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

MODELO DE FUGA Y POLÍTICAS DE RETENCIÓN EN UNA EMPRESA DE MEJORAMIENTO DEL HOGAR

MEMORIA PARA OPTAR AL TÍTULO DE INGENIERO CIVIL INDUSTRIAL

ANA ISABEL CASTILLO BELDAÑO

PROFESOR GUÍA:
ALEJANDRA PUENTE CHANDÍA

MIEMBROS DE LA COMISIÓN:
LUIS ABURTO LAFOURCADE
CLAUDIO PIZARRO TORRES

SANTIAGO DE CHILE
AGOSTO 2014

RESUMEN DE LA MEMORIA PARA OPTAR AL
TÍTIULO DE: Ingeniero Civil Industrial
POR: Ana Isabel Castillo Beldaño
FECHA: 24/08/2014
PROFESOR GUÍA: Alejandra Puente Chandía

MODELO DE FUGA Y POLÍTICAS DE RETENCIÓN EN UNA EMPRESA DE MEJORAMIENTO DEL HOGAR

El dinamismo que ha presentado la industria del mejoramiento del hogar en el último tiempo, ha llevado a que las empresas involucradas deban preocuparse por entender el comportamiento de compra de sus consumidores, ya que no solo deben enfocar sus recursos y estrategias en capturar nuevos clientes sino también en la retención de éstos.

El objetivo de este trabajo es estimar la fuga de clientes en una empresa de mejoramiento del hogar con el fin de generar estrategias de retención. Para ello se definirán criterios de fuga y se determinarán probabilidades para gestionar acciones sobre una fracción de clientes propensos a fugarse.

Para alcanzar los objetivos mencionados, se trabajará sólo con clientes que forman parte de la cartera de un vendedor y se hará uso de las siguientes herramientas: estadística descriptiva, técnica RFM y la comparación de los modelos predictivos Árbol de decisión y Random Forest, donde la principal diferencia de estos últimos es la cantidad de variables y árboles que se construyen para la predicción de las probabilidades de fuga.

Los resultados obtenidos entregan tres criterios de fuga, de manera que un cliente es catalogado como fugado cuando supera cualquiera de las cotas máximas, es decir, 180 días para el caso del recency, 20 para R/F o una variación de monto menores al -80%, por lo que la muestra queda definida con un 53,9% de clientes fugados versus un 46,1% de clientes activos. Con respecto a los modelos predictivos se tiene que el Árbol de decisión entrega un mejor nivel de certeza con un 84,1% versus un 74,7% del Random Forest, por lo que se eligió el primero obteniendo a través de las probabilidades de fuga 4 tipos de clientes: Leales (37,9%), Normales (7,8%), Propensos a fugarse (15,6%) y Fugados (38,7%).

Se tiene que las causas de fuga corresponden a largos períodos de inactividad, atrasos en los ciclos de compras y una disminución en los montos y números de transacciones al igual que un aumento en el monto de transacciones negativas aludidas directamente a devoluciones y notas de crédito, por lo que las principales acciones de retención serían promociones, club de fidelización, descuentos personalizados y mejorar gestión en despachos y niveles de stock para que el cliente vuelva efectuar una compra en un menor plazo.

Finalmente, a partir de este trabajo, se concluye que al retener 5% de clientes de probabilidades entre [0,5 y 0,75] y con el 50% de los mayores montos de transacciones se obtienen ingresos por USD $205 mil en 6 meses, representando el 5,5% de los clientes. Se propone validar este trabajo en nuevos clientes, generar alguna encuesta de satisfacción y mejorar el desempeño de los vendedores con una optimización de cartera.

AGRADECIMIENTOS

Quiero agradecer en primer lugar a mis padres, Ana y Enrique, por darme su confianza, valores y eterno amor. En especial a mi querida viejita, con quien he vivido intensamente mi etapa universitaria, apoyándome en los desafíos, tristezas y alegrías que ha llevado este camino.

También de igual manera a mis hermanos, Leo y Marco, ya que siempre han estado presente cuando he necesitado algún un consejo. Gracias Leo por decirme el 1° año de universidad "Que era capaz de ser Ingeniera y de mucho más" esas palabras fueron claves para llegar al día de hoy. A mis cuñadas, Eve e Isa, por tener siempre una palabra de aliento en nuestras conversaciones, y por supuesto a mis sobrinos, Joaquín, Francisco y Santiago, quienes me permiten sacar lo mejor de mí. A mis tíos y familia, por su eterno apoyo, amor y comprensión.

A mi profesor Luis por su infinita paciencia, ya que sin su ayuda este trabajo no hubiera sido posible, y más que todo por su apoyo en momentos difíciles, conversaciones y consejos de vida. Al profesor Marcel por la posibilidad de trabajar con él durante el último tiempo y descubrir que era capaz de grandes desafíos.

A mis compañeros de memoria Nico, Alan, Pancho y Tomás por hacer que el tiempo de este trabajo haya sido cada día más agradable y por su gran ayuda. ¡Cómo olvidar esas clases y/o presentaciones del F!

A los chicos de Penta Analytics, especialmente al área de Data Mining y Andrés, por su buena onda durante el tiempo que duro este desafío.

A mis partners industriales, Belén, Cata, Lete y Chehade, quienes me apoyaron fielmente durante este proceso, por quererme tal cual soy y acompañarme en mis locuras. Gracias también a mis amigos Alex, Pablo, Rodolfo, Rose, Jorge, Esteban y Carla, por aparecer en esta etapa. Los quiero mucho.

Y finalmente mis eternos amigos con quien comenzamos este proceso y vivimos distintas aventuras Steph, Vale, Pía, Dany, Berry, Clau, Rita, Tania, Karl y Verdugo. Se lo lejos que llegarán como profesionales y personas.

Gracias a todos aquellos con quien compartí algún momento de esta linda etapa, la cual ha sido lo mejor hasta este minuto. Lo importante es soñar en grande, porque SIEMPRE se puede.

TABLA DE CONTENIDO

1. INTRODUCCIÓN vii
1.1 Antecedentes generales del Retail 1
1.2 Antecedentes industria y empresa 2

2. DESCRIPCIÓN DEL PROYECTO Y JUSTIFICACIÓN 5

3. OBJETIVOS **6**
3.1 Objetivo General 6
3.2 Objetivos Específicos 6

4. MARCO CONCEPTUAL 6
4.1 Definición de criterio de fuga 7
4.2 Modelo de predicción de fuga de clientes 7
4.2.1 Árbol de decisión 8
4.2.2 Random Forest 9

5. METODOLOGÍA 11

6. ALCANCES 12

7. RESULTADOS ESPERADOS 13

8. SELECCIÓN Y PREPROCESAMIENTO DE DATOS 13
8.1 Descripción de la base de datos en general 13
8.2 Preprocesamiento de datos 14
8.3 Variables calculadas 14
8.3.1 Análisis univariado de variables calculadas. 16
8.3.2 Análisis multivariado de variables calculadas. 22
8.3.3 Transformación de variables 24

9. DEFINICIÓN DE CRITERIO DE FUGA 24
9.1 Criterios de fuga 25

10. MODELO PREDICTIVO DE FUGA 28
10.1 División en entrenamiento y testeo 29
10.2 Árbol de decisión 30
10.3 Random Forest 31
10.4 Análisis de sensibilidad: Random Forest 34
10.5 Probabilidades de fuga 35
10.6 Análisis de resultados 36

10.7 Resultados complementarios 37
10.7.1 Árbol de decisión sin variables RFM 37
10.7.2 Comparación de modelos con base clientes 2014 39

11. CLASIFICACIÓN DE CLIENTES 39

12. CAUSAS DE FUGA Y ACCIONES DE RETENCIÓN DE CLIENTES 41
12.1 Análisis económico de resultados 47

13. CONCLUSIONES Y RECOMENDACIONES 48
13.1 Sobre los criterios de fuga 48
13.2 Sobre los modelos de predicción de fuga 48
13.3 Sobre los resultados obtenidos 49
13.4 Recomendaciones y trabajos futuros 49

14. BIBILIOGRAFÍA 51

15. ANEXOS 53

ÍNIDICE DE TABLAS

Tabla 1: Distribución de sucursales a lo largo de Chile. 3
Tabla 2: Información sobre clientes año 2013. 5
Tabla 3: Descripción de variables básicas. 13
Tabla 4: Descripción de variables calculadas. 15
Tabla 5: Resumen del diagrama de caja de los clientes - Recency. 16
Tabla 6: Resumen del diagrama de caja de los clientes Carterizados - Frecuencia. 18
Tabla 7: Resumen del diagrama de caja de los clientes Carterizados - Monto Promedio. 20
Tabla 8: Resumen del diagrama de caja de los clientes Carterizados - R/F. 21
Tabla 9: Correlación entre variables. 23
Tabla 10: Número y distribución de clientes que experimento variaciones. 26
Tabla 11: Cotas máximas para definición de cliente fugado. 27
Tabla 12: Variables utilizadas en el modelo predictivo de fuga. 29
Tabla 13: Matriz de confusión - Testeo para Árbol de decisión. 30
Tabla 14: Matriz de confusión - Testeo para Random Forest. 31
Tabla 15: Análisis de sensibilidad del modelo Random Forest. 34
Tabla 16: Matriz de confusión - Testeo para Árbol de decisión sin RFM. 38
Tabla 17: Clasificación de clientes Carterizados de acuerdo a su probabilidad de fuga 40
Tabla 18: Principales variables de los diferentes segmentos. 41
Tabla 19: Clasificación de categorías de productos. 44
Tabla 20: Análisis económico de los clientes propensos a fugarse. 47
Tabla 21: Matriz de confusión - Entrenamiento para Árbol de decisión. 63
Tabla 22: Matriz de confusión - Entrenamiento para Random Forest. 64
Tabla 23: Análisis de sensibilidad Random Forest - Matriz de confusión 50 árboles. 65
Tabla 24: Análisis de sensibilidad Random Forest - Matriz de confusión 100 árboles. 66
Tabla 25: Análisis de sensibilidad Random Forest - Matriz de confusión 200 árboles. 66
Tabla 26: Análisis de sensibilidad Random Forest - Matriz de confusión 500 árboles. 67
Tabla 27: Matriz de confusión - Entrenamiento para Árbol de decisión sin RFM. 67

ÍNIDICE DE ILUSTRACIONES

Ilustración 1: Distribución de ventas en la industria del Retail año 2012.....1
Ilustración 2: Participación de mercado del sector de mejoramiento del hogar 2012.....2
Ilustración 3: Evolución de los clientes en empresa de mejoramiento del hogar 4
Ilustración 4: Árbol general de decisión.....8
Ilustración 5: Base de datos de clientes Carterizados.....14
Ilustración 6: Diagrama de caja de los clientes Carterizados - Recency.....16
Ilustración 7: Distribución de los clientes Carterizados - Recency. 17
Ilustración 8: Diagrama de caja de los clientes Carterizados - Frecuencia. 18
Ilustración 9: Distribución de clientes Carterizados - Frecuencia.....19
Ilustración 10: Diagrama de cajas de los clientes Carterizados - Monto Promedio.....19
Ilustración 11: Distribución de clientes Carterizados - Monto Promedio.....20
Ilustración 12: Diagrama de caja de los clientes Carterizados - R/F. 21
Ilustración 13: Distribución de clientes Carterizados - R/F. 22
Ilustración 14: Transformación de la variable Número de transacciones 24
Ilustración 15: Distribución de clientes fugados de acuerdo al recency. 25
Ilustración 16: Distribución de clientes fugados de acuerdo al R/F. 26
Ilustración 17: Distribución de clientes que experimentaron variación monetaria en ambos periodos.....27
Ilustración 18: Distribución de clientes Carterizados fugados en la muestra seleccionada. 28
Ilustración 19: Curva de ganancia - Árbol de decisión. 30
Ilustración 20: Resultados del Árbol de decisión.....31
Ilustración 21: Curva de ganancia - Random Forest. 32
Ilustración 22: MDG Random Forest. 33
Ilustración 23: MDA Random Forest. 33
Ilustración 24: Probabilidades de fuga del modelo Árbol de decisión.....35
Ilustración 25: Probabilidades de fuga del modelo Random Forest. 36
Ilustración 26: Resultados del Árbol de decisión sin RFM.....38
Ilustración 27: Curvas de ganancias con BBDD de clientes 2014.....39
Ilustración 28: Clientes activos versus fugados por segmento.....40
Ilustración 29: Reglas de asignación de probabilidades de fuga.....42
Ilustración 30: Distribución de monto total de clientes y/o fugados Carterizados. 45
Ilustración 31: Distribución de monto negativo de clientes propensos a fugarse. 46
Ilustración 32: Distribución de porcentaje de monto de notas de créditos.....46
Ilustración 33: Distribución de clientes de acuerdo a su número de transacciones.....53
Ilustración 34: Distribución de clientes de acuerdo a su número de transacciones- día.53
Ilustración 35: Distribución de clientes de acuerdo a su antigüedad.....54
Ilustración 36: Distribución de clientes de acuerdo a su máxima inactividad. 54
Ilustración 37: Distribución de clientes de acuerdo al monto total.....55
Ilustración 38: Distribución de clientes de acuerdo a la variación de monto. 55
Ilustración 39: Distribución de clientes de acuerdo a su monto total negativo.....56
Ilustración 40: Distribución de clientes de acuerdo al número de transacciones negativas.....56
Ilustración 41: Distribución de clientes de acuerdo al porcentaje de monto de transacciones negativas.....57
Ilustración 42: Distribución de clientes de acuerdo a su región.....57
Ilustración 43: Distribución de clientes de acuerdo al número de devoluciones.....58

Ilustración 44: Distribución de clientes de acuerdo al porcentaje de monto de devoluciones. 58
Ilustración 45: Distribución de clientes de acuerdo al número de notas de créditos. 59
Ilustración 46: Distribución de clientes de acuerdo al porcentaje de monto de notas de créditos....59
Ilustración 47: Distribución de clientes de acuerdo al número de transacciones negativas tipo retail....60
Ilustración 48: Distribución de clientes de acuerdo a porcentaje de monto de retail. 60
Ilustración 49: Distribución de clientes de acuerdo a su giro comercial. 61
Ilustración 50: Distribución de clientes de acuerdo al porcentaje de monto de obra gruesa. 61
Ilustración 51: Distribución de clientes de acuerdo al porcentaje de monto de obra intermedia. 62
Ilustración 52: Distribución de clientes de acuerdo al porcentaje de monto de terminaciones....62
Ilustración 53: Distribución de clientes de acuerdo al porcentaje de monto de otros. 63
Ilustración 54: Curva AUC- Entrenamiento Árbol de decisión. 64
Ilustración 55: Curva AUC- Entrenamiento Random Forest. 65
Ilustración 56: Curva AUC- Entrenamiento Árbol de decisión sin RFM. 68

1. INTRODUCCIÓN

1.1 Antecedentes generales del Retail

Durante los últimos años, la industria de retail ha experimentado un fuerte crecimiento gracias a las condiciones existentes en el país, es decir, una estabilidad política, una tasa de crecimiento de 4,1% y una tasa de inflación del 3% registrada en el año 2013, han fomentado el desarrollo de esta industria.

El PIB presentando por el sector del retail fue de USD $268 mil millones en el año 2012, el cual corresponde al 21% del PIB y se proyectan que las ventas se eleven en 14,5% entre 2011 y 2015 [11].El aporte más importante fue producido por el sector de supermercados y ferreterías, donde las ventas de los distintos sectores se observan en el Ilustración 1.

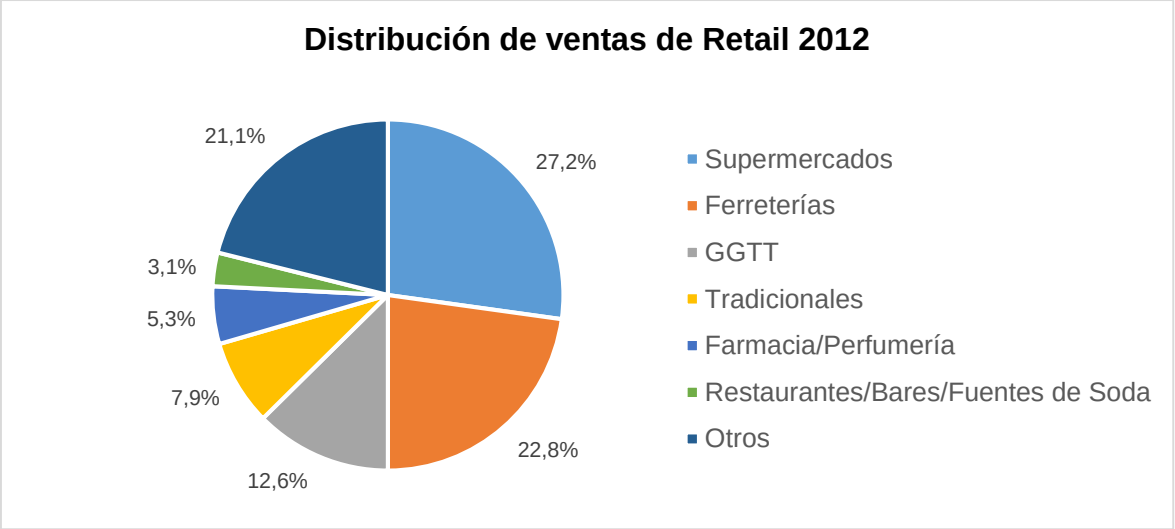

Ilustración 1: Distribución de ventas en la industria del Retail año 2012.
Fuente: Nielsen AC.

El dinamismo que presenta esta industria provoca una alta competitividad, por lo que es fundamental entender el comportamiento del consumidor con el fin de obtener ventajas frente a sus competidores. El desarrollo de modelos o herramientas que permitan generar conocimientos o acciones sobre los clientes, tal como evitar que se fuguen de algún sistema, es un gran desafío para las distintas empresas que forman parte de ella, siendo el eje central de este trabajo.

El prevenir la fuga de clientes permite generar estrategias de retención de manera que los consumidores de bienes o servicios no migren a la competencia, ya que parte del éxito de estas empresas no solo corresponde a enfocarse en captar potenciales nuevos clientes sino que también en retener y satisfacer a los que ya se encuentran presente.

Este trabajo se centrará en la industrial del retail, específicamente en el sector de mejoramiento del hogar y construcción, donde los clientes poseen una relación del tipo

no-contractual, es decir, no existe algún contrato de por medio que le indique cuando la relación terminará, por lo que la fuga de cliente no es algo conocido para la empresa.

1.2 Antecedentes industria y empresa

La industria de mejoramiento del hogar y construcción representa el 22,8% del PIB el año 2012 y vio incrementadas fuertemente sus ventas a partir del terremoto y maremoto sufrido en Chile el 27 de Febrero del año 2010 y los auges inmobiliarios registrados en el año 2011, llegando incluso a facturar más de USD $10.000 millones en el año 2012.

Este sector del retail está conformado por 6 agentes, donde el líder del mercado es Homecenter Sodimac con más de 70 sucursales a lo largo del país. La participación de mercado de los diferentes actores de esta industria se observa en la Ilustración 2.

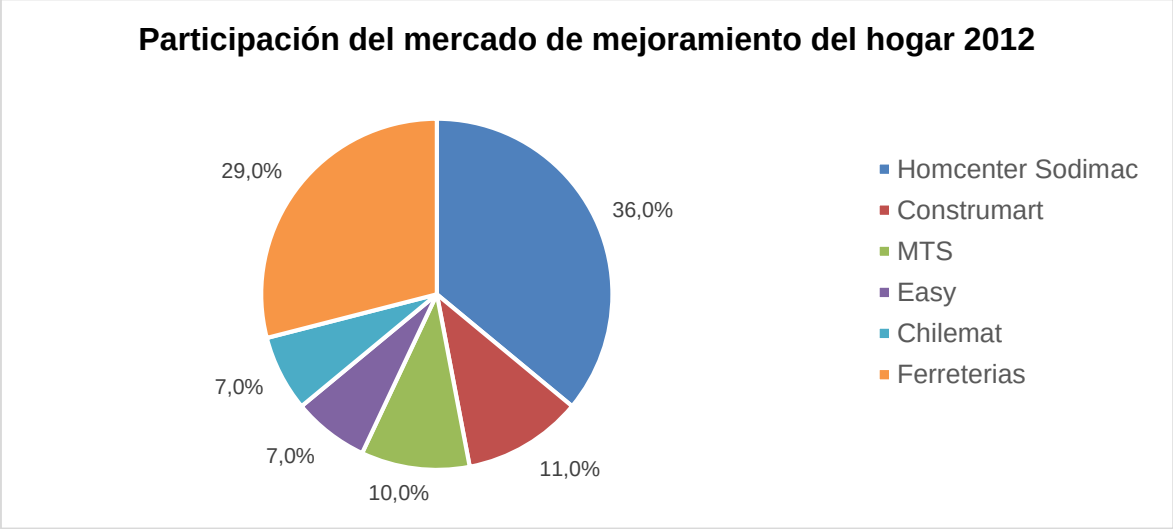

Ilustración 2: Participación de mercado del sector de mejoramiento del hogar y construcción 2012. Fuente: AC Nielsen.

La empresa de mejoramiento del hogar y construcción con la que se trabajó posee 31 sucursales a lo largo de Chile, desde Antofagasta a Puerto Montt, concentrado sus puntos de ventas en la Región Metropolitana con 12 tiendas. La distribución de las tiendas en el territorio nacional se observa en la Tabla 1.

Zona Norte	Zona Central	Región Metropolitana	Zona Sur
Calama	Valparaíso	Puente Alto	Linares
Antofagasta	Viña del Mar	Cerrillos	Chillán
Copiapó	El Belloto	El Llano	Los Ángeles
La Serena	Quillota	Ochagavía	Concepción
	Los Andes	Quilicura	Temuco
	Rancagua	Maipú	Osorno
	Curicó	Alto Las Condes	Puerto Montt
	Talca	La Florida	
		Costanera Center	
		La Dehesa	
		La Reina	

Tabla 1: Distribución de sucursales a lo largo de Chile.
Fuente: Elaboración Propia.

La superficie de estas sucursales van desde los 6.800 a 14.000 $\mathbf{m^2}$, con más de 35.000 artículos relacionados al hogar y la construcción agrupados en categorías como materiales de construcción, pinturas, herramientas, entre otros y más 3.500 empleados enfocados en brindar y asesorar a los clientes con el fin de satisfacer sus expectativas y necesidades.

Esta empresa de mejoramiento del hogar y construcción facturó más de USD $350 millones el año 2012, y observó oportunidades de negocio en el mercado extranjero, extendiendo sucursales a países como Argentina a partir del año 1993 y Colombia en el año 2008, llegando hoy a 40 y 9 sucursales respectivamente.

Tiene 4 líneas de negocios:

- Hogar: Enfocado en clientes que compran artículos en las diferentes sucursales del país, orientados a construir, remodelar, reparar, mejorar y decorar su hogar.

- Servicios: Enfocado en clientes que arriendan herramientas y maquinarias y/o contratan servicios con personal calificados como iluminación, gasfiterías, cotizaciones y asesoría de diseñadores, etcétera.

- Construcción: Enfocado en clientes que trabajan en la construcción, brindándoles bajos precios, productos y accesorios como también asesoría en proyectos, creando un espacio para ellos llamado “Mundo Experto”.

- Venta a empresas: Enfocado en clientes como empresas constructoras, contratistas y profesionales que necesitan grandes volúmenes de materiales y otros.

Dentro de la línea de negocios de ventas a empresas, existen tres clasificaciones de clientes:

- Clientes Carterizados: Este tipo de cliente corresponde a una empresa constructora, contratista o profesional que forma parte de una cartera de clientes de algún vendedor en terreno, el cual debe visitarlo en su lugar de trabajo o en la zona donde se está llevando a cabo el proyecto, con el fin de registrar el pedido de materiales y/o productos, los cuales pueden ser despachados en cualquier sucursal del país.

- Clientes Mesón: Este tipo de cliente corresponde a una empresa constructora, contratista o profesional, los cuales son atendidos por cualquier vendedor en algunas de las sucursales de la empresa.

- Clientes No Carterizados: Este tipo de cliente corresponde a una empresa constructora, contratista o profesional que no forma parte de alguna cartera de ningún vendedor, de manera que la visita a terreno puede ser hecha por cualquiera. Los clientes No Carterizados son “prospectos de clientes”, es decir, empresas que recién se están relacionando con este mercado o están en proceso de aprobación para la apertura de líneas de créditos en las sucursales de esta empresa de mejoramiento del hogar y construcción.

La evolución de estos tres tipos de clientes se observa en la Ilustración 3:

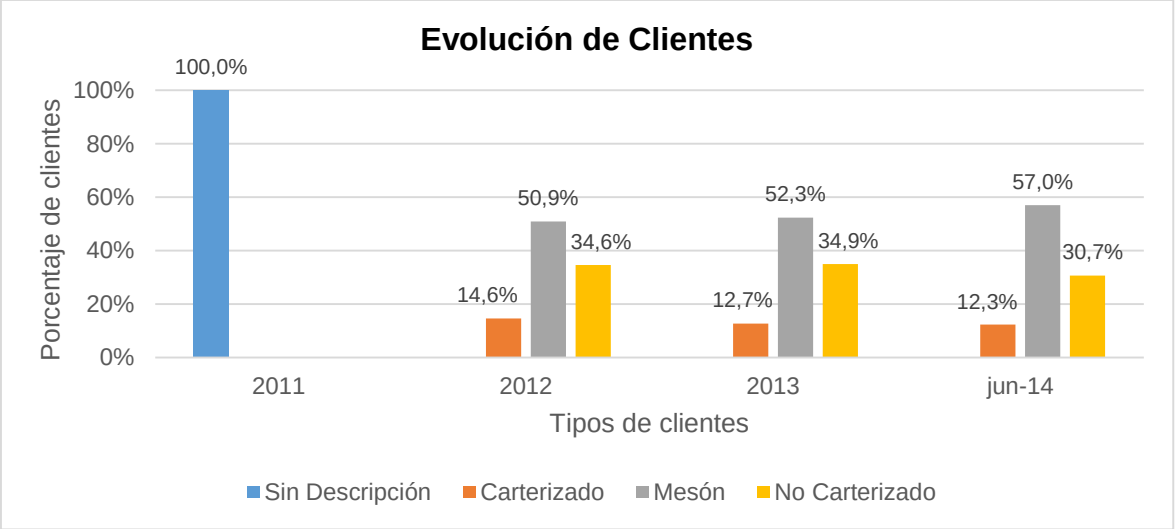

Ilustración 3: Evolución de los clientes en empresa de mejoramiento del hogar y construcción. Fuente: Elaboración Propia.

El año 2011 no registra una clasificación de los clientes a diferencia del año 2012, donde los clientes que más han experimentado un aumento, son los clientes Mesón. Durante el último tiempo el aumento de los clientes Mesón proviene de los clientes No Carterizados, ya que el número de clientes Carterizado se ha mantenido relativamente constante. La participación de estos clientes en orden decreciente corresponde a: Clientes Mesón, No Carterizados y Carterizados.

En la Tabla 2 se observa la información obtenida para estos diferentes tipos de clientes durante el año 2013.

Tipo de Cliente	Número de Clientes	Ticket promedio por transacción	Ingresos aportados
Carterizados	3.496	$ 302.804	69,6%
Mesón	14.399	$ 157.008	15,8%
No Carterizado	9.612	$ 94.843	14,5%

Tabla 2: Información sobre clientes año 2013.
Fuente: Elaboración Propia.

El total de clientes de ventas a empresas registrados el año 2013 fue de 27.507, donde la mayor participación corresponde a clientes Mesón con un 52,3%, seguido de clientes No Carterizados con un 34,9% y finalmente los clientes Carterizados con un 12,7%, a pesar de que éstos últimos representen un bajo porcentaje de clientes en comparación al resto, son quienes aportan el 69,6% de los ingresos a esta línea de negocio. Se trabajó con clientes Carterizados, ya que una de las preocupaciones latentes de la compañía es entender el comportamiento de compra de estos clientes con el fin de prevenir la fuga y migración a la competencia, aplicando distintas estrategias que permitan retenerlos.

