ESCUELA SUPERIOR POLITÉCNICA DEL LITORAL

FACULTAD DE CIENCIAS NATURALES Y MATEMÁTICAS
DEPARTAMENTO DE MATEMÁTICAS

PROYECTO DE GRADO

PREVIO A LA OBTENCIÓN DEL TÍTULO DE:

"MAGÍSTER EN CONTROL DE OPERACIONES Y GESTIÓN LOGÍSTICA"

TEMA

MODELO DE PREDICCIÓN DE DESERCIÓN DE CLIENTES DE TARJETAS DE CRÉDITO

AUTOR

BRENDA DENISSE COBEÑA TERÁN

Guayaquil - Ecuador

AÑO

2016

DEDICATORIA

Dedico este proyecto de tesis a mi familia que me apoyaron en todo momento. Depositando su confianza en cada reto que se me presentaba sin dudar ni un solo momento en mi inteligencia y capacidad.

Brenda Denisse Cobeña Terán

AGRADECIMIENTO

Valoro el tiempo y ayuda que me dieron todas las personas que me apoyaron en todo sentido sea directamente o indirectamente en especial a mi esposo.

Muchas gracias.

Brenda Denisse Cobeña Terán

DECLARACIÓN EXPRESA

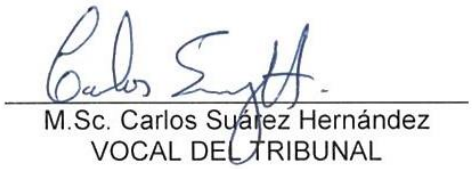
La responsabilidad por los hechos y doctrinas expuestas en este Proyecto de Graduación, me (nos) corresponde(n) exclusivamente; el patrimonio intelectual del mismo, corresponde exclusivamente a la **Facultad de Ciencias Naturales y Matemáticas, Departamento de Matemáticas** de la Escuela Superior Politécnica del Litoral.

Brenda Denisse Cobeña Terán

TRIBUNAL DE GRADUACIÓN

M.Sc. Guillermo Baquerizo Palma
PRESIDENTE DEL TIBUNAL

Ph.D. Andrés Abad Robalino
DIRECTOR DE LA TESIS



M.Sc. Carlos Suárez Hernández
VOCAL DEL TRIBUNAL

AUTOR DEL PROYECTO DE GRADO

_____

BRENDA DENISSE COBENA TERÁN

Tabla de Contenido

Índice de Figuras

Índice de Tablas

CAPÍTULO I

1. INTRODUCCIÓN

En la actualidad se ha evidenciado considerablemente el uso de tarjeta de créditos en la población ecuatoriana, convirtiéndose en una opción para costear el consumo de quienes no cuentan con la liquidez necesaria para sus actividades comerciales. En el Ecuador, en junio del 2013 se registraron 3'151.887 tarjetas de crédito, entre tarjetas principales (85%) y adicionales (15%), según datos de la Superintendencia de Bancos y Seguros (SBS) [1].

El 79% de las tarjetas de créditos en el país son emitidas por cinco entidades financieras: Banco del Pichincha, Banco de Guayaquil, Banco del Pacífico, Banco Bolivariano y Produbanco [1]. La competencia entre las principales instituciones financieras emisoras de este servicio ha requerido que las mismas desarrollen una serie de estrategias destinadas a adquirir nuevos clientes y a mantener sus clientes actuales.

Las instituciones financieras que ofertan tarjetas de crédito obtienen importantes réditos de estas operaciones. El banco en el cual se realizará el estudio pertenece a uno de los cinco bancos más importantes de Ecuador. En este banco, de las utilidades que se genera anualmente, el 23% corresponde a las tarjetas de crédito que emite dicha institución. El banco espera fortalecer la vinculación con sus clientes de tarjetas de crédito actuales, rentabilizando su relación.

Una de los principales problemas de las instituciones financieras es el abandono de los clientes. Así, si las instituciones financieras que mantienen tarjetas de crédito pudieran anticipar cuándo un cliente va a dejar de utilizar el servicio de tarjetas de crédito, podrían enfocar sus esfuerzos en retener a estos clientes.

Por tal motivo es de alto interés detectar el comportamiento y las decisiones de los clientes a futuro, para de esta forma las empresas puedan tomar decisiones, anticipándose a lo que el cliente decida [2]. Esto representaría potencial ingresos importantes para la institución.

Otra de las acciones que realizan las entidades son las campañas de retención. Una de las razones principales por las que ejecutan este tipo de estrategias es que el costo de mantener a un cliente es menor que captar a uno nuevo es cinco veces mayor que tener un nuevo cliente que ofertar un producto/servicio adicional a un cliente existente [3].

1.1 DESCRIPCIÓN DEL PROBLEMA

En el presente estudio nos enfocaremos en la deserción de los clientes de tarjeta de crédito de una de las instituciones financieras más importantes del país. ***El banco ha tenido una deserción de clientes del 11% anualmente*****,** por la que se propone desarrollar un modelo que permita detectar qué clientes son los más propensos a desertar y de esta forma se pueda desarrollar estrategias de retención y fidelización.

Dentro del portafolio de productos del Activo del Banco, lo que corresponde a Tarjeta de Crédito marca Visa tiene una participación del 40% en la *Banca Persona*. El estudio se enfoca en todos los segmentos de clientes que conforman la banca mencionada anteriormente.

1.2 SITUACIÓN ACTUAL

Dada la deserción que tiene actualmente y consciente que, dentro de su portafolio de productos del Activo, el producto Tarjeta de Crédito representa alrededor del 40%, el banco ha realizado diversas campañas de retención, logrando tan solo un 30% de efectividad. Debido a que se ha detectado que

cada segmento del cliente es diferente, por ende, las estrategias de retención deben ir enfocadas a cada segmento.

La entidad bancaria, a diciembre 2014 tiene aproximadamente 137,000 clientes con Tarjetas de Crédito de la marca Visa activas, de los cuales el 80% son titulares y el 20% adicionales.

Por otro lado, el número de tarjetas que se colocan mensualmente mediante campañas en promedio son 6,200 y mediante la red de agencias bancarias en promedio son 2,800 tarjetas.

Actualmente el total de la cartera en este producto es de $192’703,438 y la facturación mensual es de $82’147,400, de los cuales el 66% corresponde a consumos rotativos, 24% consumos diferidos y el resto a avances de efectivo.
La deuda promedio mensual por tarjeta es alrededor de $1,700.

1.3 MOTIVACIÓN

El motivo para realizar el estudio de deserción de clientes son muchos entre los cuales se menciona:

- Mejorar las relaciones con los clientes al detectar las variables que más influyen en su decisión de abandono.
- Alargar la vida media de los clientes, disminuyendo el número de desertores, logrando fidelizar al cliente. Hoy en día en promedio se deja de facturar por cliente aproximadamente $200 al mes, lo que al año representa el 10% del total de la cartera generada en Tarjeta de Crédito.
- Disminuir en un 6% el indicador de deserción de clientes.
- Posterior desarrollo e implementación de políticas comerciales de retención y estrategias a implementar para bajar el índice de deserción aplicada a los diferentes tipos de tarjetas o segmentos.

- La efectividad de las campañas comerciales al ofertar este producto es del 30%, se propone mejorar al 70%, realizando un mejor perfil de clientes.

1.4 OBJETIVOS Y GENERALIDADES

1.4.1 OBJETIVO GENERAL

Proveer de información predictiva a los Directivos del Banco para la toma de decisiones, como un mecanismo en la planificación de las estrategias de retención y fidelización de clientes.

1.4.2 OBJETIVOS ESPECÍFICOS

- Definir un modelo predictivo de deserción que permita saber los posibles clientes desertores.
- Determinar la probabilidad de deserción de los clientes y de esta forma tomar acciones preventivas.

1.4.3 ALCANCES Y LIMITACIONES

La información con la que se trabajará es confidencial y no puede ser proporcionada en claro. Específicamente, la información requerida será proporcionada con todos los datos necesarios pero cifrada.

Además, el alcance de este proyecto estaría dirigido específicamente a los clientes que desertaron voluntariamente de la ***Banca Persona*** en el grupo de clientes que tienen ***Tarjeta de Crédito VISA.***

1.5 DESCRIPCIÓN DEL CONCEPTO DE DESERCIÓN

Buckinx y Van den Poel (2005) definen un desertor parcial como alguien que va disminuyendo la frecuencia de las compras por debajo de la media y la disminución del periodo entre una compra y otra. Van den Poel y Larivière (2004) consideran al cliente que cerró sus cuentas como desertor [4].

Nosotros consideraremos a la deserción como la propensión de los clientes de dejar de hacer negocios con una empresa en un determinado período de tiempo, siguiendo la definición en Neslin et al., 2006 [5].

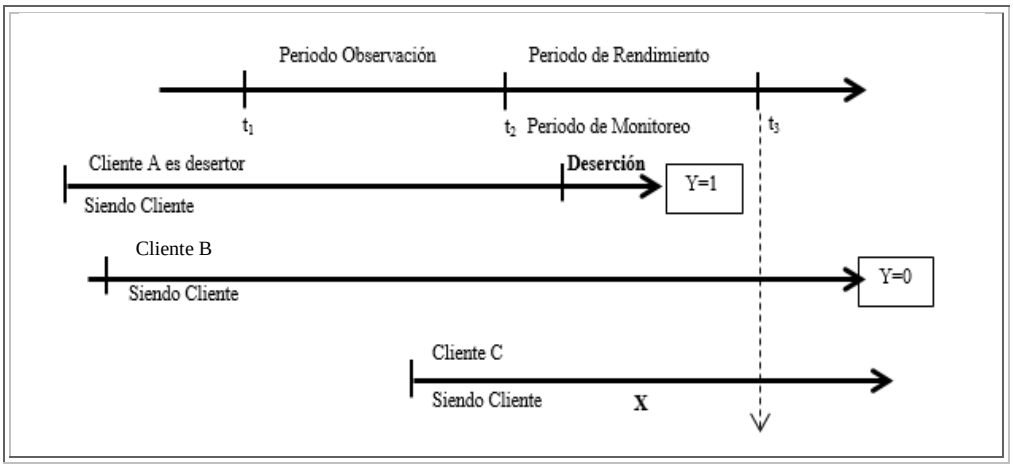
Los modelos de deserción, permiten predecir si un determinado cliente puede convertirse en un cliente desertor, en un determinado tiempo. Hoy en día este tipo de modelos son muy utilizados en entidades bancarias y empresas de prestación de suscripción.

Para definir específicamente el concepto de deserción en el presente estudio, debemos dividir la ventana de tiempo en dos fases: el período de observación y período de evaluación.

En el período de observación se diseñaron variables para analizar los comportamientos de las transacciones de un cliente, y luego se analiza si el cliente se convirtió en un desertor o no durante el período de evaluación. Las variables independientes (x_1, x_2, … x_p) son calculadas a partir la base de datos transaccional, así como las variables asociadas al comportamiento crediticio y variables sociodemográficas, información obtenida durante el periodo de observación y la variable dependiente (Y) es calculada partir de los datos almacenados de la información durante el periodo de evaluación. Con el fin de evitar efectos de meses de temporadas altas, se establece todo un año (12 meses) como el período de observación (el intervalo entre t_1 y t_2) y para el periodo de desempeño (10 meses) como (el intervalo entre t_2 y t_3). Después de observar el comportamiento de los clientes de un año entero, nuestro modelo

predice si el cliente se convierte en un desertor o no en el período de evaluación (Ver Figura 1.5). En nuestro estudio, el periodo de observación es desde enero 2013 a diciembre 2013 y el periodo de evaluación comprende desde enero 2014 a octubre 2014. Después de observar el comportamiento durante un año de los clientes, el modelo predecirá si el cliente es un desertor o no en el periodo de observación.

Figura 1.5: Ventana de Tiempo de Análisis (Adaptada de Nie, 2009)



Por esta razón para estimar la probabilidad de deserción necesitamos un conjunto de variables independientes X_1, X_2,…., X_p que explicarán una variable dependiente binaria Y_i que tomará el valor de 1 si el individuo es catalogado como Desertor y 0 si es catalogado como No desertor, así:

$$Y_i = \begin{cases} 1\ Si\ el\ cliente\ i\ es\ desertor\ durante\ el\ periodo\ de\ evaluación \\ 0\ Si\ el\ cliente\ i\ sigue\ siendo\ cliente\ durante\ el\ periodo\ de\ evaluación \end{cases}$$

1.5.1 MODELO PROPUESTO

En el presente estudio se elaborará un modelo de deserción de tarjeta de créditos para una entidad bancaria privada perteneciente al Sistema Financiero, regulado por la Superintendencia de Bancos y Seguros (SBS), la cual tiene dentro de sus actividades crediticias la emisión de tarjetas de crédito. Los datos necesarios para el desarrollo del modelo para fines académicos serán proporcionados por la entidad bancaria.

La información disponible se encuentra especificada por tarjeta, por cliente, mes, con el objetivo de tener mayor facilidad al momento que la información sea generada.

Los datos históricos son aproximadamente de 2 años, la muestra con la cual desarrollaremos el estudio serán todas las tarjetas que tuvieron movimientos durante el 2013 y 2014, con el fin de abarcar el comportamiento anual y eliminar posibles estacionalidades, ya sean por inicio de clases, pago de utilidades, festividades, etc.

Los modelos más utilizados para el problema de la predicción de clientes en la literatura son:

- Regresión Logística
- Árboles de Clasificación
- Bosques Aleatorios
- Máquinas de Soporte Vectorial
- Redes Neuronales

Específicamente, en este estudio proponemos combinar árboles de decisión mediante el AdaBoost, para predecir la deserción de clientes de tarjetas de crédito. El AdaBoost es uno de los meta-algoritmo de boosting, desarrollado por Freund y Schapire. En el AdaBoost el resultado final es una suma ponderada de los resultados individuales de los algoritmos débiles. Debido a que el proceso de entrenamiento del AdaBoost selecciona solamente aquellos predictores que incrementan el poder de predicción del modelo, se reduce la dimensión y se mejora potencialmente el tiempo de ejecución, evitando la afectación de la dimensionalidad.

El AdaBoost ya ha sido aplicado en diversas industrias, como por ejemplo en telecomunicaciones [8]. Sin embargo, para conocimiento del autor no ha sido aplicado antes al problema de predicción de deserción de clientes de tarjetas de crédito.

CAPÍTULO II

2. DESCRIPCIÓN DE METODOLOGÍA

2.1 DEFINICIÓN DE VARIABLE OBJETIVO

Para nuestro estudio la variable objetivo es la deserción. Donde el interés es conocer si un cliente es desertor o no lo es, lo cual se define deserción, la decisión del cliente de ya no tener relación comercial con la entidad y solicitar de forma voluntaria suspender el uso de la tarjeta una vez cancelado en su totalidad la tarjeta de crédito.

Podemos definir la variable objetivo como:

$$Y_i = \begin{cases} 1\ Si\ el\ cliente\ i\ es\ desertor\ durante\ el\ periodo\ de\ evaluación \\ 0\ Si\ el\ cliente\ i\ sigue\ siendo\ cliente\ durante\ el\ periodo\ de\ evaluación \end{cases}$$

2.2 RECOLECCION DE DATOS HISTÓRICOS

El análisis fue realizado sobre una base de datos que contiene información referente a todos los clientes con una antigüedad mínima de 1 año. El periodo de análisis es de 22 meses (enero 2013 hasta octubre 2014).

En el estudio se considerará una base de 5,308 clientes, de los cuales 1,732 son clientes desertores. El concepto de cliente cancelado es la forma particular que el banco clasifica a los clientes que han desertado de forma voluntaria.

