



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

**ESTIMACIÓN DEL VALOR DE CLIENTES DE UN CLUB DE
FIDELIZACIÓN DE UNA TIENDA DE RETAIL**

MEMORIA PARA OPTAR AL TÍTULO DE INGENIERO CIVIL INDUSTRIAL

MARCEL CRISTIAN GARRIDO ROSAS

PROFESOR GUÍA:
Ricardo Montoya Moreira

MIEMBROS DE LA COMISIÓN:
Manuel Reyes Jara
Luis Aburto Lafourcade

SANTIAGO DE CHILE
ABRIL 2009

**RESUMEN DE LA MEMORIA
PARA OPTAR AL TÍTULO DE
INGENIERO CIVIL INDUSTRIAL
POR: MARCEL GARRIDO ROSAS
FECHA: 21/04/2009
PROF. GUÍA: SR. RICARDO MONTOYA M.**

ESTIMACIÓN DEL VALOR DE CLIENTES DE UN CLUB DE FIDELIZACIÓN DE UNA TIENDA DE RETAIL

Las empresas están en una constante búsqueda por satisfacer a sus clientes, con el fin de generar una lealtad por parte de ellos y tener una mayor participación en su canasta de compra. Para esto, realizan diferentes estrategias de marketing enfocadas en entregarles beneficios. Sin embargo, las estrategias están dirigidas a todos los clientes por igual, y como éstos son muy distintos entre sí, las campañas tienen un impacto diferente sobre ellos, el cual no necesariamente es proporcional a los ingresos que ellos aportan a la empresa, por lo que los clientes más valiosos podrían no estar satisfechos con esto y por lo tanto buscar nuevas alternativas. Entonces la empresa tendría grandes pérdidas al no retener a éstos clientes.

El objetivo general de este trabajo de título es estimar el valor de los clientes que pertenecen al club de fidelización de una tienda de Retail. Para esto, se utiliza el modelo de clases latentes y se agrupa a los sujetos en distintos segmentos; de acuerdo a la sensibilidad de las variables de compra observadas, sobre el monto gastado por el cliente; a su vez, el modelo estima los parámetros asociados a las variables, para cada uno de los grupos encontrados. Luego, con los parámetros estimados se pronostican los montos a gastar en el futuro, a través de modelos econométricos, y se calcula el valor presente de éstos para cada individuo. Finalmente el valor del segmento es el valor promedio de los individuos de éste.

Los resultados muestran que los clientes pueden ser agrupados en 5 segmentos, con una buena calidad de ajuste según el criterio de información bayesiano (BIC), y que los parámetros estimados para cada uno de ellos pronostican los montos gastados, en general, con un error pequeño, el cual es menor al 10% en los segmentos más valiosos encontrados (grupos 3 y 5). Además, se obtiene que estos grupos, que corresponden a gente que posee un presupuesto para gastar y no compran compulsivamente (según la caracterización realizada en base a las sensibilidades de las variables), generan más de un 60% de los ingresos de la empresa, en cambio el grupo menos valioso sólo aporta con un 8% de éstos.

Por lo tanto, el tiempo, esfuerzo y dinero a invertir en cada segmento, debería ser proporcional al valor de cada uno de ellos para la empresa, y de esta forma, intentar satisfacer a los más valiosos y crear relaciones de largo plazo con ellos. Además, dadas las características encontradas de estos segmentos, las estrategias de marketing serán más focalizadas y no a todos por igual, lo que permitirá usar los recursos de manera más eficiente. Los próximos trabajos deberían seguir la misma línea e intentar conocer a cabalidad éstos grupos, por ejemplo identificando los tipos de productos que compran. Además, se deberían estudiar algunas acciones de marketing a aplicar sobre los segmentos de la empresa, y así encontrar las acciones más efectivas.

Agradecimientos

En primer lugar quiero agradecer a mis padres, Cristian Garrido y María Angélica Rosas, por la gran enseñanza, cariño y apoyo que me dan día a día, por la confianza que tienen en mí y por comprender el poco tiempo que les dedicaba debido a este trabajo. También quiero agradecerles por los valores que me han entregado durante siempre, ya que me han permitido llegar hasta donde estoy.

Quiero agradecer inmensamente a mi polola, Magdalena Contreras, quien estuvo conmigo en todo momento, quien supo comprender el momento que estaba viviendo y siempre me dio su apoyo a pesar de mi mal humor cuando las cosas no estaban resultando.

Quiero agradecer a mis amigos, que sin ellos nada de esto sería posible, ya que siempre estuvieron conmigo, en los buenos y malos momentos, y nunca me negaron una mano cuando la necesité. Quienes en muchos instantes me animaban y con quienes pasé numerosos momentos agradables. No me gustaría dejar de mencionar a ninguno, pero los que más estuvieron conmigo en estos últimos momentos universitarios fueron: Alan, Darío, Francisco, Alex, Naty, Cathy, Fernando y Gustavo. Ellos me ayudaron también a disfrutar la vida. También quiero agradecer por lo buenos momentos que pasamos juntos a: Manfred, Didier, Seba, Waren, Waraldo, Yung, Gusy, Becerra, Irra, Kena, Lacourt, Palma, Flex, Figo y Pereira entre otros. Gracias a todos y ojalá nuestra amistad crezca aún más con el tiempo.

Gracias también a Mauricio por su ayuda y tiempo dedicado en el desarrollo de esta memoria, cuando lo necesité siempre estuvo dispuesto a ayudarme.

Al profesor guía quiero agradecer por todos sus consejos, y por la gran cantidad de tiempo y apoyo que me dio. Sobre todo al principio fue indispensable su soporte, ya que tuve varias dudas y él me ayudó a aclararlas, además me dio mucho ánimo y confianza cuando lo necesité, gracias a su forma siempre positiva de decir las cosas. De verdad muchas gracias.

A las personas de la empresa donde realicé esta memoria también agradezco, ya que a pesar del poco tiempo que pasamos juntos, el ambiente de trabajo era muy grato gracias a su simpatía y disposición.

Finalmente quiero dar gracias a Dios, ya que la fe en Él me ha permitido enfrentar todas las dificultades, siempre con la confianza de salir adelante de la mejor manera.

Índice de contenidos

1	Introducción.....	9
2	Descripción del Proyecto y Justificación	11
3	Objetivos.....	12
3.1	Objetivo General.....	12
3.2	Objetivos Específicos.....	12
4	Alcances.....	13
5	Resultados Esperados.....	13
6	Metodología.....	14
6.1	Selección de Datos.....	14
6.2	Procesamiento.....	14
6.3	Transformación.....	14
6.4	Montos Gastados Agregados.....	14
6.5	Segmentación de Clientes con Modelos de Clases Latentes	15
6.6	Caracterización de Segmentos	15
6.7	Cálculo de Customer Lifetime Value (CLV) para Cada Segmento	15
6.8	Análisis de Sensibilidad a Tasa de Retención.....	16
7	Marco Conceptual.....	16
7.1	Programas de Lealtad.....	16
7.2	Programa de Puntos de la Empresa	17
7.3	CRM	18
7.4	Modelo para la Estimación de Parámetros.....	19
7.5	Segmentación con Modelos de Clases Latentes.....	22
7.6	Estimación del Modelo de Clases Latentes.....	24
7.7	Medidas para la Calidad del Ajuste.....	25
7.8	Comparación con Otros Métodos.....	27
7.9	Determinación del Valor del Cliente	28
8	Desarrollo de Metodología.....	30
8.1	Selección de Datos.....	30
8.1.1	Análisis Previo a la Selección.....	30
8.1.2	Selección	32
8.2	Procesamiento.....	33
8.3	Transformación.....	33

8.4	Montos Gastados Agregados.....	35
8.4.1	Estimación Montos Gastados.....	35
8.4.2	Validación de Estimación de Montos Gastados.....	39
8.5	Segmentación de Clientes con Modelos de Clases Latentes	41
8.5.1	Consideraciones del Modelo	42
8.5.2	Modelo Homogéneo, Una Clase Latente.....	44
8.5.3	Selección del Número de Clases.....	45
8.5.4	Estimación de Parámetros y Desarrollo del Modelo	47
8.6	Caracterización de Segmentos	53
8.6.1	Análisis de Sensibilidad de Variables sobre Monto Gastado por Segmento ..	53
8.6.2	Análisis de Variables por Segmento.....	57
8.6.3	Calificación de los Segmentos Según sus Características	61
8.7	Cálculo de Customer Lifetime Value (CLV) para Cada Segmento	63
8.7.1	CLV Real para Periodo de Estimación y Validación	63
8.7.2	CLV Estimado para Periodo de Estimación.....	64
8.7.3	CLV Estimado para Periodo de Validación.....	65
8.7.4	Validación del Modelo Estimado con Clases Latentes.....	66
8.7.5	Pronósticos	69
8.8	Análisis de Sensibilidad de Acuerdo a la Tasa de Retención	79
9	Conclusiones.....	82
9.1	Conclusiones Generales.....	82
9.2	Conclusiones con Respecto a la Segmentación.....	82
9.3	Conclusiones con Respecto al Valor del Cliente	83
9.4	Limitaciones del Modelo.....	85
9.5	Trabajos Futuros.....	85
10	Bibliografía.....	86
11	Anexos	89
11.1	Anexo 1: Cálculo del Tamaño Muestral.....	89
11.2	Anexo 2: Cálculo del Tamaño de Muestra para la Estimación de Modelos	89
11.3	Anexo 3: Caracterización Detallada de Segmentos según Sensibilidad a las Variables.....	90
11.4	Anexo 4: Análisis de Errores en Pronósticos Mes a Mes	95
11.5	Anexo 5: Variación en 1% de la Tasa de Retención.....	100

Índice de figuras y tablas

Figura 1. Ejemplo de una función logL con parámetro θ	20
Figura 2. Ejemplo de función de densidad observada y dos funciones adyacentes	23
Figura 3. Cantidad de clientes por comportamiento de canje.....	31
Figura 4. Proporción de montos gastados por clientes con distintos comportamientos de canje. 31	
Figura 5. Proporción de individuos de la población de cada segmento a priori.	49
Figura 6. Proporción de individuos de la población de cada segmento a posteriori.....	52
Tabla 1. Valores de los parámetros significativos para las distintas regresiones.....	36
Tabla 2. Criterios AIC y BIC para distintos modelos.....	37
Tabla 3. Predicción del monto gastado en el periodo de estimación.....	38
Tabla 4. Error agregado para el periodo de estimación.....	38
Tabla 5. Error cuadrático medio para el periodo de estimación.....	39
Tabla 6. Predicción del monto gastado en el periodo de validación.	39
Tabla 7. Error agregado para el periodo de validación.....	40
Tabla 8. Error cuadrático medio para el periodo de validación.....	40
Tabla 9. MAPE promedio para el periodo de estimación y validación.	41
Tabla 10. Parámetros estimados según modelo de independencia.	44
Tabla 11. Predicción del monto gastado según modelo de independencia.	45
Tabla 12. Comparación entre los distintos números de clase utilizados.....	46
Tabla 13. Parámetros estimados para cada grupo según modelo de clase latente.	48
Tabla 14. Probabilidades de pertenencia a posteriori.....	50
Tabla 15. Asignación de individuos según máxima probabilidad de pertenencia a posteriori. ...	50
Tabla 16. Calidad de asignación a cada clase (8 segmentos).....	51
Tabla 17. Calidad de asignación a cada clase (5 segmentos).....	52
Tabla 18. Sensibilidad de variables observadas en el mes anterior para el grupo 3.....	53
Tabla 19. Sensibilidad de Dummies de estacionalidad para el grupo 3.....	54
Tabla 20. Sensibilidad de variables socio-demográficas para el grupo 3.	55
Tabla 21. Resumen de la clasificación de cada grupo.	63
Tabla 22. CLV real periodo de estimación.	64
Tabla 23. CLV real periodo de validación.....	65
Tabla 24. Resultados de las predicciones de los montos gastados en el periodo de validación, según parámetros estimados en modelos de clase latente.....	67

Tabla 25. Error agregado.....	67
Tabla 26. Error cuadrático medio (MSE).....	68
Tabla 27. MAPE a nivel grupal con un grupo (homogéneo) y 5 grupos (heterogéneo).....	69
Tabla 28. Error de usar media móvil de los puntos acumulados.....	71
Tabla 29. Error de usar media móvil de NC.....	73
Tabla 30. Error agregado del pronóstico a 6 meses.....	74
Tabla 31. Error cuadrático medio (MSE) pronóstico.....	75
Tabla 32. MAPE promedio pronóstico.....	75
Tabla 33. CLV promedio monto pronosticado 6 meses.....	76
Tabla 34. Porcentaje de recursos a gastar considerando CLV de montos gastados.....	77
Tabla 35. Porcentaje de recursos a gastar considerando CLV de montos gastados y avances.....	78
Tabla 36. Variación de CLV promedio al disminuir en 5% la tasa de retención.....	80
Tabla 37. Diferencia porcentual al cambio de un 5% en la tasa de retención.....	80
Tabla 38. Disminución en CLV al disminuir 5% la tasa de retención.....	81
Tabla 39. Sensibilidad de variables observadas en el mes anterior para el grupo 4.....	90
Tabla 40. Sensibilidad de Dummies de estacionalidad para el grupo 4.....	91
Tabla 41. Sensibilidad de variables socio-demográficas para el grupo 4.....	91
Tabla 42. Sensibilidad de variables observadas en el mes anterior para el grupo 5.....	91
Tabla 43. Sensibilidad de Dummies de estacionalidad para el grupo 5.....	92
Tabla 44. Sensibilidad de variables socio-demográficas para el grupo 4.....	92
Tabla 45. Sensibilidad de variables observadas en el mes anterior para el grupo 6.....	93
Tabla 46. Sensibilidad de Dummies de estacionalidad para el grupo 6.....	93
Tabla 47. Sensibilidad de variables socio-económicas para el grupo 6.....	93
Tabla 48. Sensibilidad de variables observadas en el mes anterior para el grupo 7.....	94
Tabla 49. Sensibilidad de Dummies de estacionalidad para el grupo 7.....	94
Tabla 50. Sensibilidad de variables socio-económicas para el grupo 7.....	95
Tabla 51. Error grupal agregado mes 57.....	95
Tabla 52. MAPE y MSE mes 57.....	96
Tabla 53. Error grupal agregado mes 58.....	96
Tabla 54. MAPE y MSE mes 58.....	96
Tabla 55. Error grupal agregado mes 59.....	97
Tabla 56. MAPE y MSE mes 59.....	97
Tabla 57. Error grupal agregado mes 60.....	98
Tabla 58. MAPE y MSE mes 60.....	98
Tabla 59. Error grupal agregado mes 61.....	98

Tabla 60. MAPE y MSE mes 61.....	99
Tabla 61. Error grupal agregado mes 62.....	99
Tabla 62. MAPE y MSE mes 62.....	99
Tabla 63. Variación de CLV promedio al disminuir en 1% la tasa de retención.....	100
Tabla 64. Diferencia porcentual al cambio de un 1% en la tasa de retención.....	100
Tabla 65. Disminución en CLV al disminuir 1% la tasa de retención.....	101

1 Introducción

El Retail es un sector económico compuesto por las empresas especializadas en la comercialización masiva de productos o servicios a grandes cantidades de clientes. Es decir, empresas que cuentan con una gran diversidad tanto de productos como de clientes.

Antes, el funcionamiento de un Retail era medido a través de las ventas por departamento y tienda o ventas por hora, lo cual daba una visión limitada del negocio basada solamente en transacciones individuales y no relacionadas, dejando de lado los patrones de compra de los clientes. Pero, dado el constante aumento en la popularidad de las industrias de Retail, lo que ha provocado una fuerte lucha entre los participantes de ésta y por ende una alta competencia en la industria, se hace necesario buscar estrategias para crear valor al cliente y de esta manera tener una mayor participación en su canasta de compra. Por lo tanto, ahora existe un mayor énfasis en los clientes y en su comportamiento de compra, y el funcionamiento del Retail es medido en base al margen del cliente, al valor del ciclo de vida de éste, al valor de ventas cruzadas y a la tasa de retención de clientes, entre otros, lo cual permite cuantificar este comportamiento en el tiempo y de esta manera, obtener beneficios tales como (Andersen, 2000):

- Aumentar la lealtad del cliente
- Identificar oportunidades de venta cruzada
- Mayor efectividad en la adquisición de clientes valiosos
- Layout y localización consistente con los clientes valiosos
- Mayor efectividad y eficiencia en uso de dinero en publicidad
- Disminuir costos de inventario
- Realizar un pricing más efectivo

Algunas de las estrategias usadas por las empresas para crear valor al cliente son: tarjetas de crédito, que permiten a los clientes comprar artículos que no podrían obtener al contado; clubes de fidelización, que permiten a los clientes obtener beneficios tales como descuentos directos, tanto en la empresa como en sus alianzas, además de la acumulación de puntos por sus compras, con lo cual puede posteriormente canjear productos luego de contar con una cierta cantidad de puntos acumulados; e incluso últimamente se están creando instituciones financieras, que permiten a los clientes solicitar dinero en efectivo y otros servicios.

En esta búsqueda por satisfacer al cliente y crearle valor, las empresas realizan distintas campañas de marketing, ofreciendo nuevos productos y promociones para lograrlo. Sin embargo, al aplicar una estrategia común a todos los clientes, sólo se logra satisfacer, y no en un 100%, a algunos, dada la diversidad de éstos. No obstante, el gasto de recursos es aplicado a todos por igual, por lo que se podría decir que la empresa está mal utilizando sus recursos, ya que no sólo no satisface bien a sus clientes, sino que dedica muchos recursos a gente que no le genera mayores ganancias.

Para evitar esta poca efectividad en el uso de los recursos, las empresas buscan conocer mejor a sus clientes (así las acciones de marketing son más efectivas) e identificar a aquellos que le son realmente importantes y por lo tanto, valiosos. A estos clientes hay que dedicar los mayores esfuerzos de marketing y el gasto de recursos, para así crear relaciones de largo plazo manteniéndolos más satisfechos con la empresa. De esta manera, según Forbes (2007), al crear valor a un cliente se obtienen los siguientes beneficios:

- El cliente tiende a comprar más y con mayor frecuencia.
- Es más fácil convencer al cliente de que pruebe un producto nuevo.
- Aumenta el “boca en boca” lo que disminuye el costo de adquisición.
- Las ganancias futuras son más seguras con clientes satisfechos.

Además, es importante manejar y medir el valor de los clientes, ya que son un activo intangible muy importante para la empresa, y de esta manera, al analizar su valor, se tiene una guía para las estrategias usadas con ellos (Gupta, Lehmann y Stuart 2002). También, en base a este conocimiento, se pueden tomar mejores decisiones en los niveles de inversión, por ejemplo en publicidad y promociones, en los niveles de servicios y en los tiempos de respuesta para los distintos tipos de clientes. Igualmente, al medir el valor del cliente, se puede tener una medida aproximada de cuanto se deja de ganar frente a la pérdida de uno de ellos.

En esta memoria se agrupará a los clientes en distintos segmentos según su comportamiento de compra, con el propósito de tener en cada segmento personas que se comportan de manera similar y así poder identificar la heterogeneidad en sus comportamientos. Esto permitiría conocer lo que buscan los diversos tipos de clientes de acuerdo a las características de cada grupo. Esto a su vez ayudaría a aplicar estrategias de marketing dirigidas a cada segmento, y por lo tanto serían más efectivas.

Además, se calculará el valor presente de las transacciones futuras de los segmentos encontrados, con el objetivo de vislumbrar cuáles son los más valiosos para la empresa y así ésta, probablemente dedique la mayor cantidad de recursos en complacerlos, además de buscar una manera de crear up y cross selling (aumentar la venta y generar ventas cruzadas respectivamente) y perfeccionar sus servicios destinados a mejorar la relación con ellos. Asimismo, para los menos valiosos, se buscan formas bajas en costos de hacerlos gastar más, o simplemente no son considerados en las estrategias de la empresa (entre otras posibilidades).

El enfoque está en identificar a los clientes más valiosos, ya que como dice Blattberg y Deighton (1996), atraer y mantener a estos clientes, es el pilar de una estrategia de marketing exitosa.

Para realizar lo antes descrito se utilizará la información transaccional y demográfica de clientes que pertenecen a un club de lealtad de una tienda por departamento, en el cual los clientes pueden acumular puntos al hacer sus compras, los que podrían posteriormente ser canjeados por productos (Las características del programa serán explicadas en la memoria).

La memoria consta de una descripción del problema y la justificación de éste, para luego especificar los objetivos que se esperan alcanzar con ella, el marco conceptual utilizado y la metodología para realizarlo. Además, se definen los alcances del proyecto y los resultados esperados.

Posteriormente, se desarrollará la metodología, donde se mostrarán los resultados obtenidos con la segmentación, se caracterizarán los distintos segmentos y se calculará el valor de cada uno de ellos. Finalmente se realizan conclusiones al respecto y se harán recomendaciones para trabajos futuros.

2 Descripción del Proyecto y Justificación

La industria de Retail es cada vez más competitiva, y es por eso que las empresas están en constante búsqueda por crear valor a los clientes y de esta manera crear fidelidad. En esta búsqueda por lealtad, las empresas han creado en el último tiempo programas dedicados a este tema, específicamente, “tarjetas de fidelización”, con las cuales sus clientes no sólo se pueden sentir más identificados con la empresa, sino que también, optan a beneficios, tales como descuentos directos, posibilidades de pago en cuotas y también la opción de acumular puntos, que luego pueden ser canjeados por productos.

Al ocupar la tarjeta antes mencionada, se va guardando la información transaccional de los clientes, lo cual, junto con sus datos personales, permite identificar a los usuarios y además observar sus conductas a través del tiempo.

El proyecto consiste en analizar la información que se tiene de los clientes a lo largo del tiempo, y de esta manera conocer su actitud de compra para poder agruparlos en distintos segmentos en base a su patrón de comportamiento, el cual, no es fácilmente perceptible, pero se espera que responda a variables subyacentes no observables. Para esto se utilizarán modelos de clase latente, los cuales asumen la existencia de una variable no observada o latente que posee distintas clases o categorías, que están asociadas a una combinación de variables observadas. Y son precisamente estas clases las que definen los distintos patrones de comportamiento de los clientes.

Utilizando los resultados obtenidos se calculará el valor de cada segmento para identificar los grupos más valiosos para la empresa y así a ellos dedicar los mayores esfuerzos de marketing para satisfacerlos. Para esto se utilizarán modelos econométricos, que en base a las variables observadas (del comportamiento pasado) y a la influencia que éstas tienen sobre los montos a gastar, permitirán estimar estos últimos para los próximos periodos de tiempo y luego serán descontados para calcular su valor presente y así obtener el valor asociado a cada segmento.

De esta manera no sólo se conocerá a los clientes más valiosos y se dedicarán los mayores esfuerzos en satisfacerlos, sino que también, las campañas de marketing serán más personalizadas por tipo de segmento, lo cual las hará más efectivas.

Y como la empresa en estudio pretende ser una de las más grandes de Retail en Latinoamérica, y su tarjeta de fidelización es parte fundamental para ella, crear una estrecha relación con los clientes y conocerlos mejor es de vital importancia. Es por eso que hay que conocer a los más valiosos, que son con los que se quiere crear relaciones de largo plazo y así también los de menor valor para la empresa, para identificar a quienes dedicar muy pocos recursos o simplemente no considerar en la estrategia de marketing.

Si no se conoce el valor de los clientes para la empresa y no se diferencia entre ellos, se trata a todos por igual, y con esto se puede dejar a los clientes más valiosos insatisfechos, lo que provocaría una posible fuga a una empresa de la competencia que lo valore más. Y como en general se dice que un 80% de las utilidades de una empresa, proviene del 20% de los clientes¹, es decir, los más valiosos; la pérdida de algunos de estos clientes impactaría de manera fundamental en las ganancias de la firma. Además cabe mencionar que al tener a los clientes satisfechos, esto puede ser propagado a otros clientes que querrán llegar a ser tan valiosos como aquellos para la empresa. En caso contrario (insatisfechos), esto puede generar una propagación negativa y así perder clientes potenciales que no querrán arriesgarse con la empresa debido a las malas referencias.

3 Objetivos

3.1 Objetivo General

Calcular el valor del cliente de un club de fidelización de una tienda de Retail en base a segmentación, de acuerdo a heterogeneidad y datos transaccionales.

3.2 Objetivos Específicos

- Estudiar y definir variables relevantes que influyen en el comportamiento de compra del cliente.
- Realizar segmentación de clientes.
- Caracterizar distintos tipos de segmentos.
- Calcular el valor de los segmentos para la empresa.

¹ Principio de Pareto o regla 80-20: En general, el 20% de las actividades produce el 80% de los resultados.

- Identificar a los segmentos más valiosos
- Calcular la proporción de recursos a utilizar de acuerdo al valor del segmento.

4 Alcances

El alcance del proyecto consiste en estimar el valor de los clientes, dado su comportamiento de compra y canje, agregando la heterogeneidad de ellos, para luego identificar a los clientes más valiosos para la empresa. Esto a nivel de segmento.

Al pronosticar el comportamiento de los clientes, se necesita el valor de las variables independientes, pero éste también será desconocido, por lo que estas variables serán actualizadas con media móvil o dependerán del pronóstico en sí. No se realizará un modelo para cada variable porque no es el foco de la memoria.

No se incluye análisis de los tipos de productos que la gente compra y canjea, por lo que los segmentos están caracterizados sólo en base a la sensibilidad de las variables observadas sobre el monto gastado por los clientes.

Para calcular el valor del cliente se considera un periodo de mediano plazo (1 año aproximadamente) hacia el futuro, y dado que los clientes probablemente no cerrarán su cuenta y es muy difícil saber si realmente dejan de comprar para siempre, se considera que la tasa de retención de los clientes será 1 (100%). Por lo tanto no habrá un análisis particular de fuga de clientes.

En esta memoria no se incluye el análisis de campañas de marketing, ya que no se cuenta con la información necesaria sobre campañas realizadas anteriormente.

5 Resultados Esperados

Al final del proyecto, lo que se espera obtener es:

- Segmentos de clientes según modelos de clases latentes.
- Características de los segmentos.
- Valor promedio de los segmentos de clientes para la empresa.
- Proporción de inversión dedicada a los segmentos más valiosos.

6 Metodología

Primero se realizan los pasos necesarios para obtener los datos con los cuales se va a trabajar, ya que en general se cuenta con mucha información, de la cual sólo hay que extraer la que es relevante, revisarla y modificarla, y así tener los datos precisos para lograr los objetivos². Luego se realiza la segmentación de clientes para posteriormente caracterizar los segmentos y calcular el valor de su ciclo de vida (Customer Lifetime Value).

6.1 Selección de Datos

Se seleccionará una muestra de clientes de una tienda de Retail, los cuales están activos, es decir, tienen su cuenta habilitada para realizar cualquier compra.

Además se estudiará sólo a un tipo de clientes, los que resultan más atractivos para la empresa según su comportamiento de canje.