2. DESCRIPCIÓN DEL PROYECTO Y JUSTIFICACIÓN

El presente trabajo se realiza en la empresa Penta Analytics, la cual se encarga de llevar a cabo asesorías con respecto a la inteligencia de negocios, analizando datos, utilizando técnicas de data mining, para aportar con valiosa información a los distintas clientes con el objetivo de aumentar su rentabilidad brindando asesoría en temas de eficiencia operacional (gestión de personas, detección de fraude, detención de fuga de clientes, logística e inventario), gestión de retail (gestión de shopper, gestión de ventas, gestión de promociones), entre otros.

Particularmente, se trabajó con una empresa de mejoramiento del hogar, específicamente con su línea de negocio venta a empresas y clientes Carterizados, en una relación no contractual, donde los clientes que conforman este segmento aportaron con ingresos por más de $56 mil millones en el año 2013. La selección de los clientes Carterizados es debido a que representan aproximadamente el 70% de los ingresos registrados en el año 2013 además de poseer un ticket promedio por transacción sobre los $300.000 y al ser un número reducido de clientes en comparación a clientes Mesón y No Carterizados, es decir, el 12,7% del total de clientes, es más factible identificar características relevantes que permitan estudiar la fuga de alguno de éstos.

Esta empresa en la actualidad no posee una definición de clientes fugado ni un modelo de predicción de fuga que permita anticipar que clientes se irán del sistema, de manera que se generan pérdidas para la compañía, debido a que no logra efectuar estrategias que le permitan retenerlos y aún más identificar los motivos por los cuales migran a la competencia, de forma que si se retuviese el 1% de los clientes Carterizados registrados en el año 2013, esta empresa podría beneficiarse con ingresos por más USD $990 mil dólares anualmente.

La complejidad de esta trabajo se enfoca principalmente en que se desarrolló en la industria de mejoramiento del hogar y construcción, siendo este un sector de retail donde el comportamiento de compra no es similar al de algún cliente en supermercado o tienda por departamento, ya que depende de factores externos como estabilidad política, social, económica, entre otros, para llevar a cabo proyectos o construcciones, por lo que pueden existir largos períodos donde no se efectúen compras.

En base a lo anterior, este trabajo responde a la preguntas de ¿quiénes son estos clientes?, ¿quiénes se fugan?, ¿por qué se fugan? , ¿cómo se retienen?, de manera que se llevará cabo un modelo predictivo de fuga que permita solucionar los problemas ya mencionados. Las técnicas utilizadas para esto serán un Árbol de decisión y Random Forest, siendo este último nuevo en la resolución de problemas en data mining. Para ello se utilizarán datos del tipo transaccional tales como identificación del cliente, monto total, frecuencia, monto de transacciones negativas, entre otras.

3. OBJETIVOS

3.1 Objetivo General

Estimar fuga de clientes de una empresa de mejoramiento del hogar con el fin de generar estrategias de retención.

3.2 Objetivos Específicos

- Definir el criterio de cliente fugado en empresa de mejoramiento del hogar.
- Desarrollar un modelo predictivo de fuga de clientes.
- Definir una clasificación dentro de los clientes Carterizados y determinar la probabilidad de fuga para dicha clasificación, identificando los motivos que producen el riesgo.
- Definir acciones y campañas para retener a los clientes propensos a fugarse.

4. MARCO CONCEPTUAL

Debido a que la empresa no posee la definición de un cliente fugado es importante determinar este concepto con el objetivo de utilizarlo en el desarrollo del modelo predictivo de fuga.

4.1 Definición de criterio de fuga

La relación que posee la empresa con este tipo de clientes es de carácter no contracual, por lo que para la definición de criterio de fuga se utilizará el método RFM y la combinación de sus variables.

- RFM

Este método estudia el comportamiento de compra de los clientes, ya que se basa en el análisis de datos del pasado para identificar con alguna certeza el futuro y es una forma de segmentar a los clientes, debido a que permite reducir los datos transacciones y caracterizar en base a tres variables:

- Recency (R): Esta variable es el tiempo que transcurre desde la última compra.
- Frecuency (F): Esta variable mide la velocidad con que el cliente compra, determinada como el tiempo transcurrido entre transacciones.
- Monetary value (M): Esta variable se mide como el monto promedio de las transacciones dentro de un periodo de tiempo [2].

Con estas variables mencionadas, se diseñan tres posibles métodos para definir a un cliente como fugado:

1. Criterio Recency

Para cada cliente es posible determinar el tiempo transcurrido desde su última compra, de manera que si define un valor máximo para éste y algún cliente lo sobrepasa, este se considerará como fugado.

2. Criterio R/F

A través de este criterio se busca comparar la relación entre recency y la frecuencia, ya que si la razón es menor a 1, significa que el cliente ha realizado su última compra en un tiempo menor al habitual, por lo tanto es poco probable que se fugue [1]. Para ello se determinará un máximo valor para R/F, de manera que si algún cliente lo sobrepasa se considerará como fugado.

3. Criterio Variación de Monto

Para determinar este criterio se deberá observar la variación de monto que se obtenga por cliente al comparar los dos últimos trimestres antes de su última fecha de compra de manera que si sobrepasa un máximo valor se considerará como fugado. Este criterio se considera, ya que el cliente puede ver afectado su comportamiento de compra previamente a fugarse en la disminución de sus montos por transacciones.

4.2 Modelo de predicción de fuga de clientes

La predicción de la fuga de clientes se determinará a través del método Random Forest, el cual permitirá clasificar en base a probabilidades de fuga los distintos tipos de clientes

y entregará las variables relevantes por los cuales se producen estos motivos. Los resultados que se obtengan con este modelo se compararan con un Árbol de decisión para determinar si efectivamente esta técnica es mejor en la predicción de clientes fugados.

4.2.1 Árbol de decisión

Esta técnica consiste en la generación de un árbol que intenta explicar o predecir una variable dependiente, donde esta puede ser del tipo categórica o continua, siendo árboles de clasificación o de regresión.

El método consiste en la división sucesiva del conjunto de datos en subgrupos sobre el total de variables, ya que se genera un método recursivo donde cada grupo de datos es dividido en dos subgrupos en base a una regla que entrega dos valores. Para la división sólo se utiliza una variable que se escoge cada vez mediante la exploración exhaustiva de todas las posibilidades de forma que los subgrupos resulten lo más homogéneos posible [12].

Un árbol de decisión está compuesto por un nodo parental, el cual puede dar origen a distintos nodos hijos conectados entre sí a través de ramas. Las distintas ramas del árbol terminarán en un nodo hoja cuando se haya cumplido el criterio de parada. En la Ilustración 4 se observa la estructura base de un árbol de decisión.

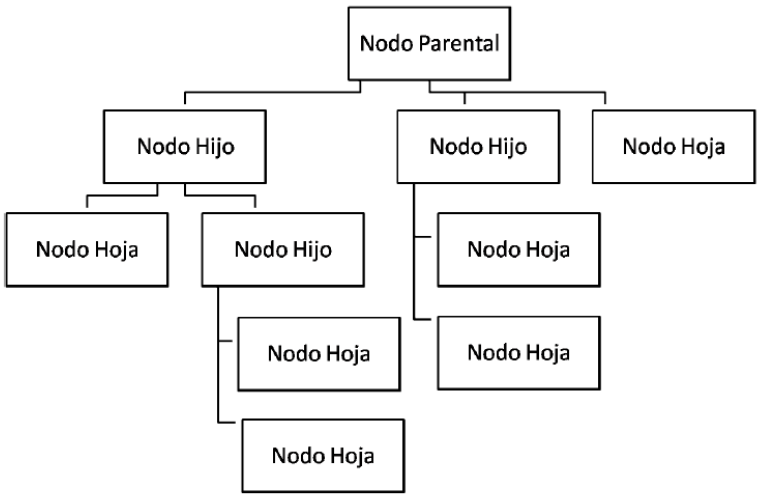

Ilustración 4: Árbol general de decisión.
Fuente: Modelos de decisión en ambiente incierto-DII.

Existen diferentes criterios que se pueden utilizar en la división de los datos, destacando:

- Índice de Gini: Este criterio se basa en la medida de impureza de cada nodo, ya que se mide la divergencia entre las distribuciones de probabilidad de los valores de los atributos objetivos [13]. La fórmula general de este índice es:

$$I_m = \sum_{k=1}^{K} p_{m,k}\,(1 - p_{m,k}) \; = 1 - \sum_{k=1}^{K} \left(p_{m,k}\right)^2$$

Donde el $k = 1, \ldots, K$ son las categorías de la variable respuesta y $p_{m,k}$ es la proporción de elementos de la categoría k en el nodo m, siendo su mínimo valor 0 cuando los datos de un nodo pertenecen a un mismo grupo.

- Ganancia de información: Este criterio utiliza la función de entropía como medida de impureza, es decir, mide la complejidad o el caos que posee la variable en cuanto a la cantidad de categorías que tiene, de manera que una variable continua resulta altamente caótica debido a su multiplicidad de valores [12].

- Ratio de Ganancia: Este índice normaliza la ganancia de información, de manera que es necesario obtener la ganancia de información de todos los atributos, y luego se calcula el ratio de ganancia solamente a aquellos atributos que posean una ganancia de información mayor al promedio de todos, por lo que cuando mayor es el ratio mejor es la división. La fórmula de este índice:

$$\text{Ratio de Ganancia} = \frac{\text{Ganancia de información } (y, S)}{\text{Entropía } (y, S)}$$

Es importante considerar otros parámetros además de que criterio utilizar, ya que de esta forma es posible generar un árbol que sea representativo y a la vez fácil de interpretar, por lo que se destaca lo siguiente:

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

- Número mínimo de observaciones por nodo final (Leaf Size): Número mínimo de observaciones que tiene que tener un nodo final para que se construya la regla.

- Número mínimo de observaciones para dividir un nodo (Split Size): Número mínimo de observaciones que tiene que tener un nodo para que se pueda cortar por la variables seleccionada [12].

4.2.2 Random Forest

Esta técnica de agregación consiste en la creación de múltiples árboles de decisión que permite mejorar la precisión de clasificación, ya que incorpora aleatoriedad en la selección de variables.

Random Forest a diferencia del método CART[1], es una versión mejorada de esta metodología ya que permite solucionar principalmente tres problemas que posee este modelo, los cuales son:

1. Cada nodo produce una división posterior, de modo que las divisiones precedentes se ven afectada por estas decisiones.

2. Los datos al experimentar pequeños cambios en su configuración pueden generar estructuras de árboles diferentes, de forma que es poco robusto.

3. Al seleccionar una variable en la división de un nodo, puede existir otra variable con características muy similares que lamentablemente quede descartada para la división de esta rama.

Este método no solo permite solucionar los problemas ya descritos, sino que también clasificar y definir ranking de importancia de variables en la predicción de la variable de respuesta, ya que ésta es de vital importancia debido a la interacción que tiene con el resto. Random Forest genera dos medida de importancia:

1. MDA (Mean Decrease Accuracy): Esta medición de importancia se basa en la contribución de la variable al error de predicción, es decir, al error de mal clasificados. Para determinar la importancia de cada una de las variables se permutan aleatoriamente los valores de esa variable en particular, dejando intacto el resto de las variables, y se vuelven a clasificar los mismos individuos según el mismo árbol pero ahora con la variable permutada. Este valor se obtiene como la media de los incrementos en todos los árboles donde actúa la variable [6].

2. MDG (Mean Decrease Gini): Esta medición se obtiene del índice de Gini, el cual mide la impureza de cada nodo una vez que se haya seleccionado la variable de división de este.

Algoritmo Random Forest

Este método consiste en árboles predictores (N árboles de decisión), donde cada árbol depende de un vector aleatorio de forma independiente pero con la misma distribución entre ellos. Adicionalmente a los parámetros de los árboles de decisión es fundamental la incorporación de dos parámetros para la construcción de este modelo, donde el primero de ellos se relaciona con determinar un número de árboles a construir y el segundo a un número de variables a seleccionar en el subconjunto de datos de cada árbol.

[1] CART: Este modelo de árboles de decisión y regresión, es una técnica no paramétrica basada en la generación de un árbol que permite explicar o predecir una determinada variable de respuesta. Leo Breiman, Machine learning, 45, 5-32, 2011.

Para la construcción de N árboles se debe realizar lo siguiente:

1. De la data existente se debe seleccionar un conjunto de datos similares a través de un bootstrap, es decir, un remuestro con reposición, de manera que se corrige el error de predicción y se escoge el $\frac{2}{3}$ de la muestra.

2. Por cada nodo se debe seleccionar al azar un subconjunto de variables del tamaño determinado previamente y se restringe la selección de la variable a este subconjunto, de manera que la variable que proporcione la mejor división de acuerdo a una función objetiva, se utiliza en una división binaria.

3. Finalmente se vuelve a repetir el proceso en todos los nodos hasta obtener nodos terminales, de manera de obtener variables categorías (clasificación) o variables continuas (regresión).

La implementación de este método está presente en las librerías del programa R y RapidMiner, de manera que se seleccionó este último, ya que posee un paquete sobre Random Forest implementado en WEKA y en el mismo programa.

En WEKA Random Forest los árboles que se construyen corresponden al tipo C4.5, en los cuales se aceptan atributos discretos y continuos, utilizando el ratio de ganancia para seleccionar el atributo de cada nodo y aplicar estrategias de poda para reducir el ruido de los datos de entrenamientos, ya que aplica un test de hipótesis que trata de responder a la pregunta de si merece la pena expandir o no una determinada rama [12]. Los parámetros que se ingresan corresponden al número de árboles y variables a considerar en la construcción a diferencia del modelo implementado en el programa RapidMiner, donde dentro de los parámetros más importantes se encuentran el número de árboles a construir, el criterio de división y los parámetros asociados a la construcción de un árbol de decisión. (Máximo profundidad, Número de observaciones por nodo final, etcétera).

Finalmente de acuerdo a la literatura se sugiere iniciar el número de selección de variables predictivas como el resultado de $\log_2 M + 1$, siendo M el número total de variables consideradas, ya que este número puede variar en función del rendimiento del modelo al igual que el número de árboles debido a que no existen valores óptimos definidos para ambos parámetros.

5. METODOLOGÍA

1. Revisión bibliográfica: Se ha realizado una investigación de los modelos que se utilizan en la industria del retail para la predicción de fuga a través de papers, memorias, libros y trabajos, con el fin de determinar el modelo que se utilizará en esta memoria. Se trabajó en el aprendizaje de este modelo, el cual ha sido descrito en el capítulo de marco conceptual.

2. Selección y preprocesamiento de datos: Para determinar este punto se hace una limpieza de aquellos datos que no entreguen valor o sean outliers, ya que afectan directamente en el desarrollo y el desempeño de cualquier método que se utilice.

Posteriormente se realiza la selección y creación de variables a través de una base datos, de forma que se utilicen en los métodos descritos anteriormente. La fuente de información maestra se llama “ventasfasct” y se obtienen datos por medio de consultas a través de lenguaje SQL (Structured Query Language). Estos datos serán denominados datos procesados.

3. Determinar el criterio de fuga para los clientes Carterizados: Este punto se abarca a través del método RFM, definiendo que criterio de fuga se utilizará en conjunto con la empresa.

4. Aplicar modelo predictivo: Una vez definido el criterio de fuga se procede a la creación y selección de variables para la aplicación del modelo predictivo Árbol de decisión y Random Forest. Para determinar que variables se utilizaran en el modelo, se verificará la correlación entre ellas, para luego normalizar aquellas que no presenten valores que no permitan comparar resultados o no posean una distribución normal y así finalmente a través del modelo determinar las probabilidades de fuga de los clientes Carterizados.

5. Caracterización de la clasificación en base a los estados y puntos de corte definidos: Luego de obtener los resultados del modelo se procede a analizar los datos, ya que se debe caracterizar a los clientes como nuevos, leales, etcétera en base a sus probabilidades y criterios de fuga definidos con anterioridad.

6. Análisis de los motivos que producen el riesgo de fuga: En base a los resultados obtenidos del modelo, se debe analizar las causas que producen que un cliente se vaya de la empresa, ya que de esta forma se podrán enfocar los recursos y estrategias en definir acciones de marketing.

7. Definir acciones de retención: Finalmente analizados lo resultados del trabajo, se debe proponer estrategias orientadas a evitar que los clientes migren a la competencia.

6. ALCANCES

- El criterio de cliente fugado y otros, se hará en conjunto con la empresa, de forma adaptarlo al negocio.

- Se determinarán criterio y probabilidades de fuga sólo a clientes que forman parte de los clientes Carterizados.

- Sólo se utilizarán los datos transaccionales otorgados por la empresa Penta Analytics, por lo que no se utilizará ninguna otra fuente de información.

- Las políticas de retención se generarán sobre clientes propensos a fugarse, valorizados por la empresa y que pertenezcan a clientes Carterizados.

7. RESULTADOS ESPERADOS

Al término de este trabajo se pretender contar con los siguientes entregables:

- Definir el criterio de fuga y otros (clientes nuevos, retornados, en fuga, etcétera), con el fin de identificar el comportamiento de compra de los actuales y futuros clientes.

- Un modelo predictivo de fuga de clientes, donde se obtengan las probabilidades de fuga de los compradores actuales.

- Determinar las principales razones del porque un cliente propenso a fugarse se va de la empresa.

- Una política de retención sobre los clientes que sean valiosos para la empresa, de forma que se pueda utilizar en futuros clientes.

8. SELECCIÓN Y PREPROCESAMIENTO DE DATOS

En este capítulo se describirán las variables a utilizar para la realización de este trabajo tanto para la definición de criterio de fuga como también para el modelo predictivo. Estos datos fueron otorgados por la empresa Penta Analytics.

8.1 Descripción de la base de datos en general

Para la base de datos utilizada se dispone de los datos transaccionales por cada tipo de cliente desde el 02 de Enero del 2012 al 28 de Febrero del 2014, por transacción que se haya efectuado en cualquier sucursal del país. La descripción de las variables básicas que se utilizaron se observan en la Tabla 3.

Variable	Tipo de Variable	Descripción
Cliente_identificación	Numérica	Indica la identificación del cliente.
Fecha_diaria	Numérica	Indica la fecha exacta en que se realizó la transacción. Está en el formato dd/mm/aaaa.
Transacción_ identificación	Numérica	Asigna un número a la transacción realizada.
Ventas_Identificación	Numérica	Indica que tipo de cliente es, es decir, Carterizado, Mesón o No Carterizado.
Monto_transacción	Numérica	Indica el monto por el cual se realizó la transacción
Documentación_identificación	Numérica	Indica que tipo de documento fue la transacción realizada. Ejemplo: Factura, boleta, nota de crédito, etcétera.
Vendedor_identificación	Numérica	Indica la identificación del vendedor que realizó la transacción.

Tabla 3: Descripción de variables básicas.
Fuente: Elaboración Propia.

De la fuente de información se obtuvieron los siguientes datos de los clientes Carterizados:

- Total de clientes registrados en este periodo: 5.065
- Número de transacciones: 407.463
- Monto total: $122.718.346.286

8.2 Preprocesamiento de datos

De la base de datos que se obtuvo, se excluyeron los clientes que presentaban solamente una transacción y suma de montos negativos debido a que podían generar errores en los resultados de la definición de fuga. En la Ilustración 5 se observa el porcentaje de datos excluidos.

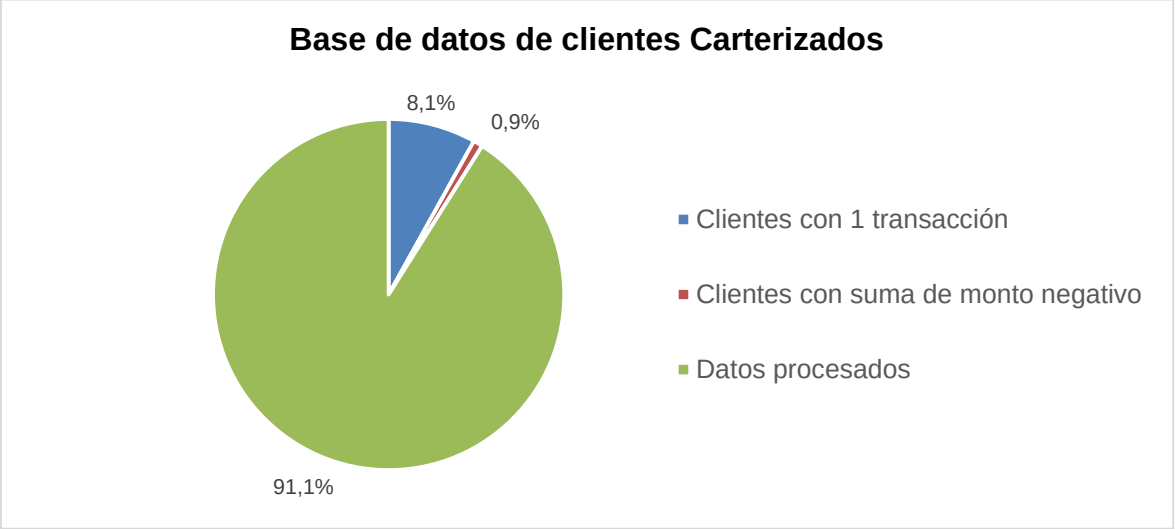

Ilustración 5 : Base de datos de clientes Carterizados.
Fuente: Elaboración Propia.

Finalmente la base de datos procesada corresponde al 91,1% de los datos de la fuente información, de manera que el nuevo resumen de datos de los clientes Carterizados es:

- Total de clientes registrados en este periodo: 4.612
- Número de transacciones: 406.661
- Monto total: $122.577.366.079

En base a lo anterior se concluye que existe una variación en el total de clientes Carterizados, números de transacciones y monto total de 9,82%, 0,19% y 0,11% respectivamente.

8.3 Variables calculadas

Para el desarrollo de este trabajo fue necesario la creación de nuevas variables que permitieran aplicar la técnica RFM descrita con anterioridad en la definición de criterio de fuga y en los modelos predictivos.

Las nuevas variables y su descripción se observan en la Tabla 4:

Variable	Tipo de Variable	Descripción
Número_transacciones	Numérica	Indica el número de transacciones total.
Número_transacciones_ Día	Numérica	Indica la cantidad de días en los cual se realizaron transacciones.
Antigüedad	Numérica	Indica la cantidad de días que el cliente lleva registrado como cliente Carterizado.
Maxima_inactividad	Numérica	Indica el máximo tiempo que transcurre entre transacciones.
Recency	Numérica	Indica el tiempo que ha transcurrido desde la última fecha de compra.
Frecuencia	Numérica	Indica el tiempo transcurrido entre transacciones.
R/F	Numérica	Indica la velocidad de compra que posee un cliente.
Monto_promedio	Numérica	Indica el monto promedio de las transacciones en un periodo determinado.
Monto_total	Numérica	Indica el monto total facturado por cliente.
Variación_monto	Numérica	Indica la variación de monto registrado en los últimos 6 meses comparando dos trimestres consecutivos.
Monto_negativo	Numérica	Indica la suma de los montos de las transacciones que presentan montos negativos.
Número_transacciones_ Negativas	Numérica	Indica el número total de transacciones que presentan montos negativos.
%Monto_ transacciones_negativas	Numérica	Indica el porcentaje que representa el monto de transacciones negativas sobre el monto total por cada cliente.
Región	Nominal	Indica a que región del país pertenece el cliente registrado.
Número_transacciones_ devoluciones	Numérica	Indica el número total de transacciones negativas que corresponden a devoluciones.
%Monto_ devoluciones	Numérica	Indica el porcentaje que representa el monto de las devoluciones sobre el monto total.
Número_transacciones_ notacredito	Numérica	Indica el número total de transacciones negativas que corresponden a notas de créditos.
%Monto_ notacredito	Numérica	Indica el porcentaje que representa el monto de las notas de créditos sobre el monto total.
Número_transacciones_ retail	Numérica	Indica el número total de transacciones negativas que corresponden a retail. (Devoluciones a través del canal mesón).
%Monto_retail	Numérica	Indica el porcentaje que representa el monto de retail sobre el monto total.
Giro_Comercial	Nominal	Indica a que rubro comercial pertenece el cliente registrado. Ejemplo: Construcción, hotelería, educación,etcétera.
%Monto_obra_ gruesa	Numérica	Indica el porcentaje que representa el monto de las transacciones que contienen productos clasificados como obra gruesa sobre el monto total.
%Monto_obra_ intermedia	Numérica	Indica el porcentaje que representa el monto de las transacciones que contienen productos clasificados como obra intermedia sobre el monto total.
%Monto_ terminaciones	Numérica	Indica el porcentaje que representa el monto de las transacciones que contienen productos clasificados como terminaciones sobre el monto total.
%Monto_otros	Numérica	Indica el porcentaje que representa el monto de las transacciones que contienen productos clasificados como otros sobre el monto total.

Tabla 4: Descripción de variables calculadas.
Fuente: Elaboración Propia.

8.3.1 Análisis univariado de variables calculadas.

De las variables mencionadas anteriormente se realizó un análisis sobre las variables recency, frecuencia, monto promedio y R/F ya que son relevantes para el método RFM. El resto del análisis de las otras variables se encuentra en Anexo 1.

- Recency

Esta variable se define como el tiempo transcurrido desde la última fecha de compra. Para el cálculo de ésta se consideró el tiempo que había transcurrió desde su última compra al 28 de Febrero del 2014 por cada cliente Carterizado, de manera que el resultado obtenido esta expresado en días.

A continuación se observa la distribución del recency de los clientes Carterizados en el Ilustración 6 y la Tabla 5.

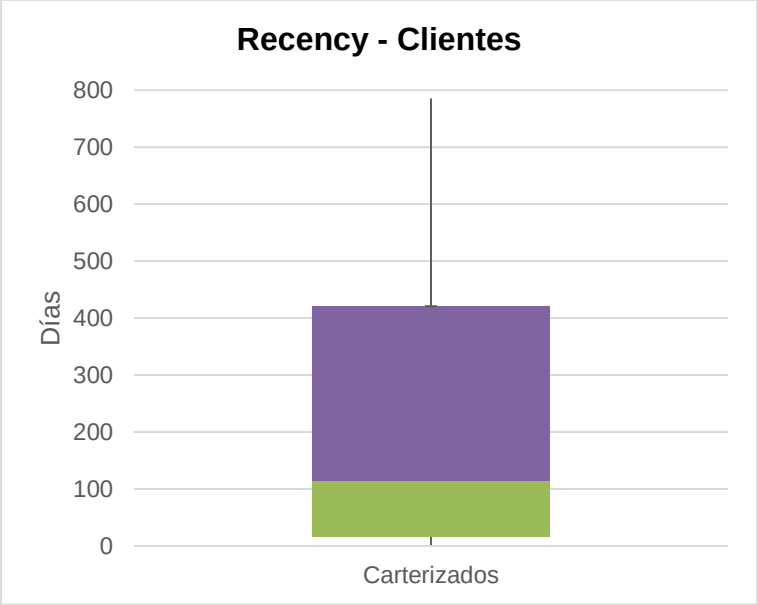

Ilustración 6: Diagrama de caja de los clientes Carterizados - Recency.
Fuente: Elaboración Propia.

Recency	Carterizados [días]
Mínimo	0
Cuartil 1	16
Mediana	115
Cuartil 2	421
Máximo	785

Tabla 5: Resumen del diagrama de caja de los clientes - Recency.
Fuente: Elaboración Propia.

De lo anterior se tiene que para el 50% de los clientes el tiempo transcurrido desde su última compra es menor o igual a 115 días en el caso de los clientes Carterizados. Es importante destacar que el tiempo transcurrido para el 75% de los clientes se obtiene cuando ya ha pasado más de un año, es decir , 365 días, por lo que el comportamiento

de compra difiere en gran medida al que se obtendría en otro sector del retail como una tienda por departamento o supermercados.

