UNIVERSIDAD DE CHILE
FACULTAD DE CIENCIAS FÍSICAS Y MATEMÁTICAS
DEPARTAMENTO DE INGENIERIA INDUSTRIAL

ESTIMACIÓN DE CUSTOMER LIFETIME VALUE MEDIANTE TÉCNICAS SUPERVISADAS DE DATA MINING EN UNA EMPRESA DE RETAIL

MEMORIA PARA OPTAR AL TÍTULO DE INGENIERO CIVIL INDUSTRIAL

PEDRO ANTONIO URZÚA SALINAS

PROFESOR GUÍA:
LUÍS ABURTO LAFOURCADE.

MIEMBROS DE LA COMISIÓN:
MANUEL REYES JARA
PABLO MARÍN VICUÑA

SANTIAGO DE CHILE
OCTUBRE 2007

RESUMEN DE LA MEMORIA
PARA OPTAR AL TÍTULO DE
INGENIERO CIVIL INDUSTRIAL
POR : PEDRO URZÚA S.
FECHA : 03/10/2007
PROF. GUÍA : LUÍS ABURTO L.

ESTIMACIÓN DE CUSTOMER LIFETIME VALUE MEDIANTE TÉCNICAS SUPERVISADAS DE DATA MINING EN UNA EMPRESA DE RETAIL

El presente trabajo de titulo tiene como objetivo utilizar técnicas de data mining para determinar una metodología que permita estimar el lifetime value de los clientes de un supermercado mayorista. El comportamiento histórico de compra y variables de georeferenciación se utilizan para estimar cómo se comportará un cliente en el futuro. Este comportamiento se define como la variación porcentual del monto que desarrollará cada cliente.

Para la estimación se utilizan técnicas probabilísticas y de data mining. En particular, se construyen cinco modelos basados en las siguientes técnicas: Pareto/NBD, Árbol de decisión y MLP. Posteriormente, se incluyen tres modelos ingenuos que permitan justificar el desarrollo de modelos sofisticados. El desempeño de los modelos indica que las técnicas de data mining, para el caso analizado, tienen mejores resultados en las estimaciones. Se concluye, que el modelo Pareto/NBD es conservador ya que tiende a asumir que un cliente mantendrá su comportamiento. No obstante, el modelo árbol de decisión es levemente más agresivo siendo el mejor a nivel de segmentos e individual con un error de precisión de un 55.8% a nivel de clientes y un 9.2% a nivel de segmentos.

El resultado final consiste en la estimación del lifetime value a nivel de segmentos e individual de los clientes, utilizando el modelo árbol de decisión, que define un flujo monetario que se descuenta a lo largo de un periodo de tiempo. De esta manera se determina el lifetime value.

Como trabajo futuro se propone definir un criterio que permita proyectar la tasa de crecimiento de los clientes a lo largo del tiempo.

ÍNDICE

1. Introducción

El proceso de planificar y ejecutar la concepción o diseño del producto, el precio, la información y la distribución de las ideas, bienes y servicios para generar transacciones que satisfagan tanto los objetivos de las personas como los de las compañías, es denominado marketing, según *Kotler* [1]. El marketing ha experimentado grandes cambios desde los años 80. En particular, ha desarrollado cambios en el paradigma dominante, transacciones hacia relaciones según Morgan y Hunt [2]. Se busca desarrollar relaciones de largo plazo con los clientes, más que simplemente transacciones puntuales con ellos.

Basado en el enfoque transaccional, las empresas actúan según un paradigma, basado en lo siguiente:

- El cliente es anónimo para la empresa, es un mero comprador.
- Cada transacción tiene que ser rentable.
- La empresa habla y el cliente escucha.
- El cliente no tienen memoria, su decisión de compra, depende exclusivamente del producto que desee adquirir y no de factores externos.
- Es más fácil y barato captar a un nuevo cliente que retener a uno antiguo.
- Existe abundancia de clientes en el mercado.

Debido a las crecientes exigencias de los clientes, dicho enfoque ha quedado obsoleto en los momentos actuales, por lo tanto, surge un nuevo enfoque del marketing llamado "relacional", cuyo paradigma se fundamenta en los puntos siguientes:

- El cliente puede tomar la iniciativa, ya sea para emitir juicios, recibir comunicaciones o iniciar una transacción.
- Las compañías deben escuchar más a sus clientes y hablar menos.
- Los clientes no son anónimos, se les debe distinguir de manera de direccionarles acciones personalizadas.
- Cada transacción queda registrada identificando al emisor y manteniendo detalle de la transacción hecha. Existe memoria.

Existen diferentes tipos de clientes; ya sea por sus diferentes intereses, motivaciones, principios, paradigmas, etc. El conocimiento que se pueda extraer sobre ellos es de gran importancia para la empresa. El futuro de una empresa no depende del número actual de sus transacciones, sino depende de la información que la empresa pueda extraer del contacto con sus clientes. Esto permite, ofrecerles un mejor servicio anticipándose a sus necesidades. Desde el punto de vista del marketing relacional, se verifica que los clientes son distintos y por ello es necesario conocer la mayor información sobre ellos y efectuar su correspondiente análisis.

Como métrica, desde un punto de vista monetario, existe el lifetime value (LTV), el cual permite hacer una distinción monetaria entre sus clientes. Éste, se puede definir como: "*el valor presente que representa para la compañía a lo largo de su vida útil como clientes*" según Pfeifer, Haskins y Conroy [3]. Mediante esta métrica es posible diferenciar entre los clientes de mayor y menor valor para la empresa, permitiéndole generar un ranking de ellos.

La información de la cual dispone una empresa, es importante para la buena gestión de esta métrica. Como ejemplo de aplicación del lifetime value se puede citar el siguiente: Actualmente existe oferta de cuentas corrientes para estudiantes universitarios, a pesar de no poseer ingresos. Los bancos llegaron a esta determinación tras evaluar sus carteras de clientes, conociendo las profesiones de ellos para seleccionar determinados perfiles de estudiantes, (carrera e institución universitaria). Esta información, le permite a la entidad bancaria proyectar, cuál será el comportamiento futuro de estos potenciales clientes y así determinar su valor, lo cual se lleva a cabo con proyectos de LTV. Desde el punto de vista de un supermercado mayorista, se puede replicar este escenario buscando patrones de los clientes con variables socio-demográficas, como por ejemplo variables de georeferenciación de ellos.

1.1 Antecedentes del sector del retail

En los últimos años la industria del retail en Chile ha experimentado un gran desarrollo, manifestado con tasas de crecimiento superiores al crecimiento del país.

La creciente globalización de los mercados ha provocado que grandes cadenas de retail de origen europeo y norteamericanas, se instalen en Latinoamérica. En varios países, estas grandes cadenas han logrado conquistar los primeros lugares en participación de mercado.

En Chile, la situación es distinta ya que grandes cadenas extranjeras han intentado instaurar sus modelos de negocio, pero han obtenido malos resultados lo que ha provocado que abandonen sus operaciones en Chile. La industria del retail chilena se caracteriza por estar constituida principalmente por empresas nacionales, las cuales han defendido su mercado fuertemente de competidores extranjeros.

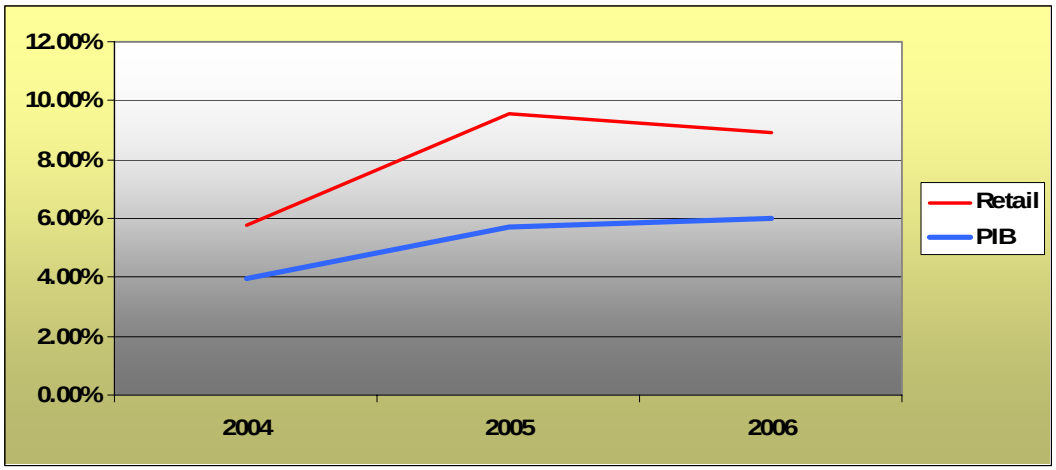


Gráfico 1.1[1]: Crecimiento de retail y PIB

Como consecuencia de la amenaza de cadenas internacionales los retails nacionales han fortalecido sus relaciones con los proveedores y han desarrollado programas para fidelizar a sus clientes, con lo que se han establecido como organizaciones legítimas debido a su gran conocimiento del mercado local. Además para potenciar su ataque a la

[1] Fuente Asociación Gremial de Supermercados de Chile www.asach.com.

competencia internacional, los retails chilenos se han preocupado de observar e imitar las mejores prácticas extranjeras, que junto al gran conocimiento adquirido del mercado local han sido la fórmula exitosa para el dominio del mercado nacional. Sin embargo, a pesar de no haber competencia internacional, existe una alta competencia nacional. Esta alta competitividad y los buenos resultados obtenidos han motivado a sus principales actores a no sólo defender sus mercados exitosamente, sino también a expandirse hacia nuevos mercados latinoamericanos.

El retail en Chile se compone principalmente de supermercados (mayoristas y minoristas), tiendas por departamento, multitiendas, tiendas de mejoramiento del hogar, tiendas de especialidad y comerciantes minoristas.

En particular, para desarrollar este proyecto de lifetime value, esta investigación se basará en el caso de un supermercado mayorista.

1.2 Investigaciones anteriores

La predicción del lifetime value se puede llevar a cabo con variables de tipo transaccional según Marín en [4] y además incorporando variables sociodemográficas según Baek en [5]. A continuación se expondrán las principales conclusiones las investigaciones nombradas.

- En la primera investigación (Marín [4]), basada en el caso particular de un supermercado mayorista, se utilizó como indicadores, la frecuencia de compra, el recency (tiempo desde la última compra) y el monto promedio de cada compra, como dimensiones suficientes para explicar el comportamiento transaccional de cada cliente. Con estas dimensiones se construyó el modelo Pareto/NBD, para descubrir patrones subyacentes. De esta manera se desarrolló una metodología de estimación de lifetime value. Los resultados obtenidos son útiles para predecir comportamientos a nivel de segmentos.
- La siguiente investigación (Baek [5]), se basó en el caso particular de una tienda de ropa. Se utilizó los indicadores de frecuencia, recency y monto promedio, incorporando variables socio-demográficas que permitan mejorar los resultados. Con las variables género, edad y región, se formaron segmentos de clientes de

manera de desarrollar un modelo individual para cada uno de ellos. Se utilizó el modelo Pareto/NBD para construir una metodología que permita estimar el lifetime value. A pesar de tener nuevas variables los resultados obtenidos no mejoran con respecto a la primera investigación, debido al bajo nivel transaccional del caso abordado.

2. Descripción de la investigación

2.1 Motivación

En la actualidad, muchas empresas poseen gran cantidad de datos relacionados con sus clientes. Estos datos pueden convertirse en información muy valiosa, siendo la correcta gestión de ésta un claro diferenciador para cada empresa.

Es importante conocer qué tan valioso es cada cliente, ya que permite desarrollar estrategias sobre como abordarlos, identificando cuales clientes son más importantes para la empresa. Una medida útil para determinar el valor de cada cliente es el lifetime value. Mediante esta métrica se pueden tomar acciones [3] sobre los clientes. En esta investigación se resolverá la problemática de estimar el valor de cada cliente, tomando como benchmarck investigaciones previas, en las cuales se utilizó el modelo Pareto/NBD.

2.2 Presentación del caso

Esta investigación es desarrollada, en una empresa perteneciente a la industria del retail. Se trata de un supermercado mayorista, es decir, venta al por mayor de productos. Es la principal empresa abastecedora de almaceneros la cual posee sucursales a lo largo de todo el país. Surge en el año 1984. Cuenta con 23 sucursales, 14 en la región metropolitana y 7 en regiones. Existen clientes socios, no socios e institucionales. La empresa desarrolló un club denominado “club del almacenero” con el fin de premiar a sus socios ofreciéndole precios más económicos.

2.3 Justificación de la investigación

Debido a los resultados poco satisfactorios obtenidos en la investigación de Marín [3] y posteriormente la de Baek [4], se hace interesante el proseguir con el estudio de esta materia. Utilizando técnicas de data mining se construirá una metodología para estimar el lifetime value de manera de analizar los resultados con este enfoque.

2.4 Diseño de la investigación

2.4.1 Objetivo general

Determinar una metodología que permita estimar el lifetime value mediante técnicas supervisadas de data mining.

2.4.2 Objetivos específicos

Como objetivos específicos se definen los siguientes:

1. Determinar grupos de clientes donde los modelos tengan mejor desempeño.
2. Rankear a los modelos desarrollados según su desempeño.
3. Precisar reglas que manifiesten un claro comportamiento de los clientes.
4. Seleccionar un modelo para estimar el lifetime value.
5. Estimar el lifetime value de de los cliente.

2.4.3 Metodología

En esta investigación se abordará una metodología basada en el proceso *KDD* y la propuesta metodológica de *M.J. Berry*. Consta de las siguientes 10 etapas:

Etapa 1: Definición el problema

Se identifica cual es la función objetivo que se quiere resolver, es decir, se define cual será la variable supervisada que se quiere pronosticar. Se plantean tres alternativas para ésta, monto futuro de compra de cada cliente, frecuencia de compra futura y variación porcentual del monto gastado en la tienda por cada cliente en un período con respecto a otro.

Etapa 2: Identificación de información valiosa

En esta etapa se observarán los datos con los que se dispone, con el fin de proponer un set de variables apropiado para calibrar los modelos posteriormente.

Etapa 3: Preprosesamiento de los datos

Se hará una limpieza de los datos, lo cual permitirá eliminar inconsistencias de posibles errores, y construir una base de datos con datos consolidados y confiables.

Etapa 4: Transformación de las variables

En esta etapa se construirán las variables propuestas. Se definirá la escala en la cual se trabajará con cada variable, para el caso de los modelos MLP, las variables deben ser normalizadas en la escala 0-1, de la siguiente manera:

$$Z_i = \frac{X_i - \min(X_i)}{\max(X_i) - \min(X_i)}$$

Etapa 5: Conocer los datos

Con el fin de conocer como se distribuye cada variable, se hará un análisis descriptivo de éstas. Además de analizar individualmente cada variable se hará un análisis multivariado, para encontrar redundancia entre las variables.

Etapa 6: Selección de variables

Con métodos sofisticados de selección de variables tales como los métodos filtros y embedded, se determinará el set de variables que se usará como input para la calibración de los modelos.

Etapa 7: Construcción de modelos

En esta etapa se llevará a cabo la implementación de los modelos. Iterativamente, se determinará los parámetros para cada modelo, que otorgue un mejor resultado. De esta forma, se conseguirá tener el mejor modelo de cada familia de técnicas utilizadas. Se definen tres tipos de modelos; regresión, clasificación_regresión e ingenuos. Cada modelo de regresión está compuesto por un único modelo (Árbol de decisión, MLP y Pareto /NBD). Cada modelo clasificación_regresión, está formado por dos modelos, uno de clasificación (MLP y Árbol de decisión) que en una primera instancia encasilla a los clientes en 5 clases (fuga, baja, mantiene, sube y multiplica), y posteriormente para los clientes asignados a cada clase, se les estima su valor continuo, con un modelo de regresión definido para cada agrupación. En el gráfico 2.1 se muestra la metodología explicada anteriormente:

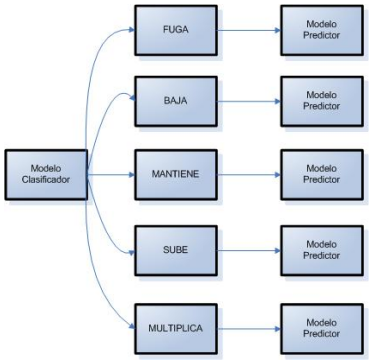


Grafico 2.1: Metodología Clasificación y regresión

Etapa 8: Comparación de modelos

En esta etapa se definen medidas de error que permitan comparar entre los modelos desarrollados, para posteriormente definir un ranking de desempeño.

Etapa 9: Interpretación de resultados

Se analizarán los resultados de los modelos identificando con que clientes se tienen resultados más confiables. Se determina cual es el mejor de los modelos desarrollados.

Etapa 10: Determinación de LTV

En esta etapa se determina la fórmula que permita proyectar a largo plazo el valor de cada cliente.

2.4.4 Resultados esperados

Mediante técnicas de data mining se definirá una metodología que permita estimar el lifetime value. Se visualizará los desempeños de cada modelo, para hacer un análisis comparativo entre todos los abordados y en especial entre los modelos de data mining con el modelo Pareto/NBD usado en otras investigaciones. Posteriormente, se indicarán reglas de negocio, que determinen segmentos de clientes, en los cuales, se tenga una buena precisión de los resultados.

2.4.5 Alcances

- En la investigación se abordará a un caso particular de la industria del retail.
- Se tiene como objetivo definir el valor de los clientes y no proponer estrategias que permitan mejorar el valor de cada uno de ellos.
- Se sacará el rendimiento a la información disponible, de manera de no utilizar técnicas de investigación de mercado para incorporar nuevas variables.

3. Marco teórico

3.1 Métodos para estimación de lifetime value

Existe una serie de enfoques o métodos para estimar el lifetime value, a continuación se resumirán los principales enfoques según Gupta, Hanssens, Hardie, Kahn, Kumar, Lin, Ravisanker, Sriram [9]:

3.1.1 Métodos RFM

Esta metodología ha sido usada por más de 30 años. Normalmente para ver tasas de respuesta a estrategias de marketing directo. Una práctica habitual consiste en crear grupos de clientes basados en tres variables:

a. R (Recency): tiempo desde la última compra hasta momento de evaluación.
b. F (Frequency): esta variable se puede estimar de dos maneras:
    - Cantidad de transacciones en un determinado período.
    - Tiempo entre transacciones
c. M (Monetary Value): monto promedio de las transacciones.

En función de las 3 variables mencionadas se definen clases que agrupan clientes dentro de cada una de estas variables. Por ejemplo, pueden generarse 5 clases para R, F y M, teniendo finalmente 125 clases dadas por las 5 clases definidas para cada una de las 3 variables. Con estas agrupaciones puede analizarse la tasa de respuesta dentro de cada clase, y definir el lifetime value para cada cliente en función de la clase a la cual pertenezca.

3.1.2 Métodos probabilísticos

Estos métodos plantean que un comportamiento se explica por un proceso estocástico, es decir, basados en una distribución de probabilidades empírica de cada variable permite predecir un valor futuro objetivo. Mediante estos modelos se busca modelar como a lo largo de una población el comportamiento de los clientes en función de una cierta distribución de probabilidades. Unos de los principales modelos probabilísticos para predecir el lifetime value es el Pareto/NBD desarrollado en el año 1987 por by Schmittlein, Morrison, and Colombo.

3.1.3 Métodos econométricos

Desde el punto de vista del lifetime value estos métodos se basan en los conceptos de adquisición, retención y el margen de los clientes. La adquisición se refiere a la primera vez que el cliente compra y para predecirla se usan normalmente regresiones logísticas del tipo Probit o Logit. La retención es la probabilidad de que un cliente se mantenga comprando. El tercer componente para estimar el lifetime value dado los modelos econométricos es el margen y crecimiento de los clientes. Una vez determinada la manera de definir estas tres dimensiones se fusionan de manera de generar un valor monetario para cada cliente. Modelos basados en estos métodos son las series de tiempo y las regresiones lineales.

3.1.4 Métodos de persistencia

Estos métodos han sido usados para ver como influyen en el lifetime value la adquisición, la retención y las ventas-cruzadas (cross-selling)[2]. Se pueden clasificar como métodos dinámicos, en los cuales se observa como los esfuerzos en adquisición y retención, y las estrategias de ventas cruzadas influyen el lifetime value. Si bien estos métodos se usan como métodos de lifetime value están alejados de las características de los otros métodos ya que estos no predicen un comportamiento futuro, sino que arrojan

[2] Relaciones entre productos en las boletas de cada cliente. Permite ofrecer un producto B, dado que el cliente compra el producto A.

un indicador de desempeño basados en la información histórica. Este indicador en la literatura financiera recibe el nombre de ROI (Retorno sobre la inversión). Este método permite evaluar como influyen las estrategias tomadas sobre los clientes en el valor que estos suponen para la empresa.

3.1.5 Métodos de data mining

Existe dos grupos de modelos de data mining. Modelos supervisados y no supervisados. La diferencia entre ambos, es que en los modelos supervisados se conoce la variable de respuesta, mientras que en los otros no se conoce.

Los métodos supervisados permiten predecir y clasificar, siendo los principales las Redes Neuronales y los Árboles de Decisión.

Estos métodos utilizan una muestra para la calibración, conocida como muestra de entrenamiento, la cual permite calibrar el modelo de tal manera de validar su efectividad con el resto de los datos, llamados muestra de testeo. De esta forma se busca iterativamente el modelo con la mayor capacidad para pronosticar el comportamiento de nuevos casos.

3.2 Técnicas para la estimación de lifetime value

Una vez abordado los principales métodos que se pueden utilizar para predecir el lifetime value se presentarán las técnicas que su utilizarán para predecir el lifetime value en esta investigación.

3.2.1 Modelo Pareto /NBD

Se basa en métodos probabilísticos y RFM. Estos modelos buscan pronosticar cuantas veces comprará cada cliente en el siguiente período, es decir la frecuencia de compra futura. Para ello utiliza las variables Recency y Frecuency. Posteriormente, se agrega la

variable de Monetary Value para obtener la estimación del monto futuro. Esta técnica tiene los siguientes supuestos[3]:

- Existe independencia entre Monetary Value y Frecuency, y Recency y Monetary Value.
- Un cliente considerado inactivo no puede volver a activarse.
- Las compras de cada cliente se representan según un proceso de Poisson de tasa λ. Estas tasas en la población se distribuyen según una distribución gamma.
- Un cliente permanece activo un periodo no observable que se distribuye según una exponencial de tasa μ. Estas tasas para la población se distribuyen según una distribución gamma.

Se estiman los parámetros de la población y posteriormente se incorporan las variables RFM de cada cliente lo que permite estimar el comportamiento individual de cada uno de ellos.

Las limitaciones de estos modelos son los siguientes:

- Las variables RFM son indicadores imperfectos del comportamiento verdadero de los clientes, ya que no toman enguanta tendencias ni aspectos más específicos de las compras de cada cliente como preferencia de productos por ejemplo.
- Sólo permiten pronosticar que el comportamiento futuro será similar al observado
- Estos modelos no toman en cuenta el hecho de que el comportamiento de compra se puede deber a otros aspectos distintos a transaccionales.

Sin embargo, Fader, Hardie, and Lee [6], muestran que con estas tres variables es posible construir un modelo de lifetime value que explique gran parte del comportamiento transaccional de los clientes.

En esta investigación se utilizarán los modelos de data mining, modelos probabilísticos y RFM.

[3] Ver [4] para mayor detalle.

3.2.2 Modelos de data mining

Para abordar un proyecto de lifetime value el tipo de modelo de data mining que se debe utilizar es de tipo supervisado, es decir, un modelo que se calibra mediante una variable de respuesta tal como comportamiento futuro de compra de los clientes. Los modelos de data mining supervisados que se usarán en esta investigación son los árboles de decisión y las redes neuronales artificiales del tipo perceptrón multicapa. A continuación se referirá en detalle a cada uno de estos modelos.

Perceptrón multicapa (MLP)

Es un método supervisado. El uso principal que se le ha dado es el de clasificar patrones y aproximar funciones continuas. Una red neuronal multicapa está compuesta por tres tipos de capas: capa de entrada, capas ocultas y capa de salida. Cada capa posee neuronas. Las neuronas de una capa están interconectadas con cada una de las neuronas de la capa anterior y también con las neuronas de la capa siguiente, por lo tanto si se tienen n neuronas en una capa y m en la siguiente se tendrán n por m conexiones y por lo tanto n por m pesos entre ambas capas. En la capa oculta existe una función de excitación que para ciertos valores de la neurona de entrada, esta función es excitada y da una determinada respuesta (neurona de salida). Cada neurona de entrada se conecta con la neurona oculta, mediante un peso asignado, y cada neurona oculta se relaciona con cada neurona de salida mediante cierto peso, definiendo un umbral para cada neurona.

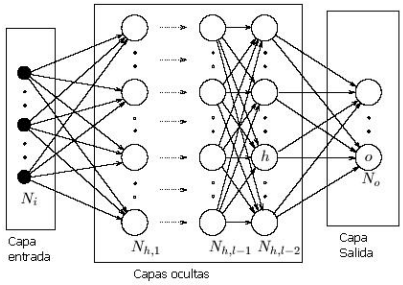


Gráfico 3.1: Perceptrón multicapa

El aprendizaje que se suele usar recibe el nombre de retropropagación del error (BP). Fijada una tasa de aprendizaje η y una condición de finalización y teniendo un vector supervisado t, el algoritmo opera de la siguiente manera:

- Paso 1: Asigna valores aleatorios a los pesos y los umbrales iniciales de cada neurona. Estos valores no deben ser nulos debido a que el aprendizaje no progresaría ya que las salidas de las neuronas serían nulas, y el incremento del peso también.
- Paso 2: Para cada patrón p del conjunto de los datos de entrenamiento:
    a. Se ejecuta la red para obtener la respuesta y^P dado ese patrón.
    b. Se calcula el error cuadrático sobre el patrón p mediante la siguiente ecuación:

$$E^P = \frac{1}{2}\sum_{i=1}^{N}(y_i^P - t_i^P)^2$$

    c. Se hallan los incrementos de los pesos y umbrales de todas las capas mediante la regla:

$$W^{(\xi)^P} = W^{(\xi-1)^P} - \eta^P\left(\frac{\partial E^P}{\partial W}\Big|_{W=W^{(\xi-1)^P}}\right)$$

- Paso 3: Se obtienen los incrementos globales una vez entrenada la red con todos los patrones, para todos los pesos y umbrales.
- Paso 4: Se actualizan los pesos.
- Paso 5: Se calcula el error cuadrático total:

$$E^P = \frac{1}{2}\sum_{p=1}^{P}\sum_{i=1}^{N}(y_i^P - t_i^P)^2 = \sum_{p=1}^{P} E^p$$

  En caso de satisfacer condición de término se vuelve a paso 2.

Para calibrar una red se debe entrenar con una muestra de entrenamiento generalmente entre un 70-90% de la muestra total, y luego se debe testear con otra muestra para estudiar el nivel de ajuste, con casos que no se usaron en el entrenamiento. Con una red neuronal se puede obtener buena precisión en los resultados, pero el problema es que con facilidad se cae en el sobre-ajuste, es decir, el modelo se ajusta mucho a la información de entrenamiento perdiendo generalidad. Con el fin de obtener los mejores resultados sin caer en sobre-ajuste se debe sensibilizar una serie de parámetros:

- Tasa de aprendizaje: indica la velocidad con que se ajustan los pesos.

- Momentun: evita los ajustes de los pesos permitiendo llegar a un óptimo global.
- Número de épocas: Cantidad de veces que se le muestran los registros a la red.
- Neuronas en la capa oculta: Cantidad de neuronas que se incorporan en la capa oculta.

No existe una manera determinada para seleccionar a priori estos parámetros, pero sí se pueden identificar ciertos rangos sobre cuales sensibilizar. Este método es considerado una caja negra, ya que es difícil de interpretar [10].

Árboles de decisión

Un árbol es un grafo formado por nodos y arcos con las siguientes restricciones:

1. Hay un sólo nodo raíz, es decir un nodo que no tiene padre.
2. Cada nodo distinto de la raíz tiene un sólo padre.

Cada ramificación del árbol se define como nivel. Las primeras ramificaciones se llevan a cabo con la variable que mejor discrimina. La metodología que usa un árbol para ramificar es clasificada como Greedy[4]. Es por esto que se da la posibilidad de que un atributo aparezca en diferentes niveles de las ramificaciones.

Son un tipo de técnica clasificada como supervisada. Se pueden usar como modelos para predecir valores continuos, árboles de regresión, o para clasificar fenómenos, árboles de clasificación.

Se debe sensibilizar una serie de parámetros, tales como que criterio utilizar para la ramificación, criterio de poda, es decir cuantos nodos terminales generar, y también cuantos casos debe haber como máximo en los nodos terminales. Mientras más nodos menor error de entrenamiento se tendrá, pero se suele incurrir en un mayor error de test lo que provoca un sobre-ajuste del modelo calibrado.

Los árboles de clasificación se pueden usar de las siguientes maneras:

[4] Greedy, metodología iterativa que selecciona en cada iteración, con la información disponible, las variables que mejor discriminan.

Árboles de regresión: Estos árboles permiten pronosticar un valor continuo, se tiene una variable supervisada conteniendo valores continuos. Hay dos criterios para determinar la ramificación del árbol

- Minimizar el error: minimiza el error general del árbol.
- Minimizar la varianza: este criterio busca minimiza la variabilidad de los registros que asignará a cada rama.