Cabe señalar que se decide entre los tipos de comportamiento de canje, debido a que la memoria está enfocada en el programa de puntos de la empresa.

6.2 Procesamiento

En esta etapa se procederá a limpiar la muestra de datos extraños o incoherentes, los cuales se deben a errores de la empresa o a clientes con comportamientos fuera de lo común, por lo que no son representativos (en la sección 8.2 será más explicado este tema).

6.3 Transformación

En este paso, se realizarán las transformaciones necesarias a los datos para obtener las variables que serán ocupadas en la memoria.

6.4 Montos Gastados Agregados

Como un primer paso para analizar a los clientes se procederá a estimar los montos que gastan a nivel agregado, es decir, tomando en cuenta que todos los clientes son iguales. Suponiendo que las variables independientes afectan a los montos.

² Esta etapa es en base a la memoria anterior realizada por M. Durán (2008)

Se estimarán los parámetros que generaron los datos en un periodo de tiempo usando modelos econométricos, y luego se validarán en un periodo posterior.

Cabe mencionar, que hasta este punto, todo será en base a la memoria anterior, realizada por Mauricio Durán (2008), los pasos son incluidos en esta memoria, ya que para entender a cabalidad lo realizado por el memorista anterior, fue necesario realizar de nuevo los pasos anteriores, además la muestra a utilizar en esta memoria será prácticamente la misma que la utilizada anteriormente.

6.5 Segmentación de Clientes con Modelos de Clases Latentes

En esta etapa se buscará agrupar a los clientes según su comportamiento de compra, suponiendo que éste está definido por las clases o categorías de una variable no observable o latente.

El número de clases y por lo tanto de segmentos, será definido según los criterios de ajuste que ocupan estos modelos.

6.6 Caracterización de Segmentos

Al obtener los distintos segmentos se procede a realizar un análisis de cada uno de ellos en base a los parámetros y a las variables que influyen en los montos gastados por los individuos. De esta manera, cada segmento es caracterizado y posteriormente bautizado con un nombre de acuerdo a sus características principales.

6.7 Cálculo de Customer Lifetime Value (CLV) para Cada Segmento

Una vez obtenidos los segmentos y los parámetros que con mayor probabilidad generaron los datos de éstos, se procederá a validar la efectividad de éstos parámetros y luego se estimará el monto gastado a nivel de segmento para un periodo de tiempo futuro.

Estos montos serán descontados según una tasa definida por juicio de expertos, para obtener el valor presente de las transacciones futuras y así calcular el valor del segmento para la empresa. Cabe mencionar que para efectos de determinar los grupos más valiosos, el valor de la tasa de descuento no es de mayor importancia, ya que afecta a todos los grupos por igual, por lo tanto, el que resulte más valioso, lo será con cualquier tasa de descuento.

Finalmente se identificará a los segmentos más valiosos para la empresa de acuerdo a su valor de CLV.

6.8 Análisis de Sensibilidad a Tasa de Retención

Para el cálculo anterior se supuso igual a 100% la tasa de retención, ya que al trabajar con una tarjeta de clientes, es poco probable que un cliente deje de usarla para siempre, por lo que siguen siendo parte de la empresa. Además los clientes observados siempre están realizando transacciones, al menos hasta el mes observado.

El análisis será realizado, para observar que tan sensibles son los segmentos al cambio en la tasa de retención. De esta manera se apreciará si es relevante este indicador para el cálculo del Lifetime Value.

7 Marco Conceptual

7.1 Programas de Lealtad

Debido a la gran competencia que hay en la industria del Retail, las empresas deben crear una relación cada vez más estrecha con sus clientes, y para esto deben identificar las características de ellos, no sólo saber el producto que compra, sino porque lo compra. Es en este punto donde se necesita desarrollar un manejo de la relación con el cliente, y para eso existe el CRM (Customer Relationship Management), que permite identificar las características de los clientes, atraerlos y retenerlos, lo cual produce un gran beneficio para la empresa, dado que en general, es más costoso obtener un nuevo cliente que mantener uno existente.

Usando estrategias para crear lealtad de clientes es que las empresas han creado la tarjeta de fidelización, donde premian a los clientes más leales, dándoles descuentos, posibilidad de pagar en cuotas y permitiéndoles acumular puntos que luego de ser acumulados pueden ser canjeados por otros productos (esto en el caso particular de la empresa donde se realiza la memoria). Uno de los pioneros en este tema fue American Airlines, quien con su programa de viajero frecuente, generó una mayor fidelidad de sus clientes y mayor rentabilidad para la empresa.

Otra estrategia usada en la actualidad, consiste en ofrecer productos con una forma de pago combinada, por ejemplo, si una persona acumula millas en una aerolínea, puede comprar un pasaje pagando una cierta cantidad de dinero y ocupando una cierta cantidad de las millas que tiene acumuladas; o bien en el caso de acumular puntos para canjear productos, se está permitiendo canjearlos con una cierta cantidad de dinero y el resto con los puntos acumulados; de esta manera los clientes tienden a percibir un menor costo por el producto o servicio, y con esto pueden eventualmente generar mayores ganancias para la empresa (Dreze y Nunes, 2004). Donde las ganancias pueden ser valoradas como disminución en el costo percibido, o como aumento en los ingresos de la empresa.

A pesar de que en un principio, la idea era crear lealtad a través de los programas mencionados anteriormente, con el tiempo estos se fueron masificando y por lo tanto el

impacto que producían en la gente ya no era tan relevante, sin embargo, los datos que se tienen gracias a este programa, pueden ser utilizados de manera muy importante para conocer al cliente. Además, el no poseer algún tipo de programa de lealtad, se ha vuelto una desventaja competitiva para las empresas, dado que la mayoría de ellas lo posee.

7.2 Programa de Puntos de la Empresa

La tarjeta de fidelización a partir de la cual se obtiene la información para esta memoria, tiene las siguientes características con respecto a los puntos, que son una variable importante en el trabajo:

- Cada vez que un cliente compra o pide un avance utilizando la tarjeta del club de fidelización de la empresa, obtiene una cantidad de puntos virtuales, cuya cantidad es igual a 1 punto por cada 120 pesos gastados con la tarjeta en el caso de los montos gastados, y a 1 punto por cada 950 pesos en el caso de pedir un avance. Los puntos acumulados en un mes, corresponden a los puntos entregados en éste, más los puntos que tenía acumulados en el mes anterior, menos los puntos que canjea en ese mes y menos los puntos que se vencen en el mismo.
- Los puntos que se entregan a un cliente tienen una duración de 12 meses, a partir del último día del mes en el cual se entregaron. En consecuencia, los puntos vencen el primer día del mes 13 posterior a la compra realizada. Por ejemplo, si una persona obtiene 100 puntos algún día en el mes de enero, en el mes de febrero del próximo año, esta persona tendrá 100 puntos menos
- Los clientes pueden canjear sólo si llegan a un nivel de canje designado por la empresa, en el cual existen productos asociados de acuerdo al nivel alcanzado. A mayor nivel, mejor premio (más caro) y más puntos debe acumular un cliente para poder canjear.
- Existen 8 niveles de canje y sólo se puede canjear al tener la cantidad de puntos acumulados indicada por cada nivel. Por ejemplo, no se pueden canjear 5500 puntos, si tiene esa cantidad tendría que canjear 5000 y el resto quedaría en su inventario de puntos. Los niveles son:

Nivel 1 = 5000 ptos.

Nivel 2 = 9000 ptos.

Nivel 3 = 12000 ptos.

Nivel 4 = 24000 ptos.

Nivel 5 = 48000 ptos.

Nivel 6 = 60000 ptos.

Nivel 7 = 90000 ptos.

Nivel 8 = 120000 ptos.

- Los puntos son descontados en base a la teoría FIFO (First In First Out), es decir, en el caso que un cliente realice el canje de un producto, se le

descontarán los primeros puntos que le fueron entregados de acuerdo a su inventario de puntos. Además se le descuenta sólo la cantidad correspondiente al nivel de canje, es decir, si un individuo tiene 6500 puntos acumulados y quiere canjear en el nivel 1, se le descontarán 5000 puntos y quedará con un inventario de 1500 puntos.

7.3 CRM

Customer Relationship Management (CRM) es el proceso de coleccionar y analizar las interacciones que un cliente tiene con una firma con el fin de mejorar el valor del cliente para la firma. Además de obtener información necesaria para hacer una gestión más eficiente tanto de los clientes como de la empresa en sí (Kamakura y Melo, 2004).

La estrategia de CRM, pretende crear una relación más cercana con el cliente, para conocerlo mejor y poder tener un trato personalizado con él. De esta manera se busca aumentar la satisfacción del cliente y así también la lealtad. Además, se pueden manejar mejor las estrategias de marketing, y así aumentar las ventas a través de up-selling y cross-selling, por mencionar algunas, dado el conocimiento del comportamiento del cliente.

Igualmente hay que señalar que en general, es menos costoso retener a un cliente que adquirir uno nuevo (Wansink, 2003), por lo que crear estas relaciones con los clientes aumenta la probabilidad de retenerlos, por ende es lo mejor para las empresas. Es más, Gupta, Lehmann y Stuart (2002) encontraron que 1% de mejora en la tasa de retención, puede incrementar el valor de la firma en un 5%.

Es por lo anterior que se crean los programas de lealtad basados en CRM, para formar relaciones estables con los clientes de la empresa. Aunque hay que señalar que estos programas con el tiempo se han vuelto tan masivos, que no tener algún tipo de programa de fidelización, es una desventaja crucial, y tenerlo no crea necesariamente la lealtad de la gente.

Para que los clientes se suscriban a estos programas, existen una serie de incentivos, tanto físicos como psicológicos. En los físicos destacan: descuentos por compra de productos, entrega de puntos por la compra, posibilidad de pagar en cuotas, etc. Entre los beneficios psicológicos destacan: la identificación del individuo con la empresa; la sensación de mayor utilidad de la compra, debido a la entrega de algo extra (por ejemplo puntos o millas); la satisfacción a medida que se acerca a la meta para obtener los beneficios (en el caso de puntos o millas acumuladas) y la obtención del beneficio; entre otros.

Finalmente cabe mencionar que el cliente tiende a comprar más a medida que se acerca a la meta (Kivetz, Urminsky y Zheng, 2006), con un beneficio asociado, es decir, si el individuo posee un avance hacia el objetivo, tiende a gastar más, por lo que las

empresas podrían proveer a sus clientes con un avance artificial hacia la meta y a la vez mover la meta en la cantidad del avance entregado. Con esto los individuos pondrían un mayor esfuerzo en alcanzar el objetivo (Nunes y Dreze, 2006), lo cual permitiría a la empresa obtener mayores beneficios

7.4 Modelo para la Estimación de Parámetros

Cuando se cuenta con datos que fueron generados por variables observadas independientes que poseen una cierta distribución de probabilidad asociada, se pueden calcular parámetros asociados a estas variables que corresponden a la incidencia de ellas en la generación de los datos.

Por ejemplo, si se tienen los datos de n clientes $Y = (y_1, \dots, y_n)$ para el mes t , entonces:

$$Y_t = f(X_t, \beta)$$

Con X_t la matriz de variables observadas y β la matriz de parámetros

Por ejemplo, en un modelo lineal se tendría lo siguiente:

$$Y_t = \beta_0 + \beta_1 X_{1t} + \beta_2 X_{2t} + \dots + \beta_m X_{mt} + \varepsilon_t$$

Donde X_{it} es el vector de la variable i observada en el tiempo t

En este caso se requiere estimar el valor de los parámetros $(\beta_0, \beta_1, \dots, \beta_m)$ de manera que se obtenga el vector Y_t

Para esto generalmente se usa la técnica de mínimos cuadrados, que minimiza la suma del error cuadrático medio, es decir, la suma de la diferencia entre el valor estimado y el valor observado, al cuadrado, dividido por el número de observaciones. Sin embargo, este tipo de estimación es usada principalmente cuando se hacen regresiones lineales, donde con este método se obtiene un estimador consistente y eficiente, pero cuando las regresiones son no lineales, este método ya no es tan eficaz, por lo que se busca una manera alternativa de obtener los parámetros.

En este contexto aparece la opción de hacerlo por medio de la máxima verosimilitud (MLE). Este método busca la combinación de parámetros que con mayor probabilidad generaron los datos, es decir, supone que las variables observadas siguen una cierta distribución y la función de densidad conjunta de estas variables es la que forma los datos, por lo que son generados con una cierta verosimilitud de los parámetros de cada distribución. Este método tiene propiedades óptimas en la estimación, tales como suficiencia, consistencia, eficiencia e invarianza en la parametrización (Myung, 2003).

La función de verosimilitud (L) se define como:

$$L = L(\beta | Y) = \prod_{i=1}^n f(y_i | \beta) = L(\beta | y_1, y_2, \dots, y_n)$$

Donde $f(y_i | \beta)$ es la función de densidad de probabilidad de observar el vector de datos Y dado el vector de parámetros β .

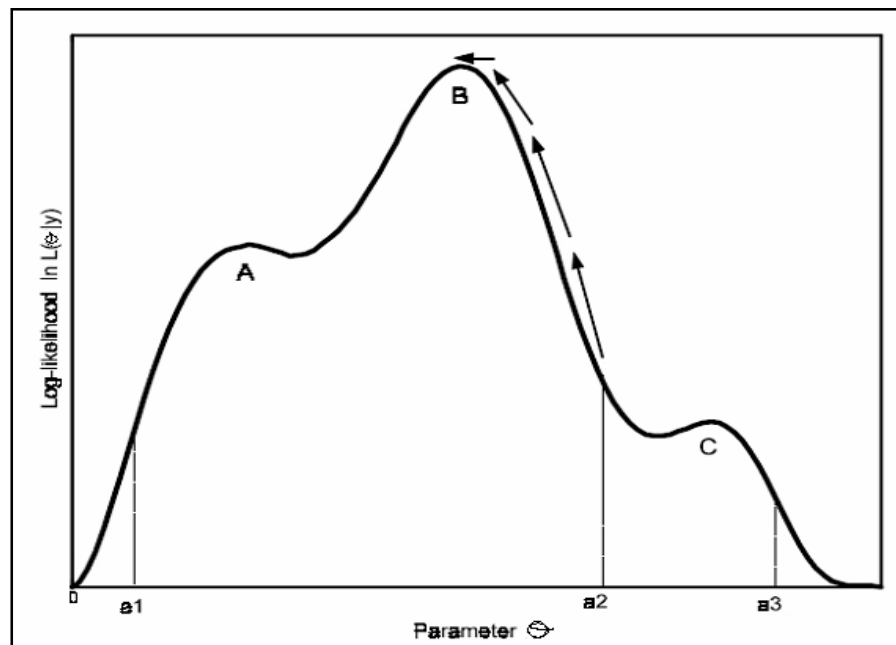
$L(\beta | Y)$ es la función de verosimilitud del vector de parámetros β dado el vector de datos observados Y , la cual corresponde a la multiplicación de las funciones de densidad de estas observaciones (densidad conjunta), asumiendo que son independientes.

Cabe mencionar que esta función (L) toma valores entre 0 y 1 dado que se trabaja con probabilidades y el producto entre ellas.

El método de máxima verosimilitud consiste en maximizar la función L con respecto a los parámetros, y de esta manera obtener aquellos que con mayor verosimilitud generaron los datos.

Además hay que señalar que a veces cuando hay muchos parámetros a estimar y las funciones de densidad son altamente no lineales, conviene hacer una modificación en el método, ésta corresponde a maximizar el logaritmo de L ($\log L$) con respecto a los parámetros. En este caso el $\log L$ toma valores que van entre $-\infty$ y 0.

Figura 1. Ejemplo de una función $\log L$ con parámetro θ .



Fuente: I. Myung (2003).

La función L o $\log L$ puede tener más de un máximo (Fig. 1), y Al trabajar con algoritmos de optimización, pueden haber algunas complicaciones, tales como que no exista un máximo global (punto B en fig. 1), o bien que al buscarlo, se encuentre primero un máximo local (punto A o C en fig. 1), dependiendo del punto de partida de la optimización. Es por eso que al trabajar con máxima verosimilitud es necesario probar con distintos puntos de partida de los parámetros, y en caso de encontrar dos óptimos se opta por él con mayor L o mayor $\log L$ según corresponda.

Cuando se trabaja con distinta cantidad de parámetros en la función L , no se puede elegir entre una y otra con el criterio de la que sea mayor, ya que al aumentar el número parámetros el ajuste de la función aumenta o se mantiene pero nunca disminuye. No obstante, existen criterios para seleccionar entre distintos tipos de parámetros usando el valor de $\log L$. Los criterios mayormente usados corresponden al criterio de información de Akaike (AIC) y el criterio de información bayesiano o de Schwarz (BIC). Según estos criterios, mientras menor sea el valor de AIC o BIC (dependiendo de cual se utilice), mejor será el ajuste y se elige ese modelo (Durán, 2008).

Los criterios mencionados se calculan como sigue:

$$AIC = 2*(k - \log L_{\max}) \quad \text{y} \quad BIC = k*\ln(n) - 2*\log L_{\max}$$

Donde:

k es el número de parámetros estimados.

n es el número de observaciones.

$\log L_{\max}$ es el valor máximo que toma $\log L$.

Estos criterios incorporan la calidad del ajuste con el número de parámetros, por lo que buscan obtener la menor cantidad de parámetros que resultan en un buen ajuste del modelo.

En el caso de tener distinto criterio de selección según AIC y BIC, Carvajal y Marco (2005) dicen que se prefiere BIC, dado que es consistente a diferencia de AIC (la varianza no decrece con el incremento del tamaño de la muestra), consecuente con esto, la variabilidad muestral de AIC para determinado orden no decrece con el incremento del tamaño muestral. Por lo tanto AIC tiende a sobreestimar el orden del proceso.

Según los mismos autores anteriores, la principal diferencia entre estos dos criterios, es que el segundo penaliza mucho más al usar un parámetro extra. Además BIC tiene más bajo sesgo y variabilidad que AIC. Asimismo BIC tiende a seleccionar modelos con menos parámetros (más parsimoniosos) e identifica el orden del proceso en forma más precisa que AIC.

7.5 Segmentación con Modelos de Clases Latentes

La segmentación es una herramienta que permite estudiar el comportamiento de los clientes, con el fin de poder diferenciarlos por éste, agrupando en un mismo segmento a personas con características más similares entre ellos que con personas que se encuentran en un segmento diferente. Así se puede discriminar entre los distintos grupos y ser abordados por las empresas de formas diferentes.

La segmentación está enfocada principalmente en lograr este trato diferenciado con los clientes, y de esta manera hacer un marketing mix mucho más efectivo sobre los segmentos a los cuales la empresa quiere satisfacer.

En general, la segmentación tiene un proceso estándar, a pesar de que existen distintos métodos de segmentación. Se pueden mencionar los siguientes pasos que serán realizados en esta memoria:

- Decidir las variables relevantes para la segmentación.
- Elegir la metodología a utilizar para el análisis de datos.
- Aplicar la metodología e identificar los segmentos relevantes.
- Describir los segmentos (Caracterizarlos).
- Seleccionar los segmentos objetivos.

Con respecto al tipo de segmentación se pueden encontrar segmentaciones a priori y segmentaciones a posteriori. Las primeras designan grupos de acuerdo a la similitud de factores conocidos con anterioridad (demográfico, geográficos, etc.). En cambio las a posteriori designan los grupos luego de los estudios del comportamiento de la gente.

En el presente trabajo no sólo se busca segmentar a los clientes, sino también, estimar los parámetros que influyen en el comportamiento de ellos en cada segmento, para así poder pronosticar posibles comportamientos futuros. En particular, se quieren pronosticar los montos a gastar por los clientes tomando en cuenta las variables que influyen en él.

Para esto se podría realizar análisis de dos pasos, donde primero se hiciera una segmentación a priori con respecto a las variables de interés, y luego se calcularan los parámetros correspondientes a cada variable, para cada grupo. O bien, se podría estimar los parámetros de las variables correspondientes a cada individuo, y luego agrupar en un mismo segmento a los que tengan parámetros similares.

El problema del primer caso, es que al segmentar de acuerdo a las variables, y luego encontrar los parámetros, se puede tener en un mismo grupo a personas que reaccionan de distinta manera frente a una variable. Por ejemplo con respecto a la variable de edad, puede haber personas que les afecte negativamente tener más edad y comprenden menos, sin embargo, otras personas compran más a medida que tienen más edad. Por lo que los parámetros no serían muy bien estimados.

El problema con el segundo caso es que estimar los parámetros para cada cliente es poco viable, ya que en general se cuenta con una gran cantidad de clientes y se necesitaría mucho tiempo para lograrlo.

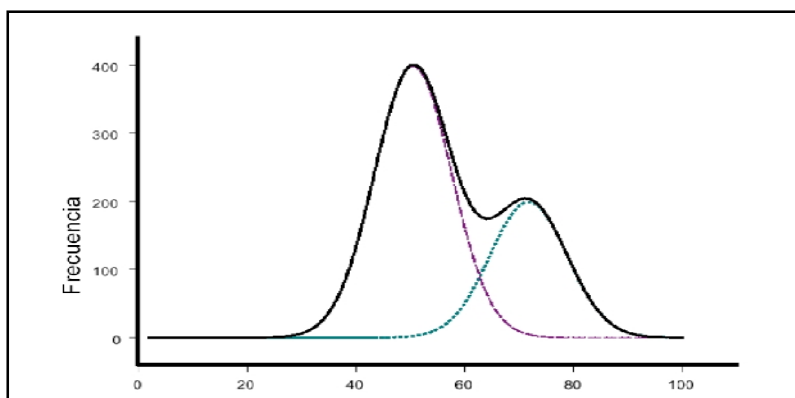
En el presente trabajo se utiliza un método que simultáneamente identifica los distintos segmentos y los parámetros correspondientes a cada uno de ellos, con el fin de agrupar por la sensibilidad de lo que se quiere pronosticar frente a las variables que se utilizan para ello. Es así como se segmenta de acuerdo a la sensibilidad de la gente frente a las variaciones en las variables y no con respecto a las variables en sí.

El método que realiza este tipo de segmentación es el llamado modelo de clases latentes, o bien regresiones de mezclas finitas.

El modelo de clases latentes es un método estadístico que estudia la existencia de una o varias variables latentes a partir de un conjunto de variables explicativas observadas; y define, a partir de sus clases (o categorías), una clasificación de los individuos analizados. Esto suponiendo la existencia de una asociación entre las variables observadas y la variable no observable directamente. De esta manera se pueden clasificar los individuos en grupos de acuerdo a su comportamiento, a pesar de que no se conoce la cantidad de individuos en cada uno de ellos, la media, ni la varianza de éstos (Anderson, 1954).

La idea es que los individuos que pertenecen a la misma clase (grupo), provienen de las mismas distribuciones de probabilidad y sus parámetros pueden ser estimados. Así cada grupo tiene sus propios parámetros y los individuos de éste se comportan de manera similar estadísticamente hablando. Consecuentemente, las personas forman una población, que es una mezcla de distintos segmentos en proporciones desconocidas, por lo que se desconoce a priori el segmento al que pertenece un sujeto (Picón, 2004). La figura 2 muestra una función de densidad observada y sus funciones subyacentes; la primera representaría a la población y cada una de las segundas representaría un grupo de la población.

Figura 2. Ejemplo de función de densidad observada y dos funciones adyacentes



Fuente: Eduardo Picón (2004).

Los modelos de clase latente se explican a continuación:

Se tienen p variables observables, X_1, \dots, X_p , además existe una variable latente X_q con J clases (categorías). Luego, al tener n individuos, se asume que ellos y sus variables asociadas $Y_n = (X_{n1}, X_{n2}, \dots, X_{np})$ provienen de una composición de J segmentos en proporciones $\pi_1, \pi_2, \dots, \pi_j$, sin conocer la probabilidad de pertenencia de cada individuo a los distintos segmentos.

De esta manera se puede decir que el vector Y_n proviene de una función de distribución de la forma $f_j(Y_n / \theta_j)$, donde θ_j es el vector de parámetros desconocidos de la distribución dada, correspondiente a la clase j .

Por lo tanto, la función de distribución final para Y_n ; compuesta por las distribuciones de cada clase y los parámetros asociados a cada una de ellas, además de las proporciones asociadas; está dada por:

$$f(Y_n / \theta, \pi) = \sum_{j=1}^J \pi_j f_j(Y_n / \theta_j)$$

Donde $\sum_{j=1}^J \pi_j = 1$ y $\pi_j \geq 0, \quad \forall j = 1, 2, \dots, J$

Como cada clase tiene sus propios parámetros, al tener una mayor cantidad de segmentos, el número de parámetros aumenta considerablemente dependiendo de la cantidad de variables observadas (independientes).

7.6 Estimación del Modelo de Clases Latentes

Para resolver este modelo existen dos alternativas mayormente utilizadas, la primera es el conocido algoritmo de Newton-Raphson, el cual necesita relativamente pocas iteraciones para converger, pero su convergencia no está garantizada. Y la segunda alternativa es el algoritmo EM (por estimación y maximización), el cual necesita una mayor cantidad de iteraciones para converger, y es muy sensible a los parámetros de partida, sin embargo, su convergencia al menos a un óptimo local está asegurada (Mallea, 2004). Este último será el método utilizado en la memoria.

Lo que se busca es encontrar los parámetros correspondientes a cada segmento y la proporción de cada uno de ellos en la población, es decir el vector $\phi = (\pi, \theta)$, para lo cual se usa la función de máxima verosimilitud de la forma:

$$L(\phi; y) = \prod_{n=1}^N f(y_n / \theta, \pi)$$

Y en la mayoría de los casos, sin pérdida de generalidad se utiliza el logaritmo de esta función, ya que simplifica los cálculos de maximización.

El algoritmo EM consta de dos etapas, una de estimación y otra de maximización. Primero, en la etapa de estimación, se calculan todos los valores esperados de los parámetros, dados los valores observados y la estimación provisional de los no observados, es decir, se estima el vector $\phi = (\pi, \theta)$, para obtener la función de verosimilitud de los parámetros. Luego en la segunda etapa, se maximiza la función de verosimilitud de todos los datos a partir de los valores esperados calculados en el paso anterior. Entonces se itera hasta alcanzar convergencia.

Así se obtienen los parámetros para cada segmento, que con mayor probabilidad generaron la muestra, y también se obtienen las proporciones de cada uno en la población.

Finalmente se asigna un individuo a la clase de la variable latente que le corresponda para lo cual se calcula la probabilidad condicionada de que el individuo de un cierto comportamiento pertenezca a la clase j . Por lo tanto se calcula:

$$p_{nj} = \frac{\pi_j f_j(y_n / \theta_j, \pi_j)}{\sum_{j=1}^J \pi_j f_j(y_n / \theta_j, \pi_j)}$$

Entonces un individuo se asigna a la clase j , donde esta última probabilidad es mayor, clasificación denominada *modal* (Picón, 2004). También existe la clasificación probabilística o solapada, la cual asigna probabilísticamente a los individuos a cada segmento. En este caso se utilizará la clasificación *modal*, ya que los resultados mostrarán que los individuos son asignados a un segmento con una probabilidad de pertenencia cercana a 1, por lo que ambos tipos de clasificaciones son similares, y la clasificación modal facilita los cálculos y posibles comparaciones.