En la Ilustración 7 se presenta la distribución de los clientes Carterizados de acuerdo al recency. Se aprecia que un 29,9% posee un recency menor o igual a un mes, esto quiere decir que los clientes han efectuado alguna compra en un tiempo menor o igual a 30 días.

También se tiene que a partir del quinto mes se obtiene menos del 5% de los clientes para los siguientes meses. Los valores mayores a 15 meses poseen muy pocas ocurrencias por lo que se decide agruparlos.

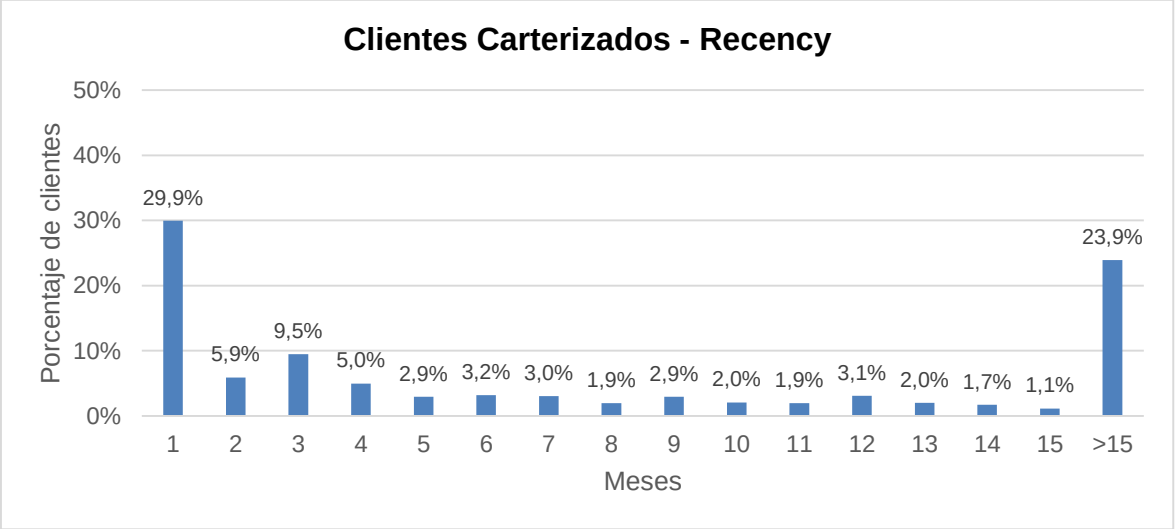

Ilustración 7: Distribución de los clientes Carterizados - Recency.
Fuente: Elaboración Propia.

- Frecuencia

Esta variable se define como el tiempo transcurrido entre transacciones. Para efectos de este trabajo se calculó como:

$$\frac{\text{última fecha de compra} - \text{primera fecha de compra}}{\text{número de transacciones} - 1}$$

A continuación se observa la distribución de los valores obtenidos para la frecuencia de los clientes Carterizados en la Ilustración 8 y la Tabla 6:

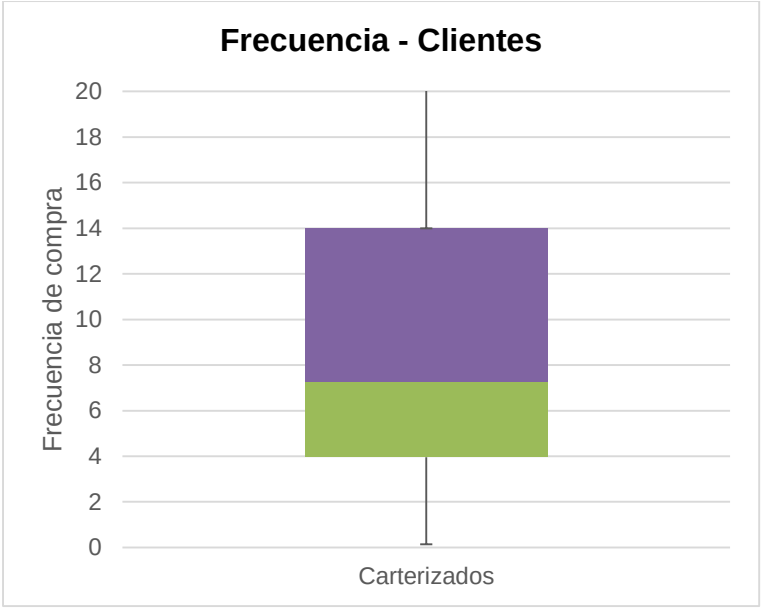

Ilustración 8: Diagrama de caja de los clientes Carterizados - Frecuencia.
Fuente: Elaboración Propia.

Frecuencia	Carterizados
Mínimo	0,13
Cuartil 1	3,94
Mediana	7,26
Cuartil 2	14,00
Máximo	613,00

Tabla 6: Resumen del diagrama de caja de los clientes Carterizados - Frecuencia.
Fuente: Elaboración Propia.

Análogo al caso del recency, se tiene que para el 50% de los clientes el tiempo transcurrido entre compras no suele ser más de 8 días para los clientes Carterizados.

En la Ilustración 9 se presenta la distribución de los clientes Carterizados de acuerdo a la frecuencia de compra. Se aprecia que solo un 2,2% posee una frecuencia menor o igual a 1, esto quiere decir que cada día por lo menos realizan una transacción, sin embargo es importante destacar que el porcentaje de clientes comienza aumentar a partir de frecuencias mayores a 1, donde el máximo de esto se obtiene entre valores superiores a 3 e iguales a 4, para luego observar que a partir de la frecuencia de valor sobre 6 la distribución de clientes comienza a disminuir. Finalmente los valores a partir de la frecuencia de compra sobre 15 tienen muy pocas ocurrencias por lo que se grafican agrupadamente.

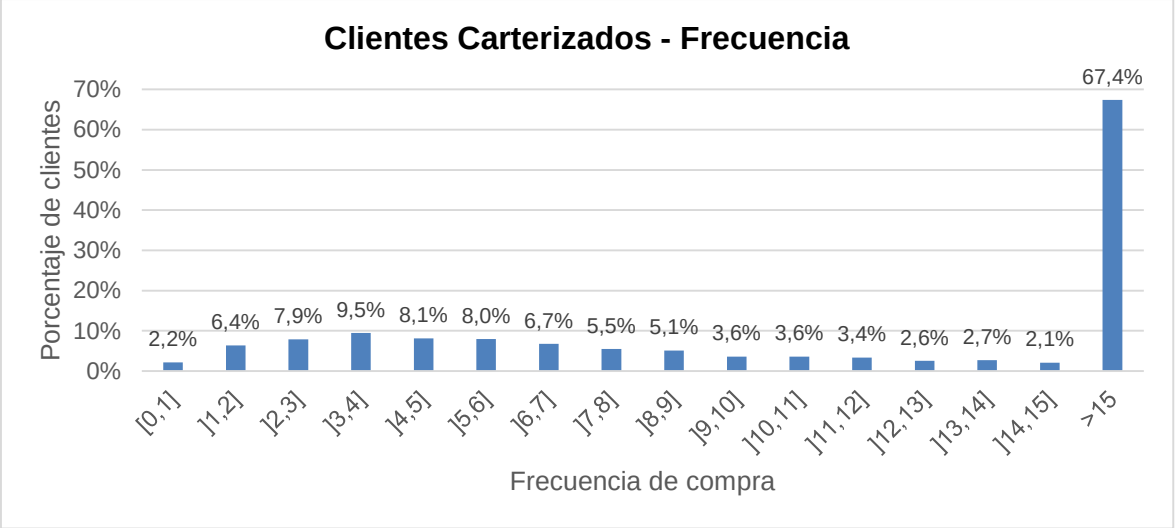

Ilustración 9: Distribución de clientes Carterizados - Frecuencia.
Fuente: Elaboración Propia.

- Monto promedio

Esta variable se define como el monto promedio de las transacciones en un periodo de tiempo determinado y su valor está expresado en pesos chilenos.

A continuación se observa la distribución de los montos obtenidos para los clientes Carterizados en la Ilustración 10 y Tabla 7.

Ilustración 10: Diagrama de cajas de los clientes Carterizados - Monto Promedio. Fuente: Elaboración Propia.

Monto	Carterizados
Mínimo	$ 2.961
Cuartil 1	$ 110.234
Mediana	$ 203.941
Cuartil 2	$ 406.736
Máximo	$ 16.362.875

Tabla 7: Resumen del Diagrama de caja de los clientes Carterizados - Monto Promedio.
Fuente: Elaboración Propia

En relación a la variable del monto, se tiene que para el 50% de los clientes el monto promedio de las transacciones para los clientes Carterizados es de $203.941 y el máximo supera los $16 millones.

En la Ilustración 11 se presenta la distribución de los montos promedios de las transacciones de los clientes Carterizados. Se aprecia que el 16,2% de los clientes realiza transacciones por un monto entre $50M y $100M, seguido con un 15,6% de los clientes en transacciones entre $150M y $100M. A partir de montos superiores a $150M el porcentaje de clientes comienza a disminuir. Las transacciones por montos sobre los $400M presentan pocas ocurrencias por lo que se muestran agrupadamente.

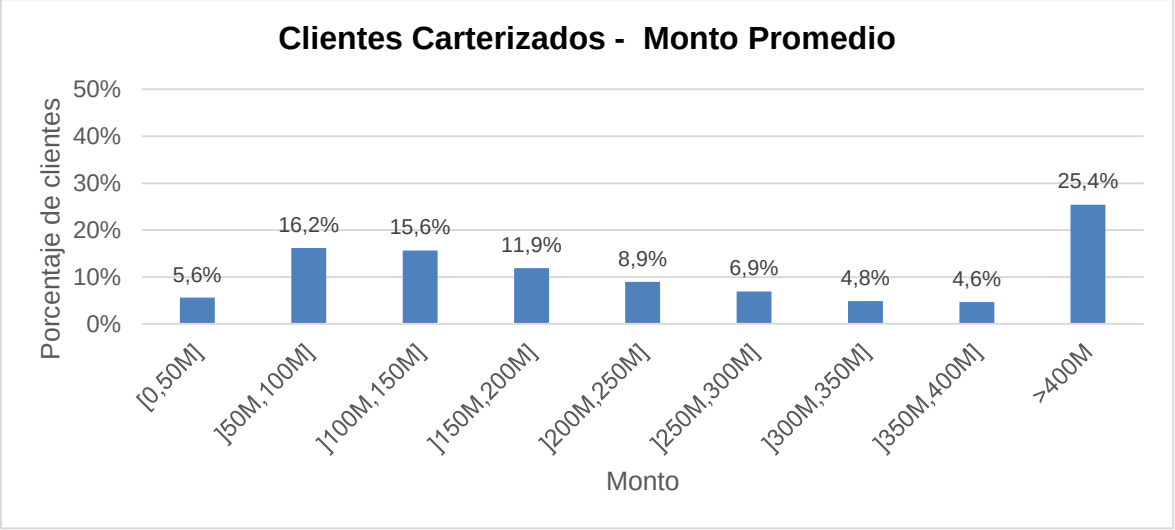

Ilustración 11: Distribución de clientes Carterizados - Monto Promedio.
Fuente: Elaboración Propia.

- R/F

Esta variable es el ratio entre recency y la frecuencia e indica cuan atrasado o adelantado se encuentra un cliente con respecto a su ciclo normal de compra.

A continuación se observa la distribución de los valores obtenidos para el R/F de los clientes Carterizados en la Ilustración 12 y la Tabla 8:

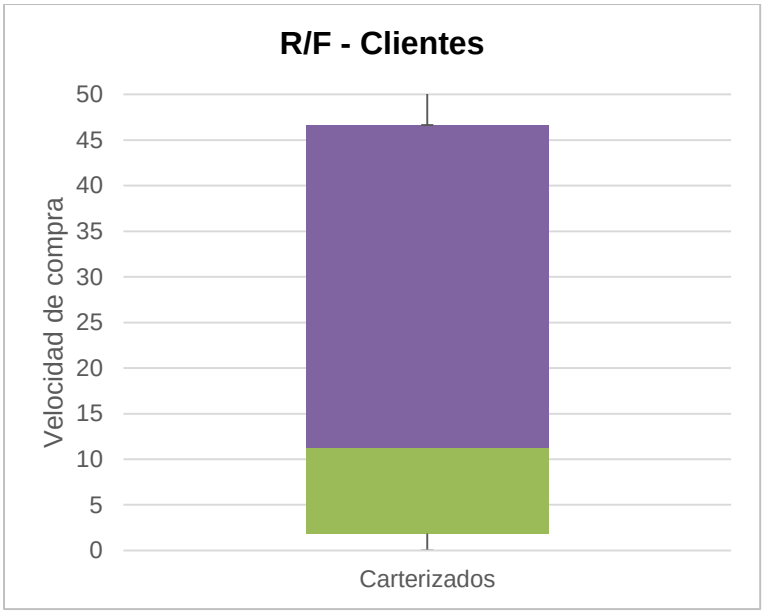

Ilustración 12: Diagrama de caja de los clientes Carterizados - R/F.
Fuente: Elaboración Propia.

R/F	Carterizados
Mínimo	0
Cuartil 1	1,85
Mediana	11,33
Cuartil 2	46,6
Máximo	1764

Tabla 8: Resumen del diagrama de caja de los clientes Carterizados - R/F.
Fuente: Elaboración Propia.

Se aprecia que para el 50% de los clientes el R/F no suele ser más de 12 ciclos para los clientes Carterizados. En la Ilustración 13 se presenta la distribución de los clientes Carterizados de acuerdo a este ratio. Se observa que el 19,1% posee un R/F menor o igual a 1, es decir, clientes que se demoran menos o igual a su ciclo de compra. Finalmente los valores a partir de R/F sobre 15 tienen muy pocas ocurrencias por lo que se grafican agrupadamente.

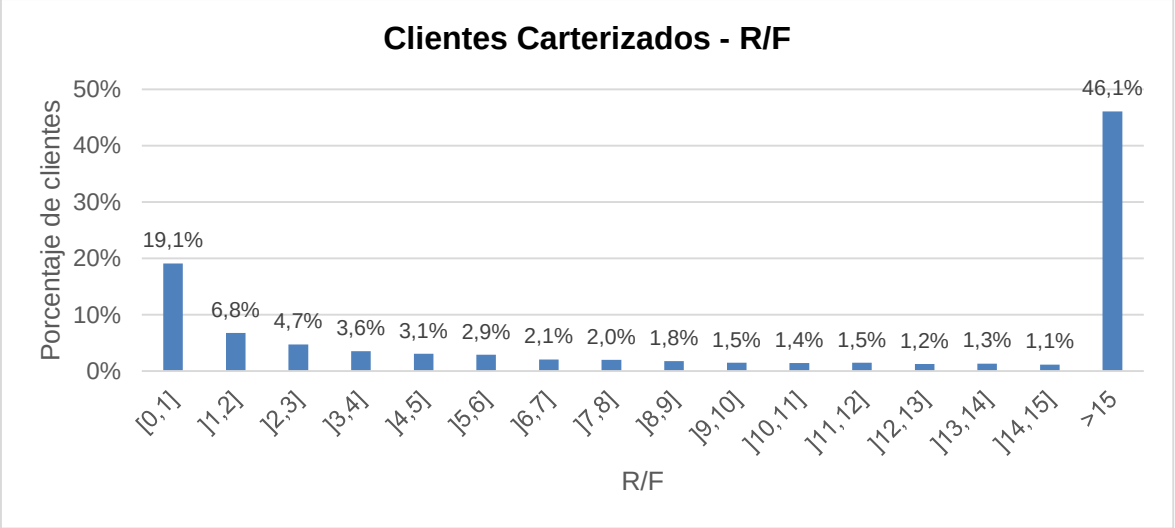

Ilustración 13: Distribución de clientes Carterizados - R/F.
Fuente: Elaboración Propia.

8.3.2 Análisis multivariado de variables calculadas.

De las variables calculadas se verificó la correlación que podría existir entre ellas, con el fin de descubrir la relación o algún comportamiento relevante que pudieran influir en los modelos predictivos.

En la Tabla 9 se observa las correlaciones entre las variables calculadas como también la variable básica de Vendedor_identificación. Se tiene que en general no poseen correlaciones muy altas a excepción de:

- Porcentaje de monto de notas de crédito con Porcentaje de monto negativo: 0,83.
- Número de transacciones con Número de transacciones por día: 0,79.
- Número de transacciones con Monto total de transacciones: 0,78.
- Número de transacciones con Número de notas de créditos: 0,75.
- Número de transacciones por día con Número de notas de créditos: 0,68.

También se observan correlaciones entre variables de valores superiores a 0,45, las cuales corresponden a:

- Porcentaje de monto de devoluciones con Porcentaje de monto negativo: 0,56.
- Monto total negativo con Número de transacciones negativo: 0,54.
- Recency con Velocidad de compra (R/F): 0,52.
- Máxima Inactividad con Frecuencia: 0,47.

Correlaciones																									
	#trans	#trandia	Anti	MI	R	F	R/F	MProm	MTotal	%V	MNeg	#Neg	%Mneg	Reg	#Dev	%Dev	#NotaC	%NotaC	#Retail	%Retail	%OG	%OI	%Otras	%T	Ven
#trans	1	**0,79**	0,21	-0,09	-0,26	-0,15	-0,06	0,22	**0,78**	0	-0,04	0,06	0,01	-0,03	0,37	0,01	**0,75**	0,01	0,27	-0,01	0,05	0,02	-0,04	-0,06	-0,09
#trandia	0,79	1	0,35	-0,11	-0,42	-0,24	-0,14	0,08	0,45	0	-0,08	0,11	0,02	-0,09	0,31	0,01	**0,68**	0,02	0,2	-0,02	0,07	0,03	-0,06	-0,07	-0,14
Anti	0,21	0,35	1	0,27	0,28	0,02	0,11	-0,02	0,11	-0,01	-0,03	0,04	-0,01	-0,15	0,12	-0,01	0,19	-0,01	0,02	-0,03	0,01	0,02	-0,06	0	-0,22
MI	-0,09	-0,11	0,27	1	-0,26	**0,47**	-0,25	-0,01	-0,05	0,03	0	-0,02	-0,01	-0,07	-0,03	-0,01	-0,07	-0,01	-0,03	-0,02	0,02	-0,02	0,02	0	-0,02
R	-0,26	-0,42	0,28	-0,26	1	0,03	0,52	-0,06	-0,13	-0,03	0,03	-0,04	-0,03	0,04	-0,12	-0,02	-0,24	-0,03	-0,07	0,01	-0,08	0,02	-0,01	0,07	-0,02
F	-0,15	-0,24	0,02	0,47	0,03	1	-0,15	0,02	-0,07	0	0,02	-0,03	-0,01	-0,01	-0,07	-0,01	-0,15	-0,01	-0,04	-0,01	-0,03	-0,04	0,08	0,04	0,06
R/F	-0,06	-0,14	0,11	-0,25	**0,52**	-0,15	1	0,01	-0,04	-0,01	0	0	0	0,05	-0,02	0	-0,06	0	-0,03	0,01	-0,05	0,04	-0,03	0,02	-0,03
MProm	0,22	0,08	-0,02	-0,01	-0,06	0,02	0,01	1	0,43	0,15	-0,02	0	0	0,07	0,04	0,01	0,16	0	0,13	-0,01	0,19	0,04	-0,04	-0,22	0,04
MTotal	0,78	0,45	0,11	-0,05	-0,13	-0,07	-0,04	0,43	1	0	-0,03	0,01	0	0,04	0,23	0	0,51	0	0,34	-0,01	0,09	0,01	-0,04	-0,08	-0,03
%V	0	0	-0,01	0,03	-0,03	0	-0,01	0,15	0	1	0	0	0	0,03	0	0	0,01	0	0	0	0,07	-0,03	-0,01	-0,03	0,03
MNeg	-0,04	-0,08	-0,03	0	0,03	0,02	0	-0,02	-0,03	0	1	-0,54	0	0,02	-0,12	0	0,01	0	0	0	-0,03	-0,01	0,02	0,03	0,02
#Neg	0,06	0,11	0,04	-0,02	-0,04	-0,03	0	0	0,01	0	**-0,54**	1	0	-0,03	0,25	0	-0,02	0	-0,01	0	0,01	0,02	-0,02	-0,02	-0,03
%Mneg	0,01	0,02	-0,01	-0,01	-0,03	-0,01	0	0	0	0	0	0	1	-0,02	0	0,56	0	0,83	0	-0,05	0,01	0	-0,03	0	0,01
Reg	-0,03	-0,09	-0,15	-0,07	0,04	-0,01	0,05	0,07	0,04	0,03	0,02	-0,03	-0,02	1	-0,08	0	0,03	-0,02	0,02	0,03	0,04	0	0,01	-0,05	0,13
#Dev	0,37	0,31	0,12	-0,03	-0,12	-0,07	-0,02	0,04	0,23	0	-0,12	0,25	0	-0,08	1	-0,02	-0,07	0,01	-0,01	-0,01	0	0,01	-0,03	0	-0,06
%Dev	0,01	0,01	-0,01	-0,01	-0,02	-0,01	0	0,01	0	0	0	0	**0,56**	0	-0,02	1	0,01	0	0	0	0,01	0,02	-0,05	-0,01	0
#NotaC	0,75	0,68	0,19	-0,07	-0,24	-0,15	-0,06	0,16	0,51	0,01	0,01	-0,02	0	0,03	-0,07	0,01	1	0	0,14	-0,01	0,06	0,01	-0,04	-0,06	-0,08
%NotaC	0,01	0,02	-0,01	-0,01	-0,03	-0,01	0	0	0	0	0	0	**0,83**	-0,02	0,01	0	0	1	0	0	0,02	-0,03	0	0,01	0,01
#Retail	0,27	0,2	0,02	-0,03	-0,07	-0,04	-0,03	0,13	0,34	0	0	-0,01	0	0,02	-0,01	0	0,14	0	1	0,03	0,04	0	-0,01	-0,03	-0,01
%Retail	-0,01	-0,02	-0,03	-0,02	0,01	-0,01	0,01	-0,01	-0,01	0	0	0	-0,05	0,03	-0,01	0	-0,01	0	0,03	1	0,28	-0,26	0	-0,02	0
%OG	0,05	0,07	0,01	0,02	-0,08	-0,03	-0,05	0,19	0,09	0,07	-0,03	0,01	0,01	0,04	0	0,01	0,06	0,02	0,04	0,28	1	-0,46	-0,17	-0,46	-0,01
%OI	0,02	0,03	0,02	-0,02	0,02	-0,04	0,04	0,04	0,01	-0,03	-0,01	0,02	0	0	0,01	0,02	0,01	-0,03	0	-0,26	-0,46	1	-0,19	-0,48	-0,02
%Otras	-0,04	-0,06	-0,06	0,02	-0,01	0,08	-0,03	-0,04	-0,04	-0,01	0,02	-0,02	-0,03	0,01	-0,03	-0,05	-0,04	0	-0,01	0	-0,17	-0,19	1	-0,07	0,04
%T	-0,06	-0,07	0	0	0,07	0,04	0,02	-0,22	-0,08	-0,03	0,03	-0,02	0	-0,05	0	-0,01	-0,06	0,01	-0,03	-0,02	-0,46	-0,48	-0,07	1	0,01
Ven	-0,09	-0,14	-0,22	-0,02	-0,02	0,06	-0,03	0,04	-0,03	0,03	0,02	-0,03	0,01	0,13	-0,06	0	-0,08	0,01	-0,01	0	-0,01	-0,02	0,04	0,01	1

Tabla 9: Correlación entre variables.
Fuente: Elaboración Propia.

8.3.3 Transformación de variables

Para utilizar de forma eficiente la base de datos en el modelo predictivo es necesario la normalización de éstos, ya que de esta manera se evita la redundancia, se es posible comparar valores y se mantiene la integridad de los datos.

En primera instancia se verificó cuan dispersos se encontraban a través de un histograma de frecuencia para luego normalizar a través de la función de Microsoft Excel *Normalización*, la cual requiere de la media y desviación estándar de las variables. Esta tipificación de la variable expresa el número de desviaciones típicas que dista de la media cada observación [14] y se calcula como:

$$Z = \frac{x-\mu}{\sigma} \sim N(0,1)$$

Donde x representa la variable, μ la media de la muestra y σ la desviación estándar de ésta. En la Ilustración 14 se observa un ejemplo de la distribución de los datos de la variable, número de transacción antes y después de la transformación.

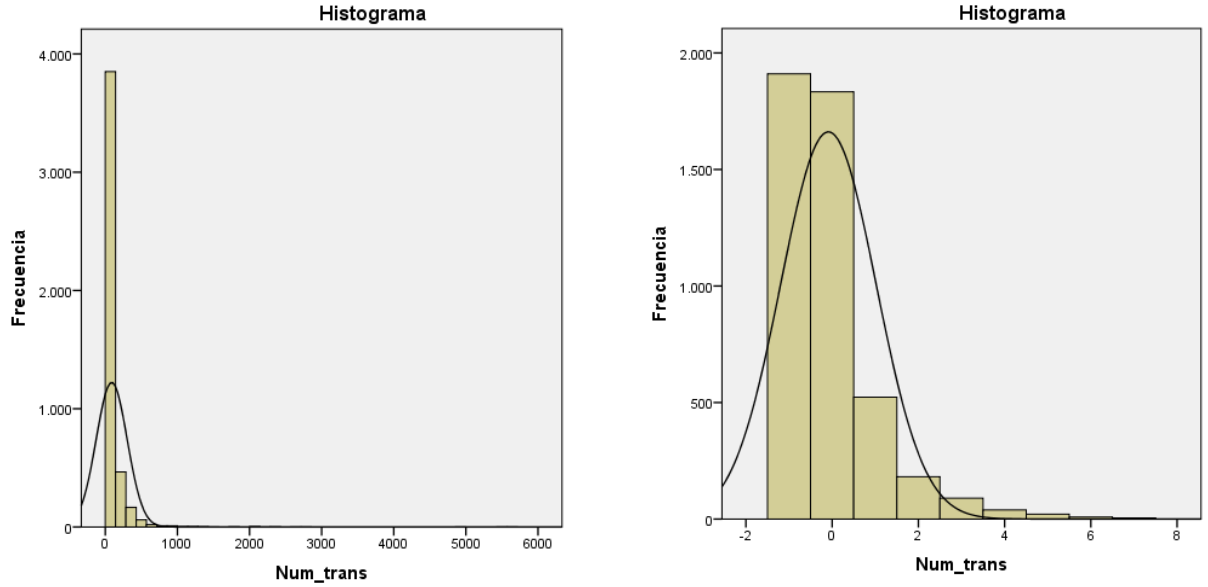

Ilustración 14: Ejemplo de la distribución de la transformación de la variable Número de transacciones. Fuente: Software SPSS.

9. DEFINICIÓN DE CRITERIO DE FUGA

Para la definición de criterio de fuga es necesario analizar las variables de recency, la velocidad de compra (R/F) y la variación de monto, tal como se indicó en el capítulo de marco conceptual a través de un análisis de sensibilidad para definir el valor máximo para cada una de ella.

9.1 Criterios de fuga

- Caso Recency:

Para calcular el porcentaje de clientes Caracterizados fugados se definió un valor máximo para un recency determinado, de manera que si un cliente sobrepasaba esta cota su estado sería de fugado. En la Ilustración 15 se observa la distribución de clientes fugados, donde se tiene que el porcentaje de cliente que posee un recency superior a 1 mes es de un 70,1% a diferencia de cuando es superior a un año donde se alcanza el 28,7%. Se esperaría que a medida de que aumente el recency la cantidad de clientes fugados fuese disminuyendo tal cual se ilustra.