El conjunto de datos a trabajar registra a todos los clientes cuya fecha de cancelación fue en el periodo de enero a octubre de 2014.

Existen tres tipos de conjunto de variables:

- **Sociodemográficas:** Las de información demográfica del cliente son las variables comúnmente más usadas, se incluye información de edad,

género, ubicación geográfica, ingresos mensuales, segmento, canal radicador, antigüedad del cliente, cantidad de productos y estado civil.

- **Externas:** Las variables externas clasificadas así dado que corresponde a la calificación del cliente en la central de riesgo y el monto de deuda que tiene el cliente.
- **Comportamiento:** El comportamiento del cliente identifica el número de transacciones realizadas, cantidad de pagos, consumos realizados sean estos dentro del país o en el exterior, conocer cuán a menudo el cliente utiliza la tarjeta de crédito, es decir el comportamiento con respecto a sus transacciones.

Ver Anexo I la descripción de las variables que se utilizarán para el estudio.

2.3 ANÁLISIS EXPLORATORIO

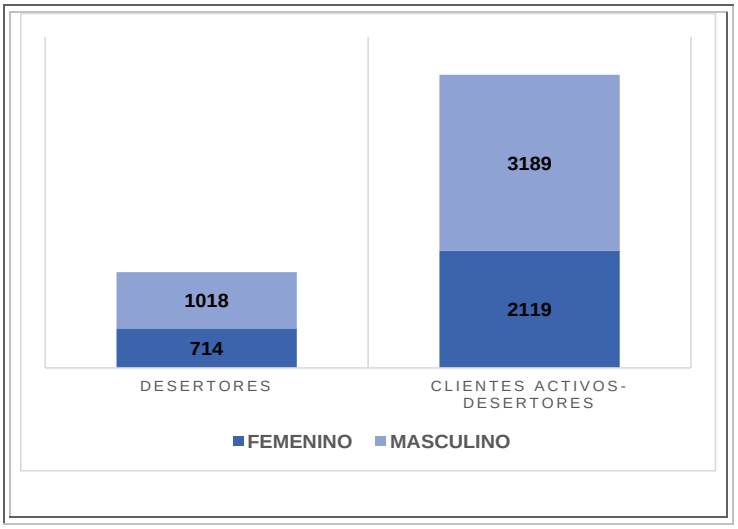
Se presentará el análisis estadístico descriptivo de las variables a considerarse en el estudio. En el caso de las variables cuantitativas, se estiman parámetros tales como media, curtosis, sesgo, varianza, desviación estándar, error estándar, cuartiles y moda.

2.3.1 CARACTERÍSTICAS DEL CLIENTE

En esta sección se analizan estadísticamente características demográficas del cliente que desertó.

Género. - De los 5,308 clientes que conforman la muestra, el 60% son hombres y 40% son del género femenino. Mientras que los clientes que han desertado (1,732), el 59% son del género masculino y el 41% son mujeres (Ver Figura 2.3.1. A).

Figura 2.3.1. A: Histograma de frecuencias *Género*



Se puede conocer que en base a lo analizado anteriormente el total de clientes del género masculino que obtuvieron la tarjeta de crédito Visa, el 19.2% desertó voluntariamente, mientras que de las mujeres el 13.5% desertaron voluntariamente del banco (Ver Tabla 2.3.1 A).

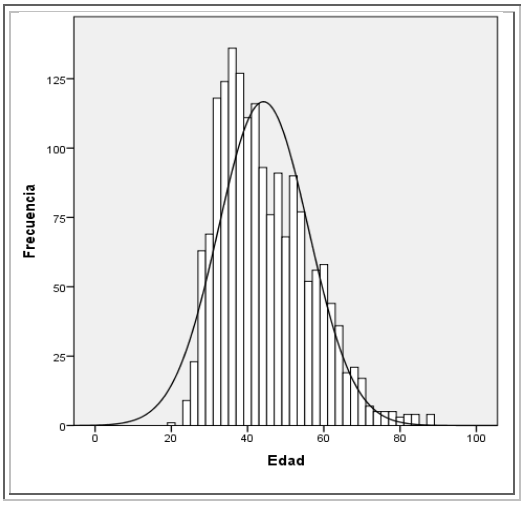
Tabla 2.3.1 A Distribución de Frecuencias de Género por Estado de Cliente

Género	Frecuencia Absoluta			Frecuencia Relativa (%)		
	Desertor	Activo	Total	Desertor	Activo	Total
Femenino	714	1,405	2,119	13.5%	26.5%	39.9%
Masculino	1,018	2,171	3,189	19.2%	40.9%	60.1%
Total	**1,732**	**3,576**	**5,308**	**32.6%**	**67.4%**	**100.0%**

***Edad*. -** de los clientes analizados al momento de desertar el 23.5% tenía edades entre 20 y 34 años; el 45.03 % declaró tener entre 35 y 49 años y el 31.47% tenía edades mayores o iguales a 50 años. La edad promedio estimada es 44.16 $\pm$ 0.28 años. El estimador de la desviación estándar es 11.84 y del error estándar 0.28 (Ver Tabla 2.3.1 B y Figura 2.3.1. B).

Tabla 2.3.1 B Estimadores Poblacionales de la Edad de los Clientes Desertores	
Media	44.16
Error Estándar	0.28
Moda	35
Desviación estándar	11.84
Varianza	140.12
Curtosis	0.16
Coeficiente de asimetría	0.69
Rango	68
Mínimo	20
Máximo	88
n	1,732
Nivel de confianza (95.0%)	0.56
Percentil 25 Percentil 50 (Mediana) Percentil 75	35 42 88

Figura 2.3.1. B: Histograma de Frecuencias *Edad*



Mediante la Prueba de Kolmogorov-Smirnov (K-S), se prueba si la función de densidad de la variable edad puede ser modelada como una normal con media μ=44.16 y varianza σ^2 =140.12 (Ver Tabla 2.3.1 C).

Tabla 2.3.1 C Prueba de bondad de Ajuste, Kolmogorov-Smirnov (K-S): Edad
H_o: La edad de los desertores tiene una distribución N (44.16, 11.84) Vs. **H_1:** No es verdad Ho Valor p= 2.2e-16

Se puede observar que en los resultados obtenidos mediante la prueba de bondad de ajuste de K-S, el valor p es 2.2e-16, lo cual significa que a un nivel de significancia α=0.05, no existe evidencia estadística para concluir que los datos observados provienen de una distribución Normal.

***Ubicación geográfica*. -** de los clientes que han desertado el 39.7% son de la ciudad de Quito, el 32.6% son de la ciudad de Guayaquil, ambas ciudades son

donde se concentran el mayor número de desertores. Mientras que el 5.6% de los desertores son de Cuenca, además el 3.3% son de Ambato, en Machala fueron el 2.7%, Portoviejo el 2%, Santo Domingo el 1.7%, el resto del país representó el 12.4% (Ver Tabla 2.3.1 D).

Tabla 2.3.1 D Distribución de Frecuencias de Ubicación geográfica

Ubicación geográfica	Frecuencia Relativa
Quito	39.7%
Guayaquil	32.6%
Cuenca	5.6%
Ambato	3.3%
Machala	2.7%
Portoviejo	2.0%
Santo Domingo	1.7%
Resto del país	12.4%
Total	**100.0%**

***Antigüedad*. -** El 18.8% de los clientes que han desertado tuvieron una antigüedad menor a 23 meses es decir menos de 2 años, el 32.97% de los clientes que han abandonado voluntariamente fueron clientes en un rango de 24 a 28 meses, el 27.87% una antigüedad de 49 a 72 meses, en estos 2 rangos es donde se concentran la mayor cantidad de desertores. Estas cifras se observan en la Tabla 2.3.1 E.

Tabla 2.3.1 E Distribución de Frecuencias de la Antigüedad

Antigüedad	Frecuencia Relativa (%)
Menor a 23 meses	18.82%
24 a 48 meses	32.97%
49 a 72 meses	27.89%
73 a 96 meses	7.27%
97 a 120 meses	3.70%
121 a 144 meses	7.85%
Mayor a 144 meses	1.50%
Total	**100.00%**

***Ingreso Mensual*. -** El 35.80% de los clientes que desertaron tienen Ingresos Mensuales de $500 a $1,499, el 28.41% tienen ingresos entre $1,500 a $2,999 y el 13.80% tienen ingresos superiores o igual a $5,000 (Ver Tabla 2.3.1 F).

Tabla 2.3.1 F Distribución de Frecuencias de Ingresos Mensuales

Rangos Ingresos Mensuales	Frecuencia Absoluta	Frecuencia Relativa (%)
Ingresos menores a $500	158	9.12%
Ingresos de $500 a $1,499	620	35.80%
Ingresos de $1,500 a $2,999	492	28.41%
Ingresos de $3,000 a $4,999	223	12.88%
Ingresos superiores o igual a $5,000	239	13.80%
Total	**1,732**	**100.00%**

***Segmento*. -** de los clientes que han desertado el 10.80% son clientes del segmento básico el 38.28% del segmento estándar, 29.04% del segmento medio, mientras que el 10.74% son del segmento medio alto y el 11.14% son del segmento alto (Ver Tabla 2.3.1 G).

Tabla 2.3.1 G Distribución de Frecuencias de Segmento

Segmento	Frecuencia Absoluta	Frecuencia Relativa (%)
Básico	187	10.80%
Estándar	663	38.28%
Medio	503	29.04%
Medio alto	186	10.74%
Alto	193	11.14%
Total	**1,732**	**100.00%**

Estado Civil vs. Actividad Económica. -

El análisis conjunto de las variables “Estado Civil” y “Actividad Económica”, se determina que el 18.71% de los clientes que son Empleados Privado son solteros. Dado los clientes que son Empleados Privados el 33.95% son Casados. Por otro lado, se puede observar que de los Empleados Públicos el 8.72% son Casados (Ver Tabla 2.3.1 H).

Tabla 2.3.1 H Estado Civil vs. Actividad Económica							
Estado Civil	***Actividad Económica***						***Marginal Estado Civil***
	Empleado privado	**Empleado público**	**Actividades profesionales**	**Jubilado**	**Comercio mayor y menor**	**Otros**	
Soltero	18.71%	5.08%	1.50%	0.81%	1.44%	2.77%	**30.31%**
Casado	33.95%	8.72%	3.35%	1.91%	4.16%	8.66%	**60.74%**
Unión libre	0.58%	0.23%	0.00%	0.23%	0.17%	0.29%	**1.50%**
Divorciado	2.54%	1.21%	0.40%	0.29%	0.58%	1.10%	**6.12%**
Viudo	0.40%	0.12%	0.06%	0.17%	0.06%	0.52%	**1.33%**
Marginal Actividad Económica	**56.18%**	**15.36%**	**5.31%**	**3.41%**	**6.41%**	**13.34%**	**100.00%**

***Canal Radicador*. -** de los clientes que han desertado el 46.07% obtuvieron la tarjeta de crédito mediante canal natural y el 53.93% mediante preventa es decir de campañas realizadas donde el Banco otorga Tarjetas de Crédito a clientes (Ver Tabla 2.3.1 I).

Tabla 2.3.1 I Distribución de Frecuencias de Canal Radicador		
Segmento	**Frecuencia Absoluta**	**Frecuencia Relativa (%)**
Mercado natural	798	46.07%
Preventa	934	53.93%
Total	**1,732**	**100.00%**

2.3.2 PRODUCTOS ADICIONALES (VINCULACIÓN)

En esta sección se analizan si los clientes que han desertado tienen otro tipo de productos adicionales a la tarjeta de crédito vinculados con el Banco.

Productos del Pasivo

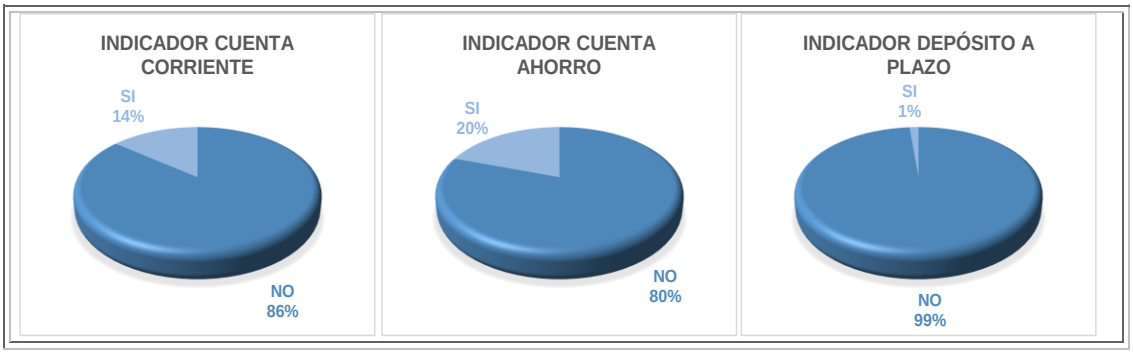
Se refiere a productos de cuenta corriente, ahorro y pólizas de acumulación (depósitos a plazo). A continuación, se analizará si los clientes que desertaron tienen otros productos con la entidad bancaria.

Indicador Cuenta Corriente. - el 14.20% de los clientes que desertaron, tienen una cuenta corriente con el Banco, mientras que del 85.80% que desertaron no tienen ningún tipo de cuenta corriente.

Indicador Cuenta Ahorro. - el 19.57% de los clientes que han desertado tienen una cuenta de ahorro con el Banco, mientras que del 80.43% no tienen ningún tipo de cuenta de ahorro.

Indicador Depósito a Plazo. - de los clientes que han desertado solo el 1.5% tienen una cuenta depósito a plazo (Ver Figura 2.3.2. A).

Figura 2.3.2. A Indicador de Productos



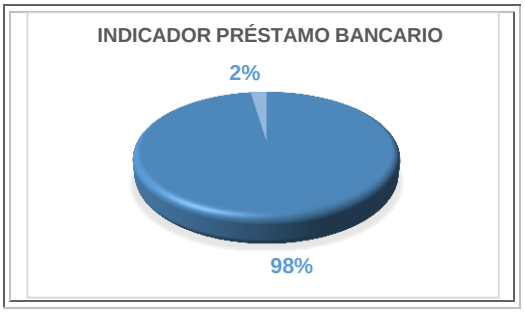
Productos del Activo

Se refiere a productos tales como préstamos bancarios que el cliente tiene vigente en la entidad bancaria, estos préstamos incluyen: créditos de consumo, vehículo, educativo, micro empresarial o de vivienda.

A continuación, se analizará si los clientes que desertaron tienen algún préstamo con el Banco.

Préstamo. - de los clientes que han desertado solo el 2% tienen préstamo con el Banco (Ver Figura 2.3.2. B).

Figura 2.3.2. B Histograma de frecuencias
Ingreso Mensual



2.3.3 COMPORTAMIENTO EXTERNO

En esta sección se analiza la calificación y monto de deuda que los clientes tienen en total en el sistema financiero registrado en el país (central de riesgo).

Monto de Riesgo: en la Tabla 2.3.3 A se muestra en promedio el monto de riesgo en dólares que los clientes que han desertado tienen en central de riesgo.

Tabla 2.3.3 A Promedio de Monto de deuda en Central de Riesgo por Segmento	
Segmento	**Promedio (US$)**
Básico	17,488.39
Estándar	12,013.91
Medio	20,970.37
Medio Alto	33,304.07
Alto	64,132.99
Promedio(US$)	**23,122.33**

En la tabla anterior se observa que los clientes del segmento básico tienen en promedio un monto de deuda de $17,488, los del segmento estándar de $12,014, mientras que los de segmento medio y medio alto el promedio de deuda es de $20,970 y $33,304 respectivamente. Los del segmento alto registran en promedio un monto de deuda superior a $60,000.