Árboles de clasificación: Estos árboles permiten clasificar registros. La variable supervisada está compuesta por valores nominales que indican la pertenencia de cada registro a cada clase. Se destacan los siguientes tipos de árbol de regresión:

- *Chi Squared Automatic Interaction Detection* (CHAID): fue diseñado por J. A. Hartigan. Utiliza la prueba chi-cuadrado para determinar si se debe continuar con la ramificación y en caso afirmativo identifica que variable debe ramificar.
- C4.5: basada en ID3 (Interactive Dichotomiser 3) de J. Ross Quinlan. Este tipo selecciona las variables para la ramificación de acuerdo a indicadores de entropía o desorden, es decir, utiliza la variable que proporcione mayor información en cada ramificación. Otro criterio para determinar como efectuar la ramificación es el índice de GINI el cual se define de la siguiente manera:

$$GINI(V) = \sum_{i=j} p(v) \sum_{i \neq j} p(i \mid v) p(j \mid v)$$

v: nodo
p(v): probabilidad de estar en el nodo v.
p(i|v): probabilidad de pertene-cer a la clase i dado que estoy en el nodo j

3.2.3 Comparación entre los modelos expuestos

En esta investigación se implementarán tres modelos, árboles de decisión, redes neuronales del tipo perceptrón multicapa y el modelo Pareto /NBD. Se expondrán las principales diferencias entre estos modelos:

- Explicabilidad: Las redes neuronales son consideradas una caja negra, es decir no es fácil interpretar los resultados de esta. Los árboles de decisión tienen una muy buena manera de interpretar, ya que a través de cada ramificación se definen reglas. El modelo pareto no otorga unas reglas claras como para determinar a nuevos clientes sus pronósticos.
- Exactitud: Las redes neuronales predicen con una mejor precisión que los árboles de decisión [10]. El modelo pareto no es muy exacto a nivel individual debido a que los parámetros se calibran a nivel muestral.
- Período de análisis: los modelos de minería de datos no son dependientes del período de análisis. El modelo pareto si depende del período debido a que para valores mayores a 120 no logra retornar resultados.
- Calidad de los datos: el MLP es muy sensible a la calidad de los datos a diferencia de los árboles de decisión.
- Sobre ajuste: en un MLP se cae con mayor facilidad en este problema que en un árbol. En el modelo pareto no se cae en sobre ajuste ya que determina los parámetros óptimos encontrando la solución global. Los modelos de minería de datos son dependientes de los parámetros que se eligen a diferencia del modelo pareto.

A continuación una tabla resumen de las principales diferencias de las técnicas expuestas:

Comparación	MLP	Árbol	Pareto
Largo del período de análisis	Sin limitación	Sin limitación	Limitado
Explicabilidad	Caja Negra	Reglas	Mala
Exactitud	Buena	Media	Media
Datos	Sensible	Indiferente	Sensible
Parámetros	Dependiente	Dependiente	Único
Peligro de Sobre ajuste	Alto	Medio	Bajo

Tabla 3.1: Análisis comparativo de los modelos

3.3 Metodologías para el desarrollo de estudios de data mining

3.3.1 Metodología Knowledge Discovery in Databases (KDD)

Es un proceso no trivial de identificación de patrones de comportamiento previamente desconocidos en grandes volúmenes de datos. Esta metodología consta de las siguientes etapas:

Etapa 1: Selección

- Elección de los datos necesarios para desarrollar el proceso.
- Determinación la información más relevante.

Etapa 2: Preprocesamiento

- Exploración de los datos mediante análisis descriptivos
- Detección de presencia de errores y falta de información.
- Preparación de los datos para el modelamiento.

Etapa 3: Transformación

- Se convierten los datos en información valiosa, es decir, información que explique de mejor manera de problemática.

Etapa 4: Data mining

- Construcción e implementación de los modelos.

Etapa 5: Interpretación y evaluación de los resultados

- Determinación de la calidad de los resultados obtenidos.

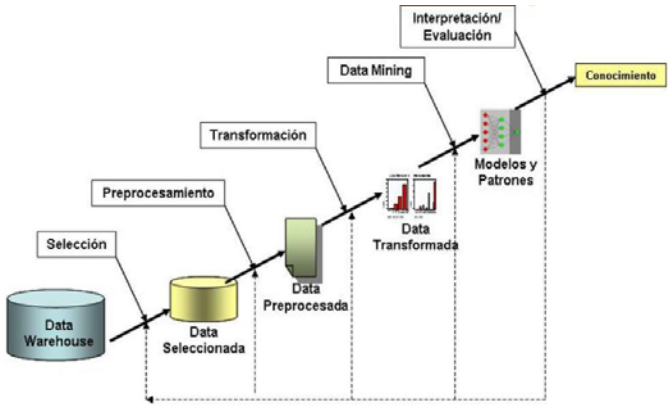


Gráfico 3.2: Metodología KDD

3.3.2 Metodología propuesta por Michael J. Berry

Esta metodología surge como una mejora a la metodología KDD [19], incorporando más etapas al proceso. Las etapas son las siguientes:

1. **Definir el problema de negocio:** Se determina cual es el fin comercial, objeto de la investigación.
2. **Traducir el problema en un problema abordable por la minería de datos:** Se visualiza el objetivo comercial de tal manera de ser abordable por el enfoque de la minería de datos.
3. **Seleccionar datos apropiados:** Se identifica potenciales datos a considerar para determinar el objetivo en estudio, de acuerdo a la disponibilidad, requerimientos y dimensionalidad.
4. **Conocer los datos:** Se llevan a cabo análisis descriptivos de tal manera de conocer su variabilidad y sus distribuciones, de tal manera de identificar posibles falencias presentes en los datos.
5. **Crear set de modelos:** Se indaga en la literatura para determinar cuales son los modelos más apropiados para abordar la problemática. Además se determinan las muestras de testeo y de entrenamiento.
6. **Resolver problemas de los datos:** Se debe tomar decisiones para definir criterios de limpieza de los datos por posibles inconsistencias o falencias presentes en ellos.
7. **Trasformar los datos a información valiosa:** Se manipulan los datos seleccionados de tal manera de obtener información valiosa para tener mejores resultados.
8. **Construir modelos:** Se seleccionan herramientas para la programación y desarrollo de los modelos.
9. **Evaluar modelos:** Se indaga en la precisión y en la generalidad de cada modelo por separado.
10. **Comparar modelos:** Se selecciona medidas de error de manera de poder comparar a los modelos.
11. **Evaluar resultados:** Se determinan la calidad de las soluciones obtenidas, sacando conclusiones al respecto.

En el gráfico 3.2, se especifican las etapas de la metodología propuesta por M.J. Berry:

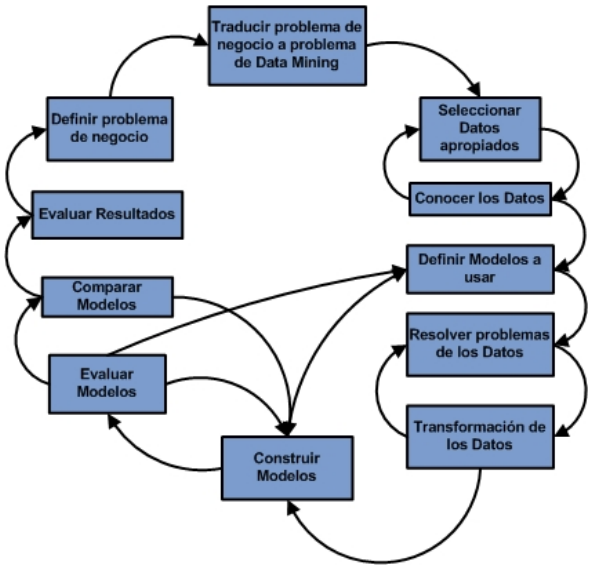


Gráfico 3.3: Metodología de M.J. Berry.

3.4 Métodos para la selección de variables

3.4.1 Métodos Filtros

Estos métodos permiten hacer análisis entre las relaciones existentes entre las variables. Se destacan los siguientes métodos filtros:

a. **Análisis de componentes principales (PCA)**: Reduce el espacio en nuevas variables llamadas componentes. Estas se determinan por combinaciones lineales de las variables iniciales. Estos componentes explican la varianza total de las variables iniciales, siendo el primer componente el que explica la mayor cantidad de ella. La cantidad de componentes es determinada por el total de la varianza que se desee explicar. Este método para el caso de seleccionar variables se puede usar de la siguientes maneras:

    i. **Reducción de variables**: Usar los componentes obtenidos como variables para explicar el fenómeno.
    ii. **Generación de un ranking de variables**: *Montoya* [11] propone una metodología que mediante PCA se genera un ranking de variables. Mediante la suma ponderada entre la varianza explicada por cada componente y la correlación entre la variable y cada componente se determina un índice para cada una de las variables. Mientras mayor sea el índice, mayor es la importancia de la variable, lo cual genera un ranking de las variables.

b. **Test de ANOVA**: Este test compara las medias de las variables entre los distintos grupos, de manera de indicar si cada una de las variables continuas hacen diferencia dentro y entre los grupos. Si la diferencia dentro de los grupos es mucho menor que la diferencia entre los grupos se determina que la variable discrimina entre los grupos.

c. **Test Chi-cuadrado**: Este test evalúa el grado de dependencia entre dos variables de tipo categóricas.

d. **Índice de correlación de Pearson**: Se utiliza este índice para verificar la relación lineal existente entre un par de variables de tipo continuo.

3.4.2 Métodos Wrapper

Estos métodos utilizan un modelo para discriminar entre las variables. Se le muestran distintos subconjuntos de variables al modelo. El subconjunto que retorna el mejor resultado en el modelo es el subconjunto de variables seleccionado. La cantidad de subconjuntos es de orden 2^N donde N es la cantidad de variables que se consideran.

Operacionalmente no es abordable evaluar todos los subconjuntos ya que en algunos casos la cantidad de variables puede ser grande. Con 20 variables ya se superan el millón de subconjuntos a analizar. Es por esto que existen los siguientes métodos iterativos para reducir este subconjunto:

1. Backward: Se presentan todas las variables al modelo y se evalúa el modelo. Luego en cada iteración siguiente se saca una variable y se evalúa el modelo. Se itera hasta que el sacar la siguiente variable signifique una reducción en la eficiencia del modelo. Una vez terminada la iteración la variables seleccionadas son las que no han sido eliminadas.
2. Forward: Se presenta una sola variable al modelo y se evalúa. En cada iteración se incorpora una nueva variable y se evalúa el modelo. Se termina de iterar cuando el agregar una nueva variable significa una reducción en la eficiencia del modelo.
3. Stepwise: Es una combinación del método Backward y Forward, en términos de que variable que salió del modelo puede volver a incorporarse, y una variable que entró al modelo puede salir.

3.4.3 Métodos Embedded

Existen modelos que en su operatoria manejan de manera implícita la selección de variables. Destacan los árboles de decisión ya que al buscar ajustarse a una variable supervisada seleccionan las mejores variables para ello, determinando así un "ranking automáticante" de las variables. El árbol de decisión por si sólo decide en cada ramificación que variable es la propicia para discriminar entre sus ramas.

3.5 Tipos de problemas con los datos

Al manejar grandes volúmenes de información surgen constantemente problemas e inconsistencia en los datos que se posee. Esto puede provocar que los resultados de un análisis se vean influenciados por este efecto originando resultados incorrectos. Los principales problemas que se pueden encontrar en los datos son los siguientes:

1. Valores fuera de rango, denominado en la literatura como **outlier**. Son valores que se encuentran muy alejados de la masa principal de la concentración de valores, es decir, datos cuyo valor cae fuera de los límites que encierran a la mayoría de los valores correspondientes de la muestra.
2. Valores perdidos, denominados en la literatura como **missing-values**. Son campos que se encuentran sin dato. El origen de este inconveniente puede ser por las siguientes razones:
    a. Los valores perdidos dependen del valor de la variable. Por ejemplo si un cliente no tiene correo electrónico se deja ese campo vacío.
    b. Los valores perdidos se debe a la relación con otra variable. Por ejemplo, si un individuo en la variable cantidad de hijos tiene 0, en el campo nombre del hijo estará vacío.
    c. Los valores son perdidos son consecuencia de un error, y no existe una relación como en los casos anteriores.

3. Valores erróneos. Este inconveniente ocurre habitualmente con variables de tipo nominal. Principalmente tienen su origen por errores de tipeo o diferencia entre los formatos.

3.6 Medidas de error

Para determinar la precisión de los resultados obtenidos por cada modelo las siguientes medidas de efectividad son indicadores que permiten comparar entre los modelos y a su vez determinar el desempeño de cada uno de ellos.

a. MAE: Error absoluto medio. Esta medida permite ver el error lineal en la unidad de medida de trabajo, se define de la siguiente manera:

$$MAE = \frac{1}{N}\sum_{i=1}^{N}\left|\mathrm{Re}\,al - \mathrm{Pr}\,edicción\right|$$

b. MAPE: Error porcentual absoluto medio. Este error mide de manera porcentual el error lineal producido por las predicciones de cada modelo.

$$MAPE = \frac{1}{N}\sum_{i=1}^{N}\frac{\left|\mathrm{Re}\,al - \mathrm{Pr}\,edicción\right|}{\left|\mathrm{Re}\,al\right|}$$

c. MSE: Error cuadrado medio. Este error mide de manera cuadrática la falta de precisión de los modelos. Este tipo de error castiga de manera más severa a los grandes errores de pronóstico.

$$MSE = \frac{1}{N}\sum_{i=1}^{N}(\mathrm{Re}\,al - \mathrm{Pr}\,edicción)^2$$

d. NMSE: Error cuadrático medio v/s error cuadrático del promedio

$$NMSE = \frac{\sum_{i=1}^{N}(\mathrm{Re}\,al - \mathrm{Pr}\,edicción)^2}{\sum_{i=1}^{N}(\mathrm{Re}\,al - \mathrm{Pr}\,omedio)^2}$$

4. Desarrollo de la metodología

4.1 Definición del problema

Se cuenta con la información de dos períodos, años 2005 y 2006. El volumen transaccional correspondiente a tales períodos es cercano a los 100 millones.

Se debe identificar la función objetivo que se quiere resolver para enfocar los esfuerzos de desarrollo. Para ello, se determinará el problema que se quiere resolver, es decir determinar la cuál será la variable a supervisar. Como objetivo comercial se plantea pronosticar el lifetime value de cada cliente. Para enfocar este propósito mediante técnicas de data mining, se proponen las siguientes alternativas:

1. **Estimar el monto futuro:** es decir, pronosticar el monto que comprará cada cliente en el año 2006 con variables del año 2005.
2. **Frecuencia futura:** es decir, pronosticar cuál será la frecuencia de compra de cada cliente en el año 2006 utilizando variables del año 2005.
3. **Variación porcentual de monto:** es decir, pronosticar la variación del monto gastado que efectuará cada cliente entre el año 2005 y 2006 mediante el uso de variables del año 2005.

En esta investigación la variable a predecir será la variación porcentual del monto, ya que permite definir una tasa de crecimiento ó decrecimiento para cada cliente a lo largo del tiempo, permitiendo proyectar con una tasa de descuento definida, el monto que gastará cada cliente en cada espacio de tiempo futuro, para estimar el lifetime value.

Numéricamente, la tasa de crecimiento de un cliente se ponderará con el monto conocido para estimar el monto futuro teniendo un menor error en las estimaciones que las otras alternativas expuestas.

Se define un año completo para los cálculos de las variables predictoras y la variable supervisada debido a la estacionalidad que se presenta a lo largo de un año completo. En el gráfico

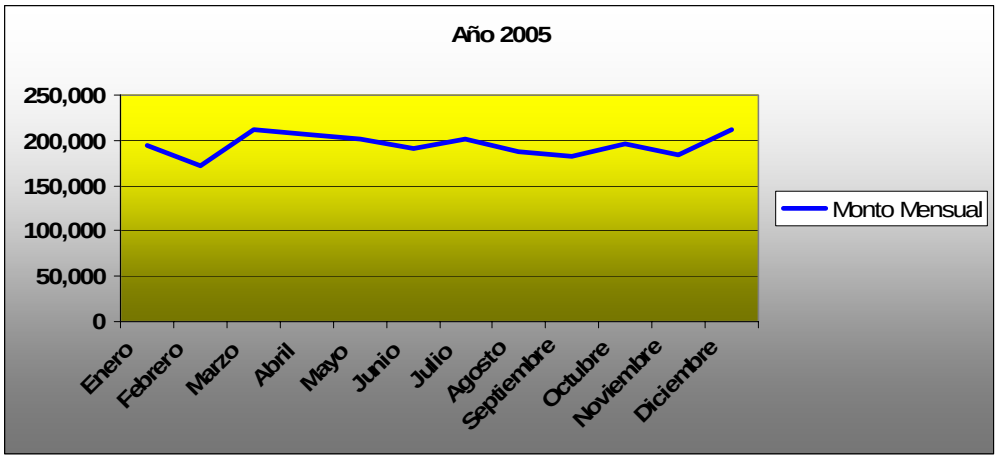


Grafico 4.1: Monto promedio mensual de los clientes

Se observa en el gráfico 4.1 que existe una pequeña estacionalidad en el monto promedio mensual gastado por cada cliente, lo que implícitamente manifiesta una estacionalidad en la variación de este monto.

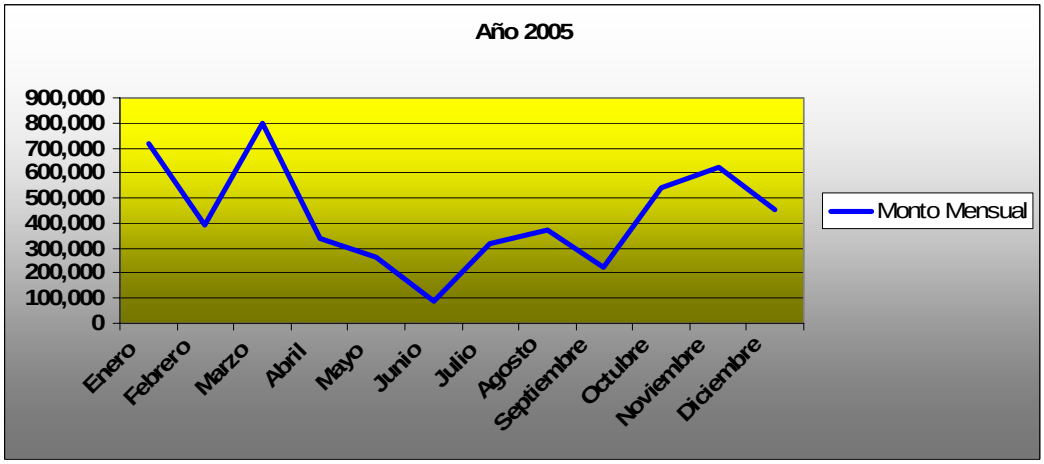


Gráfico 4.2: Cliente estacional

En el grafico 4.2, se presenta un cliente ejemplo con una gran estacionalidad en su comportamiento a lo largo del año.

4.2 Identificación de información valiosa

Se posee una amplia base de información transaccional y una mínima información sociodemográfica de los clientes de la empresa.

De la **información transaccional** que se dispone es posible inferir información como la siguiente:

- Productos que compra cada cliente
- Cantidad de unidades que lleva
- Capacidad de compra
- Momento en que se efectúa cada transacción
- Costo de cada transacción
- Sucursal en la cual compra

En la información que contiene los **datos de los clientes** es posible rescatar información como:

- Antigüedad de los clientes
- Dirección del cliente
- Tipo de cliente (SOCIO, NO SOCIO, INSTITUCIONAL)

En el diagrama siguiente se puede observar los campos seleccionados para el análisis

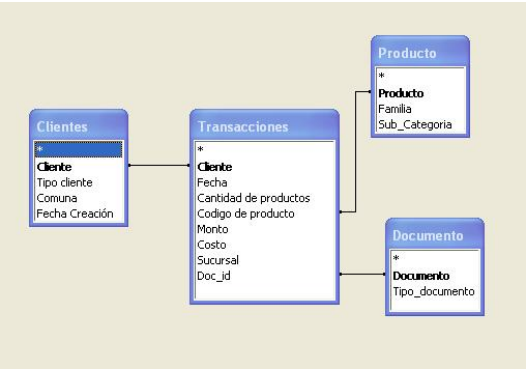


Grafico 4.3: Diagrama lógico de datos.

En el gráfico 4.3 se presenta el diseño lógico de los datos utilizados. A continuación se presentará cada uno de ellos:

Tabla Clientes, esta tabla contiene la lista de todos los clientes con los siguientes campos seleccionados para el análisis:

- Cliente: es un código que se le asigna a cada cliente con el fin de identificarlo en las diferentes tablas de datos.
- Tipo de cliente: indica si un cliente pertenece a la categoría de socio, no socio o institucional.
- Comuna: indica el domicilio comunal perteneciente a cada cliente
- Fecha de creación: Indica la fecha en que cada cliente se incorporó como cliente a la empresa.

Tabla Transacciones, esta tabla posee cada transacción efectuada por todos los clientes, los campos necesarios para esta investigación son los siguientes:

- Clientes: código que se le asigna a cada cliente con el fin de identificarlo en las diferentes tablas de datos.
- Fecha: indica el instante en que se efectúa la transacción.
- Cantidad de productos: este campo indica cuantas unidades fueron objeto de una transacción.
- Código de producto: este campo permite enlazar esta tabla con la tabla producto para identificar atributos de pertenencia de cada uno de estos.
- Monto: este campo indica el valor monetario de la transacción.
- Costo: indica el costo que supone para la empresa la transacción.
- Sucursal: indica en que sucursal se efectuó la transacción.
- Docu id: campo que permite cruzar la tabla con la tabla documento.

Tabla Producto, esta tabla contiene la identificación de cada uno de los productos que se transan en las transacciones, y consta de los siguientes campos:

- Producto: código definido para cada producto
- Familia: indica en que familia se clasifica a cada producto
- Sub-Categoría: indica en que sub-categoría se clasifica cada producto.

Tabla Documento, contiene la descripción de cada tipo de documento con los siguientes campos:

- Documento: código que se le asigna a cada tipo de documento.
- Tipo documento: indica si una transacción es objeto de una boleta factura o devolución[5].

4.3 Preprocesamiento de los datos

Para obtener resultados confiables es necesario hacer una limpieza de los datos, mejorando inconsistencias que se detecten.

Se utilizarán las transacciones que son objeto de compras, filtrando así las transacciones que son objeto de una devolución de la tabla transaccional, ya se quiere conocer comportamientos de compra.

Se sospecha que la distancia entre el domicilio del cliente y la ubicación del supermercado pueden tener correlación con la frecuencia transaccional. Es por ello, que se decidió utilizar esa variable. Existen comunas escritas de manera errónea, o con diminutivos como por ejemplo Santiago y Stgo. En estos casos se quiere que los clientes se asocien a la misma comuna a pesar de que esté escrita de manera distinta. Para abordar este problema se utiliza una metodología llamada (Distancia de) **Distancia de Levenshtein**[6], la cual, calcula la distancia mínima que necesita una cadena de caracteres (palabra) para transformarse en otra, de manera que agrupa las palabras que tienen distancia baja entre ellas. Para corregir de manera eficiente este problema se creó una base de datos, en la cual están todas las comunas de manera correcta escritas, de tal manera de poder consolidarlas. Esto se implementó en Microsoft Excel, con un tiempo de ejecución inferior a 4 minutos.

Tras aplicar el algoritmo se rescatan 163 comunas de las 330 que aparecen en la tabla originaria de los clientes con transacciones. Con las comunas en correcto estado, existe

[5] En términos de retail a una devolución se le denomina nota de crédito.
[6] En el Anexo 1 se detalla algoritmo.

un 5.86% de los clientes que no tienen comuna asignada. Un 98.58% de los clientes sin comuna corresponde a clientes NO SOCIOS.

Del total de clientes hay un 93.15% que corresponde a clientes SOCIOS. Los clientes SOCIOS son los que poseen tarjeta de socio, de tal manera que se tiene seguridad que la mayor parte de sus transacciones en el supermercado han sido procesadas, lo cual no se puede hacer con los clientes NO SOCIOS. Por lo tanto se filtran los clientes que no pertenezcan a la categoría SOCIO, quedando un total de 93.15% de los clientes y filtrando los que no tienen comunas queda un total de 93.06%. Se infiere que el hecho de que algunos clientes no tengan comuna asignada es consecuencia de que son NO SOCIOS. .

Existen 25 clientes con fecha de activación posterior al año 2005, lo que corresponde a un 0.06%, estos clientes serán extraídos de la base, es decir, no se considerarán para el análisis.

- **Filtro antigüedad**

Para decidir la normalización de la variable supervisada, se mostrará a continuación una tabla con los principales estadísticos por antigüedad de clientes.

Antigüedad	Monto 2005	Desviación del monto 2005	Monto 2006	Desviación del monto 2006	% clientes
1	1,871,173	2,509,645	1,481,896	2,604,512	1.44%
2	1,716,901	2,556,958	1,640,366	2,586,113	1.40%
3	1,732,791	5,905,063	2,249,066	14,648,295	2.97%
4	1,169,189	1,489,158	1,264,291	2,110,799	2.56%
5	1,076,368	1,745,246	1,496,645	2,275,429	2.90%
6	901,385	1,278,656	1,390,371	2,254,221	2.45%
7	778,193	784,618	1,291,105	1,709,995	1.94%
8	719,678	760,120	1,360,255	1,932,561	1.93%
9	703,786	699,720	1,649,609	2,139,973	1.43%
10	511,584	437,921	1,644,408	1,895,008	1.09%
11	475,192	455,674	2,224,781	2,306,844	0.93%
12	2,267,335	5,097,774	2,097,385	7,099,317	78.96%

Tabla 4.1: Estadísticos dado antigüedad.

Se observa en la tabla 4.1, que los clientes con antigüedad entre 6 y 11 meses, aumentan su gasto en promedio considerablemente, es decir, existe una inestabilidad entre los clientes con antigüedad menor a 12 meses. El problema principal es la presencia de clientes estacionales, efecto que no es posible detectarlo con clientes que tienen antigüedad menor al período de análisis, es decir, menor a 12 meses. Para lograr captar el posible efecto estacional de los clientes, se filtrarán del análisis aquellos clientes que tengan una antigüedad inferior a 12 meses, quedando un total de 78.96% de los clientes restantes, siendo un 74.13% del total de los clientes iniciales.

- **Análisis de clientes con muy bajo monto y baja frecuencia**

Debido a la existencia de clientes con muy poco valor en sus montos de compras, se deben efectuar filtros para no considerar a los clientes que sean despreciables. Para esto se deberá determinar que clientes se pueden considerar como clientes que incorporen grandes errores a los modelos. Para determinar a clientes dentro de este grupo se analizarán las siguientes variables:

1. Visitas del año 2005.
2. Monto acumulado año 2005.

Para determinar un umbral de corte para la variable frecuencia se analizarán los principales estadísticos de algunas variables de los clientes que cumplan con determinada frecuencias de compra.

Visitas 2005	Clientes	Total %	Monto 2005 promedio	Monto 2006 promedio	Variación promedio	Desviación promedio
1	4122	10.21%	51,391	152,681	1038.08%	12214.16%
2	2283	5.68%	80,123	247,165	524.09%	3374.25%
3	1558	3.88%	106,064	234,923	240.30%	1098.94%
4	1386	3.45%	132,072	243,301	175.98%	855.57%
5	1114	2.77%	158,364	331,692	188.87%	683.65%
6	1001	2.49%	179,574	353,194	132.97%	609.76%

Tabla 4.2: Estadísticos dado frecuencia

Se puede observar en la tabla 4.2, que los clientes con una frecuencia menor o igual a 6 visitas en el año 2006, presentan mucha variabilidad en su comportamiento. Esto se ve

reflejado en la alta desviación estándar de la variación del comportamiento entre el año 2005 y 2006.

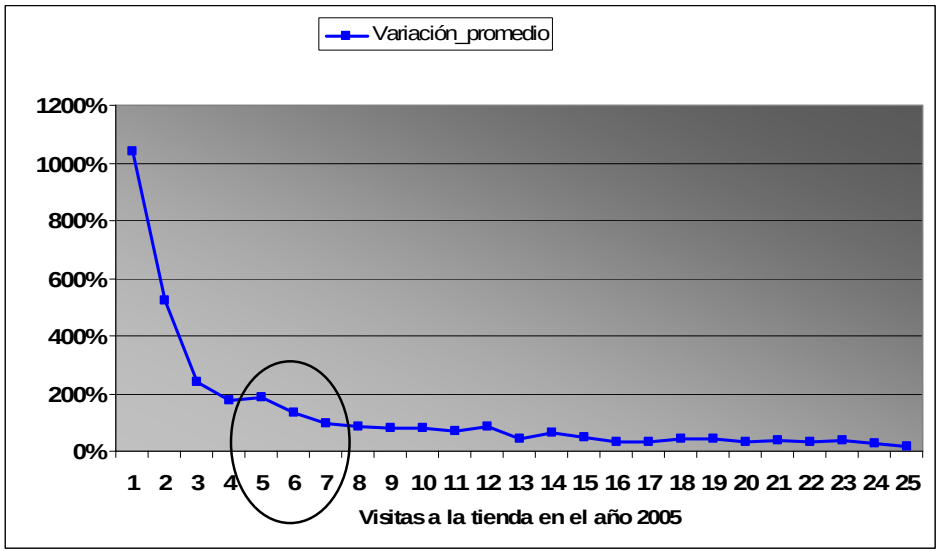


Gráfico 4.4: Variación promedio

En el gráfico 4.4 se puede apreciar que con una frecuencia de visitas al supermercado menor a 6 hay altas variaciones. Los incrementos se hacen más pequeños a medida que aumenta la frecuencia. Por lo tanto, serán filtrados los clientes que tienen menos de 7 visitas en el año al supermercado, correspondiente a 7.096 clientes de los restantes, quedando un 56.5% de los clientes iniciales.

Otra variable para decidir el corte de los clientes será su monto gastado en el año 2005. Para determinar este se verá el comportamiento de estos, con montos menores a $300.000 (en montos de $50.000) analizando su variación promedio y la desviación estándar de esta variación.