Adicionalmente se calculó el porcentaje de clientes que se habían considerado fugados que vuelven a realizar transacción dentro de sus registros históricos, de manera de determinar el error que se producía con algún criterio de recency. Se observa en la misma Ilustración este valor mencionado donde el 68,5% de los clientes que se habían considerados fugados vuelven a efectuar alguna compra cuando su recency es superior a 1, de manera que este valor es un mal indicador debido a que prácticamente todos los clientes retornan a la empresa a diferencia de cuando el recency es mayor a 6 meses donde el porcentaje de clientes es bajo el 10%.

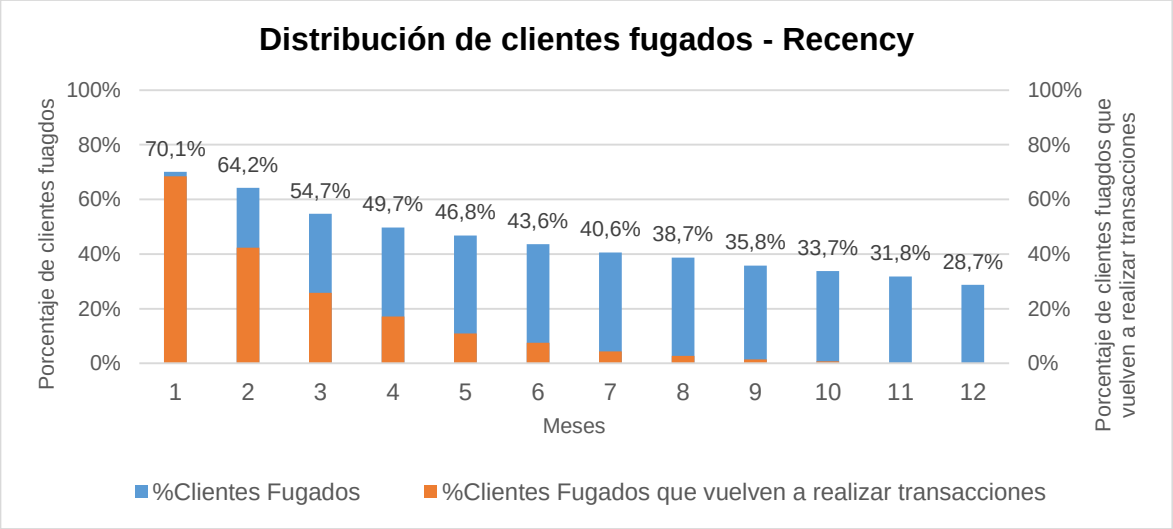

Ilustración 15: Distribución de clientes fugados de acuerdo al recency.
Fuente: Elaboración Propia.

- Caso R/F:

En la Ilustración 16 se observa el porcentaje de clientes que se considera fugado de acuerdo a cierto R/F. Se tiene que cuando se ha superado una cota máxima de 10 se tiene que el 52,6% de los clientes se encuentra fugado, y así sucesivamente, de manera que se esperaría que medida que se amplía el valor de R/F se obtiene un menor número de clientes fugados, el cual se ve reflejado en la disminución de clientes fugados en la ilustración.

Se calculó el error asociado a esta distribución de clientes fugados como aquellos clientes que se consideraron fugados pero volvieron a realizar alguna transacción al igual que en el caso del recency. Este valor se observa en la misma ilustración donde para un R/F de valor 10, el porcentaje de clientes que efectúa nuevamente una compra es 25,4%, por lo que sería una mala cota máxima debido al alto número de clientes en que se falla. A partir de valores de R/F sobre 20 se obtiene errores asociados a menos del 10%.

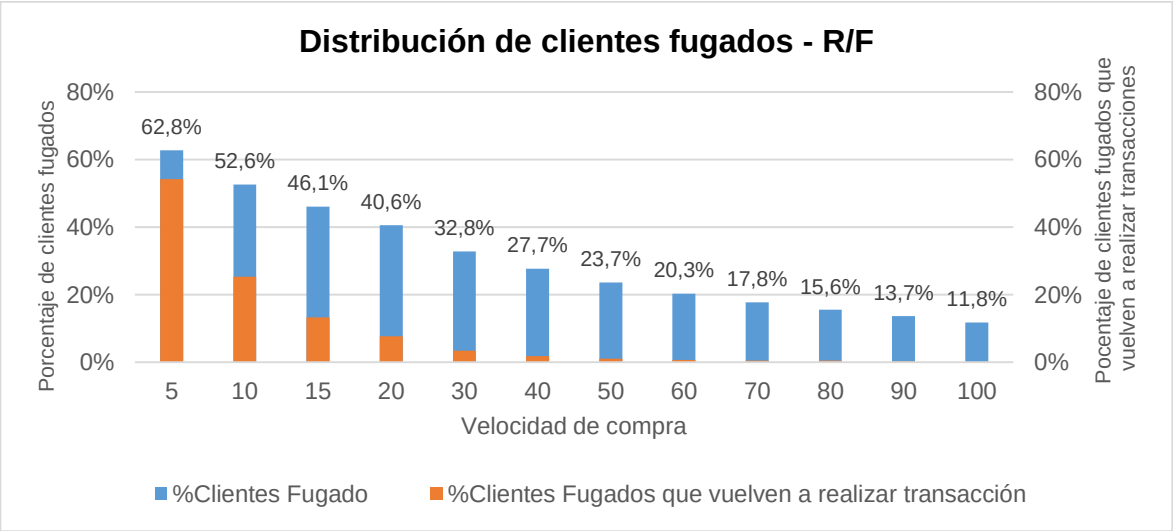

Ilustración 16: Distribución de clientes fugados de acuerdo al R/F.
Fuente: Elaboración Propia.

- Caso Variación de Monto

Esta variable mide la variación de monto que posee un cliente en los últimos 6 meses antes de su última compra. Se consideró contrastar los dos últimos trimestres consecutivos de los datos procesados, debido que este tipo de clientes no efectúa compras todos los meses, por lo que al no presentar compras o variaciones monetarias muy altas podría dar indicio de que algo sucede en su comportamiento de compra, el cual en caso de no ser capturado por este criterio, podría será abarcado por los otros dos (recency o R/F).

En la Tabla 10 se observa que un 23,7% de los clientes no presenta variación dado que su última fecha de compra fue antes de 6 meses o sólo registra montos en un trimestre, por lo que no forman parte de este análisis.

Descripción	Número de Clientes	% Clientes
Clientes con variación negativa	1758	38,1%
Clientes con variación positiva	1761	38,2%
Clientes sin variación	1093	23,7%

Tabla 10: Número y distribución de clientes que experimento variaciones.
Fuente: Elaboración Propia.

Finalmente se calculó la distribución de clientes que experimentaron variaciones monetarias negativas durante este periodo. En la Ilustración 17 se observa la distribución acumulada de esta variable y se aprecia que cuando se posee una variación entre [0%,

-80%] se captura el 86,9% de los clientes que presentaron una variación negativa en este periodo.

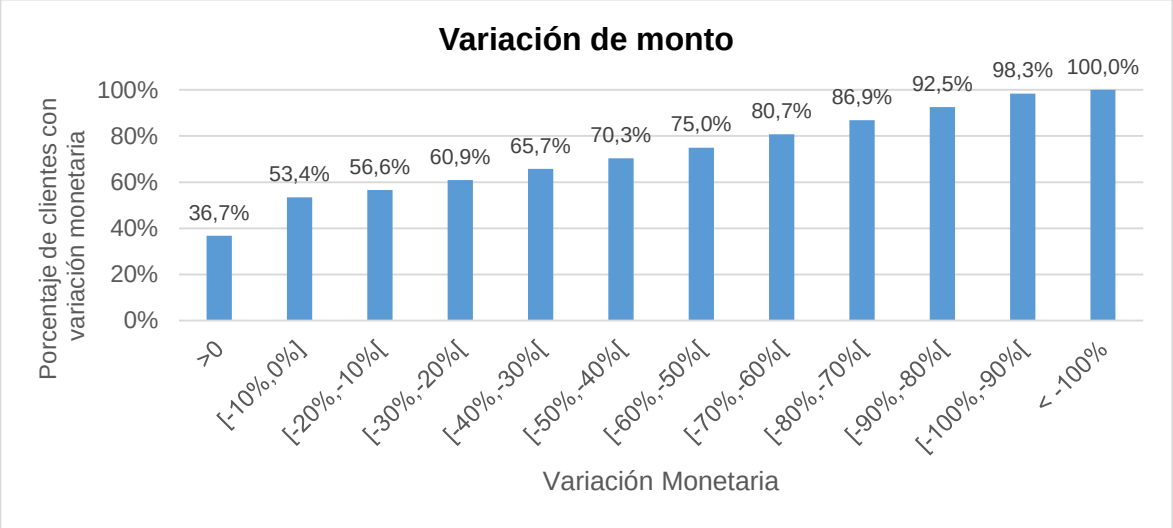

Ilustración 17: Distribución de clientes que experimentaron variación monetaria en ambos periodos. Fuente: Elaboración Propia.

- Criterio de fuga definido

Una vez analizado cada criterio por separado, se decidió junto con la empresa que para que un cliente sea considerado fugado debe sobrepasar la cota máxima para cualquiera de estos tres criterios.

Para la definición de la cota máxima se observó el análisis de sensibilidad por cada caso y el error asociado a cada uno de estos entendiéndose para el recency y R/F como el porcentaje de clientes que se consideraron fugados que vuelven a realizar transacciones. Para determinar cuanta variación monetaria negativa se aceptaría se observó el grafico de distribución acumulada de variación monetaria de manera de no capturar un porcentaje de los clientes con las variaciones negativas más significativas. No existe porcentaje de clientes fugados que vuelven a realizar transacciones, ya que se consideraron los 6 últimos meses de cada cliente antes de su última fecha de compra de los datos procesados.

En la Tabla 11 se observa las cotas máximas para cada criterio donde el error asociado tanto para el recency y R/F es menor al 10%.

Criterio	Cota Máxima	%Clientes fugados	%Clientes que vuelven a realizar transacciones
Recency	>180 días	43,6%	7,6%
R/F	>20	40,6%	7,7%
Variación Monetaria	<-80%	10,0%	-

Tabla 11: Cotas máximas para definición de cliente fugado. Fuente: Elaboración Propia.

Determinados los criterios de fuga junto con sus respectivas cotas máximas se aplicó en los datos procesados. En la Ilustración 18 se observa la distribución de clientes fugados de acuerdo a la combinación de los diferentes criterios, donde los clientes fugados representan el 53,9% versus los clientes activos que son el 46,1%. Se tiene que la mayor cantidad de fugados proviene de clientes donde su recency y R/F es mayor a la cota máxima y presentan variación monetaria mayor a -80%, representando estos un 24,1% de los clientes fugados. Adicionalmente se tiene que los clientes Carterizados que cumplen con los tres criterios determinados para considerarlo fugado representan el 3,4%.

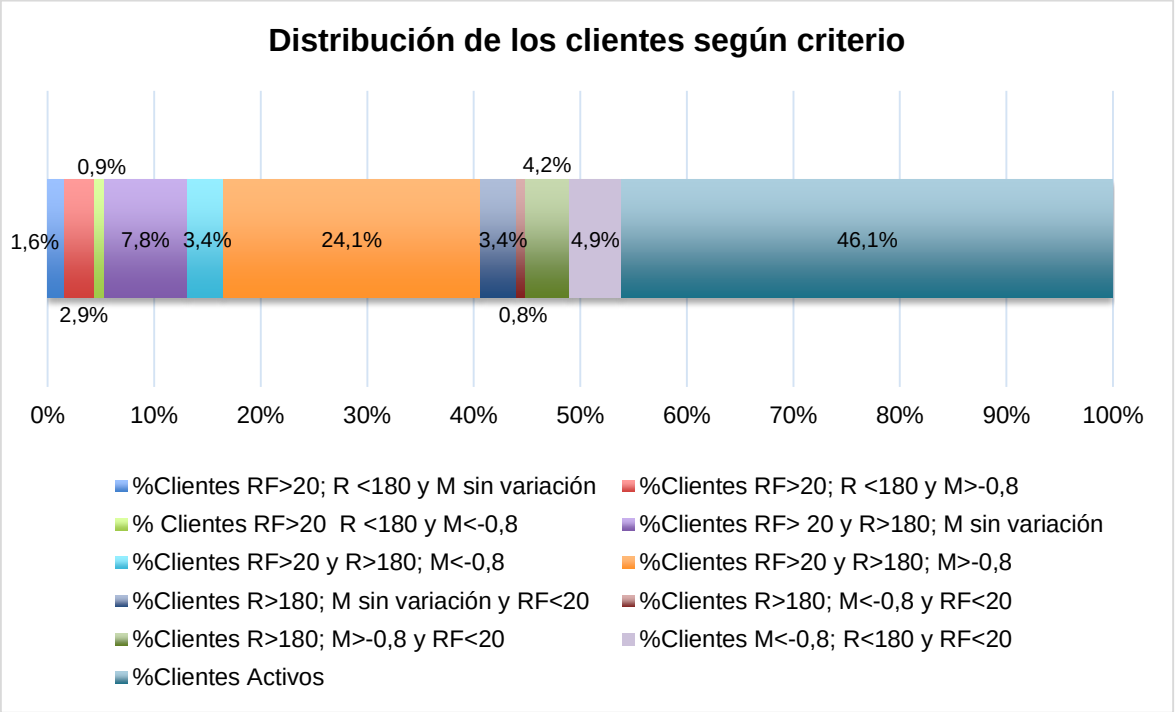

Ilustración 18: Distribución de clientes Carterizados fugados en la muestra seleccionada. Fuente: Elaboración Propia

10. MODELO PREDICTIVO DE FUGA

Las variables que forman parte del criterio de fuga, particularmente recency y R/F fueron originalmente calculadas al 28 de Febrero del 2014, si bien es cierto, permiten identificar si un cliente es fugado o no para el modelo predictivo es importante entender el comportamiento previo que tenía un cliente a fugarse, por lo que el valor de estas variables al igual que el resto de las otras, para los clientes catalogados como fugados fueron nuevamente calculadas pero esta vez en relación a la penúltima fecha de compra. De esta forma la base de datos que se utilizó finalmente en el modelo fue de 4456 clientes, es decir, 3,4% menos de datos que la base de datos procesados, debido a que existían clientes que poseían solamente dos transacciones o su R/F se indefine.

Las variables que se utilizaron en el modelo se observan en la Tabla 12.

N°	Variable	Tipo	Media	Desviación Estándar
1	Frecuencia	Continua	12,2	22,7
2	Recency	Continua	27,4	41,3
3	R/F	Continua	3,3	10,0
4	Variación Monto	Continua	2,2	46,3
5	Número transacciones	Continua	90,7	218,1
6	Máxima Inactividad	Continua	92,1	103,7
7	Monto Total	Continua	27.394.010	149.558.957
8	Monto Promedio	Continua	287.600	577.798
9	Monto Negativo	Continua	-2.838.697	25.404.209
10	%Monto Negativo	Continua	-0,2	3,4
11	Número transacciones negativo	Continua	7,0	21,7
12	Vendedor identificación	Categórica	-	-
13	Región	Categórica	-	-
14	%Monto Devoluciones	Continua	0,0	1,9
15	%Monto Nota Crédito	Continua	-0,1	2,8
16	%Monto Retail	Continua	0,0	0,0
17	Giro Comercial	Categórica	-	-
18	%Monto Obra Gruesa	Continua	0,2	0,3
19	%Monto Obra Intermedia	Continua	0,4	0,3
20	%Monto Otros	Continua	0,1	0,1
21	%Monto Terminaciones	Continua	0,3	0,3
22	Fugado	Discreta	-	-

Tabla 12: Variables utilizadas en el modelo predictivo de fuga.
Fuente: Elaboración Propia.

10.1 División en entrenamiento y testeo

Para la obtención de las probabilidades de fuga se ocupó una base con 21 variables independientes donde 18 de ellas son variables continuas y 3 categóricas. También se tiene una variable dependiente del tipo discreta que indica si el cliente está fugado o no.

Para este tipo de problemas es necesario entrenar y testear el modelo por lo que se utilizó la técnica cross validation o validación cruzada, en la que se decide un número fijo de particiones del conjunto de datos n, luego se separa el conjunto en n particiones iguales y en cada iteración se utiliza cada una de ellas para probar mientras que el resto se usa para entrenar el modelo, en la que se puede tomar cada partición de forma estratificada de tal manera que cada partición represente el conjunto de datos original [12].Para este caso se utilizó un n =10 y la ejecución de ambos modelos se realizó en el programa RapidMiner donde el modelo Random Forest utilizó el paquete Random Forest y su extensión en WEKA.

10.2 Árbol de decisión

Este modelo (CART) se llevó a cabo con los siguientes parámetros:

- Nodo parental: 45.
- Nodo filial: 15.
- Criterio de división: Gini.
- Ramas de profundidad: 6
- Nivel de confianza: 97,5%.

Una vez ejecutado el modelo en Rapid Miner, se obtuvieron resultados después 20 segundos donde en la Tabla 13 se observa la matriz de confusión de la partición de testeo. Los resultados de la partición de entrenamiento se observan en el Anexo 2.

Matriz de confusión			
	True No fugado	True Fugado	Class precisión
Pred. NoFugado	173	39	18,4%
Pred. Fugado	32	202	86,3%
Class recall	15,6%	83,8%	

Tabla 13: Matriz de confusión - Testeo para Árbol de Decisión.
Fuente: Rapid Miner.

El nivel de certeza en la partición de testeo fue de 84,1% y se tiene que de los 241 fugados reales, se detectaron 202 por lo que se predice el 83,8% de los fugados reales (Sensibilidad) y de los 234 clientes que el modelo clasifica como fugados, 202 lo eran, lo cual corresponde al 86,3%(Precisión)

La curva de ganancia indica que porcentaje de clientes declarados como fugados se captura con cierto porcentaje del total de los clientes. Se observa en la Ilustración 19 que con 50% del total de clientes se obtiene el 81,2% de los clientes fugados.

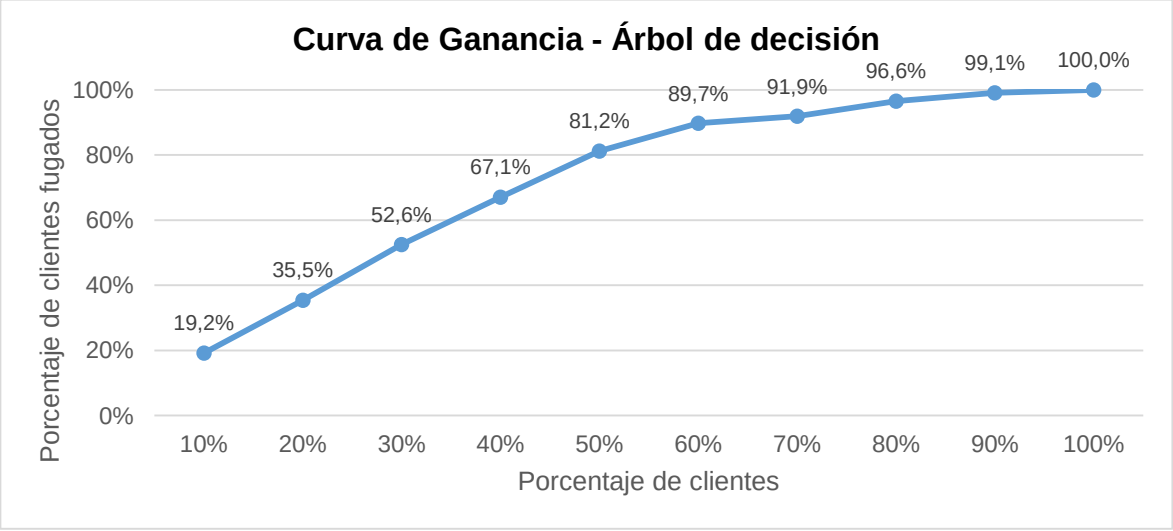

Ilustración 19: Curva de ganancia - Árbol de decisión.
Fuente: Elaboración Propia.

Finalmente el Árbol que se genera se observa en la Ilustración 20, donde se tiene que las variables más relevantes en la clasificación de los clientes son: recency, variación de monto, número de transacciones, máxima inactividad y frecuencia.

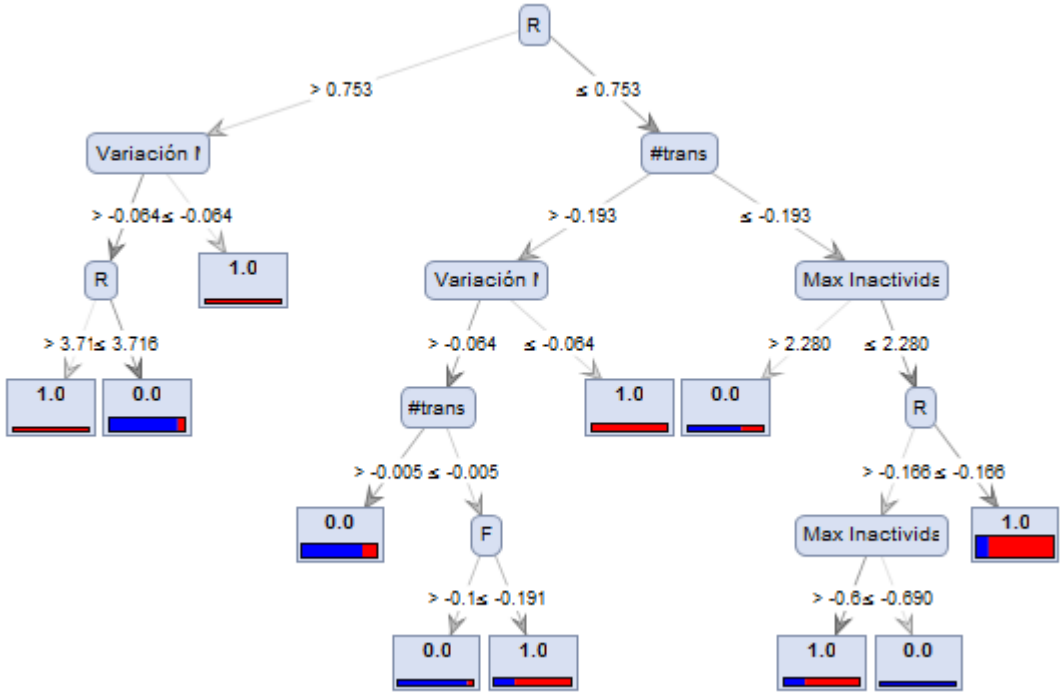

Ilustración 20: Resultados del Árbol de decisión.
Fuente: Rapid Miner.

10.3 Random Forest

Este modelo se llevó a cabo los mismo parámetros que el Árbol de decisión pero agregando el parámetro de construcción de 100 árboles en el paquete Random Forest de Rapid Miner.

Una vez ejecutado el modelo, se obtuvieron resultados después de 3 minutos y 42 segundos, donde en la Tabla 14 se observa la matriz de confusión de la partición de testeo. Los resultados de la partición de entrenamiento se observan en el Anexo 3.

Matriz de confusión			
	True No fugado	True Fugado	Class precisión
Pred. NoFugado	148	64	30,2%
Pred. Fugado	49	185	79,1%
Class recall	24,9%	74,3%	

Tabla 14: Matriz de confusión - Testeo para Random Forest.
Fuente: Rapid Miner.

De lo anterior se tiene que el nivel de certeza para el modelo predictivo fue de 74,7% donde de los 234 fugados reales, se detectaron 185 por lo que se predice el 79,1% de

los fugados reales (Sensibilidad) y de los 249 clientes que el modelo clasifica como fugados, 185 lo eran, lo cual corresponde al 74,3%(Precisión).

En la Ilustración 21 se tiene que la curva de ganancia indica que con 50% del total de los clientes se captura el 73,1% de los clientes fugados.

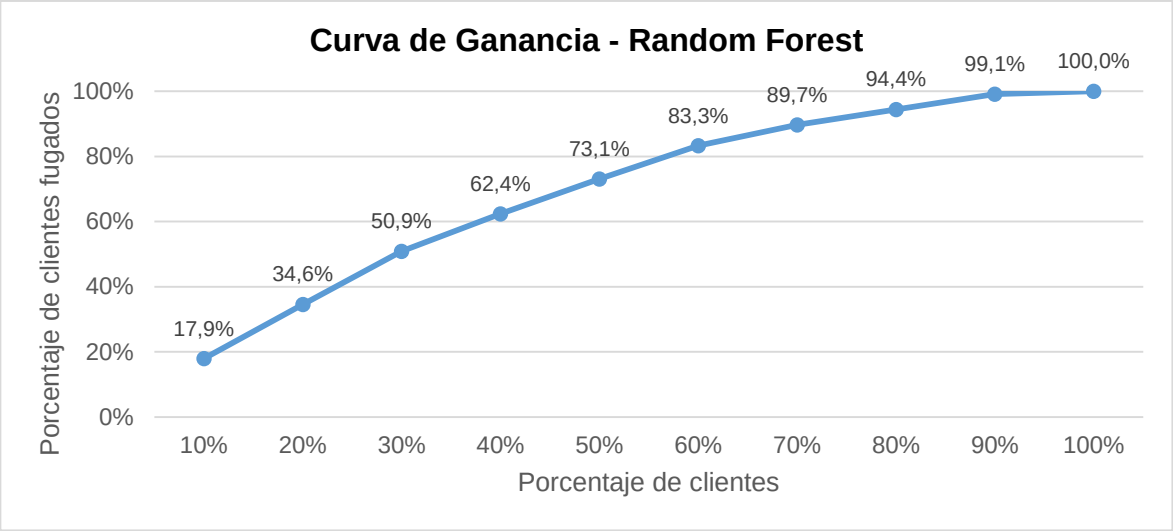

Ilustración 21: Curva de ganancia - Random Forest.
Fuente: Elaboración Propia.

Finalmente en relación a los dos indicadores que arroja el modelo Random Forest se tiene el Mean Decrease Accuracy (MDA) que indica la contribución de la variable al error de clasificación y el Mean Decrease Gini (MDG) que indica la contribución de la variable a la construcción del modelo. En las Ilustraciones 22 y 23 se observa estos indicadores, donde las variables están ordenadas en forma decreciente.

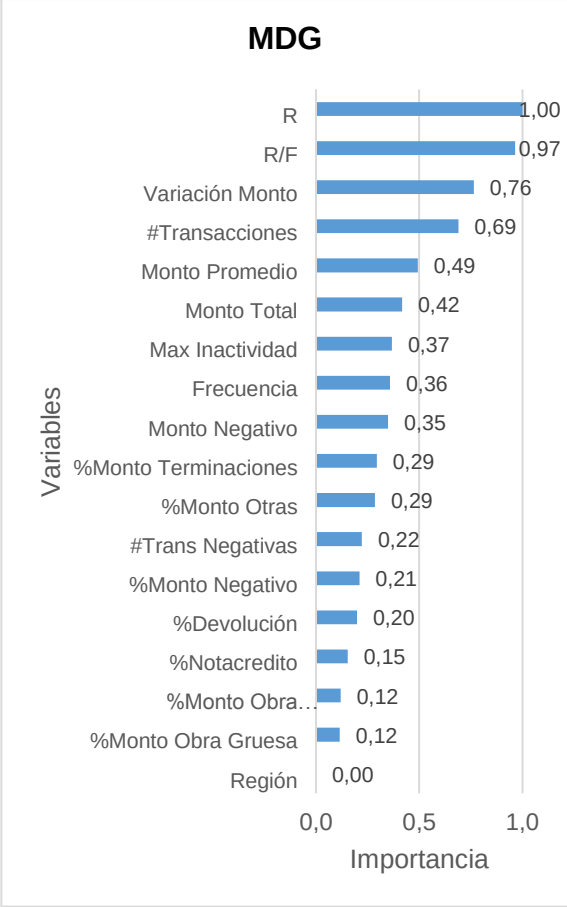

Ilustración 22: MDG Random Forest. Fuente: Elaboración Propia.