Calificación Central de Riesgo: el 93.82% de los clientes que desertaron tuvieron una excelente calificación en la central de riesgo es decir “A”, mientras que el 3,87% registraron una calificación de B, el 1.27% de los clientes tuvieron la peor calificación “E” (Ver Tabla 2.3.3 B).

Tabla 2.3.3 B Distribución de Frecuencias de Calificación Central Riesgo		
Calificación	**Frecuencia Absoluta**	**Frecuencia Relativa**
A	1,625	93.82%
B	67	3.87%
C	6	0.35%
D	12	0.69%
E	22	1.27%
Total	**1,732**	**100.00%**

2.3.4 COMPORTAMIENTO TRANSACCIONAL

En esta sección se analiza las variables de tipo transaccional, el periodo de análisis de comportamiento de enero a diciembre de 2013.

***Montos de Consumo*. -** El 9.35% de los desertores nunca realizaron consumos con la tarjeta de crédito, por otro lado, se puede apreciar que el 15.07% tuvieron en promedio consumos mensuales entre $0 a $100, mientras que el 13.68% realizaron consumos de $100 a $300, el 12.41% hicieron transacciones mensuales con montos en promedio de $300 a $500. Además, existe un grupo considerable de desertores en los cuales sus consumos mensuales son superiores a $1,100 lo que representa el 23.33%, del total de desertores que se analizó. Estas cifras se observan en la siguiente Tabla 2.3.4 A.

Tabla 2.3.4 A Distribución de Frecuencias de Montos de Consumo

Rangos Monto Consumos	Frecuencia Absoluta	Frecuencia Relativa
Igual a 0	162	9.35%
(0 a $100]	261	15.07%
($100 a $300]	237	13.68%
($300 a $500]	215	12.41%
($500 a $700]	195	11.26%
($700 a $900]	146	8.43%
($900 a $1,100]	112	6.47%
Mayor a $1,100	404	23.33%
Total	**1,732**	**100.00%**

***Cantidad Transacciones Facturadas*. -** el 24.42% de los clientes que desertan realizan a lo más una transacción por mes en algún tipo de establecimiento, lo que indica que la tarjeta de crédito no es de uso frecuente en las transacciones que realizan. Mientras que el 17.21% de los clientes desertores registran transacciones de 15 a 25 en total al año, es decir en promedio han realizado 2 transacciones por mes (Ver Tabla 2.3.4 B).

Tabla 2.3.4 B Distribución de Frecuencias de Cantidad de Transacciones Facturadas al Año		
Rangos Cantidad Transacciones Facturadas	**Frecuencia Absoluta**	**Frecuencia Relativa**
Igual a 0	288	16.63%
(0 a 5]	212	12.24%
(5 a 15]	423	24.42%
(15 a 25]	298	17.21%
(25 a 35]	186	10.74%
(35 a 45]	91	5.25%
(45 a 55]	66	3.81%
(55 a 65]	48	2.77%
(65 a 75]	33	1.91%
Mayor a 75	87	5.02%
Total	**1,732**	**100.00%**

***Monto de Sobregiro.* -** El 75.92% de los clientes que desertaron nunca estuvieron sobregirados, sin embargo, existe un 8.49% que tuvieron al mes en promedio sobregiros entre 0 a $10. (Ver Tabla 2.3.4 C).

Tabla 2.3.4 C Distribución de Frecuencias de Monto de Sobregiros		
Rangos Monto de Sobregiro	**Frecuencia Absoluta**	**Frecuencia Relativa**
Igual a 0	1,315	75.92%
(0 a 10]	147	8.49%
(10 a 20]	87	5.02%
(20 a 30]	53	3.06%
(30 a 40]	19	1.10%
(40 a 50]	15	0.87%
Mayor a 50	96	5.54%
Total	**1,732**	**100.00%**

***Números de Meses Sobregirado.* -** como se analizó en la variable anterior se pudo observar que el 75.92% de los clientes que desertaron nunca estuvieron sobregirados, sin embargo existe un 24.08% que si lo estuvieron, de ese grupo, el 7.68% estuvo sobregirado en el año 1 vez, el 4.10% estuvieron 2 veces sobregirados y el 12% estuvieron sobregirados más de 3 veces en el año (Ver Tabla 2.3.4 D).

Tabla 2.3.4 D Distribución de Frecuencias de Números de Meses Sobregirado - Cliente Desertor		
Cantidad de Meses con Transacción	**Frecuencia Absoluta**	**Frecuencia Relativa**
0	1,315	75.92%
1	133	7.68%
2	71	4.10%
3	46	2.66%
4	30	1.73%
5	31	1.79%
6	27	1.56%
7	18	1.04%
8	14	0.81%
9	7	0.40%
10	17	0.98%
11	8	0.46%
12	15	0.87%
Total	**1,732**	**100.00%**

Cantidad Transacciones Masivas vs. Indicador Transacciones Masivas. -

Se puede observar que el 6.24% de los clientes que desertaron solo utilizaron la tarjeta de crédito para transacciones masivas es decir solo para débitos programados ya sea para pago de Planes de Internet, TV Pagadas, Servicios Básicos, etc. De este total existe un 2.42% de clientes desertores que en el año tienen más de 1 transacción de tipo masivo que registró la tarjeta de crédito que utilizaron.

Tabla 2.3.4 E Transacciones Masivas vs. Indicador Solo Transacciones Masivas			
Rango Transacciones Masivas	**Indicador Transacciones Masivas**		**Marginal Transacciones Masivas**
	No	**Si**	
Igual a 0	63.74%	0.17%	**63.91%**
(0 a 5]	10.85%	2.02%	**12.88%**
(5 a 10]	5.83%	1.04%	**6.87%**
(10 a 15]	9.87%	2.42%	**12.30%**
Mayor a 15	3.46%	0.58%	**4.04%**
Marginal Indicador Transacciones Masivas	**93.76%**	**6.24%**	**100.00%**

CAPÍTULO III

3. ANÁLISIS DE SUPERVIVENCIA

En esta sección se presenta la metodología de análisis de supervivencia aplicado al problema de deserción de los clientes en Tarjeta de Créditos de una importante institución bancaria. Se utilizará la metodología de Kaplan-Meier y el modelo de regresión de Cox para identificar las variables que nos permitan detectar los mayores riesgos de abandono. Así mismo identificaremos si el riesgo de un cliente con unas características dadas es mayor que el de otro cliente con otras características.

3.1 ANÁLISIS DE SUPERVIVENCIA CON ESTIMADOR DE KAPLAN - MEIER

El estimador de Kaplan – Meier fue introducido por Edward L. Kaplan y Paul Meier en 1958[5], es un método no paramétrico ya que no asume que la distribución de probabilidad del tiempo pertenece a una familia de funciones paramétrica.

El estimador utiliza casos censurados y no censurados, para realizar la estimación de la función de supervivencia. El estimador en cualquier instante del tiempo es obtenido de la multiplicación de una secuencia de probabilidades condicionales de supervivencia estimadas. Cada probabilidad condicional estimada se obtiene del número de casos observados en riesgo y el número de "muertes" en un instante de tiempo.

Supóngase una muestra de N observaciones independientes y sean $t_1 \leq t_2 \leq t_3 \leq \ldots \leq t_N$ los tiempos de vida observados y t_i el tiempo de vida de la observación de mayor duración en la muestra. Se define entonces [6]:

- n_i= Número de sujetos en riesgo antes del instante t_i. De no haber censura, n_i es el número de supervivientes inmediatamente antes del momento t_i. Con censura es el número de supervivientes menos el

número de casos censurados: sólo se observan los sujetos vivos que no se han caído del estudio en el momento en que ocurre una muerte.

- d_i= Número de muertes en el instante t_i

El estimador de la función de supervivencia de Kaplan-Meier $\hat{S}(t)$ se calcula de la siguiente manera [5]:

$$\hat{S}(t) = \prod_{t_i<t} \frac{n_i - d_i}{n_i}$$

En esta parte se presentarán los resultados obtenidos de la aplicación de la metodología de Kaplan-Meier para clientes de Tarjeta de Crédito Visa. Se asumirá el tiempo de supervivencia de un cliente como el tiempo durante el cual permanece en la entidad bancaria, la mortalidad corresponderá a la deserción.

Se pretende determinar cuáles de las variables analizadas en este estudio están asociadas a la supervivencia de los clientes.

A continuación, se detalla el análisis de supervivencia, los cuales fueron realizados mediante el software R.

Por Género

En la Tabla 3.1 A muestra las diferentes estadísticas y curvas de supervivencia por género y sus comparaciones.

Tabla 3.1 A Estimador Kaplan Meier – Género

Cod	Género	N	Observed	Expected	(O-E)^2/E	(O-E)^2/V
0	Masculino	3217	1018	1009	0.0821	0.2040
1	Femenino	2091	714	723	0.1146	0.2040

De la anterior, *N* se refiere a la cantidad de clientes analizados perteneciente a la institución financiera, en total 3,217 son hombres, *Observed* corresponde a la cantidad de clientes desertores durante el periodo de análisis, para el caso de los hombres, se tiene que han desertado 1,018 y en las mujeres la deserción ha sido 2,091.

Tabla 3.1 B
Estimador Kaplan Meier Tiempo de Supervivencia – Género

Cod	Género	records	n.max	n.start	events	median	0.95LCL	0.95LCL
0	Masculino	3217	3217	3217	1018	56	53	59
1	Femenino	2091	2091	2091	714	58	55	61

La Tabla 3.1 B se refiere al tiempo de supervivencia por género de los clientes analizados en el estudio, así para los que son del género masculino la *mediana* de permanencia como clientes en la institución financiera es de 56 meses; *Intervalos* de confianza se refiere a los límites entre los cuales se espera este el valor medio del tiempo de supervivencia de los clientes, para los del género femenino está entre 55 y 61 meses y los del género masculino entre 53 y 59 meses; *Confianza* se refiere a la probabilidad con la cual se espera que el intervalo de confianza contenga a la media de supervivencia, para el estudio se utilizó una confianza del 95%.

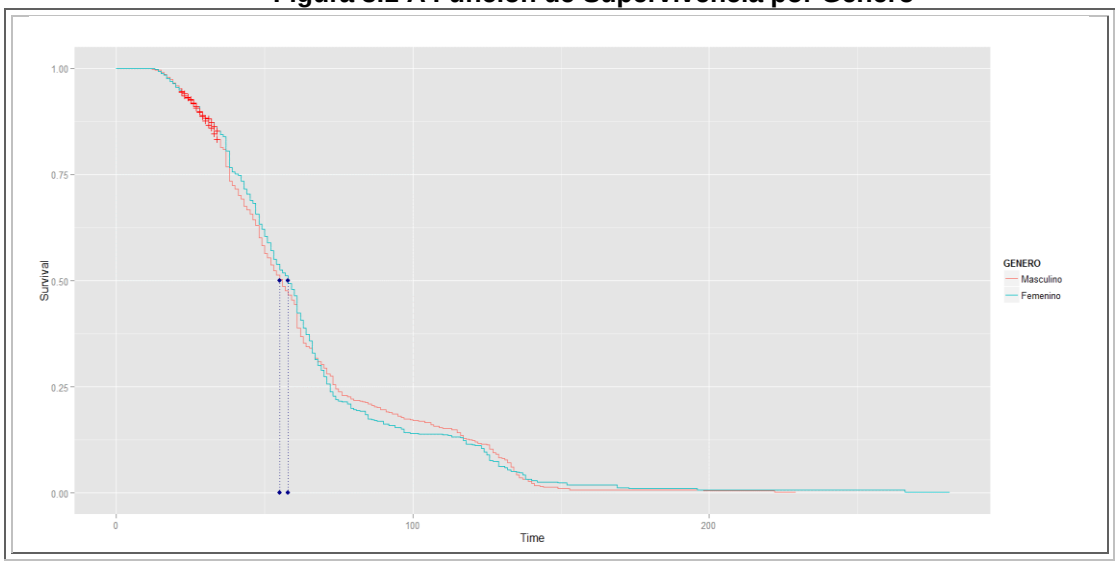
Tabla 3.1 C Estadístico de Prueba Estimador Kaplan

Chi-cuadrado	gl	Sig.
0.2	1	6.51E-01

Se puede observar que en los resultados obtenidos mediante la prueba de Chi - cuadrado, el valor p=6.51E-01, lo cual significa que a un nivel de significancia α=0.05, no existe evidencia estadística entre las funciones de supervivencia.

En la Figura 3.1 A se muestran las diferentes curvas de supervivencia por género de los clientes.

Figura 3.1 A Función de Supervivencia por Género



Por Estado Civil

En la Tabla 3.1 D se presentan los resultados, donde muestra las diferentes estadísticas y curvas de supervivencia por estado civil y sus comparaciones.

Tabla 3.1 D Estimador Kaplan Meier – Estado Civil						
Cod	**Estado**	**N**	**Observed**	**Expected**	**(O-E)^2/E**	**(O-E)^2/V**
1	Soltero	1,687	523	513.7	0.1694	0.2500
2	Casado	3,101	1,044	1,040.5	0.0119	0.0311
3	Unión Libre	84	26	22.8	0.4389	0.4567
4	Divorciado	333	107	119.6	1.3321	1.4770
5	Viudo	66	25	28.4	0.4056	0.4262

De la anterior, el *N* se refiere a la cantidad de clientes analizados perteneciente a la institución financiera, en total 1,687 son solteros, *Observed* corresponde a la cantidad de clientes desertores durante el periodo de análisis, para el caso de los solteros, se tiene que han desertado 523, así mismo se observa que del total de clientes de estado civil casado de los 3,101 que estuvieron en el periodo de análisis, 1,044 han desertado voluntariamente.