7.7 Medidas para la Calidad del Ajuste

Una medida interesante es la mencionada por Bartholomew, Steele y otros (2002), que compara el ajuste de un modelo con una variable latente de j clases, y el modelo de independencia, este último indica que no existe una variable latente que tenga distintas asociaciones de las variables observadas. Esta medida ($\%L^2$) expresa el grado de asociación explicado por la variable latente.

$$\%L^2 = \frac{L_0^2 - L_q^2}{L_0^2}$$

El criterio de decisión para esta medida es que mientras más cercano a 1 sea ρ^2 , mejor será el modelo.

Otra medida para observar la calidad del ajuste del modelo, es el estadístico mencionado por K. Train (2003), llamado “the likelihood ratio index” (ρ), el cual mide que tan bien funciona el modelo con los parámetros estimados versus un modelo donde todos los parámetros son 0. El valor de éste índice se calcula como sigue:

$$\rho = 1 - \frac{LL(\hat{\beta})}{LL(0)},$$

Donde $LL(\hat{\beta})$ es el valor de la función de log-likelihood con los parámetros estimados, y $LL(0)$ es el valor de la función cuando todos los parámetros son 0. Este indicador toma valores entre 0 y 1; siendo 0 si los parámetros estimados no mejoran de los valores 0, como si no hubiera modelo; y tomando el valor 1 si los parámetros estimados se ajustaran perfectamente a los datos, con lo cual $LL(\hat{\beta})$ tomaría el valor 0 y por lo tanto $\rho = 1$.

Cabe mencionar que como los individuos son asignados a una clase de la variable latente, dado un patrón de respuesta y de acuerdo a una probabilidad, existe la posibilidad de una mala clasificación. Por lo tanto, se calcula la proporción esperada de los errores de clasificación según la fórmula expuesta a continuación:

$$E = \sum_{i_1, \dots, i_p, j} |1 - \max(\pi_{j/i_1, \dots, i_p})| \hat{\pi}_{i_1, \dots, i_p}$$

Se espera que este valor sea cercano a cero, lo que indica que se espera que los errores de asignación sean pequeños. Con esto el porcentaje de buena clasificación es:

$$PBC = (1-E)*100$$

Asimismo puede calcularse un indicador (Goodman & Kruskal) con respecto a la calidad de asignación a una clase concreta, de la siguiente forma:

$$\lambda_j = \frac{|1 - \max(\hat{\pi}_j) - E}{1 - \max(\hat{\pi}_j)}$$

Si este indicador es cercano a 1 (100%), entonces la calidad de asignación a la clase j es buena porque los errores cometidos en ésta, son muy pequeños.

El error de clasificación y el indicador de Goodman & Kruskal son mencionados por A. González y M Monteavaro (2003).

7.8 Comparación con Otros Métodos

- A diferencia de las técnicas estándares de segmentación, la de clase latente asume que una mezcla de las distribuciones de probabilidad generan los datos y es un modelo estadístico, por lo que el criterio de segmentación es menos arbitrario (Magidson y Vermunt, 2002).
- Dado que es un modelo estadístico, la elección del número de segmentos puede ser determinada de acuerdo a criterios formales, tales como BIC, AIC y L^2 entre otros, a diferencia de K-means, donde es menos asistida la elección.
- No se necesita saber la probabilidad de pertenencia a priori a un grupo, dado que se trabaja asumiéndola desconocida y estimándola, a diferencia de análisis discriminante y regresiones logísticas que necesitan conocer estas pertenencias.
- En este modelo los individuos contribuyen a la media de un grupo en particular con un peso igual a la probabilidad de pertenencia a posteriori a este grupo, en cambio otros métodos, como K-means, contribuyen con un peso igual a 0 o 1, lo que no es apropiado en caso de clasificaciones mixtas. Cabe mencionar que el método Fuzzy set (Fuzzy-means), también trabaja con probabilidades de pertenencia a los grupos, sin embargo no posee variable dependiente, por lo que no es posible calcular la sensibilidad de los individuos en una variable dependiente, frente a cambios en otras variables explicativas.
- Permite incluir variables exógenas simultáneamente en el modelo, a diferencia de otros que tienen que describir diferencias entre grupos con respecto a estas variables usando un análisis discriminante posterior a la selección de los grupos.
- Es capaz de realizar simultáneamente la segmentación de los clientes y encontrar los parámetros que mejor se ajustan a los datos correspondientes a cada segmento. En cambio para llegar a resultados similares otros métodos tendrían que segmentar en base a las variables y luego estimar el valor de los parámetros correspondientes a cada segmento. O bien, estimar los parámetros de manera individual y luego agrupar en segmentos los individuos con valores similares de parámetros. Estos procedimientos bifásicos optimizan dos criterios diferentes y no necesariamente relacionados entre sí (Picón, 2004), a diferencia del método de clases latentes, que intenta optimizar simultáneamente de acuerdo a un mismo criterio de interés, que normalmente corresponde a la maximización de la función de verosimilitud logística.

Finalmente, hay que señalar que el modelo de clases latentes trabaja con la sensibilidad de la variable dependiente frente a las variables independientes, a diferencia de otros métodos que dividen en base a las variables independientes en sí mismas, lo que podría provocar que personas que tienen distintos tipos de sensibilidad frente a una variable, se encuentren en el mismo grupo.

Cada método tiene sus ventajas y desventajas, y éste no es la excepción. A pesar de las ventajas mencionadas anteriormente, estos modelos también tienen sus limitaciones, como por ejemplo suponen la existencia de una variable latente, la cual, no necesariamente tiene que existir en todos los casos. Además dado que se usa un método de maximización para estimar estos modelos, el máximo podría no existir, o simplemente podrían existir varios máximos, por lo cual el modelo sería difícil de estimar.

La elección de uno u otro método depende principalmente de los objetivos buscados y de la decisión personal del investigador. En esta memoria se trabaja con modelos de clases latentes, ya que lo que se busca es segmentar y calcular los valores de los parámetros asociados a cada segmento y dadas las características mencionadas anteriormente; especialmente la realización de estos dos pasos simultáneamente con un criterio común, y el trabajo con la sensibilidad de los parámetros frente a una variable dependiente; esta técnica es apropiada para los objetivos y se espera que entregue buenos resultados.

7.9 Determinación del Valor del Cliente

Los clientes tienen un ciclo de vida asociado a la relación que llevan con la empresa, y este ciclo tiene un valor para ella, es decir, como sea el comportamiento de compra del individuo durante el tiempo que dure su relación con la empresa, definirá cuánto vale para ésta.

El valor del ciclo de vida de un cliente (Customer Lifetime Value o CLV) se define como el valor presente de las ganancias futuras que el cliente generará para la empresa en un periodo de tiempo.

Este indicador es fundamental para las empresas y permite desarrollar una mejor estrategia de CRM, dado que al conocer el valor de los clientes se puede determinar cuántos recursos utilizar en ellos, cuántos esfuerzos de marketing brindarles y cuánto tiempo dedicarles.

Para calcular el CLV existen varias técnicas distintas (Gupta, Hanssens y Hardie, 2006), a continuación se muestran dos formas alternativas, generalmente utilizadas para realizar este cálculo:

$$(1) \quad CLV = \sum_{t=0}^T m_t \frac{r^t}{(1+d)^t} - AC \quad \text{o} \quad (2) \quad CLV = (\hat{M} - \hat{C}) \cdot VP(d_{anual}, \hat{x})$$

Donde:

- T: horizonte de tiempo para estimar CLV
- m_t : margen generado por un cliente en el tiempo t
- r: tasa de retención de un cliente

d: tasa de descuento

AC: costo de adquisición de un cliente

\hat{M} : Monto promedio pronosticado por transacción

\tilde{C} : Estimación de costo asociado al cliente

$VP(d_{\text{anual}}, \hat{x}) = \text{DET}$: valor presente de transacciones futuras con \hat{x} el número de transacciones pronosticadas y d_{anual} la tasa de descuento anual.

Este último elemento (DET) incluye la probabilidad de un cliente de estar activo en el tiempo.

La diferencia entre los dos modelos, es que el primero (Gupta, Lehmann y Stuart, 2002) considera una tasa de retención de clientes y el segundo (Fader y Hardie, 2004) ocupa la probabilidad de un cliente de estar activo en el tiempo.

La forma más común de estimar el CLV es mediante modelos de RFM (Recency, Frequency y Monetary value), pero existen modelos probabilísticos, econométricos y de crecimiento entre otros, los cuales, junto a otros tipos de modelos, son explicados a grandes rasgos por S. Gupta, D. Hanssens y B. Hardie (2006). En general se usa el modelo probabilístico de Pareto/NBD utilizado por Fader y Hardie (2004), pero en esta memoria se optará por modelos econométricos debido a las siguientes razones:

- Modelo de Pareto/NBD asume una serie de condiciones que no necesariamente se consideran en este caso, por ejemplo que si un individuo deja de comprar en un periodo de tiempo entonces ya no es cliente de la empresa (es inactivo permanentemente)
- Modelo de Pareto/NBD sólo considera variables de Recency y Frequency para estimar el número de transacciones esperadas, sin embargo, en el modelo econométrico se pueden incluir otro tipo de variables (por ejemplo demográficas).
- Modelo Pareto/NBD tiende a sobrestimar el número de transacciones futuras para clientes con baja frecuencia, obteniéndose un alto valor presente del número de estas transacciones (DET del modelo). Asimismo, tiende a subestimar este número de transacciones para clientes que tienen una alta frecuencia (Marín, 2005). Esto, debido a que supone que un cliente que compra siempre, debe mantener este nivel, por lo que si deja de comprar en un periodo corto de tiempo, asume que el cliente está inactivo. Análogamente los clientes que compran solo a veces, si dejan de comprar por un periodo corto de tiempo, no se puede asumir que están inactivos. Es decir, castiga más a los clientes frecuentes.
- Modelo Pareto/NBD estima el monto por transacción según una distribución gamma-gamma, lo cual incluye asumir varias condiciones de los parámetros necesarios. En cambio el modelo econométrico a aplicar estima el monto gastado considerando todas las variables independientes que resulten significativas, incluyendo tanto variables de R, F y M, como variables de estacionalidad y variables que representan características personales de los clientes.

A pesar de lo anterior, cabe destacar que los modelos consisten en distintos enfoques, en consecuencia, es difícil decidir entre uno y otro como el más correcto. Por lo que finalmente depende de cada investigador el modelo a utilizar y en este caso serán modelos econométricos.

Con esta decisión, el cálculo del CLV será como el mostrado en la ecuación (1), donde el m_t corresponderá al monto que el cliente gasta en el tiempo t (se considerará que el costo de adquisición del cliente es un costo hundido, por lo que no se incluye en el cálculo), r será la tasa de retención que se considerará constante para los distintos periodos de tiempo y su valor dependerá exclusivamente de lo designado por la empresa (aunque habrá un análisis de sensibilidad al respecto), y en un principio será considerado como 1, ya que en general los clientes que poseen la tarjeta no cierran nunca la cuenta, y los casos que lo hacen son despreciables.

8 Desarrollo de Metodología

8.1 Selección de Datos

8.1.1 Análisis Previo a la Selección

Para hacer este estudio se cuenta con la información de 2.363.621 clientes, pero se trabajará con una muestra aleatoria de ellos según el criterio mostrado en el anexo 1. Entonces se trabaja con 59.000 clientes aproximadamente.

Como un primer paso para seleccionar los datos a ocupar, se realiza un análisis del comportamiento de los clientes³, separándolos en tres grupos, de acuerdo al uso de los beneficios de la tarjeta de clientes, particularmente con respecto a los puntos que se obtienen al comprar con ella. Estos grupos consisten en los clientes canjeadores, los que no canjean a pesar de que tienen los puntos necesarios y los no canjeadores.

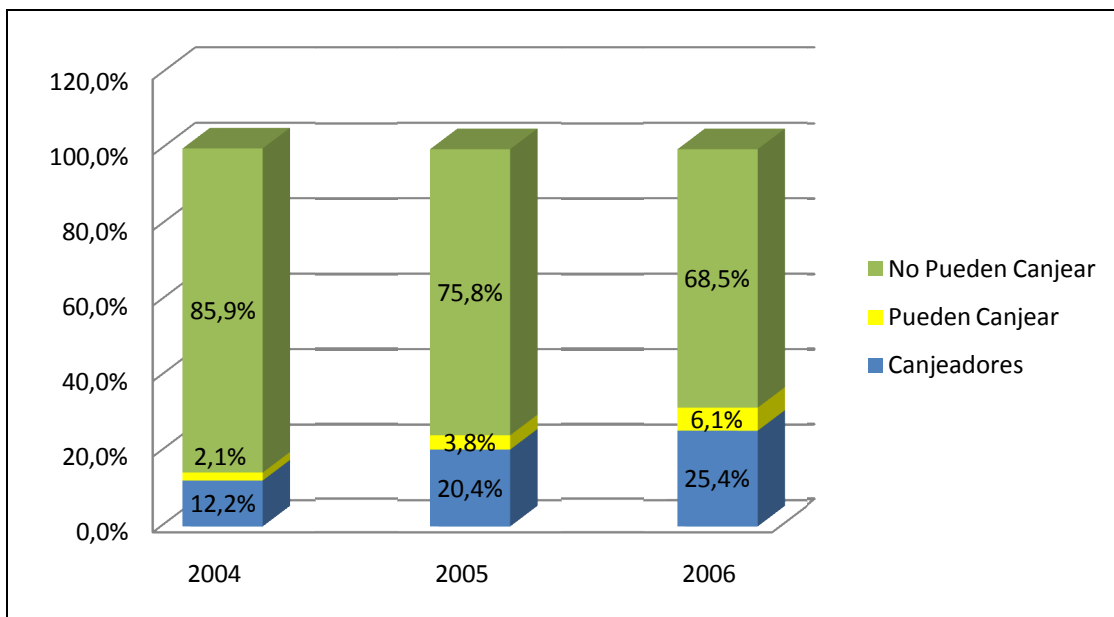
El grupo de los canjeadores está formado por clientes que en el periodo de estudio (desde marzo del 2003 hasta abril del 2008), realizaron al menos un canje de producto de acuerdo a la cantidad de puntos que tenía acumulados. Por su parte, el grupo de los no canjeadores que podrían haber canjeado, correspondía a los clientes que no realizaron canje alguno en este periodo, a pesar de que tenían acumulados suficientes puntos para hacerlo. Finalmente el grupo de los no canjeadores corresponde a los clientes que no acumularon suficientes puntos para canjear.

Como se puede apreciar en la figura 3, la cantidad de clientes que no canjean es mucho mayor que la cantidad de clientes que si lo hacen, sin embargo, para las empresas esto es llamado el "80-20", donde el 80% de los clientes generan el 20% de las ganancias y 20% restante de los clientes genera el 80% de las ganancias

³ Trabajo realizado por M. Durán (2008).

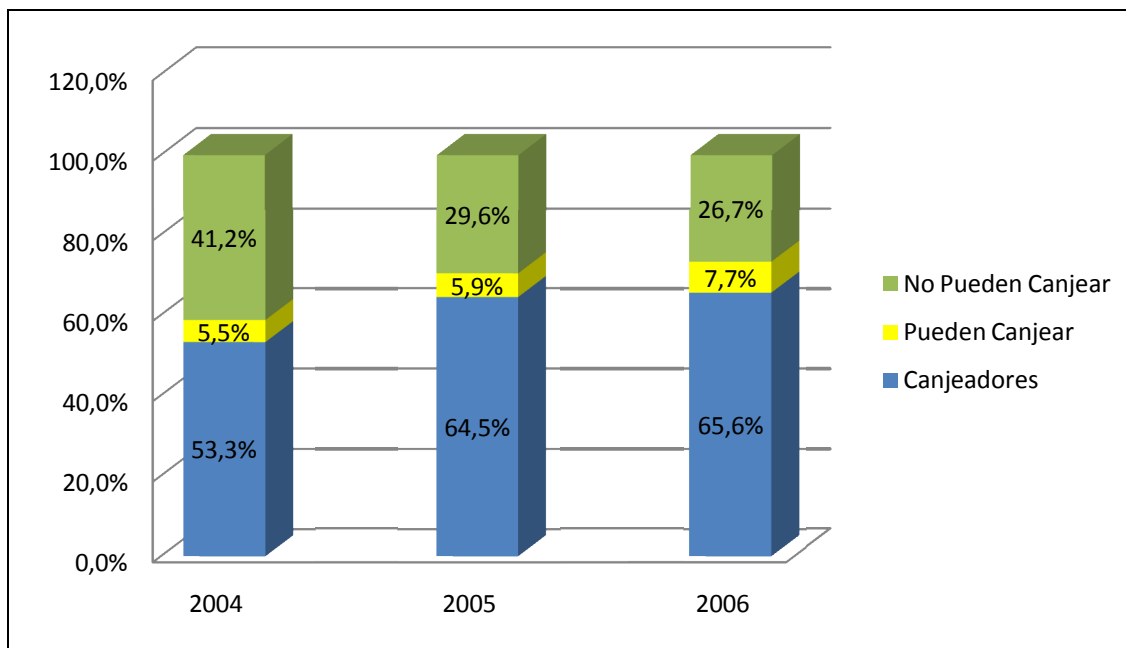
(Andersen, 2000). En la figura 4 se puede apreciar que la teoría no es tan distante de la realidad, ya que al observar la proporción de los montos gastados por cada tipo de cliente, se ve que aproximadamente el 50% y más de los montos gastados provienen de los clientes canjeadores.

Figura 3. Cantidad de clientes por comportamiento de canje



Fuente: M. Durán (2008)

Figura 4. Proporción de montos gastados por clientes con distintos comportamientos de canje.



Fuente: M. Durán (2008)

De los gráficos mostrados anteriormente se puede apreciar que el grupo de los clientes canjeadores es muy atractivo y será el estudiado en esta memoria. Además esto coincide con los intereses de la empresa, quienes desean estudiar este tipo de clientes.

8.1.2 Selección

En esta memoria se utilizan datos transaccionales de clientes de una tienda de Retail, las transacciones fueron hechas con su tarjeta de fidelización en un periodo de 62 meses, desde marzo de 2003 hasta abril de 2008. Además se cuenta con información de características personales tales como edad, sexo y años que tiene la cuenta.

La información corresponde a clientes que tienen la cuenta habilitada (2.363.621 clientes), los cuales pueden ser separados en canjeadores y no canjeadores, donde el primer grupo representa el 38.5% (910.855 clientes) del total y el segundo corresponde al 61.5% restante (1.452.736 clientes).

Por simplicidad y sin pérdida de generalidad se seleccionará una muestra de 800 clientes obtenidos aleatoriamente del grupo de los canjeadores, los cuales tienen su cuenta activa, es decir, a la fecha no han tenido problemas de morosidad, por lo que no les han cerrado su cuenta. Además tienen la particularidad de ser canjeadores, porque es el segmento de mayor interés para la empresa, dado que son los que representan mayores ingresos para ésta.

Se considera la muestra como suficiente según el cálculo del Anexo 2, y se realiza esta selección, debido a que trabajar con demasiados⁴ datos resulta engorroso y difícil de realizar.

Para cada mes t se tiene la siguiente información por cliente:

- Monto Gastado (MG_t)
- Puntos Entregados (PE_t)
- Puntos Canjeados (PC_t)
- Puntos Vencidos (PV_t)
- Número de compras (NC_t)

Además se cuenta con la información mensual sobre el sexo, la edad y los años que el cliente tiene la cuenta.

⁴ Son 23 meses de datos y 19 variables, por lo que si el número de clientes es muy grande, por ejemplo 10.000, se debe trabajar con $10.000 \times 23 \times 19 = 4.370.000$ datos.

Cabe mencionar que como los datos están de manera mensual, si un cliente realiza más de una transacción en el mismo mes, se considera la suma de las transacciones en el monto gastado.

8.2 Procesamiento

Previo a la muestra de 800 clientes, los datos fueron procesados para eliminar los casos considerados como outliers. En este caso se excluyeron a las personas que tenían puntos entregados o canjeados negativos, lo cual se debía probablemente a una devolución del producto comprado o canjeado respectivamente.

También se eliminaron de la muestra los clientes que tenían puntos canjeados mayores que los puntos acumulados, lo cual consistía en un error, o bien existe la posibilidad de que las personas se aprovecharan del sistema y compraran un producto para llegar a acumular los puntos necesarios para el nivel de canje, y luego de canjear un producto, devolvían el producto comprado, con lo cual se le quitaban los puntos que se le habían entregado, a pesar de que el producto ya había sido canjeado y no se devolvía.

8.3 Transformación

Con los datos entregados anteriormente se procedió a obtener las variables que se utilizarán en el desarrollo de esta memoria, las cuales se mencionan a continuación, y serán explicadas posteriormente.

Las variables utilizadas para cada cliente en el mes t son:

- Puntos acumulados (PA_t)
- Tiempo pasado desde la última compra (Recency de compra, $Rcomp_t$)
- Frecuencia de compra ($Fcomp_t$)
- Monto promedio de compra ($Mcomp_t$)
- Variables de mes (corresponden a 11 dummies)
- Variable de sexo (sex)
- Edad
- Años que tiene la cuenta (A_{cta})
- Número de compras (NC_t)
- Distancia al nivel de canje más próximo ($Dist_t$)

A continuación se explica el cálculo de cada variable de la cual no se tenga la información directa.

Los puntos acumulados para cada cliente se calculan directamente de los datos observados mensuales, donde la cantidad acumulada para el mes t es:

$$PA_t = PE_t - PC_t - PV_t + PA_{t-1} \quad Y \quad PA_0 = PE_0 - PC_0 - PV_0$$

Para las variables de R, F y M se crea un indicador (α) que corresponde al mes en que el consumidor realizó la primera transacción, es decir, α puede tomar valores entre 1 y 62. Este indicador se crea para evitar el distinto criterio que se tendría con clientes que empiezan a comprar desde un mes posterior al inicio del periodo de estudio, con lo cual en los primeros meses tendrían por ejemplo todos los montos igual a 0, y al calcular el promedio de los montos gastados, éste sería más bajo de lo que realmente fue. De esta manera, al tener este indicador se puede considerar el periodo del cliente desde la primera vez que realiza la transacción.

Rcomp_t será igual a 61 si en el mes t el cliente no ha comprado ninguna vez, es decir, $t < \alpha$, y si lo ha hecho será el tiempo que ha transcurrido desde la última compra hasta el mes t.

Para calcular la frecuencia se crea una dummy (S_t) que sea igual a 1 si el cliente compra en el mes t y 0 si no.

Luego Fcomp_t será igual a 0 si el cliente no ha comprado ninguna vez hasta el mes t, es decir, $t < \alpha$, y si lo ha hecho, será como se muestra en seguida:

$$Fcomp_t = \frac{\sum_{k=\alpha}^t S_k}{t - \alpha + 1} \quad \text{con } S_k = 1 \text{ si el cliente compra en el mes } k$$

Al igual que para la frecuencia, Mcomp_t será igual a 0 si el cliente no ha comprado ninguna vez hasta el mes t, es decir, $t < \alpha$, y si lo ha hecho será igual a:

$$Mcomp_t = \frac{\sum_{k=\alpha}^t MG_k}{t - \alpha + 1} \quad \text{con } MG_k \text{ igual al monto gastado en el mes } k$$

Las variables de mes corresponden a dummies d2, d3,..., d12, donde d2 corresponde a febrero, d3 a marzo,..., d12 a diciembre. Estas variables toman el valor 1 si el mes en el que se encuentra corresponde al de la variable y 0 en el caso contrario. Se omite la dummy de enero para evitar colinealidad de las variables.

La variable de sexo toma el valor 1 si el cliente es hombre y 0 si es mujer.

Finalmente la variable dist_t es una variable continua que toma valores entre -1 y 1. Esta variable mide la distancia que hay entre la cantidad de puntos acumulados en el mes t, y el nivel de canje más cercano. Si la cantidad de puntos acumulados en el mes t coincide con el nivel de canje inferior, Dist_t=-1, si coincide con el nivel superior, dist_t=1. Si dist_t=0 quiere decir que la cantidad de puntos acumulados en el mes t está a la misma distancia del nivel de canje inferior y el superior.

La variable está definida sólo para puntos acumulados mayores que 4000, por lo que en caso contrario el valor de la variable no tiene sentido y no es considerado (-1 es el valor puesto a la variable en estos casos).

$$\text{El cálculo de } dist_t \text{ será: } Dist_t = \frac{2 * (PA_t - n_inf)}{n_sup - n_inf} - 1$$

Donde n_inf es el nivel de canje inferior y n_sup el superior, con respecto a los puntos acumulados en el mes t .

8.4 Montos Gastados Agregados

Para tener una visión global de cuánto gastan los clientes mensualmente y cuáles son las variables más significativas en este gasto, se realiza una estimación de los montos gastados para un mes, tomando en cuenta las variables que influyen en ellos. Es decir, se toma en cuenta el comportamiento de un cliente en base a las variables observadas y se pronostica cuanto será el monto que el cliente gastará al mes siguiente.

Para esto se usan modelos econométricos, en los cuales se tiene una variable dependiente (MG) y una serie de variables independientes que se asume que generaron la anterior.

El método consiste en calibrar el valor de los parámetros, para lo cual se considera un periodo de estimación de 23 meses utilizando distintos tipos de regresiones, para identificar la que calcula mejor los parámetros. Luego se utiliza un periodo de validación de 6 meses para observar que tan preciso se es en la predicción.

8.4.1 Estimación Montos Gastados

Para estimar los montos futuros primero se estiman los parámetros de las variables observadas que con mayor probabilidad los generaron en el pasado. Para esto se utiliza el método de máxima verosimilitud, con el cual se estima la variable de montos gastados (MG) para cada cliente, usando las combinaciones de los regresores que son más significativos.

Los regresores significativos dependen del tipo de regresión ocupada, y fueron identificados al momento de realizar la regresión, por lo que al encontrar valores no significativos, se volvía a estimar la regresión sin tomar en consideración estas variables. Esto se realizaba para obtener los parámetros asociados solamente a variables significativas.

La variable dependiente corresponde a los montos gastados en el mes t y las variables independientes son las encontradas en la parte de transformación de variables, descrita anteriormente. Así, se consideran las variables de tiempo pasado desde la última compra (Rcomp), frecuencia (Fcomp) y monto promedio gastado (Mcomp), junto con los puntos acumulados (PA), el número de compras (NC) y la distancia al canje más cercano (Dist); todas éstas correspondientes al mes $t-1$. A su vez, las variables de estacionalidad (d_2, d_3, \dots, d_{12}), la de años que tiene la cuenta (A_cta), el sexo (sex) y la edad, son consideradas en el mes t , ya que estas variables afectan en ese mes.