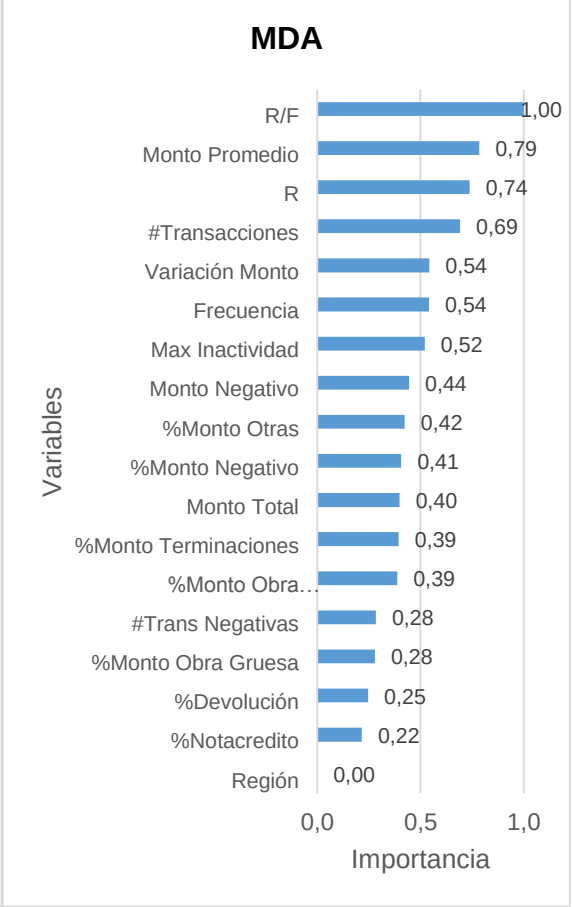

Ilustración 23: MDA Random Forest. Fuente: Elaboración Propia.

De acuerdo a lo anterior se esperaría que las variables más relevantes en la predicción del modelo tengan directa relación con aquellas que están involucradas en el criterio de fuga, es decir, recency, R/F y variación de monto. Esta situación que se ve reflejada en ambos gráficos, ya que estas variables se repiten dentro de los primeros lugares en ambos indicadores pero en distinto orden.

Destaca la influencia del MDA en la clasificación de un cliente donde las cinco variables más relevantes son: R/F, monto promedio, recency, número de transacciones y variación de monto. Se hubiese esperado que las variables que se relacionan con porcentajes de devoluciones o notas de créditos estuviesen dentro de los primeros lugares en la clasificación de un cliente, sin embargo esto no sucedió, pero es importante destacar la variable monto negativo que figura en octavo lugar dentro del ranking proporcionado por el MDA.

Las variables vendedor y giro comercial no aparecen en estos indicadores ya que al tomar los parámetros de nodo parental y filial de valores 45/15, no posee un número de casos que permita clasificar con estos valores. (Ejemplo: La variable vendedor, posee muchos identificadores de vendedores sin embargo el número de clientes es reducido por cada uno de ellos).

10.4 Análisis de sensibilidad: Random Forest

Para realizar el análisis de sensibilidad del modelo Random Forest se utilizó la extensión WEKA que posee el programa Rapid Miner, ya que permite variar dos parámetros fundamentales de este modelo, el número de árboles y el número de variables a considerar. Es importante destacar que con esta extensión no es posible ejecutar las dos medidas que entrega el modelo, es decir, el MDA y MDG, ya que para clasificar utiliza las variables más importantes de acuerdo al número de variables que se indique.

De esta forma se consideraron 12 situaciones para observar el rendimiento del modelo al modificar estos parámetros, debido a que la tasa de error de clasificación depende de la fuerza y la correlación entre los clasificadores individuales de árboles, ya que la determinación de un número de variables son cruciales en la reducción de la correlación entre árboles individuales en el aprendizaje [17] y la fuerza del modelo se ve mejorada al generar un rendimiento robusto (más árboles) porque baja la tasa del error global.

En la Tabla 15 se observa el rendimiento del modelo al variar el número de árboles (50, 100, 200 y 500) y el número de variables (22, 10, 5) en la partición de testeo de validación cruzada. Los resultados de las matrices de confusión se observan en el Anexo 4 para las diferentes situaciones.

Número de Árboles	Número de Variables	Tiempo de ejecución [seg]	Nivel de Certeza	Error Global	Sensibilidad	Precisión
50	22	3’19’’	79,6%	20,4%	80,2%	81,2%
	10	1’37’’	81,4%	18,6%	81,1%	84,2%
	5	**54’’**	**83,0%**	**17,0%**	**82,4%**	**85,9%**
100	22	6’52’’	80,0%	20,0%	80,6%	81,6%
	10	3’12’’	81,4%	18,6%	80,6%	85,0%
	5	1’51’’	82,7%	17,3%	82,3%	85,5%
200	22	12’44’’	79,6%	20,4%	80,2%	81,2%
	10	6’43’’	81,4%	18,6%	80,9%	81,2%
	5	3’36’’	81,4%	18,6%	81,1%	84,2%
500	22	33’38’’	79,4%	20,6%	79,8%	81,2%
	10	15’02’’	81,6%	18,4%	80,4%	85,9%
	5	8’46’’	82,3%	17,7%	81,6%	85,5%

Tabla 15: Análisis de sensibilidad del modelo Random Forest.
Fuente: Elaboración Propia.

De lo anterior se tiene que, al aumentar el número de árboles a construir, el tiempo operacional aumenta y aún más si se consideran un mayor número de variables. Este punto es muy importante a considerar cuando se posee una base de datos con un gran número de registros. Con respecto al nivel de certeza, se ve mejorado al disminuir el número de variables, ya que se ve disminuida la correlación entre los árboles construidos al momento de clasificar.

El modelo con 200 árboles presenta una situación anómala, ya que se hubiese esperado que obtuviera mejores resultados que el modelo con 100 árboles, por lo que se infiere que sobre esta cota no presenta mejoras en el rendimiento.

Finalmente todos los modelos para sus diferentes números de árboles obtienen mejores resultados al seleccionar 5 variables, donde el mejor de todos fue el modelo de 50 árboles de decisión, con un nivel de certeza de 83,0%, sensibilidad de 82,4% y precisión de 85,9%.

10.5 Probabilidades de fuga

Las probabilidades de fuga permiten identificar cuán propenso es un cliente a que se vaya del sistema, ya a qué medida que ésta es más alta es más probable que un cliente se fugue.

En las Ilustraciones 24 y 25 se observan la distribución de las probabilidades de fuga obtenidas por ambos modelos y como éstos se relacionan con los clientes catalogados como fugados sobre el total de clientes por probabilidad. Se esperaría que a medida que aumentan las probabilidades de fuga, se obtengan un mayor porcentaje de clientes fugados, ya que el modelo debería catalogarlo con una probabilidad más alta. Efectivamente esto sucede en ambos modelos, ya que a partir de la probabilidad de fuga mayor o igual a 0,6 se concentra un mayor porcentaje de los clientes fugados.

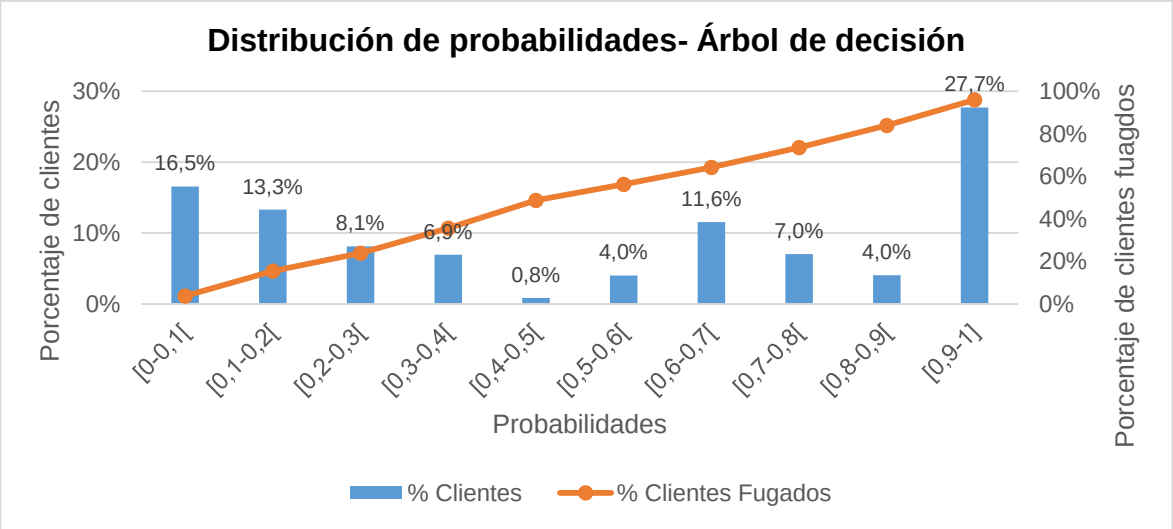

Ilustración 24: Probabilidades de fuga del modelo Árbol de decisión.
Fuente: Elaboración Propia.

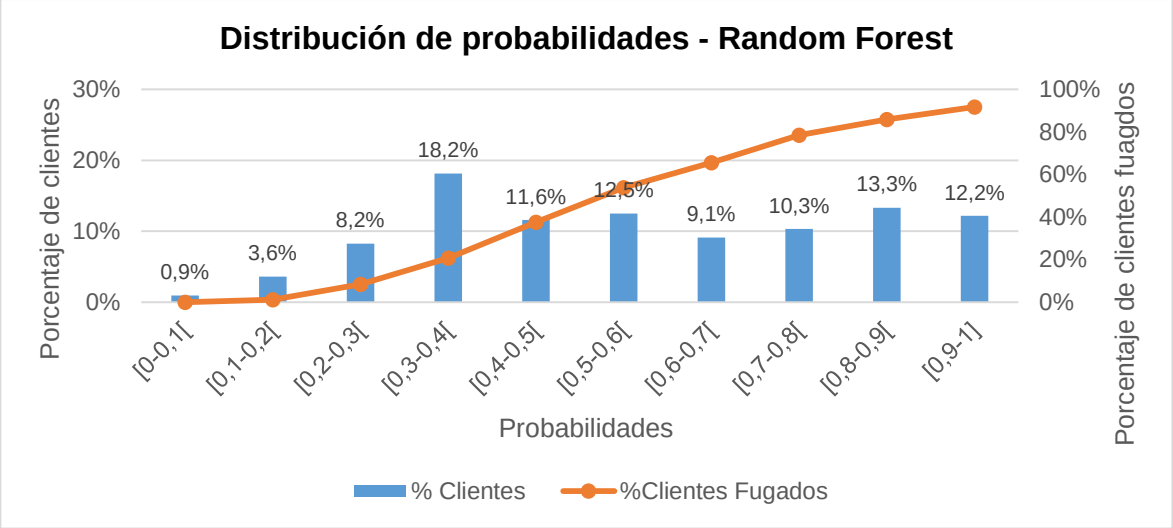

Ilustración 25: Probabilidades de fuga del modelo Random Forest.
Fuente: Elaboración Propia.

Se observa que en el caso del modelo del Árbol de decisión las probabilidades de fuga se concentran en los extremos, ya que existe un bajo número de clientes que presenten probabilidades mayor o igual a 0,3 y menor o igual a 0,6. El alto porcentaje de clientes con probabilidades sobre 0,9 puede verse explicado por aquellos clientes que de acuerdo al criterio se encuentran catalogados como fugados dado que su última fecha de compra fue en el año 2012-2013.

Para el caso del modelo Random Forest las probabilidades de fuga se encuentran distribuidas mucho más homogénea que el modelo anterior, ya que existe un incremento de clientes a partir de las probabilidades mayores a 0,2, alcanzando su peak en el rango mayor o igual a 0,3 y menor a 0,4 con 18,2%.

Finalmente existe una diferencia entre ambos modelos para clientes que poseen baja probabilidad de fuga (menor o igual a 0,2), ya que el Árbol de decisión considera mucho más clientes que el Random Forest.

10.6 Análisis de resultados

Una vez analizados los dos modelos se tiene que:

El Árbol de decisión presenta un mayor nivel de certeza para clasificar a un cliente frente al Random Forest a pesar de haber realizado un análisis de sensibilidad en relación a la variación de sus dos parámetros fundamentales, el número de árboles y el número de variables.

El modelo Random Forest entrega el MDA y MDG que permiten identificar las variables más relevantes en la clasificación de un cliente y en el Árbol de decisión se observan a través de los nodos de construcción, de manera que dentro de los primeros lugar se repiten en ambos modelos 5 variables: recency, variación de monto, número de transacciones, máxima Inactividad y frecuencia, por lo que se tiene que estas variables son claves para la determinación de la probabilidad de fuga de un cliente.

El modelo Random Forest se ve afectado al variar sus parámetros, ya que se tiene que a mayor número de árboles mayor fuerza del modelo debido a la robustez que alcanza, llegando a una cota donde la variación de resultados es poco significativa, siendo en este caso un n=100 árboles. También se tiene que a medida que el número de árboles es mayor el tiempo operacional va aumentando, siendo el máximo tiempo de ejecución de este trabajo 33’38’’ para la construcción de 500 árboles. En relación al número de variables se obtuvo que a menor número de variables mejor predicción del modelo, ya que se evita la correlación entre los distintos árboles que se construyen, entregando mejores resultados con un m=5.

Con respecto a las distribuciones de probabilidad de fuga existen diferencias entre ambos modelos, principalmente en aquellos clientes que poseen un baja probabilidad, donde el Árbol de decisión entrega un mayor porcentaje, sin embargo esto podría explicarse ya que al definir los criterios de fuga, existen clientes con características o patrones determinantes para clasificarlos como fugado versus a alguien que no lo está, además de poseer un 35,6% de clientes donde su última compra fue efectuada en Enero o Febrero 2014.

Finalmente en base a los resultados obtenidos se decidió optar por el Árbol de decisión, debido al rendimiento del modelo (diferencia aproximadamente del 2%) y la ventaja que ofrece sobre Random Forest en la interpretación de 1 árbol versus 100 de estos.

10.7 Resultados complementarios

10.7.1 Árbol de decisión sin variables RFM

Este modelo (CART) se llevó a cabo sin las variables transaccionales que se vinculan a la técnica RFM, es decir, se excluyeron: recency, R/F, variación de monto, frecuencia, máxima inactividad y número de transacciones con el objetivo de identificar cuan valiosas son las otras variables para determinar algún motivo de fuga. De esta forma el modelo se ejecutó con los siguientes parámetros:

- Nodo parental: 45.
- Nodo filial: 15.
- Criterio de división: Gini.
- Ramas de profundidad: 6
- Nivel de confianza: 97,5%.

Se obtuvieron resultados después 48 segundos donde en la Tabla 16 se observa la matriz de confusión de la partición de testeo. Los resultados de la partición de entrenamiento se observan en el Anexo 5.

Matriz de confusión			
	True No fugado	True Fugado	Class precisión
Pred. NoFugado	163	74	31,2%
Pred. Fugado	96	113	54,1%
Class recall	37,1%	60,4%	

Tabla 16: Matriz de confusión - Testeo para Árbol de Decisión sin RFM.
Fuente: Rapid Miner.

El nivel de certeza en la partición de testeo fue de 61,9% y se tiene que de los 187 fugados reales, se detectaron 113 por lo que se predice el 60,4% de los fugados reales (Sensibilidad) y de los 209 clientes que el modelo clasifica como fugados, 113 lo eran, lo cual corresponde al 54,1%(Precisión)

Finalmente el Árbol que se genera se observa en la Ilustración 26, donde se tiene que las variables que más inciden en esta clasificación corresponden a: monto negativo y %monto de notas de créditos.

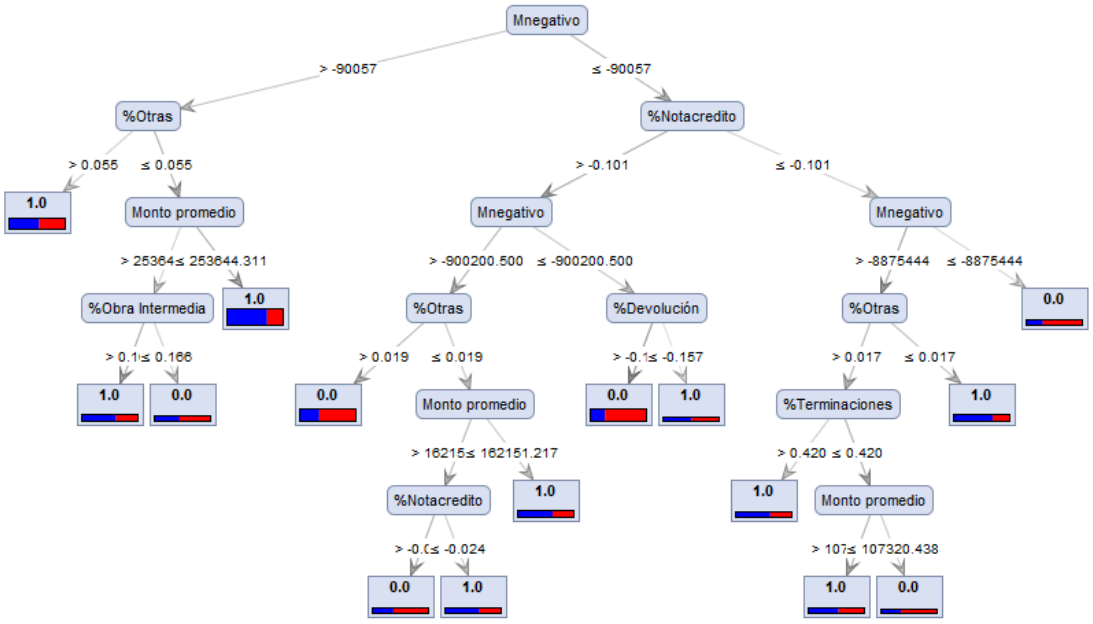

Ilustración 26: Resultados del Árbol de decisión sin RFM.
Fuente: Rapid Miner.

Si se comparan este modelo con el Árbol de decisión que incluye las variables RFM, se tiene que las variables recency, R/F y variación de monto tienen un alto valor predictivo, ya que existe una diferencia de más de un 20% en el nivel de certeza, sin embargo este árbol permite generar indicios de que estas variables explican razones de fuga de los clientes.

10.7.2 Comparación de modelos con base clientes 2014

Adicionalmente se comparó ambos modelos pero con una nueva base de datos con las mismas variables, sin embargo esta vez sólo se consideraron aquellos clientes cuya última compra fuese en Enero o Febrero del 2014. De esta forma la nueva base de datos posee 1582 clientes donde 182 son clientes fugados de acuerdo a los criterios de fuga definidos, el cual representa el 11,5%.

Para ambos modelos se consideró los siguientes parámetros:

- Nodo parental: 45.
- Nodo filial: 15.
- Criterio de división: Gini.
- Ramas de profundidad: 6
- Nivel de confianza: 97,5%.

En la Ilustración 27 se observan las curvas de ganancias del modelo Árbol de decisión versus Random Forest (100 Árboles de decisión).

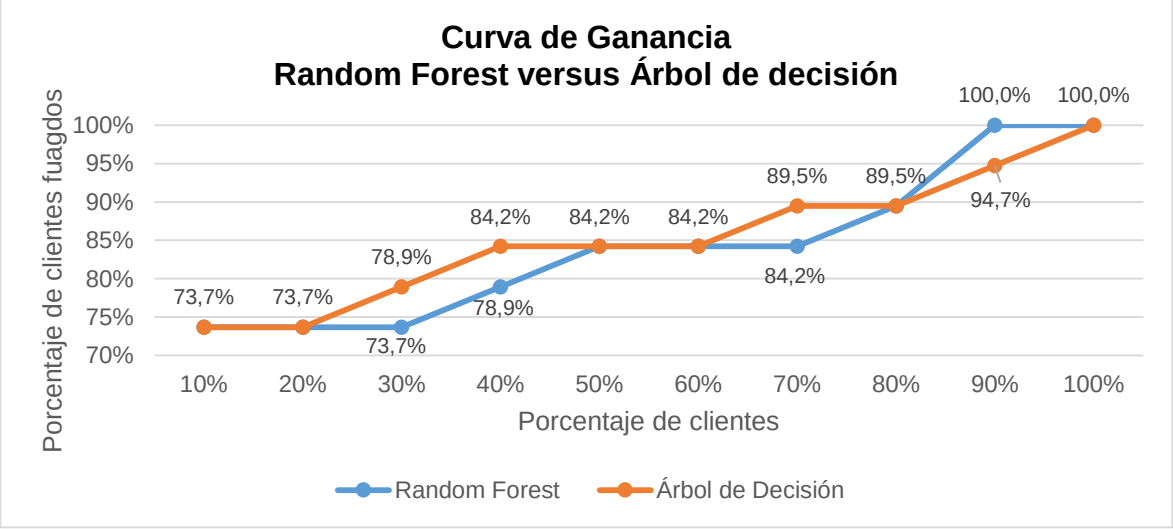

Ilustración 27: Curvas de ganancias con BBDD de clientes 2014.
Fuente: Elaboración Propia.

De lo anterior se tiene que con 30% del total de los clientes se captura el 73,7% de los clientes catalagodos como fugados en el caso del modelo Random Forest y el 78,9% con el Árbol de decisión, sin embargo al 90% del total de los clientes el modelo Random Forest captura el 100% de los clientes fugados mientras que el Árbol de decisión sólo el 94,7%.

11. CLASIFICACIÓN DE CLIENTES

Finalmente como el modelo escogido fue el Árbol de decisión se realizó una clasificación de los clientes Carterizados a través de las probabilidades de fuga definiendo rango de probabilidades que caractericen a los clientes que pertenezcan a esos segmentos.

Para ello se observó la distribución de las probabilidades de los clientes y la curva de ganancia mencionada anteriormente, de manera que se obtuvieron 4 segmentos de clientes: Leales, Normales, Propensos a fugarse y Fugados.

El rango de las probabilidades de fuga de esto segmentos se observa en la Tabla 17, junto con la distribución de clientes de los datos procesados.

Nombre del segmento	Rango de probabilidades	Porcentaje de clientes
Leales	[0-0,3[	37,9%
Normales	[0,3-0,5[	7,8%
Propensos	[0,5-0,7[	15,6%
Fugados	[0,7-1]	38,7%

Tabla 17: Clasificación de clientes Carterizados de acuerdo a su probabilidad de fuga.
Fuente: Elaboración Propia.

En la Ilustración 28 se observan la distribución de los clientes activos versus los clientes catalogados fugados por los distintos criterios. Se tiene que el mayor porcentaje de clientes activos se concentra en los clientes leales donde más del 87% se encuentra activo, no así el segmento fugado que sólo presenta un 9,4% de los clientes activos. Se esperaría que a medida que las probabilidades aumenten los clientes activos comiencen a disminuir, situación que se verifica en la Ilustración.

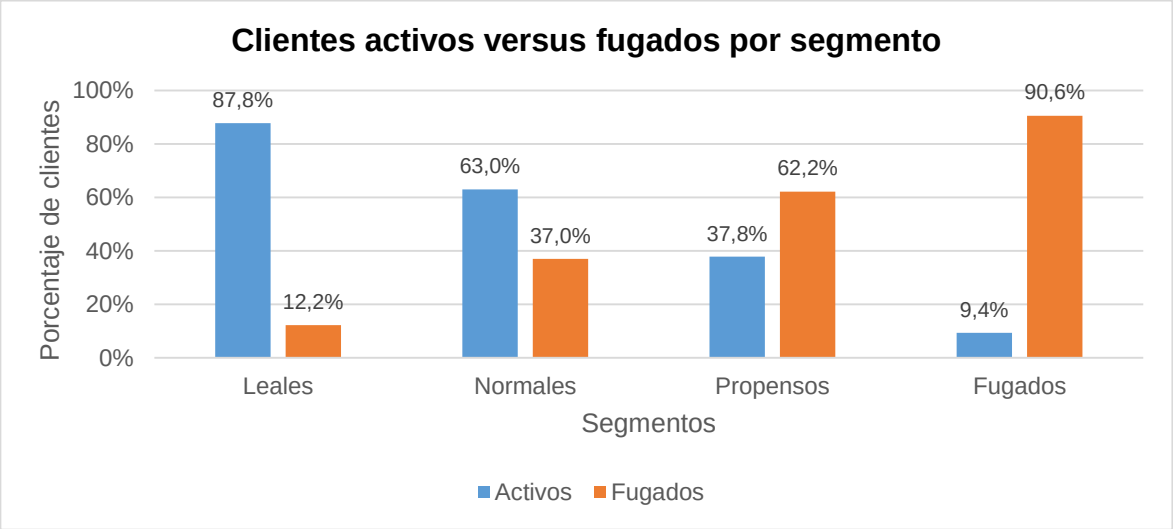

Ilustración 28: Clientes activos versus fugados por segmento.
Fuente: Elaboración Propia.

En la Tabla 18 se observa el promedio de las principales variables de los diferentes segmentos, donde se tiene que los valores tanto del recency y R/F van aumentando a medida que las probabilidades de fuga aumentan, no así el resto de las variables, ya que el mayor número de transacciones las concentran aquellos clientes que poseen las probabilidades de fuga más bajas. En el caso de la variable de máxima inactividad destaca el segmento Propenso a Fugarse debido a que existen períodos donde no se efectúan compras cercanas a los 5 meses y el monto promedio más alto lo posee el segmento Leal con $371.860.

	R	R/F	Número de Transacciones	Máxima Inactividad	Monto Promedio
Leales	35,661	5,256	240	97	$ 371.860
Normales	66,283	9,512	124	121	$ 331.056
Propensos a fugarse	101,606	12,225	28	144	$ 285.057
Fugados	419,834	67,497	16	53	$ 192.456

Tabla 18: Principales variables de los diferentes segmentos.
Fuente: Elaboración Propia.

Finalmente analizados los segmentos y las probabilidades de fuga es importante destacar que se debe poner énfasis en aquellos clientes que tienen probabilidades mayores a 0,5, ya que se son los clientes más propensos a fugarse, por lo que los recursos y acciones de retención que se apliquen deben enfocarse en estos clientes.

12. CAUSAS DE FUGA Y ACCIONES DE RETENCIÓN DE CLIENTES

Para definir acciones de retención es fundamental el comprender los motivos por los cuales un cliente se fuga de la empresa, ya de que esta forma es más efectiva la acción que se realice sobre él.

Las principales causas de fugas se asocian a las variables que el modelo indicó como relevantes para determinar las probabilidades de fuga. Estas variables ordenadas en orden de importancia fueron: recency, variación de monto, número de transacciones, máxima inactividad y frecuencia, de manera que en la Ilustración 29 se observa las reglas más importantes para determinar si a un cliente le otorgó una probabilidad alta de fuga o no.

```
R > 0.753
|   Variación M > -0.064
|   |   R > 3.716: 1.0 {0.0=0, 1.0=22}
|   |   R ≤ 3.716: 0.0 {0.0=694, 1.0=71}
|   Variación M ≤ -0.064: 1.0 {0.0=1, 1.0=54}
R ≤ 0.753
|   #trans > -0.193
|   |   Variación M > -0.064
|   |   |   #trans > -0.005: 0.0 {0.0=753, 1.0=191}
|   |   |   #trans ≤ -0.005
|   |   |   |   F > -0.191: 0.0 {0.0=172, 1.0=17}
|   |   |   |   F ≤ -0.191: 1.0 {0.0=66, 1.0=181}
|   |   Variación M ≤ -0.064: 1.0 {0.0=0, 1.0=223}
|   #trans ≤ -0.193
|   |   Max Inactividad > 2.280: 0.0 {0.0=67, 1.0=28}
|   |   Max Inactividad ≤ 2.280
|   |   |   R > -0.166
|   |   |   |   Max Inactividad > -0.690: 1.0 {0.0=109, 1.0=281}
|   |   |   |   Max Inactividad ≤ -0.690: 0.0 {0.0=26, 1.0=0}
|   |   |   R ≤ -0.166: 1.0 {0.0=238, 1.0=1262}
```

Ilustración 29: Reglas de asignación de probabilidades de fuga.
Fuente: Rapid Miner.