Tabla 3.1 E Estimador Kaplan Meier Tiempo de Supervivencia – Estado Civil								
Cod	**Estado**	**records**	**n.max**	**n.start**	**events**	**median**	**0.95 LCL**	**0.95 LCL**
1	Soltero	1,687	1,687	1,687	523	57	53	60
2	Casado	3,101	3,101	3,101	1,044	56	53	59
3	Unión Libre	84	84	84	26	52	39	73
4	Divorciado	333	333	333	107	58	53	67
5	Viudo	66	66	66	25	61	45	73

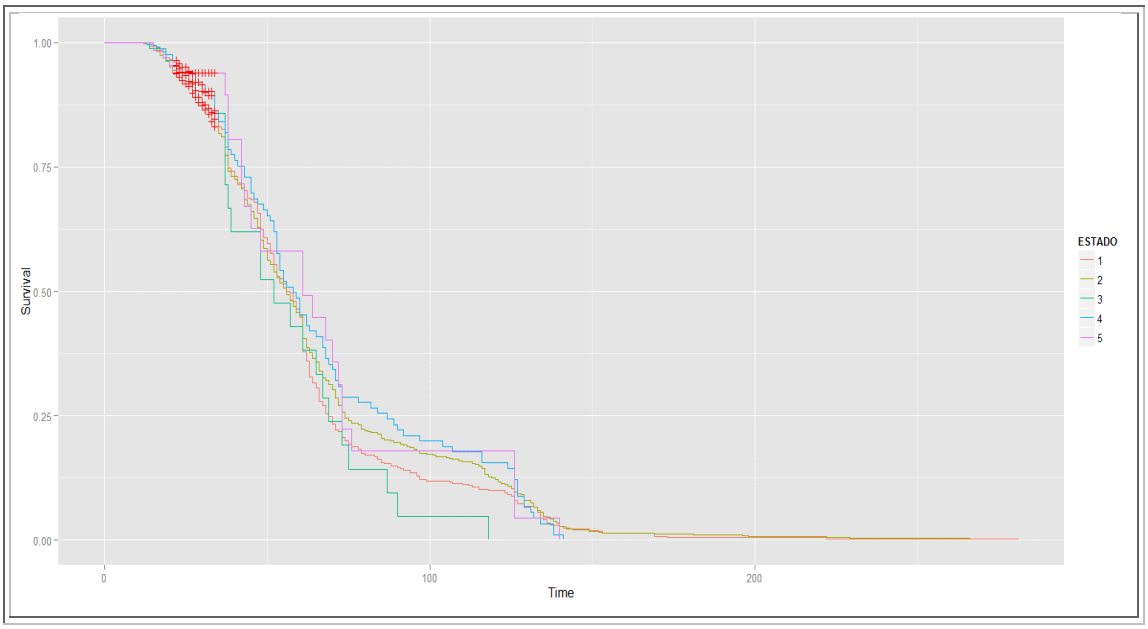
En la Tabla 3.1 E se refiere al tiempo de supervivencia por estado civil de los clientes analizados en el estudio, así para los casados la *mediana* de permanencia como clientes en la institución financiera es de 56 meses; *Intervalos* de confianza se refiere a los límites entre los cuales se espera este el valor medio del tiempo de supervivencia de los clientes, para los casados está entre 53 y 59 meses; *Confianza* se refiere a la probabilidad con la cual se espera que el intervalo de confianza contenga a la media de supervivencia, para el estudio se utilizó una confianza del 95%.

Tabla 3.1 F Estadístico de Prueba – Estimador Kaplan Meier		
Chi-cuadrado	gl	Sig.
2.4	4	0.656

Se puede observar que en los resultados obtenidos mediante la prueba de Chi - cuadrado, el valor p=6.56E-01, lo cual significa que a un nivel de significancia α=0.05, no existe evidencia estadística entre las funciones de supervivencia.

En la Figura 3.1 B se muestran las diferentes curvas de supervivencia por estado civil de los clientes.

Figura 3.1 B Función de Supervivencia por Estado Civil



Por Rangos de Edad

En la siguiente Tabla 3.1 G se muestra, en el campo *Observed* corresponde a la cantidad de clientes desertores durante el periodo de análisis, los que tienen edades comprendidas entre 26 a 30 años han desertado 145, así mismo se observa que del total de clientes con edades entre 31 a 35 años de los 985 que estuvieron en el periodo de análisis, 1,320 han desertado voluntariamente.

Tabla 3.1 G Estimador Kaplan Meier – Rangos de Edad						
Cod	**Rangos de Edad**	**N**	**Observed**	**Expected**	**(O-E)^2/E**	**(O-E)^2/V**
1	Menor a 25	110	20	16.2	0.8909	0.9185
2	26 a 30	620	145	124.7	3.3074	3.7194
3	31 a 35	985	320	232.2	33.2366	40.3117
4	36 a 40	900	296	298.8	0.0256	0.0322
5	41 a 45	768	251	218.4	4.8724	5.7710
6	46 a 50	623	193	182.1	0.6490	0.7478
7	51 a 55	511	190	195.5	0.1536	0.1816
8	56 a 60	398	143	159.4	1.6783	1.9322
9	61 a 65	231	89	126.3	11.0038	12.4626
10	Mayor a 65	133	85	178.6	49.0249	61.4236

En la Tabla 3.1 H se refiere al tiempo de supervivencia por rangos de edad clientes analizados en el estudio, así para los que tienen Menor a 25 años la *mediana* de permanencia como clientes en la institución financiera es de 47 meses; los que tienen entre 26 a 30 años la mediana de permanencia es de 53 meses. *Intervalos* de confianza se refiere a los límites entre los cuales se espera este el valor medio del tiempo de supervivencia de los clientes, para los que tienen edades entre 31 a 35 años el tiempo de permanencia es entre 48 y 55 meses.

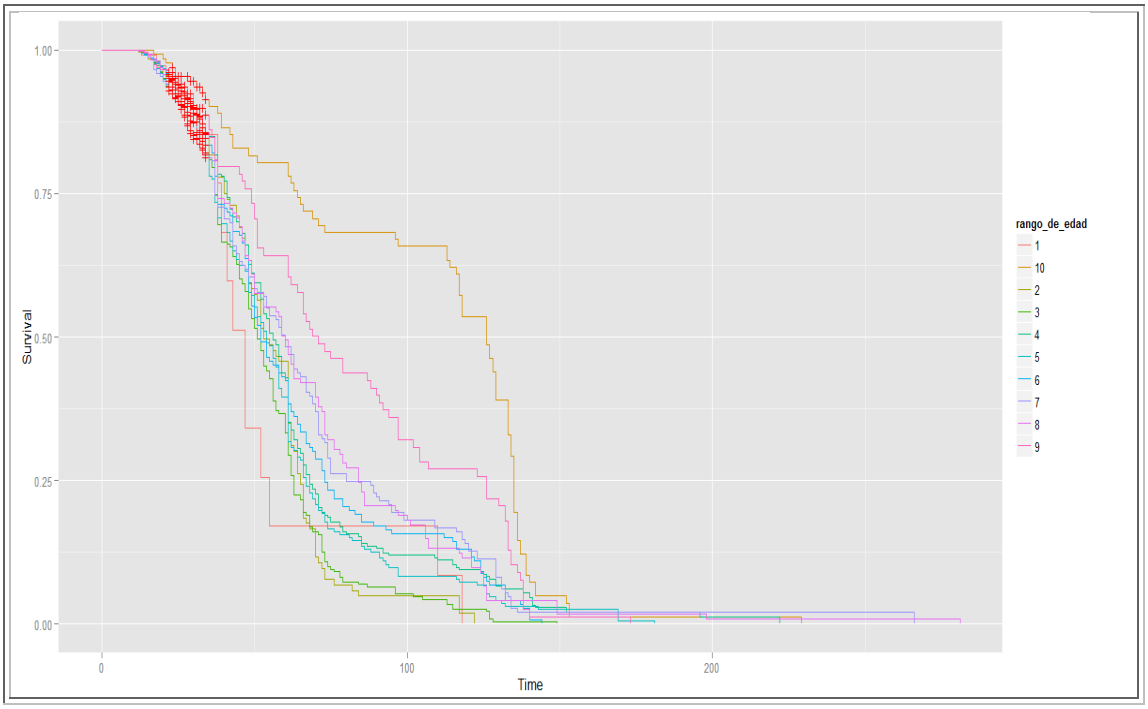
Tabla 3.1 H Estimador Kaplan Meier Tiempo de Supervivencia – Rangos de Edad						
Cod	**Rangos de Edad**	**n**	**events**	**median**	**0.95LCL**	**0.95LCL**
1	Menor a 25	110	20	47	39	NA
2	26 a 30	620	145	53	50	61
3	31 a 35	985	320	51	48	55
4	36 a 40	900	296	56	53	60
5	41 a 45	768	251	54	50	58
6	46 a 50	623	193	52	49	61
7	51 a 55	511	190	60	54	67
8	56 a 60	398	143	60	50	70
9	61 a 65	231	89	71	62	92
10	Mayor a 65	133	85	126	117	133

Tabla 3.1 I Estadístico de Prueba – Estimador Kaplan Meier		
Chi-cuadrado	gl	Sig.
122	9	0

Se puede observar que en los resultados obtenidos mediante la prueba de Chi - cuadrado, el valor p=0, lo cual significa que a un nivel de significancia α=0.05, existe evidencia estadística entre las funciones de supervivencia.

En la Figura 3.1 C se muestran las diferentes curvas de supervivencia por Rango de Edad. Se puede apreciar que, a mayor edad, mayor tiempo de permanencia tendrá el cliente con la entidad bancaria.

Figura 3.1 C Función de Supervivencia por Rango de Edad



Por Segmento

A continuación, se muestran las diferentes estadísticas y curvas de supervivencia por segmento.

Tabla 3.1 J Estimador Kaplan Meier – Segmento						
Cod	**Segmento**	**N**	**Observed**	**Expected**	**(O-E)^2/E**	**(O-E)^2/V**
1	Básico	587	187	289	35.77	46.59
2	Estándar	2,323	663	567	16.34	25.93
3	Medio	1,425	503	468	2.69	3.83
4	Medio Alto	509	186	169	1.67	1.91
5	Alto	464	193	240	9.16	11.24

En la Tabla 3.1 J, se tiene que han desertado 187 clientes en el segmento básico; se puede observar que en el segmento estándar es donde se dan el mayor número de casos censurados es decir desertores, donde existen 663, otro segmento donde se observan un número considerable de clientes desertores es en el segmento medio que en total fue 503, durante el periodo de análisis.

Tabla 3.1 K Estimador Kaplan Meier Tiempo de Supervivencia – Segmento								
Cod	**Segmento**	**records**	**n.max**	**n.start**	**events**	**median**	**0.95LCL**	**0.95LCL**
1	Básico	587	587	587	187	68	64	73
2	Estándar	2323	2323	2323	663	53	51	55
3	Medio	1425	1425	1425	503	56	52	61
4	Medio Alto	509	509	509	186	56	50	61
5	Alto	464	464	464	193	61	59	64

En la Tabla 3.1 K se puede observar que los clientes del segmento estándar tienen en promedio menor tiempo de supervivencia, con una mediana de 53 meses y con intervalos de confianza al 95% entre 51 a 55 meses.

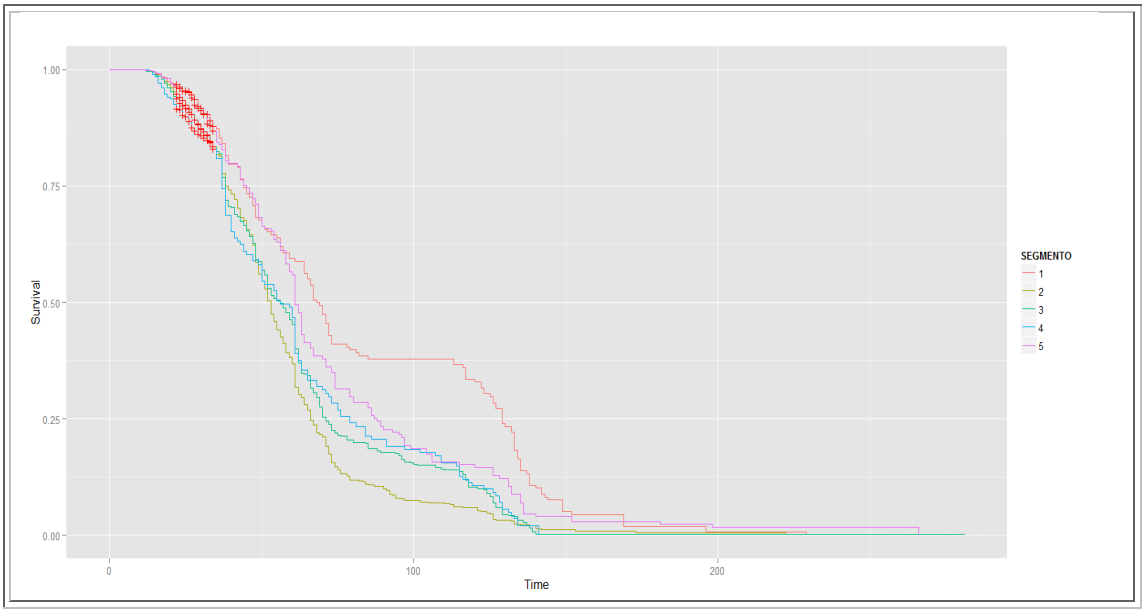
Tabla 3.1 L Estadístico de Prueba – Estimador Kaplan Meier		
Chi-cuadrado	gl	Sig.
72.5	4	6.77E-15

Se puede observar que en los resultados obtenidos mediante la prueba de Chi - cuadrado, el valor p=6.77E-15, lo cual significa que a un nivel de significancia α=0.05, existe evidencia estadística entre las curvas de supervivencia.

En el Figura 3.1 D se muestran las diferentes curvas de supervivencia por segmento. Dentro de los segmentos que son 5 categorías, se nota clara superioridad en término de supervivencia de los clientes de los segmentos:

básico, medio, medio alto y alto. El segmento estándar tiene el peor nivel de supervivencia.

Figura 3.1 D Función de Supervivencia por Segmento



Por Canal Radicador

A continuación, se presentan los resultados obtenidos y las curvas de supervivencia por canal radicador y sus comparaciones, es decir de los clientes que son parte del estudio el medio por el cual obtuvieron la tarjeta de crédito sea por canal preventa o canal natural.

Tabla 3.1 LL Estimador Kaplan Meier – Canal Radicador						
Cod	**Canal**	**N**	**Observed**	**Expected**	**(O-E)^2/E**	**(O-E)^2/V**
0	Natural	1,492	798	1138	102	423
1	Preventa	3,816	934	594	195	423

Se puede apreciar en la Tabla 3.1 LL que los clientes que adquirieron la tarjeta de crédito por el canal preventa son los que tienen la mayor cantidad de desertores, siendo en total 934, mientras que por el canal natural fueron 798 desertores.

Tabla 3.1 M Estimador Kaplan Meier Tiempo de Supervivencia – Canal Radicador								
Cod	**Canal**	**records**	**n.max**	**n.start**	**events**	**median**	**0.95LCL**	**0.95LCL**
0	Natural	1492	1492	1492	798	66	63	69
1	Preventa	3816	3816	3816	934	47	45	48

En la Tabla 3.1 M se muestra una clara diferencia en los tiempos de supervivencia entre los clientes de canal natural y preventa, los que obtuvieron la tarjeta por el canal natural presentan un intervalo de confianza al 95% para la media de supervivencia mucho mayor que el de los clientes del canal preventa, es decir con intervalo entre 63 y 69 meses, mientras que para los clientes del canal preventa es entre 45 y 48 meses.

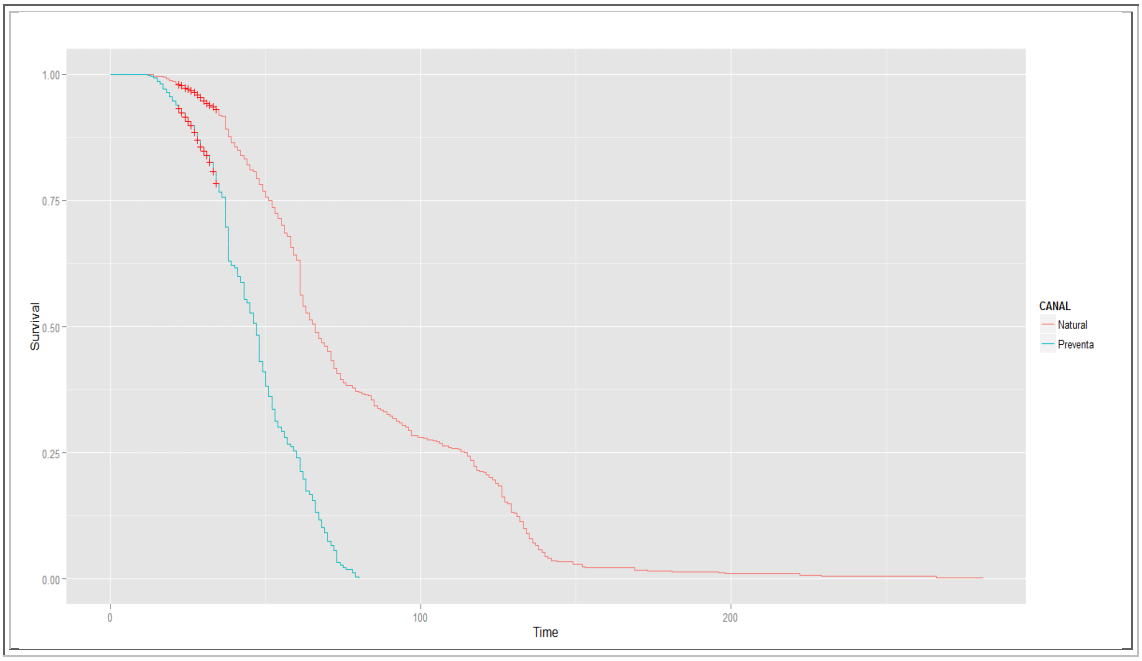
Tabla 3.1 N Estadístico de Prueba – Estimador Kaplan Meier

Chi-cuadrado	gl	Sig.
424	1	0.00

Se puede observar que en los resultados obtenidos mediante la prueba de Chi - cuadrado, el valor p=0, lo cual significa que a un nivel de significancia α=0.05, existe suficiente evidencia estadística entre las curvas de supervivencia de canal radicador.