Rango Monto	Promedio variación	Desviación variación
[0,50.000)	949.74%	10313.04%
[50.000,100.000)	150.11%	710.56%
[100.000,150.000)	125.97%	530.31%
[150.000,200.000)	83.34%	378.46%
[200.000,250.000)	60.30%	296.23%
[250.000,300.000)	48.54%	262.71%

Tabla 4.3: Estadísticos de tramos por monto

Se puede observar que existe una alta variabilidad para clientes con montos menores a $50.000, luego se filtrarán a los clientes que tengan un monto acumulado menor a $50.000, correspondiente a 252 clientes de los restantes, quedando 22.439 clientes para el análisis. A continuación se expondrá la cantidad de clientes que se mantienen tras cada filtro.

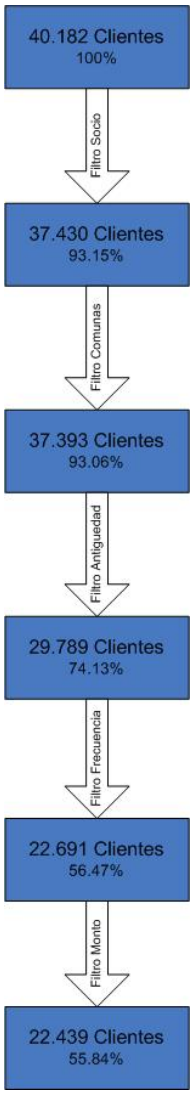


Grafico 4.5: Cantidad de clientes tras filtros

4.4 Transformación de las variables

En esta etapa se propondrá un set de variables potenciales a ser incluidas en los modelos. Este set está compuesto por variables los siguientes tipos:

1. **Tendencia:** Con este subset de variables, se busca captar la pendiente, en la cual se encuentra un cliente, mostrando como ha variado acumuladamente el desempeño de cada cliente en el período de análisis.
2. **Familias[7] de productos:** Con estas variables se buscará identificar la preferencia de compra de los productos por los clientes.
3. **Tier de precios:** Estas variables muestran si un cliente compra productos de alto, medio ó bajo valor.
4. **Variables estáticas:** Estas variables son variables provistas a los largo del período, como por ejemplo las variables RFM.
5. **Georeferenciación:** Con el propósito de incorporar variables demográficas, se observa que la única variable disponible tiene que ver con la ubicación comunal de los clientes.

A continuación se expondrá el detalle de cada variable perteneciente a cada subset definido anteriormente:

4.4.1 Variables de tendencia

Para definir estas variables se utilizaron dos horizontes de tiempo distintos: mensual y trimestral. Para cada horizonte se calcularon las siguientes variables:

- Frecuencia de compra: cantidad de veces que compra cada cliente en el supermercado.
- Recency: días desde la última compra de cada cliente.
- Monto acumulado gastado del período.
- Cantidad de productos diferentes comprados por cada cliente, independiente de la familia de productos al cual pertenezcan.

[7] Agregación de productos dado atributos en común, por ejemplo LACTEOS, ABARROTES.

- Cantidad de unidades independiente de la familia de productos al cual pertenezcan cada unidad.

Una vez obtenidas estas variables para cada período por separado se procedió a hacer una nueva transformación, con el fin de obtener la tendencia llamando a cada variable de la siguiente manera:

- Delta Monto
- Delta Frecuencia
- Delta Productos
- Delta Unidades

Cada una de estas variables se determina con la siguiente transformación:

$$Delta\_X = \cdot \sum_{t=n}^{N} \frac{X(t+1) - X(t)}{X(t)}$$

Donde $X \in$ {Frecuencia, Recency, Monto, Unidades, Productos}. Con los clientes, que tienen antigüedad menor al período de análisis, se encuentra una condición de borde, debido a que un cliente que tiene antigüedad menor al período, iniciará su primera compra con un valor de 100% en todas sus variables Delta, ya que en los períodos anteriores a su primera compra tendrá asignado un 0 en su comportamiento. Es por esto, que para subsanar este problema, se comenzará el cálculo desde la data de antigüedad, es decir, se comienza donde X(t) sea mayor a 0.

En caso de que X(t+1) sea igual a 0 y X(t) sea mayor a 0 se asigna una caída del -100% esto debido a que el cliente compró en el período t pero no lo hizo en el período siguiente. Si el cliente no compró en el período t pero si lo hizo en el período t+1 y además cumple la restricción de antigüedad se asigna una subida del 100%.

4.4.2 Variables de familias de productos

Mediante estas variables se pretende captar la preferencia de cada cliente, por cada una de las diferentes familias. Dado la data disponible existe una agregación de los productos a nivel de sub-línea, línea y familia, siendo cada una de las anteriores, subconjunto de la siguiente.

Detalle	Cantidad
Familias	12
Línea	83
Sub-Línea	360
SKU	4821

Tabla 4.4: Agregación de productos

Se determinó rescatar variables a nivel de familias ya que estas manifiestan una mejor preferencia por parte de los clientes, y además por las cantidades que supone cada agregación. Para cada familia se propuso un subset de variables a nivel de clientes, las cuales se especificarán a continuación:

- Cantidad de productos distintos comprados en el período por familia.
- Cantidad de unidades compradas en el periodo por familia.
- Cantidad de boletas emitidas en el periodo, que contienen a cada familia.
- Monto total gastado en cada familia en el período

A fin de hacer comparables estas variables entre los clientes se deben trasformar a variables porcentuales:

- Porcentaje de unidades compraras sobre el total de las familias
- Presencia de familias en boletas (porcentaje en que cada familia aparece en boletas)
- Porcentaje de monto total gastado por familia
- Porcentaje de productos comprados sobre el total de productos de cada familia.

4.4.3 Variables Tier de precios

Lo que se pretende capturar con esta variable es qué tipo de productos compra cada cliente, es decir, si consume productos de alto, medio ó bajo valor. Para definir un producto como barato, caro o medio se toman todos los SKU agrupados por sub-linea, dividiendo cada sub-linea en 3 grupos, siendo los de mayor valor productos denominados TIER ALTO, los de menor valor denominados de TIER BAJO, y el resto denominados TIER MEDIO. Cada tipo de TIER tendrá la misma cantidad de productos.

Se decidió hacer la agrupación a nivel de sub-línea, de manera de conocer la preferencia de los clientes por productos comparables, por ejemplo un yogurt de una determinada marca con otro de otra marca, y no interesa comparar, por ejemplo, si el cliente decide comprar un yogurt o un queso. Las variables que se determinaron para este subset fueron las siguientes para cada uno de los clientes:

- Cantidad de productos comprados por tier en el período.
- Cantidad de transacciones efectuadas por cada tier en el período.
- Cantidad de unidades (SKU) por tier en el período.
- Monto gastado por cliente por tier en el período.
- Porcentajes de productos comprados por tier en el período.

Estas variables de la misma manera que las variables de familias, se transformaron a variables porcentuales con el fin de hacer comparables a los diversos clientes.

4.4.4 Variables Estáticas

En este subset de variables se incorporan las siguientes variables:

- Promedio de Frecuencia (Mensual y trimestral): Cuantas veces compró cada cliente en cada mes en promedio.
- Promedio de Recency (Mensual y trimestral): esta variable determina cuantos días hay desde la última compra de cada mes, hasta el último día de cada mes.
- Promedio de Monto (Mensual y trimestral): cuanto dinero gastó cada cliente.
- Cantidad de sucursales: en cuantas sucursales distintas compró un cliente.
- Antigüedad, días desde la primera compra, es decir, cuantos días han pasado desde la primera compra del período hasta el último día del período (T).
- Días entre la primera y la última compra: periodo en que el cliente se mantuvo comprando (tx).

4.4.5 Variables de georeferenciación

Con esta variable se pretende incorporar la presencia de la ubicación geográfica de cada cliente con el fin de determinar si influyó en su comportamiento de compra.

4.5 Conocimiento de las variables

En esta etapa se analizan las variables para conocer sus distribuciones. Se hará por cada uno de los subset propuestos en la etapa de transformación.

4.5.1 Variables de familias de productos

Las familias son clasificaciones de los productos que la empresa maneja. En el período de análisis son 12 Familias las que se transan.

Familia de productos	Porcentaje productos vendidos en el período
ABARROTES	23.57%
ALIMENTOS PERECIBLES	19.81%
BAZAR Y PAQUETERIA	3.46%
CAJA INSTITUCIONAL	0.01%
CAJA PRODUCTOS	0.02%
CONFITES	14.25%
CUIDADO DEL AUTOMOVIL	0.07%
CUIDADO PERSONAL	14.51%
GENERICO	0.01%
HOGAR	10.03%
LIQUIDOS	12.94%
MASCOTAS	1.30%

Tabla 4.5: Volumen de familias

Existen familias que poseen un muy bajo volumen transaccional y además existen muy pocos productos en esas clasificaciones de familias. Estas familias son "CAJA INSTITUCIONAL", "GENÉRICO", "CAJA DE PRODUCTO" y "CUIDADO DEL AUTOMOVIL". Las variables asociadas a estas familias no se tomarán en cuentas para el análisis debido al bajo nivel transaccional que presentan.

Se puede observar en la tabla 4.5, cuáles son las familias con mayor cantidad de productos tales como “ABARROTES” y “ALIMENTOS PERECIBLES”, las cuales también tienen presencia transaccional, esto se debe a la cantidad de productos en estas clasificaciones de familias.

En el anexo 6.2 se presentan los principales estadísticos de las variables pertenecientes a las familias de productos. Se infiere que existen familias que presentan valores fuera de rango manifestado por el estadístico Skewness. Esto se debe a que hay pocos clientes que en determinadas familias tienen un consumo alto con respecto al cliente promedio de la empresa. En las familias predominantes se observa lo contrario, es decir, la cantidad de clientes que tienen baja presencia en sus variables de éstas es ínfima.

4.5.2 Variables de tendencia

Como se expuso anteriormente, este tipo de variables se agregan con el propósito de captar la tendencia de compra de los clientes. En la tabla 14, se pueden ver los principales estadísticos de las variables, que buscan captar la pendiente de los clientes, mediante las variables delta.

Variable	Mínimo	Máximo	Promedio	Desviación estándar
Delta F mes	-3.33	25.16	1.08	2.03
Delta monto mes	-4.22	6,581.69	3.03	35.13
Delta prod mes	-4.35	1,179.08	2.70	10.20
Delta unid mes	-4.45	7,598.46	5.11	48.96
Delta F trim.	-2.73	65.65	0.90	2.71
Delta monto trim	-2.89	1,151.74	1.69	12.99
Delta prod trim	-2.87	372.42	1.59	8.29
Delta unid trim	-2.93	3,055.76	2.52	30.57

Tabla 4.6: Descriptivos variables tendencia

Se puede observar que las variables “Deltas” pueden obtener valores bastantes altos. Esto se debe a que son muy sensibles a variaciones con valores pequeños. Por ejemplo, si un cliente compra $5.000 y en otro período compra $50.000 tendrá una variación o delta de 1000%.

4.5.3 Variables de tier de precios

Se puede observar que en promedio, los clientes llevan más productos de tier bajo, a pesar de que en cada agrupación existe una cantidad igual de productos. Esto se debe principalmente a la disponibilidad de productos en góndola.

Variable	Mínimo	Máximo	Promedio	Desviación estándar
% Cant prod tier alto	0%	100%	20.63%	11.54%
% Cant proa tier bajo	0%	100%	52.85%	14.72%
% Cant proa tier medio	0%	100%	26.52%	11.06%
% Presencia tier alto	0%	100%	81.66%	23.46%
% Presencia tier bajo	0%	100%	93.13%	15.95%
% Presencia tier medio	0%	100%	84.02%	22.12%
% Cant unid tier alto	0%	100%	8.64%	10.20%
% Cant unid tier bajo	0%	100%	69.76%	15.94%
% Cant unid tier medio	0%	100%	21.60%	12.49%
% Monto tier alto	0%	100%	24.98%	14.61%
% Monto tier bajo	0%	100%	48.63%	16.48%
% Monto tier medio	0%	100%	26.39%	12.71%

Tabla 4.7: Descriptivos variables de tier de precios

En la tabla 4.7 se pueden ver los principales estadísticos. La presencia en boletas de cada tipo de agrupación de productos es homogénea, aunque la presencia de productos baratos es mayor, y posee una variabilidad inferior entre clientes, esto quiere decir que un cliente promedio en el total de sus boletas lleva de los tres tipos de productos, pero la cantidad de unidades que lleva de productos de tier bajos es superior.

4.5.4 Variables estáticas

El la tabla 4.8 se presentan los principales estadísticos de estas variables:

Variable	Mínimo	Máximo	Promedio	Desviación estandar
Frecuencia/mes	0.08	30.08	3.85	4.60
Recency mensual	0.00	30.50	13.26	9.90
Monto/mes	13.83	34,437,598	135,602	350,785
Frecuencia/trimestre	0.25	90.25	11.10	13.62
Recency trimestral	0.00	91.50	25.92	27.16
Monto/trimestre	41.50	103,312,794	388,372	1,020,980

Tabla 4.8: Descriptivos variables estáticas

Los clientes tienen tendencia a venir a mediados de mes, eso se manifiesta en el recency promedio de 13,26. Un cliente promedio compra $135.602 por mes. Existen 67 clientes con montos superiores a $50 millones, lo que equivale a un 0.17%. Se presume que estos clientes sean institucionales, a pesar de no estar marcados como tal.

4.5.5 Variable Georeferenciación

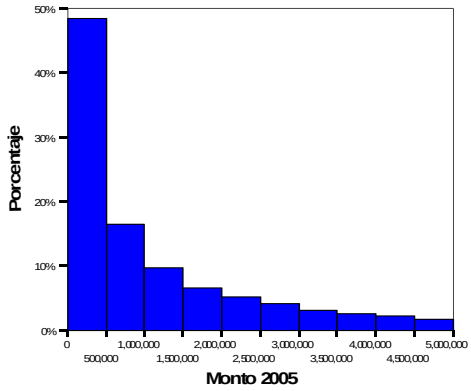
Del total de comunas presentes, existen algunas que tienen una concentración de clientes muy baja. Debido a esto se tomó la decisión de filtrar a las comunas con menos de 50 clientes, quedando un total de 43 comunas a considerar para el análisis.
En la tabla se muestran las 10 comunas que poseen la mayor concentración de clientes. La comuna de SANTIAGO, es la que concentra el mayor número de clientes, un 23.04%.

Comuna	Total
SANTIAGO	23.04%
PUENTE ALTO	9.28%
MAIPU	4.80%
LA PINTANA	4.70%
PUDAHUEL	4.36%
PENALOLEN	4.06%
SAN BERNARDO	3.88%
LA FLORIDA	3.73%
EL BOSQUE	3.63%
CERRO NAVIA	3.56%

Tabla 4.9: Top 10 concentración de comunas

4.5.6 Variable supervisada

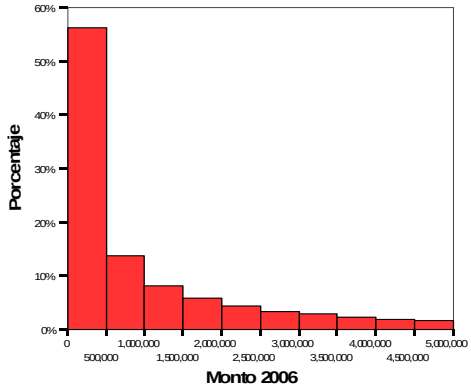
Antes de analizar la variación porcentual del monto acumulado de un año con respecto a otro se analizarán individualmente estos montos.
En el gráfico 10 se puede observar que existe una clara acumulación de clientes, entre montos menores a $500.000 por año. Es decir, clientes que gastan en promedio menos de $42.000 mensuales. Se puede observar la presencia de clientes con montos bastante alto, puesto de que la mediana es la mitad del valor promedio.



Promedio:	$	2,267,335
Desviación:	$	5,097,774
Mediana:	$	1,136,773

Gráfico 4.6: Monto 2005 acumulado por clientes

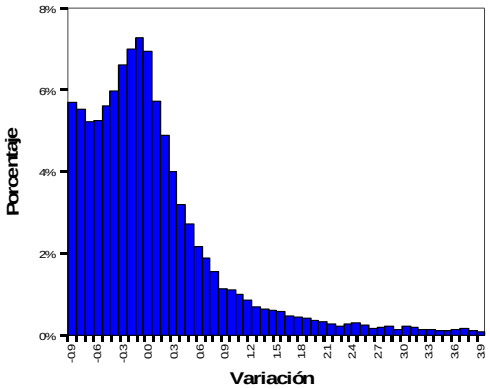
En el gráfico 4.6 se puede observar que aumenta la cantidad de clientes con montos bajos y en general la mayoría baja su comportamiento. Comparando el grafico 4.6 con el 4.7 se observa que hay una tendencia a la baja en los montos de compra acumulados de un año de cada cliente, además se puede ver este efecto en los estadísticos.



Promedio:	$	2,097,385
Desviación:	$	7,099,317
Mediana:	$	890,825

Gráfico 4.7: Monto 2005 acumulado por clientes

En el gráfico 4.7 se puede observar las distribución de los clientes dado por su variación porcentual del monto. Se infiere que la mayoría de los clientes disminuye sus compras de un año con respecto al siguiente. Existe un 15,02% de clientes que en el año 2006, no visitan el supermercado a pesar de si haberlo hecho el año anterior, a estos clientes se les denomina clientes fugados. De los clientes que si visitan la tienda en el año 2006, se infiere que mas del 30% de clientes no compra más de $600.000 en el año es decir, son clientes que no gastan mas de $50.000 mensuales.



Promedio:	0.05
Desviación:	1.46
Mediana:	-0.12

Gráfico 4.8: Variable supervisada: variación monto

En el gráfico 4.8, se puede observar la presencia de clientes que aumentan en grandes proporciones de un periodo con respecto a otro, esto se manifiesta con la mediana de la variable.

En la tabla 4.10 se puede observar que los clientes presentan grandes variaciones.

Variación porcentual	Cantidad clientes	% clientes
Menor 0.1	11683	35.2%
Entre -0.1 y 0.1	3478	10.5%
Mayor 0.1	7278	21.9%

Tabla 4.10: Variación porcentual del monto

4.6 Definición de grupos de clientes

Se definen grupos de clientes para analizar diferentes comportamientos y desempeños de los modelos en cada grupo. Se harán tres diferentes segmentaciones de los clientes, la primera basada en la variable supervisada, la segunda según el tamaño monetario de cada cliente y como tercera se usará la agrupación que utiliza el supermercado. A continuación se expondrán las condiciones que se deben cumplir para pertenecer a cada grupo de cada segmentación:

4.6.1 Variación transaccional

Existen clientes que cambian significativamente su comportamiento de un año con respecto al siguiente. Por ejemplo, un cliente que en el año 1 compra $350.000 en el supermercado, y el año 2 su compra en el mismo supermercado aumenta a $5.400.000. Esta variación corresponde a un 1443%. Es aquí, donde mediante técnicas de clasificación, es posible, bajo un umbral de holgura determinar que clientes tienen mayor propensión a tener cambios abruptos en su comportamiento.

El cambio de comportamiento de los clientes tiene una cota inferior de -100%, es decir, que en el año 2 no efectúe compra alguna en el supermercado en análisis. En cambio, no existe una cota superior definible, ya que, no existe una restricción al monto que un determinado cliente pueda gastar.

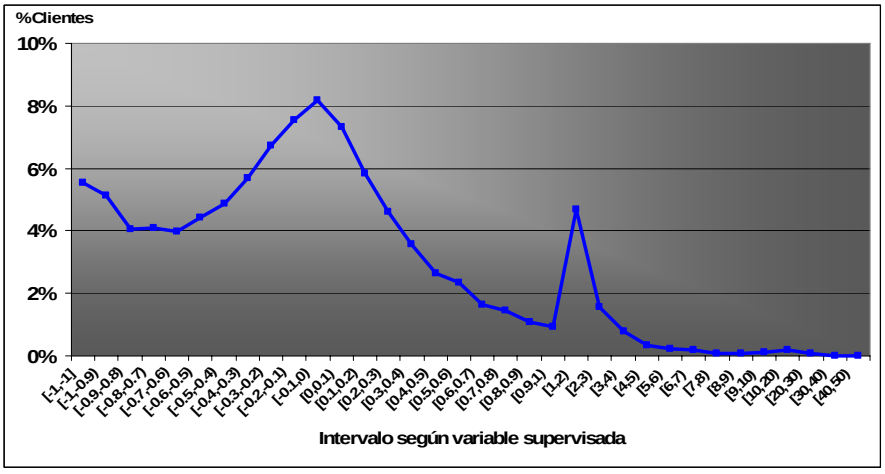


Gráfico 4.9: Concentración de clientes por intervalo de variación

La mayor cantidad de clientes se encuentra concentrada en el intervalo [-1,0]. Se puede ver que en el intervalo [1,2] hay una alta concentración de clientes. Una vez visualizado el contexto en el cual se encuentra este problema, se debe proceder a determinar como predecir si un cliente va a caer en cierto intervalo de variación.

Dado la visión general de la variable, se definen 5 clases:

- **Fuga**: Si un cliente no vuelve a comprar en el año 2006, es decir su variación es -100%.
- **Baja**: si un cliente si visita el supermercado, pero el monto que compran es inferior al del año anterior. Si su variación es superior a -100% e inferior a -35%.
- **Mantiene**: Si un cliente compra un monto similar al del año anterior. Si su variación es mayor o igual a -35% y inferior a 5%.
- **Sube**: Si un cliente compra más que en el año anterior. Si su variación es igual o superior a 5% e inferior o igual a 100%.
- **Multiplica**: Si un cliente aumenta el valor monetario de sus compras del año 2006 considerablemente, con respecto a su comportamiento en el año 2005, es decir, si su variación es mayor a 100%.

Clase	Clientes	%
FUGA	1244	5.54%
BAJA	6568	29.27%
MANTIENE	6604	29.43%
SUBE	6146	27.39%
MULTIPLICA	1877	8.36%

Tabla 4.11: Clasificación según variación

Los puntos de corte entre las clases “baja”, “mantiene” y “sube” se determinaron de manera que en cada una de esas tres clases haya una cantidad de clientes lo más homogénea posible.

4.6.2 Tamaño de cliente

Mediante el método no supervisado para clasificación denominado K-medias se procedió a segmentar a los clientes, tomando como variable discriminadora el monto acumulado del año 2005. Luego, usando los cortes definidos, para el año 2005, se

incorporó otra clasificación dado el monto del año 2006 de cada cliente. Mediante este algoritmo se definieron 3 clases, Grande, mediano y pequeño.

Monto 2005				
Tipo Cliente	**Mínimo**	**Máximo**	**Promedio**	**Cantidad clientes**
PEQUEÑO	50,067	1,694,755	635,706	13,567
MEDIANO	1,695,041	5,080,464	3,012,030	6,236
GRANDE	5,085,053	413,251,177	8,903,300	2,636

Tabla 4.12: Cortes para las cluster dado el tamaño

En la tabla 4.12 se muestran los criterios de corte para cada clase definida según el algoritmo de clasificación K-medias.

	Año 2005		**Año 2006**	
Tipo Cliente	**Cantidad clientes**	**% clientes**	**Cantidad clientes**	**% clientes**
PEQUEÑO	13567	60.46%	14481	64.53%
MEDIANO	6236	27.79%	5510	24.56%
GRANDE	2636	11.75%	2448	10.91%

Tabla 4.13: Concentración según año

Se observa en la tabla 4.13, que en el año 2006 existen una mayor cantidad de clientes menores que en el año anterior. Más adelante se analizarán las transiciones que suceden.

4.6.3 Clientes Apóstoles

Las reglas que debe cumplir un cliente para ser definido como Apóstol son las siguientes dentro de un período de tres meses:

1. Monto promedio mensual del período, sea mayor a $100.000.
2. Monto promedio de los dos últimos meses, sea mayor a $100.000.
3. Frecuencia de compra mensual mayor a 3 veces.

Dentro del año 2005 se tienen 10 períodos en los cuales se puede evaluar si un cliente cumple la condición de apóstol, para el análisis se considerarán apóstoles, aquellos clientes que lo han sido en algún período

Los clientes que cumplen alguna vez las condiciones de Apóstol corresponden a un 56,38 % de los clientes en análisis.

Clasificación	Cantidad	%
Nunca Apóstol	9787	43.62%
Apóstol alguna vez en el período	12652	56.38%

Tabla 4.14: Tipos de clientes

La mayor parte de los clientes que han sido apóstoles, lo han sido en todos los períodos que se han evaluado a lo largo del año 2005. En el gráfico 4.9 se puede apreciar la distribución de los clientes dado su condición de apóstol.

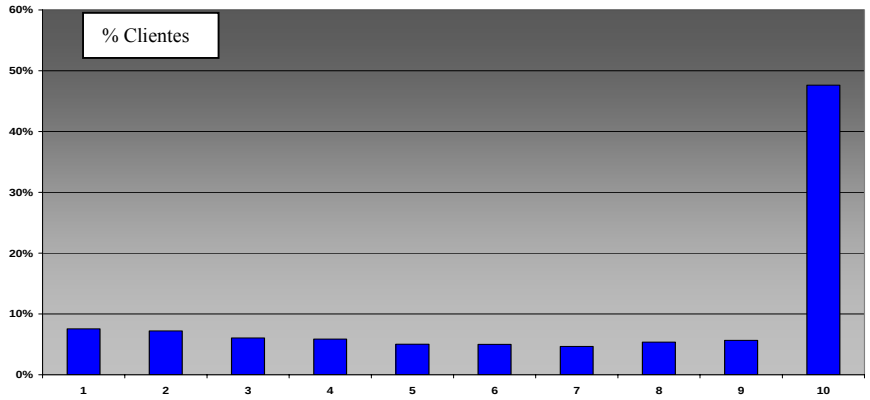


Gráfico 4.10: Cantidad de clientes por veces Apóstol

4.7 Mapeo de clientes según variables

A continuación se hará un análisis bi-variado, para las variables de tipo tier de precios, las variables estáticas y las variables deltas con el fin de visualizar en un espacio de dos dimensiones como las variables discriminan entre cada cliente dado por la agrupación del tamaño de éstos.

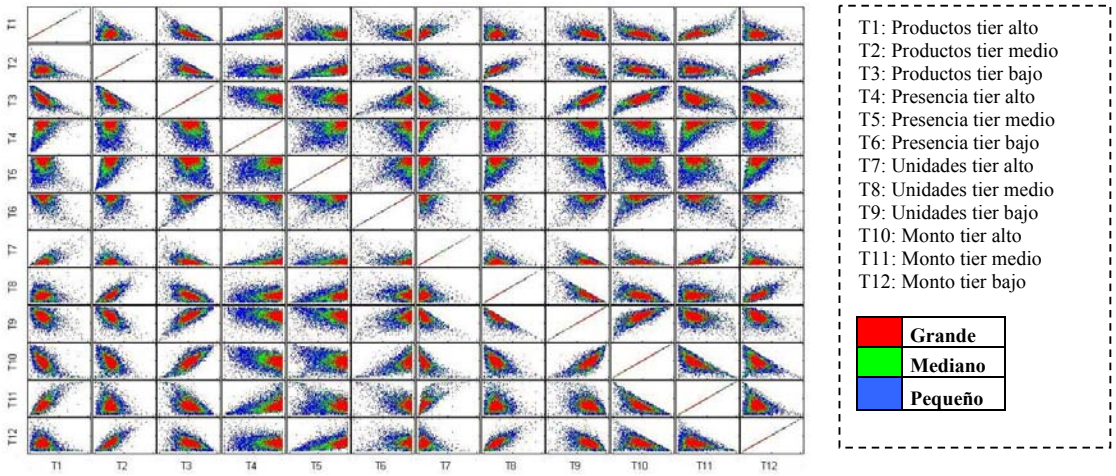


Grafico 4.11[8]: Variables tier

Se puede observar en el gráfico 4.11 que en todas las relaciones que se forman con las variables tier es posible ver que cada grupo se encuentra concentrados, lo cual manifiesta la importancia de este tipo de variables.

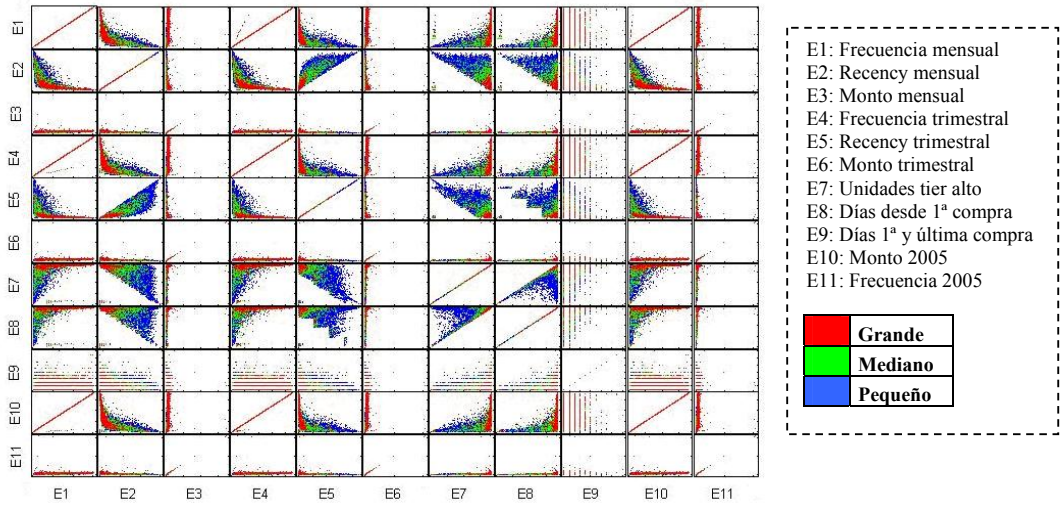


Grafico 4.12: Variables estáticas

[8] Este tipo de grafico permite mostrar el poder discriminador en dos dimensiones de las variables sobre grupos definidos.

En el gráfico 4.12 se puede observar que existe al igual que en las variables de tier de precios una clara discriminación entre los grupos. Esto manifiesta la importancia de este tipo de variables.