Debido a que las variables se encuentran transformadas no se observa que valores son determinantes a la hora de definir probabilidades y clasificación, por lo que se filtró la base de datos procesados para encontrar clientes que presentaran estas características, las cuales son:

- Recency > 100 días.
  Variación de monto ≤ -0,6.
  $\bar{x}$ Probabilidad de fuga: 0,95.

- Recency > 100 días.
  Variación de monto > -0,6.
  Recency > 145 días.
  $\bar{x}$ Probabilidad de fuga: 0,93.

- Recency ≤ 100 días.
  Número de transacciones > 49 transacciones.
  Variación de monto ≤ -0,6.
  $\bar{x}$ Probabilidad de fuga: 0,86.

- Recency $\leq$ 100 días.
  Número de transacciones $\leq$ 49 transacciones.
  Máxima inactividad $\leq$ 65 días
  Recency $\leq$ 67 días (aproximadamente 2 meses).
  $\bar{x}$ Probabilidad de fuga: 0,81.

- Recency $\leq$ 100 días.
  Número de transacciones $\leq$ 49 transacciones.
  Máxima inactividad $\leq$ 65 días
  Recency $>$ 67 días (aproximadamente 2 meses).
  Máxima inactividad $>$ 7 días
  $\bar{x}$ Probabilidad de fuga: 0,73.

- Recency $\leq$ 100 días.
  Número de transacciones $>$ 49 transacciones.
  Variación de monto $>$ -0,6.
  Número de transacciones $\leq$ 87 transacciones.
  Frecuencia $\leq$ 9,53.
  $\bar{x}$ Probabilidad de fuga: 0,67.

De manera que las acciones de retención se deben hacer previamente a que el cliente se etiquete en el segmento fugado, ya que la probabilidad de que se fugue del sistema es altísima. Bajo este concepto, las políticas de marketing propuestas serían:

- El tiempo que ha transcurrido desde la última compra es fundamental en la retención de un cliente, por lo que al presentar un rango de [60-75] días sin actividad se debe emitir una alerta de posible fuga para que el vendedor genere un llamado o una visita al respectivo cliente. También se puede emitir un descuento en su próxima compra exigiendo que ésta sea en un tiempo menor a 60 días, ya que es posible identificar en que clasificación de categorías el cliente efectúa mayores montos de compra (ver Tabla 19), por lo que este descuento sería personalizado, acción que también aplica para la variable máxima inactividad.

Clasificación Categorías			
Obra Gruesa	**Obra Intermedia**	**Terminaciones**	**Otras**
Construcción	Electricidad	Aperturas /Puertas y ventanas	Automotor
	Ferretería	Ampolletas y tubos	Bolsas Calientes
	Herramientas	Baños y cocinas	Concepto Finanzas
	Maderas	Electro-Hogar	Jardín y mascotas
	Plomería	Flooring	Material de empaque
		Iluminación	Material de uso interno
		Materiales de aseo	Menaje y decoración
		Muebles	Outdoor
		Organizadores	Uniformes
		Pinturas	Servicios
		Textil Hogar	Otros

Tabla 19: Clasificación de categorías de productos.
Fuente: Elaboración Propia.

- Las variables frecuencia y número de transacciones están ligadas directamente al cliente e indirectamente al vendedor, ya que es este último quien incentiva en primera instancia la compra, por lo que es interesante de realizar gestiones sobre éste, de manera de aumentar y potenciar la presencia del cliente en la empresa.
  De acuerdo a estas variables se debe aumentar la frecuencia de visita o de llamado a los clientes, con el fin de monitorear la situación actual por la cual están pasando (comienzo, desarrollo o final de un proyecto), de manera de intensificar el número de transacciones asociados a los montos de compras cuando se está iniciando o se está en pleno desarrollo de algún proyecto.

- La variable variación de monto se relaciona directamente con el monto de compra de las transacciones efectuadas, por lo que se observa en la Ilustración 30 que los clientes son más propensos a fugarse cuando han alcanzado un monto de compra cercano a los $ 5MM, por lo que es necesario generar metas por montos obtenidos otorgando descuentos o regalos, de manera que el cliente se sienta considerado por la empresa, de esta forma se podría genera un club de fidelización que haga diferencias entre los clientes de acuerdo a cuan valiosos son para ésta, de manera que los que obtengan mayores montos transados puedan optar a beneficios que otros no tienen y así sentir que son valorados por la empresa.

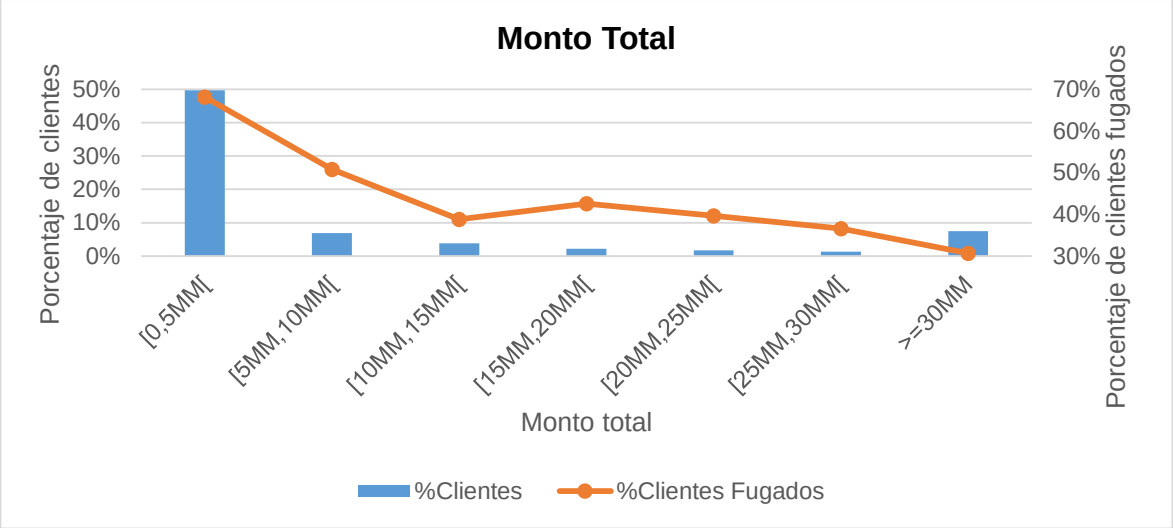

Ilustración 30: Distribución de monto total de clientes y/o fugados Carterizados. Fuente: Elaboración Propia.

Existen otras variables que son relevantes de considerar para definir las causas de fuga a pesar de no formar parte de la construcción del Árbol de decisión, siendo estas, monto negativo y porcentaje de monto de notas de créditos. Estas variables fueron las más significativas para definir si un cliente era fugado o no cuando se ejecutó el modelo sin las variables RFM y además se tenía un recency menor a 30 días, es decir, se trataban de clientes que habían efectuado compras en menos de un mes.

La variable monto negativo tiene directa incidencia con las notas de créditos, devolución y retail (devoluciones por canal Mesón) hechos por los clientes. Si bien es cierto una devolución es una nota de crédito, la empresa específica en su descripción cuando se trata de la devolución de materiales.

En las Ilustraciones 31 y 32 se observan ambas variables de los clientes propensos a fugarse clasificado por este modelo, los cuales corresponden al 32,5% de la muestra (4456 clientes).

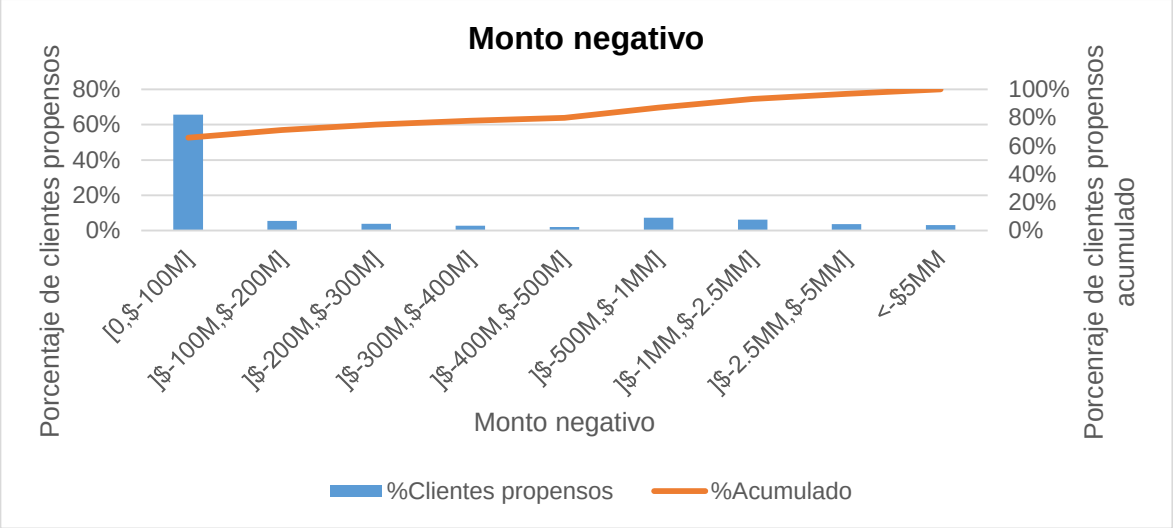

Ilustración 31: Distribución de monto negativo de clientes propensos a fugarse.
Fuente: Elaboración Propia.

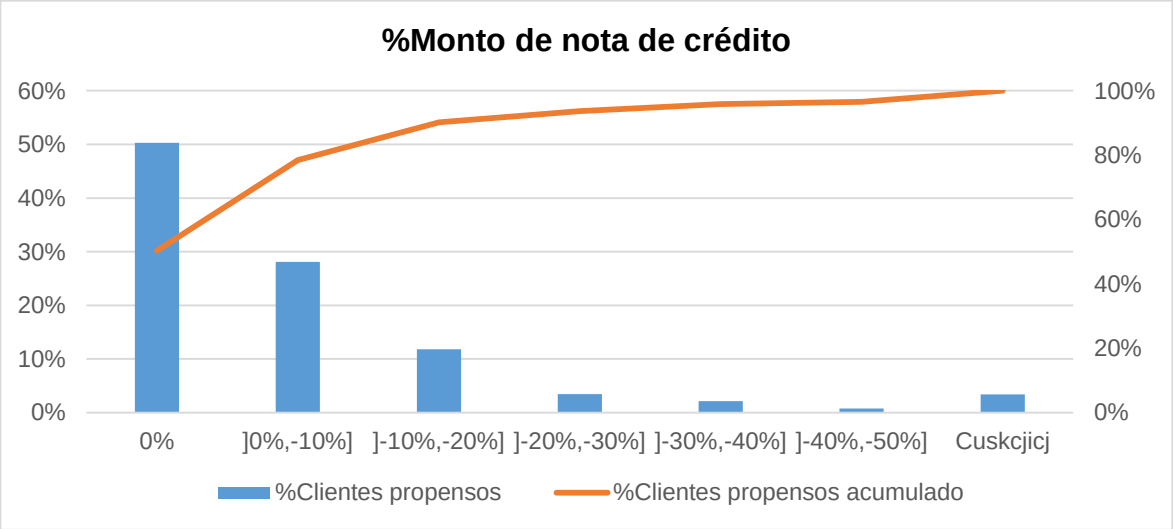

Ilustración 32: Distribución de porcentaje de monto de notas de créditos.
Fuente: Elaboración Propia.

Se tiene que con respecto a la variable monto negativo el mayor porcentaje de clientes propensos a fugarse se concentra en el rango de]0,$-100M] con un 68,5% de estos, siendo ya más de la mitad, de manera que cuando el monto negativo comienza a aumentar existe un alta probabilidad de que se fugue.

Finalmente se observa la variable porcentaje de monto de nota de crédito, donde la mayor concentración de clientes propensos a fugarse se produce cuando se tiene 0% y]0%, -10%] siendo 50,3% y 28,1% respectivamente. A pesar de que existe la mitad de clientes que no se ven representando por esta variable el restante si, por lo que es destacable de considerar. De esta manera los clientes son más propensos a fugarse a medida que esta variable aumenta, es decir, cuando más del 10% de su monto negativo está asociado a las notas de créditos, donde los principales motivos corresponde a una mala gestión en el despacho de productos (atrasos) o disponibilidad de estos cuando se requiere enviar,

por lo que claramente la acción recae en la empresa, quien podría mejorar sus niveles de stock de productos para no producir quiebres y determinar políticas de envíos asociadas a plazos de entrega para no retrasar los pedidos, considerando que si fallase se podría otorgar algún regalo o beneficio para suplir la insatisfacción del cliente.

12.1 Análisis económico de resultados

Dentro de cada segmento es posible encontrar clientes que poseen altos montos de compras versus otros que son muy bajos, de manera que los recursos que posee la empresa se deben enfocar sobre aquellos clientes considerados como “valiosos”, ya que generan importantes ingresos para la compañía.

De esta forma el análisis económico se realizó sobre una fracción de los clientes, específicamente aquellos que poseen una probabilidad de fuga entre [0,5-0,75], considerando el 25% y 50% de aquellos que posean los mayores montos transados de los últimos 6 meses antes de su última fecha de compra. En la Tabla 20 se observa un análisis de sensibilidad suponiendo la retención desde el 1% hasta el 10% de los clientes ya sea para el 50% o el 25% de los clientes con mayores montos.

Retención de clientes	50% de clientes		25% de clientes	
	Monto 6 últimos meses	USD	Monto 6 últimos meses	USD
1%	$ 22.734.817	$ 41.112	$ 17.847.019	$ 32.273
2%	$ 45.469.634	$ 82.224	$ 35.694.038	$ 64.546
3%	$ 68.204.452	$ 123.335	$ 53.541.056	$ 96.819
4%	$ 90.939.269	$ 164.447	$ 71.388.075	$ 129.092
5%	$ 113.674.086	$ 205.559	$ 89.235.094	$ 161.365
6%	$ 136.408.903	$ 246.671	$ 107.082.113	$ 193.639
7%	$ 159.143.720	$ 287.782	$ 124.929.131	$ 225.912
8%	$ 181.878.538	$ 328.894	$ 142.776.150	$ 258.185
9%	$ 204.613.355	$ 370.006	$ 160.623.169	$ 290.458
10%	$ 227.348.172	$ 411.118	$ 178.470.188	$ 322.731
Dólar	$553 al 30 de Junio 2014			

Tabla 20: Análisis económico de los clientes propensos a fugarse.
Fuente: Elaboración Propia.

Los clientes que presentan probabilidades en ese rango corresponden a 494 que representa el 11,15% de los clientes totales. Se tiene que para el 50% de los clientes el corte del monto se realizó en $ 2.745.147 a diferencia del 25% donde fue de $ 6.058.982.

Se observa que al retener el 5% de los clientes se podría obtener como ingresos USD $205.559 en el caso de considerar el 50% de los clientes versus USD $161.365 en el 25% de los clientes con mayores montos.

13. CONCLUSIONES Y RECOMENDACIONES

13.1 Sobre los criterios de fuga

Se definieron tres criterios de fuga donde las variables recency y R/F capturan el mayor porcentaje de clientes fugados versus la variable variación de monto. Esto se explica ya que este tipo de clientes compra por periodos de acuerdo al tiempo determinado para proyectos u obras, dejando de compra o atrasando su ciclo de compra bruscamente , sin embargo existe un grupo de clientes que presenta una disminución de sus montos de transacciones lo que da indicio a que es un cliente propenso a fugarse.

Una vez definidos los criterios se observó que en los datos procesados existían un porcentaje importante de clientes catalogados como fugados donde su fecha de fuga fue entre el año 2012 y 2013, representando un 48,8%% de los clientes totales.

La definición de un clientes fugado es totalmente aplicable a nuevos datos transaccionales, ya que al definir cotas máximas se puede dar alguna etiqueta a los futuros clientes que se registren.

13.2 Sobre los modelos de predicción de fuga

Si bien es cierto el modelo elegido fue el Árbol de decisión en relación al rendimiento de este versus Random Forest, su superioridad fue aproximadamente del 2% comparado con el mejor modelo una vez que se realizó el análisis de sensibilidad, por lo que la mejora es poco significativa.

Se hubiese esperado que el modelo Random Forest fuese mejor que un Árbol de decisión debido a que modela con muchos más árboles, sin embargo esto no fue así. Esto se produce ya que existían variables que eran muy relevantes para la clasificación de un cliente fugado propias de los criterios por lo que el modelo falla debido a la correlación que se produce entre la construcción de estos árboles aportando directamente en el error global. Este modelo es sensible en la cantidad de variables y árboles a construir afectando al rendimiento, indicadores (MDA y MDG) y al tiempo de ejecución cuando se está entrenando para su testeo, por lo que la optimización de estos parámetros es bastante relativo, incluso la literatura no tiene información concreta sobre esto. De acuerdo a las fuentes de información se tiene que este modelo entrega buenos resultados cuando se trabaja con una gran cantidad de datos y variables, ya que examina e incluye a todas éstas, sin embargo la desventaja que presenta y se observó en este trabajo es que es difícil de interpretar debido al alto número de árboles versus solo uno.

Al igual que la definición de criterios de fuga estos modelos son replicables para nuevos datos transaccionales, ya que de esta forma se pueden obtener las probabilidades de fugas para futuros clientes.

13.3 Sobre los resultados obtenidos

Dentro de los resultados obtenidos por ambos modelos la gran diferencia que se produce es la distribución de probabilidades de fuga, ya que el Árbol de decisión registró un mayor porcentaje de clientes con baja probabilidad que el modelo Random Forest como también una distribución más homogénea para este último, sin embargo para ambos casos a medida que aumentaba la probabilidad de fuga se obtiene el mayor porcentaje de clientes catalogados como fugados.

Debido a que el modelo escogido fue el Árbol de decisión, se determinaron 4 grupos de clientes dentro de los clientes Carterizados, es decir, Leales, Normales, Propensos a fugarse y Fugados, donde el mayor número de clientes se tiene en clientes Leales y Fugados.

Las acciones de retención se deben enfocar en una fracción de clientes que aporten altos ingresos para la empresa y que presenten una probabilidad entre [0,5 y 0,75], ya que un valor superior a este es prácticamente un cliente fugado de acuerdo a los criterios definidos, por lo que es fundamental entender las principales causas de fuga para orientar estas acciones.

En ambos modelos se obtuvo que dentro de las variables más relevantes asociadas a la determinación de las probabilidades de fuga se encuentran: recency, variación de monto, número de transacciones, máxima inactividad y frecuencia, de manera que las principales acciones se asociaron a generar descuentos o promociones personalizadas (categorías donde realiza transacciones) dentro de un plazo para que el cliente vuelva a efectuar una compra como también un posible club de fidelización al igual que otras industrias.

También es importante destacar las variables de monto negativo y de porcentaje de monto de nota de crédito, ya que aportan a la identificación de un cliente propenso a fugarse a medida que su monto y porcentaje comienzan a aumentar significativamente en un periodo menor a 6 meses.

Los beneficios que se podrían optar dada la factibilidad de llevar a cabo las acciones de retención debido a su baja complejidad, son considerables, ya que en primera instancia aportan dinero y adicionalmente permiten mantener y cuidar a los clientes, quienes son fundamentales en este negocio.

13.4 Recomendaciones y trabajos futuros

Una vez obtenidas las conclusiones de este trabajo se desprenden recomendaciones y acciones futuras que se relacionan con la empresa y el comportamiento de compra de los clientes Carterizados, con el fin de mejorar la situación actual que presentan, de manera que se tiene lo siguiente:

- Los vendedores que interactúan con los clientes son fundamentales para esta línea de negocio, por lo que se podría estandarizar el concepto de cliente Carterizado, ya que se indicó que un cliente forma parte de este segmento cuando el jefe o el vendedor lo deciden, de manera que se podrían determinar

características o patrones relacionados a su comportamiento de compra que lo clasifiquen en clientes Carterizados, Mesón o No Carterizados.

- Se observó a través de trabajo que la asignación de clientes a los vendedores es al azar, por lo que no existe la optimización de cartera, de manera que se tiene vendedores con muchos clientes y otros con muy pocos, por lo que se podría generar una optimización de esta.

- Los vendedores no llevan registros de los pedidos de los clientes, por lo que se podría incorporar la tecnología con el fin de registrar estas solicitudes de manera de sugerir productos, promociones y/o descuentos relacionados a las categorías donde los clientes compran para potenciar las ventas.

- Existen jefes de vendedores por sucursales de manera que se podría inspeccionar a través de visitas o llamadas a los clientes cuando deberían estar siendo atendidos por los vendedores, de igual forma se debería efectuar encuestas sobre satisfacción a los clientes sobre su relación con la empresa y el desempeño de los vendedores de forma periódica.

- Sería interesante determinar la métrica CLV o Customer Lifetime Value, ya que ayuda a potenciar la relación con el cliente Carterizado debido a que determina el valor que representa para la empresa la relación con algún cliente a lo largo de la vida.[18]. De esta forma trataría de aumentar la cuota por cliente y ayudaría en potenciar y dimensionar las campañas de retención al igual que revisar la propuesta de valor de la empresa.

- Finalmente no se tiene un registro de reclamos, por lo que generar uno podría beneficiar a la empresa para detectar otras causas de fuga de los clientes y así mejorar en la relación contractual con ellos.

14. BIBILIOGRAFÍA

- [1]. Bastías Larraín, Priscilla Denise, “Metodología para realizar predicción de fuga de clientes en una empresa de retail”, Memoria Ingeniería Civil Industrial, Universidad de Chile, 2009.

- [2]. Segovia Riquelme, Carolina Andrea “Caracterización del proceso de fuga de clientes en retail banking utilizando información transaccional”, Memoria Ingeniería Civil Industrial, Universidad de Chile, 2005.

- [3]. Videla Araya, María Inés, “Metodología para diseñar acciones de retención de clientes no contractuales en una empresa de retail”, Memoria Ingeniería Civil Industrial, Universidad de Chile, 2011.

- [4]. Segovia, C, Aburto, L; Goic M., “Caracterización del proceso de fuga de clientes utilizando información transaccional” Penta Analytics, 2005.

- [5]. Neslin, Scott A., Gupta, Sunil; Kamakura, Wagner; Lu, Junxiang; Mazon, Charolette H, “Defection Detection: Measuring and Understanding the predictive accuracy or customer churns model”, Journal of Maketing research, v18, 204-211, 2006.

- [6]. Andy Liaw, Mathhew Wiener, “Classification and regression by Random Forest”, vol 2/3, 2002.

- [7]. Leo Breiman, “Random Forest, Machine learning, 45, 5-32, 2011.

- [8]. Tin Kam Ho, “Random decision forest”, USA, 2004.

- [9]. Yaya Xie, Xiu Li, E.W.T Ngai , Weiyun Ying, “Customer churn prediction using improved balanced random forest”,Expert system with applications, 2008.

- [10]. J.Burez, D.van den Poel, “Handling class imbalance in customer churn prediction”, Expert system with applications, 2009

- [11]. Terra. 2014. Banco Central: PIB DE Chile anotó alza de 4,1% en 2013. [en línea] 18 de Marzo 2014,http://economia.terra.cl/banco-central-pib-de-chile-anoto-alza-de-4.html[consulta: 6 de Abril 2012]

- [12]. Barrientos Inostroza, Francisco Javier, “Diseño e implementación de na metodología de predicción de fuga de clientes en una compañía de telecomunicaciones.”, Memoria Ingeniería Civil Industrial, Universidad de Chile, 2011.

- [13]. García Fernández, Raquel “Método de predicción de fuga con grandes volúmenes de datos”, Facultad de Ciencias, Universidad de Valladolid, 2010.

- [14]. Universidad Carlos II de Madrid. Transformación de variables. [en línea] http://halweb.uc3m.es/esp/Personal/personas/jmmarin/esp/EDescrip/tema4.pdf. [consulta: 9 de Mayo 2014]

- [15]. Departamento de la Computación, Universidad de Chile. Tutorial Rapid Miner [en línea] http://users.dcc.uchile.cl/~nbaloian/DSS-DCC/Software/Tutorial%20RapidMiner/Tutorial_I%20.pdf [consulta: 3 de Mayo 2014]

- [16]. M. J. Tapia, Ventas de retail crecerán 14,5% en Chile a 2012 y bajarán peso en la región a 3,1%, La Tercera, p. 25, 21 Febrero 2012.

- [17]. Taghi M.Khoshgoftaar, Moiz Golawala, Jason Van Hulse, “An emperical study of learning from imbalanced data using Random Forest”, USA.

- [18].Enrique Dans, Área de Sistemas de información en el instituto de empresa. [en línea] http://profesores.ie.edu/enrique_dans/download/clv.pdf [consulta: 24 de Agosto 2014].

15. ANEXOS

15.1 Anexo 1: Análisis de las variables calculadas.

- Número de transacciones: Indica el número total de transacciones.

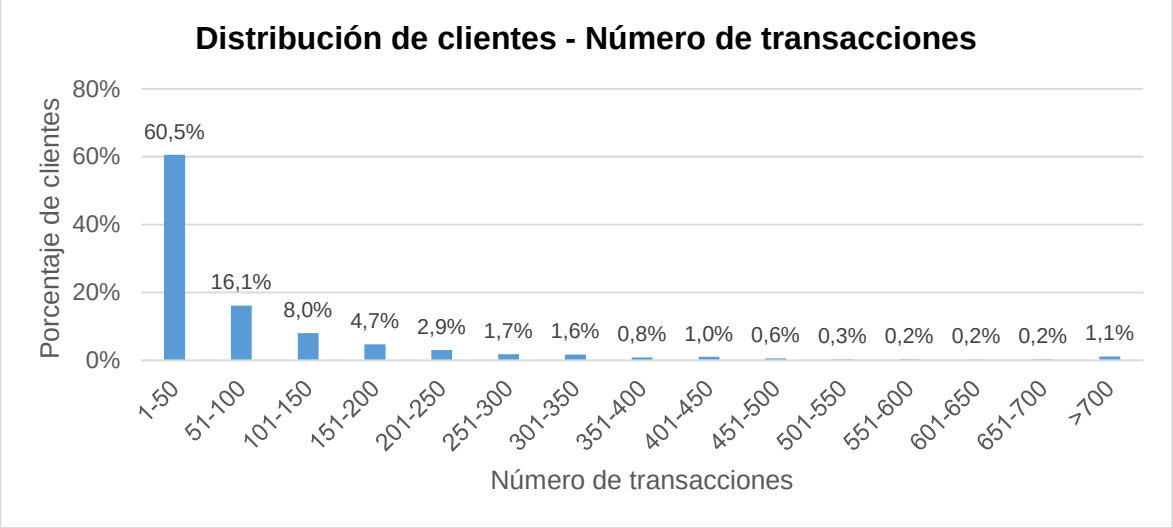

Ilustración 33: Distribución de clientes de acuerdo a su número de transacciones.
Fuente: Elaboración Propia.

- Número de transacciones por día: Indica el número de días en el cual se realizaron transacciones.