Figura 3.1 E Función de Supervivencia por Canal Radicador

.



En el Figura 3.1 E se muestran las diferentes curvas de supervivencia por canal, se nota en términos de supervivencia que los clientes de canal natural tienen más probabilidad de supervivencia, mientras que los clientes que obtuvieron la tarjeta de crédito mediante el canal preventa tienen menor tiempo de supervivencia.

Por Tipo de Tarjeta

Se muestra que el total de clientes desertores con tarjeta de crédito tipo Clásica durante el periodo de análisis fue 623, los clientes que desertaron con el tipo de tarjeta Oro fue 534, no hubo desertores en tarjetas de crédito tipo Corporativa (Ver Tabla 3.1 O)

Tabla 3.1 O Estimador Kaplan Meier – Tipo de Tarjeta

Cod	Tipo de Tarjeta	N	Observed	Expected	(O-E)^2/E	(O-E)^2/V
1	Oro	1951	534	520.146	0.369	0.550
2	Clásica	1552	623	680.997	4.939	8.575
3	Platinum	780	218	191.082	3.792	4.436
4	Signature	748	144	139.848	0.123	0.142
5	Nacional	145	94	97.176	0.104	0.115
6	De afinidad y marcas compartidas	123	116	98.918	2.950	3.347
7	Corporativa	6	0	0.833	0.833	0.841

En la siguiente Tabla 3.1 P, se puede observar que los clientes que tienen la tarjeta tipo Oro tienen un tiempo de supervivencia con una mediana de 58 meses, en la categoría Clásica es de 57 meses, en Platinum y Signature es de 50 y 49 meses respectivamente, mientras que en la de tipo Nacional es de 62 meses.

Tabla 3.1 P Estimador Kaplan Meier Tiempo de Supervivencia – Tipo de Tarjeta

Cod	Tipo de Tarjeta	records	n.max	n.start	events	median	0.95LCL	0.95LCL
1	Oro	1951	1951	1951	534	58	53	61
2	Clásica	1552	1552	1552	623	57	54	61
3	Platinum	780	780	780	218	50	47	54
4	Signature	748	748	748	144	49	40	85
5	Nacional	145	145	145	94	62	60	70
6	De afinidad y marcas compartidas	123	123	123	116	59	57	62
7	Corporativa	6	6	6	0	NA	NA	NA

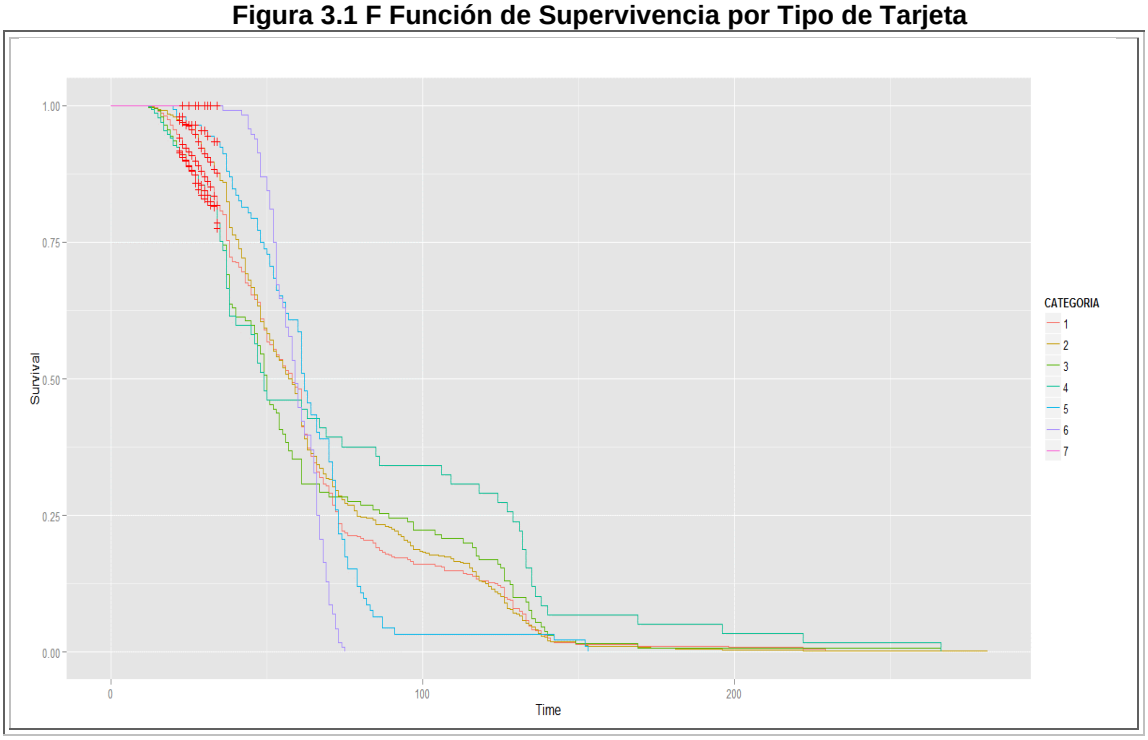
Se puede observar que en los resultados obtenidos mediante la prueba de Chi - cuadrado, el valor p=0.0338, lo cual significa que a un nivel de significancia α=0.05, existe evidencia estadística entre las curvas de supervivencia de tipo de tarjeta.

Tabla 3.1 Q Estadístico de Prueba – Estimador Kaplan Meier

Chi-cuadrado	gl	Sig.
13.7	6	0.0338

En la Figura 3.1 F se muestra como las tarjetas tipo: Oro, Clásica, Nacional y de Afinidad son las que tienen un mayor nivel de supervivencia contra las que son de tipo Platinum y Signature que muestran un nivel de supervivencia más bajo.

Figura 3.1 F Función de Supervivencia por Tipo de Tarjeta



Por Actividad Económica

A continuación, se presentan los resultados obtenidos y las curvas de supervivencia por actividad económica y sus comparaciones.

Tabla 3.1 R Estimador Kaplan Meier– Actividad Económica						
Cod	**Actividad Económica**	**N**	**Observed**	**Expected**	**(O-E)^2/E**	**(O-E)^2/V**
1	Empleado privado	2686	973	1058.3	6.870	18.627
2	Empleado público	760	266	213.2	13.100	15.443
3	Actividades profesionales	382	92	78	2.500	2.725
4	Jubilado	405	59	33.8	18.740	20.097
5	Comercio mayor y menor	348	111	128.1	2.280	2.603
6	Otros	727	231	220.6	0.490	0.579

Se puede apreciar en la Tabla 3.1 R que los clientes que son Empleados Privados son los mayores desertores, siendo en total 973, mientras los que son Empleados Públicos fueron 266 desertores, por otro lado, se puede observar que hubo 111 desertores que se tienen como actividad económica Comercio mayor y menor.

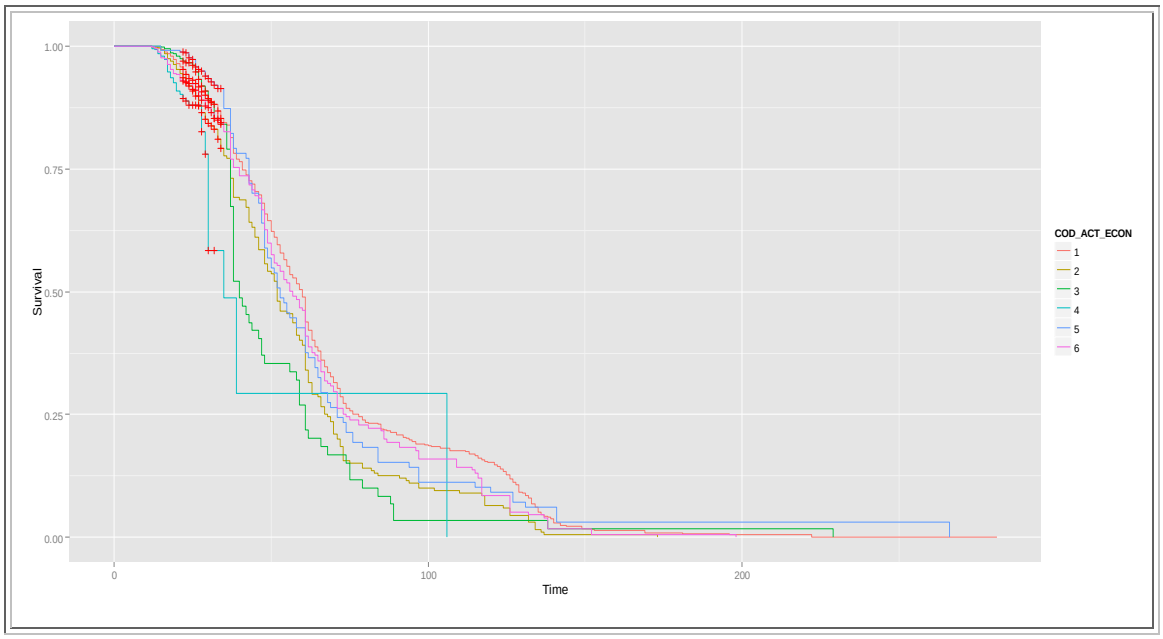
Tabla 3.1 S Estimador Kaplan Meier Tiempo de Supervivencia – Actividad Económica						
Cod	**Actividad Económica**	**N**	**events**	**median**	**0.95LCL**	**0.95LCL**
1	Empleado privado	2686	973	60	57	61
2	Empleado público	760	266	52	48	58
3	Actividades profesionales	382	92	40	38	48
4	Jubilado	405	59	35	30	NA
5	Comercio mayor y menor	348	111	53	49	61
6	Otros	727	231	57	51	61

En la Tabla 3.1 S se muestra los tiempos de supervivencia de los clientes por actividad económica, los que son empleados privados presentan un intervalo de confianza al 95% para la media de supervivencia mayor que los demás clientes con otro tipo de actividad económica, es decir con intervalo entre 57 y 61 meses.

Tabla 3.1 T Estadístico de Prueba – Estimador Kaplan Meier		
Chi-cuadrado	gl.	Sig.
47	5	5.77E-09

Se puede observar que en los resultados obtenidos mediante la prueba de Chi - cuadrado, el valor p=5.77E-09, lo cual significa que a un nivel de significancia α=0.05, existe evidencia estadística entre las curvas de supervivencia.

Figura 3.1 G Función de Supervivencia por Actividad Económica



En la Figura 3.1 G se muestran las diferentes curvas de supervivencia por actividad económica, se nota en términos de supervivencia que los clientes que son Empleados Privados o que se tienen Actividades de Comercios por mayor y menor tienen más probabilidad de supervivencia, mientras que los clientes que son Jubilados o tienen de actividad Servicios Profesionales tienen menor tiempo de supervivencia.

Por Indicador Transacciones Masivas

A continuación, se presentan los resultados obtenidos del estimador y las curvas de supervivencia de los clientes que solo utilizaron la tarjeta de crédito para realizar transacciones masivas.

Tabla 3.1 U Estimador Kaplan Meier– Indicador Transacciones Masivas

Cod	Indicador Transacciones Masivas	N	Observed	Expected	(O-E)^2/E	(O-E)^2/V
0	No	5059	1624	1680	1.87	64.00
1	Si	249	108	52	60.41	64.00

Se puede apreciar en la Tabla 3.1 U de los clientes que desertaron y que solo realizaron transacciones masivas fueron 108.

Tabla 3.1 V Estimador Kaplan Meier Tiempo de Supervivencia – Indicador Transacciones Masivas

Cod	Indicador Transacciones Masivas	N	events	median	0.95LCL	0.95LCL
0	No	5059	1624	57	55	59
1	Si	249	108	43	36	54

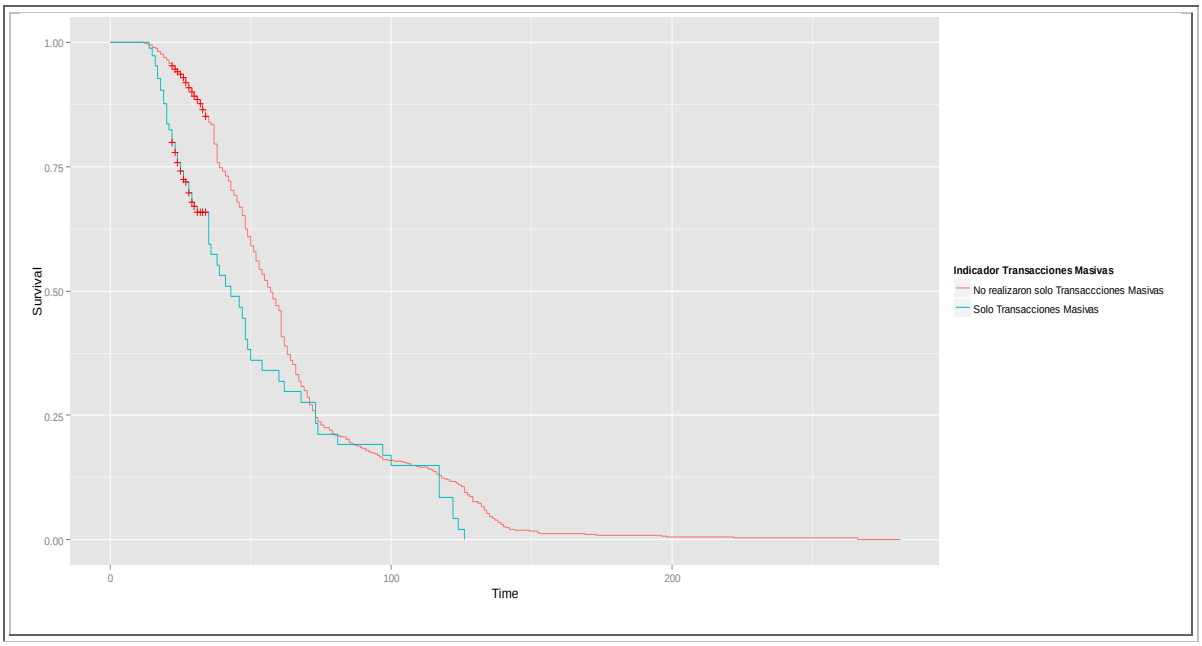
En la Tabla 3.1 V se muestra los tiempos de supervivencia de los clientes que solo realizaron transacciones masivas, presentan un intervalo de confianza al 95% para la media de supervivencia menor que los demás clientes que realizaron otro tipo de transacciones además de solo masivo, es decir con intervalo entre 36 y 54 meses.