Tabla 1. Valores de los parámetros significativos para las distintas regresiones.

MG	Lineal	Log-normal	Poisson
Cte	23.723	9,906344	10,56371
PA	1,987703	5,04E-06	-4,58E-06
Dist			0,02206
Rcomp			-0,1195312
Fcomp	-44.662,03	0,9742795	0,2278407
Mcomp	0,7591172	0,0000016	2,53E-06
d2			0,0528447
d3	6.026	0,1615381	0,1591353
d4			0,00700
d5		0,17297	0,07570
d6			0,13556
d7			0,0963862
d8		-0,13103	0,0336402
d9		0,21874	
d10	7.633	0,14274	0,2449858
d11			0,1550182
d12	24.694,49	0,331	0,3901375
Sex		0,0731283	0,0404796
A_cta	-439		0,0034227
Edad			-0,0005844
NC	6.988	0,02764	0,0404507

Fuente: Mauricio Durán (2008).

El periodo considerado para la estimación es desde diciembre del 2005 hasta octubre del 2007, es decir desde el mes 34 al 56, y por lo tanto, el periodo de las variables consideradas en el mes “t-1” se utiliza desde el mes 33 al 55.

Los parámetros son calculados a nivel agregado para todos los clientes, por lo que se asume homogeneidad entre ellos, a pesar de que esto no es correcto por inspección, permite observar como las variables inciden en el comportamiento y cuáles son las variables relevantes o significativas.

En la tabla 1 se muestran los resultados obtenidos para los parámetros significativos, según un modelo de regresión lineal, uno log-normal y uno de Poisson. Sólo son consideradas las variables asociadas que son significativas al menos para alguno de los modelos utilizados, el resto de las variables al ser no significativas se ignoran en los resultados.

Cabe mencionar que estos tipos de regresiones, al igual que las variables particulares, son consideradas en base a la memoria anterior realizado por Mauricio Durán (2008), en la cual se mostró que éstas regresiones eran las más adecuadas para predecir los montos gastados de los individuos y que estas variables a su vez podían ser ocupadas para predecir el comportamiento de éstos.

Según los criterios de la calidad del ajuste de estos modelos, se observa que el modelo lineal es el que tiene mejor ajuste, como muestra la tabla 2, dado que tiene el menor valor tanto de AIC como de BIC.

Tabla 2. Criterios AIC y BIC para distintos modelos.

	LogL	AIC	BIC
Lineal	-240.575	481.168	481.238
Log-normal	-241.265	482.554	482.648
Poisson	-5,E+08	1,E+09	1,E+09

Para obtener los valores de los montos gastados, se utiliza el valor de los parámetros estimados en cada modelo. De este modo, la suma de la multiplicación de cada parámetro por su variable correspondiente, equivale al valor pronosticado para la variable de monto gastado en el caso de una regresión lineal. Para los otros dos casos, la suma antes descrita equivale al logaritmo de los montos gastados, por lo que es necesario aplicar exponencial a este valor, para obtener los montos gastados por los clientes.

Por lo tanto, para la regresión lineal se tiene $MG = \hat{\beta}_0 + \hat{\beta}_1 X_1 + \hat{\beta}_2 X_2 + \dots + \hat{\beta}_n X_n$. Y para la regresión de Poisson y log-normal se obtiene el monto de la ecuación

$\ln(MG) = \hat{\beta}_0 + \hat{\beta}_1 X_1 + \hat{\beta}_2 X_2 + \dots + \hat{\beta}_n X_n$. En ambos casos X_i es el valor de la variable i y β_i el valor del parámetro estimado asociado a dicha variable.

Los resultados de las predicciones para el periodo de estimación se muestran en la tabla 3, donde aparece el promedio mensual del monto gastado por cada cliente y el total gastado en el periodo. Esto para cada tipo de regresión y también para el monto real gastado de modo de poder comparar.

Tabla 3. Predicción del monto gastado en el periodo de estimación.

	Promedio	Total
Reg. Lineal	64.360	1.184.215.941
Log Normal	65.676	1.208.446.978
Poisson	73.479	1.352.016.372
MG real	64.328	1.183.628.436

Al ver la tabla 3 se observa que las estimaciones no distan mucho del valor real, por lo que éstas, a nivel agregado, son de buena calidad para los tres tipos de regresiones.

Luego, para observar con mayor precisión la capacidad predictiva y las medidas de ajuste de los modelos, se procede a calcular el error agregado de la estimación para cada una de las regresiones utilizadas, este error es ilustrado en la tabla 4 y se calcula como muestra la siguiente fórmula:

$$Error\%Agg = \frac{abs(\sum MG_{REAL} - \sum MG_{ESTIMADO})}{\sum MG_{REAL}} * 100$$

Tabla 4. Error agregado para el periodo de estimación.

	Error agregado
Reg. Lineal	0,05%
Log Normal	2,10%
Poisson	14,23%

Asimismo, la tabla 5 muestra el error cuadrático medio de la estimación para cada tipo de regresión, el cual se calcula según la fórmula:

$$MSE = \sqrt[2]{\sum (MG_{REAL} - MG_{ESTIMADO})^2}$$

Tabla 5. Error cuadrático medio para el periodo de estimación.

	MSE
Reg. Lineal	15.646.828
Log Normal	16.287.205
Poisson	20.214.210

Al observar las medidas de error se aprecia que las estimaciones a nivel agregado son bastante buenas, y que el modelo de regresión lineal es el que genera los datos de forma más precisa.

8.4.2 Validación de Estimación de Montos Gastados

A continuación se procede a validar el modelo usando los parámetros estimados para, a partir de los datos de las variable significativas, calcular el valor de los montos gastados para el mes siguiente al de los datos observados. Para esto se utilizan datos desde el mes de noviembre de 2007 hasta abril de 2008, es decir entre los meses 57 y 62, para las variables en el mes t, y desde el mes 56 al 61 para las variables del mes t-1, análogamente al periodo de estimación.

Los resultados de las predicciones para el periodo de validación se muestran en la tabla 6, donde aparece el promedio del monto gastado por cada cliente en cada mes y el total gastado en el periodo. Esto para cada tipo de regresión y también para el monto real gastado de modo de poder comparar.

Tabla 6. Predicción del monto gastado en el periodo de validación.

	Promedio	Total
Reg. Lineal	72.357	347.313.512
Log Normal	67.929	326.059.934
Poisson	81.070	389.134.366
MG real	80.767	387.681.356

Nuevamente se aprecia que a nivel agregado las estimaciones son bastante precisas, sin embargo, ahora es la regresión de Poisson la que se ajusta mejor a los datos.

En la tabla 7 y 8 se muestran los errores agregados y los de mínimos cuadrados para los distintos modelos de modo de nuevamente comparar entre ellos y apreciar la calidad de los modelos.

Tabla 7. Error agregado para el periodo de validación.

	Error agregado
Reg. Lineal	10,41%
Log Normal	15,89%
Poisson	0,37%

Tabla 8. Error cuadrático medio para el periodo de validación.

	MSE
Reg. Lineal	12.363.899
Log Normal	13.086.186
Poisson	14.467.893

Al analizar el error cuadrático medio se observa que nuevamente el modelo de regresión lineal es el mejor, lo cual era de esperarse porque este modelo consiste precisamente en minimizar este error. No obstante, al observar el error agregado, se aprecia que es el modelo de regresión de Poisson el de mejor calidad.

Al examinar los resultados obtenidos con la estimación de parámetros según los distintos tipos de modelos, es difícil decidir qué modelo es mejor para la estimación de éstos. Si bien, en el periodo de estimación, es el modelo de regresión lineal el que claramente es mejor, dado que tiene un mejor ajuste y tanto su error cuadrático medio como su error agregado son menores; en el periodo de validación, los parámetros estimados con el modelo de regresión de Poisson son los que presentan un menor error agregado. Además hay una teoría detrás de los modelos de Poisson que indican que son los más adecuados en caso de tener datos que son mayores o iguales a cero. Esto último no lo cumple el modelo de regresión lineal, el que en este caso fue reparado tomando como cero todos los valores estimados negativos.

Al mismo tiempo, se pudo comprobar que los modelos tienen una buena estimación a nivel agregado tanto de clientes como de los meses involucrados, sin

embargo, a nivel individual la estimación de los parámetros es muy mala, lo que se puede apreciar en la tabla 9, donde se presenta el MAPE promedio de los individuos para cada regresión, tanto en el periodo de estimación, como en el de validación. Esto era de esperarse dado que se está trabajando con clientes que tienen características muy distintas y los parámetros al ser iguales para todos no captan esta diferencia.

Tabla 9. MAPE promedio para el periodo de estimación y validación.

	MAPE Estimación	MAPE Validación
Reg. Lineal	901,17%	853,90%
Log Normal	1305,80%	1238,85%
Poisson	1422,03%	1390,91%

Por lo tanto, se hace necesario estimar el valor de los montos gastados pero a clientes que tengan comportamientos de compra similares, agrupándolos en un mismo segmento, para así apreciar si la estimación a nivel de segmento es de buena calidad y con esto se podría conocer mejor los tipos de clientes e inferir como se comportarán en el futuro.

Con respecto a las variables, la influencia de cada una de éstas sobre el monto gastado es explicada por Mauricio Durán en su memoria para obtener el título de ingeniero civil industrial (2008), aunque cabe mencionar que depende del signo que tenga el parámetro asociado a las variables el que define si la influencia es positiva o negativa. Por ejemplo, para la distribución de Poisson, la variable que indica el tiempo que ha pasado desde la última compra (Rcomp) tiene un parámetro asociado de valor negativo (-0,1195312), por lo tanto, mientras mayor sea el valor de la variable Rcomp, menor será el valor del monto gastado. A su vez, la variable de frecuencia tiene un parámetro asociado de valor positivo (0,2278407), por lo que a mayor frecuencia de compra, mayor será el monto gastado.

8.5 Segmentación de Clientes con Modelos de Clases Latentes

Para realizar la segmentación de clientes existen varios métodos realizados y probados, cada uno con sus ventajas y desventajas, por lo que la elección de uno de ellos depende tanto del problema como de la decisión personal del investigador. Dado lo que se busca en esta memoria; que es estimar el valor de los clientes a nivel de segmento y se necesitan obtener los parámetros correspondientes de cada uno de ellos, que sirven para predecir de la mejor manera lo que los clientes harán en el futuro; se ocupará el método de clases latentes explicado anteriormente (punto 7.5), que realiza simultáneamente la segmentación de los clientes y la estimación de los parámetros correspondientes a cada uno de los segmentos encontrados. Además al

realizar estos dos pasos de manera conjunta, se disminuye el error cometido al realizarlos por separado, ya que en ese caso se ocupan distintos criterios de optimización en cada uno y no necesariamente relacionados entre sí (Picón, 2004), lo que puede provocar que gente que está en el mismo segmento tenga una sensibilidad distinta frente a las variaciones de las variables.

Para esto, primero se harán unas consideraciones del modelo, en seguida se realizará el modelo para una clase latente (modelo de independencia), lo cual es equivalente a no realizar segmentación. Luego se seleccionará el número de clases o segmentos a utilizar, para seguir con la estimación de parámetros y el desarrollo del modelo. Finalmente se caracterizarán los distintos segmentos obtenidos.

8.5.1 Consideraciones del Modelo

Para realizar el modelo, fue utilizado el software Matlab, por lo que previamente hubo que familiarizarse con el programa, con el fin de poder programar satisfactoriamente los pasos a seguir por el método. La elección de este software se debe principalmente a lo flexible y potente del programa, además que no es “una caja negra”, como la mayoría de los softwares utilizados, ya que la persona que lo utiliza programa paso a paso lo que quiere que el software realice. En cambio, los otros softwares sólo necesitan datos iniciales y entregan resultados, por lo que no se sabe mayormente acerca de los pasos utilizados. Asimismo, utilizando Matlab el autor de esta memoria no sólo obtiene los resultados de ésta, sino que también obtiene el beneficio personal de conocer una herramienta muy potente que puede ser importante en el futuro laboral.

Las variables utilizadas en el modelo son las mencionadas en la parte de estimación del monto gastado en la sección 8.4.1, para la regresión de Poisson, las que serán utilizadas en el resto de la memoria dados los buenos resultados que se obtuvieron con ésta. Además que debido a los datos con los que se cuenta y lo que se quiere predecir, los modelos de Poisson son bastante adecuados, ya que lo que se quiere predecir sólo puede tomar valores positivos, y además los datos son bastante aleatorios y no tienen la forma típica de una normal, por lo que el modelo de regresión lineal y log normal no parecen apropiados para esto.

Cabe mencionar que al igual que en la parte de estimación del monto gastado, los datos corresponden a los mismos 23 meses de información. Aunque en este caso la muestra es de 760 personas. Esta cantidad resultó luego de hacer una nueva limpieza de los 800 datos, ya que existían personas muy fuera de lo común e impredecibles, dado que no tenían un patrón de comportamiento y eran muy difíciles de estimar. Estas personas gastaban en la mayoría de los meses un valor bajo (menos de 5 mil pesos en promedio), pero, aleatoriamente, en un mes llegaban a gastar millones de pesos (hasta 7 millones), y esto hacía que el modelo divergiera; por lo tanto, fueron sacadas de la muestra.

Dado que suponemos que la variable a predecir (monto gastado) sigue una distribución de Poisson de tasa λ , entonces la probabilidad de que el monto gastado tome un valor k es:

$$P(MG = k) = \frac{\lambda^k * e^{-\lambda}}{k!}$$

Donde $\lambda = e^{X*\beta}$, con X la matriz de variables por individuo y β el vector de los parámetros asociados a éstas.

Como los valores de k son muy altos en algunas oportunidades, sin pérdida de generalidad se divide la variable MG por 10.000^5 , para tener datos manejables y que ni el valor de “ $k!$ ” ni el de “ λ^k ” diverjan. Además, para evitar el mismo problema en el cálculo de λ , se divide la variable de monto promedio por 1000 , y la variable de los puntos acumulados por 100 , de manera que al aplicar exponencial el valor no diverja.

Cabe mencionar que si bien los datos son alterados, para efectos de la estimación, esto no debería influir en gran medida, ya que la variación en las variables debería ser compensada con el valor de los parámetros estimados.

Luego de estas consideraciones, se fija un criterio de convergencia para la función evaluada (función de máxima verosimilitud de los parámetros) y se realiza la operación partiendo de distintas combinaciones de parámetros iniciales, y así intentar buscar un máximo global. Si se obtienen distintos valores de la función, se elige la que tiene el mayor valor de loglikelihood (logaritmo de la función de máxima verosimilitud) y por lo tanto, el mejor BIC. Por ejemplo, para 6 clases se partió de un valor aleatorio de los 125 parámetros (20 por cada grupo y 5 por las probabilidades de pertenencia a priori), multiplicado por 0.002^6 y también se partió de los valores obtenidos por 5 clases, es decir, 104 parámetros, y colocando los otros 21 valores igual a 0 . Con estos dos puntos de partida se obtuvo distintos puntos de convergencia, el primero tuvo un valor de loglikelihood de $-456111,7$ con un BIC de 92445 y el segundo un valor de loglikelihood de $-44699,6$ con un BIC de 90620 , por lo que se elige el segundo caso.

El criterio de convergencia utilizado, es que la función no varíe más de un 0.01% entre una iteración y otra. No se puso un máximo de iteraciones como criterio de convergencia, ya que dada la multidimensionalidad de la función a medida que aumentaba el número de clases y por ende el número de parámetros, ésta debía iterar cada vez una mayor cantidad de veces para encontrar un óptimo.

⁵ Se elige este valor por ser adecuado y no exagerado, si fuera menor, no se podría trabajar con valores altos del monto gastado, y si fuera mayor, se perdería mucho el efecto de los montos gastados pequeños, para pronosticar luego con los parámetros estimados, hay que multiplicar por este valor los resultados.

⁶ El valor es pequeño para evitar divergencia de λ .

Para encontrar el óptimo de la función, se utiliza la función `fminsearch` de Matlab, la cual busca el mínimo de una función basado en las evaluaciones de ésta. Existen otras funciones en el software, como por ejemplo, `fminunc`, pero se utiliza la antes mencionada debido a que las otras divergían antes de encontrar el óptimo. Como `fminsearch` busca el mínimo de la función objetivo, y en este caso lo que se busca es un máximo, simplemente la función objetivo se ingresa con un signo negativo, es decir, el inverso aditivo de la función.

El proceso realizado anteriormente debía ser repetido para cada número de clases con las que se quería trabajar, para luego comparar los ajustes de los modelos, mediante los valores del criterio de información de Akaike (AIC) y el criterio de información Bayesiano (BIC). Además del indicador que muestra el grado de asociación explicado por la variable latente ($%L^2$) y el radio de verosimilitud (ρ). De esta manera se puede elegir el número de clases que presenta un mejor ajuste de los datos.

8.5.2 Modelo Homogéneo, Una Clase Latente

Al considerar sólo una clase latente, se tienen los mismos parámetros para toda la población, por lo que es similar a lo realizado en la sección 8.4. Sin embargo los resultados serán distintos dadas las consideraciones hechas en este modelo. No obstante, las variables utilizadas seguirán siendo las mismas ocupadas en el modelo anterior para seguir una misma línea.

Los parámetros estimados según el modelo de clases latentes, para el caso de una sola clase latente se muestran en la tabla 10.

Tabla 10. Parámetros estimados según modelo de independencia.

Variables	Parámetros	Variables	Parámetros
Cte	1,1383	d2	0,1446
PA	0,0003	d3	0,1626
Dist	0,0859	d4	0,0847
Rcomp	-0,0222	d5	0,0861
Fcomp	0,0532	d6	0,0577
Mcomp	0,0057	d7	0,0594
Sex	-0,0108	d8	0,0140
A_cta	0,0017	d10	0,1125
Edad	-0,0018	d11	0,0997
NC	0,0458	d12	0,4445

Luego, en la tabla 11, se presentan los resultados del pronóstico, dados los parámetros utilizados en el modelo de independencia. En ella se puede observar que en el total los valores de los montos reales y los montos pronosticados son bastante similares y por lo tanto a nivel agregado el pronóstico es bastante bueno. De hecho, el error agregado de éste es de 0.96%, lo cual indica una muy buena estimación.

Tabla 11. Predicción del monto gastado según modelo de independencia.

	Promedio	Total
MGreal	65.527	298.804.209
MGpro	64.899	295.939.686

Finalmente, cabe mencionar que el error promedio a nivel individual (MAPE promedio), es muy alto en este caso (1066,65%), debido a que se considera homogeneidad entre los clientes. Por lo tanto, hay una gran motivación a incluir la heterogeneidad en los comportamientos de los clientes y por lo tanto separarlos en grupos o segmentos, que es lo que se hace a continuación.

8.5.3 Selección del Número de Clases

Uno de los primeros pasos al trabajar con clases latentes, es determinar el número de clases a utilizar, para esto es necesario ejecutar el modelo con distintos números de clases y luego de encontrar el valor óptimo, utilizar los criterios AIC y BIC para determinar el número que representa un mejor ajuste del modelo.

Para efectos prácticos del uso de los resultados obtenidos, se supone un número de clases manejable, por lo que se realiza el modelo de clases latentes para un máximo de 10 clases, partiendo del modelo de independencia que corresponde al de una clase.

La tabla 12 muestra los valores obtenidos para cada clase, donde aparece el número de parámetros estimados en cada una de ellas y el valor máximo del logaritmo de la función de máxima verosimilitud, además de los criterios antes mencionados para medir la calidad del ajuste (AIC, BIC, $\%L^2$ y ρ), para poder comparar entre las distintas clases y elegir el número de ellas que tiene el mejor ajuste.

Como se puede apreciar, el valor de la máxima loglikelihood fue aumentando a medida que aumentaba el número de clases, lo cual era de esperarse ya que mientras más parámetros están explicando los datos, mayor será el valor de la función de likelihood, dado que se cuenta con más información.

Tabla 12. Comparación entre los distintos números de clase utilizados.

Nºclases	Nº parámetros	Loglikemax	AIC	BIC	%L ²	ρ
1	20	-94167	188375	188530	-	0,5174
2	41	-56864	113810	114129	0,6354	0,7086
3	62	-48295	96715	97197	0,7370	0,7525
4	83	-45667	91500	92145	0,7648	0,7660
5	104	-45076	90360	91168	0,7709	0,7690
6	125	-44700	89649	90620	0,7747	0,7709
7	146	-44531	89353	90487	0,7764	0,7718
8	167	-44420	89174	90471	0,7775	0,7724
9	188	-44367	89110	90570	0,7780	0,7726
10	209	-44325	89068	90692	0,7784	0,7728

Además se observa lo mencionado anteriormente con respecto al número de parámetros, los cuales van aumentando considerablemente a medida que aumenta el número de clases, y por ende, la multidimensionalidad de la función de máxima verosimilitud también aumenta.

Con respecto a los criterios de información utilizados (AIC y BIC), se observa que el Criterio de Información de Akaike (AIC) va disminuyendo su valor a medida que aumenta el número de clases, por lo que sería el modelo de 10 clases el que presentaría un mayor ajuste según este criterio. Sin embargo, hay que tener en cuenta que sólo se realiza el modelo hasta este número de clases, por lo que el criterio no sería una variable de decisión clara. Ahora bien, el criterio de información bayesiano (BIC) también va disminuyendo a medida que aumenta el número de clases, pero luego de 8 clases comienza a aumentar. Por lo tanto, según este último criterio de información, el modelo que presenta el mayor ajuste es el de 8 clases, ya que posee el menor valor del BIC y mientras menor sea este valor, más parsimonioso (con menos parámetros) es el modelo.

Como se tienen distintos resultados según los criterios de información, hay que elegir cuál de ellos utilizar y si trabajar con un modelo de 8 clases o con uno de 10 en este caso. En general, se prefiere BIC sobre AIC, ya que entre otras cosas tiende a seleccionar modelos con menos parámetros (BIC penaliza mucho más que AIC). Además, según von Storch y Zweirs (1999), el BIC normalmente identifica el orden del proceso en forma más precisa que AIC, mejorando con el incremento de la muestra. Asimismo tiene más bajo sesgo y variabilidad. Finalmente cabe mencionar que Hannan (1980) demuestra que el BIC es un criterio consistente para determinar el orden.

Por lo tanto el modelo de 8 clases sería el elegido. Además se puede notar que a partir de 8 clases, el valor del AIC comienza a variar en una cantidad menor y además según este criterio hay incertidumbre si realmente 10 clases es mejor, ya que podría seguir disminuyendo el valor, por lo tanto no será considerado.

Asimismo al observar el grado de asociación explicado por la variable latente (%L2), se aprecia que va aumentando poco a poco y cada vez está más cerca de 1, por lo que según este criterio, mientras más clases, mayor es el ajuste, sin embargo, se puede notar que a partir de 5 clases, el valor comenzó a aumentar en menor medida, y luego de las 8 clases varía menos de 0,0006, por lo que este número de clases resulta con un gran ajuste y por lo tanto, es adecuado para trabajar.

Finalmente al observar en la última columna el radio de likelihood (ρ) se observa que con esta cantidad de clases el valor de este índice es relativamente alto, por lo que los parámetros estimados predicen con una alta probabilidad los datos observados.

En consecuencia, dado todo lo anterior, el número de clases elegido, y con lo que se trabajará en el modelo de clases latentes, es 8.

8.5.4 Estimación de Parámetros y Desarrollo del Modelo

Luego de elegir el número de clases a utilizar, se estiman los parámetros para cada una de ellas, además de las probabilidades a priori de pertenencia a cada grupo, las cuales son el resultado de la maximización de la función de máxima verosimilitud entre ellos. La tabla 13 muestra los valores de los parámetros estimados para cada una de las variables utilizadas, incluyendo la constante. Estos parámetros son significativos estadísticamente con un intervalo de confianza del 95%, según un análisis t-student utilizando el método de re-muestreo llamado “jackknife” (Smith, 2006).

Los parámetros muestran las sensibilidades de las variables observadas (independientes) sobre la variable dependiente, que en este caso es el monto gastado. Esto para cada uno de los grupos encontrados, ya que cada segmento tiene su propia sensibilidad frente a las variables observadas. Si el valor es positivo significa que mientras mayor sea el valor de la variable observada, mayor será el valor de la variable dependiente; y si es negativo, mientras mayor sea el valor de la variable independiente, menor será el valor del monto gastado (variable dependiente). Es decir, el cliente gastará más si aumenta una variable asociada a un parámetro positivo, y gastará menos si aumenta el valor de una variable asociada a un parámetro negativo.

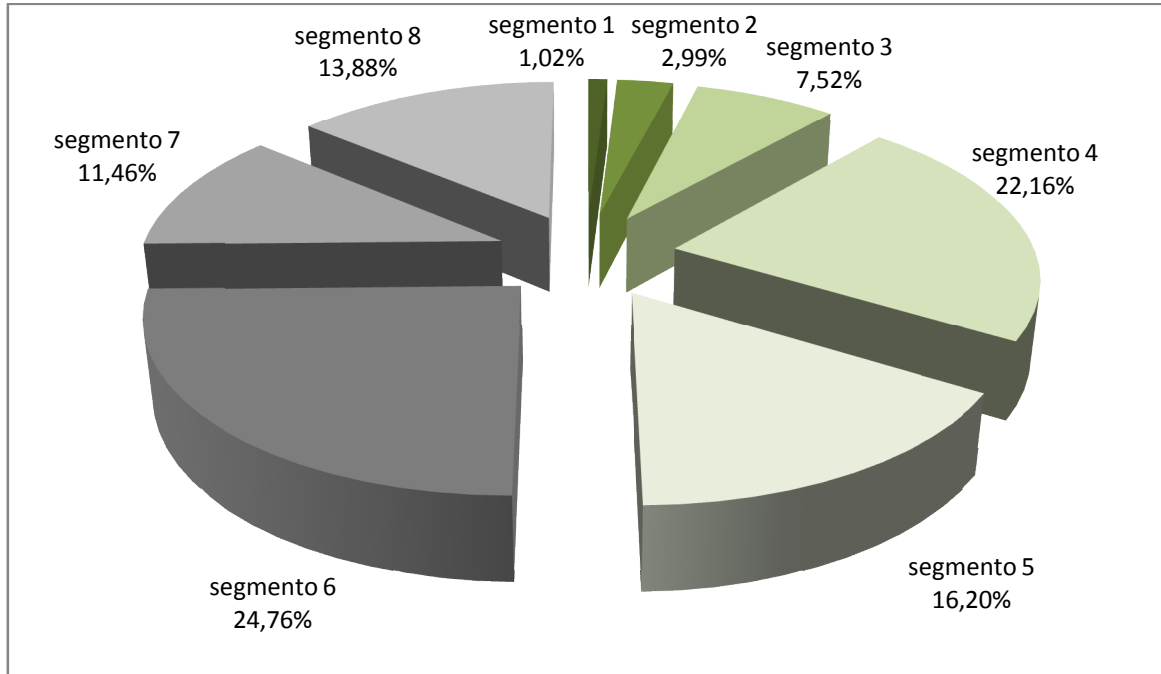
Además hay que mencionar que la intensidad con que la variable observada influye sobre la variable dependiente, también es diferente en cada grupo, y finalmente es la combinación de estos parámetros estimados los que definen las características de cada uno de ellos.