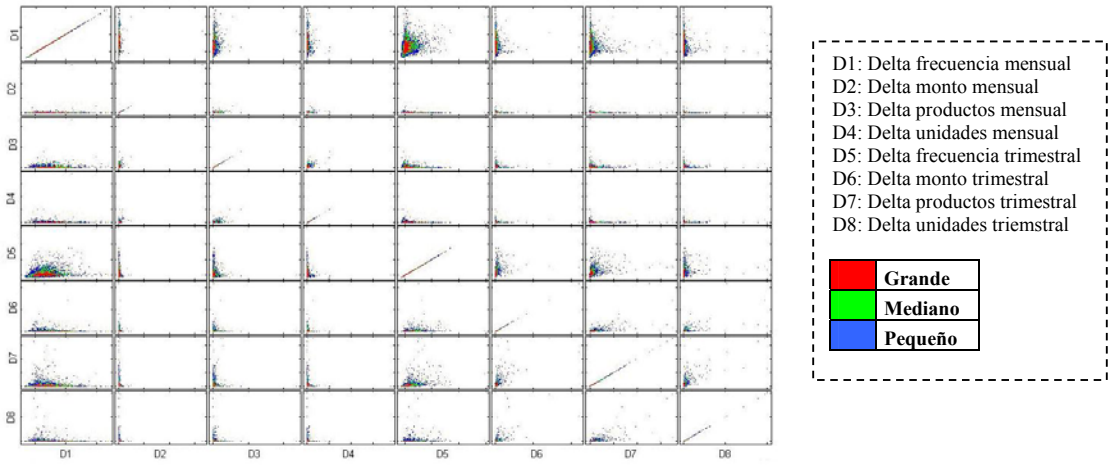


Grafico 4.13: Variables delta

En el gráfico 4.13 se puede ver como las variables delta discriminan entre los grupos, dado el tamaño de los clientes. Se infiere que en la mayoría de las relaciones los clientes se encuentran muy concentrados en rangos muy pequeños, es decir, se observa diferencia entre los grupos, pero con menos definición dado que varios clientes presentan los mismos resultados en sus variables.

4.7.1 Análisis de transiciones de los clientes

Una vez definido las clasificaciones para los clientes se procederá a ver las transiciones en los períodos de análisis. Existen clientes, que en un período compran en el supermercado un monto y en el año siguiente cambian ese comportamiento, en la tabla 4.15 se muestra la cantidad de clientes que se involucran en una transición:

		2006			
		PEQUEÑO	MEDIANO	GRANDE	Cantidad clientes
2005	PEQUEÑO	91.51%	8.06%	0.43%	13,567
	MEDIANO	30.66%	61.59%	7.75%	6,236
	GRANDE	5.84%	21.81%	72.34%	2,636

Tabla 4.15: Porcentaje de clientes en cada transición.

Se puede observar que existe una pequeña cantidad de clientes que pasan de ser clientes Grandes a Pequeños y viceversa. En general los clientes mientras más pequeños son más propensos a mantener su estado.

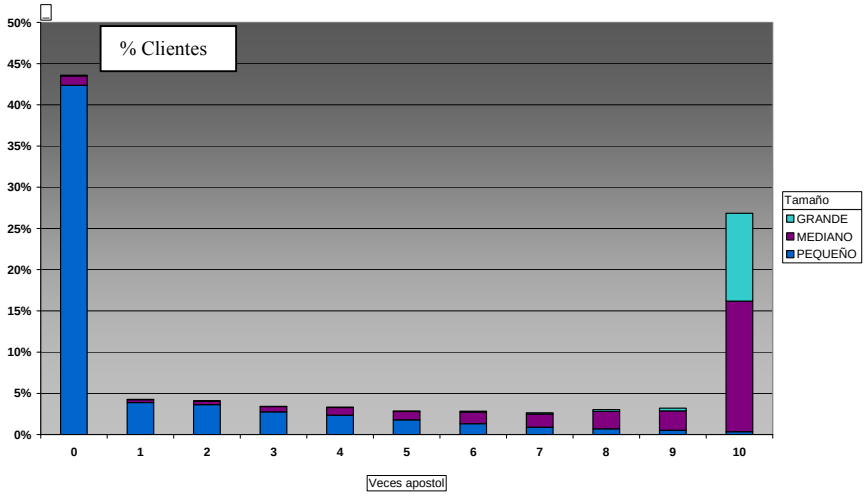


Gráfico 4.14: Clientes según tamaño y veces apóstol

En el gráfico 4.14, se puede observar que los clientes considerados pequeños no han sido nunca apóstoles y que los clientes que han sido apóstoles todo el período, son Medianos ó Grandes. Esto ocurre por el criterio de definición de apóstoles.

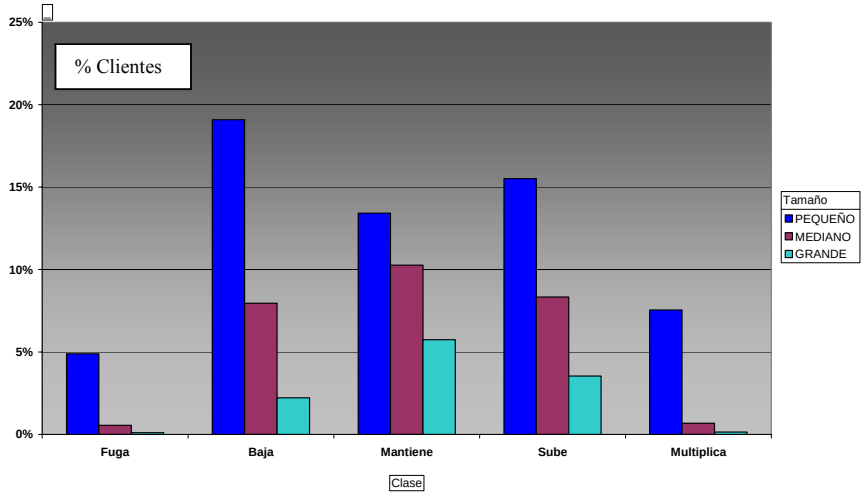


Gráfico 4.15: Clientes según tamaño y clase

En el gráfico 4.15 se puede observar que los clientes Grandes y Medianos tienden a concentrar entre las clases más estables es decir Baja, Mantiene y Sube.

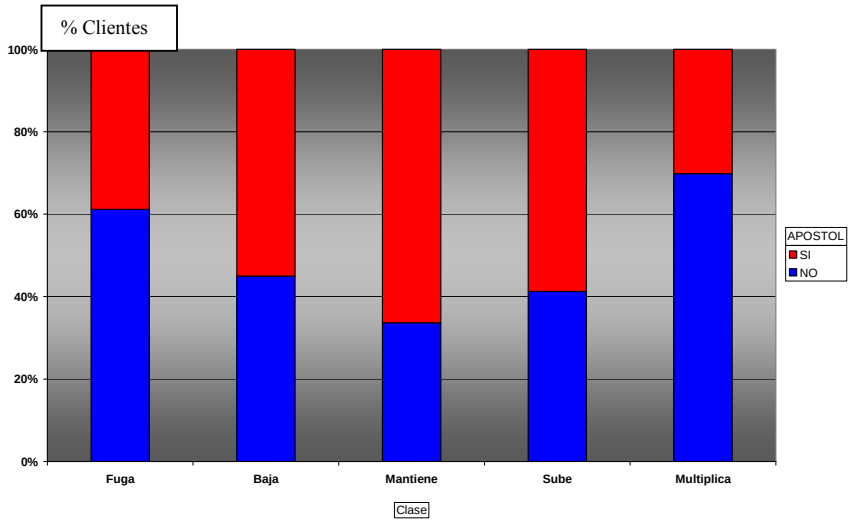


Gráfico 4.16: Clientes según tamaño y categoría apóstol

Predomina la concentración de clientes apóstoles a medida que la clase es más estable. En el gráfico 4.16 se puede observar que los más estables son los clientes que han sido todo el período apóstoles.

4.7.2 Clientes fugados

La mayor concentración de clientes fugados corresponde a clientes de bajo monto, es decir, clientes considerados pequeños. En el gráfico 4.17, se puede observar que el 60% de los clientes que se fugan, corresponde a clientes que no son apóstoles, y además que son de bajo monto.

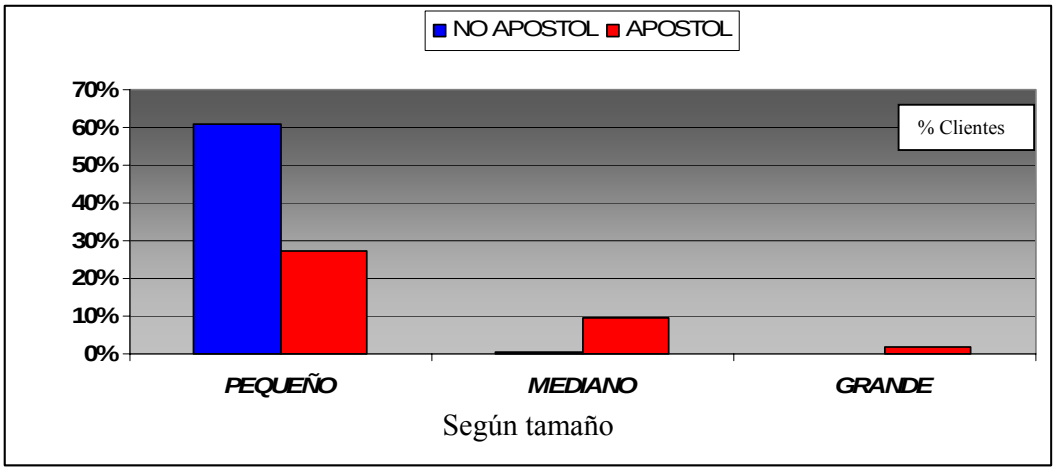


Gráfico 4.17: Clientes Fugados según tamaño y condición de Apóstol.

En resumen, existe una masa de clientes, que tienen cambios en su comportamiento de compra. Este comportamiento, debe ser captado por los modelos para poder identificar clientes que son potenciales a tener cambios bruscos de comportamiento. Para ello, es imprescindible encontrar las variables que mejor puedan captar este fenómeno, eliminando aquellas que generen ruido en los resultados.

4.8 Selección de variables

Se tiene un total de 117 variables. Cada una de éstas pertenece a un subset de variables definido. En esta etapa se busca reducir el espacio propuesto, de manera que las variables que se seleccionen, sean las que expliquen de mejor manera el fenómeno.

Inicialmente se utilizarán métodos filtros[9] para analizar de manera uni-variada las relaciones, entre las variables predictores y la variable supervisada, y posteriormente se utilizará un método embedded, que permita seleccionar las variables.

4.8.1 Análisis de correlaciones[10]

Este análisis permite ver la dependencia lineal que existe entre dos variables. Mediante el coeficiente de correlación de Pearson, es posible analizar la correlación entre dos variables, sin que influya el rango en el que oscila cada variable. Sólo es posible hacer este análisis con variables numéricas.

En las variables de precios se observa, que existe una alta correlación positiva entre la cantidad de productos y la cantidad de unidades. Además, existe una alta correlación negativa, entre la cantidad y unidades de productos de tier alto con respecto a los de tier bajo, esto quiere decir, que un cliente que compra muchos productos de tier alto, su comportamiento de compra de productos de tier bajo es ínfimo, y viceversa.

[9] Ver marco teórico
[10] Ver anexo para detalle de correlaciones.

En las variables de tipo familia ocurre el mismo fenómeno, es decir, existe una mayor correlación positiva, entre la cantidad de productos con la cantidad de unidades de cada familia. De esto se infiere, que un cliente que compra gran variedad de productos en una familia, también lo hace en grandes cantidades.

En las variables delta se observa, que existe una alta correlación entre las variables monto cantidad de unidades y cantidad de productos distintos para los horizontes de tiempo mensual y trimestral. No ocurre lo mismo con la variable delta frecuencia, la cual no tiene una alta correlación con sus pares.

En las variables estáticas se observa una alta correlación negativa entre las variables frecuencia y recency, esto se debe a que un cliente que tienen una alta frecuencia de compra, se mantiene visitando la tienda en todo el período, con lo que tendrá un recency bajo, es decir, hace poco tiempo que visitó por última vez el supermercado. Otra correlación que se observa, es entre los días desde la primera compra y los días entre la primera y la última compra. Esto se debe a que si un cliente vino por primera vez hace un largo tiempo existe mayor probabilidad que su período de actividad sea grande.

Una vez analizadas las correlaciones, es posible determinar a priori, que dos variables que tienen alta correlación, probablemente están explicando lo mismo. Esto permite seleccionar sólo una de ellas. Para tomar la decisión de cuál elegir, más adelante se usarán métodos Embedded que permitan definir que variable es mejor.

4.8.2 Análisis Chi-Cuadrado

Mediante este test es posible analizar la dependencia entre variables nominales. Es por ello que este contraste de hipótesis se usará para evaluar la significancia de las variables de georeferenciación, con el fin de generar un ranking. La variable comuna se testea con la clasificación del tamaño de clientes. Mediante esa clasificación los resultados de este test son los siguientes:

Test	Valor	g.l.	p-valor
Chi-Cuadrado	2681	162	0

Tabla 4.16: Resultado Chi Cuadrado

Los grados de libertad (g.l.) indican cuantas comunas diferentes se están analizando. Se observa en la tabla 4.16, que el valor del Chi-Cuadrado permite rechazar la hipótesis nula de independencia de las comunas con el tamaño de los clientes. Es decir, esta variable permite discriminar a los clientes según su tamaño monetario. Para ver cuales son las comunas que mejor discriminan se procedió a separar estas variables en Dummies[11].

	Chi-cuadrado			Promedio monto	
Comuna	**Estadístico**	**g.l.**	**p-valor**	**No pertenece a comuna**	**Pertenece a comuna**
SANTIAGO	809.44	2	0.00	2,252,102	1,424,718
PUENTE ALTO	140.90	2	0.00	2,026,870	2,636,616
MAIPU	132.09	2	0.00	2,133,803	1,272,692
MALLOCO	101.14	2	0.00	2,074,918	2,580,066
LA PINTANA	35.84	2	0.00	2,072,275	2,461,931
PUDAHUEL	34.01	2	0.00	2,085,511	2,304,082
LA FLORIDA	32.78	2	0.00	2,081,662	2,456,597
LO PRADO	23.73	2	0.00	2,091,595	2,438,037
RENCA	21.97	2	0.00	2,095,887	2,227,674
LA GRANJA	20.91	2	0.00	2,085,139	2,467,864

Tabla 4.17: Comunas que discriminan dado el chi-cuadrado

En la tabla 4.17 se muestra el TOP 10[12] de las 22 comunas que rechazan la hipótesis nula. Se observa que este test es sensible a la cantidad de casos asociados a cada comuna, es decir, a mayor cantidad de clientes, existe mayor probabilidad de rechazar la hipótesis nula de independencia. Para cada comuna existen 2 grados de libertad, es decir, presencia o no de un cliente en una comuna.

4.8.3 Análisis ANOVA de una vía

Mediante el análisis ANOVA de una vía, se procederá a ver si existe diferencia entre las medias de los clientes pertenecientes a cada una de las comunas, con respecto a la variable variación del monto. Existen 11 comunas que según el test ANOVA

[11] Variables que toman valor 1 si es que el cliente pertenece a la comuna cuyo nombre toma la variable y en caso de no pertenecer toma valor 0. Cabe mencionar que un cliente sólo pertenece a una comuna.

[12] En el anexo se pueden ver los resultados de todas las comunas que rechazan la hipótesis nula con un 5% de confianza.

discriminan entre grupos. En la siguiente página se exponen los resultados de aquellas comunas que discriminan según la variable supervisada.

	ANOVA		Promedio monto		
COMUNA	**F**	**p-valor**	**No pertenece a comuna**	**Pertenece a comuna**	**Clientes**
MAIPU	32.35	0.00	4.09%	31.69%	949
RECOLETA	26.57	0.00	4.68%	49.41%	288
SANTIAGO	22.68	0.00	7.49%	-4.45%	4196
PENALOLEN	13.45	0.00	4.83%	35.33%	314
LAS CONDES	12.41	0.00	5.15%	117.74%	21
CONCHALI	11.35	0.00	4.61%	22.15%	820
RENCA	7.75	0.01	5.55%	-20.12%	255
LA REINA	6.57	0.01	5.17%	80.26%	25
INDEPENDENCIA	5.52	0.02	5.08%	34.66%	136
MALLOCO	5.07	0.02	5.73%	-4.95%	998
MACUL	4.69	0.03	5.19%	74.38%	21

Tabla 4.18: Comunas que discriminan dado ANOVA

4.8.4 Análisis de componentes principales

Mediante la metodología propuesta por Montoya en [11], es posible determinar un ranking de variables. Este ranking se determina por la cantidad de varianza que explica cada variable.

Al efectuar componentes principales con 20 factores se explica el 86.16% de la varianza total[13]. Con estos factores se determina un rating, el cual permite generar un ranking de las variables cuantitativas.

Variable	Rating
Productos tier bajo	13.26
Monto Alimentos Perecibles	13.23
Monto tier bajo	13.21
Productos Alimentos Perecibles	12.91
Presencia Alimentos Perecibles	12.70
Unidades tier bajo	12.66
Unidades Alimentos Perecibles	12.30
Transacciones tier bajo	12.23
Unidades tier alto	11.97
Presencia Hogar	11.58

Tabla 4.19[14]: Top 10 según PCA

[13] Mediante SPSS 13.0 se determinan los componentes principales necesarios.

Este método no incorpora la relación con la variable supervisada, pero aún así permite rankear a las variables, según la variabilidad que explica cada una de ellas, eliminando aquellas variables que se pueden expresar como una combinación lineal de las otras.

4.8.5 Método Embedded

Mediante la calibración de un árbol de decisión se procede a reducir el espacio de las variables seleccionadas. Dentro de cada subset de variables se calibran árboles de decisión, de manera de generar un ranking. El árbol de decisión genera un valor para cada variable. Valores mayores indican que la variable tiene mayor importancia relativa versus otras con menor valor. La suma de este valor, (de los valores) que asigna cada árbol a cada variable es 100.

Para las variables de tier de precios se validaron 3 árboles de decisión, supervisando la variable objetivo (variación porcentual del monto). Uno distinto para cada tipo de tier. De esta manera, el valor promedio de cada variable de los árboles construidos indicará el ranking de este tipo de variables.

	Tier de precios				
	Alto	**Medio**	**Bajo**	**Promedio**	**Ranking**
Monto	28.7	29.4	27.9	28.7	**1**
Presencia	26.1	20.3	27.9	24.8	**2**
Unidades	22.5	23.8	23.0	23.5	**3**
Productos	22.7	26.5	21.2	23.1	**4**

Tabla 4.20: Ranking de variables tier según árbol de decisión

En la tabla 4.20 se puede observar que el monto es la mejor variable para el tier de precios, seguida por la presencia. De la misma manera, para las variables de familias se construye un árbol para cada una de las familias, de manera de generar otro ranking para estas variables.

	Familias de productos									
	Abarrotes	**Alimentos Perecibles**	**Bazar y Banquetería**	**Confites**	**Cuidado Personal**	**Hogar**	**Líquidos**	**Mascotas**	**Promedio**	**Ranking**
Presencia	26.1	21.9	46.0	26.1	20.8	25.5	26.0	23.2	27.0	**1**
Monto	23.7	27.3	16.5	26.8	29.5	26.1	23.8	40.2	26.7	**2**
Unidades	24.0	27.4	22.7	24.8	25.5	24.3	25.8	20.8	24.4	**3**
Productos	26.2	23.4	14.9	22.4	24.3	24.1	24.4	15.8	21.9	**4**

Tabla 4.21: Ranking de variables familia de productos

[14] En el anexo se muestra el ranking completo de las variables.

Dado los resultados de los árboles de decisión, mostrados en las tablas 4.20 y 4.21, las variables que tienen mejor poder discriminador para los subset de variables tier de precios y familias de productos, son el monto y la presencia.

Para seleccionar el horizonte de tiempo, se evaluará cada árbol de manera que, compare cada variable de tipo mensual con la misma de tipo trimestral. Esto se aplicará a las variables estáticas y a las variables Delta.

	Delta					
	Frecuencia	**Monto**	**Productos**	**Unidades**	**Promedio**	**Ranking**
Trimestral	81.0	72.3	85.9	78.7	79.5	**1**
Mensual	19.0	27.7	14.1	21.3	20.5	**2**

Tabla 4.22: Ranking de variables delta según horizonte de tiempo

	Estáticas		
	Frecuencia	**Monto**	**Recency**
Trimestral	48.4	55.0	33.6
Mensual	51.6	45.0	66.4

Tabla 4.23: Ranking de variables estáticas

En la tabla 4.23 se puede observar que las variables de tipo trimestral, siempre son superiores a las de tipo mensual, mientras que en la tabla 24 la variable frecuencia y recency, son mejores mensualmente que trimestralmente, a diferencia de lo que ocurre con la variable monto. A continuación se hará el análisis con todas las variables de manera de generar un ranking general de estas.

Variable	Rating	Ranking
Antigüedad del año (T)	36.60	**1**
Delta Monto Trim	9.30	**2**
Delta Prod Trim	6.90	**3**
Monto Trim	6.00	**4**
Dias desde la 1ª a la última compra (tx)	5.50	**5**
Delta Unid Trim	2.00	**6**
Delta Monto Mes	1.70	**7**
Delta F Mes	1.40	**8**
Recency Trim	1.10	**9**
Recency Mes	1.00	**10**

Tabla 4.24[15]: Top 10 Ranking de todas las variables

[15] En el anexo se expone el ranking completo

Se puede observar en la tabla 4.24 la importancia de las variables días desde la última compra y días entre la primera y última compra de cada cliente. Con las variables seleccionadas se calibrarán en primera instancia el árbol de decisión, de manera que las variables seleccionadas por éste sean las variables inputs para los modelos MLP's.

En resumen, mediante los métodos empleados para analizar la importancia de cada una de las variables, se tiene una visión clara de la importancia relativa que tiene cada variable dentro de cada subset de éstas. Por otro lado, el método de filtros PCA busca las variables que explican la mayor variabilidad total de los datos. Dado los criterios usados se han seleccionado 31 variables continuas de las 62 que se tenían inicialmente.

Subset	Observación	Cantidad de variables
Variables de tier de precios	Monto y Presencia	6
Variables de familias	Monto y Presencia	16
Variables estáticas	F. mensual, M. trimestral, R. Mensual, tx y T	5
Variables delta	Trimestrales	4
	Total General	**31**

Tabla 4.25: Variables seleccionadas

En el capitulo de elaboración de modelos se construirá un árbol que incluya todos todas las variables seleccionadas de cada subset, es decir, las 31 variables que se exponen en la tabla 4.25 más las 11 variables de georreferenciación.

4.9 Construcción de modelos

Una vez definidas las variables predictoras, se procede a desarrollar los modelos con el fin de pronosticar de la manera más precisa la variable supervisada. Para calibrar los modelos, se usará una muestra de clientes llamada Train y para testear los resultados de la calibración se usará una muestra de testeo llamada Test. La muestra Train que se usará[16] para calibrar a los modelos es del 80% de los clientes y la muestra Test corresponde al 20% de los clientes restantes.

Muestra	Clientes	%
Train	17925	80%
Test	4514	20%

Tabla 4.26: Tamaño de muestras

Para asegurar la generalización de los modelos construidos, se debe hacer validación cruzada[17], de manera que cada modelo se construirá a partir de tres muestras distintas de train, con sus respectivas muestras de test.

4.9.1 Modelo Pareto /NBD

Este modelo, mediante distribuciones de probabilidad, predice la frecuencia futura de cada cliente. Para calibrar este modelo se necesitan 3 variables, tx T[18] y frecuencia de compra de los clientes, para posteriormente, con el monto promedio de cada boleta estimar el monto que gastará el cliente en el período futuro. Las variables se deberán transformar de diarias a mensuales, debido a problemas de divergencia en el cálculo numérico al aumentar el número de transacciones de un cliente sobre 120. La presencia de valores tan elevados se debe al horizonte de tiempo seleccionado para el análisis, correspondiente a un año completo. Este modelo, a diferencia de los de data mining, se calibra con la muestra seleccionada y no requiere una muestra de testeo. Para la

[16] Aleatoriamente se definirá la asignación de cada cliente a cada muestra.
[17] Validación cruzada, es decir, probar los modelos con diferentes muestras.
[18] T: Días desde la primera compra de cada cliente.
tx: días entre la primera y la última compra de cada cliente.

calibración, se deben usar todos los clientes que se quiere analizar, convergiendo a un óptimo.

4.9.2 Modelos de data mining

1. **Árboles de decisión:** Este modelo mediante la calibración de los parámetros ,permite clasificar o predecir un valor continuo, árbol de clasificación o árbol de predicción.
    a. **Árbol de clasificación:** Los parámetros a calibrar son los siguientes: Criterio de ramificación, número de casos mínimos en los nodos terminales y largo del árbol.

    Para definir el número de casos en el último nodo se hizo un análisis de sensibilidad entre 10 y 160 nodos.

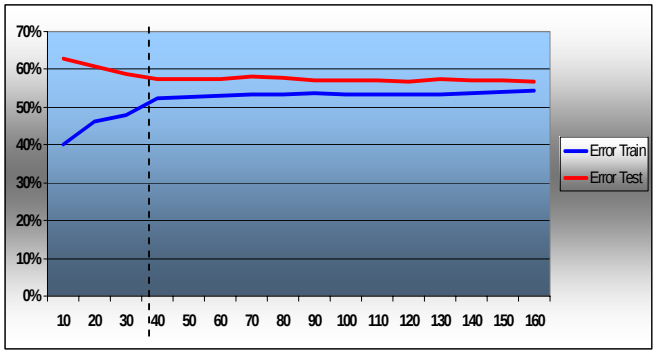


Gráfico 4.1823: Errores dado casos en nodos terminales

Se puede observar en el gráfico 4.18, que definiendo como mínimo 40 casos en los nodos terminales los errores se estabilizan.

A continuación se procederá a calibrar el criterio de ramificación y el largo máximo del árbol. Para definir el criterio de ramificación se usarán los más comunes: Ganancia de información, Chi-cuadrado y Gini.

Chi-cuadrado				
Largo	**Error Train**	**Error Test**	**Casos**	**Nodos**
11	52.91%	57.73%	40	219
10	53.31%	56.91%	40	169
9	53.99%	57.07%	40	137
8	**54.60%**	**55.91%**	**40**	**93**
7	56.01%	58.09%	40	45
6	56.99%	58.70%	40	33
5	58.88%	58.99%	40	17
4	61.42%	61.32%	40	11
3	63.20%	64.40%	40	7

Tabla 4.27: Resultado Chi-cuadrado

Ganancia de información				
Largo	**Error Train**	**Error Test**	**Casos**	**Nodos**
10	52.32%	57.31%	40	253
9	53.47%	57.80%	40	173
8	54.77%	57.95%	40	115
7	55.56%	57.71%	40	57
6	**56.41%**	**57.16%**	**40**	**31**
5	57.27%	57.82%	40	19
4	58.05%	58.75%	40	11
3	61.05%	62.94%	40	7

Tabla 4.28: Resultado Ganancia de información

Gini				
Largo	**Error Train**	**Error Test**	**Casos**	**Nodos**
15	47.43%	59.97%	40	495
14	47.79%	59.23%	40	441
13	48.25%	59.64%	40	409
12	48.95%	58.46%	40	353
11	49.89%	58.37%	40	303
10	51.10%	58.66%	40	249
9	52.43%	57.60%	40	181
8	53.25%	58.19%	40	111
7	**54.63%**	**56.62%**	**40**	**69**
6	55.73%	56.80%	40	35
5	56.44%	56.07%	40	19

Tabla 4.29: Resultado Gini

En las tablas 4.27, 4.28 y 4.29, se pueden ver los resultados de la sensibilización para cada uno de los tres criterios de poda. Se observa que el mejor resultado, se obtiene con el criterio Chi-cuadrado con un largo 8 y con 40 clientes como mínimo en los nodos terminales.

b. Árbol de regresión: Para calibrar este tipo de árboles bajo el criterio de mínima varianza, se debe seleccionar el número de casos mínimos en los nodos terminales y el largo máximo del árbol.

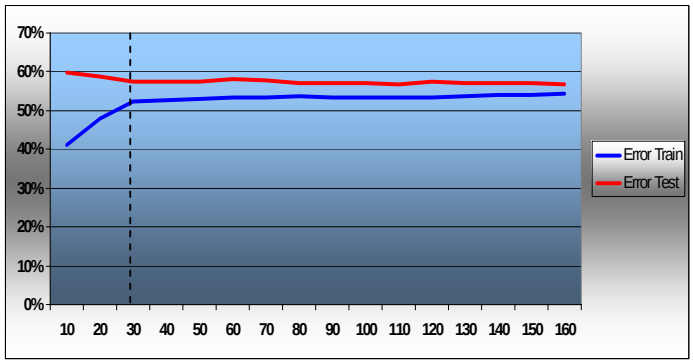


Gráfico 4.19: Errores dado casos en nodos terminales

Se puede observar en el gráfico 4.19, que definiendo como mínimo 30 casos en los nodos terminales los errores se estabilizan. Mediante el error MAPE se define el largo del árbol. Se observa en el gráfico 25, que con un largo 8 para el árbol se obtiene el error MAPE más bajo.

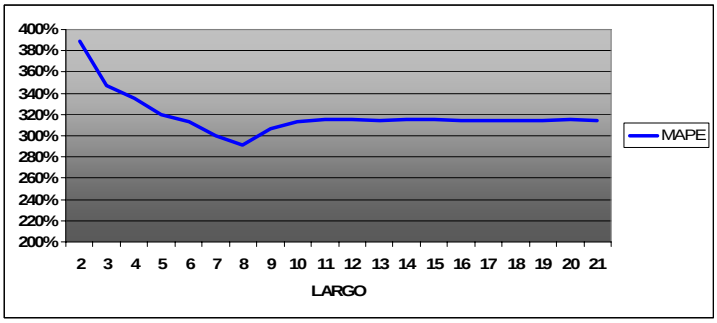


Gráfico 4.20: Sensibilidad de largo del árbol

Las variables usadas por los árboles se muestran en la tabla 4.30.