Tabla 3.1 W Estadístico de Prueba – Estimador Kaplan Meier

Chi-cuadrado	gl.	Sig.
64	1	1.22E-15

Se puede observar que en los resultados obtenidos mediante la prueba de Chi - cuadrado, el valor p=1.22E-15, lo cual significa que a un nivel de significancia α=0.05, existe evidencia estadística entre las curvas de supervivencia.

Figura 3.1 H Función de Supervivencia por Indicador de Transacciones Masivas



En la Figura 3.1 H se muestran las diferentes curvas de supervivencia de los clientes que solo realizaron transacciones masivas, se nota en términos de supervivencia este grupo de clientes tienen menos probabilidad de supervivencias, a diferencia de los clientes que utilizan la tarjeta de crédito en transacciones en establecimientos.

Por Cantidad de Transacciones (Rango)

A continuación, se presentan los resultados obtenidos y las curvas de supervivencia por cada rango del total de transacciones facturadas.

Tabla 3.1 X Estimador Kaplan Meier– Total Transacciones Facturadas Anual

Cod	Rango Total Transacciones Facturadas	N	Observed	Expected	(O-E)^2/E	(O-E)^2/V
1	Igual a 0	772	288	339.6	7.8379	10.3389
2	(0 a 5]	376	212	219	0.2207	0.2678
3	(5 a 15]	991	423	331.9	25.0235	32.0299
4	(15 a 25]	935	298	272.3	2.4317	3.0129
5	(25 a 35]	678	186	179.1	0.2641	0.3043
6	(35 a 45]	445	91	89.2	0.0375	0.0409
7	(45 a 55]	289	66	79.1	2.1752	2.3487
8	(55 a 65]	225	48	52.8	0.4347	0.4626
9	(65 a 75]	156	33	47.8	4.5936	4.8861
10	Mayor a 75	441	87	121.3	9.6968	10.8201

Se puede apreciar en la Tabla 3.1 X que en total 288 de los clientes que desertaron no realizaron algún tipo de transacción, 212 clientes desertores al menos tuvieron en el año 5 transacciones, mientras que 423 de los desertores registraron transacciones en un rango de 5 a 15 en total al año.

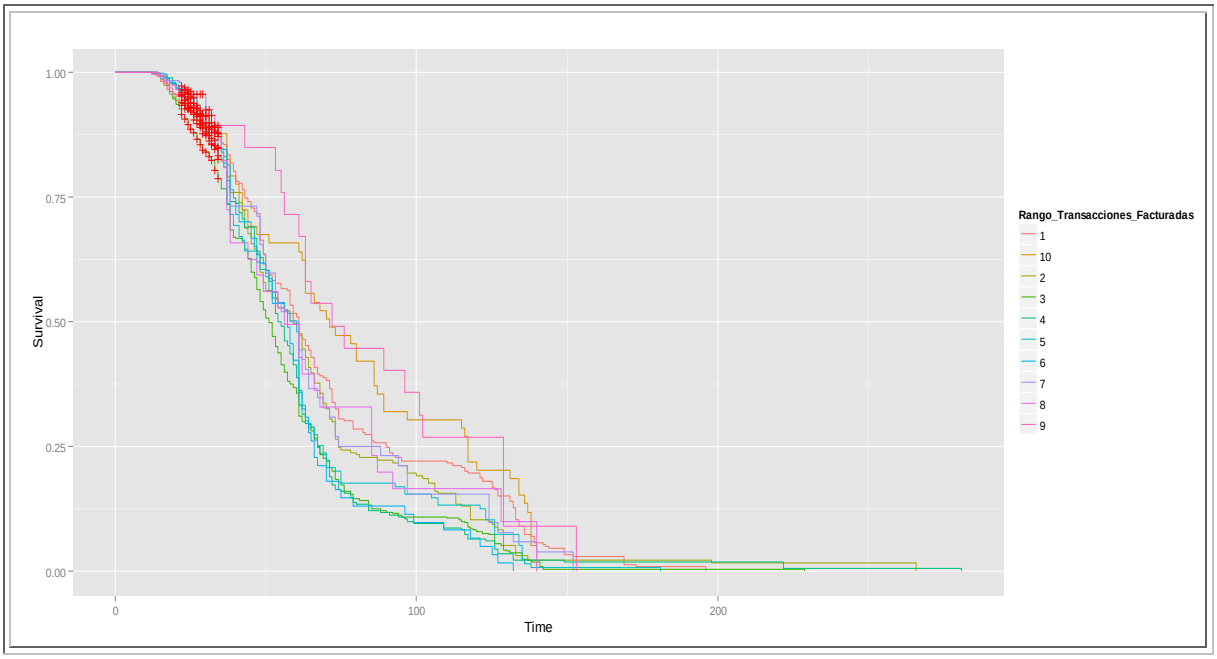
Tabla 3.1 Y Estimador Kaplan Meier Tiempo de Supervivencia – Total Transacciones Facturadas Anual (Rangos)						
Cod	**Rango Total Transacciones Facturadas**	**N**	**events**	**median**	**0.95LCL**	**0.95LCL**
1	Igual a 0	772	288	61	58	65
2	(0 a 5]	376	212	59	50	64
3	(5 a 15]	991	423	51	48	54
4	(15 a 25]	935	298	54	52	59
5	(25 a 35]	678	186	60	51	61
6	(35 a 45]	445	91	57	50	61
7	(45 a 55]	289	66	58	50	67
8	(55 a 65]	225	48	56	44	85
9	(65 a 75]	156	33	72	63	129
10	Mayor a 75	441	87	71	63	87

En la Tabla 3.1 Y se muestra los tiempos de supervivencia de los clientes por rango de transacciones facturadas al año, los que registraron transacciones de rangos mayores es decir mayor a 65, presentan un intervalo de confianza al 95% para la media de supervivencia mayor que los demás clientes realizaron transacciones inferiores, es decir con intervalo entre 63 y 129 meses.

Tabla 3.1. Z Estadístico de Prueba – Estimador Kaplan Meier		
Chi-cuadrado	gl.	Sig.
55.3	9	1.09E-08

Se puede observar que en los resultados obtenidos mediante la prueba de Chi - cuadrado, el valor p=1.09E-08, lo cual significa que a un nivel de significancia α=0.05, existe evidencia estadística entre las curvas de supervivencia.

Figura 3.1 I Función de Supervivencia por Rango Total Transacciones Facturadas Anual



En la Figura 3.1 I se muestran las diferentes curvas de supervivencia por rango de total transacciones facturadas en el año, se nota en términos de supervivencia que los clientes que tienen mayor número de transacciones en el año tienen más probabilidad de supervivencia.

3.1.1 VARIABLES EXPLICATIVAS CONSIDERANDO ESTIMADORES KAPLAN - MEIER

Se obtuvo de los resultados proporcionados por el Estimador Kaplan Meier las variables que inciden en la deserción de los clientes.

Tabla 3.1.1 Variables Inciden Deserción	
Variable	**Significancia (valor p)**
X_1.- Edad	0.000
X_2.- Segmento	6.77E-15
X_3.- Canal Radicador	0.000
X_4.- Tipo de Tarjeta	0.0338
X_5.- Actividad Económica	5.77E-09
X_6.- Indicador solo Transacción Masiva	1.22E-15
X_7.- Total de Transacciones Facturadas	1.09E-08

Como se muestra en la Tabla 3.1.1, se puede observar que en los resultados obtenidos mediante la prueba de Chi - cuadrado, se presenta el valor p de las variables, lo cual significa que a un nivel de significancia α=0.05, se concluye que existe suficiente evidencia estadística para determinar que son variables explicativas que influyen para predecir la deserción.

3.2 ANÁLISIS DE SUPERVIVENCIA CON REGRESION DE COX

Este modelo fue propuesto por el estadístico británico David Cox en 1972, es un modelo usado para modelar los riesgos que afectan a la supervivencia de una población de sujetos [7].

El modelo de Cox expresa la función de riesgo instantáneo de muerte λ en función del tiempo t y de variables X_1, X_2, X_3,…, X_n así:

$$\lambda(t, X_1, \dots, X_n) = \lambda_0(t)\exp(\sum_{i=1}^{n} \beta_i X_i)$$

Sea $\lambda_0(t)$ la función de riesgo base y corresponde al riesgo de muerte cuando todas las variables tienen valor 0. Es la única parte de la expresión que depende del tiempo, la otra parte $\exp(\sum_{i=1}^{n} \beta_i X_i)$, sólo depende del resto de las variables.

3.2.1 LA HIPÓTESIS DE LOS RIESGOS PROPORCIONALES

El modelo parte de una hipótesis fundamental: los riesgos son proporcionales. Para comprender esta noción, puede estudiarse el caso de dos individuos, i y j que sólo se diferencian en la k-ésima variable. Supóngase que vale 0 para i y 1 para j. Entonces, para cualquier tiempo t,

$$\frac{\lambda(t,j)}{\lambda(t,i)} = \frac{\lambda_0(t)\exp(\beta_1 X_1' + \cdots + \beta_{k-1}X_{k-1}' + \beta_k * 1 + \beta_{k+1}X_{k+1}'\beta_n X_n')}{\lambda_0(t)\exp(\beta_1 X_1' + \cdots + \beta_{k-1}X_{k-1}' + \beta_k * 0 + \beta_{k+1}X_{k+1}'\beta_n X_n')} = \exp(\beta_k)$$

La fórmula muestra que el cociente es independiente del tiempo [8].

Al utilizar la regresión de Cox es necesario verificar que se cumple dicha hipótesis. Para ello es necesario comprobar que el efecto de cada variable es constante en el tiempo.

En la sección anterior se realizó un análisis exploratorio de la situación de supervivencia de los clientes mediante el estimador de Kaplan-Meier. En esta sección se busca estimar un modelo que permita explicar la supervivencia de los clientes a partir de una serie de variables sociodemográficas.

Las variables a analizar son las disponibles en la base de datos para la elaboración de este estudio (edad, estado civil, segmento, canal radicador y productos). Para la estimación del modelo se utilizó el paquete de supervivencia mediante el software R [9].

Para el análisis, en el caso de las variables explicativas de índole nominal, con más de dos categorías (politómicas), recibieron un tratamiento especial, para incluirlas en el modelo. En el caso de las variables como estado civil, segmento, canal radicador, productos, actividad económica, calificación central riesgo y Ubicación geográfica con más de 2 categorías, se incluyó en el modelo de regresión de Cox como variable categórica, de manera que a partir de ella se crean variables dicotómicas, llamadas dummy o ficticias.

Se escogerá un modelo apropiado que contenga el número mínimo de variables explicativas posible del conjunto de variables.

A continuación, se muestra la Regresión de Cox los 3 grupos de variables que se consideró: demográficas, externas y comportamiento transaccional, con lo que obtenemos el siguiente resultado que se muestra en el Tabla 3.2.1

Tabla 3.2.1 Modelo de Regresión de Cox						
Variables		**coef**	**exp(coef)**	**se(coef)**	**z**	**p**
Canal		1.19E+00	3.27E+00	7.72E-02	15.36	< 2e-16
Edad		-1.67E-02	9.83E-01	2.72E-03	-6.12	9.3E-10
Segmento	Básico	-3.57E-01	7.00E-01	1.59E-01	-2.25	2.57E-02
	Estándar	2.34E-01	1.26E+00	1.36E-01	1.72	8.56E-02
	Medio	2.13E-01	1.24E+00	1.32E-01	1.61	1.07E-01
	Medio Alto	2.66E-01	1.31E+00	1.43E-01	1.87	6.19E-02
	Alto	NA	NA	0.00E+00	NA	NA
Estado Civil	Soltero	7.73E-01	2.17E+00	4.58E-01	1.69	9.18E-02
	Casado	7.73E-01	2.17E+00	4.56E-01	1.70	9.00E-02
	Unión Libre	7.71E-01	2.16E+00	5.21E-01	1.48	1.14E-01
	Divorciado	7.85E-01	2.19E+00	4.69E-01	1.67	9.42E-02
	Viudo	7.86E-01	2.19E+00	5.06E-01	1.55	1.20E-01
Productos	Cuenta Corriente	5.06E-02	1.05E+00	1.64E-01	0.31	7.58E-01
	Cuenta Ahorro	-1.20E-01	8.87E-01	1.75E-01	-0.68	4.94E-01
	Deposito Plazo	-1.27E-01	8.81E-01	2.59E-01	-0.49	6.24E-01
	Préstamos	-2.74E-02	9.73E-01	2.39E-01	-0.11	9.09E-01
	No Tiene	-4.94E-01	6.10E-01	1.97E-01	-2.51	1.20E-02
Monto Deuda Riesgo		5.86E-08	1.00E+00	4.81E-07	0.12	9.03E-01
Calificación Central Riesgo	A	-4.78E-02	9.53E-01	1.23E-01	-0.39	6.97E-01
	B	-2.56E-02	9.75E-01	1.88E-01	-0.14	8.92E-01
	C	7.17E-02	1.07E+00	4.70E-01	0.15	8.79E-01
	D	-4.68E-01	6.26E-01	3.80E-01	-1.23	2.18E-01
	E	-1.73E-01	8.41E-01	2.98E-01	-0.58	5.62E-01
Ingreso Neto		1.05E-04	1.00E+00	2.35E-05	4.48	7.6E-06
Ingreso Mensual		-1.01E-04	1.00E+00	2.22E-05	-4.56	5.2E-06
Ubicación geográfica	Guayaquil	-1.80E-01	8.35E-01	8.55E-02	-2.10	3.53E-02
	Quito	5.13E-02	1.05E+00	8.29E-02	0.62	5.36E-01
	Cuenca	1.59E-01	1.17E+00	1.45E-01	1.10	2.71E-01
	Machala	7.69E-02	1.08E+00	2.06E-01	0.37	7.09E-01
Actividad Económica	Empleado Privado	-4.00E-01	6.70E-01	9.40E-02	-4.26	2.10E-05
	Empleado Público	-2.19E-01	8.03E-01	1.14E-01	-1.92	5.43E-02
	Actividades Profesionales	-2.94E-01	7.45E-01	1.55E-01	-1.90	5.80E-01
	Jubilado	1.05E-01	1.11E+00	1.90E-01	0.55	5.83E-01
	Comercio Mayor y Menor	-2.78E-01	7.57E-01	1.48E-01	-1.87	6.09E-02
Total Monto Consumido		-1.00E-05	1.00E+00	1.61E-05	-0.62	5.33E-01
Saldo Consumido		3.06E-04	1.00E+00	2.16E-04	1.42	1.57E-02
Saldo Rotativo		-2.66E-04	1.00E+00	9.48E-05	-2.80	5.00E-03
Cantidad Meses Sobregirado		-3.12E-03	9.97E-01	1.98E-02	-0.16	8.74E-01
Indicador Sobregiro		6.46E-02	1.07E+00	1.28E-01	0.50	6.14E-01
Monto Sobregiro		-1.51E-04	1.00E+00	6.66E-04	-0.23	8.20E-01
Total Transacciones Facturadas		-1.48E-03	9.99E-01	1.31E-03	-1.13	2.59E-03
Indicador Transacciones Masivas		-3.01E-01	7.40E-01	9.53E-02	-3.16	1.60E-03
Total Transacciones Masivas		8.06E-03	1.01E+00	4.73E-03	1.70	8.87E-02

Canal tiene un p-valor=< 2e-16 menor que α=0.05, lo que indica que el canal radicador es un factor pronóstico de la deserción de clientes. La regresión de Cox también nos aporta la razón de riesgo(exp(coef)) que tiene por valor 3.27, lo cual significa que los clientes que obtuvieron la Tarjeta de Crédito mediante canal preventa tienen un riesgo 3.27 veces mayor de deserción que los clientes que obtienen la tarjeta de crédito por canal mercado natural.