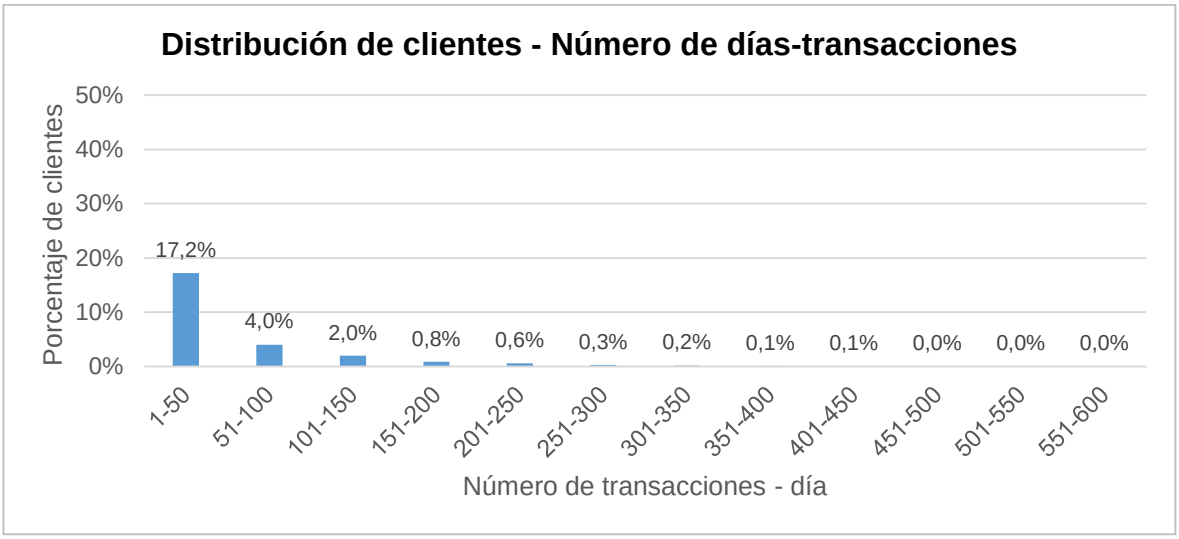

Ilustración 34: Distribución de clientes de acuerdo a su número de transacciones- día.
Fuente: Elaboración Propia.

- Antigüedad: Indica la cantidad de días que el cliente lleva registrado como cliente Carterizado.

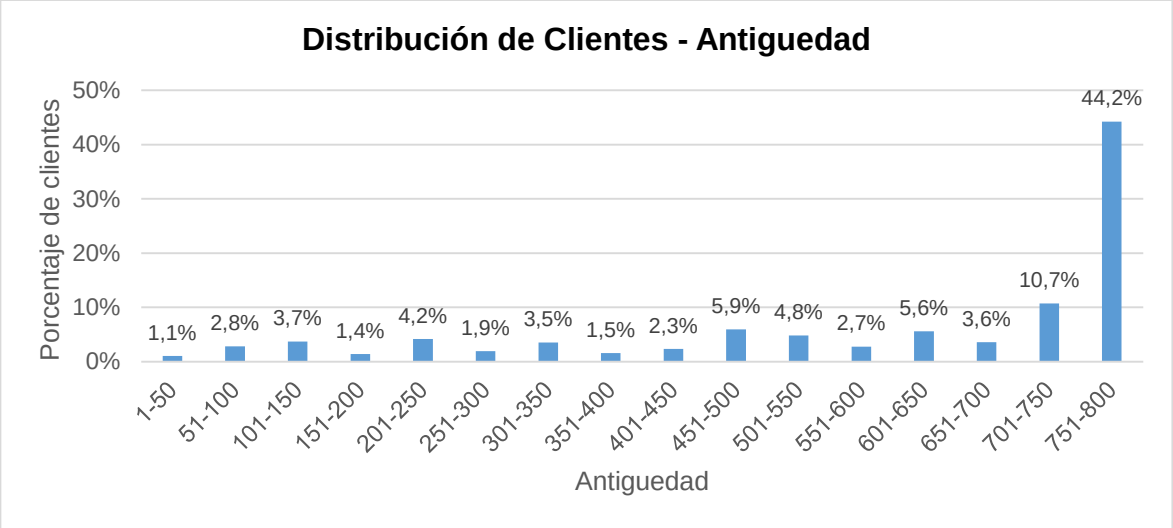

Ilustración 35: Distribución de clientes de acuerdo a su antigüedad.
Fuente: Elaboración Propia.

- Máxima Inactividad: Indica el máximo tiempo que transcurre entre transacciones.

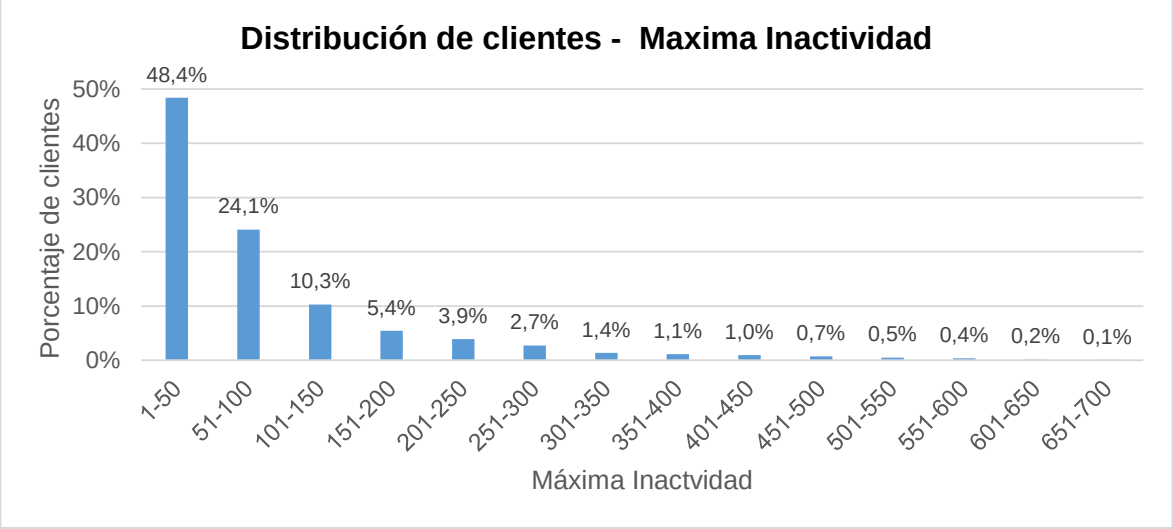

Ilustración 36: Distribución de clientes de acuerdo a su máxima inactividad.
Fuente: Elaboración Propia.

- Monto total: Indica el monto total facturado por cliente.

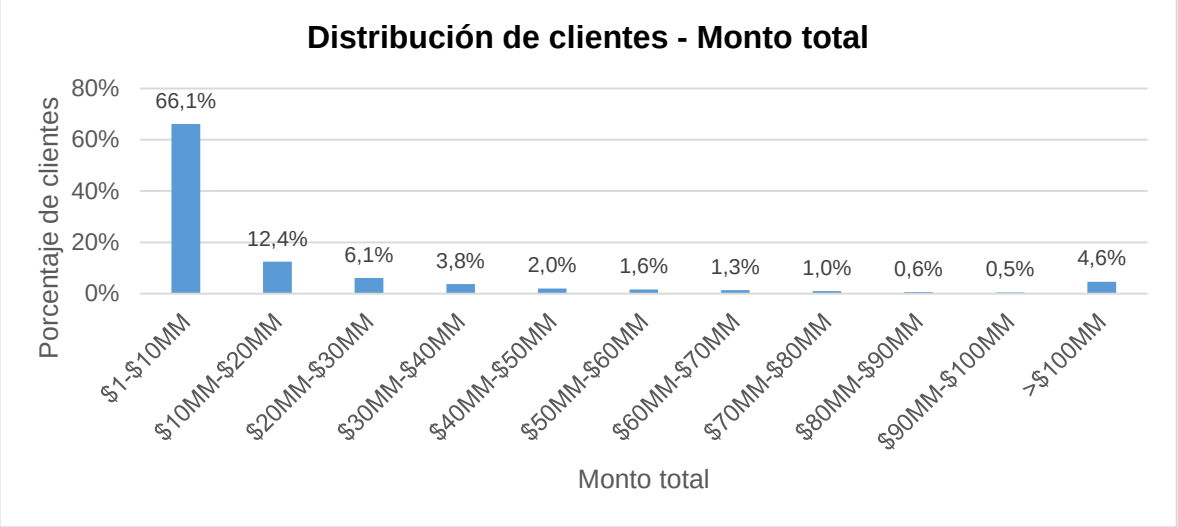

Ilustración 37: Distribución de clientes de acuerdo al monto total.
Fuente: Elaboración Propia.

- Variación de monto: Indica la variación de monto registrado en los últimos 6 meses, para ello se comparan dos trimestres consecutivos.

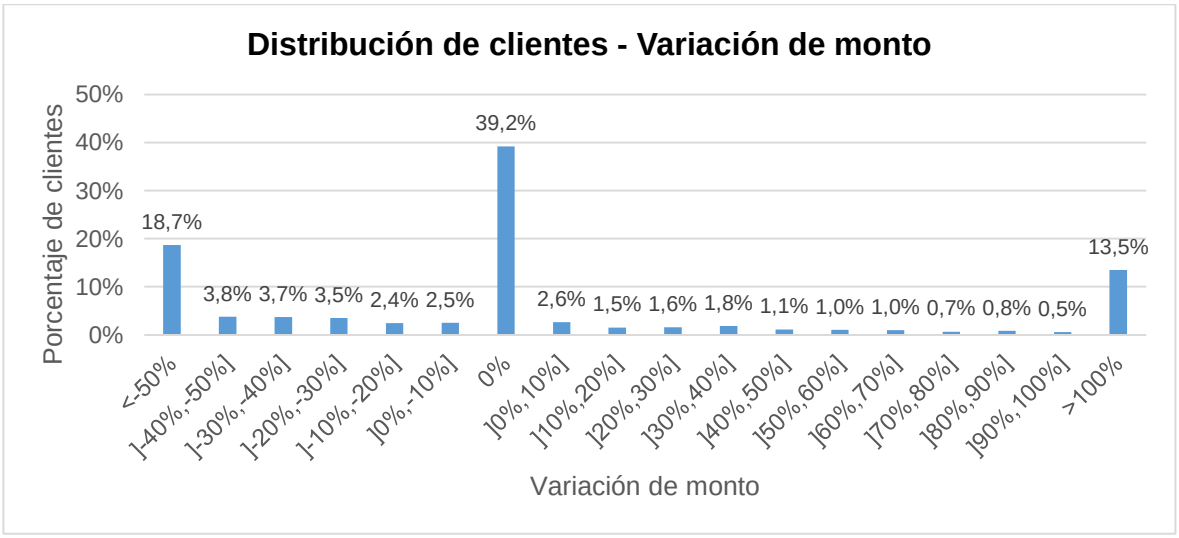

Ilustración 38: Distribución de clientes de acuerdo a la variación de monto.
Fuente: Elaboración Propia.

- Monto negativo total: Indica la suma de los montos de las transacciones que presentan montos negativos.

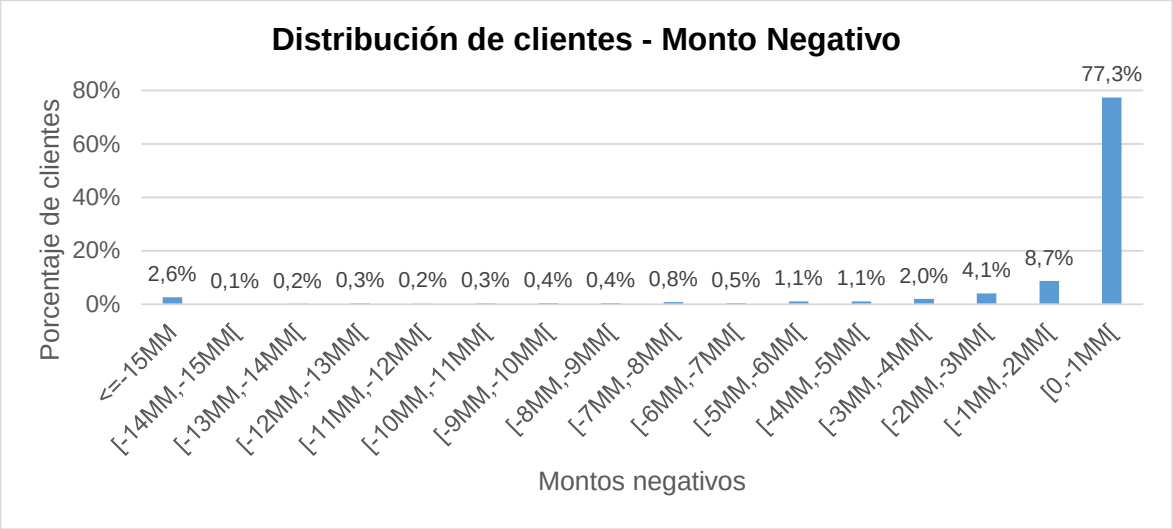

Ilustración 39: Distribución de clientes de acuerdo a su monto total negativo.
Fuente: Elaboración Propia.

- Número de transacciones negativas: Indica el número total de transacciones que presentan montos negativos.

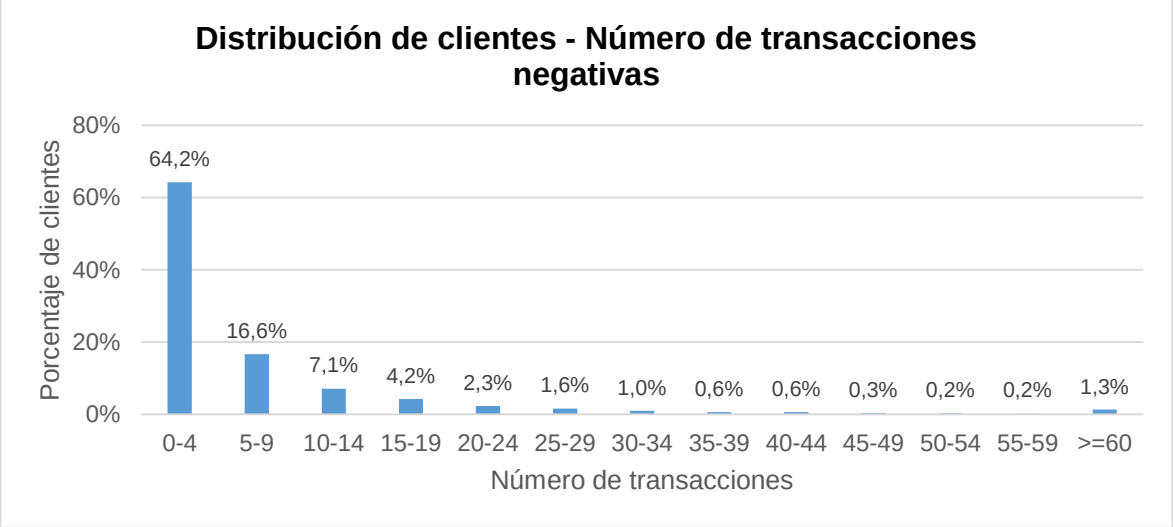

Ilustración 40: Distribución de clientes de acuerdo al número de transacciones negativas.
Fuente: Elaboración Propia.

- Porcentaje de monto de transacciones negativas: Indica el porcentaje que representa el monto de transacciones negativas sobre el monto total por cada cliente.

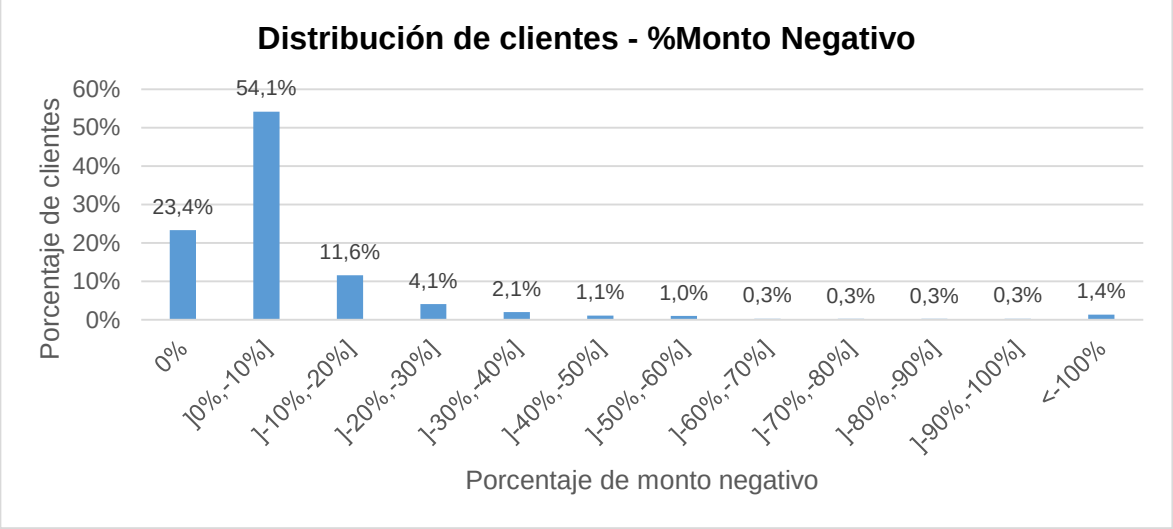

Ilustración 41: Distribución de clientes de acuerdo al porcentaje de monto de transacciones negativas. Fuente: Elaboración Propia.

- Región: Indica a que región del país pertenece el cliente registrado.

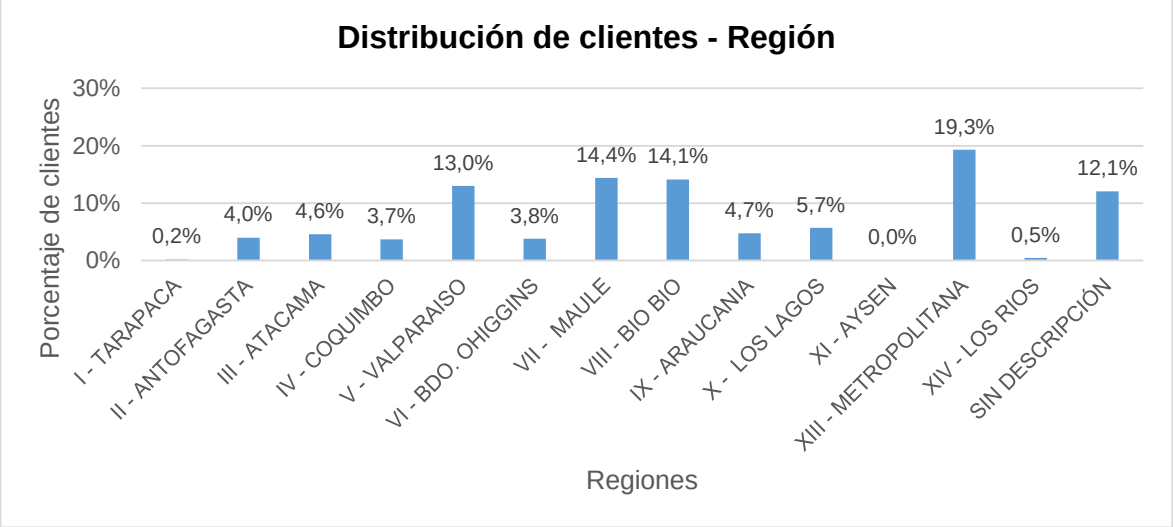

Ilustración 42: Distribución de clientes de acuerdo a su región. Fuente: Elaboración Propia.

- Número de devoluciones: Indica el número total de transacciones negativas que corresponden a devoluciones.

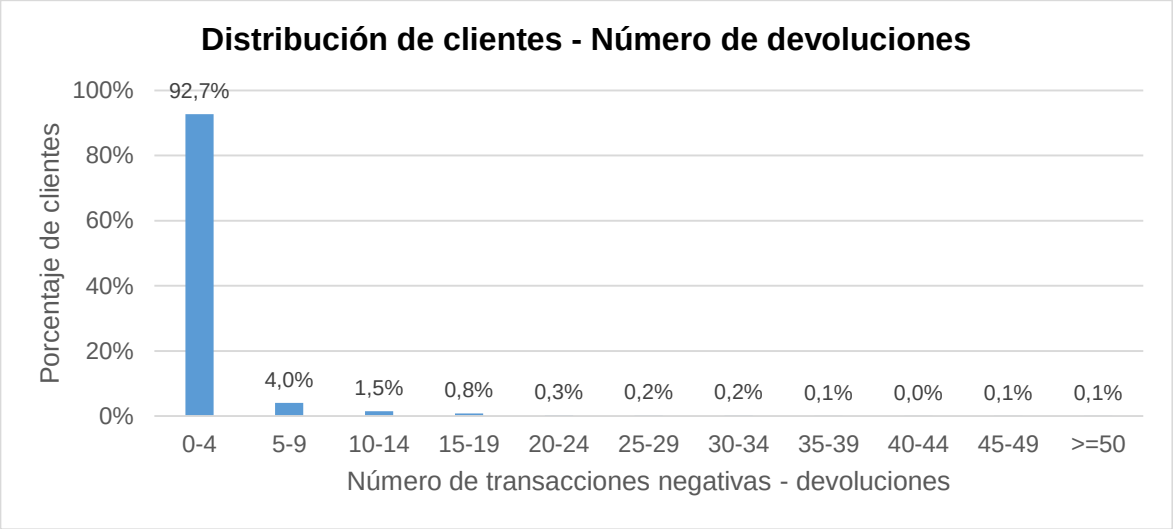

Ilustración 43: Distribución de clientes de acuerdo al número de devoluciones.
Fuente: Elaboración Propia.

- Porcentaje de monto de devoluciones: Indica el porcentaje que representa el monto de las devoluciones sobre el monto total.

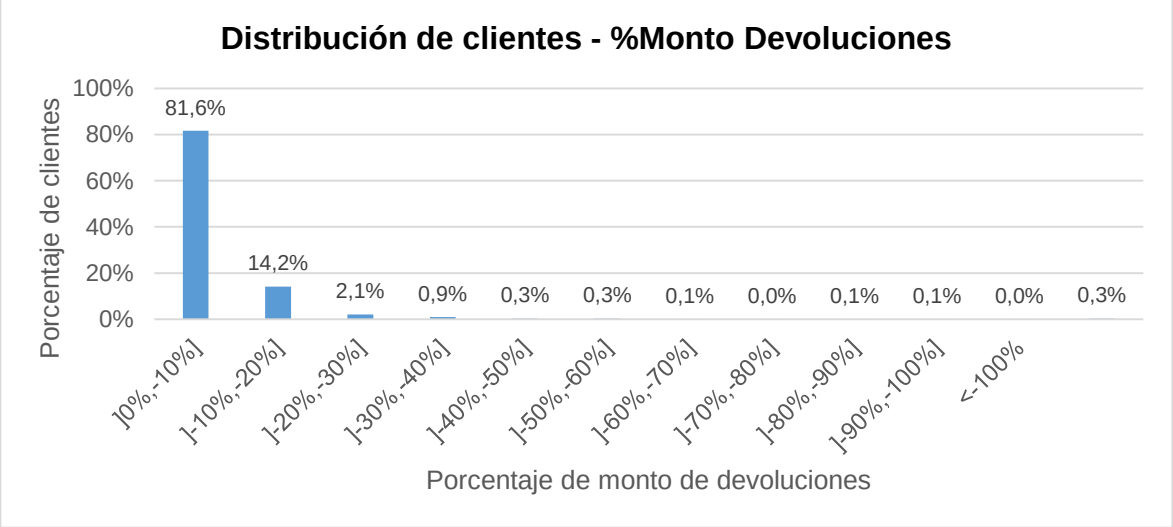

Ilustración 44: Distribución de clientes de acuerdo al porcentaje de monto de devoluciones.
Fuente: Elaboración Propia.

- Número de notas de créditos: Indica el número total de transacciones negativas que corresponden a notas de créditos.

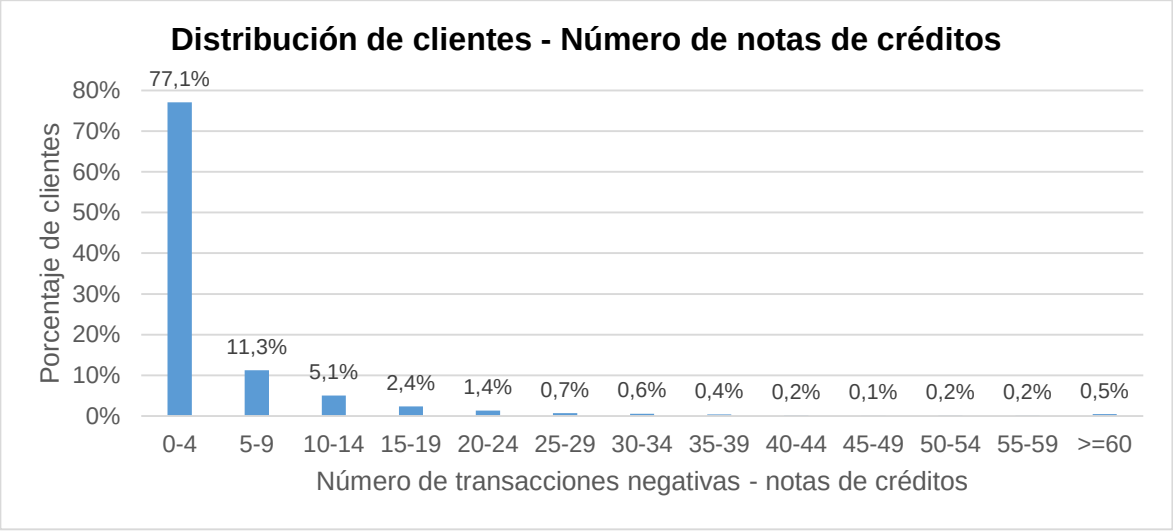

Ilustración 45: Distribución de clientes de acuerdo al número de notas de créditos.
Fuente: Elaboración Propia.

- Porcentaje de monto de notas de créditos: Indica el porcentaje que representa el monto de las notas de créditos sobre el monto total.

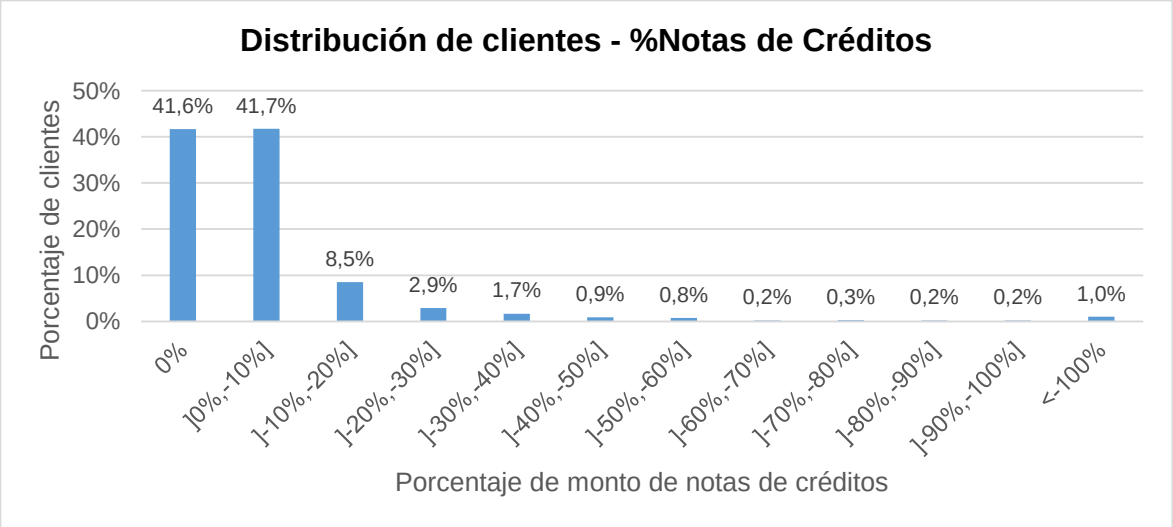

Ilustración 46: Distribución de clientes de acuerdo al porcentaje de monto de notas de créditos.
Fuente: Elaboración Propia.

- Número de retail: Indica el número total de transacciones negativas que corresponden a retail. Este tipo de transacciones son devoluciones a través del canal mesón.