Posición	Variable
1	Antigüedad de compra (T)
2	Delta Productos trimestral
3	Delta Frecuencia trimestral
4	Monto trimestral
5	Delta Unidades trimestral
6	Días entre primera y ultima compra (tx)
7	Delta Monto trimestral
8	Monto Cuidado Personal
9	Frecuencia Mensual
10	Presencia Hogar
11	Transacciones Tier Alto
12	Monto Alimentos Perecibles
13	Monto Tier Alto
14	MAIPU
15	SANTIAGO

Tabla 4.30: Variables usadas por árbol

Estas serán las variables que serán incorporadas en los modelos MLP's, cabe destacar la reducción de variables efectuada gracias al modelo de árbol de decisión.
La posición indica descendentemente la importancia relativa entre las variables que el árbol en su operatoria manifestó.

2. **MLP:** Al igual que los árboles de decisión, con una red neuronal del tipo perceptrón multicapa (MLP), es posible clasificar o predecir un valor continuo. Para calibrar la red se deben estandarizar los valores. Por lo tanto, se normalizaron todas las variables, al igual que la variable supervisada en el rango 0 y 1 mediante la siguiente trasformación:

$$Z_i = \frac{X - \min(X_i)}{\max(X_i) - \min(X_i)}$$

   Una vez estandarizadas todas las variables, es posible proceder a la calibración. Se analizará(los donde) tipos de MLP, es decir, clasificación y regresión.

   a. **MLP clasificación:** Para seleccionar la mejor red se calibraron los siguientes parámetros: cantidad de neuronas en la capa oculta, número de épocas y la tasa de aprendizaje. La clasificación corresponde a las 5 clases que se han definido (Fuga, Baja, Mantiene, Sube y Multiplica),

por lo tanto, se debe tener 5 neuronas de salida, cada una correspondiente a cada clase. Se hizo análisis de sensibilidad de la cantidad de neuronas (5 a 20), de la tasa de aprendizaje (0.001, 0.01 y 0.1) y de las épocas (500 y 1000). En total se evaluaron 96 redes. Las funciones que se usaron fueron sinusoidal debido al escalamiento de las variables entre 0 y 1.

Parámetro	MLP CLASIFICACION
EPOCAS	1000
NEURONAS	10
TASA APRENDIZAJE	0.1

Tabla 4.31: Configuración seleccionada

Como medida, para seleccionar la mejor configuración, se usó el porcentaje de acierto, que en este caso para la mejor red fue bastante bajo (37,75%). De todas maneras, este es un paso intermedio para posteriormente predecir el valor continuo. En el punto 3 de este apartado se mostrará dicha metodología.

b. MLP de regresión: Mediante esta red, se estimará el valor exacto de la variable supervisada para cada cliente. Para conseguir la mejor red se deben evaluar diferentes combinaciones de los parámetros que se deben configurar, ya sea la tasa de aprendizaje, las épocas y las neuronas en la capa oculta. Se usará una neurona de salida, la cual está formada por el vector de salida (respuesta de cada uno de los clientes). Mediante un análisis de sensibilidad de los parámetros, cantidad de neuronas (5 a 20), tasa de aprendizaje (0.001, 0.01 y 0.1) y de las épocas (500 y 1000) se determina el modelo MLP regresión con mejores resultados. En total se deben evaluar 96 redes, es decir, todas las combinaciones posibles. Una vez obtenido los resultados, la variable supervisada se debe convertir a su escala natural, de tal manera, de poder hacer una comparación con el valor real.

Parámetro	MLP REGRESIÓN
EPOCAS	1000
NEURONAS OCULTAS	15
TASA APRENDIZAJE	0.01

Tabla 4.32: Configuración seleccionada

En la tabla 33, se muestra la mejor configuración. Para seleccionar ésta se usó la medida de error MAPE, teniendo para la muestra de Test un error MAPE de 272.21%.

3. **Clasificación y regresión:** Este modelo se divide en dos etapas. En la primera etapa, mediante un modelo de clasificación (MLP o árbol de decisión) se clasifica a los clientes, en las clases definidas dado la variable supervisada[19]. Posteriormente, en una segunda etapa, para cada clasificación por separado se predice el valor continuo de cada cliente.

   Los modelos de predicción se calibran con los clientes pertenecientes a las clases definidas. La ventaja de esta metodología, es que si un cliente en primera instancia es clasificado de manera correcta, con mayor certeza se puede asegurar una correcta predicción.

   En los modelos explicados anteriormente se expusieron los métodos para la clasificación. Una vez clasificados los clientes, se debe predecir dentro de cada una de las clasificaciones el valor continuo de cada cliente, con el fin de conocer certeramente su comportamiento futuro.

   a. **MLP regresión:** Tomando cada uno de los clientes que pertenecen a cada una de las clases, se hizo un análisis de sensibilidad, de tal manera de obtener las mejores configuraciones de cada red.

Parámetro	MLP BAJA	MLP MANTIENE	MLP SUBE	MLP MULTIPLICA
EPOCAS	1000	500	500	1000
NEURONAS	14	15	13	13
T. APRENDIZAJE	0.001	0.001	0.01	0.01

Tabla 4.33: Mejores combinaciones

En la Tabla 4.33, se muestran las mejores configuraciones obtenidas para cada una de las predicciones dentro de cada una de las clases. Para seleccionar la mejor configuración, se usó la medida de error MAPE, la cual arrojó los siguientes resultados:

	MLP BAJA	MLP MANTIENE	MLP SUBE	MLP MULTIPLICA
$MAPE_{Monto}$	261.30%	12.95%	13.10%	37.37%

Tabla 4.34: MAPE de cada configuración seleccionada

[19] Fuga, Baja, Mantiene, Sube y Multiplica.

Se puede observar, que se obtuvo un MAPE de monto bajo, para todas las clasificaciones, excepto para los clientes que bajan su comportamiento. Esto manifiesta el hecho de que con una correcta clasificación a priori se obtendrían resultados muy satisfactorios.

b. Árbol Regresión: Tras una clasificación, se debe predecir el valor continuo del comportamiento futuro de cada cliente. Para ello se debe calibrar los árboles. Para los árboles de regresión se determinó que con 30 casos como mínimo en los nodos terminales se obtienen los mejores resultados. Definido este parámetro se debe seleccionar el largo de cada uno de los árboles.

	BAJA	MANTIENE	SUBE	MULTIPLICA
Largo	6	6	5	7
$MAPE_{Monto}$	**290.90%**	**11.47%**	**23.28%**	**18.49%**

Tabla 4.35: Largo de cada configuración seleccionada

En la tabla 4.35, se muestra el largo seleccionado para cada uno de los árboles. Se usó el error MAPE para seleccionar este largo, el cual se expone además en la tabla.

Se observa que el árbol de clasificación y el MLP de clasificación tienen muy malos resultados. No ocurre lo mismo con los árboles de regresión y los MLP's de regresión para cada una de las clases. Pero, como uno es dependiente del otro, se construirá cada secuencia con modelos de la misma naturaleza, es decir, un árbol de clasificación, seguido de árboles de regresión y otro modelos formado por un MLP de clasificación seguido de MLP's de regresión para cada una de las clases. A la secuencia de árboles se le denominará Árbol_clasi_reg y a la secuencia de MLP's se le llamará MLP_clasi_reg.

4.9.3 Modelos ingenuos

De manera de analizar la relevancia de construir modelos sofisticados, se incorporarán al análisis modelos que no requieren esfuerzo. A continuación, se mostrará como se construirán estos modelos:

1. **Promedio General:** Este modelo asigna a cada cliente la variación promedio de todos los clientes que se están analizando. Por lo tanto, este modelo asignará para cada cliente una variación de 5,3% correspondiente al promedio general.
2. **Promedio en el cluster según tamaño (promedio cluster):** Este modelo asigna a cada cliente como variación futura, la variación promedio dada por la clasificación según tamaño al que pertenece. La asignación seguida por este modelo será la siguiente:
   - Cliente de monto bajo: variación porcentual del monto= 15,9%
   - Cliente de monto: variación porcentual del monto=-11.8%
   - Cliente de monto alto: variación porcentual del monto=-9.3%
3. **Mantiene:** Este modelo asume que los clientes se van a comportar siempre igual, es decir en el siguiente período van a gastar lo mismo que gastan ahora. Por lo tanto, la variación de cada cliente será 0%.

4.10 Análisis comparativo entre los modelos

Una vez desarrollados los modelos y determinado la mejor configuración de cada uno de estos, se expondrá los resultados obtenidos. En una primera instancia se utilizará el error MAPE del monto, como métrica para medir los resultados, de manera de mostrar la ineficiencia de esta medida. Posteriormente, se utilizarán otras medidas basadas en la variación porcentual del monto, que sí permiten tener una evaluación asertiva de los resultados de los modelos.

4.10.1 Error MAPE dado predicción del monto

En el gráfico 4.21, se muestra el porcentaje de clientes, que explica cada porcentaje del error total de en cada modelo.

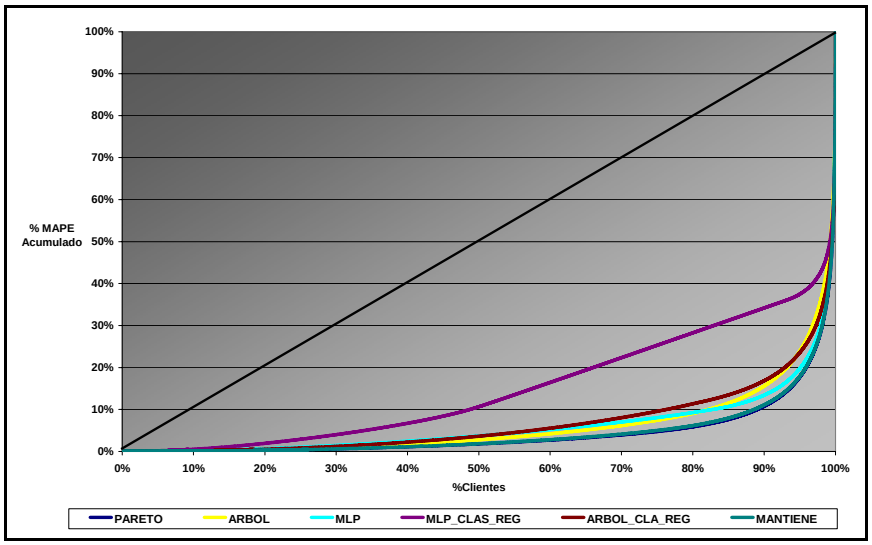


Grafico 4.21: Distribución del error de cada modelo.

Se puede ver en el gráfico 27, que todos los modelos, excepto el modelo MLP de clasificación y regresión con el 80% de los clientes explican menos del 20% del error, esto quiere decir que un 20% de los clientes explican el 80% del error.

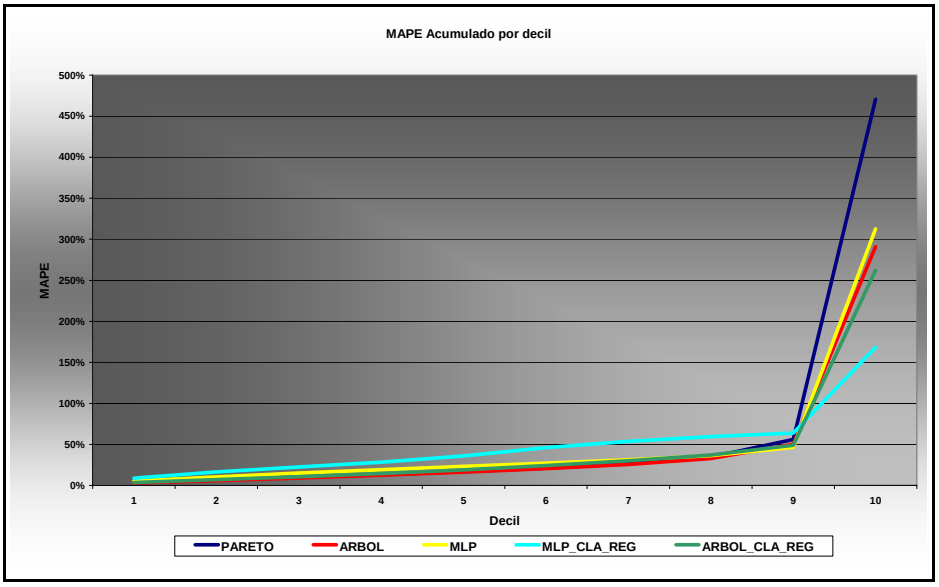


Grafico 4.22: Error promedio MAPE acumulado por decil

Se observa en el gráfico 4.22, que el modelo que presenta error MAPE más alto, de los modelos expuestos, es el modelo Pareto. Cada decil está ordenado de menor a mayor, dado el error MAPE[20] de cada cliente. Además se infiere del gráfico, que todos los modelos, salvo el modelo MLP_Clasi_Reg presentan errores menores al 50% hasta el 8º percentil.

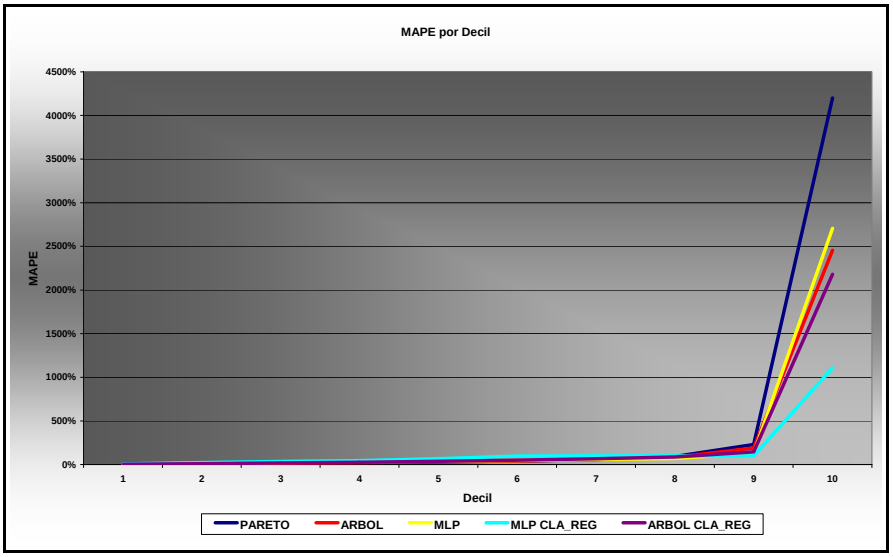


Gráfico 4.23: Error promedio por decil

[20] Ver página 27 para definición de errores.

En el gráfico 4.23 se muestra el error MAPE promedio en cada percentil. Este gráfico no acumula los errores de los deciles, a diferencia del gráfico 28. Se puede apreciar que en el último percentil los modelos de data mining superan ampliamente al modelo Pareto.

CLASIFICACIÓN	MODELO	$MAPE_{MONTO}$	$RANKING_{MAPE}$
PROBABILÍSTICO	PARETO	470.70%	6
DATA MINING	ÁRBOL	290.93%	4
	MLP	312.62%	3
	MLP_CLASI_REG	**168.32%**	1
	ÁRBOL_CLASI_REG	262.21%	2
INGENUO	MANTIENE	456.78%	5
	PROMEDIO	481.85%	8
	PROMEDIO CLUSTER	473.81%	7

Tabla 4.36: Ranking de modelos según MAPE

En la tabla 4.36, se puede observar que los modelos de data mining en términos de error porcentual para el monto, tienen mejores desempeños que los modelos ingenuos y el modelo probabilístico abordado. El modelo con menor error, es el modelo MLP con clasificación y regresión, el cual tiene un error MAPE inferior a la mitad del que tienen las demás clasificaciones.

La medida de error MAPE es útil debido a que incorpora el tamaño monetario de cada cliente. Esto hace posible tener una medida que permite evaluar el error general. No obstante, esta métrica no puede incorporar a los clientes que han sido fugados, debido a que se indefine. Además un modelo, que en general prediga que los clientes van a empeorar con respecto a su estado inicial, siempre tendrá un error MAPE menor o igual a 100%. En vista de que esta medida no incorpora a los clientes fugados, estos se evaluarán con otras medidas de error.

De los clientes seleccionados para el análisis un 5.54% corresponde a clientes fugados, por lo tanto su monto futuro real será 0. Por lo tanto se analizará el monto promedio estimado para estos clientes por cada modelo.

CLIENTES FUGADOS			
CLASIFICACIÓN	**Modelo**	**MAE_{MONTO}**	**RANKING**
PROBABILÍSTICO	**PARETO**	767.679	**5**
DATA MINING	**ÁRBOL**	430.622	**3**
	MLP	552.572	**4**
	MLP_CLASI_REG	181.967	**1**
	ARBOL_CLASI_REG	271.777	**2**
INGENUO	**MANTIENE**	864.387	**6**
	PROMEDIO	2.267.523	**8**
	PROMEDO_CLUSTER	1.032.352	**7**

Tabla 4.37: Error MAE de modelos para clientes fugados

En la tabla 4.37, se observa que, los modelos de data mining tienen mejores resultados que los otros modelos, en especial los modelos Clasi_Reg, son los mejores, esto es, debido a que estos modelos en la etapa de clasificación, clasifican correctamente a clientes que realmente han sido fugados, cosa que los otros modelos no logran hacer.

CLIENTES FUGADOS		
CLASIFICACIÓN	**Modelo**	**Variación Promedio**
PROBABILÍSTICO	**PARETO**	-7.48%
DATA MINING	**ÁRBOL**	-39.62%
	MLP	-35.35%
	MLP_CLASI_REG	-87.29%
	ARBOL_CLASI_REG	-68.47%
INGENUO	**MANTIENE**	0.00%
	PROMEDIO	723.54%
	PROMEDO_CLUSTER	143.15%

Tabla 4.38: Predicción de la variación para fugados dado cada modelo

Un cliente, en esta investigación ha sido considerado fugado cuando su variación, de un período con respecto a otro, es igual a -100%, bajo ese punto de vista, se puede observar en la tabla 4.38, que los modelos que incorporan clasificación y regresión tienen muy buenos resultados, y además se infiere que los modelos de data mining son los que más se aproximan al resultado real (-100%). Cabe destacar al modelo MLP_Clisi_Reg ,que con una predicción promedio de un -87.29% tiene una precisión bastante cercana a la meta.

CLIENTES FUGADOS										
	PARETO		ÁRBOL		MLP		MLP_CLASI_REG		ARBOL_CLASI_REG	
Percentil	Promedio	Std	Promedio	Std	Promedio	Std	Promedio	Std	Promedio	Std
1	-90.84%	3.99%	-96.31%	2.97%	-53.46%	4.04%	-100.00%	0.00%	-100.00%	0.00%
2	-74.29%	5.57%	-89.69%	0.00%	-47.64%	0.70%	-100.00%	0.00%	-100.00%	0.00%
3	-44.59%	10.85%	-85.34%	4.18%	-45.36%	0.58%	-99.80%	0.09%	-100.00%	0.00%
4	-16.18%	7.37%	-73.81%	2.38%	-43.30%	0.68%	-99.70%	0.01%	-100.00%	0.00%
5	-6.81%	1.89%	-66.69%	4.50%	-40.81%	0.73%	-99.67%	0.01%	-100.00%	0.00%
6	-0.90%	1.17%	-48.50%	4.71%	-38.31%	0.72%	-99.65%	0.01%	-89.62%	6.42%
7	5.83%	2.32%	-32.71%	9.78%	-35.63%	1.03%	-99.63%	0.01%	-80.28%	2.82%
8	17.65%	2.55%	-6.86%	4.91%	-32.01%	1.19%	-99.60%	0.01%	-70.39%	3.69%
9	37.41%	7.23%	18.53%	8.05%	-25.28%	3.21%	-99.57%	0.01%	-32.05%	41.52%
10	89.94%	40.75%	76.14%	74.31%	5.31%	45.08%	17.54%	227.25%	77.21%	87.87%

Tabla 4.39: Variación porcentual del monto por decil de clientes fugados

En la tabla 4.39, se puede observar el desempeño de cada modelo en cada decil de clientes según la estimación de cada modelo. Se observa que el modelo MLP_Clasi_Reg, mencionado anteriormente como el de mejor desempeño tiene una precisión casi exacta para alrededor del 90% de los clientes fugados.

El error MAPE, permite evaluar conjuntamente a clientes de montos relativos distintos, debido a esto, es interesante usar esta medida. En los modelos evaluados, se observó que el modelo MLP_CLASI_REG tiene un error MAPE más bajo que los otros modelos abordados. Además, en el caso de los clientes fugados también tiene mejor desempeño.

El modelo MLP_CLASI_REG, estima que la mayoría de los clientes, va a comprar menos e incluso serán fugados en el año 2006. Debido a los resultados obtenidos de todos los modelos, un modelo que estime que todos los clientes serán fugados, será el mejor modelo utilizando como el MAPE como medida de comparación. Por lo tanto, el error MAPE, no es un buen indicador para evaluar a los modelos, a pesar de haber sido el usado en otras investigaciones. Es por ello, que se usarán otras medidas de desempeño en el capítulo siguiente.

4.10.2 Errores de precisión

Debido al mal desempeño del MAPE como medida para evaluar a los modelos, se utilizarán otras medidas de errores (MAE, MSE, NMSE y R^2). Estas métricas permitirán evaluar el nivel de ajuste con respecto a la variable supervisada[21] de cada modelo. Al evaluar estos errores en la variación porcentual, y no en el monto, se tiene una medida de error que permite hacer comparaciones entre los clientes incluyendo a los clientes fugados:

Modelo	MAE	MSE	NMSE	R^2
PARETO	62.90%	180.86%	84.39%	0.16
ARBOL	**58.11%**	**165.05%**	**77.01%**	0.23
MLP	59.97%	172.24%	80.37%	0.20
MLP_CL_REG	81.67%	273.06%	127.40%	-0.27
ARBOL_CL_REG	61.91%	186.91%	87.21%	0.13
MANTIENE	100.00%	214.60%	100.13%	0.00
PROMEDIO	65.29%	214.32%	100.00%	0.00
PROMEDIO CLUSTER	65.99%	212.58%	99.19%	0.01

Tabla 4.40: Errores de la predicción

El error MAE permite comparar en términos lineales las predicciones de los modelos, a diferencia del error MSE, que lo hace en términos cuadráticos. Es por ello, que el error MSE castiga más cuando el diferencial entre la predicción y el valor real es alto. Ambos errores, permiten incorporar a los clientes fugados a diferencia del error MAPE, y además, al ser calculados sobre la variación porcentual es posible evaluar a todos los clientes con las mismas medidas de error sin necesidad de una normalización. Se observa en la tabla 4.40, que el Árbol, es el modelos de mayor desempeño y los modelos de data mining, a excepción del modelos "ingenuo" MLP_clasi_reg, superan a los demás modelos.

[21] Variación porcentual del monto.

MAE_{M1}/MAE_{M2}	PARETO	ARBOL	MLP	MLP CLASI REG	ARBOL_ CLASI REG	MANTIENE	PROM	PROM CLUSTER
PARETO	1.00	1.08	1.05	0.77	1.02	0.63	0.96	0.95
ARBOL	0.92	1.00	0.97	0.71	0.94	0.58	0.89	0.88
MLP	0.95	1.03	1.00	0.73	0.97	0.60	0.92	0.91
MLP_CLA_REG	1.30	1.41	1.36	1.00	1.32	0.82	1.25	1.24
ARBOL_CLA_REG	0.98	1.07	1.03	0.76	1.00	0.62	0.95	0.94
MANTIENE	1.59	1.72	1.67	1.22	1.62	1.00	1.53	1.52
PROMEDIO	1.04	1.12	1.09	0.80	1.05	0.65	1.00	0.99
PROM CLUSTER	1.05	1.14	1.10	0.81	1.07	0.66	1.01	1.00

Tabla 4.41: Comparación entre modelos según error MAE

En la tabla 4.41, se muestra en color azul a los modelos de la columna que superan a los modelos de la columna dado el error MAE.

MSE_{M1}/MSE_{M2}	PARETO	ARBOL	MLP	MLP CLASI REG	ARBOL_ CLASI REG	MANTIENE	PROM	PROM CLUSTER
PARETO	1.00	1.10	1.05	0.66	0.97	0.84	0.84	0.85
ARBOL	0.91	1.00	0.96	0.60	0.88	0.77	0.77	0.78
MLP	0.95	1.04	1.00	0.63	0.92	0.80	0.80	0.81
MLP_CLA_REG	1.51	1.65	1.59	1.00	1.46	1.27	1.27	1.28
ARBOL_CLA_REG	1.03	1.13	1.09	0.68	1.00	0.87	0.87	0.88
MANTIENE	1.19	1.30	1.25	0.79	1.15	1.00	1.00	1.01
PROMEDIO	1.19	1.30	1.24	0.78	1.15	1.00	1.00	1.01
PROM CLUSTER	1.18	1.29	1.23	0.78	1.14	0.99	0.99	1.00

Tabla 4.42: Comparación entre modelos según error MSE

En la tabla 4.42, se destaca en color azul, cuando un modelo de la fila supera a un modelo de la columna por el error cuadrático (MSE). Se puede observar que los modelos Árbol y MLP son los que tienen los mejores desempeños.

En términos lineales y cuadráticos, el Árbol seguido del MLP, son los modelos con mejores resultados en términos lineales y cuadráticos. Además se infiere que todos los modelos, a excepción del modelo MLP_Clasi_Reg, superan al modelo Pareto /NBD.

4.11 Interpretación de los resultados

4.12.1 Clasificaciones

A continuación, se expondrá el desempeño de cada modelo, dentro de cada grupo definido. Este desempeño se analizará de manera lineal (MAE), permitiendo evaluar la precisión de cada estimación dentro de cada grupo.

MAE	TAMAÑO CLIENTE		
MODELOS	PEQUEÑO	MEDIANO	GRANDE
PARETO	81.06%	38.30%	27.66%
ARBOL	**75.91%**	**33.57%**	**24.53%**
MLP	76.06%	38.41%	28.16%
MLP_CLASI_REG	106.32%	46.73%	37.49%
ARBOL_CLASI_REG	80.01%	37.03%	27.61%
MANTIENE	82.03%	39.67%	28.61%
PROMEDIO	83.14%	41.26%	30.26%
PROMEDIO CLUSTER	86.21%	38.16%	27.74%

Tabla 4.43: Error MAE según tamaño de clientes

Según el tamaño, los modelos tienen mejor precisión en clientes medianos y grandes, teniendo un error inferior a la mitad del que tienen en los clientes pequeños.

MAE	Apóstol en algún período	
MODELOS	NO	SI
PARETO	86.11%	44.96%
ARBOL	83.45%	**38.51%**
MLP	**80.28%**	44.26%
MLP_CLASI_REG	115.63%	55.41%
ARBOL_CLASI_REG	87.94%	41.77%
MANTIENE	86.03%	46.92%
PROMEDIO	87.02%	48.47%
PROMEDIO CLUSTER	89.63%	47.70%

Tabla 4.44: Error MAE según clientes Apóstol

En la clasificación según Apóstol, el desempeño de los modelos supera en el doble, al desempeño que tienen estos mismos, en los clientes no Apóstoles. Se puede concluir que los modelos tienen mejor precisión en los clientes Apóstoles.

MAE	TRANSICIÓN				
MODELOS	FUGADO	BAJA	MANTIENE	SUBE	MULTIPLICA
PARETO	92.52%	69.83%	15.42%	40.69%	258.85%
ARBOL	60.38%	59.19%	**23.86%**	37.42%	**241.07%**
MLP	66.97%	64.14%	17.71%	37.46%	263.17%
MLP_CLASI_REG	**12.71%**	**37.97%**	44.51%	99.16%	353.78%
ARBOL_CLASI_REG	31.53%	54.97%	38.14%	43.49%	250.28%
MANTIENE	100.00%	67.12%	14.50%	36.11%	294.45%
PROMEDIO	105.25%	72.38%	19.10%	**30.86%**	289.20%
PROMEDIO CLUSTER	112.67%	73.59%	19.07%	33.09%	281.21%

Tabla 4.45: Error MAE según transición.

Según la transición, que efectúan los clientes de un año con respecto al otro, se tiene mayor precisión en los clientes que mantendrán su comportamiento. Cabe destacar a los modelos Clasi_Reg, el buen desempeño que tienen en los clientes que serán fugados, pero no olvidar que estos modelos son propensos a clasificar a todos los clientes como fugados, en especial el MLP_CLASI_REG sobre el Arbol_CLASI_REG.

4.12.2 Análisis del ranking generado por los modelos

Mediante este análisis, será posible evaluar cada propuesta de ranking de los clientes, dado por su monto, generada por cada uno de los modelos. Conocido el ranking de los clientes dado por su monto futuro (año 2006), se analizará la precisión de cada modelo en términos del ranking generado, dado por el monto de la predicción. Para generar una medida de error que permita evaluar entre los rankings generados por cada modelo se determinó la siguiente relación:

$$Error\_Ranking = \frac{\Pr omedio(\left| POSICIÓN_{REAL_{cliente}} - POSICIÓN_{MODELO_{cliente}} \right|)}{ERROR_{MÁXIMO_{total}}}$$

El error máximo es el peor resultado que se podrá haber obtenido de un modelo.

MODELO	ERROR CLASIFICACIÖN
PARETO	26.03%
ARBOL	21.88%
MLP	25.93%
MLP CLASI REG	26.96%
ARBOL CLASI REG	22.54%
MANTIENE	26.48%

Tabla 4.46: Error de clasificación por ranking generado por modelos

Se puede observar en la tabla 4.46, que los árboles de decisión tienen mejor precisión al generar el ranking de los clientes, dado por su monto predicho, en comparación con el ranking real.