Edad si influye en la deserción de clientes, es decir al aumentar la edad de las personas, el riesgo de deserción disminuye en (1-0.983) =0.017=1.7%. De esta forma se reduciría la deserción de clientes por cada incremento de edad. Es decir, las personas a mayor edad es menos probable que deserten.

Estado Civil, en esta variable politómica se la transformó en dummy para cada una de las categorías existentes, se puede observar que los clientes que tienen estado civil soltero, casado, divorciado y unión libre son factores de pronóstico de la deserción, el valor p es estadísticamente significativo en cada categoría.

Productos, en los clientes que no tienen ningún otro tipo de producto además de la tarjeta de crédito es un factor predictor porque el p valor es menor que α=0.05 lo que sí es estadísticamente significativo, tiene una razón de riesgo del 6.10% es decir que deben enfocarse en clientes que ya poseen por lo menos un producto/servicio con el Banco, ya que los clientes que solo tienen Tarjeta de Crédito y ningún otro producto son los más probables a desertar.

La variable **Segmento**, se puede concluir que los segmentos Básico, Estándar, y Medio Alto son estadísticamente significativo dado que el valor p es menor a α=0.05 por lo tanto son factores de predicción.

Ingreso Mensual e Ingreso Neto, si influye en la deserción de clientes, es decir al percibir menor ingresos mensuales, el riesgo de deserción aumenta. De esta forma se reduciría la deserción de clientes en personas que tienen mejores ingresos.

Ubicación geográfica los que son de la Ubicación geográfica Guayaquil tiene un p-valor 3.53E-02 menor que α=0.05, lo que indica que es un factor pronóstico de la deserción de clientes.

La variable **Actividad Económica**, se puede concluir que los que son Empleados Privados, Privados o los que se dedican a actividades de Comercios Mayor y Menor son estadísticamente significativo dado que el valor p es menor a α=0.05 por lo tanto son factores de predicción.

Total transacciones facturadas si influye en la deserción de clientes, es decir al incrementarse el número de transacciones, el riesgo de deserción disminuye en (1-0.999) =0.001=0.1%.

Indicador de Transacciones Masivas tiene un p-valor= 1.60E-03 menor que α=0.05, lo que indica que las personas que solo utilizaron la T/C para realizar transacciones de tipo masivo es un factor pronóstico de la deserción de clientes.

3.2.2 VARIABLES EXPLICATIVAS APLICANDO REGRESIÓN DE COX

A continuación se realizará la verificación de los supuestos del modelo de Cox, es decir el supuesto de riesgos proporcionales se lo realizó mediante el contraste de hipótesis en donde se obtuvo los siguientes resultados.

Tabla 3.2.2 Supuestos de Riesgos Proporcionales				
Variables		**rho**	**chisq**	**p**
Canal		0.03180	1.35613	0.2442
Edad		-0.04799	2.77972	0.0955
Segmento	Básico	0.00420	0.02164	0.8830
	Estándar	0.01612	0.32782	0.5669
	Medio	0.00391	0.01919	0.8898
	Medio Alto	-0.02163	0.58823	0.4431
Producto	CDP	-0.02589	0.82321	0.3642
Ubicación geográfica	Quito	-0.02947	1.08126	0.2984
	Cuenca	-0.01515	0.28799	0.5915
Estado Civil	Soltero	-0.01684	0.34188	0.5587
	Casado	-0.02030	0.49376	0.4823
	Unión Libre	-0.02030	0.49376	0.4823
	Divorciado	-0.01057	0.13409	0.7142
	Viudo	-0.01606	0.31110	0.5770
Actividad Económica	Empleado Privado	0.01942	0.46169	0.4968
	Actividades Serv. Profesional	0.02170	0.59961	0.4387
	Jubilado	0.00225	0.00617	0.9374
Indicador Transacciones Masivas		0.00885	0.09726	0.7551
GLOBAL		**NA**	**21.36222**	**0.2615**

De donde se concluye que no existe evidencia significativa con $\propto = 0.05$ de que se viole el supuesto de riesgos proporcionales, el resultado GLOBAL el valor p=0.2615, entonces desde el punto de vista global y para cada covariable se cumple el supuesto de proporcionalidad.

Para la construcción del modelo se utilizó una base de 5,308 clientes, de los cuales 1,732 son clientes desertores como se mencionó en el Capítulo II.

Mediante el empleo de muestreo aleatorio simple se seleccionó dos conjuntos de datos, de desarrollar el modelo con el conjunto de datos de entrenamiento que representó el 70% y probarlo con la segunda que corresponde al conjunto de datos de prueba que corresponde el 30% de la población.

De los resultados proporcionados del modelo de regresión de Cox se puede concluir que el canal radicador, edad, estado civil, segmento (excepto alto), actividad económica sea empleado privado, público servicios profesionales o jubilado, los de la Ubicación geográfica Quito o Cuenca e indicador transacciones masivas son las variables que más inciden en el riesgo de desertar.

El siguiente cuadro muestra el desempeño del modelo de Cox aplicado, mediante el conjunto de datos de prueba. Se puede observar que el modelo presenta un error del 45.19% y una precisión del 54.81% lo cual nos indica que el modelo no es lo suficientemente poderoso en la predicción.

Rendimiento Modelo	**Modelo Cox**
Error Promedio %	45.19%
Precisión (Accuracy) %	54.81%
Precisión Negativa	723
Falso Negativo	366
Precisión Positiva	165
Falso Positivo	366

Sin embargo, el uso de modelos de supervivencia nos permite estimar con mejor precisión las variables demográficas y comportamiento del cliente a través del tiempo en su intención de desertar.

Cabe recalcar que los modelos de supervivencia utilizados en este capítulo corresponden a modelos semiparamétricos. En el capítulo IV se construirá un modelo en el cual se tendrá a partir de una serie de variables una mayor precisión en el modelo de predicción de deserción de clientes mediante la aplicación de árboles de decisión y adaboost.

CAPÍTULO IV

4. MODELOS DE PREDICCIÓN Y BOOSTING

4.1 INTRODUCCIÓN A MODELOS DE PREDICCIÓN

En la actualidad existen numerosas técnicas de predicción. Muchas de ellas se han aplicado con éxito en la industria de telecomunicaciones, instituciones financieras, entre otras [12]. La particularidad de estas técnicas donde se han implementado modelos de predicción es que tienen un enfoque de Data Mining, que consiste en procesar y extraer datos para buscar patrones recurrentes desde los cuales se puedan encontrar reglas para la toma de decisiones. En nuestro caso, predecir al abandono.

Cuando la variable a predecir es discreta entonces el problema de clasificación consiste en predecir el estado del cliente dado el conjunto de atributos que lo caracteriza.

Los modelos más utilizados para el problema de la predicción de la deserción de clientes en la literatura son:

- **Regresión Logística:** es una técnica de modelización estadística con resultados dicotómicos y que transforma la variable respuesta de $\mathbb{R}$ a [0,1] [10].
- **Árboles de Decisión:** es una técnica de clasificación que puede representar el conocimiento extraído en un conjunto de reglas de decisión de fácil entendimiento, además que puede considerar atributos continuos y discretos [10].
- **Bosques Aleatorios:** introducido por Breiman (2001), es una combinación de árboles predictores tal que cada árbol depende de los valores de un vector aleatorio muestreado independientemente y con la misma distribución para cada uno de estos [11].
- **Máquinas de Soporte Vectorial:** son modelos de aprendizaje basado en las estadísticas de la teoría del aprendizaje y que puede resolver la

no linealidad, la máxima dimensión, y los problemas de minimización locales [12].

- **Redes Neuronales:** es un modelo matemático no lineal, de naturaleza computacional. Frente a los métodos estadísticos estándar, no requiere de ningún conocimiento previo acerca del fenómeno de estudio, ni hipótesis sobre su distribución, modelizando de modo natural cualquier tipo de relación no lineal con la variable respuesta, así como la existencia de posibles interacciones entre las variables explicativas. Este modelo ha sido utilizado ampliamente en muchos problemas de negocios [13].

En este estudio se utilizará técnicas de Árboles de Decisión y Boosting, aplicadas a Árboles de decisión. Boosting es un conjunto de técnicas en Aprendizaje de Máquina para incrementar significativamente la precisión de predicción sin sobreajustar.

4.2 ÁRBOLES DE DECISIÓN

Los árboles de decisión son una técnica bastante conocida y ha tenido muchas aplicaciones exitosas a problemas reales (Tsai y Chiou, 2009). Debido a que el modelo no pertenece a una familia de modelos parametrizables.

Consideraremos la variable dependiente binaria (Desertor/No desertor) y las variables explicativas (cualitativa y cuantitativas), la técnica del árbol de decisión es un algoritmo recursivo que consiste en particionar la población de estudio en segmentos homogéneos mediante la utilización de reglas de partición basadas en los valores que tomen las variables explicativas.

El proceso iterativo que se sigue para generar los segmentos es el siguiente: Primero se particiona la población de dos subconjuntos homogéneos, luego cada uno de estos subconjuntos es particionado nuevamente en dos subconjuntos más homogéneos, el proceso es repetido recursivamente y termina si el subconjunto presenta una cantidad de clientes menor o igual a la

mínima requerida (criterio de parada), finalmente se establece el tipo de subconjunto (Desertor/No desertor).

En el presente estudio emplearemos los árboles de decisión para identificar y construir características que permitan generar las variables de decisión basadas en los clientes que desertaron.

4.2.1 VARIABLES DE DECISIÓN A UTILIZAR

Tabla 4.2.1 Variables Decisión a utilizar

Var	Descripción	Mínimo	Máximo
X_1	Ubicación geográfica (categórica)	1	6
X_2	Antigüedad(meses)	12	281
X_3	Canal radicador (categórica)	0	1
X_4	Edad	18	113
X_5	Género (categórica)	1	2
X_6	Estado civil (categórica)	1	5
X_7	Actividad económica (categórica)	1	6
X_8	Segmento (categórica)	1	5
X_9	Dummy, si tiene cuenta corriente es 1 sino 0	0	1
X_{10}	Dummy, si tiene cuenta de ahorro es 1 sino 0	0	1
X_{11}	Dummy, si tiene depósito a plazo es 1 sino 0	0	1
X_{12}	Dummy, si tiene préstamo es 1 sino 0	0	1
X_{13}	Ingreso neto	0	150,000
X_{14}	Ingreso mensual	0	204,000
X_{15}	Categoría tarjeta Visa (categórica)	1	7
X_{16}	Monto deuda sistema financiero	2.01	8,114,040
X_{17}	Calificación central riesgo	1	5
X_{18}	Total monto consumo anual	0	244,378
X_{19}	Monto Promedio consumo	0	20,365
X_{20}	Monto Promedio consumo rotativo	0	19,231
X_{21}	Monto Promedio consumo diferido	0	3,438
X_{22}	Cantidad veces sobregirado	0	12
X_{23}	Monto Promedio sobregiros	0	1,374
X_{24}	Total transacciones facturadas	0	647
X_{25}	Total transacciones masivas	0	82
X_{26}	Dummy, si solo utiliza la T/C pagos recurrentes es 1 sino 0	0	1

Como se muestra en la Tabla 4.2.1 existe disponibilidad de 26 variables, de los cuales se construyeron 5 modelos con diferentes combinaciones de variables para evaluar el poder de los diferentes tipos de información.

Modelo 1: el primer modelo es construido en base a las variables predictoras que dieron como resultado aplicando el estimador de Kaplan Meier que se realizó en el capítulo 2. Estas variables son:

Var	Descripción
X_3	Canal radicador
X_4	Edad
X_7	Actividad económica
X_8	Segmento
X_{24}	Total transacciones facturadas
X_{26}	Dummy, si solo utiliza la T/C pagos recurrentes es 1 sino 0

Modelo 2: el segundo modelo es construido en base a las variables predictoras aplicando el estimador Regresión de Cox en el capítulo 2.

Var	Descripción
X_1	Ubicación geográfica
X_3	Canal radicador
X_4	Edad
X_6	Estado civil
X_7	Actividad económica
X_8	Segmento
X_{25}	Total transacciones masivas

Modelo 3: el tercer modelo se construyó en base a variables de comportamiento (transaccional), externas (monto deuda en el sistema financiero) e ingresos netos.

Var	Descripción
X_{13}	Ingreso neto
X_{16}	Monto deuda sistema financiero
X_{18}	Total monto consumo anual
X_{19}	Monto Promedio consumo
X_{20}	Monto Promedio consumo rotativo
X_{21}	Monto Promedio consumo diferido
X_{22}	Cantidad veces sobregirado
X_{23}	Monto Promedio sobregiros
X_{24}	Total transacciones facturada
X_{25}	Total transacciones masivas
X_{26}	Dummy, si solo utiliza la T/C pagos recurrentes es 1 sino 0

Modelo 4: el cuarto modelo se construyó en base a variables de comportamiento (transaccional), externas y demográficas. En total las 26 variables que se mencionaron en Tabla 4.2.1.

Modelo 5: el quinto modelo consiste en las variables demográficas y externas. Como se muestra a continuación:

Var	Descripción
X_1	Ubicación geográfica (categórica)
X_2	Antigüedad(meses)
X_3	Canal radicador (categórica)
X_4	Edad
X_5	Género (categórica)
X_6	Estado civil (categórica)
X_7	Actividad económica (categórica)
X_8	Segmento (categórica)
X_9	Dummy, si tiene cuenta corriente es 1 sino 0
X_{10}	Dummy, si tiene cuenta de ahorro es 1 sino 0
X_{11}	Dummy, si tiene depósito a plazo es 1 sino 0
X_{12}	Dummy, si tiene préstamo es 1 sino 0
X_{13}	Ingreso neto
X_{14}	Ingreso mensual
X_{16}	Monto deuda sistema financiero
X_{17}	Calificación central riesgo

4.3 ADABOOST

En este estudio proponemos usar el AdaBoost, para predecir la deserción de clientes de tarjetas de crédito. El AdaBoost es el principal meta-algoritmo de boosting, desarrollado por Freund y Schapire [14]. En el AdaBoost el resultado final es una suma ponderada de los resultados individuales de los algoritmos débiles (*). Debido a que el proceso de entrenamiento del AdaBoost selecciona solamente aquellos predictores que incrementan el poder de predicción del modelo, se reduce la dimensión y se mejora potencialmente el tiempo de ejecución, evitando la maldición de la dimensionalidad conocido como efecto Hughes.

(*) Algoritmo débil: Las estrategias de Boosting pretenden elevar el desempeño de un algoritmo de aprendizaje débil combinando varias hipótesis adecuadamente y generando un algoritmo de aprendizaje fuerte

Para conocimiento del autor, el AdaBoost con árboles de clasificación no ha sido aplicado antes al problema de predicción de deserción de clientes de tarjetas de crédito.

A continuación se menciona como funciona el algoritmo de Adaboost [7]

Entrada:

Conjunto de Entrenamiento (training set) $S = (x_1, y_1), \dots, (x_m, y_m)$

Donde $S = \{(x_\mathrm{i}, y_\mathrm{i}) : x_i \epsilon \mathbb{R}^d, y_i \epsilon \{-1, +1\}, i = 1, \dots, \mathrm{m}\}$

Por ejemplo, x_i podría representar el canal radicador de los clientes del banco y y_i sería si este cliente es desertor o no de acuerdo al canal radicador.