Tabla 13. Parámetros estimados para cada grupo según modelo de clase latente.

	g1	g2	g3	g4	g5	g6	g7	g8
Cte	4,181	3,225	2,618	-6,037	1,881	0,410	0,008	-2,058
PA	0,001	0,001	0,0004	0,0001	-0,0006	-0,0003	0,0037	0,005
Dist	0,001	-0,012	-0,004	0,556	0,068	0,149	0,091	-0,612
Rcomp	-0,00000005	-0,030	-11,141	-10,787	0,008	-0,504	0,012	-12,880
Fcomp	-0,213	-0,204	-0,355	4,635	-0,389	0,002	1,701	1,011
Mcomp	0,002	0,005	0,006	0,004	0,006	0,008	0,011	0,003
d2	0,308	0,194	0,093	0,165	0,093	0,107	0,184	-0,948
d3	0,221	0,113	0,009	-0,018	0,080	0,076	0,077	0,671
d4	0,141	0,101	0,010	0,357	0,003	0,036	-0,278	-0,001
d5	0,252	0,205	0,1289	0,1701	0,1265	0,1674	-0,0002	0,301
d6	0,089	0,039	0,102	0,761	0,000	0,146	0,158	0,218
d7	0,121	-0,056	-0,0004	-0,0855	-0,1032	0,0745	0,4510	0,837
d8	0,075	0,010	-0,0788	0,2961	0,0006	-0,0004	0,0781	0,000
d10	0,237	0,172	0,136	0,337	0,092	0,108	-0,019	-0,715
d11	-0,043	-0,032	0,023	-0,00003	0,118	0,169	-0,001	0,292
d12	0,213	0,251	0,316	0,157	0,310	0,369	0,967	0,540
Sex	-0,127	-0,044	-0,035	-0,409	0,014	-0,038	-0,161	-0,340
A_cta	0,006	0,006	0,007	0,015	0,001	-0,009	0,032	-0,023
Edad	-0,005	-0,004	-0,002	0,008	0,0004	0,008	-0,034	-0,041
NC	0,006	0,009	0,011	0,264	0,026	0,024	0,046	0,236

En la figura 5, se muestran las proporciones de cada segmento en base a las probabilidades de pertenencia a priori de cada grupo, que representan las probabilidades de pertenecer a un grupo antes de incorporar la información de compras a mes de los individuos. Donde se puede apreciar que los individuos tendrían una menor probabilidad de pertenecer a los grupos 1 y 2 con una probabilidad a priori de 0,0102 y 0,0299 respectivamente. Y pertenecerían más probablemente con una probabilidad a priori de 0,2216 y 0,2476 a los grupos 4 y 6 respectivamente. Es decir, un 25% de la población correspondería al segmento 6, y sólo un 1% de ésta correspondería al segmento 1.

Figura 5. Proporción de individuos de la población de cada segmento a priori.



Luego, en base a la proporción a priori encontrada de cada segmento, es necesario calcular la probabilidad de pertenencia a posteriori, es decir, luego de conocer la información de cada cliente en el transcurso de los 23 meses en que se realiza la estimación. De esta manera, dado el comportamiento de compra de los clientes en el transcurso del tiempo, las probabilidades de pertenecer a un segmento u otro, dependerán de las variaciones en las variables independientes y del monto gastado por el cliente mes a mes, además de los parámetros estimados para cada grupo.

La probabilidad de pertenecer al segmento i , luego de los 23 meses de información, está dada por:

$$P_i = \frac{\prod_{t=1}^{23} \pi_i * Pit}{\sum_{i=1}^8 \prod_{t=1}^{23} \pi_i * Pit}$$

Donde π_i corresponde a la probabilidad a priori mostrada anteriormente y Pit es la probabilidad de pertenecer al segmento i en el mes t .

En la tabla 14 se muestran las probabilidades a posteriori obtenidas para cada segmento, tomando en cuenta los 23 meses de información. Hay que considerar que un cliente pertenece a cada grupo con alguna probabilidad.

Tabla 14. Probabilidades de pertenencia a posteriori.

Grupo	1	2	3	4	5	6	7	8
Pbb	5,0971E-164	1,000006	17,96	20,41	313,47	244,63	156,96	5,56
%	0,00%	0,13%	2,36%	2,69%	41,25%	32,19%	20,65%	0,73%

La fila “Pbb” muestra la suma de las probabilidades de pertenencia al grupo, de todos los individuos de la muestra, y la fila “%” es el porcentaje de personas de la muestra que pertenecería al grupo de la columna correspondiente, por lo que se puede apreciar que los grupos 1, 2 y 8, no representan una gran probabilidad de pertenencia, ya que no alcanza ni el 1% de los individuos.

A pesar de que los individuos pertenecen con cierta probabilidad a cada grupo, se asignaron al grupo que pertenecían con mayor probabilidad, para tener un mejor manejo e interpretación de los resultados. La tabla 15 muestra la cantidad de personas asignadas a cada grupo y por lo tanto el porcentaje de la población correspondiente a cada grupo luego de esta asignación.

Se puede apreciar que la cantidad de personas pertenecientes a cada grupo es casi la misma al asignar de acuerdo a la suma de las probabilidades y al asignar de acuerdo a la mayor probabilidad. Esto se explica debido a que los individuos pertenecen a un grupo con una probabilidad muy alta, en muchos casos casi 1, y a los otros grupos con una probabilidad mucho más pequeña (muy cercana a 0).

Tabla 15. Asignación de individuos según máxima probabilidad de pertenencia a posteriori.

Grupo	1	2	3	4	5	6	7	8
Personas	0	1	17	21	313	246	157	5
%	0,00%	0,13%	2,24%	2,76%	41,18%	32,37%	20,66%	0,66%

Al asignar a los individuos a la probabilidad más alta de pertenencia se comete un error, y la proporción esperada de los errores de clasificación, corresponde a 0,01285 según la fórmula vista en la sección 7.7 de esta memoria, lo cual corresponde a un error muy pequeño y por lo tanto, el porcentaje de buena clasificación es 98,72%. En consecuencia, se observa una clasificación muy buena de los individuos a cada segmento. Lo mismo se puede apreciar en la tabla 16, al calcular el indicador de Goodman & Kruskal (λ de la sección 7.7 también) para la calidad de asignación a una clase concreta, donde se observa que ningún grupo o clase tiene una calidad de asignación menor al 97%, lo cual demuestra la buena calidad de las clasificaciones.

Tabla 16. Calidad de asignación a cada clase (8 segmentos).

Segmento	Error (E)	Lamda
1	-	-
2	0,00000	100,00%
3	0,00006	99,99%
4	0,00197	99,74%
5	0,01001	98,67%
6	0,01691	97,75%
7	0,01543	97,95%
8	0,00087	99,88%

El segmento 1 no tiene error de clasificación, ya que ningún individuo tenía una mayor probabilidad de pertenecer a ese segmento frente a los otros.

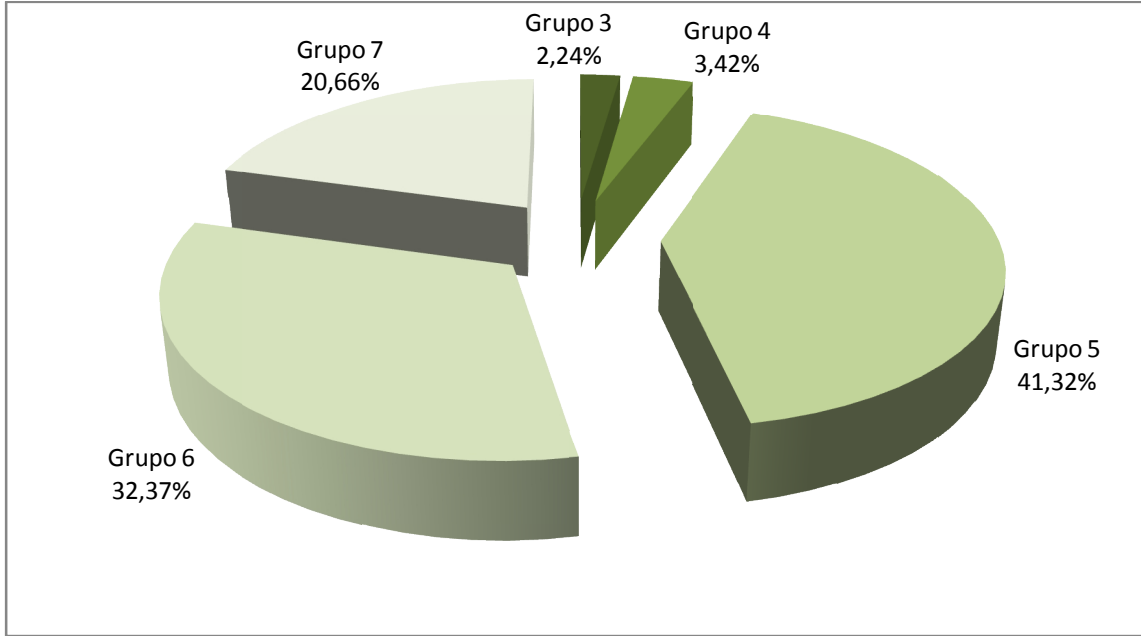
Para efectos prácticos, dado que hay tres grupos que prácticamente no tienen personas, el 1, el 2 y el 8, éstos no fueron considerados para el análisis posterior de los segmentos, y las personas que pertenecían a esos grupos con alguna probabilidad, fueron asignadas al segundo grupo que pertenecían con mayor probabilidad. Es así como la única persona que pertenecía al grupo dos, fue asignada al grupo 5, y las 5 personas que pertenecían al grupo 8, fueron asignadas al grupo 4, ya que cada una de ellas pertenecía con mayor probabilidad a ese grupo que a los otros.

La figura 6, muestra las proporciones de la población correspondientes a cada segmento, luego de tomar en cuenta la información correspondiente a los individuos en el tiempo de estimación. Esto considerando a los 5 grupos de interés mencionados anteriormente. En esta figura se percibe que los grupos 3 y 4 son los más pequeños contando con menos personas (2,24% y 3,42% de la población respectivamente), y el grupo 5 es el más grande con un 41,32% de la población. Los grupos 6 y 7 también tienen una alta proporción en la población, con un 32,37% y un 20,66% respectivamente.

Con esto concluye la segmentación de clientes y se obtienen finalmente 5 grupos de clientes, los cuales tienen comportamiento de compra diferente y que serán analizados en la siguiente sección.

Análogamente al caso de 8 segmentos, se calculan los errores de asignación, y así comprobar que con los 5 segmentos el error de asignación sigue siendo pequeño, y por lo tanto el porcentaje de buena asignación es alto.

Figura 6. Proporción de individuos de la población de cada segmento a posteriori.



La proporción esperada de los errores de clasificación, corresponde a 0,01286, un cambio de 0,00001 con respecto al anterior, por lo cual, claramente el error sigue siendo muy pequeño y el porcentaje de buena clasificación muy alto. Lo mismo se puede apreciar en la tabla 17, al calcular la calidad de asignación a una clase concreta.

En ella se puede apreciar que el porcentaje de buena asignación es superior al 97% en cada una de las clases, por lo que la asignación de los individuos a cada segmento es realmente de buena calidad.

Tabla 17. Calidad de asignación a cada clase (5 segmentos).

Segmento	Error (E)	Lamda
3	0,00006	99,99%
4	0,00221	99,71%
5	0,00998	98,67%
6	0,01691	97,75%
7	0,01543	97,95%

8.6 Caracterización de Segmentos

8.6.1 Análisis de Sensibilidad de Variables sobre Monto Gastado por Segmento

Para caracterizar el comportamiento de compra de los distintos grupos es necesario observar el valor de los parámetros estimados para cada uno de ellos, ya que como se dijo anteriormente, el método usado (Clase latente) trabaja con las sensibilidades de las variables independientes sobre la variable dependiente, por lo que son éstas sensibilidades (parámetros), las que caracterizan a los distintos grupos.

A continuación se presenta un análisis a cada grupo de personas, mostrando las sensibilidades de cada variable sobre el monto gastado. Si el valor del parámetro es positivo, la influencia de la variable asociada es positiva, análogamente si el valor del parámetro es negativo, la influencia de la variable asociada es negativa.

Grupo 3

Se comenzará con las variables que son observadas en el mes anterior al mes que se quiere predecir, éstas se refieren a los puntos que acumula la persona y a las variables RFM de la compra de productos, además del número de compras realizadas. En la tabla 18 se ilustra lo antes descrito.

Tabla 18. Sensibilidad de variables observadas en el mes anterior para el grupo 3.

PA	Dist	Rcomp	Fcomp	Mcomp	NC
0,000359	-0,0038	-11,1413	-0,35472	0,005629	0,01057

Se puede observar que los puntos acumulados (PA) influyen positivamente en este grupo, por lo que mientras más puntos acumulados tenga un individuo en un mes, mayor será el monto gastado en el mes siguiente.

La variable Dist, que indica la distancia al canje más cercano, tiene una influencia negativa, sin embargo, esto no tiene mucho sentido, ya que es ilógico que los individuos compren menos al mes siguiente sólo por el hecho de estar cerca de un canje superior, y compren más sólo por el hecho de estar en un canje inferior. Esto se puede explicar tomando en cuenta que probablemente las personas pertenecientes a este grupo en realidad no prestan mayor atención a la distancia que tienen a un canje, sino que son las otras variables las que hacen inferir el próximo comportamiento.

La tercera variable (Rcomp) muestra como las personas de este grupo tienden a gastar menos a medida que pasa el tiempo sin comprar, es decir, un individuo de este grupo probablemente gastará menos en el próximo mes si lleva 5 meses sin comprar,

que si lleva 3 meses sin hacerlo, obviamente considerando “ceteris paribus” para las demás variables explicativas.

También se puede observar gracias a la variable de frecuencia (Fcomp), que los individuos de este segmento se controlan en sus gastos, ya que a medida que su frecuencia de compra aumenta, es decir, están yendo muy seguido a comprar, es probable que vayan gastando menos. Por lo tanto, un individuo que ha ido más veces a comprar, probablemente gastará menos que uno que lo ha hecho con menor frecuencia.

Con respecto al monto promedio, este influye positivamente en el monto gastado, lo que significa que mientras mayor sea el monto promedio en un mes, es más probable que al mes siguiente el individuo gaste más que si hubiera tenido un monto promedio menor. Cabe mencionar que esta variable afecta a todos los grupos positivamente, aunque no exactamente con el mismo valor, por lo que esta variable no marca mucha diferencia entre un grupo y otro, de manera que no es relevante mencionarla en cada uno de ellos.

La variable NC (número de compras), también influye positivamente en todos los grupos, por lo cual, mientras más compras un individuo haya realizado en un mes, es más probable que en el mes siguiente el individuo gaste más dinero que si hubiera realizado menos compras.

A continuación, en la tabla 19, se observan las dummies de mes, en las cuales se dejó afuera el mes de enero para evitar la colinealidad entre estas variables. Como el mes de septiembre resultó ser no significativo en el análisis de monto gastado realizado en un comienzo, éste también será omitido para obtener los resultados. En consecuencia, las variables son consideradas con respecto a estos dos meses, por lo tanto, cuando el individuo se encuentre en el mes de enero o de septiembre, el monto gastado no se verá afectado ni positiva, ni negativamente por las variables de estacionalidad (mes).

Estas variables son consideradas en el mismo mes que se quiere predecir, por lo que si se quiere saber cuánto gastará el individuo en noviembre por ejemplo, la dummy d11 tomará el valor uno, y las demás serán 0.

Tabla 19. Sensibilidad de Dummies de estacionalidad para el grupo 3.

d2	d3	d4	d5	d6	d7	d8	d10	d11	d12
0,0933	0,0091	0,0104	0,1289	0,1024	-0,0004	-0,0787	0,1356	0,0231	0,3163

Se observa que este grupo disminuye los gastos en los meses de julio y agosto, y el mes de diciembre es el que lo afecta en mayor medida, aumentando el monto a

gastar en ese mes. En los demás meses los individuos también tienden a gastar más que en los meses de control (enero y septiembre).

En seguida, en la tabla 20 se muestran las influencias de las variables socio-demográficas, correspondientes al sexo, la edad, y los años que tiene la cuenta un individuo. Estas variables al igual que las dummies de estacionalidad, son medidas en el mismo mes que se quiere predecir, aunque cabe destacar, que el sexo siempre va a ser igual, y para efectos de pronóstico, tanto de un mes como de un año, la edad y los años que posee la cuenta pueden variar máximo en 1.

Se observa que la variable de sexo, influye negativamente a los individuos de este grupo, por lo que si la compra la hace una mujer, es más probable que gaste más que un hombre en las mismas condiciones.

Tabla 20. Sensibilidad de variables socio-demográficas para el grupo 3.

Sex	A_cta	Edad
-0,03541	0,006652	-0,00185

Las variables de edad y años que posee la cuenta influyen de distinta manera en al monto a gastar de los individuos, lo cual puede parecer poco intuitivo, ya que mientras más años tiene una persona, lógicamente más años poseerá la cuenta. Sin embargo, analizando sólo estas dos variables y dejando las demás igual, por ejemplo si dos personas tienen la misma edad, el que tenga hace más tiempo la tarjeta, será el que gaste más. De la misma forma, si dos personas tienen hace el mismo tiempo la tarjeta, el más joven será el que gaste más.

En resumen, se puede decir que este grupo se caracteriza principalmente por ser controlado en sus gastos (Fcomp), por no tomar tanto en cuenta la posibilidad de canjear (Dist), por gastar menos en julio y agosto y gastar un poco más en diciembre y porque mientras más joven y más años tenga la cuenta, gasta más.

El análisis detallado variable por variable de los otros grupos es análogo al grupo descrito anteriormente, y se encuentra en el anexo 3, por lo que a continuación sólo se mostrará el resumen de cada grupo.

Grupo 4

Este grupo se caracteriza principalmente porque le influye más que a todos la distancia al canje más cercano, por lo que aumentan su gasto si están cerca del canje superior y se conforman si están cerca del inferior (Dist).

También son más consumistas, y mientras mayor sea su frecuencia de compra (Fcomp), más gastarán en el mes siguiente. Sin embargo, si no han ido a comprar en un tiempo, gastarán menos que si hubieran ido a comprar más seguido (Rcomp).

En este grupo los individuos aproximadamente cada 4 meses gastan menos, siendo los meses de julio y octubre donde gasta un poco más.

Es el grupo más sensible a la variable de sexo, siendo las mujeres las que gastan más que los hombres.

La gente de más edad y con más años de cuenta, son los que tienden a gastar más.

Grupo 5

Este grupo también controla sus gastos, ya que mientras más seguido ha ido a comprar, es más probable que gaste menos. Sin embargo, si llevan mucho tiempo sin comprar, es probable que gasten más que si lo han hecho todos los meses. Es como si tuvieran un presupuesto para comprar, si van muy seguido compran menos, pero si no van en un tiempo compran más.

No le dan tanta importancia a los puntos acumulados, sin embargo si están cerca del canje superior gastan un poco más, y un poco menos si están cerca del inferior.

En general gastan más que en los meses de control, excepto en junio y julio donde gastan menos. Hay un leve aumento del gasto en diciembre.

Es el único grupo donde los hombres gastan más que las mujeres.

Al igual que el grupo 4, mientras más años tiene el individuo, y mientras más años tiene la cuenta, es más probable que gaste más que uno más joven o con menos años que tenga la cuenta.

Grupo 6

Este grupo se parece un poco al grupo 4, ya que mientras más han ido a comprar, es probable que gasten más al mes siguiente, pero si han dejado de comprar en un tiempo es probable que gasten menos que si hubieran comprado el mes anterior. Sin embargo, las influencias de las variables descritas (Fcomp y Rcomp) es mucho menor que en ese grupo.

A pesar de que en este grupo la distancia al canje tiene una influencia positiva sobre el monto gastado, los individuos no le dan tanta importancia a los puntos acumulados en el mes anterior.

Además este grupo gasta menos sólo en agosto, y en los otros meses gasta más que en los de control, siendo el mes de diciembre cuando los individuos de este grupo gastan más.

También este grupo es el único donde las personas que tiene menos años la cuenta gastan más que los que la tienen por más tiempo. A pesar de lo anterior, si las personas tienen mayor edad tienden a gastar más.

Grupo 7

Este es el único grupo donde todas las variables del mes anterior (PA, Dist, R, F) influyen positivamente sobre el monto gastado al mes siguiente, es decir, mientras más puntos tienen acumulados, mientras más cerca están del canje superior, mientras más tiempo ha pasado desde la última compra y mientras mayor sea la frecuencia de compra, mayor será el monto gastado al mes siguiente.

Estos individuos siempre quieren gastar y no se controlan, no pueden estar sin ir a comprar, ya que mientras más tiempo pasan sin comprar, más gastarán, mientras más van a comprar, más gastarán.

Al mirar la estacionalidad se observa que a pesar de lo dicho anteriormente, en este grupo se puede decir que ahorran un poco antes de los gastos más grandes, ya que gastan más a mediados de mes (julio) y a fin de mes (diciembre), posiblemente por navidad y vacaciones de invierno. Sin embargo, en los meses anteriores a éstos, los individuos tienden a gastar menos. Entonces se tiene que en abril y mayo gastan menos para gastar más en julio, y en octubre y noviembre gastan menos para gastar mucho más en diciembre.

En este grupo la gente joven tiende a gastar más, no obstante, mientras más años tienen la cuenta los individuos, gastan más.

8.6.2 Análisis de Variables por Segmento

A pesar de que el modelo de clases latentes trabaja con las sensibilidades de las variables independientes sobre la dependiente, se realiza un análisis de las variables en cada grupo, para complementar el análisis anterior y permitir caracterizar aún mejor a cada uno de ellos. Para esto fueron asignados los grupos de acuerdo a la mayor probabilidad de pertenencia de los individuos a éste.

Además en este análisis se incluye una variable que no fue considerada en el modelo por falta de información, pero cuando ya se tenían los resultados del modelo se pudo acceder a ella, por lo que será mencionada con el fin de conocer aún mejor a los individuos de cada grupo. Esta variable corresponde al avance que las personas piden con la tarjeta.

Como los grupos están separados por sensibilidades y no por las variables en sí, sólo se puede hablar en promedio. Por ejemplo, si nos fijamos en la edad, podemos decir que en un grupo en promedio (o en general) hay más gente entre los 30 y 40 años, a pesar de que en realidad hay gente de todas las edades en cada grupo. Lo mismo ocurre para todas las variables, sin embargo, se realiza este análisis con el fin de tener una caracterización más completa de los segmentos e incorporar la variable de avances ignorada hasta este momento. Además se incluyen los valores de la variable dependiente, lo que permite identificar mejor los tipos de clientes en cuanto a esta variable.

Grupo 3

En su mayoría son personas entre 40 y 50 años, con pocos años de poseer la cuenta (de 3 a 5 años), y son más mujeres que hombres, con un 65% de mujeres en el grupo.

Los individuos compran bastante seguido, la mayoría lo hace todos los meses, obteniendo una frecuencia de compra promedio de 98% en el periodo, considerando que si una persona va todos los meses a comprar tiene una frecuencia de 100%, y el promedio es sacado entre todas las personas del grupo.

Con respecto al monto que gastan, los individuos de este grupo gastan en promedio más de 150 mil pesos mensuales, y no gastan menos de 80 mil mensual en promedio.

En este grupo las personas piden alrededor de 45 mil pesos mensuales en avances, considerando todos los individuos de este grupo, sin embargo, los que piden son aproximadamente un 48% de las personas, por lo que éstos pedirían en promedio, cerca de 96 mil pesos mensuales.

Cabe mencionar que son promedios mensuales, pero en realidad las personas piden sólo en algunos meses.

Gracias a los montos gastados y a los avances, la gente llega a acumular en promedio 18 mil puntos (promedio de los máximos acumulados), canjeando principalmente en el primer nivel de canje, aunque también lo hacen en el segundo y en el tercero con frecuencia.

Los canjes son realizados principalmente en los meses de diciembre, enero, abril, mayo y julio.

Además este grupo realiza en general entre 2 y 5 compras mensuales.

Grupo 4

Hay de todas las edades entre 30 y 50 años principalmente, teniendo la cuenta por varios años, alrededor de 7 en general. Son más hombres que mujeres, con un 65% aproximadamente de hombres.

No hay característica clara de frecuencia, hay personas que van todos los meses, otros que van temporalmente y otros que cada vez han ido menos a comprar, con lo que llegan a una frecuencia promedio de 84% en el periodo de estimación.

En promedio gastan alrededor de 10 mil pesos mensuales, aunque hay algunos de este grupo que gastan mucho para épocas de fiesta, como navidad (alrededor de 100 mil pesos), pero son muy pocos.

El promedio de avances pedidos por este grupo es de 15 mil mensual, sin embargo considerando sólo los que piden, que son alrededor de un 54%, éstos solicitan en promedio 27 mil pesos mensuales.

Los individuos de este grupo llegan a acumular máximo en promedio 5500 puntos, con lo que canjean casi sólo en el nivel 1, y sólo unos pocos a veces llegan a canjear en el nivel 2 o 3. Canjeando principalmente en diciembre, enero y septiembre.

Generalmente compran una sola vez al mes aunque hay algunos que lo hacen 2 veces.

Grupo 5

En este grupo hay gran diversidad de gente, hay de todas las edades, y se pueden encontrar personas entre los 30 y 60 años no existiendo una gran tendencia de edad. Están bastante iguales en cuanto al sexo, con un 46% y un 54% de mujeres y hombres respectivamente. La cuenta la tienen principalmente 7 años, aunque hay varias personas que la tienen hace más tiempo.

Hay gente que compra siempre, otra que no compra casi nunca y otra que compra esporádicamente, por lo que no es clara la frecuencia de compra, sin embargo, en promedio de todos sería alrededor de un 90%.

Los individuos de este grupo gastan en promedio entre 70 y 80 mil pesos mensuales, aunque son muy volátiles, ya que en algunos meses gastan mucho (400 mil pesos es el promedio de los máximos gastados en algún mes), y en otros muy poco.

Los avances son en promedio 23 mil pesos, tomando en cuenta todos los individuos del grupo. Al considerar el promedio sólo del 49% que pide avances, éstos alcanzan una media de 47 mil pesos mensuales.

En este grupo llegan a acumular en promedio, alrededor de 11 mil puntos como máximo, canjeando principalmente en el nivel 1, y alcanzando también el nivel 2 y 3 en varias oportunidades y unos pocos el nivel 4. Canjeando la mayoría en los meses de enero, junio y octubre, a pesar de que hay canjeadores en todos los meses.