4.12.3 Análisis de reglas de negocio

Para encontrar segmentos de clientes con mejores resultados en los modelos, se definen clases determinadas por las 15 variables usadas en los modelos. Esto, permite clasificar a un nuevo cliente, conociendo cual será el nivel de precisión que se tendrá sobre él. Se determinan 3 clases proporcionales para cada variable, de ésta manera se tiene una clasificación ALTA, MEDIA y BAJA en cada una de las variables. La cantidad de segmentos, utilizando las 15 variables, es de 3 elevado a la quinceava potencia (3^N do de N es la cantidad de variables).

Debido a la gran cantidad de segmentos, con las 15 variables, se utilizarán menos variables, de manera de hacer abordable el problema. Con 4 variables se tienen 81 grupos, desminuyendo gradualmente la cantidad de segmentos. Por lo tanto, se usarán las 4 mejores variables (Antigüedad de compra, Delta de productos, Delta frecuencia y Monto trimestral), definidas por los métodos de selección de variables, para hacer este análisis. En el anexo 6.16, se puede mostrar las tablas con los resultados, utilizando 3 y 4 variables con los modelos de data mining y Pareto.

Regla	Antigüedad de compras	Delta Productos	Delta Frecuencia	Monto trmestral	% Clientes	REAL	PARETO	ÁRBOL	MLP	MLP Clasi Reg	ARBOL Clasi Reg
Regla 1	ALTO	BAJO	BAJO	MEDIO	3.90%	-49.70%	-10.32%	**-48.42%**	-16.02%	-75.62%	-64.86%
Regla 2	MEDIO	BAJO	BAJO	MEDIO	3.97%	-44.31%	-8.96%	**-45.47%**	-13.72%	-75.39%	-66.37%
Regla 3	BAJO	ALTO	ALTO	BAJO	10.21%	76.98%	44.46%	**74.64%**	41.09%	-0.02%	84.16%
Regla 4	BAJO	ALTO	ALTO	MEDIO	5.20%	67.07%	43.79%	**60.10%**	52.44%	43.74%	84.88%
Regla 5	MEDIO	ALTO	ALTO	BAJO	1.77%	51.71%	6.26%	**50.25%**	1.71%	-84.22%	69.47%
Regla 6	ALTO	BAJO	BAJO	ALTO	4.63%	-36.82%	-9.69%	**-32.17%**	-8.40%	-41.79%	-43.67%

Tabla 4.47: Estimación de los modelos según algunas reglas

El árbol de decisión, genera pronósticos en promedio bastante parecidos a la variación real. Esto implica, que en promedio el árbol de decisión, tiene muy buen desempeño en la tabla. Un modelo ingenuo, tal como el mantiene, hubiese tenido una variación de 0%. Por ejemplo, la regla 3 indica una variación real promedio de los clientes que satisfacen

esa regla es de un 76.98%. El árbol estima una variación del 74.64%. Esta regla esta formada por un 10% de los clientes. Por lo tanto, se puede concluir que con una probabilidad de un 10% es posible tener un error de tan sólo un 2.34% en un cliente.

A continuación se mostrará el desempeño general a nivel de segmentos:

MODELO	ERROR REGLAS
PARETO	17.08%
ARBOL	9.20%
MLP	17.27%
MLP CLASI REG	61.49%
ÁRBOL CLASI REG	19.89%
MANTIENE	19.92%
PROMEDIO	20.89%
PROMEDIO TAMAÑO	19.80%

Tabla 4.48: Error promedio de precisión dado por las reglas

El modelo árbol, es el que tiene mejor desempeño a nivel de segmento, teniendo una precisión casi del doble mejor al que lo precede. Además en los análisis, del capítulo pasado, también resultó ser este modelo el mejor. Por lo tanto, para estimar el lifetime value, será el árbol el modelo seleccionado.

4.13 Estimación de Lifetime Value

El modelo para estimar el lifetime value, será el árbol de decisión, debido a la calidad de los resultados obtenidos.

4.13.1 Definición de parámetros

La variación porcentual del monto, permite incorporar la tasa de crecimiento/decrecimiento de un cliente a lo largo de los años. Además se debe usar una tasa de descuento, que permita descontar los flujos generados por cada cliente a lo largo de los años, esta tasa se definirá de manera empírica como un 5%, cercana al crecimiento del país en los últimos años. Es decir, se le exige como mínimo a un cliente que crezca igual que el crecimiento económico del país. De la misma manera se considerará un plazo de 5 años como tiempo de vida de un cliente de la empresa para evaluar el valor de ellos.

Se define gi como la tasa de crecimiento/decrecimiento para cada cliente i, estimada por el modelo seleccionado (árbol).

4.13.2 Determinación de LTV

Una vez definido los parámetros para la tasa de crecimiento/decrecimiento, se determinará la metodología para estimar el lifetime value de cada cliente.

$$LTV_i = \sum_{t=1}^{T} \frac{Monto_{i_0}(1+g_i)}{(1+r)^t}$$

Esta construcción permite mantener las tasas de error, debido a que se mantiene la dispersión de los resultados. Esta dispersión es posible mantenerla ya que el numerador es constante para cada cliente, dividiendo a todos por una factor común, la tasa de descuento.

	Promedio	Desviación estándar	Mediana
LTV	**$ 9,071,960**	**$ 16,670,868**	**$ 4,395,276**

Tabla 4.50: Estadísticos del lifetime value general.

En la tabla 4.50 se especifican los estadísticos de lifetime value a nivel general. A continuación se estimará el lifetime value para cada uno de los grupos definidos.

Tamaño	LTV	Clientes
BAJO	2,749,180	13,567
MEDIO	11,828,831	6,236
ALTO	35,092,184	2,636

Tabla 4.51: Lifetime value según tamaño

Se puede observar en la tabla 4.51, que los clientes que en el año 2005 son de tamaño alto son los que tienen un mayor valor. Ocurre lo mismo, con los clientes que son apóstoles. En promedio, un cliente apóstol tiene una valorización de $ 14 millones, como se muestra en la tabla 4.52.

Apóstol	LTV	Clientes
SI	14,247,870	12,652
NO	2,380,879	9,787

Tabla 4.52: Lifetime value según condición de apóstol

En el anexo 6.16, se exponen los resultados de las estimaciones del lifetime value para cada regla.

Mediante estos valores estimados, la empresa podrá saber cual es la inversión que podrá destinar a cada uno de los clientes.

4.14 Conclusiones

- Existe un alto nivel transaccional en caso de estudio lo que ha permitido tener un gran volumen de transacciones de los clientes. Esto, junto al empleo de algoritmos de limpieza de datos, permitió construir una base consolidada de los datos que se utilizaron para el análisis.
- Los modelos de data mining permiten tener mayor grado de libertad en la definición de la función objetivo que se quiere calibrar a diferencia del modelo Pareto /NBD.
- Como variable supervisada para los modelos de data mining se usó la variación porcentual del monto, ya que esta variable permite definir una pendiente de crecimiento/decrecimiento al proyectar los flujos en el futuro.
- Los modelos de data mining tienen mejor desempeño que las otras técnicas utilizadas.
- En general, los modelos tienen un mejor desempeño en los clientes que tienen un alto monto acumulado. Además, en los clientes definidos como apóstoles los modelos son más exactos que en los otros clientes.
- El modelo árbol de decisión es el seleccionado para estimar el lifetime value debido a que tiene un mejor desempeño sobre los otros desarrollados con un error MAE de 58.11%.
- Mediante la discretización de las variables utilizadas por los modelos fue posible definir reglas de negocio para identificar patrones, con el fin de definir un indicador de la precisión al incorporar un nuevo cliente al análisis.
- A nivel de segmentos el modelo árbol de decisión tiene un error de precisión (MAE) de un 9.2%, superando ampliamente a los demás.
- Mediante las estimaciones del lifetime value expuestas, la empresa podrá saber cual es la inversión que podrá hacer tanto a nivel de clientes como a nivel de segmentos. No obstante, también conocerá el riesgo en el que incurre, determinado por el error. Cabe destacar que a nivel de segmento es posible tener una mayor diversificación del dinero invertido por la empresa.

4.15 Trabajos futuros

- Estimar un factor de descuento α de manera de suavizar la proyección de los flujos que permita permutar la tasa de crecimiento g a lo largo de los periodos.

$$LTV_i = \sum_{t=1}^{T} \frac{Monto_{i_0}(1+g_i)^t \alpha^t}{(1+r)^t}$$

- Realizar una investigación de mercado para determinar el tiempo de vida de un cliente de un supermercado mayorista.
- Generar un modelo dinámico que actualice la tasa de crecimiento a medida que se tenga nueva información.

5. Bibliografía

[1] Dirección de Marketing 2000 Kotler P. Edición del Milenio, pag. 8.

[2] Morgan, R. and Hunt, S. (1994), The Commitment-Trust Theory of Relationship Marketing, Journal of Marketing, July, pp.20-38.

[3] Pfeifer P., Haskins M., Conroy R. (2004), Customer Lifetime Value, Customer Profitability, and the Treatment of Acquisition Spending.

[4] Marín, P. 2005. Estimación de lifetime value basado en comportamiento transaccional. Memoria de ingeniería civil industrial.

[5] Baek H. 2006. Estimación de customer lifetime value a nivel de clientes usando variables transaccionales y socio-demográficas. Memoria de ingeniería civil industrial.

[6] Fader, Peter S. Hardie, Bruce G., Ka Lok Lee: Counting Your Customers" the Easy Way: An Alternative to the Pareto/NBD Model. Marketing Science; Spring2005, Vol. 24 Issue 2, p275-284, 10p

[7] Thomas, Jacquelyn (2001), "A Methodology for Linking Customer Acquisition to Customer Retention," Journal of Marketing Research, 38 (2), 262-68.

[8] Blattberg, Robert, Deighton, John Deighton (1996), "Managing Marketing by the Customer Equity Test," *Harvard Business Review*, 75 (4), 136-44.

[9] Gupta S., Hanssens D., Hardie B., Kahn W., Kumar V., Lin N., Ravisanker N., Sriram S. 2006 "Modeling Customer Lifetime Value".

[10] Okell J., "Neural Network versus CHAID". A White paper from smartFOCUS

[11] Montoya R. 2002, Programación matemática para data mining: utilización de Support Vector Machines para selección de atributos. Memoria de Magister en Gestión de Operaciones.

[12] Carvajal M. 2006, Implementación de metodología para la selección de variables. Memoria de ingeniería Industrial.

[13] Mani, D. R., Drew, J., Betz, A., Datta, P. Statistics and data mining techniques for lifetime value modeling. Proceedings of the Fifth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining; 94-103. San Diego. ACM Press: New York, NY, 1999.

[14] H. Hyunseok, J. Jung, S. Suh: An LTV model and customer segmentation based on customer value: a case study on the wireless telecommunication industry. Expert Systems with Applications 26 (2004) 181–188

[15] Ya-Yeuh Shih, Chung-Yuan Liu: A method for customer lifetime value ranking -- Combining the analytic hierarchy process and clustering analysis. Journal of Database Marketing & Customer Strategy Management; Dec2003, Vol. 11 Issue 2, p159-172, 14p

[16] Haenlein, M., Kaplan, Andreas M., Schoder, Detlef: Valuing the Real Option of Abandoning Unprofitable Customers When Calculating Customer Lifetime Value. Journal of Marketing; Jul2006, Vol. 70 Issue 3, p5-20,

[17] H. Bauer, M. Hammerschmidt, M. Braehler. The customer lifetime value concept and its contribution to corporate valuation. Yearbook of Marketing and Consumer Research, Vol. 1 2003.

[18] R. Venkatesan, V. Kumar. A Customer Lifetime Value Framework for Customer Selection and Resource Allocation Strategy. Journal of Marketing. 2004.

[19] Michael J.A. Berry, Gordon S. Linoff. 2004. Data mining techniques: for marketing, sales, and customer relationship management

[20] T. Hastie, R Tibshirani, J. Friedman 2001: The elements of statistical learning: data mining, inference and prediction.

[21] D. Hand, H. Mannila, P. Smyth: Principles' of data mining.

[22] J. Han, M. Kamber.: Data mining: concepts and techniques

[23] Raab, David: Using Lifetime Value. DM Review; Aug2006, Vol. 16 Issue 8, p10

6. ANEXO

6.1 Distancia secuencial (Levenshtein).

Levenshtein propuso un algoritmo que calcula la distancia entre dos palabras. Este algoritmo toma una palabra y calcula la menor cantidad de variaciones que necesita una palabra para convertirse en la otra. El número de variaciones requeridas se denomina distancia de Levenshtein.

Este algoritmo forma una matriz, situando en la posición de las filas valores desde el 0 hasta el largo de la palabra, en la primera columna, se hace de igual manera, asignando valores desde el 0 hasta el largo de la segunda palabra. Luego inmediatamente anterior a los valores colocados se coloca la palabra 1 en la fila anterior, y en la columna de la misma manera. A modo de ejemplo se comparará la palabra AHA con AGUA:

		A	H	A
	0	1	2	3
A	1	0	1	1
G	2	1	1	2
U	3	2	2	2
A	4	2	3	2

Para rellenar esta matriz, la metodología es la siguiente:

1) Se compara la letra situada en la columna con las letras situadas en las filas.
2) Si la letra de la columna coincide con la de la fila, se obtiene un costo 0. En caso de no coincidencia, se tiene un costo igual a 1.
3) Para asignar el valor a cada celda se le debe sumar el costo al valor mínimo entre el valor que se encuentra antecedente diagonalmente, antecede en la fila, y antecede en la columna.
4) La distancia de Levinstein se obtiene al llenar toda la matriz. Esta es el valor ubicado en la celda en la posición (n.m), en el caso del ejemplo, la distancia es 2.

Para llevar a cabo esta metodología se tomó un maestro con todas las palabras bien escritas, donde, la comparación de las palabras no se realizará entre las disponibles, sino, que con unas que ya se encuentran en buen estado. Esto se hizo para poder agrupar de correcta manera casos como SANTIAGO y STGO.

Se notó que el cálculo de la distancia de Levinstein no es conmutativo, es decir, no es lo mismo comparar palabra1 con palabra2, que palabra2 con palabra1, es por esto, que se hizo los dos cálculos y se tomó la asociación que otorga la mínima distancia.

6.2 Descriptivos de variables familias

Variable	Mínimo	Máximo	Media	Desviación	Skewness	Kurtosis
cant_prod_ABARROTES	0.00	1.00	0.28	0.15	1.05	2.66
SKU_ABARROTES	0.00	1.00	0.29	0.17	1.12	2.44
en_bol_ABARROTES	0.00	1.00	0.77	0.25	-1.43	1.49
Monto_ABARROTES	0.00	1.00	0.31	0.18	0.93	1.51
cant_prod_ALIMENTOS_PERECIBLES	0.00	1.00	0.32	0.18	0.54	0.88
SKU_ALIMENTOS_PERECIBLES	0.00	1.00	0.31	0.19	0.69	0.84
en_bol_ALIMENTOS_PERECIBLES	0.00	1.00	0.80	0.27	-1.66	1.94
Monto_ALIMENTOS_PERECIBLES	0.00	1.00	0.30	0.19	0.75	0.96
cant_prod_BAZAR_Y_PAQUETERIA	0.00	0.33	0.00	0.00	45.58	3,104.48
SKU_BAZAR_Y_PAQUETERIA	0.00	0.32	0.00	0.00	70.24	7,144.06
en_bol_BAZAR_Y_PAQUETERIA	0.00	1.00	0.01	0.03	13.69	275.08
Monto_BAZAR_Y_PAQUETERIA	0.00	0.36	0.00	0.01	38.03	1,988.06
cant_prod_CONFITES	0.00	1.00	0.08	0.10	3.25	16.56
SKU_CONFITES	0.00	1.00	0.07	0.10	3.63	20.39
en_bol_CONFITES	0.00	1.00	0.38	0.30	0.48	-0.79
Monto_CONFITES	0.00	1.00	0.05	0.09	4.58	31.13
cant_prod_CUIDADO_PERSONAL	0.00	1.00	0.04	0.06	7.42	86.94
SKU_CUIDADO_PERSONAL	0.00	1.00	0.02	0.05	13.71	238.56
en_bol_CUIDADO_PERSONAL	0.00	1.00	0.33	0.28	0.65	-0.42
Monto_CUIDADO_PERSONAL	0.00	1.00	0.04	0.07	6.73	65.39
cant_prod_HOGAR	0.00	1.00	0.13	0.12	3.14	16.27
SKU_HOGAR	0.00	1.00	0.15	0.14	2.75	11.18
en_bol_HOGAR	0.00	1.00	0.61	0.31	-0.51	-0.80
Monto_HOGAR	0.00	1.00	0.16	0.14	2.39	9.34
cant_prod_LIQUIDOS	0.00	1.00	0.15	0.14	2.19	6.95
SKU_LIQUIDOS	0.00	1.00	0.18	0.16	1.88	4.97
en_bol_LIQUIDOS	0.00	1.00	0.57	0.31	-0.40	-0.91
Monto_LIQUIDOS	0.00	1.00	0.13	0.16	2.45	7.33
cant_prod_MASCOTAS	0.00	1.00	0.00	0.02	30.14	1,184.13
SKU_MASCOTAS	0.00	1.00	0.00	0.01	51.30	3,023.62
en_bol_MASCOTAS	0.00	1.00	0.03	0.10	5.84	42.54
Monto_MASCOTAS	0.00	1.00	0.00	0.03	19.08	471.20

Descriptivos variables familias

6.3 Descriptivos de variables Tier de precios

Variable	Mínimo	Máximo	Media	Desviación	Skewness	Kurtosis
%_prod_tier_alto	0	1	0.20	0.11	2.41	13.07
%_Tran_tier_alto	0	1	0.81	0.23	-1.66	2.62
%_SKU_tier_alto	0	1	0.08	0.09	5.38	42.38
%_Monto_tier_alto	0	1	0.25	0.14	1.78	6.04
%_prod_tier_medio	0	1	0.26	0.10	1.27	7.46
%_Tran_tier_medio	0	1	0.84	0.21	-1.91	3.78
%_SKU_tier_medio	0	1	0.21	0.12	1.94	8.58
%_Monto_tier_medio	0	1	0.26	0.12	1.31	5.62
%_prod_tier_bajo	0	1	0.53	0.14	-0.47	2.18
%_Tran_tier_bajo	0	1	0.93	0.15	-3.72	16.37
%_SKU_tier_bajo	0	1	0.70	0.15	-1.70	5.36
%_Monto_tier_bajo	0	1	0.49	0.16	-0.25	1.26

Descriptivos variables tier de precios

6.4 Correlaciones entre variables tier

	Tier_1	Tier_2	Tier_3	Tier_4	Tier_5	Tier_6	Tier_7	Tier_8	Tier_9	Tier_10	Tier_11	Tier_12
Tier_1	1.00	0.02	-0.72	0.40	-0.17	-0.43	0.72	0.11	-0.48	-0.63	0.80	-0.09
Tier_2	0.02	1.00	-0.71	0.02	0.41	-0.22	0.07	0.82	-0.68	-0.58	-0.02	0.83
Tier_3	-0.72	-0.71	1.00	-0.30	-0.17	0.45	-0.56	-0.64	0.81	0.84	-0.55	-0.51
Tier_4	0.40	0.02	-0.30	1.00	0.55	0.23	0.23	0.06	-0.17	-0.28	0.36	-0.04
Tier_5	-0.17	0.41	-0.17	0.55	1.00	0.40	-0.18	0.33	-0.17	-0.13	-0.21	0.43
Tier_6	-0.43	-0.22	0.45	0.23	0.40	1.00	-0.41	-0.25	0.41	0.50	-0.46	-0.14
Tier_7	0.72	0.07	-0.56	0.23	-0.18	-0.41	1.00	0.10	-0.62	-0.62	0.76	-0.04
Tier_8	0.11	0.82	-0.64	0.06	0.33	-0.25	0.10	1.00	-0.85	-0.60	0.06	0.77
Tier_9	-0.48	-0.68	0.81	-0.17	-0.17	0.41	-0.62	-0.85	1.00	0.81	-0.45	-0.59
Tier_10	-0.63	-0.58	0.84	-0.28	-0.13	0.50	-0.62	-0.60	0.81	1.00	-0.71	-0.54
Tier_11	0.80	-0.02	-0.55	0.36	-0.21	-0.46	0.76	0.06	-0.45	-0.71	1.00	-0.21
Tier_12	-0.09	0.83	-0.51	-0.04	0.43	-0.14	-0.04	0.77	-0.59	-0.54	-0.21	1.00

Correlaciones de las variables Tier de precios

Tier_1	Prod tier alto	Tier_4	Presencia tier alto	Tier_7	Unid tier alto	Tier_10	Monto tier bajo
Tier_2	Prod tier medio	Tier_5	Prsencia tier medio	Tier_8	Unid tier medio	Tier_11	Monto tier alto
Tier_3	Prod tier bajo	Tier_6	Presencia tier bajo	Tier_9	Unid tier bajo	Tier_12	Monto tier medio

Nombre de Variables tier

6.5 Correlación entre variables delta

	D_mes_1	D_mes_2	D_mes_3	D_mes_4	D_trim_1	D_trim_2	D_trim_3	D_trim_4
D_mes_1	1.00	0.13	0.32	0.13	0.40	0.12	0.16	0.06
D_mes_2	0.13	1.00	0.75	0.87	0.03	0.03	0.03	0.02
D_mes_3	0.32	0.75	1.00	0.83	0.11	0.08	0.14	0.07
D_mes_4	0.13	0.87	0.83	1.00	0.03	0.05	0.05	0.05
D_trim_1	0.40	0.03	0.11	0.03	1.00	0.36	0.49	0.22
D_trim_2	0.12	0.03	0.08	0.05	0.36	1.00	0.69	0.64
D_trim_3	0.16	0.03	0.14	0.05	0.49	0.69	1.00	0.69
D_trim_4	0.06	0.02	0.07	0.05	0.22	0.64	0.69	1.00

Correlaciones de las variables delta

D_mes_1	Delta F mes	**D_trim_1**	Delta F trim
D_mes_2	delta monto mes	**D_trim_2**	delta monto trim
D_mes_3	delta prod mes	**D_trim_3**	delta prod trim
D_mes_4	delta SKU mes	**D_trim_4**	delta SKU trim

Nombre de Variables delta

6.6 Correlación entre variables familia

	F 1.1	F 1.2	F 1.3	F 1.4	F 2.1	F 2.2	F 2.3	F 2.4	F 3.1	F 3.2	F 3.3	F 3.4	F 4.1	F 4.2	F 4.3	F 4.4	F 5.1	F 5.2	F 5.3	F 5.4	F 6.1	F 6.2	F 6.3	F 6.4	F 7.1	F 7.2	F 7.3	F 7.4	F 8.1	F 8.2	F 8.3	F 8.4
F 1.1	1.00	0.89	0.53	0.90	0.36	0.32	0.10	0.34	0.00	0.01	0.00	0.02	0.22	0.22	0.16	0.23	0.06	0.07	0.02	0.08	0.04	0.00	0.10	0.01	0.42	0.35	0.29	0.42	0.03	0.03	0.00	0.04
F 1.2	0.89	1.00	0.48	0.87	0.32	0.35	0.10	0.33	0.02	0.00	0.01	0.01	0.14	0.17	0.11	0.15	0.05	0.07	0.01	0.08	0.00	0.06	0.05	0.07	0.38	0.38	0.28	0.39	0.02	0.02	0.00	0.03
F 1.3	0.53	0.48	1.00	0.50	0.06	0.07	0.40	0.06	0.00	0.01	0.09	0.00	0.07	0.08	0.26	0.09	0.10	0.13	0.36	0.13	0.05	0.07	0.51	0.03	0.31	0.25	0.22	0.32	0.10	0.07	0.03	0.10
F 1.4	0.90	0.87	0.50	1.00	0.27	0.27	0.12	0.35	0.01	0.02	0.02	0.03	0.19	0.18	0.17	0.20	0.11	0.10	0.07	0.14	0.04	0.08	0.01	0.11	0.37	0.33	0.31	0.42	0.04	0.03	0.02	0.05
F 2.1	0.36	0.32	0.06	0.27	1.00	0.91	0.60	0.90	0.05	0.02	0.06	0.04	0.22	0.16	0.11	0.19	0.28	0.20	0.18	0.20	0.46	0.41	0.24	0.40	0.32	0.33	0.15	0.28	0.05	0.04	0.05	0.04
F 2.2	0.32	0.35	0.07	0.27	0.91	1.00	0.55	0.89	0.04	0.02	0.04	0.03	0.18	0.15	0.08	0.17	0.24	0.17	0.15	0.16	0.43	0.43	0.23	0.41	0.31	0.39	0.14	0.29	0.03	0.02	0.01	0.01
F 2.3	0.10	0.10	0.40	0.12	0.60	0.55	1.00	0.60	0.03	0.02	0.06	0.01	0.11	0.06	0.24	0.10	0.21	0.19	0.21	0.18	0.30	0.29	0.26	0.23	0.27	0.25	0.25	0.26	0.10	0.07	0.03	0.08
F 2.4	0.34	0.33	0.06	0.35	0.90	0.89	0.60	1.00	0.04	0.02	0.03	0.03	0.14	0.10	0.04	0.15	0.29	0.21	0.17	0.23	0.43	0.42	0.22	0.42	0.28	0.32	0.11	0.29	0.05	0.03	0.04	0.05
F 3.1	0.00	0.02	0.00	0.01	0.05	0.04	0.03	0.04	1.00	0.69	0.67	0.82	0.05	0.03	0.05	0.03	0.06	0.04	0.04	0.05	0.05	0.04	0.03	0.04	0.03	0.02	0.01	0.02	0.00	0.00	0.02	0.00
F 3.2	0.01	0.00	0.01	0.02	0.02	0.02	0.02	0.02	0.69	1.00	0.34	0.84	0.03	0.03	0.02	0.02	0.03	0.03	0.01	0.03	0.02	0.02	0.00	0.02	0.01	0.02	0.01	0.01	0.00	0.00	0.01	0.00
F 3.3	0.00	0.01	0.09	0.02	0.06	0.04	0.06	0.03	0.67	0.34	1.00	0.54	0.07	0.04	0.16	0.05	0.07	0.04	0.17	0.06	0.05	0.04	0.13	0.05	0.04	0.03	0.09	0.04	0.00	0.00	0.06	0.00
F 3.4	0.02	0.01	0.00	0.03	0.04	0.03	0.01	0.03	0.82	0.84	0.54	1.00	0.06	0.04	0.05	0.05	0.05	0.04	0.04	0.04	0.03	0.02	0.02	0.03	0.02	0.01	0.00	0.01	0.00	0.00	0.01	0.00
F 4.1	0.22	0.14	0.07	0.19	0.22	0.18	0.11	0.14	0.05	0.03	0.07	0.06	1.00	0.88	0.68	0.93	0.01	0.00	0.00	0.02	0.15	0.14	0.10	0.11	0.05	0.04	0.09	0.09	0.01	0.01	0.03	0.00
F 4.2	0.22	0.17	0.08	0.18	0.16	0.15	0.06	0.10	0.03	0.03	0.04	0.04	0.88	1.00	0.62	0.88	0.04	0.02	0.04	0.00	0.19	0.19	0.13	0.16	0.07	0.02	0.11	0.09	0.01	0.00	0.03	0.02
F 4.3	0.16	0.11	0.26	0.17	0.11	0.08	0.24	0.04	0.05	0.02	0.16	0.05	0.68	0.62	1.00	0.62	0.01	0.01	0.34	0.03	0.11	0.11	0.28	0.05	0.00	0.00	0.42	0.02	0.03	0.02	0.11	0.01
F 4.4	0.23	0.15	0.09	0.20	0.19	0.17	0.10	0.15	0.03	0.02	0.05	0.05	0.93	0.88	0.62	1.00	0.02	0.01	0.01	0.02	0.16	0.15	0.10	0.12	0.07	0.05	0.11	0.09	0.01	0.01	0.02	0.00
F 5.1	0.06	0.05	0.10	0.11	0.28	0.24	0.21	0.29	0.06	0.03	0.07	0.05	0.01	0.04	0.01	0.02	1.00	0.84	0.59	0.89	0.33	0.37	0.22	0.34	0.14	0.09	0.08	0.13	0.00	0.01	0.03	0.00
F 5.2	0.07	0.07	0.13	0.10	0.20	0.17	0.19	0.21	0.04	0.03	0.04	0.04	0.00	0.02	0.01	0.01	0.84	1.00	0.38	0.83	0.22	0.25	0.10	0.20	0.11	0.10	0.10	0.11	0.03	0.01	0.03	0.03
F 5.3	0.02	0.01	0.36	0.07	0.18	0.15	0.21	0.17	0.04	0.01	0.17	0.04	0.00	0.04	0.34	0.01	0.59	0.38	1.00	0.52	0.28	0.27	0.63	0.33	0.17	0.10	0.33	0.17	0.04	0.03	0.10	0.04
F 5.4	0.08	0.08	0.13	0.14	0.20	0.16	0.18	0.23	0.05	0.03	0.06	0.04	0.02	0.00	0.03	0.02	0.89	0.83	0.52	1.00	0.22	0.25	0.13	0.23	0.12	0.09	0.07	0.13	0.00	0.00	0.02	0.00
F 6.1	0.04	0.00	0.05	0.04	0.46	0.43	0.30	0.43	0.05	0.02	0.05	0.03	0.15	0.19	0.11	0.16	0.33	0.22	0.28	0.22	1.00	0.92	0.55	0.93	0.23	0.18	0.15	0.23	0.00	0.01	0.01	0.01
F 6.2	0.00	0.06	0.07	0.08	0.41	0.43	0.29	0.42	0.04	0.02	0.04	0.02	0.14	0.19	0.11	0.15	0.37	0.25	0.27	0.25	0.92	1.00	0.52	0.93	0.21	0.20	0.15	0.22	0.01	0.01	0.01	0.00
F 6.3	0.10	0.05	0.51	0.01	0.24	0.23	0.26	0.22	0.03	0.00	0.13	0.02	0.10	0.13	0.28	0.10	0.22	0.10	0.63	0.13	0.55	0.52	1.00	0.59	0.21	0.15	0.35	0.21	0.06	0.04	0.08	0.06
F 6.4	0.01	0.07	0.03	0.11	0.40	0.41	0.23	0.42	0.04	0.02	0.05	0.03	0.11	0.16	0.05	0.12	0.34	0.20	0.33	0.23	0.93	0.93	0.59	1.00	0.21	0.17	0.07	0.23	0.02	0.02	0.00	0.03
F 7.1	0.42	0.38	0.31	0.37	0.32	0.31	0.27	0.28	0.03	0.01	0.04	0.02	0.05	0.07	0.00	0.07	0.14	0.11	0.17	0.12	0.23	0.21	0.21	0.21	1.00	0.89	0.56	0.92	0.03	0.02	0.05	0.03
F 7.2	0.35	0.38	0.25	0.33	0.33	0.39	0.25	0.32	0.02	0.02	0.03	0.01	0.04	0.02	0.00	0.05	0.09	0.10	0.10	0.09	0.18	0.20	0.15	0.17	0.89	1.00	0.53	0.88	0.04	0.03	0.05	0.03
F 7.3	0.29	0.28	0.22	0.31	0.15	0.14	0.25	0.11	0.01	0.01	0.09	0.00	0.09	0.11	0.42	0.11	0.08	0.10	0.33	0.07	0.15	0.15	0.35	0.07	0.56	0.53	1.00	0.51	0.07	0.04	0.05	0.06
F 7.4	0.42	0.39	0.32	0.42	0.28	0.29	0.26	0.29	0.02	0.01	0.04	0.01	0.09	0.09	0.02	0.09	0.13	0.11	0.17	0.13	0.23	0.22	0.21	0.23	0.92	0.88	0.51	1.00	0.03	0.02	0.04	0.03
F 8.1	0.03	0.02	0.10	0.04	0.05	0.03	0.10	0.05	0.00	0.00	0.00	0.00	0.01	0.01	0.03	0.01	0.00	0.03	0.04	0.00	0.00	0.01	0.06	0.02	0.03	0.04	0.07	0.03	1.00	0.86	0.52	0.85
F 8.2	0.03	0.02	0.07	0.03	0.04	0.02	0.07	0.03	0.00	0.00	0.00	0.00	0.01	0.00	0.02	0.01	0.01	0.01	0.03	0.00	0.01	0.01	0.04	0.02	0.02	0.03	0.04	0.02	0.86	1.00	0.30	0.68
F 8.3	0.00	0.00	0.03	0.02	0.05	0.01	0.03	0.04	0.02	0.01	0.06	0.01	0.03	0.03	0.11	0.02	0.03	0.03	0.10	0.02	0.01	0.01	0.08	0.00	0.05	0.05	0.05	0.04	0.52	0.30	1.00	0.58
F 8.4	0.04	0.03	0.10	0.05	0.04	0.01	0.08	0.05	0.00	0.00	0.00	0.00	0.00	0.02	0.01	0.00	0.00	0.03	0.04	0.00	0.01	0.00	0.06	0.03	0.03	0.03	0.06	0.03	0.85	0.68	0.58	1.00