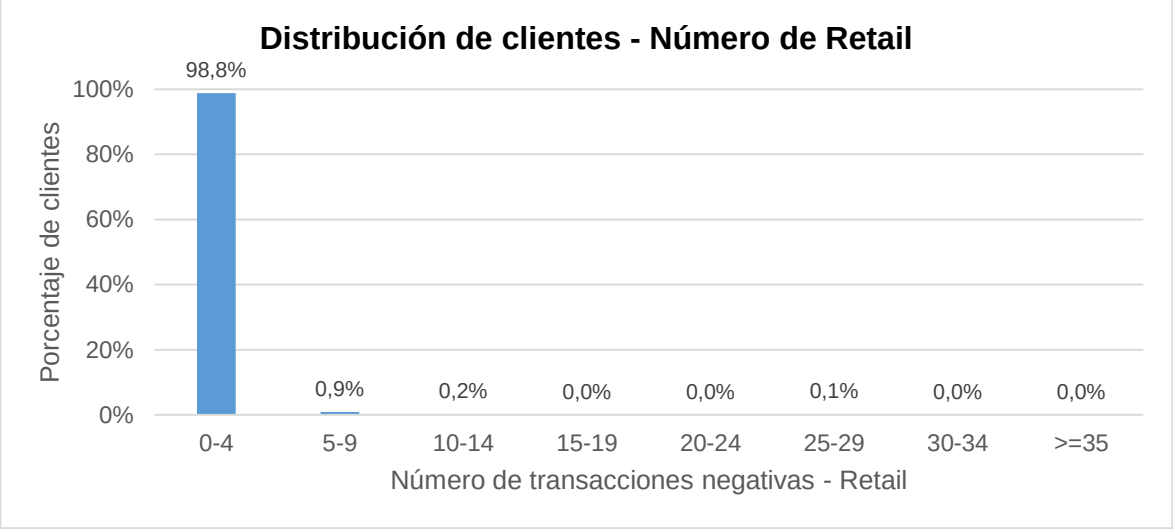

Ilustración 47: Distribución de clientes de acuerdo al número de transacciones negativas tipo retail. Fuente: Elaboración Propia.

- Porcentaje de monto de retail: Indica el porcentaje que representa el monto de retail sobre el monto total.

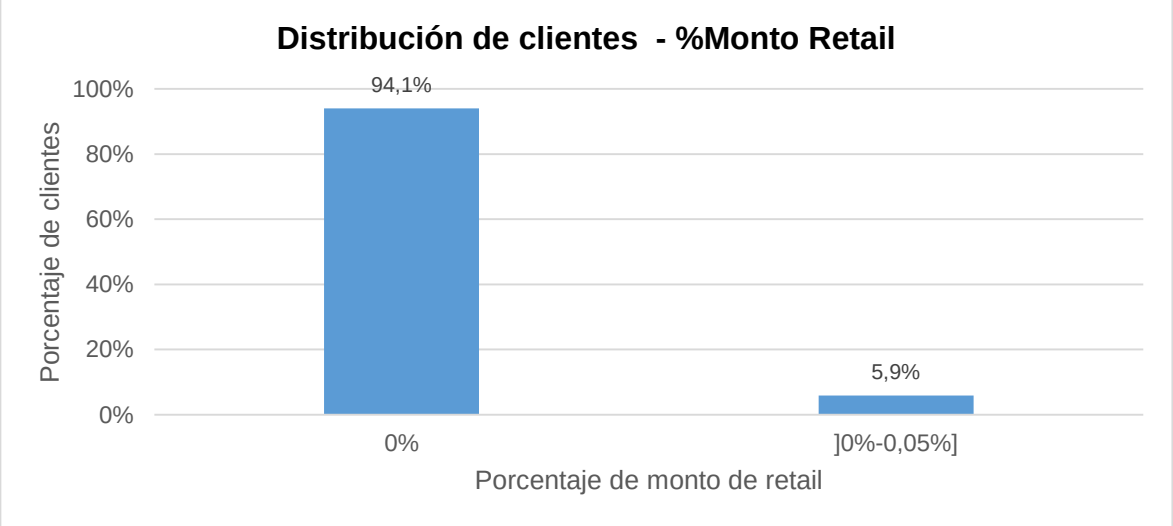

Ilustración 48: Distribución de clientes de acuerdo a porcentaje de monto de retail. Fuente: Elaboración Propia.

- Giro Comercial: Indica a que rubro comercial pertenece el cliente registrado.

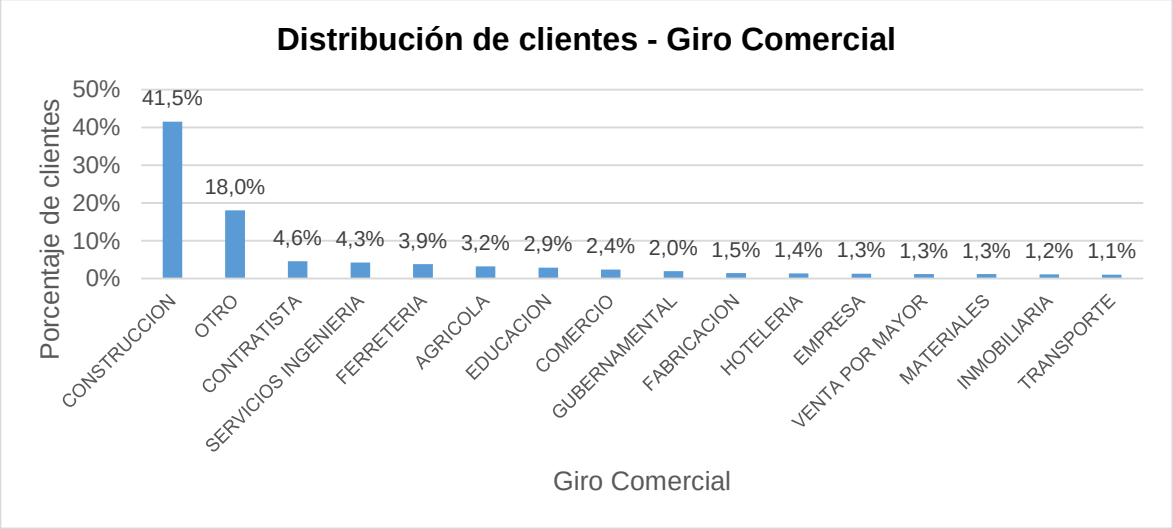

Ilustración 49: Distribución de clientes de acuerdo a su giro comercial.
Fuente: Elaboración Propia.

- Porcentaje de monto de obra gruesa: Indica el porcentaje que representa el monto de las transacciones que contienen productos clasificados como obra gruesa sobre el monto total.

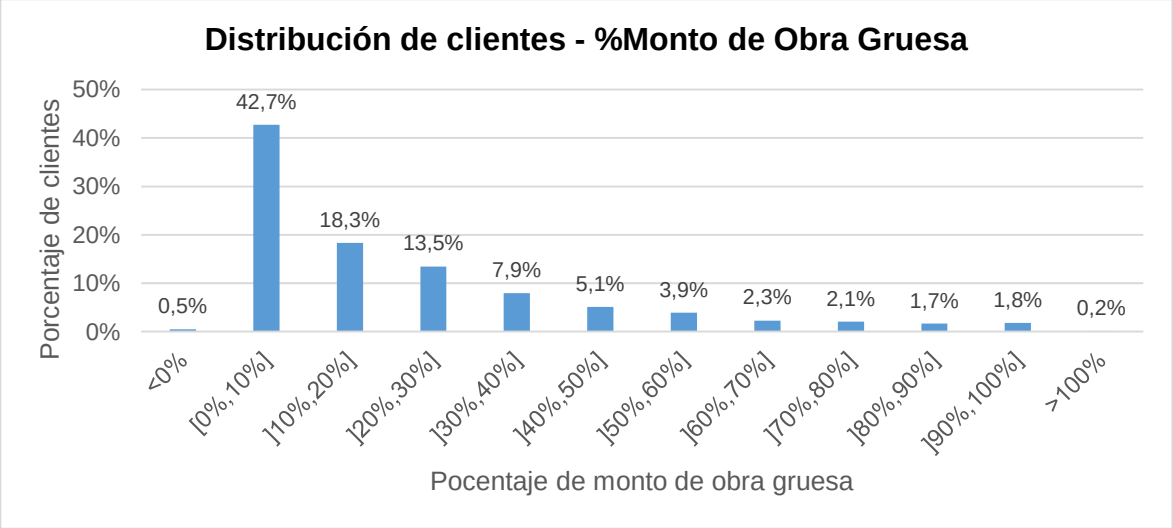

Ilustración 50: Distribución de clientes de acuerdo al porcentaje de monto de obra gruesa.
Fuente: Elaboración Propia.

- Porcentaje de monto de obra intermedia: Indica el porcentaje que representa el monto de las transacciones que contienen productos clasificados como obra intermedia sobre el monto total.

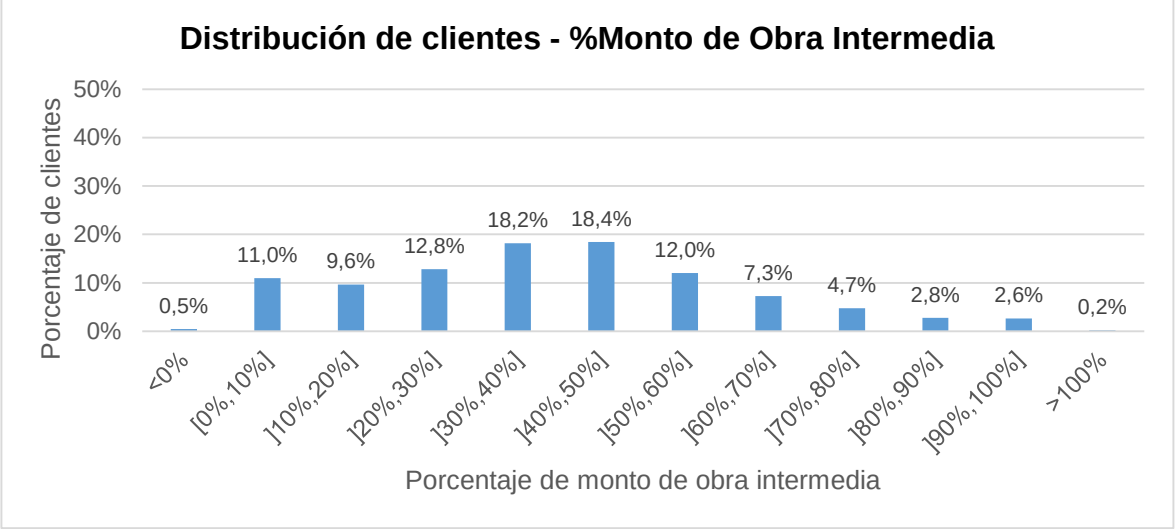

Ilustración 51: Distribución de clientes de acuerdo al porcentaje de monto de obra intermedia.
Fuente: Elaboración Propia.

- Porcentaje de monto de terminaciones: Indica el porcentaje que representa el monto de las transacciones que contienen productos clasificados como terminaciones sobre el monto total.

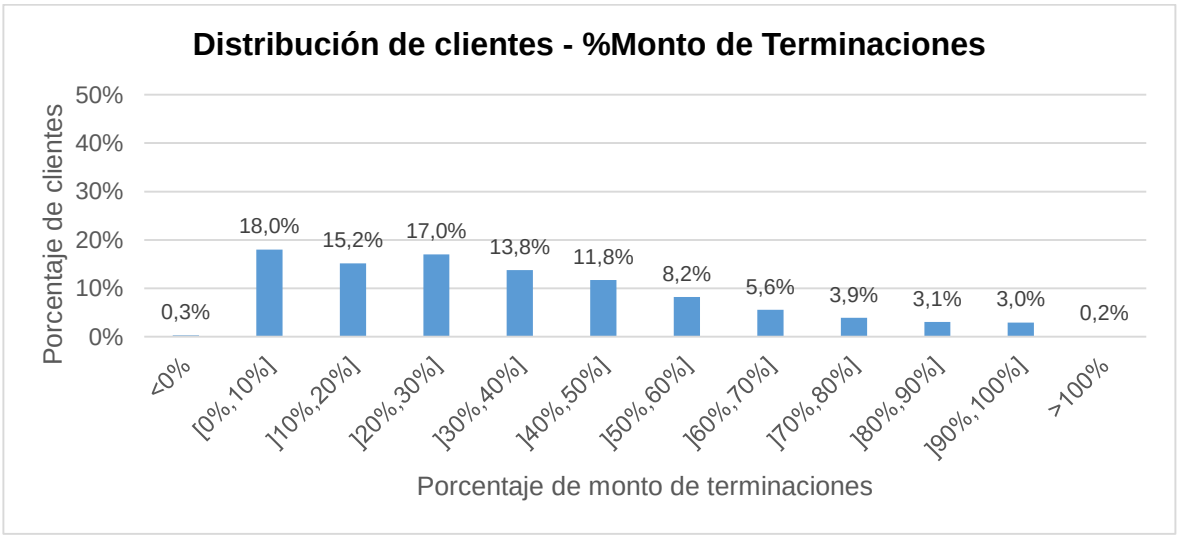

Ilustración 52: Distribución de clientes de acuerdo al porcentaje de monto de terminaciones.
Fuente: Elaboración Propia

- Porcentaje de monto de otro: Indica el porcentaje que representa el monto de las transacciones que contienen productos clasificados como otros sobre el monto total.

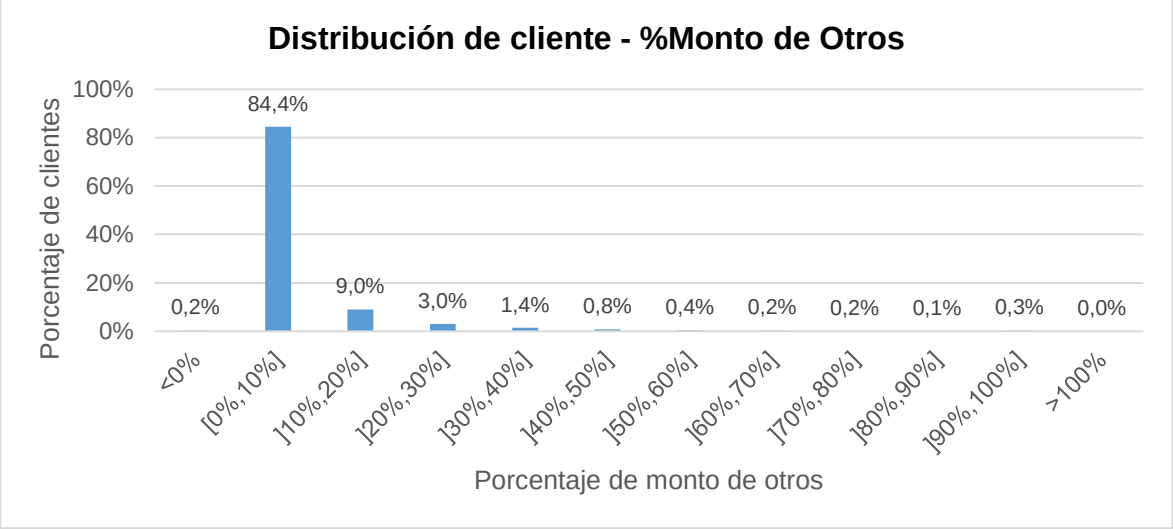

Ilustración 53: Distribución de clientes de acuerdo al porcentaje de monto de otros. Fuente: Elaboración Propia.

15.2 Anexo 2: Árbol de decisión - Entrenamiento

El nivel de certeza para el modelo predictivo en la partición de entrenamiento fue de 82,2%, donde la matriz de confusión se observa en la Tabla 21.

Matriz de confusión			
	True No fugado	True Fugado	Class precisión
Pred. NoFugado	1698	366	82,3%
Pred. Fugado	428	1964	82,1%
Class recall	79,9%	84,3%	

Tabla 21: Matriz de confusión - Entrenamiento para Árbol de decisión. Fuente: Rapid Miner.

Se observa que de los 2330 fugados reales, se detectaron 1964 por lo que se predice el 84,3% de los fugados reales (Sensibilidad) y de los 2392 clientes que el modelo clasifica como fugados, 1964 lo eran, lo cual corresponde al 82,1%(Precisión).

El área bajo la curva, es decir, AUC fue de 85,9% donde se observa en la Ilustración 54 que para obtener el 75% de un cliente fugado que el modelo predice como fugado se tendrá menos del 20% de un cliente no fugado que el modelo predice como fugado.

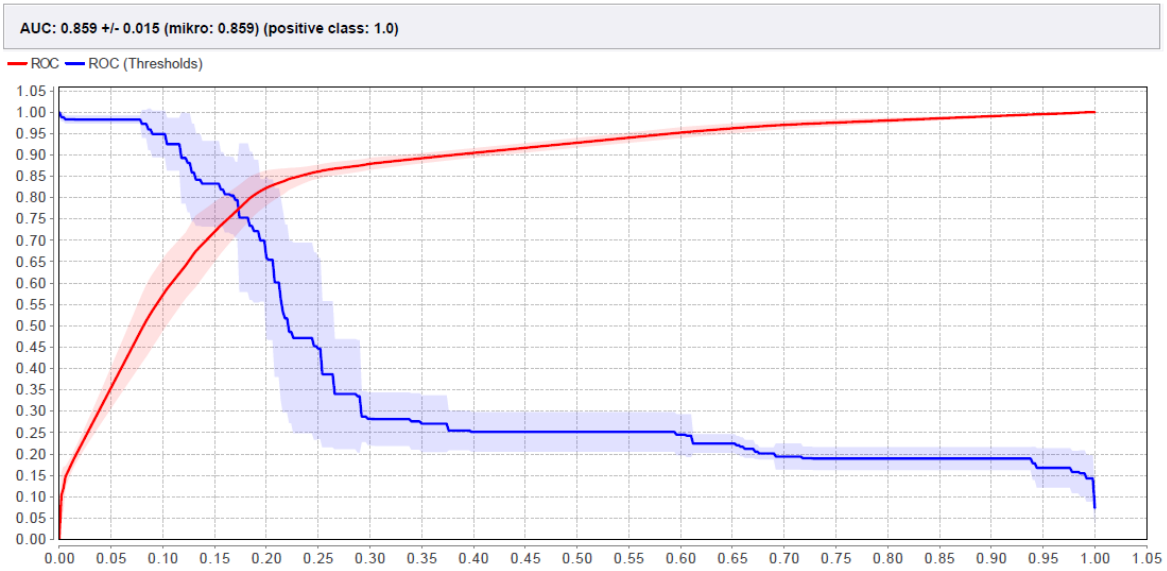

Ilustración 54: Curva AUC- Entrenamiento Árbol de decisión.
Fuente: Rapid Miner.

15.3 Anexo 3: Random Forest - Entrenamiento

El nivel de certeza para el modelo predictivo en la partición de entrenamiento fue de 73,7%, donde la matriz de confusión se observa en la Tabla 22.

Matriz de confusión			
	True No fugado	True Fugado	Class precisión
Pred. NoFugado	1465	509	74,2%
Pred. Fugado	661	1821	73,4%
Class recall	68,9%	78,2%	

Tabla 22: Matriz de confusión - Entrenamiento para Random Forest.
Fuente: Rapid Miner.

Se observa que de los 2330 fugados reales, se detectaron 1821 por lo que se predice el 78,2% de los fugados reales (Sensibilidad) y de los 2482 clientes que el modelo clasifica como fugados, 1821 lo eran, lo cual corresponde al 73,4%(Precisión).

El área bajo la curva, es decir, AUC fue de 82,2% donde se observa en la Ilustración 55 que para obtener el 65% de un cliente fugado que el modelo predice como fugado se tendrá alrededor del 15% de un cliente no fugado que el modelo predice como fugado.

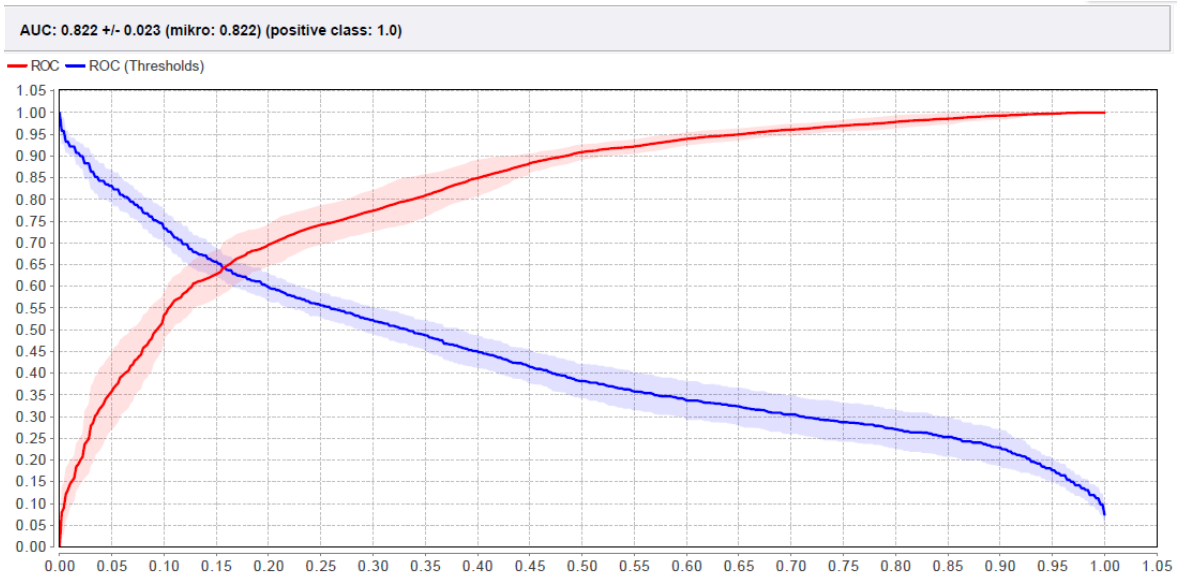

Ilustración 55: Curva AUC- Entrenamiento Random Forest.
Fuente: Rapid Miner.

15.4 Anexo 4: Matriz de confusión - Análisis de sensibilidad Random Forest.

- 50 Árboles de decisión.

Matriz de confusión: 22 variables.			
	True No fugado	True Fugado	Class precisión
Pred. NoFugado	165	47	22,2%
Pred. Fugado	44	190	81,2%
Class recall	21,1%	80,2%	
Matriz de confusión: 10 variables.			
	True No fugado	True Fugado	Class precisión
Pred. NoFugado	166	46	21,7%
Pred. Fugado	37	197	84,2%
Class recall	18,2%	81,1%	
Matriz de confusión: 5 variables.			
	True No fugado	True Fugado	Class precisión
Pred. NoFugado	169	43	20,3%
Pred. Fugado	33	201	85,9%
Class recall	16,3%	82,4%	

Tabla 23: Análisis de sensibilidad Random Forest - Matriz de confusión 50 árboles.
Fuente: Elaboración Propia.

- 100 Árboles de decisión.

Matriz de confusión: 22 variables.			
	True No fugado	True Fugado	Class precisión
Pred. NoFugado	166	46	21,7%
Pred. Fugado	43	191	81,6%
Class recall	20,6%	80,6%	
Matriz de confusión: 10 variables.			
	True No fugado	True Fugado	Class precisión
Pred. NoFugado	164	48	22,6%
Pred. Fugado	35	199	85,0%
Class recall	17,6%	80,6%	
Matriz de confusión: 5 variables.			
	True No fugado	True Fugado	Class precisión
Pred. NoFugado	169	43	20,3%
Pred. Fugado	34	200	85,5%
Class recall	16,8%	82,3%	

Tabla 24: Análisis de sensibilidad Random Forest - Matriz de confusión 100 árboles.
Fuente: Elaboración Propia.

- 200 Árboles de decisión.

Matriz de confusión: 22 variables.			
	True No fugado	True Fugado	Class precisión
Pred. NoFugado	165	47	22,2%
Pred. Fugado	44	190	81,2%
Class recall	21,1%	80,2%	
Matriz de confusión: 10 variables.			
	True No fugado	True Fugado	Class precisión
Pred. NoFugado	165	47	22,2%
Pred. Fugado	36	198	84,3%
Class recall	17,9%	80,8%	
Matriz de confusión: 5 variables.			
	True No fugado	True Fugado	Class precisión
Pred. NoFugado	166	46	21,7%
Pred. Fugado	37	197	84,2%
Class recall	18,2%	81,1%	

Tabla 25: Análisis de sensibilidad Random Forest - Matriz de confusión 200 árboles.
Fuente: Elaboración Propia.

- 500 Árboles de decisión.

Matriz de confusión: 22 variables.			
	True No fugado	True Fugado	Class precisión
Pred. NoFugado	164	48	22,6%
Pred. Fugado	44	190	81,2%
Class recall	21,2%	79,8%	
Matriz de confusión: 10 variables.			
	True No fugado	True Fugado	Class precisión
Pred. NoFugado	163	49	23,1%
Pred. Fugado	33	201	85,9%
Class recall	16,8%	80,4%	
Matriz de confusión: 5 variables.			
	True No fugado	True Fugado	Class precisión
Pred. NoFugado	167	45	21,2%
Pred. Fugado	34	200	85,5%
Class recall	16,9%	81,6%	

Tabla 26: Análisis de sensibilidad Random Forest - Matriz de confusión 500 árboles.
Fuente: Elaboración Propia.

15.5 Anexo 5: Árbol de decisión sin RFM – Entrenamiento.

El nivel de certeza para el modelo predictivo en la partición de entrenamiento fue de 65,5%, donde la matriz de confusión se observa en la Tabla 27.

Matriz de confusión			
	True No fugado	True Fugado	Class precisión
Pred. NoFugado	1197	607	33,6%
Pred. Fugado	929	1723	65,0%
Class recall	43,7%	73,9%	

Tabla 27: Matriz de confusión - Entrenamiento para Árbol de decisión sin RFM.
Fuente: Elaboración Propia.

Se observa que de los 2330 fugados reales, se detectaron 1723 por lo que se predice el 73,9% de los fugados reales (Sensibilidad) y de los 2652 clientes que el modelo clasifica como fugados, 1723 lo eran, lo cual corresponde al 65,0%(Precisión).

El área bajo la curva, es decir, AUC fue de 71,3% donde se observa en la Ilustración 56 que para obtener el 60% de un cliente fugado que el modelo predice como fugado se tendrá menos del 30% de un cliente no fugado que el modelo predice como fugado.

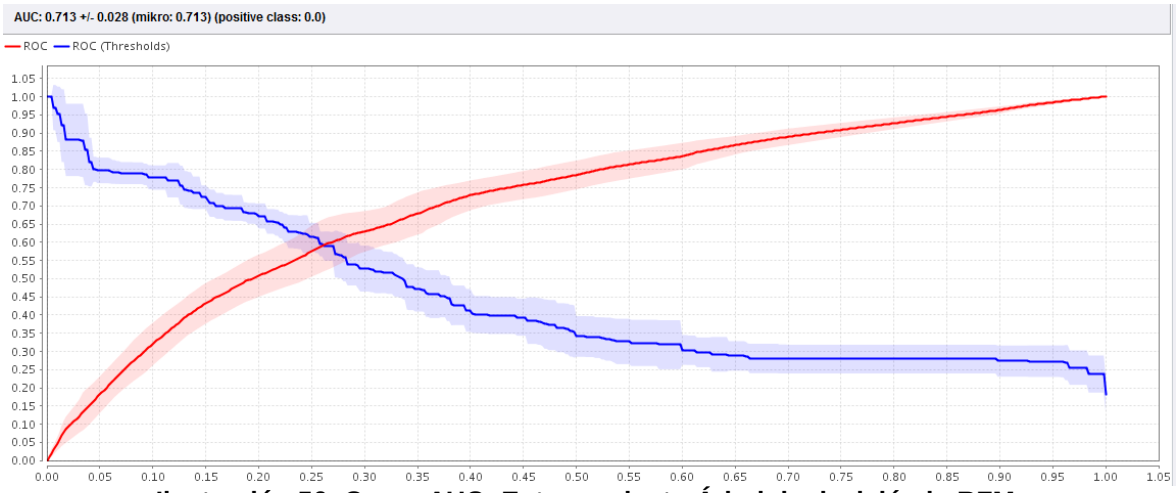

Ilustración 56: Curva AUC- Entrenamiento Árbol de decisión in RFM.
Fuente: Elaboración Propia.