La mayoría realiza más de 3 compras mensuales, sin embargo también hay muchos meses en que realizan una compra.

Grupo 6

Al igual que el grupo anterior, hay gente de todas las edades, concentrando la mayoría entre los 35 y 55 años principalmente. Son casi un 50% de hombres y un 50% de mujeres, donde la mayoría tiene pocos años la cuenta (3 años), sin embargo, al igual que en las edades, hay gente que tiene 5, 6, 7 y hasta más de 15 años la cuenta.

Existen en este grupo individuos que compran todos los meses, pero hay otros que compran muy ocasionalmente y otros que compran cada cierta cantidad de tiempo. Como promedio se encuentra una frecuencia de compra de aproximadamente 88%.

El monto gastado en promedio, mensualmente, varía entre 35 y 40 mil pesos, aunque en algunos meses pueden llegar a gastar alrededor de 190 mil pesos y otros meses gastar casi nada.

El promedio mensual de avances pedidos es cerca de 25 mil pesos entre todos los individuos. Si se considera sólo el 60%, que son los individuos que piden avances, se tiene que éstos piden alrededor de 41 mil pesos mensuales.

En este grupo los individuos acumulan en promedio, máximo 8000 puntos, con lo cual canjean casi siempre en el nivel 1, a pesar de que algunos poco llegan a canjear en el nivel 2 o en el 3, pero esto ocurre en muy pocas oportunidades. La mayoría de los canjes ocurren en los meses de marzo y abril, aunque hay que destacar que en todos los meses hay canjes de este grupo, al igual que todos.

El número de compras mensuales son en su mayoría 1 o 2, aunque como siempre hay individuos que realizan en algunos meses una mayor cantidad de éstas.

Grupo 7

La mayoría son personas entre 30 y 50 años. Hay un poco más de mujeres que hombres en este grupo, siendo éstas un 53% de los individuos. La mayoría posee hace poco tiempo la cuenta, sin embargo, hay varios también que la poseen hace más tiempo.

Al considerar su frecuencia de compra se observa que hay gente que va muy seguido a comprar, pero hay otros que van sólo en algunas oportunidades. Con esta diversidad, en promedio tienen cerca de un 79% de frecuencia de compra considerando todos los individuos en los 23 meses.

En promedio gastan alrededor de 37 mil pesos mensuales, aunque cabe mencionar que en navidad gastan mucho más, llegando algunos a gastar por sobre los 200 mil pesos, con lo que se nota que estos individuos son más estacionales (navideños).

Con respecto a los avances solicitados por los individuos de este grupo, ellos piden en promedio 21 mil pesos mensuales, considerando todos los individuos en todos los meses. Sin embargo, al tomar en cuenta sólo los que solicitan avances de este grupo, que corresponden a un 37% de los individuos, el monto varía a 33 mil pesos mensuales.

Hay que recordar que la gente no pide todos los meses avances y cuando lo hace es por un monto mucho más alto, sin embargo, el promedio es para los 23 meses.

Los puntos máximos acumulados por las personas de este grupo corresponden a 8000 mil puntos en promedio, con lo cual canjean principalmente en el nivel 1 y algunos lo hacen en el nivel 3, algunos pocos lo hacen en el nivel 2. Canjean principalmente en los meses de enero, febrero, marzo y agosto.

Se puede inferir que estas personas compran mucho en navidad, por lo que acumulan bastantes puntos, los cuales son canjeados prontamente en los meses siguientes.

El número de compras es en su mayoría hasta 4 compras mensuales, no obstante, también son realizadas más compras en algunos meses por los individuos de este grupo.

8.6.3 Calificación de los Segmentos Según sus Características

En esta sección, se bautizará cada segmento con un nombre, en base a los análisis realizados anteriormente. Cabe mencionar que es el análisis del punto 8.6.1 el que más influye en esta calificación, debido a que los segmentos fueron separados de acuerdo a las sensibilidades y no de acuerdo a las variables en sí, a pesar de ello, el análisis del punto 8.6.2 sirve como complemento para asignar el nombre a cada segmento.

En un principio se diferencian dos tipos de grupos, los que se caracterizan por su sensibilidad a la frecuencia de compra. Éstos son los “**consumistas**” (grupos 4, 6 y 7) y los “**controlados**” (grupos 3 y 5). Los primeros mientras más seguido van a comprar,

más van gastando, en cambio los segundos, mientras más seguido van a comprar, van gastando menos.

Entre los consumistas, el grupo 4 es en general el que se ve más afectado por variaciones en las variables, ya que los valores de los parámetros son más altos. Por lo tanto son **“vulnerables”**. Además este grupo se caracteriza por gastar poco en promedio mensual, sin embargo en ocasiones especiales como navidad, algunos tienden a gastar más, aunque no es una tendencia de todo el grupo. Por lo tanto son **“de gastos bajos”**.

El grupo 6 es parecido al grupo anterior, con la diferencia que no es tan vulnerable a las variables. La principal característica de este grupo es que a diferencia de los demás, mientras más años tiene la cuenta, menos gasta. Por lo que serán **“nuevos en la cuenta”**. Además gastan más que el grupo anterior, bastante más, por lo que son **“de gastos medios”**.

El último grupo de los consumistas es el grupo 7, el cual se caracteriza principalmente por su estacionalidad, ya que en algunos meses es negativa la influencia de estas variables, pero luego de éstos, la influencia de los otros meses es positiva. Se puede interpretar como que ahorran en algunos meses para poder comprar más al mes siguiente. Son los **“planificados”**. Este grupo gasta de manera similar al grupo 6, por lo que también son de **“de gastos medios”**. Además, este es el grupo que se ve más afectado por la dummy de diciembre, y al ver los montos que gastaban en este mes se notó que gastaban mucho, por lo que también serán llamados **“navideños”**

Entre los controlados, el grupo 3 se caracteriza por la gran sensibilidad negativa que tiene frente a la variable de Recency, es decir, mientras más tiempo pasa sin comprar es más probable que gaste menos, además de no gastar más por estar cerca del canje. Se llamarán **“despreocupados del canje”**. Asimismo, dado el alto monto que gastan mensualmente los clientes de este grupo, son también **“gastadores”**.

Finalmente el grupo 5 se caracteriza principalmente por tener una sensibilidad positiva frente al Recency, esto, junto con las influencia positiva de la frecuencia podría interpretarse como personas que si dejan de comprar es probable que después gasten más pero si están comprando muy seguido gastan menos. Por lo tanto son **“presupuestados o personas con presupuesto”**. Además, dado su comportamiento de compra, estos individuos gastan mucho en algunos meses y muy poco en otros, por lo que son **“volátiles en los gastos”**.

A continuación, la tabla 21 muestra el nombre con el que se bautiza a cada grupo, de acuerdo a las clasificaciones antes descritas. A pesar de que se tiene el nombre de cada grupo, en secciones consecutivas a ésta se mencionará a los grupos solamente por su número, con el propósito de que sea más entendible para el lector.

Tabla 21. Resumen de la clasificación de cada grupo.

Grupo	Nombre
3	Controlados, despreocupados del canje y gastadores
4	Consumistas, vulnerables y de gastos bajos
5	Controlados, presupuestados y volátiles en los gastos
6	Consumistas, nuevos en la cuenta y de gastos medios
7	Consumistas, planificados, de gastos medios y navideños

8.7 Cálculo de Customer Lifetime Value (CLV) para Cada Segmento

8.7.1 CLV Real para Periodo de Estimación y Validación

Si bien, el cálculo del CLV en general se realiza hacia el futuro con datos pronosticados, en este caso se realizará primero con datos reales para tener una idea de los segmentos más valiosos en el periodo de trabajo. Por lo tanto se realizará tanto para el periodo de estimación como para el periodo de validación, con el fin de observar si los segmentos valiosos cambian en estos dos periodos de tiempo.

Como se dijo en la sección 7.9, para calcular este valor se descuentan los montos hacia el futuro y se tiene una estimación de éstos en el presente. Además se ocupa una tasa de descuento de 1% mensual y una tasa de retención de 100%.

La tasa de descuento mensual es elegida de acuerdo a juicio de expertos, quienes en general usan una tasa de descuento entre un 10% y un 15% anual, y esto al transformarlo a un periodo mensual, resulta aproximadamente en un 1%, entonces $d=0.01$ en la ecuación (1) de la sección 7.9. Para esto se despeja el valor de la tasa mensual, dado el valor de la tasa anual, de la siguiente ecuación:

$$(1 + \text{tasa_anual}) = (1 + \text{tasa_mensual})^{12}$$

Dado que se tienen los individuos asignados a cada grupo, es posible calcular un valor de las transacciones realizadas por éstos en el futuro, a nivel grupal, para así poder comparar entre los distintos grupos de acuerdo a las transacciones efectuadas. El valor de cada grupo corresponderá al promedio de los valores de cada individuo perteneciente a éste.

8.7.2 CLV Estimado para Periodo de Estimación

El periodo de estimación consta de 23 meses a partir del mes 33, por lo cual el valor que se obtendrá correspondería al mes 32, es decir, es equivalente a que se hubieran pronosticado en ese mes los datos con los que se cuenta actualmente del mes 33 al 56 y se hubiera calculado el CLV en aquel mes.

A continuación en la tabla 22 se muestran los valores promedio de CLV correspondientes a cada segmento, los cuales fueron calculados de acuerdo a las transacciones realizadas con respecto a los montos gastados y a los avances realizados por los individuos de cada grupo, con lo cual se obtiene un total para cada uno de éstos.

Se observa que en cuanto al monto gastado, el grupo 3 es el más valioso, seguido por el grupo 5, siendo el menos valioso el grupo 4, lo cual era de esperarse, ya que cuando se analizaron estos grupos de acuerdo a la variable monto gastado, el grupo 3 gastaban en promedio más de 150 mil pesos mensuales y el grupo 4 gastaba en promedio 10 mil pesos mensuales. Además el grupo 5 gastaba alrededor de 75 mil pesos en promedio, lo cual lo ubicaba en el segundo lugar en cuanto a los montos gastados por lo clientes.

Tabla 22. CLV real periodo de estimación.

Segmento	CLV Promedio Monto	CLV Promedio Avance	CLV Promedio Total
3	3.412.045	899.000	4.311.045
5	1.582.393	466.353	2.048.746
7	766.702	428.414	1.266.926
6	754.489	512.437	1.195.115
4	212.268	312.545	524.813

Ahora bien, al mirar los avances realizados por las personas de cada grupo, se observa que el orden del valor de los grupos cambia, pasando el grupo 6 a ser más valioso que el 5 y el 7. No obstante, el grupo 3 y 4 mantienen sus lugares en el valor, siendo el mejor y el peor respectivamente.

A pesar del cambio en el valor de los segmentos en el cálculo de los avances, se puede apreciar que las diferencias entre los grupos no es tan grande en este ámbito (monto de avance), por lo que al sumar los dos tipos de transacciones, se observa que el orden de los montos gastados es el que prima, y por lo tanto, el grupo número 3 es el más valioso, seguido por el grupo 5, luego el grupo 7 y en seguida el grupo 6, siendo

finalmente el menos valioso el grupo 4. Esto es lo que aparece en la columna de CLV Promedio Total.

8.7.3 CLV Estimado para Periodo de Validación

Análogamente a lo hecho con el periodo de estimación, ahora se realiza un Customer Lifetime Value para el periodo de validación, es decir los 6 meses siguientes a partir del mes 57. Para esto se utiliza la misma tasa de retención y la misma tasa de descuento que para el periodo de estimación.

Con esto se quiere observar si los grupos siguen siendo igual de valiosos, o si cambian en este ranking, dados los nuevos datos.

En la tabla 23 se tiene el CLV para los montos gastados, los avances, y la suma de ambos, con lo cual se tendría el valor de cada segmento en el mes 56, es decir, antes del periodo de validación.

Tabla 23. CLV real periodo de validación.

Segmento	CLV Promedio Monto	CLV Promedio Avance	CLV Promedio Total
3	1.226.581	253.839	1.480.420
5	500.549	185.901	686.450
7	284.362	181.440	465.802
6	259.447	189.805	449.252
4	95.193	173.642	268.835

Se puede observar que se tienen los mismos patrones que en el periodo de estimación, ya que en cuanto a los montos gastados, en el periodo de validación, el grupo más valioso es el número 3, seguido por el grupo 5, luego el grupo 7 y el grupo 6, y el grupo 4 sigue siendo el menos valioso.

Asimismo, en cuanto a los avances, el grupo 6 pasa a ser más valioso que el 5 y el 7, y el grupo 3 y 4 siguen en la misma posición, el más valioso y el menos valioso respectivamente.

Finalmente, al igual que en el periodo de estimación, en este periodo los avances son muy parecidos entre los grupos, por lo que prima el valor del monto gastado, y en consecuencia se tiene que el grupo más valioso considerando estos dos tipos de

transacciones, es el grupo 3, seguido por el grupo 5, 7 y 6 (en ese orden), siendo el menos valioso nuevamente el grupo 4.

8.7.4 Validación del Modelo Estimado con Clases Latentes

Siguiendo con el modelo, al tener los parámetros de cada variable para cada grupo, se procede a validarlos en el periodo de 6 meses siguientes a los de la estimación. Es así como se predice el monto gastado al mes siguiente, tomando en cuenta la información real del mes anterior de las variables R, F, M, NC, PA y Dist; y la del mismo mes del resto de las variables observadas. Esto se compara con el monto real gastado en el mes correspondiente. Es decir, por ejemplo, se toma la información de las primeras variables mencionadas en el mes 56 y la información de las demás en el mes 57, y se predice el monto a gastar en el mes 57, esto se realiza en los 6 meses de validación y luego se comparan con el monto real gastado en cada mes.

Para esto, el monto pronosticado corresponde a la exponencial de la suma de la multiplicación de las variables por sus parámetros asociados, ya que se está trabajando con una distribución de Poisson, pero además hay que multiplicar el resultado por 10.000, ya que se debe recordar que para estimar los parámetros, la variable MG fue dividida por este valor. Por lo tanto se tiene:

$$\ln\left(\frac{MGp}{10.000}\right) = \sum_i \beta_i * X_i \quad \Rightarrow \quad MGp = 10.000 * e^{\sum_i \beta_i * X_i}$$

Donde MGp es el monto a gastar pronosticado, X_i es la información (valor) de la variable i y β_i es el parámetro correspondiente a la variable i .

Los resultados de las predicciones son presentados en la tabla 24, donde aparece el promedio del monto gastado por cada cliente en cada mes de la validación y el total gastado en este periodo. Esto para cada uno de los 5 grupos, además de en total todos los grupos. De esta manera se puede tener una idea de la precisión de la estimación.

Se aprecia al mirar esta tabla, que los montos pronosticados no difieren mucho de los montos reales, por lo que las estimaciones tanto a nivel agregado, como a nivel grupal, son de buena calidad, aunque claramente hay que considerar un error no tan despreciable en éstas, el cual será calculado posteriormente.

Se puede observar también que los grupos 3 y 5, que corresponden a clientes controlados, fueron sobreestimados, a diferencia de los otros tres grupos, correspondientes a clientes consumistas, los cuales fueron subestimados.

Tabla 24. Resultados de las predicciones de los montos gastados en el periodo de validación, según parámetros estimados en modelos de clase latente.

		Promedio	Total
Grupo 3	Real	211.752	21.598.670
	Pronóstico	229.824	23.442.041
Grupo 4	Real	16.477	2.570.386
	Pronóstico	8.188	1.277.302
Grupo 5	Real	86.234	162.464.313
	Pronóstico	93.944	176.990.113
Grupo 6	Real	44.705	65.983.869
	Pronóstico	36.485	53.852.180
Grupo 7	Real	49.031	46.186.971
	Pronóstico	44.736	42.141.479
Total	Real	65.527	298.804.209
	Pronóstico	65.286	297.703.116

A su vez, en la tabla 25 se muestran los errores agregados para cada grupo de clientes, y el error total agregado para todos ellos, tomando en cuenta el valor real y el valor pronosticado por el modelo. En esta tabla se puede observar que el error agregado total es bastante pequeño (0,4%), lo cual supone una buena estimación.

Al analizar los errores a nivel grupal (Tabla 25), se observa que el error agregado en los grupos 3 y 5, catalogados como controlados en sus gastos y que tienen un presupuesto para gastar, no supera el 10%, por lo que los pronósticos para estos grupos son de buena calidad.

Tabla 25. Error agregado.

Grupo	Monto Real	Monto Pronosticado	Error Agregado
3	21.598.670	23.442.041	8,5%
4	2.570.386	1.277.302	50,3%
5	162.464.313	176.990.113	8,9%
6	65.983.869	53.852.180	18,4%
7	46.186.971	42.141.479	8,8%
Total	298.804.209	297.703.116	0,4%

Se puede observar que el error es bastante grande en el grupo 4, lo cual se debe a que en este grupo los individuos en el periodo de estimación gastaban muy poco y en el periodo de validación aumentaron sus gastos de manera considerable. Hay que recordar que este era el grupo más vulnerable a los cambios en las variables externas, por lo que una posible explicación es que en ese periodo hayan existido cambios externos que provocaron tal aumento en los gastos (No hay información al respecto). Además, dado que es un grupo pequeño, el error en un cliente es importante en el cálculo del error total del grupo.

También se puede apreciar que en los grupos 6 y 7, que también son catalogados como consumistas (al igual que el grupo 4) en el análisis de los parámetros, los montos pronosticados son menores que los reales, debido probablemente a la misma razón explicada anteriormente.

Luego, para seguir complementando el análisis de la validación del pronóstico realizado, en la tabla 26 se muestra el valor del error cuadrático medio, tanto para cada grupo, como en total, comparando estos resultados con los obtenidos sin realizar la segmentación (modelo de independencia).

Si bien al no realizar segmentación, caso homogéneo, no existen grupos como para comparar, se tomó en cuenta el valor pronosticado de este caso para cada individuo y se asignaron al grupo encontrado posteriormente con la segmentación. De esta manera la diferencia en cada caso son los parámetros usados para el pronóstico, pero los clientes considerados son los mismos

Tabla 26. Error cuadrático medio (MSE).

Grupo	MSE homogéneo	MSE Heterogéneo
3	1.783.773	1.737.503
4	562.442	449.147
5	5.729.019	5.043.461
6	3.316.760	3.442.691
7	2.854.341	3.578.823
Total	7.447.685	7.301.866

Se observa que el error cuadrático medio de la muestra disminuye al agregar heterogeneidad, sin embargo, la diferencia es muy pequeña. Además, a nivel grupal se observa que los grupos que hasta ahora son de mayor interés, porque según el análisis de variables son los que más gastan, grupo 3 y 5, tienen un MSE menor que el obtenido con los parámetros iguales para cada grupo, lo cual podría significar un beneficio importante para la empresa al pronosticar mejor los clientes que son más importantes. Asimismo el grupo menos importante para la empresa (Grupo 4), también tiene un MSE

menor que en el caso homogéneo, por lo que esto igualmente podría ayudar a la empresa a discriminar a las personas de este grupo, ya que lo pronosticado no se aleja mucho de la realidad.

A continuación, en la tabla 27, se muestra el error individual promedio de cada grupo (MAPE), para observar el error que hay al pronosticar el monto gastado de cada individuo. Esto se compara con el obtenido en el caso homogéneo, y al igual que en el caso anterior, se asignaron los individuos a cada grupo para poder comparar.

Tabla 27. MAPE a nivel grupal con un grupo (homogéneo) y 5 grupos (heterogéneo).

Grupo	MAPE homogéneo	MAPE Heterogéneo
3	765,38%	261,51%
4	2.456,95%	80,31%
5	783,97%	162,29%
6	1.295,86%	117,12%
7	1.140,23%	118,01%
Total	1.066,65%	139,55%

Se observa que claramente el error a nivel individual disminuye considerablemente al agregar la heterogeneidad al problema, esto es así tanto a nivel de grupos, como a nivel global, por lo que la estimación es mucho más precisa.

Cabe mencionar que un MAPE promedio de un 100% significa que el error en la estimación fue del doble del valor, sin embargo, la disminución en este valor es notable (disminuye en un orden de magnitud), por lo que se puede decir que agregar la heterogeneidad, resulta en una mayor calidad de la estimación.

Luego de validar los valores de los parámetros estimados, se procede a utilizarlos, multiplicándolos por el valor de las variables independientes (observadas), para así poder hacer pronósticos futuros para la variable de montos gastados (dependiente).

8.7.5 Pronósticos

Hasta el momento sólo se ha pronosticado para un mes, considerando datos del mismo mes y datos del mes anterior, dependiendo de la variable observada. Sin embargo, para calcular el CLV se necesitan las transacciones de un periodo de tiempo de interés, por lo cual hay que hacer el pronóstico para el periodo y no sólo para el mes.

Para realizar estos pronósticos en un periodo de tiempo (dado que se puede predecir sólo para un mes), se pronostica para un segundo mes a partir de la predicción

para el primero, por lo cual es preciso actualizar el valor de las variables necesarias para el pronóstico, proceso que será explicado más adelante en la sección 8.7.5.1. A partir de los datos de las variables actualizadas, se predice el valor de la variable dependiente para el segundo mes. Luego, se realiza el mismo proceso anterior y se predice el monto gastado para un tercer mes. Así se continúa sucesivamente actualizando los valores de las variables para pronosticar el mes siguiente, y así tener la predicción para un periodo de tiempo y no sólo de un mes, utilizando sólo la información real del pasado.

En este caso se comienza con el pronóstico a partir del mes 56, es decir, a partir de los datos referentes al último mes de la estimación. Y así se pronostica el monto a gastar para el mes 57, luego se actualiza la información de las variables para este mes y se procede a pronosticar el valor del monto para el mes 58, lo mismo para pronosticar el mes 59 y así sucesivamente.

Los datos fueron pronosticados a partir del último mes de la estimación para poder comparar los montos pronosticados por el método, con los montos reales gastados en el periodo de validación, es decir, desde el mes 57 hasta el mes 62.

8.7.5.1 Actualización de Variables

A continuación se explicará la actualización una a una de las variables usadas.

Puntos acumulados (PA): A pesar de que los puntos acumulados en un mes dependen directamente del monto gastado, hay que recordar que también dependen de los puntos canjeados y los puntos vencidos en el mes en cuestión, por lo que se necesitaría un registro de un año hacia atrás de los puntos entregados al individuo, para así poder ir calculando los puntos vencidos 13 meses después de ser entregados. Hay que considerar también los puntos canjeados en el mes y la teoría FIFO del stock de puntos, dado que los puntos canjeados son los que se entregaron primero, lo cual provocaría un cálculo complicado de realizar para esta variable. Es por eso, que por simplicidad, y sin ser una idea muy fuera de la realidad, se decide actualizar esta variable usando una media móvil de los puntos acumulados hasta 6 meses antes del mes que se quiere obtener. De esta manera los puntos acumulados en un mes en particular, sería el promedio de los puntos acumulados en los 6 meses anteriores a éste.

Cabe mencionar que al aplicar una media móvil de los datos, se incurre en un error con respecto a lo que sucede realmente, ya que ésta se va haciendo constante, por lo tanto afectaría los pronósticos posteriores. Sin embargo, en la tabla 28 se muestra el error cometido en el periodo al hacer este supuesto, tanto a nivel grupal como a nivel agregado y se aprecia que los errores no son tan grandes por lo que al menos para el periodo de validación no es tan absurdo asumir este supuesto.

Tabla 28. Error de usar media móvil de los puntos acumulados.

	PA Real	PA Móvil	Error agregado
3	1.521.503	1.216.275	20,06%
4	341.789	351.179	2,75%
5	12.663.496	11.826.789	6,61%
6	6.147.777	6.066.013	1,33%
7	4.035.710	3.985.335	1,25%
Total	24.710.275	23.445.592	5,12%

El grupo que más error posee al utilizar el supuesto de media móvil para los puntos acumulados, es el grupo 3, por lo que se podría sobreestimar este grupo al momento de pronosticar los montos gastados. Esto último debido a la influencia positiva de esta variable frente a la que se quiere pronosticar. Cabe mencionar que como es un grupo pequeño, el error de estimación para un individuo afecta en gran medida el error del grupo entero.

Dist: La distancia al canje más cercano se actualiza usando los puntos acumulados actualizados que se tienen en el mes en cuestión (explicados recientemente), y con la misma fórmula para calcular esta variable explicada en la parte de transformación (sección 8.3).

Las variables de Recency (Rcomp), Frequency (Fcomp) y Monetary value (Mcomp), dependen directamente del monto pronosticado. Dado que la variable es estimada según la distribución de Poisson, entonces el valor es una exponencial, por lo que nunca será 0, a menos que la suma de la multiplicación de los parámetros estimados por los valores de las variables fuera $-\infty$. Por lo tanto, para efectos de la actualización de estas variables, será considerado como que un cliente no gasta, cuando el valor pronosticado sea menor que 1.

Con la consideración mencionada anteriormente las variables quedan como sigue:

Rcomp: El número de meses que han pasado desde la última compra será 0 si el valor pronosticado del monto gastado es mayor o igual a 1, y será igual al número de meses que han pasado en el mes anterior más uno ($Rcomp(t-1)+1$), si el valor pronosticado es menor que 1.

Fcomp: La frecuencia de compra será actualizada de manera de mantener la forma en que ha sido calculada anteriormente, por lo cual será considerada desde el primer mes de estimación. En consecuencia, si el monto pronosticado es mayor o igual a 1, la frecuencia será igual a la frecuencia en el mes anterior multiplicada por el número de meses que habían transcurrido hasta el mes anterior, más uno, todo esto dividido por el número de meses transcurridos actualmente. Si el valor pronosticado es

menor que 1, la frecuencia será igual la frecuencia en el mes anterior multiplicada por el número de meses que habían transcurrido hasta el mes anterior y dividida por el número de meses que han transcurrido hasta el mes actual.

Para entender mejor lo descrito anteriormente, se muestra el cálculo de la frecuencia para el mes 57.

$$Fcomp_{57} = \begin{cases} \frac{Fcomp_{56} * 56 + 1}{57} & \text{Si } MG_{57} \geq 1 \\ \frac{Fcomp_{56} * 56}{57} & \text{Si } MG_{57} < 1 \end{cases}$$

Mcomp: El monto promedio gastado en el mes t, se actualiza de manera muy similar a la frecuencia, con la diferencia que en vez de sumarle uno en caso de que el monto pronosticado sea mayor o igual que uno, en este caso se suma tal monto. Por lo tanto si el monto pronosticado es mayor o igual a 1 Mcomp será igual al monto promedio obtenido en el mes anterior multiplicado por la cantidad de meses transcurridos hasta el mes anterior, más el monto pronosticado multiplicado por 10.000⁷, todo esto dividido por la cantidad de meses transcurridos hasta el mes t. En caso contrario, Mcomp será igual al monto promedio obtenido en el mes anterior multiplicado por la cantidad de meses transcurridos hasta el mes anterior y dividido por el número de meses transcurridos hasta el mes actual.