Correlaciones de las variables familias

F 1.1	Prod ABARROTES	**F 5.1**	Prod CUIDADO PERSONAL
F 1.2	Unid ABARROTES	**F 5.2**	Unid CUIDADO PERSONAL
F 1.3	Presencia ABARROTES	**F 5.3**	Presencia CUIDADO PERSONAL
F 1.4	Monto ABARROTES	**F 5.4**	Monto CUIDADO PERSONAL
F 2.1	Prod ALIMENTOS PERECIBLES	**F 6.1**	Prod HOGAR
F 2.2	Unid ALIMENTOS PERECIBLES	**F 6.2**	Unid HOGAR
F 2.3	Presencia ALIMENTOS PERECIBLES	**F 6.3**	Presencia HOGAR
F 2.4	Monto ALIMENTOS PERECIBLES	**F 6.4**	Monto HOGAR
F 3.1	Prod BAZAR Y PAQUETERIA	**F 7.1**	Prod LIQUIDOS
F 3.2	Unid BAZAR Y PAQUETERIA	**F 7.2**	Unid LIQUIDOS
F 3.3	Presencia BAZAR Y PAQUETERIA	**F 7.3**	Presencia LIQUIDOS
F 3.4	Monto BAZAR Y PAQUETERIA	**F 7.4**	Monto LIQUIDOS
F 4.1	Prod CONFITES	**F 8.1**	Prod MASCOTAS
F 4.2	Unid CONFITES	**F 8.2**	Unid MASCOTAS
F 4.3	Presencia CONFITES	**F 8.3**	Presencia MASCOTAS
F 4.4	Monto CONFITES	**F 8.4**	Monto MASCOTAS

Nombre variables familias

6.7 Correlación entre variables estáticas

	E_1	**E_2**	**E_3**	**E_4**	**E_5**	**E_6**	**E_7**	**E_8**
E_1	1.00	-0.70	0.37	1.00	-0.52	0.37	0.37	0.24
E_2	-0.70	1.00	-0.30	-0.70	0.88	-0.30	-0.67	-0.40
E_3	0.37	-0.30	1.00	0.37	-0.23	1.00	0.17	0.11
E_4	1.00	-0.70	0.37	1.00	-0.52	0.37	0.39	0.27
E_5	-0.52	0.88	-0.23	-0.52	1.00	-0.23	-0.80	-0.38
E_6	0.37	-0.30	1.00	0.37	-0.23	1.00	0.18	0.12
E_7	0.37	-0.67	0.17	0.39	-0.80	0.18	1.00	0.66
E_8	0.24	-0.40	0.11	0.27	-0.38	0.12	0.66	1.00

Correlaciones variables estáticas

E_1	Frecuencia mensual	**E_5**	Recency trimestral
E_2	Recency mensual	**E_6**	Monto trimestral
E_3	Promedio monto mes	**E_7**	Días entre primera y última compra
E_4	Frecuencia trimestral	**E_8**	Días desde primera compra

Nombre variables estáticas

6.8 Test chi cuadrado para variables comunas

	Chi-cuadrado			Promedio	
Comuna	χ^2	**g.l.**	**p-valor**	**No Pertenece**	**Pertenece**
SANTIAGO	809.44	2	0.00	2,252,102	1,424,718
PUENTE ALTO	140.90	2	0.00	2,026,870	2,636,616
MAIPU	132.09	2	0.00	2,133,803	1,272,692
MALLOCO	101.14	2	0.00	2,074,918	2,580,066
LA PINTANA	35.84	2	0.00	2,072,275	2,461,931
PUDAHUEL	34.01	2	0.00	2,085,511	2,304,082
LA FLORIDA	32.78	2	0.00	2,081,662	2,456,597
LO PRADO	23.73	2	0.00	2,091,595	2,438,037
RENCA	21.97	2	0.00	2,095,887	2,227,674
LA GRANJA	20.91	2	0.00	2,085,139	2,467,864
QUILICURA	18.54	2	0.00	2,088,157	2,564,166
RECOLETA	12.50	2	0.00	2,088,018	2,817,806
CERRO NAVIA	10.63	2	0.00	2,089,718	2,276,919
PIRQUE	9.34	2	0.01	2,096,230	2,906,073
SAN MIGUEL	8.58	2	0.01	2,101,070	1,840,203
CONCHALI	6.98	2	0.03	2,102,766	1,955,507
PENALOLEN	6.77	2	0.03	2,093,334	2,382,817
TALAGANTE	6.21	2	0.04	2,097,472	2,085,879
COLINA	5.41	2	0.07	2,096,809	2,224,751
LA REINA	5.39	2	0.07	2,097,158	2,301,135
SAN JOAQUIN	4.33	2	0.11	2,096,038	2,163,958
SAN RAMON	3.42	2	0.18	2,098,495	2,059,530
PEDRO AGUIRRE CERDA	3.37	2	0.19	2,089,293	3,887,023
NUNOA	3.22	2	0.20	2,098,880	1,582,842
VITACURA	3.16	2	0.21	2,098,003	557,663
HUECHURABA	3.14	2	0.21	2,092,132	2,862,540
EL BOSQUE	2.64	2	0.27	2,110,711	1,815,815
LA CISTERNA	2.54	2	0.28	2,096,516	2,174,211
CERRILLOS	2.45	2	0.29	2,098,414	1,973,596
MACUL	2.31	2	0.31	2,097,472	2,004,423
INDEPENDENCIA	2.28	2	0.32	2,094,273	2,607,708
LAS CONDES	1.91	2	0.38	2,094,612	5,057,621
SAN BERNARDO	1.84	2	0.40	2,104,505	1,957,664
QUINTA NORMAL	1.78	2	0.41	2,095,384	2,203,050
PROVIDENCIA	1.75	2	0.42	2,098,472	1,600,658
ESTACION CENTRAL	1.61	2	0.45	2,099,978	1,390,300
PENAFLOR	1.58	2	0.45	2,097,504	2,065,331
LAMPA	1.08	2	0.58	2,096,900	2,269,535
BUIN	0.36	2	0.83	2,097,689	1,931,188
SAN JOSE DE MAIPO	0.11	2	0.95	2,097,061	2,283,542
LO ESPEJO	0.08	2	0.96	2,098,574	1,842,081

Resultado chi-cuadrado comunas

6.9 ANOVA de 1 vía para las comunas

	ANOVA		Promedio		
COMUNA	**F**	**p-valor**	**No pertenece**	**Pertenece**	**Cantidad**
MAIPU	32.35	0.00	4.09%	31.69%	949
RECOLETA	26.57	0.00	4.68%	49.41%	288
SANTIAGO	22.68	0.00	7.49%	-4.45%	4196
PENALOLEN	13.45	0.00	4.83%	35.33%	314
LAS CONDES	12.41	0.00	5.15%	117.74%	21
CONCHALI	11.35	0.00	4.61%	22.15%	820
RENCA	7.75	0.01	5.55%	-20.12%	255
LA REINA	6.57	0.01	5.17%	80.26%	25
INDEPENDENCIA	5.52	0.02	5.08%	34.66%	136
MALLOCO	5.07	0.02	5.73%	-4.95%	998
MACUL	4.69	0.03	5.19%	74.38%	21
CERRILLOS	2.28	0.13	5.39%	-10.93%	185
LO PRADO	2.21	0.14	5.44%	-5.90%	375
LA GRANJA	1.56	0.21	5.03%	11.97%	718
HUECHURABA	1.50	0.22	5.16%	19.71%	153
SAN RAMON	1.33	0.25	5.45%	-1.32%	639
PIRQUE	1.00	0.32	5.29%	-20.55%	32
ESTACION CENTRAL	0.94	0.33	5.31%	-10.38%	82
LA PINTANA	0.70	0.40	5.04%	8.37%	1446
SAN BERNARDO	0.60	0.44	5.43%	1.90%	1088
LO ESPEJO	0.59	0.44	5.31%	-5.76%	104
BUIN	0.53	0.46	5.29%	-11.44%	41
PEDRO AGUIRRE CERDA	0.53	0.47	5.30%	-5.36%	101
LAMPA	0.48	0.49	5.22%	18.07%	63
QUINTA NORMAL	0.48	0.49	5.16%	10.18%	417
QUILICURA	0.47	0.49	5.16%	10.03%	435
PUENTE ALTO	0.31	0.58	5.45%	3.76%	2595
PROVIDENCIA	0.26	0.61	5.23%	15.90%	49
LA FLORIDA	0.22	0.64	5.35%	3.05%	941
CERRO NAVIA	0.22	0.64	5.35%	3.03%	919
VITACURA	0.14	0.71	5.26%	-12.88%	9
SAN MIGUEL	0.12	0.73	5.21%	8.09%	317
NUNOA	0.10	0.76	5.27%	-0.36%	65
PUDAHUEL	0.02	0.88	5.29%	4.66%	1219
LA CISTERNA	0.02	0.89	5.24%	6.56%	251
COLINA	0.01	0.91	5.25%	6.89%	101
PENAFLOR	0.01	0.92	5.25%	6.86%	83
TALAGANTE	0.01	0.93	5.25%	6.28%	168
SAN JOAQUIN	0.00	0.95	5.26%	4.81%	445
EL BOSQUE	0.00	0.98	5.26%	5.15%	1014
SAN JOSE DE MAIPO	0.00	0.99	5.25%	5.48%	39

Resultado ANOVA comunas

6.10 Análisis de componentes principales

Pos	Variable	Rating
1	prod tier bajo	13.26
2	Monto ALIMENTOS PERECIBLES	13.23
3	Monto tier bajo	13.21
4	Prod ALIMENTOS PERECIBLES	12.91
5	en bol ALIMENTOS PERECIBLES	12.7
6	SKU tier bajo	12.66
7	SKU ALIMENTOS PERECIBLES	12.3
8	Tran tier bajo	12.23
9	SKU tier alto	11.97
10	en bol HOGAR	11.58
11	Monto HOGAR	11.27
12	Monto tier alto	11.12
13	Monto tier medio	10.93
14	Monto LIQUIDOS	10.83
15	en bol LIQUIDOS	10.74
16	Tran tier medio	10.54
17	Prod LIQUIDOS	10.43
18	prod tier medio	10.36
19	SKU HOGAR	10.32
20	SKU tier medio	10.18
21	en bol ABARROTES	9.84
22	Prod HOGAR	9.73
23	Tran tier alto	9.67
24	prod tier alto	9.66
25	en bol CUIDADO PERSONAL	9.56
26	SKU LIQUIDOS	9.43
27	Monto ABARROTES	8.88
28	Prod ABARROTES	8.47
29	SKU ABARROTES	8.29
30	en bol CONFITES	7.44

	Variable	Rating
31	Prod CUIDADO PERSONAL	6.56
32	RECENCY MENSUAL	6.38
33	SKU CONFITES	6.12
34	Frec mensual	6.1
35	Frec prom trim	6.1
36	Monto CUIDADO PERSONAL	5.73
37	Prod CONFITES	5.63
38	SKU CUIDADO PERSONAL	5.6
39	Monto CONFITES	5.57
40	en bol BAZAR Y PAQUETERIA	5.26
41	R prom trim	5.2
42	Prod MASCOTAS	4.78
43	Monto prom trim	4.74
44	Promedio Monto Mes	4.74
45	Monto MASCOTAS	4.72
46	delta prod mes	4.58
47	Dias 1ª Compra	4.58
48	en bol MASCOTAS	4.42
49	Dias entre 1ª ultima compra	4.4
50	SKU MASCOTAS	4.38
51	Prod BAZAR Y PAQUETERIA	4.35
52	Delta F mes	4.25
53	Monto BAZAR Y PAQUETERIA	4.14
54	Delta F trim	4.1
55	SKU BAZAR Y PAQUETERIA	3.93
56	delta SKU mes	3.69
57	delta monto mes	3.34
58	delta prod trim	3.12
59	delta monto trim	2.63
60	delta SKU trim	2.42

Ranking según PCA

6.11 Ranking de variables

Variable	Rating	Ranking
Dias 1ª Compra	36.60	1
Delta Monto Trim	9.30	2
Delta Prod Trim	6.90	3
Monto Trim	6.00	4
Dias entre 1ª ultima compra	5.50	5
Delta Unid Trim	2.00	6
Delta Monto Mes	1.70	7
Delta F Mes	1.40	8
Recency Trim	1.10	9
Recency Mes	1.00	10
Delta Unid Mes	0.92	11
Frec trim	0.80	12
Presencia Tier Medio	0.68	13
Unid Tier Alto	0.67	14
Unid Tier Bajo	0.63	15
Monto Cuidado Personal	0.63	16
Unid Liquidos	0.62	17
Monto Liquidos	0.62	18
Monto Tier Alto	0.60	19
Delta Prod Mes	0.60	20
Unid Tier Medio	0.57	21
Monto Alimentos Perecibles	0.52	22
Prod Abarrotes	0.50	23
Prod Líquidos	0.49	24
Prod Tier Bajo	0.48	25
Presencia Hogar	0.46	26
Presencia Alimentos Perecibles	0.44	27
Prod Hogar	0.43	28
Presencia Cuidado Personal	0.41	29
Presencia Tier Alto	0.41	30
Presencia Líquidos	0.34	31
Unid Confites	0.32	32
Unid Alimentos Perecibles	0.32	33
Prod Alimentos Perecibles	0.31	34
Monto Abarrotes	0.31	35
Prod Tier Alto	0.29	36
Monto Tier Medio	0.28	37
Prod Tier Medio	0.27	38
Presencia Confites	0.24	39
Prod Cuidado Personal	0.24	40
Monto Confites	0.22	41
Unid Cuidado Personal	0.21	42
Prod Confites	0.21	43
Monto Hogar	0.20	44
Monto Tier Bajo	0.20	45
Presencia Tier Bajo	0.18	46
Unid Abarrotes	0.17	47
Unid Bazar y Banquetería	0.17	48
Cantidad de sucursales	0.15	49
Presencia Abarrotes	0.15	50
Unidades Abarrotes	0.10	51
Prod Mascotas	0.08	52

Ranking Embedded con todas las variables

6.12 Clientes Fugados

	Tipo Cliente		
Tamaño	NO APOSTOL	APOSTOL	Total
PEQUEÑO	60.69%	27.41%	88.10%
MEDIANO	0.48%	9.49%	9.97%
GRANDE	0.00%	1.93%	1.93%
Total	61.17%	38.83%	

Cross tamaño/condición de apóstol

6.13 Árboles de decisión

Largo	MAPE
2	389.40%
3	347.35%
4	335.59%
5	319.40%
6	313.73%
7	299.63%
8	290.93%
9	306.57%
10	312.83%
11	314.96%
12	314.97%
13	314.61%
14	315.10%
15	315.21%
16	314.53%
17	314.51%
18	314.14%
19	314.04%
20	315.05%
21	314.04%

Sensibilización largo Árbol de Regresión.

SUBE	
largo	MAPE
14	25.17%
13	25.08%
12	24.85%
11	24.85%
10	24.71%
9	24.43%
8	24.18%
7	23.86%
6	23.70%
5	**23.28%**
4	23.37%

Sensibilización largo Árbol de Regresión para clase sube.

MULTIPLICA	
Largo	**MAPE**
17	19.48%
16	19.46%
15	19.47%
14	19.50%
13	19.38%
12	19.34%
11	19.28%
10	19.20%
9	19.05%
8	18.98%
7	18.73%
6	**18.49%**
5	18.65%
4	18.73%
3	18.88%

Sensibilización largo Árbol de Regresión para clase Multiplica.

BAJA	
Largo	**MAPE**
11	296.03%
10	295.30%
9	295.80%
8	294.40%
7	**290.90%**
6	291.10%
5	293.20%
4	294.00%
3	295.80%

Sensibilización largo Árbol de Regresión para clase Baja.

MANTIENE	
Largo	**MAPE**
16	11.92%
15	11.93%
14	11.95%
13	11.92%
12	11.91%
11	11.85%
10	11.83%
9	11.73%
8	11.61%
7	11.52%
6	**11.47%**
5	11.74%
4	11.81%

Sensibilización largo Árbol de Regresión para clase Mantiene.

6.14 Errores dado tamaño

MAPE monto			
Modelo	**PEQUEÑO**	**MEDIANO**	**GRANDE**
PARETO	440.44%	443.68%	678.40%
ARBOL	285.50%	271.06%	363.39%
MLP	281.63%	327.02%	426.89%
MLP_CLASI_REG	143.47%	206.64%	197.31%
ARBOL_CLASI_REG	227.61%	247.69%	461.33%
MANTIENE	415.06%	478.37%	605.43%
PROMEDIO	2581.11%	412.44%	228.16%
PROMEDIO CLUSTER	684.45%	549.37%	726.70%

Error MAPE según tamaño de los clientes.

CLIENTES FUGADOS			
MAE monto			
Modelo	**PEQUEÑO**	**MEDIANO**	**GRANDE**
PARETO	450,734	2,235,263	7,658,974
ARBOL	252,222	1,113,219	5,050,809
MLP	322,524	1,668,889	5,290,464
MLP_CLASI_REG	58,944	555,207	3,871,616
ARBOL_CLASI_REG	148,797	649,192	3,937,845
MANTIENE	506,342	2,578,114	8,360,866
PROMEDIO	2,267,523	2,267,523	2,267,523
PROMEDIO CLUSTER	636,017	3,012,030	8,903,300

Error MAE según tamaño de los clientes

	PARETO	**Cluster 2006**		
	MAPE monto	**PEQUEÑO**	**MEDIANO**	**GRANDE**
Cluster 2005	**PEQUEÑO**	480.04%	50.21%	74.25%
	MEDIANO	1455.51%	23.48%	39.64%
	GRANDE	13031.57%	68.96%	20.04%

	ARBOL	**Cluster 2006**		
	MAPE monto	**PEQUEÑO**	**MEDIANO**	**GRANDE**
Cluster 2005	**PEQUEÑO**	309.77%	45.63%	72.13%
	MEDIANO	865.35%	24.30%	33.34%
	GRANDE	6753.49%	62.04%	18.64%

	MLP	Cluster 2006		
	MAPE monto	PEQUEÑO	MEDIANO	GRANDE
Cluster 2005	PEQUEÑO	303.82%	62.59%	83.63%
	MEDIANO	1036.02%	31.36%	53.63%
	GRANDE	7911.12%	40.95%	33.05%

	MLP_ Clasi_Reg	Cluster 2006		
	MAPE monto	PEQUEÑO	MEDIANO	GRANDE
Cluster 2005	PEQUEÑO	150.01%	79.23%	79.96%
	MEDIANO	604.14%	40.15%	59.20%
	GRANDE	3255.82%	38.09%	36.82%

	Árbol_Clasi_Reg	Cluster 2006		
	MAPE monto	PEQUEÑO	MEDIANO	GRANDE
Cluster 2005	PEQUEÑO	246.05%	45.54%	64.41%
	MEDIANO	768.93%	32.61%	28.47%
	GRANDE	8615.01%	69.77%	23.56%

MAPE de transiciones de clientes entre cluster por modelos

	PARETO	APOSTOL	
	MAPE monto	NO	SI
Cluster 2005	PEQUEÑO	368.76%	609.26%
	MEDIANO	164.15%	455.46%
	GRANDE	39728.16%	301.03%

	ARBOL	APOSTOL	
	MAPE monto	NO	SI
Cluster 2005	PEQUEÑO	280.95%	296.20%
	MEDIANO	106.72%	277.98%
	GRANDE	11942.26%	251.49%

	MLP	APOSTOL	
	MAPE	NO	SI
Cluster 2005	PEQUEÑO	221.68%	422.81%
	MEDIANO	116.45%	335.89%
	GRANDE	19019.02%	247.22%

	MLP_Clasi_Reg	APOSTOL	
	MAPE monto	NO	SI
Cluster 2005	PEQUEÑO	132.40%	169.55%
	MEDIANO	91.06%	211.51%
	GRANDE	166.51%	197.61%

	Árbol_Clasi_Reg	APOSTOL	
	MAPE monto	NO	SI
Cluster 2005	PEQUEÑO	225.50%	232.60%
	MEDIANO	96.94%	254.04%
	GRANDE	31474.41%	161.63%

Cruce de MAPE dado tipo y tamaño de clientes

	PARETO	CLASE			
	MAPE	BAJA	MANTIENE	SUBE	MULTIPLICA
Cluster 2005	PEQUEÑO	1216.21%	27.02%	28.44%	60.33%
	MEDIANO	1464.87%	13.30%	28.96%	60.92%
	GRANDE	3489.11%	11.78%	26.16%	60.24%

	ARBOL	CLASE			
	MAPE	BAJA	MANTIENE	SUBE	MULTIPLICA
Cluster 2005	PEQUEÑO	755.47%	42.68%	28.80%	56.21%
	MEDIANO	872.29%	19.66%	23.76%	54.08%
	GRANDE	1836.60%	14.12%	20.71%	54.15%

	MLP	CLASE			
	MAPE	BAJA	MANTIENE	SUBE	MULTIPLICA
Cluster 2005	PEQUEÑO	735.39%	23.33%	47.79%	73.93%
	MEDIANO	1039.33%	20.00%	45.37%	69.46%
	GRANDE	2116.78%	20.59%	43.20%	67.09%

	MLP_C_R	CLASE			
	MAPE	BAJA	MANTIENE	SUBE	MULTIPLICA
Cluster 2005	PEQUEÑO	255.34%	78.59%	85.15%	95.78%
	MEDIANO	603.42%	31.01%	54.22%	78.00%
	GRANDE	891.97%	24.66%	47.66%	76.33%

	Árbol_C_R	CLASE			
	MAPE	BAJA	MANTIENE	SUBE	MULTIPLICA
Cluster 2005	PEQUEÑO	562.19%	65.20%	37.67%	60.58%
	MEDIANO	775.49%	32.33%	24.56%	50.59%
	GRANDE	2329.37%	23.81%	18.97%	48.33%

Cruce de MAPE dado tamaño y clase variación de clientes

	Cantidad Clientes	APOSTOL	
		NO	SI
Cluster 2005	PEQUEÑO	41.30%	17.54%
	MEDIANO	1.17%	27.67%
	GRANDE	0.12%	12.21%

Cruce de MAPE dado tamaño y clase variación de clientes

	Cantidad Clientes	CLASE			
		BAJA	MANTIENE	SUBE	MULTIPLICA
Cluster 2005	PEQUEÑO	20.21%	14.22%	16.42%	7.99%
	MEDIANO	8.43%	10.86%	8.83%	0.72%
	GRANDE	2.34%	6.08%	3.75%	0.15%

Cantidad de clientes dado cruce tamaño vs clase

	Cantidad Clientes	Cluster 2006		
		PEQUEÑO	MEDIANO	GRANDE
Cluster 2005	PEQUEÑO	53.40%	5.16%	0.27%
	MEDIANO	8.44%	18.12%	2.28%
	GRANDE	0.61%	2.71%	9.00%

Cantidad de clientes dado cruce tamaño vs tamaño futuro

6.15 Análisis Cross con error MAE de la variación

PARETO	TAMAÑO		
CLASE	PEQUEÑO	MEDIANO	GRANDE
FUGA	93.19%	86.86%	91.02%
BAJA	77.21%	57.05%	52.24%
MANTIENE	21.58%	10.65%	9.54%
SUBE	42.08%	40.70%	34.56%
MULTIPLICA	268.97%	167.49%	157.13%

PARETO	APOSTOL	
TAMAÑO	NO	SI
PEQUEÑO	87.39%	66.24%
MEDIANO	40.35%	38.21%
GRANDE	61.39%	27.34%

PARETO	APOSTOL	
CLASE	NO	SI
FUGA	98.09%	83.74%
BAJA	80.41%	61.18%
MANTIENE	24.25%	10.94%
SUBE	40.31%	40.96%
MULTIPLICA	285.43%	197.26%

Error MAE para modelo Pareto

ARBOL	TAMAÑO		
CLASE	**PEQUEÑO**	**MEDIANO**	**GRANDE**
FUGA	62.43%	42.55%	58.74%
BAJA	66.63%	45.53%	44.12%
MANTIENE	34.95%	16.22%	11.60%
SUBE	41.84%	33.41%	27.48%
MULTIPLICA	251.07%	150.47%	142.47%

ARBOL	APOSTOL	
TAMAÑO	**NO**	**SI**
PEQUEÑO	84.75%	55.19%
MEDIANO	37.44%	33.41%
GRANDE	53.25%	24.26%

ARBOL	APOSTOL	
CLASE	**NO**	**SI**
FUGA	72.72%	40.93%
BAJA	74.28%	46.83%
MANTIENE	38.36%	16.51%
SUBE	42.84%	33.61%
MULTIPLICA	265.35%	184.84%

Error MAE para modelo Árbol

MLP	TAMAÑO		
CLASE	**PEQUEÑO**	**MEDIANO**	**GRANDE**
FUGA	66.65%	67.99%	76.33%
BAJA	66.42%	61.34%	54.58%
MANTIENE	22.96%	14.47%	11.21%
SUBE	40.38%	34.21%	32.33%
MULTIPLICA	274.44%	159.18%	160.57%

MLP	APOSTOL	
TAMAÑO	**NO**	**SI**
PEQUEÑO	81.42%	63.50%
MEDIANO	38.95%	38.39%
GRANDE	64.03%	27.81%

MLP	APOSTOL	
CLASE	**NO**	**SI**
FUGA	67.65%	65.91%
BAJA	65.30%	63.19%
MANTIENE	23.81%	14.61%
SUBE	40.61%	35.25%
MULTIPLICA	293.88%	192.03%

Error MAE para modelo MLP

MLP_C_R	TAMAÑO		
CLASE	**PEQUEÑO**	**MEDIANO**	**GRANDE**
FUGA	10.87%	20.92%	53.97%
BAJA	42.37%	30.49%	26.93%
MANTIENE	66.81%	27.62%	22.57%
SUBE	121.20%	74.17%	61.42%
MULTIPLICA	369.62%	210.43%	196.59%

MLP_C_R	APOSTOL	
TAMAÑO	**NO**	**SI**
PEQUEÑO	117.27%	80.67%
MEDIANO	58.91%	46.22%
GRANDE	66.38%	37.22%

MLP_C_R	APOSTOL	
CLASE	**NO**	**SI**
FUGA	8.81%	18.85%
BAJA	44.33%	32.77%
MANTIENE	71.75%	30.70%
SUBE	127.16%	79.49%
MULTIPLICA	390.50%	268.73%

Error MAE para modelo MLP_CLASI_REG

ARBOL_C_R	TAMAÑO		
CLASE	**PEQUEÑO**	**MEDIANO**	**GRANDE**
FUGA	32.28%	24.13%	35.55%
BAJA	59.88%	45.89%	45.28%
MANTIENE	54.32%	27.09%	20.07%
SUBE	53.36%	33.18%	24.48%
MULTIPLICA	262.36%	141.30%	128.93%

