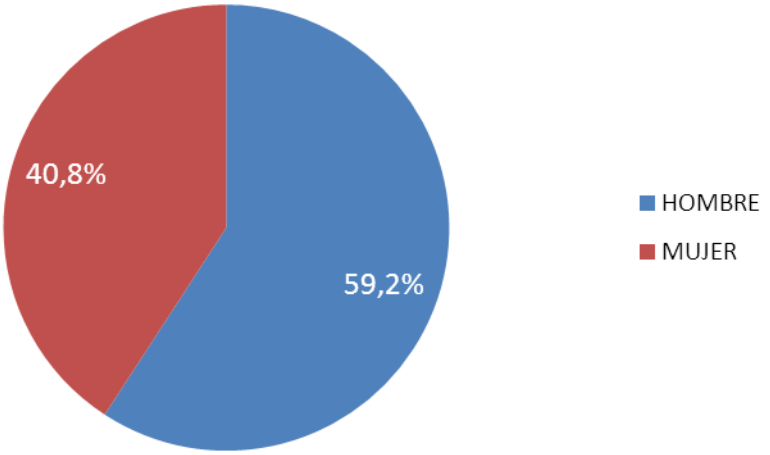
NUM_DIAS_CAT: Número de días transcurridos desde el pago del último recibo: 1= recientemente / 0= tienen que pagar dentro de poco



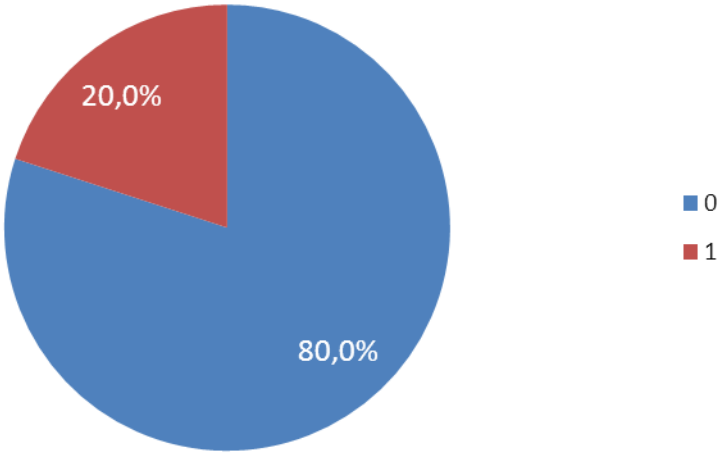
RECIB_PROMOCION: Toma el valor 1 si al cliente le ofrecieron un descuento especial en el último recibo. 0 en caso de precio normal



SEXO_CAT: Sexo



TARGET: Variable objetivo en árboles de decisión y regresión logística



VINCULACION_CAT: Vinculación