Entonces se tiene:

$$Mcomp_{57} = \begin{cases} \frac{Mcomp_{56} * 56 + 10.000 * pron_{MG_{57}}}{57} & \text{Si } MG_{57} \geq 1 \\ \frac{Mcomp_{56} * 56}{57} & \text{Si } MG_{57} < 1 \end{cases}$$

Dummies_mes: Para las dummies de estacionalidad, simplemente se cambia el valor a 1 en la dummy que corresponde al mes actual en que se está haciendo el pronóstico, por ejemplo si el mes es abril, la dummy “d4” tomará el valor 1 y las demás serán 0. En el caso de que se esté en el mes de enero o septiembre (Dummies de control), todas las dummies de estacionalidad tomarán el valor 0.

Variables Socio-demográficas: Estas variables no serán actualizadas, y se tomarán como constantes para cada individuo, ya que para efectos del pronóstico en 6 meses y hasta en un año, la variación de la edad y de los años que posee la cuenta, Edad y A_cta respectivamente, será a lo más en una unidad, por lo que no afectaría en

⁷ Recordar que para realizar el pronóstico, el monto gastado fue dividido por 10000 para estimar los parámetros, por lo que hay que multiplicarlo para saber el valor después de la estimación.

gran medida los resultados del pronóstico. Además con respecto al sexo de los individuos (variable sex) éste será siempre el mismo.

NC: El número de compras realizadas en el mes t será actualizado de manera similar a los puntos acumulados, por lo que se realiza un promedio de los valores de los 6 meses anteriores.

Esta variable, al igual que para el caso de PA, tiene un error asociado al asumir una media móvil, y este es ilustrado en la tabla 29, donde se puede apreciar que el error corresponde alrededor de un 10% para cada grupo.

Hay que señalar que al actualizar las variables se comete un error constantemente, ya que algunas son actualizadas de acuerdo al pronóstico del monto gastado y otras de acuerdo a una media móvil, ambos tienen un error asociado. Sin embargo, dado que es prácticamente imposible predecir el futuro y no hay un método estándar para hacerlo, es necesario asumir estos errores para llegar a los resultados finales.

Tabla 29. Error de usar media móvil de NC.

	NC Real	NC Móvil	Error agregado
3	1160	995	14,21%
4	380	349	8,06%
5	9626	8818	8,40%
6	5070	4631	8,67%
7	2822	2593	8,13%
Total	19058	17385,471	8,78%

8.7.5.2 Validación del Pronóstico

Una vez actualizadas las variables independientes, se tienen los datos para ser multiplicados por los parámetros y así, luego de aplicar la exponencial a la suma de éstos y multiplicar por 10.000, obtener el monto pronosticado para el mes siguiente.

A continuación se muestran los resultados obtenidos luego del pronóstico de los montos gastados, y éstos son comparados con los montos reales para tener una medida de error de lo pronosticado.

En la tabla 30 se muestran los errores agregados para cada grupo de clientes, y el error total agregado para todos ellos, tomando en cuenta el valor real y el valor pronosticado con la actualización mes a mes. Esto es realizado para un periodo de 6 meses partiendo de los datos del mes 56 en la estimación.

Tabla 30. Error agregado del pronóstico a 6 meses.

Grupo	Monto Real	Monto Pronosticado	Error Agregado
3	21.598.670	23.060.346	6,77%
4	2.570.386	261.647	89,82%
5	162.464.313	170.956.301	5,23%
6	65.983.869	48.340.514	26,74%
7	46.186.971	36.464.127	21,05%
Total	298.804.209	279.082.935	6,60%

Se observa que el error a nivel agregado es bastante pequeño, por lo que el pronóstico para todos los individuos es de una calidad aceptable. Sin embargo, al observar los resultados a nivel grupal, se tiene que el grupo 4 está muy subestimado, obteniéndose un 89,82% de error agregado en ese grupo. Este error se podría explicar porque es un grupo pequeño, por lo que un error individual influye mucho en el error grupal, además de la vulnerabilidad de este grupo mencionada anteriormente.

También se nota nuevamente que los grupos 6 y 7, fueron, al igual que el grupo 4, subestimados, y su error es mayor al 20%. Por lo tanto se podría decir que el pronóstico para los grupos considerados como consumistas no es de muy buena calidad.

A pesar de lo anterior, se puede notar que los grupos de mayor interés (controlados), según lo visto en “CLV de validación”, siguen siendo bastante bien estimados, con un error menor al 10% en su pronóstico. Por lo tanto, lo que se pronostica que gasten los individuos es similar a lo que realmente gastan, entonces las personas de estos grupos seguirían siendo las más valiosas y el pronóstico para estos grupos es de buena calidad.

Luego, al igual que en la sección de validación, se puede analizar el error cuadrático medio del pronóstico realizado hacia el futuro de los grupos obtenidos en la segmentación y compararlo con los datos reales en el caso de no tener segmentación (asignando los individuos a los grupos igual que en el caso de validación). Este indicador se muestra en la tabla 31, donde se observa que los MSE heterogéneos son menores en los grupos 3, 4 y 5, pero son mayores en los grupos 6 y 7. En cuanto al total, el MSE también es menor por lo que las estimaciones agregando la heterogeneidad, siguen siendo más precisas que a nivel agregado y además permiten discriminar entre los distintos tipos de clientes.

Tabla 31. Error cuadrático medio (MSE) pronóstico.

Grupo	MSE homogéneo	MSE Heterogéneo
3	1.783.773	1.736.911
4	562.442	451.115
5	5.729.019	4.884.464
6	3.316.760	3.361.194
7	2.854.341	2.970.218
Total	7.447.685	6.870.092

A continuación, en la tabla 32, se ilustra el error individual (MAPE promedio), tanto a nivel grupal como a nivel total, con el fin de visualizar el error que se comete al pronosticar el monto gastado para cada individuo. Para poder comparar con el caso homogéneo nuevamente se asignan los individuos a un grupo como en la parte anterior.

Tabla 32. MAPE promedio pronóstico.

Grupo	MAPE homogéneo	MAPE Heterogéneo
3	765,38%	265,97%
4	2456,95%	96,57%
5	783,97%	151,65%
6	1295,86%	110,82%
7	1140,23%	110,24%
Total	1066,65%	147,05%

Se aprecia que aunque los errores a nivel individual siguen siendo bastante grandes, la gran disminución en el error asociado luego de aplicar la heterogeneidad, permite decir que trabajar con la diversidad de clientes aumenta en gran medida la calidad del conocimiento de las personas, bajando en el error de predicción individual, desde alrededor de un 1000%, a un valor alrededor de un 120%. Algo que podría diferenciar en gran medida el trabajo con los clientes de una empresa versus otra.

Cabe mencionar que estos errores son calculados en todo el periodo pronosticado, sin embargo existe un error cometido mes a mes, según lo que se predice y lo real. Estos análisis mensuales se muestran en el Anexo 4.

8.7.5.3 CLV para el Pronóstico

Al igual que para el periodo de estimación y para el de validación, ahora se procede a calcular el Customer Lifetime Value para los montos pronosticados en todo el periodo de validación, es decir, se calcula el valor presente de las transacciones futuras de los individuos de cada grupo, luego se hace un promedio de estos valores en cada grupo y eso representa el valor del segmento correspondiente.

Para esto se usa una tasa de descuento de 1% mensual y una tasa de retención del 100%, con lo cual se obtienen los resultados ilustrados en la tabla 33. Además estos son comparados con los de los datos reales, para observar que tan cercano a la realidad es el cálculo de este indicador en base al pronóstico del futuro.

Al obtener los valores presentes de los montos gastados en los próximos 6 meses, se consiguen los mismos resultados que en los cálculos anteriores, al menos en el orden de importancia de los grupos, siendo el más valioso el grupo número 3, seguido por los números 5, 7 y 6 (en ese orden), quedando en el último lugar el grupo menos valioso que sigue siendo el número 4.

Tabla 33. CLV promedio monto pronosticado 6 meses.

Segmento	CLV Promedio Monto Real	CLV Promedio Monto Pronosticado
3	1.226.581	1.310.866
4	95.193	9.796
5	500.549	526.363
6	259.447	190.029
7	284.362	225.118

Como se vio anteriormente, los avances son muy parecidos entre los grupos, por lo que al agregarlos al pronóstico el ranking del valor de éstos, sigue teniendo el mismo orden.

También se puede apreciar que el CLV pronosticado es muy cercano al que se obtendría con los datos reales, por lo que el pronóstico sería de gran utilidad para calcular este indicador e identificar a los segmentos más valiosos

8.7.5.4 Utilización de Recursos

Dado que se tiene una estimación del valor del cliente a futuro, a continuación se analiza la cantidad de recursos que se deberían utilizar en cada grupo, de acuerdo a los resultados obtenidos, con el fin de realizar un uso más eficiente de éstos, dedicando una mayor cantidad a los más valiosos.

Se mencionó anteriormente que los avances pedidos por las personas no marcaban una gran diferencia en los distintos grupos, y que para el cálculo del valor del cliente, éstos no alteraban el orden del valor de cada segmento para la empresa. Por lo tanto, en la tabla 34 se muestra el porcentaje que debería ser dedicado a cada segmento, tomando en cuenta el CLV en el periodo de validación, sólo de los montos gastados reales y los pronosticados. En esta tabla se puede apreciar que el porcentaje de los recursos que se dedicaría a cada segmento tomando en cuenta el monto pronosticado, no es muy distinto de lo que realmente debería dedicarse, dados los montos reales.

Se observa que los porcentajes de los grupos 3 y 5 son sobreestimados y los de los otros grupos son subestimados. Este error en la diversificación de recursos no debería ser considerado como grave, ya que se beneficia más a los clientes más valiosos para la empresa y se perjudica a los que tienen menor valor. Además, una posible estrategia de la empresa sería no considerar al segmento que le genera menores ingresos o dedicar los recursos sólo para los segmentos valiosos, por lo que los errores cometidos no perjudicarían en gran medida a las estrategias de la empresa.

Por lo tanto, según el criterio del monto pronosticado, en el grupo más valioso para la empresa (grupo 3) deberían ser ocupados alrededor de un 55% de los recursos y en el segundo más valioso (grupo 5), se deberían utilizar alrededor de un 20% de los recursos. El 25% de los recursos restantes deberían ser divididos entre los otros segmentos de clientes, dedicando al menos valioso, cerca de un 1% de los recursos.

Tabla 34. Porcentaje de recursos a gastar considerando CLV de montos gastados.

Grupo	% a Gastar Real	% a Gastar Pronosticado
3	51,84%	57,95%
4	4,02%	0,43%
5	21,15%	23,27%
6	10,97%	8,40%
7	12,02%	9,95%

A pesar de que los avances no fueron considerados, vale la pena observar la distribución de los recursos si incluimos esta variable en las transacciones de los clientes. Es así como en la tabla 35 se muestran los porcentajes de recursos dedicados

a cada grupo, considerando el cálculo del Customer Lifetime Value en el periodo de validación, incluyendo tanto los montos gastados como los avances pedidos por los clientes. En este caso cabe mencionar que los montos de avances son reales en ambos casos (Real y Pronóstico), ya que no hay un modelo para pronosticar los avances, y suponer una media móvil para esta variable se aleja mucho de la realidad, ya que al observar los datos, los avances pedidos por los clientes sólo son en algunos meses.

Obviamente, como los valores de los avances son los mismos en el caso real y pronosticado, la diferencia sólo será por parte de los montos gastados, es por eso que el error cometido seguirá siendo bajo, pero hay que apreciar que los recursos estimados para los grupos 3 y 5 disminuyen, en cambio para los otros grupos aumentan. A pesar de esto, la proporción de recursos que deberían ser destinados al grupo más valioso sigue siendo alta, con alrededor de un 45%; y para el segundo más valioso se dedicaría un porcentaje cercano al 20% (lo mismo anterior). El 35% restante sería dedicado a los otros tres grupos, y hay que notar que ahora la proporción dedicada al segmento número 4 aumenta a un 6%. Esto último debido a que, a pesar de que el grupo 4 gasta poco en cuanto a los montos gastados, no tiene una diferencia tan marcada en cuanto a los avances solicitados (igualmente es el grupo que pide menos avances).

Tabla 35. Porcentaje de recursos a gastar considerando CLV de montos gastados y avances.

Grupo	% a Gastar Real	% a Gastar Pronosticado
3	44,18%	48,19%
4	8,02%	5,65%
5	20,49%	21,94%
6	13,41%	11,70%
7	13,90%	12,52%

Dado que las empresas deberían considerar todos los tipos de ingresos recibidos por los clientes para designar la cantidad de recursos que dedicarían para satisfacerlos, son los resultados obtenidos en la tabla 35 los más adecuados para este tema. Sin embargo, como se trata de pronosticar hacia el futuro, no se tendrían los avances reales pedidos por los clientes de la empresa, a menos que se realizara un modelo que pronosticara esta variable, para lo cual habría que hacer un nuevo estudio al respecto y no es parte de esta memoria.

Pese a todo, al comparar el porcentaje real a gastar tomando en cuenta los avances (que es lo que se debería hacer) y el porcentaje pronosticado a gastar sin considerarlos, se puede apreciar que a excepción de los grupos 3 y 4, el porcentaje pronosticado considerando sólo los montos gastados, es bastante cercano a lo que realmente debería ser. Además como se sabe que los montos para el grupo 3 y 5 son

sobreestimados y para los otros grupos son subestimados, se deberían asignar los recursos teniendo en cuenta esto, con lo que al grupo 5 se le concedería un 20% y a los grupos 6 y 7 se les adjudicaría un poco más de un 10%. Al considerar los errores de pronóstico de los montos gastados se tiene una asignación bastante cercana a lo que se debería hacer, al menos para los grupos 5, 6 y 7.

Ahora bien, para los grupos 3 y 4, el error que se cometería en la asignación de recursos sería mucho mayor, destinando muchos recursos al primero, y demasiado pocos al segundo. No obstante, si lo que la empresa quiere es preocuparse de los clientes que le entregan la mayor cantidad de ingresos, no sería contraproducente dedicar aún más recursos de los que deberían para estos clientes. Asimismo, los clientes que generan una pequeña cantidad de los ingresos de la empresa quizás deberían no ser considerados, o bien, tener una muy pequeña proporción de los recursos.

Por lo tanto, la asignación de recursos de acuerdo a los montos gastados parece bastante adecuada para la empresa, ya que se busca la fidelización de los clientes más valiosos. En consecuencia, se dedicaría alrededor de un 55% de los recursos a los clientes más valiosos, tomando en cuenta que este grupo también es sobreestimado.

8.8 Análisis de Sensibilidad de Acuerdo a la Tasa de Retención

Para el cálculo del valor presente de las transacciones futuras (CLV), se realizó el supuesto de una tasa de retención de 100%, es decir, de un mes a otro se mantienen todos los clientes y por lo tanto no es necesario multiplicar por un factor “r” en la fórmula de CLV (ver sección 7.9).

A continuación se realiza un análisis de sensibilidad de la tasa de retención, variando el valor de ella de acuerdo a lo intuido por la empresa, mínimo un 85%. La variación se realiza disminuyendo en intervalos de 5% el valor de la tasa. Por lo tanto se tendrán los resultados para 95%, 90% y 85%, ya que el valor para 100% fue calculado anteriormente según el supuesto.

El análisis se efectúa en el periodo de validación, ya que este es el periodo de mayor interés y los resultados se muestran en la tabla 36. Usando una tasa de retención de 95%, 90% y 85% respectivamente como muestra esta tabla, se puede apreciar que el orden de los segmentos sigue siendo el mismo. Donde el grupo número 3 es el de mayor valor, seguido por el grupo 5, consecutivamente los grupos 7 y 6, dejando al grupo número 4 como el menos valioso. Se comprueba entonces que para distintas tasas de retención, el ranking en cuanto a valor de los grupos sigue siendo el mismo.

Tabla 36. Variación de CLV promedio al disminuir en 5% la tasa de retención.

Segmento	CLV (95%)	CLV (90%)	CLV (85%)
3	1.027.922	860.511	719.748
5	423.818	358.372	302.633
7	239.829	202.123	170.247
6	219.531	185.573	156.726
4	78.710	64.978	53.573

Los resultados se obtienen al considerar que todos los grupos tienen la misma tasa de retención, lo cual no es tan claro en la realidad, por lo que más que observar el orden que tienen aplicando una misma tasa de retención a todos, hay que observar el porcentaje de variación que tienen los grupos frente a un cambio porcentual en la tasa de retención.

Es así que se puede calcular la diferencia porcentual, obtenida al variar la tasa de retención en un cierto porcentaje, como:

$$DIF\% = \frac{Valor(95\%) - Valor(90\%)}{Valor(95\%)}$$

Con lo que se obtendría el porcentaje de variación en el valor del CLV, dada una disminución en un 5% de la tasa de retención. La tabla 37 muestra las diferencias porcentuales calculadas con la fórmula anterior, además de un promedio de éstas para cada grupo. Entonces, para el grupo 3, el CLV disminuirá en un 16,28%; para el grupo 5, en un 15,44%; para el grupo 7, en un 15,72%; para el grupo 6, en un 15,47%; y para el grupo 4, en un 17,44%.

Tabla 37. Diferencia porcentual al cambio de un 5% en la tasa de retención.

Segmento	DIF	DIF	DIF	Promedio
	100%-95%	95%-90%	90%-85%	
3	16,20%	16,29%	16,36%	16,28%
5	15,33%	15,44%	15,55%	15,44%
7	15,66%	15,72%	15,77%	15,72%
6	15,38%	15,47%	15,54%	15,47%
4	17,32%	17,45%	17,55%	17,44%

Análogamente se puede ver en el anexo 5, la diferencia porcentual para una variación de 1% en la tasa de retención.

Los resultados muestran que todos los grupos tienen una sensibilidad similar con respecto a la variación en la tasa de retención, sin embargo, el grupo más sensible a ésta es el grupo número 4, y el menos sensible es el grupo número 5. Esto podría explicarse por la gran cantidad de gente que tiene el grupo 5, junto con los altos valores de sus montos gastados. En cambio el grupo 4 está compuesto por mucho menos individuos, por lo que perder un porcentaje de clientes afecta más a este grupo al ser más pequeño.

Cabe mencionar que a pesar de la sensibilidad mencionada de los distintos grupos frente al cambio en la tasa de retención, dado que lo que gastan es muy distinto entre grupos, es de utilidad mostrar la diferencia que hay en pesos (tabla 36), tomando en cuenta el valor del CLV al variar un 5% la tasa de retención, partiendo de 100% y hasta 85%. También se puede ver el anexo 5 para observar el cambio con 1% de variación en esta tasa.

Tabla 38. Disminución en CLV al disminuir 5% la tasa de retención.

Segmento	100%-95%	95%-90%	90%-85%
3	198.659	167.411	140.763
5	76.731	65.446	55.739
7	44.533	37.706	31.876
6	39.916	33.958	28.847
4	16.483	13.732	11.405

Se puede apreciar en la tabla 38 que los grupos más valiosos disminuyen en mayor cantidad su valor, dado que gastan más, lo que es una clara prueba de que es muy importante retenerlos.

Observando las tablas anteriores, principalmente la tabla 36, se nota que para determinar el valor de los clientes, es importante conocer mejor la tasa de retención, ya que por ejemplo el grupo 6, que con una tasa de 95% de retención, es menos valioso que el grupo 7 con la misma tasa, si este último tiene una tasa de retención de 90% y el otro mantiene su tasa de 95%, es el grupo 6 el que pasa a ser el más valioso entre los dos, es decir, si la tasa de retención del grupo 7 es un 5% menor que la del grupo 6, este último es el más valioso entre los dos.

9 Conclusiones

9.1 Conclusiones Generales

En esta memoria se puede concluir que, como era de esperarse, las predicciones con respecto a los montos gastados por los clientes, son más precisas al considerar heterogeneidad entre ellos. Es decir, tomando en cuenta distintos grupos de clientes en base a su comportamiento de compra, las predicciones para cada uno de estos grupos son de buena calidad, lo que se observa en los errores agregados de los pronósticos. Además, al agrupar a los clientes, la predicción a nivel individual también mejora considerablemente, lo que se puede apreciar en el MAPE promedio de los individuos. Esto permite conocer mejor el comportamiento de compra de los clientes y con esto ser más eficaz al momento de crear estrategias para satisfacerlos.

También se puede concluir que el comportamiento de compra histórico, medido a través de variables transaccionales, junto con variables socio-demográficas, puede explicar de una manera aceptable el comportamiento futuro de los clientes, por medio de modelos de predicción, obteniéndose pronósticos cercanos a la realidad (ya que los errores de predicción son pequeños). Esto permite estimar el valor del ciclo de vida de un cliente (CLV) para la empresa.

Los modelos de predicción son adecuados a nivel de segmento, aunque tienden a sobreestimar clientes que son considerados como controlados en sus gastos (grupos 3 y 5). En cambio para clientes considerados como consumidores (grupos 4, 6 y 7) el modelo tiende a subestimar el monto a gastar por ellos.

Cabe mencionar también que al estimar un modelo que predice sólo un mes con los datos entregados en el mes anterior, las estimaciones son de buena calidad, sin embargo, al hacerlo para un periodo de tiempo futuro, esta precisión es mucho menor. Esto es de esperarse, ya que al pronosticar para un periodo de tiempo se incurre en un error mes a mes, el cual va aumentando mientras mayor es el periodo.

Pese a lo anterior, la estimación para los grupos más valiosos es bastante buena en el periodo de tiempo, por lo que si la empresa se quiere preocupar de estos segmentos de clientes, el estudio le permitiría tener una idea de cuánto valen estos clientes para ella y por lo tanto tener un indicio de cuánto deberían invertir en ellos.

9.2 Conclusiones con Respecto a la Segmentación

El método de clase latente, con el cual se realiza la segmentación, permite calcular simultáneamente los distintos tipos de segmentos y los parámetros, asociados a cada variable, para cada uno de ellos, por lo cual se realizan estos dos pasos con el mismo criterio de optimización (Picón, 2004). Por lo tanto, dado que se necesitaban realizar estos dos pasos, este modelo resultó bastante útil y permitió obtener resultados satisfactorios.

Gracias al uso de modelos de clase latente fue posible encontrar 5 grupos de personas de acuerdo a su comportamiento de compra, los cuales se diferencian en la sensibilidad, que tienen los parámetros asociados a las variables independientes, sobre el monto gastado. De esta manera las personas pertenecientes a un mismo grupo reaccionan de la misma manera frente a un cambio en el valor de las variables independientes, pero de manera distinta a las personas de otro grupo.

Los grupos se diferencian principalmente en su frecuencia de compra, aunque cabe destacar que son diferentes en la influencia de cada uno de los parámetros asociados a las variables independientes. En algunos casos es diferente sólo el valor del parámetro, pero en otros la influencia es totalmente distinta, ya que el signo del parámetro es opuesto.

En un principio se identifican dos tipos de grupos, los consumidores y los controlados, caracterizados por la influencia del parámetro asociado a la variable de frecuencia. El primero corresponde a las personas que mientras más van a comprar, más gastan, y estos son los grupos 4, 6 y 7; el segundo, grupo de los controlados, corresponde a personas que si están comprando muy frecuentemente, comienzan a disminuir el monto de la compra, éstos son los grupos 3 y 5. A su vez, los grupos son caracterizados de acuerdo a la influencia de los otros parámetros, junto con los montos gastados por cada uno de ellos, obteniéndose una clasificación y un nombre para cada uno de los cinco grupos como muestra la tabla 21.

Cabe mencionar, que las sensibilidades de los parámetros asociados a las variables, muestran que en general son las mujeres las que gastan más (en todos los grupos excepto el 5), que la gente gasta más en diciembre, y que las personas en general se sienten motivadas a gastar más cuando están cerca de un premio, en este caso cerca de un canje superior, y que tiende a conformarse con lo que tiene cuando ya lo tiene, es decir, está más cerca del canje inferior (excepto el grupo 3). Estas características son las que se esperaban por lógica, lo que hace inferir que el estudio interpreta de buena manera el comportamiento de las personas y no es irracional.

9.3 Conclusiones con Respecto al Valor del Cliente

El valor del cliente es muy importante para las empresas, y el conocerlo les permite identificar a los clientes más valiosos y en ellos enfocar los mayores esfuerzos de marketing para crear relaciones de largo plazo con ellos.

Se puede concluir de acuerdo a los resultados, que los grupos más valiosos son los que se controlan en los gastos, los que probablemente tienen un presupuesto de compra y que en este caso corresponden a los grupos 3 y 5. Esto puede ser debido a que la gente que mantiene el control de sus gastos tiene el dinero pronosticado para gastar en la tienda, por lo que siempre se mantendrá gastando una suma más o menos constante cada año. Por lo tanto, estos clientes tienden a ser más fieles a las empresas

y compran en general en los mismos lugares, siempre que se sientan conformes con los productos y servicios brindados.

Es a los clientes mencionados anteriormente a quienes hay que mantener contentos y dedicar los mayores esfuerzos de marketing, además, de acuerdo al CLV calculado, estos dos segmentos generan más de un 60% de los ingresos de la empresa, por lo que es necesario satisfacerlos, para evitar una pérdida significativa por la posible fuga del cliente a una empresa de la competencia. Como estos clientes tienen al parecer un presupuesto para gastar, sería recomendable observar los productos que compran, y así percibir si el dinero dedicado a la empresa es siempre para el mismo tipo de producto. De esta manera se podría identificar los productos en los cuales hay que mantener contento al cliente y en cuáles hay que tratar de atraerlos. Es decir, más que buscar un up-selling, hay que buscar un cross-selling con este tipo de clientes y así aumentar la participación de la empresa en la canasta de compra de ellos, ya sea con la firma en sí, o al menos con alguna de sus alianzas. También sería recomendable hacer encuestas a estos tipos de clientes con el fin de conocerlos mejor e identificar patrones de comportamiento para poder satisfacerlos de manera más precisa.

Cabe mencionar que la tasa de retención cumple un papel importante al momento de calcular el CLV, sin embargo dado los resultados, los grupos considerados como más valiosos, lo seguirán siendo aunque su tasa de retención sea menor que la de los otros grupos, a menos que la diferencia entre éstas fuera demasiado grande, lo cual no tendría sentido. Además, la estrategia de la empresa tiene que enfocarse en retener a estos clientes, por lo que la tasa de retención debe ser alta. De no serlo las pérdidas serían muy altas.

Sobre este mismo punto, es interesante notar que al disminuir la tasa de retención de los más valiosos (grupo 3) en un 5%, el valor de éstos disminuiría en más de 100 mil pesos inmediatamente, versus una disminución de alrededor de 10 mil pesos en el segmento menos valioso. Por lo tanto, claramente no retener a estos clientes valiosos es una gran pérdida para la empresa. Ellos representan más de 10 veces lo que vale un cliente del segmento menos valioso.