ARBOL_C_R	APOSTOL	
TAMAÑO	**NO**	**SI**
PEQUEÑO	89.04%	58.84%
MEDIANO	48.89%	36.53%
GRANDE	64.45%	27.26%

ARBOL_C_R	APOSTOL	
CLASE	**NO**	**SI**
FUGA	35.01%	26.04%
BAJA	65.09%	46.68%
MANTIENE	58.44%	27.85%
SUBE	55.95%	34.73%
MULTIPLICA	282.09%	176.60%

Error MAE para modelo Árbol_CLASI_REG

MANTIENE	TAMAÑO		
CLASE	**PEQUEÑO**	**MEDIANO**	**GRANDE**
FUGA	100.00%	100.00%	100.00%
BAJA	68.75%	64.81%	61.46%
MANTIENE	15.03%	14.35%	13.53%
SUBE	40.46%	32.59%	25.39%
MULTIPLICA	308.71%	165.44%	152.64%

MANTIENE	APOSTOL	
TAMAÑO	**NO**	**SI**
PEQUEÑO	87.32%	69.64%
MEDIANO	40.00%	39.65%
GRANDE	63.40%	28.28%

MANTIENE	APOSTOL	
CLASE	**NO**	**SI**
FUGA	100.00%	100.00%
BAJA	67.58%	66.75%
MANTIENE	15.04%	14.23%
SUBE	40.49%	33.04%
MULTIPLICA	327.96%	216.83%

Error MAE para modelo Mantiene

PROMEDIO	TAMAÑO		
CLASE	**PEQUEÑO**	**MEDIANO**	**GRANDE**
FUGA	105.25%	105.25%	105.25%
BAJA	74.00%	70.06%	66.71%
MANTIENE	19.66%	18.92%	18.14%
SUBE	35.20%	27.33%	20.13%
MULTIPLICA	303.46%	160.19%	147.39%

PROMEDIO	APOSTOL	
TAMAÑO	**NO**	**SI**
PEQUEÑO	88.28%	71.08%
MEDIANO	41.73%	41.24%
GRANDE	66.75%	29.91%

PROMEDIO	APOSTOL	
CLASE	**NO**	**SI**
FUGA	105.25%	105.25%
BAJA	72.84%	72.00%
MANTIENE	19.71%	18.80%
SUBE	35.24%	27.78%
MULTIPLICA	322.71%	211.57%

Error MAE para modelo Promedio

PROM_TAMAÑO	TAMAÑO		
CLASE	**PEQUEÑO**	**MEDIANO**	**GRANDE**
FUGA	115.92%	88.21%	90.70%
BAJA	84.67%	53.01%	52.16%
MANTIENE	30.33%	9.68%	9.54%
SUBE	26.65%	44.38%	34.68%
MULTIPLICA	292.80%	177.24%	161.94%

PROM_TAMAÑO	APOSTOL	
TAMAÑO	**NO**	**SI**
PEQUEÑO	91.07%	74.82%
MEDIANO	38.26%	38.15%
GRANDE	59.98%	27.43%

PROM_TAMAÑO	APOSTOL	
CLASE	**NO**	**SI**
FUGA	115.70%	107.90%
BAJA	82.71%	66.13%
MANTIENE	29.46%	13.81%
SUBE	27.53%	36.99%
MULTIPLICA	312.17%	209.49%

Error MAE para modeloPromedio tamaño

6.16 Resultados de las reglas determinadas (segmentos)

6.16.1 Utilizando 3 variables

ANTIG.	Delta_Prod	Delta_F	N	REAL	PARETO	Árbol	MLP	MLP_C_R	ARB._C_R
BAJO	**ALTO**	**ALTO**	**16.99%**	69.05%	42.22%	67.14%	43.58%	14.85%	82.10%
MEDIO	**ALTO**	**ALTO**	**5.28%**	29.57%	-1.35%	29.71%	2.21%	-64.41%	51.98%
ALTO	**ALTO**	**ALTO**	**3.62%**	25.69%	-3.47%	23.94%	-1.76%	-59.21%	40.19%
BAJO	**ALTO**	**BAJO**	**0.63%**	19.56%	14.34%	34.68%	-5.42%	-90.44%	-21.88%
BAJO	**MEDIO**	**ALTO**	**2.35%**	18.27%	12.07%	13.56%	11.09%	-55.12%	15.15%
BAJO	**ALTO**	**MEDIO**	**2.23%**	16.69%	9.67%	25.80%	3.68%	-69.86%	24.23%
ALTO	**ALTO**	**MEDIO**	**1.67%**	15.90%	-7.16%	13.82%	-4.35%	-51.24%	35.90%
ALTO	**MEDIO**	**ALTO**	**1.53%**	13.11%	-8.10%	3.08%	-3.05%	-49.47%	11.70%
BAJO	**MEDIO**	**MEDIO**	**3.26%**	12.16%	10.31%	7.96%	4.41%	-59.35%	7.05%
MEDIO	**ALTO**	**MEDIO**	**1.95%**	11.12%	-5.13%	18.50%	-0.36%	-60.74%	31.18%
ALTO	**ALTO**	**BAJO**	**0.45%**	9.70%	3.87%	20.21%	-9.15%	-75.12%	2.83%
MEDIO	**MEDIO**	**ALTO**	**2.18%**	1.37%	-6.44%	6.24%	-0.34%	-54.92%	13.06%
ALTO	**MEDIO**	**MEDIO**	**9.34%**	0.05%	-8.66%	-2.26%	-6.05%	-40.82%	4.01%
MEDIO	**MEDIO**	**MEDIO**	**8.06%**	-0.97%	-7.86%	-0.51%	-2.18%	-44.51%	4.47%
BAJO	**MEDIO**	**BAJO**	**1.14%**	-2.50%	3.95%	2.02%	-2.87%	-74.39%	-29.47%
ALTO	**BAJO**	**ALTO**	**0.25%**	-4.89%	0.99%	-8.71%	-5.87%	-69.46%	-32.57%
ALTO	**MEDIO**	**BAJO**	**3.01%**	-6.83%	-8.15%	-6.44%	-6.77%	-47.30%	-6.37%
BAJO	**BAJO**	**ALTO**	**0.69%**	-7.48%	17.71%	2.11%	1.57%	-88.17%	-25.37%
MEDIO	**MEDIO**	**BAJO**	**2.46%**	-8.50%	-7.56%	-5.45%	-2.80%	-53.37%	-10.81%
MEDIO	**ALTO**	**BAJO**	**0.52%**	-10.22%	7.03%	19.68%	-9.78%	-82.71%	-15.49%
BAJO	**BAJO**	**MEDIO**	**1.31%**	-10.85%	8.70%	-1.90%	0.85%	-79.18%	-39.74%
MEDIO	**BAJO**	**ALTO**	**0.44%**	-12.91%	1.75%	-8.21%	-1.40%	-74.78%	-32.32%
ALTO	**BAJO**	**MEDIO**	**3.01%**	-14.39%	-8.61%	-19.70%	-5.45%	-44.33%	-28.76%
BAJO	**BAJO**	**BAJO**	**4.75%**	-19.54%	3.81%	-28.34%	-17.84%	-89.28%	-66.63%
MEDIO	**BAJO**	**MEDIO**	**2.51%**	-20.98%	-6.63%	-17.87%	-2.37%	-54.79%	-31.29%
MEDIO	**BAJO**	**BAJO**	**9.93%**	-39.48%	-8.24%	-39.20%	-17.06%	-75.60%	-64.39%
ALTO	**BAJO**	**BAJO**	**10.44%**	-42.16%	-9.76%	-39.83%	-15.47%	-64.29%	-57.56%

Resultado reglas con 3 variables

6.16.2 Estimación de lifetime value en segmentos con 3 variables.

T	Delta_Prod	Delta_F	Clientes	LTV
BAJO	BAJO	BAJO	1,065	$ 1,905,720
BAJO	BAJO	MEDIO	294	$ 2,803,071
BAJO	BAJO	ALTO	155	$ 2,582,115
BAJO	MEDIO	BAJO	255	$ 4,029,346
BAJO	MEDIO	MEDIO	731	$ 5,218,919
BAJO	MEDIO	ALTO	527	$ 5,644,810
BAJO	ALTO	BAJO	141	$ 3,004,372
BAJO	ALTO	MEDIO	500	$ 3,314,839
BAJO	ALTO	ALTO	3,812	$ 5,458,073
MEDIO	BAJO	BAJO	2,228	$ 4,353,329
MEDIO	BAJO	MEDIO	564	$ 9,283,323
MEDIO	BAJO	ALTO	98	$ 4,993,464
MEDIO	MEDIO	BAJO	553	$ 10,540,815
MEDIO	MEDIO	MEDIO	1,809	$ 13,444,723
MEDIO	MEDIO	ALTO	490	$ 11,129,905
MEDIO	ALTO	BAJO	117	$ 4,483,550
MEDIO	ALTO	MEDIO	437	$ 8,689,306
MEDIO	ALTO	ALTO	1,184	$ 8,571,140
ALTO	BAJO	BAJO	2,343	$ 8,267,666
ALTO	BAJO	MEDIO	676	$ 14,548,319
ALTO	BAJO	ALTO	57	$ 7,814,198
ALTO	MEDIO	BAJO	676	$ 16,110,858
ALTO	MEDIO	MEDIO	2,095	$ 20,206,141
ALTO	MEDIO	ALTO	344	$ 13,486,334
ALTO	ALTO	BAJO	102	$ 7,282,351
ALTO	ALTO	MEDIO	374	$ 14,274,473
ALTO	ALTO	ALTO	812	$ 12,948,096

6.16.3 Utilizando 4 variables

ANT	Delta_P	Delta_F	Monto	N	REAL	PARETO	Árbol	MLP	MLP_C_R	ARB._C_R
BAJO	ALTO	ALTO	BAJO	10.21%	76.98%	44.46%	74.64%	41.09%	-0.02%	84.16%
ALTO	MEDIO	MEDIO	ALTO	7.01%	-3.06%	-9.88%	-3.97%	-6.65%	-35.72%	4.00%
BAJO	ALTO	ALTO	MEDIO	5.20%	67.07%	43.79%	60.10%	52.44%	43.74%	84.88%
ALTO	BAJO	BAJO	ALTO	4.63%	-36.82%	-9.69%	-32.17%	-8.40%	-41.79%	-43.67%
MEDIO	MEDIO	MEDIO	ALTO	4.36%	-3.71%	-9.73%	-3.87%	-3.53%	-35.98%	5.04%
MEDIO	BAJO	BAJO	MEDIO	3.97%	-44.31%	-8.96%	-45.47%	-13.72%	-75.39%	-66.37%
ALTO	BAJO	BAJO	MEDIO	3.90%	-49.70%	-10.32%	-48.42%	-16.02%	-75.62%	-64.86%
MEDIO	BAJO	BAJO	BAJO	3.50%	-34.58%	-7.12%	-35.37%	-27.31%	-95.76%	-74.56%
BAJO	BAJO	BAJO	BAJO	3.12%	-10.81%	6.59%	-21.20%	-20.74%	-95.79%	-67.18%
MEDIO	MEDIO	MEDIO	MEDIO	2.73%	0.46%	-8.38%	-0.84%	-0.30%	-47.55%	3.55%
MEDIO	BAJO	BAJO	ALTO	2.46%	-38.67%	-8.68%	-34.52%	-7.87%	-47.32%	-46.78%
MEDIO	ALTO	ALTO	MEDIO	2.25%	24.05%	-3.73%	21.64%	3.59%	-61.40%	46.73%
ALTO	BAJO	MEDIO	ALTO	2.01%	-23.35%	-9.75%	-19.74%	-5.99%	-36.33%	-24.01%
ALTO	BAJO	BAJO	BAJO	1.91%	-39.72%	-8.80%	-40.90%	-31.50%	-95.74%	-76.38%
ALTO	MEDIO	BAJO	ALTO	1.89%	-7.69%	-9.66%	-6.19%	-7.13%	-36.90%	-0.84%
ALTO	MEDIO	MEDIO	MEDIO	1.84%	5.52%	-8.62%	-0.45%	-3.43%	-50.37%	3.38%
MEDIO	ALTO	ALTO	BAJO	1.77%	51.71%	6.26%	50.25%	1.71%	-84.22%	69.47%
BAJO	MEDIO	MEDIO	BAJO	1.60%	30.89%	24.45%	21.87%	5.59%	-66.11%	13.53%
BAJO	ALTO	ALTO	ALTO	1.58%	24.45%	22.59%	41.92%	30.46%	15.85%	59.69%
BAJO	ALTO	MEDIO	BAJO	1.50%	25.60%	15.16%	35.72%	3.88%	-73.62%	24.67%
ALTO	ALTO	ALTO	MEDIO	1.47%	27.59%	-4.66%	20.10%	-0.70%	-61.89%	37.37%
BAJO	BAJO	BAJO	MEDIO	1.39%	-36.57%	-0.80%	-43.78%	-13.39%	-81.00%	-67.28%
BAJO	MEDIO	ALTO	BAJO	1.27%	33.38%	22.51%	27.98%	14.70%	-55.77%	22.83%
MEDIO	ALTO	ALTO	ALTO	1.26%	8.21%	-7.83%	15.14%	0.46%	-41.82%	36.67%
ALTO	ALTO	ALTO	ALTO	1.25%	12.27%	-9.00%	14.82%	-3.02%	-39.06%	34.33%
BAJO	MEDIO	MEDIO	MEDIO	1.15%	-6.90%	-1.28%	-6.06%	3.70%	-57.78%	-2.84%
MEDIO	BAJO	MEDIO	ALTO	1.05%	-26.74%	-9.81%	-20.92%	-2.49%	-39.26%	-23.58%
MEDIO	MEDIO	BAJO	ALTO	1.03%	-10.94%	-9.77%	-8.50%	-4.53%	-36.85%	-1.38%
MEDIO	MEDIO	MEDIO	BAJO	0.97%	7.39%	2.02%	15.58%	-1.37%	-74.39%	4.50%
MEDIO	BAJO	MEDIO	MEDIO	0.96%	-14.13%	-7.49%	-23.15%	-0.26%	-55.72%	-36.33%
MEDIO	MEDIO	ALTO	ALTO	0.92%	-6.52%	-9.48%	0.18%	-2.21%	-37.48%	15.38%
MEDIO	MEDIO	BAJO	MEDIO	0.90%	41.34%	6.23%	43.00%	-1.74%	-82.96%	53.01%
ALTO	ALTO	ALTO	BAJO	0.90%	-12.68%	-8.34%	-9.13%	-0.69%	-57.33%	-9.30%
MEDIO	MEDIO	ALTO	MEDIO	0.83%	4.46%	-7.57%	-0.55%	1.61%	-58.36%	7.47%
BAJO	MEDIO	ALTO	MEDIO	0.82%	1.05%	1.25%	-2.20%	8.63%	-57.12%	6.29%
ALTO	ALTO	MEDIO	ALTO	0.81%	11.84%	-9.40%	8.11%	-5.32%	-37.68%	34.04%
ALTO	BAJO	MEDIO	MEDIO	0.81%	-18.74%	-8.22%	-24.24%	-4.00%	-54.29%	-37.75%
ALTO	MEDIO	BAJO	MEDIO	0.79%	9.74%	-7.53%	10.84%	0.73%	-57.77%	35.08%
MEDIO	ALTO	MEDIO	MEDIO	0.79%	-9.02%	-7.66%	-9.21%	-4.76%	-59.68%	-12.00%
ALTO	MEDIO	ALTO	ALTO	0.77%	0.76%	-9.69%	0.18%	-4.15%	-37.19%	13.53%
BAJO	BAJO	MEDIO	BAJO	0.76%	-8.06%	18.11%	13.46%	-0.27%	-91.62%	-36.33%
ALTO	ALTO	MEDIO	MEDIO	0.65%	18.07%	-6.17%	15.43%	-3.09%	-61.07%	36.91%
MEDIO	ALTO	MEDIO	BAJO	0.64%	16.64%	1.31%	34.91%	-2.28%	-80.89%	26.94%
BAJO	ALTO	MEDIO	MEDIO	0.62%	0.29%	11.13%	12.29%	-6.08%	-89.00%	-36.91%
BAJO	MEDIO	BAJO	BAJO	0.62%	-0.52%	-0.29%	5.49%	3.92%	-65.57%	21.72%
ALTO	MEDIO	ALTO	MEDIO	0.56%	12.34%	-7.54%	3.70%	-1.86%	-56.66%	12.35%
MEDIO	MEDIO	BAJO	BAJO	0.54%	3.01%	-2.11%	6.35%	-3.03%	-77.99%	-31.06%
MEDIO	ALTO	MEDIO	ALTO	0.52%	6.45%	-9.34%	10.03%	0.35%	-40.61%	30.45%
MEDIO	BAJO	MEDIO	BAJO	0.50%	-22.00%	1.65%	-1.47%	-6.15%	-85.45%	-37.83%
BAJO	MEDIO	MEDIO	ALTO	0.50%	-3.99%	-8.35%	-4.33%	2.26%	-41.28%	9.06%
ALTO	MEDIO	MEDIO	BAJO	0.49%	23.96%	8.60%	15.39%	-7.28%	-77.78%	6.57%
BAJO	BAJO	ALTO	BAJO	0.48%	-1.30%	23.48%	13.22%	0.19%	-91.99%	-21.63%
BAJO	BAJO	MEDIO	MEDIO	0.45%	29.12%	17.63%	46.64%	-6.78%	-93.65%	-28.04%
BAJO	ALTO	BAJO	BAJO	0.45%	-15.23%	-3.59%	-23.58%	3.17%	-65.02%	-45.95%
MEDIO	MEDIO	ALTO	BAJO	0.43%	12.16%	2.20%	32.21%	-0.16%	-85.32%	18.91%
BAJO	MEDIO	BAJO	MEDIO	0.33%	2.56%	-3.86%	-12.86%	1.37%	-63.47%	-24.88%
ALTO	MEDIO	BAJO	BAJO	0.33%	3.35%	-0.60%	-1.24%	-9.51%	-77.42%	-24.65%
MEDIO	ALTO	BAJO	BAJO	0.29%	-19.66%	14.28%	27.70%	-13.03%	-92.74%	-36.34%
BAJO	MEDIO	ALTO	ALTO	0.26%	-0.76%	-4.51%	-6.77%	1.25%	-45.58%	5.85%
BAJO	BAJO	BAJO	ALTO	0.25%	-34.27%	-5.39%	-31.71%	-6.04%	-53.34%	-56.06%
ALTO	ALTO	MEDIO	BAJO	0.21%	24.90%	-1.54%	30.95%	-4.48%	-73.43%	39.98%
ALTO	MEDIO	ALTO	BAJO	0.20%	30.39%	14.00%	33.12%	-13.52%	-89.53%	-1.35%
ALTO	ALTO	BAJO	BAJO	0.20%	62.75%	-3.57%	12.47%	-2.14%	-76.54%	2.86%
ALTO	BAJO	MEDIO	BAJO	0.19%	98.13%	1.62%	-0.19%	-5.99%	-86.51%	-40.76%
MEDIO	ALTO	BAJO	MEDIO	0.19%	5.10%	-1.20%	9.78%	-5.69%	-75.77%	7.20%

BAJO	**MEDIO**	**BAJO**	**ALTO**	0.18%	-21.77%	-6.52%	-6.03%	0.37%	-43.70%	-12.04%
BAJO	**BAJO**	**ALTO**	**MEDIO**	0.17%	-15.32%	5.84%	-22.86%	5.67%	-86.18%	-30.79%
MEDIO	**BAJO**	**ALTO**	**MEDIO**	0.17%	-34.95%	-3.64%	-26.05%	-1.79%	-71.27%	-42.61%
MEDIO	**BAJO**	**ALTO**	**BAJO**	0.17%	17.69%	12.93%	17.16%	-1.44%	-95.02%	-30.46%
ALTO	**ALTO**	**BAJO**	**MEDIO**	0.16%	-10.37%	-2.16%	11.07%	-4.82%	-75.46%	0.14%
BAJO	**ALTO**	**BAJO**	**MEDIO**	0.15%	-3.65%	7.88%	5.52%	-2.00%	-89.16%	-7.70%
BAJO	**ALTO**	**MEDIO**	**ALTO**	0.11%	-7.68%	-9.18%	5.29%	-0.45%	-42.42%	32.67%
BAJO	**BAJO**	**MEDIO**	**ALTO**	0.10%	-12.25%	-6.82%	-20.20%	-1.03%	-49.48%	-37.67%
MEDIO	**BAJO**	**ALTO**	**ALTO**	0.09%	-0.23%	-7.49%	8.23%	-7.23%	-43.63%	16.39%
ALTO	**ALTO**	**BAJO**	**ALTO**	0.09%	-27.33%	-8.48%	-20.99%	-0.61%	-44.69%	-16.58%
ALTO	**BAJO**	**ALTO**	**MEDIO**	0.09%	12.04%	15.33%	8.57%	-9.57%	-96.31%	-37.99%
ALTO	**BAJO**	**ALTO**	**BAJO**	0.09%	-19.09%	-5.10%	-17.91%	-3.59%	-62.71%	-36.54%
ALTO	**BAJO**	**ALTO**	**ALTO**	0.08%	-8.10%	-8.72%	-18.24%	-4.19%	-45.82%	-21.50%
MEDIO	**ALTO**	**BAJO**	**ALTO**	0.04%	-12.52%	-7.75%	7.05%	-5.08%	-41.56%	31.53%
BAJO	**BAJO**	**ALTO**	**ALTO**	0.04%	-46.93%	0.63%	-21.68%	0.27%	-51.36%	-46.34%
BAJO	**ALTO**	**BAJO**	**ALTO**	0.03%	-8.93%	-2.68%	-0.46%	-1.97%	-50.19%	0.21%

Resultado reglas con 4 variables

6.16.4 Estimación de lifetime value en segmentos con 4 variables

T	Delta_Prod	Delta_F	Monto_prom	Clientes	LTV
BAJO	**BAJO**	**BAJO**	**BAJO**	699	**$ 835,409**
BAJO	**BAJO**	**BAJO**	**MEDIO**	311	**$ 2,664,201**
BAJO	**BAJO**	**BAJO**	**ALTO**	55	**$ 11,219,540**
BAJO	**BAJO**	**MEDIO**	**BAJO**	170	**$ 1,242,728**
BAJO	**BAJO**	**MEDIO**	**MEDIO**	101	**$ 3,531,579**
BAJO	**BAJO**	**MEDIO**	**ALTO**	23	**$ 11,136,946**
BAJO	**BAJO**	**ALTO**	**BAJO**	107	**$ 1,156,087**
BAJO	**BAJO**	**ALTO**	**MEDIO**	39	**$ 3,728,894**
BAJO	**BAJO**	**ALTO**	**ALTO**	9	**$ 14,566,627**
BAJO	**MEDIO**	**BAJO**	**BAJO**	140	**$ 1,244,329**
BAJO	**MEDIO**	**BAJO**	**MEDIO**	75	**$ 4,128,402**
BAJO	**MEDIO**	**BAJO**	**ALTO**	40	**$ 13,591,176**
BAJO	**MEDIO**	**MEDIO**	**BAJO**	360	**$ 1,430,674**
BAJO	**MEDIO**	**MEDIO**	**MEDIO**	259	**$ 4,506,917**
BAJO	**MEDIO**	**MEDIO**	**ALTO**	112	**$ 19,041,924**
BAJO	**MEDIO**	**ALTO**	**BAJO**	284	**$ 1,552,841**
BAJO	**MEDIO**	**ALTO**	**MEDIO**	185	**$ 4,402,555**
BAJO	**MEDIO**	**ALTO**	**ALTO**	58	**$ 29,643,711**
BAJO	**ALTO**	**BAJO**	**BAJO**	101	**$ 1,586,726**
BAJO	**ALTO**	**BAJO**	**MEDIO**	33	**$ 4,255,017**
BAJO	**ALTO**	**BAJO**	**ALTO**	7	**$ 17,563,083**
BAJO	**ALTO**	**MEDIO**	**BAJO**	336	**$ 1,575,246**
BAJO	**ALTO**	**MEDIO**	**MEDIO**	140	**$ 5,002,232**
BAJO	**ALTO**	**MEDIO**	**ALTO**	24	**$ 17,826,017**
BAJO	**ALTO**	**ALTO**	**BAJO**	2290	**$ 1,921,777**
BAJO	**ALTO**	**ALTO**	**MEDIO**	1167	**$ 6,443,994**
BAJO	**ALTO**	**ALTO**	**ALTO**	355	**$ 25,028,630**
MEDIO	**BAJO**	**BAJO**	**BAJO**	785	**$ 818,149**
MEDIO	**BAJO**	**BAJO**	**MEDIO**	890	**$ 2,935,110**
MEDIO	**BAJO**	**BAJO**	**ALTO**	553	**$ 11,654,110**
MEDIO	**BAJO**	**MEDIO**	**BAJO**	113	**$ 1,357,302**
MEDIO	**BAJO**	**MEDIO**	**MEDIO**	215	**$ 4,273,575**
MEDIO	**BAJO**	**MEDIO**	**ALTO**	236	**$ 17,642,375**

MEDIO	**BAJO**	**ALTO**	**BAJO**	38	**$ 1,390,221**
MEDIO	**BAJO**	**ALTO**	**MEDIO**	39	**$ 3,559,354**
MEDIO	**BAJO**	**ALTO**	**ALTO**	21	**$ 14,176,962**
MEDIO	**MEDIO**	**BAJO**	**BAJO**	122	**$ 1,630,861**
MEDIO	**MEDIO**	**BAJO**	**MEDIO**	201	**$ 4,693,117**
MEDIO	**MEDIO**	**BAJO**	**ALTO**	230	**$ 20,377,345**
MEDIO	**MEDIO**	**MEDIO**	**BAJO**	217	**$ 1,656,135**
MEDIO	**MEDIO**	**MEDIO**	**MEDIO**	613	**$ 5,615,151**
MEDIO	**MEDIO**	**MEDIO**	**ALTO**	979	**$ 20,960,199**
MEDIO	**MEDIO**	**ALTO**	**BAJO**	97	**$ 1,638,385**
MEDIO	**MEDIO**	**ALTO**	**MEDIO**	187	**$ 5,566,536**
MEDIO	**MEDIO**	**ALTO**	**ALTO**	206	**$ 20,649,455**
MEDIO	**ALTO**	**BAJO**	**BAJO**	66	**$ 1,695,204**
MEDIO	**ALTO**	**BAJO**	**MEDIO**	42	**$ 5,157,609**
MEDIO	**ALTO**	**BAJO**	**ALTO**	9	**$ 21,785,817**
MEDIO	**ALTO**	**MEDIO**	**BAJO**	143	**$ 1,701,555**
MEDIO	**ALTO**	**MEDIO**	**MEDIO**	177	**$ 5,889,840**
MEDIO	**ALTO**	**MEDIO**	**ALTO**	117	**$ 21,464,980**
MEDIO	**ALTO**	**ALTO**	**BAJO**	398	**$ 1,950,178**
MEDIO	**ALTO**	**ALTO**	**MEDIO**	504	**$ 6,100,263**
MEDIO	**ALTO**	**ALTO**	**ALTO**	282	**$ 22,331,651**
ALTO	**BAJO**	**BAJO**	**BAJO**	429	**$ 807,257**
ALTO	**BAJO**	**BAJO**	**MEDIO**	874	**$ 2,988,941**
ALTO	**BAJO**	**BAJO**	**ALTO**	1040	**$ 15,781,243**
ALTO	**BAJO**	**MEDIO**	**BAJO**	43	**$ 1,389,255**
ALTO	**BAJO**	**MEDIO**	**MEDIO**	181	**$ 4,622,930**
ALTO	**BAJO**	**MEDIO**	**ALTO**	452	**$ 19,774,724**
ALTO	**BAJO**	**ALTO**	**BAJO**	20	**$ 1,462,780**
ALTO	**BAJO**	**ALTO**	**MEDIO**	20	**$ 4,227,610**
ALTO	**BAJO**	**ALTO**	**ALTO**	17	**$ 19,505,970**
ALTO	**MEDIO**	**BAJO**	**BAJO**	74	**$ 1,563,891**
ALTO	**MEDIO**	**BAJO**	**MEDIO**	177	**$ 5,059,397**
ALTO	**MEDIO**	**BAJO**	**ALTO**	425	**$ 23,246,350**
ALTO	**MEDIO**	**MEDIO**	**BAJO**	110	**$ 1,649,339**
ALTO	**MEDIO**	**MEDIO**	**MEDIO**	413	**$ 5,949,563**
ALTO	**MEDIO**	**MEDIO**	**ALTO**	1572	**$ 25,250,170**
ALTO	**MEDIO**	**ALTO**	**BAJO**	45	**$ 1,823,736**
ALTO	**MEDIO**	**ALTO**	**MEDIO**	126	**$ 6,000,532**
ALTO	**MEDIO**	**ALTO**	**ALTO**	173	**$ 21,972,044**
ALTO	**ALTO**	**BAJO**	**BAJO**	45	**$ 1,468,484**
ALTO	**ALTO**	**BAJO**	**MEDIO**	36	**$ 5,740,195**
ALTO	**ALTO**	**BAJO**	**ALTO**	21	**$ 22,384,333**
ALTO	**ALTO**	**MEDIO**	**BAJO**	47	**$ 1,937,847**
ALTO	**ALTO**	**MEDIO**	**MEDIO**	145	**$ 6,577,500**
ALTO	**ALTO**	**MEDIO**	**ALTO**	182	**$ 23,592,508**
ALTO	**ALTO**	**ALTO**	**BAJO**	201	**$ 2,052,556**
ALTO	**ALTO**	**ALTO**	**MEDIO**	330	**$ 6,550,266**
ALTO	**ALTO**	**ALTO**	**ALTO**	281	**$ 28,255,169**

Lifetime value para reglas con 4 variables