Finalmente se pudo concluir que, a pesar de ser importante incluir la variable de los montos de avances para estimar el verdadero valor del cliente, los resultados obtenidos utilizando sólo los montos, ofrecen una buena alternativa para saber cuánto invertir en cada uno de los segmentos encontrados por el modelo. Y con respecto al valor de los segmentos, el ranking de ellos es el mismo con y sin incorporar la información de los avances.

9.4 Limitaciones del Modelo

El modelo considera la media móvil de 6 meses para pronosticar algunas variables tales como, los puntos acumulados y los números de compras realizados. Dado que al aplicar media móvil, el valor de la variable se va haciendo constante, el valor de la variable podría ser muy distinto a lo real. Además la variable de la distancia al canje más cercano depende directamente de los puntos acumulados, por lo que también sufren el error de usar media móvil.

Asimismo, las variables de RFM dependen del monto pronosticado y se definen de acuerdo a éste, por lo que mes a mes se incurre en un error que viene también por parte de los montos pronosticados.

Finalmente, la variable de avance no fue incluida en el modelo y esta podría afectar a los montos gastados, por lo que debería ser incluida. Además habría que hacer un modelo también para pronosticar esta variable.

Esto limita el modelo realizado, y se deben asumir estos errores. Podría ser realizado un modelo para predecir el valor de cada variable en el tiempo, pero eso está fuera del alcance de esta memoria.

9.5 Trabajos Futuros

Para conocer aún mejor a los clientes, sería recomendable investigar los tipos de productos que compran los clientes de los distintos segmentos, además de los tipos de productos que canjean éstos cuando llegan a acumular los puntos necesarios.

Otro posible trabajo futuro podría incluir los montos de avances como variable independiente y realizar el modelo con esta variable internalizada. También sería bueno realizar un modelo para estimar esta variable y quizás también las otras, y así tener pronósticos más precisos.

Por otra parte, dado que se observó que la tasa de retención influye bastante en la estimación de los montos gastados, sería útil realizar un análisis sobre la fuga de los clientes de la empresa, para notar cual es el verdadero porcentaje de clientes que está dejando la empresa e incluirlos en el cálculo del valor del cliente, para tener una estimación mejor de este indicador.

Finalmente, dado que se tienen distintos tipos de clientes y lo que se busca es satisfacerlos, para crear relaciones de largo plazo con ellos; se deberían estudiar algunas acciones de marketing a aplicar sobre los segmentos de la empresa, y así encontrar las acciones más efectivas.

10 Bibliografía

- Andersen, A. "New Rules, New Realities". Advertisement. 2000.
- Anderson, T. "On Estimation of Parameters in Latent Structure Analysis". Psychometrika. Vol 19. 1954. 1-10.
- Bartholomew, D., F. Steele, I. Moustaki y J. Galbraith. "The Analysis and Interpretation of Multivariate Data for Social Scientists". Chapman and Hall. Boca Raton. 2002
- Blattberg, R. y J. Deighton. "Management Marketing by the Customer Equity Test". Harvard Business Review. 1996.
- Desarbo, W. "Customer Value Analysis in a Heterogeneous Market". Strategic Management Journal. Vol 22. 2001. 845-857.
- Durán, M. "Estimación del Valor del Programa de Puntos, en Función del Comportamiento de Canje y Compra de los Clientes". Memoria para optar al título de Ingeniero Civil Industrial, Universidad de Chile. Facultad de Ciencias Físicas y Matemáticas, Departamento de Ingeniería Industrial. 2008.
- Espinoza, I. "Estimación de Demanda de Productos, Dado un Cierta Nivel de Puntos de Canje, para un Club de Fidelización que se Basa en la Acumulación de Puntos". Memoria para optar al título de Ingeniero Civil Industrial, Universidad de Chile. Facultad de Ciencias Físicas y Matemáticas, Departamento de Ingeniería Industrial, 2008.
- Fader, P., B. Hardie y P. Berger. "Customer-Base Analysis with Discrete-time Transaction Data". 2004. Disponible en SSRN: <http://ssrn.com/abstract=596801>.
- Fader, P., B. Hardie y K. Lee. "RFM and CLV: Using Iso-value Curves for Customer Base Analysis". Journal of Marketing Research. Vol. 42. 2005. 415-430.
- Forbes, T. "Valuing Customers". Journal of Database Marketing & Customer Strategy Management. 2007. 4-10.
- González, A. y M. Monteavaro. "Algunos Problemas en la Determinación de la Confiabilidad de Indicadores Cualitativos Vía Análisis de Clases Latentes". Revista de Investigación Operacional. Vol. 24. No. 3. 2003.
- Gupta, S., D. Lehmann y J. Stuart. "Valuing Customers". Journal of Marketing Research. Vol. 41. 2002. 7-18.

- Gupta, S., D. Hanssens y B. Hardie. "Modeling Customer Lifetime Value". *Journal of Service Research*. Volume 9. No.2. November 2006. 139-155.
- Hannan, E. "The Estimation of the Order of an ARMA Model Process". *Ann Statist.* Vol 8. 1980. 1071-1081.
- Kamakura, W. y G. Russell. "A Probabilistic Choice Model for Market Segmentation and Elasticity Structure". *Journal of Marketing Research*. 1989. 379-390.
- Kamakura, W. y C. Mela. "Choice Models and Relationship Management". Summary Paper for the sixth Choice Symposium. 2004.
- Kivetz, R., Y. Zheng y O. Urminsky. "The Goal-Gradient Hypothesis Resurrected: Purchase Acceleration, Illusionary Goal Progress, and Customer Retention". *Journal of Marketing Research*. 2006.
- Magidson, J. y J. Vermunt. "Latent Class Models for Clustering: A Comparison with K-means". *Canadian Journal of Marketing Research*. Volume 20. 2002. 37-44.
- Mallea, J. "Aplicación de los Modelos de Clase Latente a la Segmentación: Caso Economax". Memoria para optar al título de Ingeniero Civil Industrial, Universidad de Chile. Facultad de Ciencias Físicas y Matemáticas, Departamento de Ingeniería Industrial. 2004.
- Marín, P. "Estimación de Lifetime Value Basada en Comportamiento Transaccional", Memoria para optar al título de Ingeniero Civil Industrial, Universidad de Chile. Facultad de Ciencias Físicas y Matemáticas, Departamento de Ingeniería Industrial. 2005.
- Myung, I. "Tutorial on Maximum Likelihood Estimation". *Journal of Mathematical Psychology*. Vol. 47. 2003. 90-100.
- Nunes, J. y X. Drèze. "Pricing in Combinations of Currencies to Lower Consumers' Perceived Cost". *Journal of Marketing Research*. Vol. 41. 2004. 59-72.
- Nunes, J. y X. Drèze. "The Endowed Progress Effect: How Artificial Advancement Increases Effort". *Journal of Consumer Research*. Vol. 32. 2006. 504-512.
- Picón, E. "Una Comparación Montecarlo de 3 Métodos Métricos de Segmentación con Análisis Conjunto". Facultad de Psicología, Universidad de Salamanca. 2004. 231-252.
- Smith, C. y J. Pontius. "Jackknife Estimator of Species Richness with S-PLUS". *Journal of Statistical Software*. 2006.

- Train, K. "Discrete Choice Methods with Simulation". Cambridge University Press. 2003.
- Von Storch, H. and F. Zwiers. "Statistical Analysis in Climate Research". Cambridge University Press. 1999.
- Wansink, B. "Developing a Cost Effective Brand Loyalty Program". Journal of Advertising Research. 2003. 301-311.

11 Anexos

11.1 Anexo 1: Cálculo del Tamaño Muestral

Se utiliza la fórmula:

$$n = \frac{N \cdot Z_{\alpha/2}^2 \cdot P \cdot (1 - P)}{(N - 1) \cdot e^2 + Z_{\alpha/2}^2 \cdot P \cdot (1 - P)}$$

Donde:

N = tamaño de la población = 2.363.621 clientes. Se asume que se distribuye normalmente.

$Z_{\alpha/2}$ = nivel de confianza elegido. Se utiliza un 99% de confianza, con lo cual $Z_{99\%} = 2,576$

P = proporción de una categoría. Se elige P = 0,5 para maximizar el tamaño muestral.

e = error máximo. Se elige un 0,6%.

Luego, $n(99\%) = 45201$

11.2 Anexo 2: Cálculo del Tamaño de Muestra para la Estimación de Modelos

Se utiliza la misma fórmula del anexo 1.

N= 910.885

$Z_{\alpha/2}$ (95% confianza)=1,96

P=0,5

e (error muestral)=0,036

Se busca el tamaño adecuado para la muestra (n), con un intervalo de confianza de 95% y un error muestral de 3,6%:

$$n = \frac{N \cdot Z_{\alpha/2}^2 \cdot P \cdot (1 - P)}{(N - 1) \cdot e^2 + Z_{\alpha/2}^2 \cdot P \cdot (1 - P)} = \frac{446333.7}{570.3} = 740$$

También se puede considerar un intervalo de confianza de 99% con un error muestral de 4,7%. Entonces $n=750$

11.3 Anexo 3: Caracterización Detallada de Segmentos según Sensibilidad a las Variables

Grupo 4

En la tabla 39 se muestra la influencia de los parámetros asociados a las variables que son consideradas en el mes anterior al mes que se quiere pronosticar. Las influencias de estas variables son explicadas en seguida.

Tabla 39. Sensibilidad de variables observadas en el mes anterior para el grupo 4.

PA	Dist	Rcomp	Fcomp	Mcomp	NC
0,000102	0,555912	-10,7865	4,635044	0,003673	0,264244

Los puntos acumulados influyen positivamente en este grupo, por lo que mientras más puntos tenga acumulado un individuo en un mes, mayor será el monto gastado en el mes siguiente.

La distancia al canje más cercano influye positivamente en este grupo, mientras más cerca del canje superior se encuentra el individuo, es más probable que gaste más con el fin de poder canjear. Análogamente, mientras más cerca está el individuo del canje inferior, es probable que no se motive para gastar más y se conforme con lo que ya tiene, por lo tanto gasta menos.

Al igual que en el grupo 3, los individuos probablemente gastarán menos al próximo mes si lleva 5 meses sin comprar, que si lleva 3, es decir, mientras más tiempo lleve sin comprar, menos gastará el siguiente mes (según Rcomp).

En este grupo, las personas no se controlan en lo que gastan, y mientras más seguido van a comprar, es más probable que gasten más. Esto se puede observar en la influencia positiva de la frecuencia de compra. Además es el grupo que tiene una mayor influencia de esta variable.

El número de compras también influye positivamente en el monto gastado, lo que indica que mientras más compras realice en un mes, es más probable que gaste más el próximo mes. Y cabe mencionar que a pesar de que en todos existe una influencia positiva del número de compras, este es el grupo más sensible con respecto a esta variable.

La tabla 40 muestra las sensibilidades de los parámetros asociados a las variables de estacionalidad, es decir, las dummies correspondientes a los meses, que toman el valor 1 si el mes a pronosticar corresponde a la dummy relacionada a ese mes, y 0 en caso contrario.

Tabla 40. Sensibilidad de Dummies de estacionalidad para el grupo 4.

d2	d3	d4	d5	d6	d7	d8	d10	d11	d12
0,1654	-0,0177	0,3566	0,1701	0,7607	-0,0855	0,2961	0,3371	-0,00003	0,1575

Con respecto a la estacionalidad se observa que este grupo cada 4 meses tiende a gastar menos, y esto ocurre en los meses de marzo, julio y noviembre. En cambio los otros meses tienen una influencia positiva con respecto a los meses de control. Se nota un leve aumento de la influencia en los meses de julio y octubre.

Finalmente, la tabla 41 muestra las sensibilidades de los parámetros asociados a las variables socio-demográficas correspondientes al mes que se quiere predecir.

Tabla 41. Sensibilidad de variables socio-demográficas para el grupo 4.

Sex	A_cta	Edad
-0,40931	0,014948	0,008149

En este grupo gastan más las mujeres que los hombres, ya que el valor del parámetro Sex es negativo. Es el grupo más sensible a esta variable.

Además, mientras más años tienen los individuos y más años poseen la cuenta, es más probable que gasten más. Esto se debe a la influencia positiva tanto de la edad como de los años de la cuenta.

Grupo 5

Tabla 42. Sensibilidad de variables observadas en el mes anterior para el grupo 5.

PA	Dist	Rcomp	Fcomp	Mcomp	NC
-0,00056	0,067586	0,00751	-0,38909	0,00634	0,02631

Los puntos acumulados influyen negativamente en este grupo, lo cual resulta poco intuitivo, ya que es poco probable que una persona gaste menos por el hecho de tener más puntos acumulados. Sin embargo esto se puede explicar tomando en cuenta que quizás este grupo no está tan pendiente de los puntos que va acumulando, y es en realidad debido a las otras variables por lo que gasta menos cuando tiene más puntos acumulados.

A pesar de lo dicho anteriormente, este grupo si se ve afectado positivamente por la distancia al canje más cercano, por lo que si están cerca del canje superior tienden a gastar más, y si están cerca del canje inferior, tienden a gastar menos.

En este grupo hay una influencia positiva por parte de la cantidad de meses que llevan sin comprar (Rcomp), es decir, mientras más meses lleven sin comprar, es más probable que gasten más. Por ejemplo una persona gastará más el próximo mes si lleva 4 meses sin comprar, que si compró el mes pasado.

La frecuencia nos indica que este grupo tiende a controlar sus gastos, por lo que si llevan muchos meses seguidos comprando es probable que gasten menos al mes siguiente.

Como en todos los grupos el monto promedio y el número de compras tienen una influencia positiva sobre el monto gastado.

Tabla 43. Sensibilidad de Dummies de estacionalidad para el grupo 5.

d2	d3	d4	d5	d6	d7	d8	d10	d11	d12
0,0933	0,0804	0,0029	0,1265	-0,000002	-0,1032	0,0006	0,0918	0,1177	0,3103

Con respecto a la estacionalidad (tabla 43), este grupo gasta un poco más en diciembre, pero en casi todos los meses hay una influencia positiva sobre el gasto, excepto a mediados de año, es decir, junio y julio, donde los individuos tienden a gastar menos.

Tabla 44. Sensibilidad de variables socio-demográficas para el grupo 4.

Sex	A_cta	Edad
0,014065	0,001331	0,000438

La variable de sexo influye positivamente en el monto a gastar, por lo que si el individuo es hombre, probablemente gastará más que una mujer con las mismas características.

Las variables de edad y años cuenta tienen una influencia positiva sobre el monto a gastar. Por lo que a mayor edad y años que tienen la cuenta tienden a gastar más.

Grupo 6

Tabla 45. Sensibilidad de variables observadas en el mes anterior para el grupo 6.

PA	Dist	Rcomp	Fcomp	Mcomp	NC
-0,00026	0,14862	-0,50389	0,001749	0,00772	0,024406

Al igual que el grupo 5, los individuos de este grupo probablemente no se fijan demasiado en los puntos que llevan acumulados mes a mes (PA negativo), y si están cerca de un canje superior tienden a gastar más, ya que la variable Dist tiene una influencia positiva sobre el monto gastado.

En este grupo, a medida que va pasando el tiempo sin comprar, es más probable que cuando vuelvan a comprar gasten menos, ya que el Rcomp tiene una influencia negativa.

La frecuencia influye positivamente sobre el monto a gastar, por lo que mientras más seguido han ido a comprar, es más probable que gasten más al próximo mes.

El monto promedio y el número de compras influyen positivamente como en todos, por lo que mientras mayor sea el monto promedio y el número de compras, mayor será el monto gastado al mes siguiente.

Tabla 46. Sensibilidad de Dummies de estacionalidad para el grupo 6.

d2	d3	d4	d5	d6	d7	d8	d10	d11	d12
0,1069	0,0763	0,0361	0,1674	0,1458	0,0745	-0,0004	0,1083	0,1689	0,3695

Este grupo tiende a gastar menos en agosto, para los otros meses la influencia de las dummies es siempre positiva, siendo el mes de diciembre en el que la influencia positiva sobre el monto gastado es mayor.

Tabla 47. Sensibilidad de variables socio-económicas para el grupo 6.

Sex	A_cta	Edad
-0,0376	-0,00936	0,008414

Al igual que en la mayoría de los grupos, la influencia de la variable sex es negativa, lo que indica que las mujeres tienden a gastar más que los hombres en este grupo.

Es el único grupo donde la variable de años cuenta tiene una influencia negativa sobre el monto gastado, por lo que la gente gasta menos al tener más años la cuenta. Sin embargo la variable edad tiene una sensibilidad positiva, por lo que mientras más años tiene el individuo, es más probable que gaste más.

Grupo 7

Tabla 48. Sensibilidad de variables observadas en el mes anterior para el grupo 7.

PA	Dist	Rcomp	Fcomp	Mcomp	NC
0,003674	0,09084	0,012189	1,700684	0,011302	0,046298

Hay una influencia positiva de los puntos acumulados sobre el monto gastado, por lo que a medida que tienen más puntos acumulados, los individuos de este grupo tienden a gastar más. Lo mismo ocurre para la distancia al canje más cercano, ya que mientras más cerca están del canje superior, probablemente gasten más, y mientras más cerca estén del canje inferior, gasten menos.

Con respecto a los meses que llevan sin comprar, mientras más meses lleven sin comprar, es más probable que al volver a comprar gasten más, ya que hay una influencia positiva también de esta variable.

Asimismo la frecuencia influye positivamente en el monto a gastar, por lo que mientras mayor sea ésta, mayor será el gasto a realizar el mes siguiente.

El monto y el número de compras también tienen una influencia positiva sobre el monto a gastar, por lo que mientras mayor sea el monto promedio y el número de compras el mes anterior, mayor será el monto gastado el mes siguiente.

Tabla 49. Sensibilidad de Dummies de estacionalidad para el grupo 7.

d2	d3	d4	d5	d6	d7	d8	d10	d11	d12
0,1841	0,0774	-0,2782	-0,0002	0,1576	0,4510	0,0781	-0,0191	-0,0014	0,9673

En este grupo los individuos gastan más a mediados de mes (julio) y a fin de mes (diciembre), donde mayor es la influencia positiva de la dummy, pero cabe mencionar que en los meses anteriores a éstos, tienden a gastar menos. Esto ocurre en los meses de abril–mayo y octubre–noviembre, donde las sensibilidades son negativas.

Tabla 50. Sensibilidad de variables socio-económicas para el grupo 7.

Sex	A_cta	Edad
-0,16135	0,032433	-0,03426

En este grupo también son las mujeres las que gastan más, ya que hay una influencia negativa de la variable de sexo.

Con respecto a la edad se observa que hay una influencia negativa, por lo que mientras menos años tiene el individuo probablemente gaste menos. A pesar de esto, los años de la cuenta tienen una influencia positiva sobre el monto, por lo que mientras más años posee la cuenta, es más probable que gaste más.

11.4 Anexo 4: Análisis de Errores en Pronósticos Mes a Mes

A continuación se adjuntan las tablas con los distintos errores cometidos en el pronóstico del monto gastado mes a mes. A partir del mes 57 y hasta el mes 62 que corresponde al periodo de validación del modelo.

Mes 57

Es el primer mes de la estimación y se destaca que para el grupo 4 es el que tiene menor MAPE promedio y menor error agregado de todos los meses. En cuanto a MSE destaca el grupo 6, ya que es el mes en que tiene mayor valor en este indicador. Cabe señalar que el MSE en los grupos 5,6 y 7, es más alto ya que estos grupos tienen muchas más personas que los otros dos, por lo que la suma de los errores cuadráticos medios, lógicamente debe ser mayor.

Tabla 51. Error grupal agregado mes 57.

Grupo	Monto Real	Monto Pronosticado	Error Agregado
3	3.522.133	3.563.167	1,2%
4	294.472	113.809	61,4%
5	24.355.389	28.542.546	17,2%
6	10.948.430	8.240.722	24,7%
7	7.859.682	4.794.618	39,0%
Total	46.980.106	45.254.862	3,7%

Tabla 52. MAPE y MSE mes 57.

Grupo	MAPE Individual	MSE Heterogéneo
3	133,60%	504.653
4	82,40%	116.008
5	157,47%	1.947.991
6	114,65%	1.485.331
7	94,71%	1.266.097
Total	116,56%	2.805.710

Mes 58

Tabla 53. Error grupal agregado mes 58.

Grupo	Monto Real	Monto Pronosticado	Error Agregado
3	3.532.875	4.720.829	33,6%
4	483.270	34.798	92,8%
5	39.316.425	34.740.853	11,6%
6	14.776.290	10.183.577	31,1%
7	9.854.057	12.233.096	24,1%
Total	67.962.917	61.913.154	8,9%

Tabla 54. MAPE y MSE mes 58.

Grupo	MAPE Individual	MSE
3	429,82%	669.003
4	99,02%	177.917
5	143,59%	2.481.851
6	113,07%	1.431.795
7	125,30%	1.393.335
Total	182,16%	3.260.402

Este mes corresponde a diciembre, y se puede apreciar que todos los grupos son subestimados, a excepción del grupo 3, lo que demuestra que en este mes los individuos gastan mucho más.

Cabe mencionar que el MSE en este mes es el más alto tanto para el grupo 5 y 7, como a nivel global, lo que indica que el error de pronóstico para estos grupos y a nivel global es mayor en este mes.

A nivel individual, se aprecia que el MAPE promedio en el grupo 3 es muy alto, por lo que la predicción de los individuos no es muy precisa en este mes, lo que a su vez provoca el error agregado de pronóstico mencionado anteriormente, que es el más alto observado para este grupo con un 33,6%.

Mes 59

Tabla 55. Error grupal agregado mes 59.

Grupo	Monto Real	Monto Pronosticado	Error Agregado
3	3.646.773	3.523.125	3,4%
4	304.482	23.010	92,4%
5	25.736.076	25.660.925	0,3%
6	10.160.864	7.064.494	30,5%
7	6.207.715	4.767.363	23,2%
Total	46.055.910	41.038.917	10,9%

Tabla 56. MAPE y MSE mes 59.

Grupo	MAPE Individual	MSE
3	288,39%	629.682
4	99,46%	125.331
5	156,74%	1.962.982
6	104,02%	1.168.316
7	98,44%	801.858
Total	149,41%	2.504.685

Mes 60

Tabla 57. Error grupal agregado mes 60.

Grupo	Monto Real	Monto Pronosticado	Error Agregado
3	3.292.433	3.916.178	18,9%
4	464.127	27.661	94,0%
5	26.228.754	28.173.148	7,4%
6	9.877.936	7.873.475	20,3%
7	7.079.435	5.756.260	18,7%
Total	46.942.685	45.746.722	2,5%

Tabla 58. MAPE y MSE mes 60.

Grupo	MAPE Individual	MSE
3	366,14%	948.931
4	99,75%	158.921
5	135,80%	1.706.663
6	118,24%	1.455.721
7	116,87%	1.109.594
Total	167,36%	2.681.184

Mes 61

Tabla 59. Error grupal agregado mes 61.

Grupo	Monto Real	Monto Pronosticado	Error Agregado
3	3.621.851	3.643.195	0,6%
4	368.577	24.582	93,3%
5	23.146.753	27.920.947	20,6%
6	10.155.017	7.631.553	24,8%
7	7.295.425	5.229.045	28,3%
Total	44.587.623	44.449.321	0,3%

Tabla 60. MAPE y MSE mes 61.

Grupo	MAPE Individual	MSE Heterogéneo
3	294,90%	684.589
4	99,10%	201.944
5	158,55%	1.910.077
6	102,87%	1.216.320
7	129,13%	1.297.419
Total	156,91%	2.705.655

Mes 62

Tabla 61. Error grupal agregado mes 62.

Grupo	Monto Real	Monto Pronosticado	Error Agregado
3	3.982.605	3.693.852	7,3%
4	655.458	37.787	94,2%
5	23.680.916	25.917.881	9,4%
6	10.065.332	7.346.692	27,0%
7	7.890.657	3.683.745	53,3%
Total	46.274.968	40.679.958	12,1%

Tabla 62. MAPE y MSE mes 62.

Grupo	MAPE Individual	MSE Heterogéneo
3	82,99%	740.939
4	99,66%	276.851
5	157,72%	1.867.979
6	112,09%	1.441.478
7	96,97%	1.311.616
Total	109,89%	2.813.037

Se observa en general que el grupo número 3, que es el de mayor interés por ser el más valioso para la empresa, presenta el mayor error de estimación a nivel individual (MAPE), sin embargo, a nivel grupal es el más preciso (Error Agregado).

El MSE de este grupo es sólo mayor que el del grupo 4, y esto es debido a la gran diferencia que hay en la cantidad de montos gastados entre estos grupos de pocos individuos y a la mala estimación de los individuos a nivel individual.

11.5 Anexo 5: Variación en 1% de la Tasa de Retención

Se calcula la variación del CLV para distintas tasas de retención, en este caso serán 99%, 98%, 97% y 96%, con el fin de verificar la diferencia porcentual de una variación de 1% en la tasa de retención. Los resultados se muestran en la tabla 63.

Tabla 63. Variación de CLV promedio al disminuir en 1% la tasa de retención.

Segmento	CLV (99%)	CLV (98%)	CLV (97%)	CLV (96%)
3	1.184.108	1.143.048	1.103.358	1.064.996
5	484.217	468.393	453.061	438.207
7	274.857	265.660	256.763	248.155
6	250.942	242.706	234.731	227.008
4	91.653	88.240	84.947	81.771

Usando la misma fórmula para calcular la diferencia porcentual, mostrada en el cuerpo de esta memoria (sección 11), se obtienen los resultados que aparecen en la tabla 64, para los mismos porcentajes de la parte anterior, es decir, la diferencia entre 99% y 98%, entre 98% y 97% y entre 97% y 96%, finalmente aparece un promedio de estas tres diferencias.

Tabla 64. Diferencia porcentual al cambio de un 1% en la tasa de retención.

Segmento	DIF 99%-98%	DIF 98%-97%	DIF 97%-96%	Promedio
3	3,47%	3,47%	3,48%	3,47%
5	3,27%	3,27%	3,28%	3,27%
7	3,35%	3,35%	3,35%	3,35%
6	3,28%	3,29%	3,29%	3,29%
4	3,72%	3,73%	3,74%	3,73%

Donde se observa que en promedio, al disminuir en 1% la tasa de retención, el grupo 3 disminuye en 3.47% su valor, el grupo 5 en un 3.27%, el grupo 7 en un 3.35%, el grupo 6 en un 3.29% y el grupo 4 en un 3.73%.

Además en la tabla 65 se muestra que los grupos considerados más valiosos disminuyen en mayor cantidad el valor de su ciclo de vida (CLV) en el caso de una disminución en su tasa de retención.

Tabla 65. Disminución en CLV al disminuir 1% la tasa de retención.

Segmento	99%-98%	98%-97%	97%-96%
3	41.060	39.690	38.362
5	15.824	15.332	14.854
7	9.197	8.898	8.607
6	8.236	7.975	7.723
4	3.414	3.293	3.175