Weak learner WL (clasificadores débiles)

Número de iteraciones T

Inicializa Necesitamos definir $\boldsymbol{D}^{(t)}$ sobre el conjunto de datos (entrenamiento) S, es decir se entrena el modelo usando el set de entrenamiento.

$$\boldsymbol{D}^{(1)} = (\frac{1}{m}, \dots, \frac{1}{m}).$$

Para $t = 1, \dots, T$:

Recurre a los Weak learner $h_t = WL(\boldsymbol{D}^{(t)}, S)$

Calcular $\epsilon_t = \sum_{i=1}^{m} D_i^{(t)} \mathbb{1}_{[y_i \neq h_t(x_i)]}$

Definimos $w_{t=} \frac{1}{2} \log(\frac{1}{\epsilon_t} - 1)$

Actualizar los pesos para la iteración (t+1)

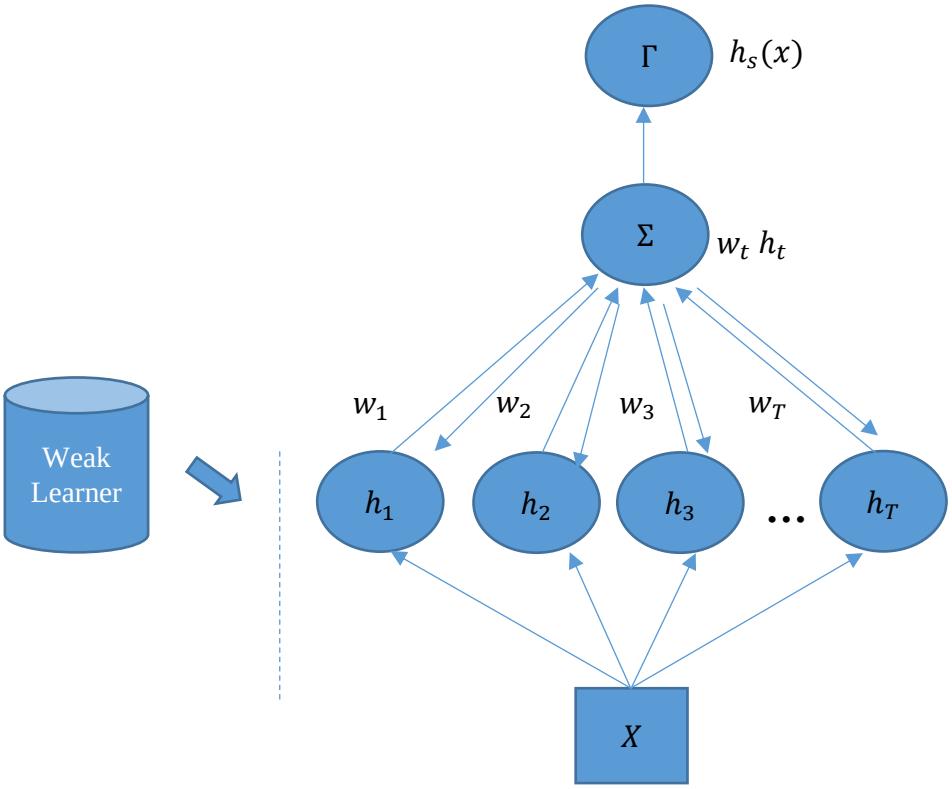
$$D_i^{(t+1)} = \frac{D_i^{(t)} \exp\left(-w_t y_i h_t(x_j)\right)}{\sum_{j=1}^{m} D_j^{(t)} \exp\left(-w_t y_j h_t(x_j)\right)} \quad \forall i = 1, \dots, m$$

Salida La salida del sistema será la votación ponderada de todos los h_t .

Hipótesis $h_s(x) = sign(\sum_{t=1}^{T} w_t h_t\,(x))$

Nuestro objetivo es elegir $h_t(x)$ y w_t para minimizar el error de clasificación empírica a partir de un fuerte clasificador.

A continuación se muestra una ilustración del clasificador de Adaboost [15]:



4.4 CRITERIOS DE EVALUACIÓN DE ALGORITMOS

4.4.1 MATRIZ DE CONFUSIÓN

En el campo de minería de datos una matriz de confusión es una herramienta que permite la visualización del desempeño de un algoritmo que se emplea en aprendizaje supervisado. Cada columna de la matriz representa el número de predicciones de cada clase, mientras que cada fila representa a las instancias en la clase real. Uno de los beneficios de las matrices de confusión es que facilitan ver si el sistema está confundiendo dos clases [16].

La siguiente tabla muestra la matriz de confusión para un clasificador de dos clases en este caso Desertor y No Desertor[17].:

Matriz de Confusión		Predicción	
		No Desertor *(Negativo)*	**Desertor** *(Positivo)*
Valor Real	**No Desertor** *(Negativo)*	a *Precisión Negativa*	b *Falso Positivo*
	Desertor *(Positivo)*	c *Falso Negativo*	d *Precision Positiva*

La terminología y derivaciones a partir de una matriz de confusión se muestran a continuación:

Sensitive (True Positive Rate *TPR*): es la proporción de casos positivos que fueron identificados correctamente, tal como se calcula usando la ecuación:

$$TPR = d/(c + d)$$

Specificity (True Negative Rate *TNR*): es la proporción de casos negativos que fueron identificados correctamente, tal como se calcula usando la ecuación:

$$TNR = a/(a + b)$$

Accuracy (ACC): es la proporción de casos positivos clasificados incorrectamente como negativos, tal como se calcula utilizando la ecuación:

$$ACC = (a + d)/(a + b + c + d)$$

Type I Error : es la proporción de casos negativos que fueron clasificados incorrectamente como positivos, tal como se calcula utilizando la ecuación:

$$Type\ I\ Error = b/(a + b)$$

Type II Error : es la proporción de casos positivos que fueron clasificados incorrectamente como negativos, tal como se calcula utilizando la ecuación:

$$Type\ II\ Error = c/(c+d)$$

Average Error : es la proporción de casos negativos que fueron clasificados incorrectamente como positivos y la proporción de casos positivos que fueron clasificados incorrectamente como negativos tal como se calcula utilizando la ecuación:

$$Average\ Error = (b+c)/(a+b+c+d)$$

4.4.2 FUNCIÓN DE PÉRDIDA

Además de analizar el desempeño del modelo mediante la Matriz de Confusión, se incorporará como indicador de rendimiento en base a la Función de pérdida que no es más que los costos que representa una mala claficación causada por los errores del modelo. [10]

La función de pérdida es contruida a partir de los siguientes factores como: Tasa de deserción, Type I Error, Type II Error mencionados en la sección de Matriz de Confusión y Factor Económico.

La fórmula de la función de pérdida está dada por [10]:

$$Misclassification\ Cost = Be_sP_{ave} + \frac{M}{Ge_f + B(1-e_s)}Ge_f$$

Donde,

$B = número\ de\ clientes\ desertores$
$e_f = Type\ I\ Error\ (Matrix\ Confusion)$
$e_s = Type\ II\ Error\ (Matrix\ Confusion)$
$P_{ave} = Utilidad\ promedio\ por\ cliente$
$M = Presupuesto\ de\ marketing$
$G = número\ de\ clientes\ no\ desertores$

La primera parte de la función de costo, es decir Be_sP_{ave}, es la pérdida causada por el segundo tipo de error.

La segunda parte de la función , es decir $\frac{M}{Ge_f+B(1-e_s)}Ge_f$, es la pérdida causada por el primer tipo de error, la parte izquierda denota el costo promedio de marketing por la predicción de cada cliente. Ge_f, es el número de no desertores que no fueron identificados.
Los costos causados por el Type II Error son considerados 5% - 20% más altos que los Type I Error (Lee, Cjiu, Chou, & Lu, 2006).

En el siguiente capítulo V se analizará y discutirá los algoritmos aplicados con Árboles de Decision y Adaboost para predecir la deserción de clientes, donde se evaluará el rendimiento de los diferentes modelos en base a los indicadores que proporciona la Matriz de Confusión y también considerando el factor económico en la función de pérdida.

CAPÍTULO V

5. ANÁLISIS Y DISCUSIÓN DE RESULTADOS

5.1 CONSTRUCCIÓN DEL MODELO DE DESERCIÓN

Para la construcción del modelo se utilizó una base de 5,308 clientes, de los cuales 1,732 son clientes desertores como se mencionó en el Capítulo II.
Mediante el empleo de muestreo aleatorio simple se seleccionó dos conjuntos de datos, con el objetivo de desarrollar el modelo con el conjunto de datos de entrenamiento que representó el 70% y validarlo con la segunda que corresponde a conjunto de datos de prueba que corresponde el 30% de la población.

En la siguiente tabla se muestra la distribución:

	Desertor	**No Desertor**	**Total**
Población Total	1,732	3,576	5,308
Muestra de Entrenamiento	1,185	2,503	3,688
Muestra de validación	547	1073	1,620

De acuerdo a lo definido en el capítulo II la variable objetivo es la deserción, lo cual se definió de la siguiente forma:

$$Y_i = \begin{cases} 1\ Si\ el\ cliente\ i\ es\ desertor\ durante\ el\ periodo\ de\ evaluación \\ 0\ Si\ el\ cliente\ i\ sigue\ siendo\ cliente\ durante\ el\ periodo\ de\ evaluación \end{cases}$$

Una vez definido el conjunto de datos con el cual se construirá y validará el modelo, en este capítulo nos centraremos en presentar varios resultados obtenidos de los 5 modelos que se mencionaron en el capítulo III de los cuáles analizaremos el rendimiento y precisión de cada uno y mediante los indicadores de Matriz de Confusión y Función de Pérdida se discutirá cuál es el mejor modelo.

5.2 RENDIMIENTO DE PREDICCIÓN DEL MODELO

5.2.1 MODELO DE DESERCIÓN CON ÁRBOLES DE DECISIÓN

En el capítulo IV se mencionó el Error Tipo I, Error Tipo II y Promedio del Error los cuáles son generalmente utilizados para medir la calidad y precisión de un modelo. En este estudio se incorporará un nuevo indicador que muestre el rendimiento del modelo, el cual corresponde a la Función de Costo (Pérdida)

En la tabla 5.2.1 muestra el desempeño de los 5 modelos aplicados, mediante el conjunto de datos de prueba, los cuales fueron realizados en el Software R (en el Anexo I se muestra paquetes, librerías y códigos empleados para la ejecución del modelo).

Para el cálculo de la función de pérdida se consideraron los siguientes valores:

$B = 547\ que\ es\ el\ número\ de\ desertores$
$P_{ave} = \$90\ que\ es\ la\ Utilidad\ promedio\ por\ cliente$
$M = \$32{,}000\ Presupuesto\ de\ marketing\ que\ es\ (\$20) * 1{,}620\ clientes$
$G = 1{,}073\ número\ de\ clientes\ no\ desertores$

Tabla 5.2.1 Rendimiento de Predicción mediante Árboles de Decisión

	Modelo 1	**Modelo 2**	**Modelo 3**	**Modelo 4**	**Modelo 5**
e_f = Error Tipo I %	6.80%	7.36%	0.84%	1.12%	0.00%
e_s =Error Tipo 2 %	71.48%	71.48%	52.65%	45.52%	18.65%
Error Promedio %	28.64%	29.01%	18.33%	16.11%	6.30%
Loss= Costo	$ 45,518.38	$ 46,081.91	$27,008.06	$ 23,664.19	$ 9,180.00
ACC Precisión (Accuracy) %	71.36%	70.99%	81.67%	83.89%	93.70%

El *Modelo I* es construido con las 6 variables que dio como predictoras aplicando el Estimador de Kaplan Meier que son: canal radicador, edad, actividad económica, segmento, transacciones facturadas y si solo utiliza la tarjeta para pagos recurrentes, en la Tabla 5.2.1 muestra que el modelo I el primer tipo de error es 6.80% y el segundo tipo de Error es 71.48%, por otro lado, se puede observar que los costos por una mala clasificación de los clientes dan como resultado $45,518.38.

En el *Modelo 2* las variables empleadas para la construcción del modelo son las variables que aplicando Regresión de Cox dieron como predictoras, estas son: ubicación geográfica, canal radicador, edad, estado civil, actividad económica, segmento y transacciones masivas, como se muestra en cuadro anterior el promedio del error del Modelo II es 29.01%, es el modelo que presentó mayor error en la predicción de la deserción.

Por otro lado, en el *Modelo 3* es construido en base a variables solo de tipo transaccional, lo cual resulta ser un buen modelo, en comparación con los Modelos I y II el error promedio es 18.33% y los costos por una mala clasificación en la predicción es $27,008.06 son menores.

En el *Modelo 4* es una combinación de variables del Modelo I, II y III, el % del error promedio dio como resultado 16.11 lo cual es menor que los Modelos mencionados anteriormente, siendo un buen modelo dada la precisión fue del 83.89 % y los costos de una mala clasificación son menos altos incurriendo en costos de $23,664. El siguiente cuadro muestra la Matriz de Confusión del Modelo 4.

Matriz de Confusión Modelo 4		**Predicción**	
		No Desertor *(Negativo)*	**Desertor** *(Positivo)*
Valor Real	**No Desertor** *(Negativo)*	1061 *Precisión Negativa*	12 *Falso Positivo*
	Desertor *(Positivo)*	249 *Falso Negativo*	298 *Precision Positiva*

A continuación, se mencionan las principales reglas extraídas del árbol de decisión del modelo 4, los cuales indican que son desertores si cumplen tal regla.

Si $X_2 \leq 34.5$, $X_{21} \leq 0.42$, $X_{26}=1$, X_{15}=Quito y Cuenca, $X_{24} \leq 21.5$, $X_4 \leq 67.5$ es desertor

Si $X_{21}<1.07$, $X_2 \leq 21.5$, $X_{20} \leq 334.14$, X_{15}=Quito y Cuenca, $X_{26}=1$ es desertor

Si $X_2 \leq 21.5$, $X_{26}=1$, $X_{20}<271.3$, $X_{19}<3{,}255.59$, es desertor

Acorde a las reglas que se mencionaron, se puede observar cuáles son las variables más importantes para la predicción de la deserción de clientes, corresponden a: ubicación geográfica (X_1), antigüedad del cliente (X_2), edad (X_4), monto promedio de consumo (X_{19}), monto promedio consumo rotativo (X_{20}), monto promedio sobregiros (X_{21}), si solo utiliza la T/C transacciones recurrentes (X_{26}) y total transacciones facturadas (X_{24}).

Finalmente, el modelo que mejor poder de predicción resultó mediante la aplicación de árboles de decisión fue el modelo 5, dado que el promedio del error es menor que los demás modelos siendo 6.30%, incluso fue el modelo que dio los costos más bajos, obteniendo un valor de $9,180 y la precisión del modelo fue del 93.70%. Para la construcción del modelo 5 solo se utilizaron las variables demográficas y externas. A continuación, se muestra la matriz de confusión.

Matriz de Confusión Modelo 5		**Predicción**	
		No Desertor *(Negativo)*	**Desertor** *(Positivo)*
Valor Real	**No Desertor** *(Negativo)*	1073 *Precisión Negativa*	0 *Falso Positivo*
	Desertor *(Positivo)*	102 *Falso Negativo*	445 *Precision Positiva*

A continuación, se mencionan las principales reglas extraídas del árbol de decisión del modelo 5, los cuales indican que son desertores si cumplen tal regla.

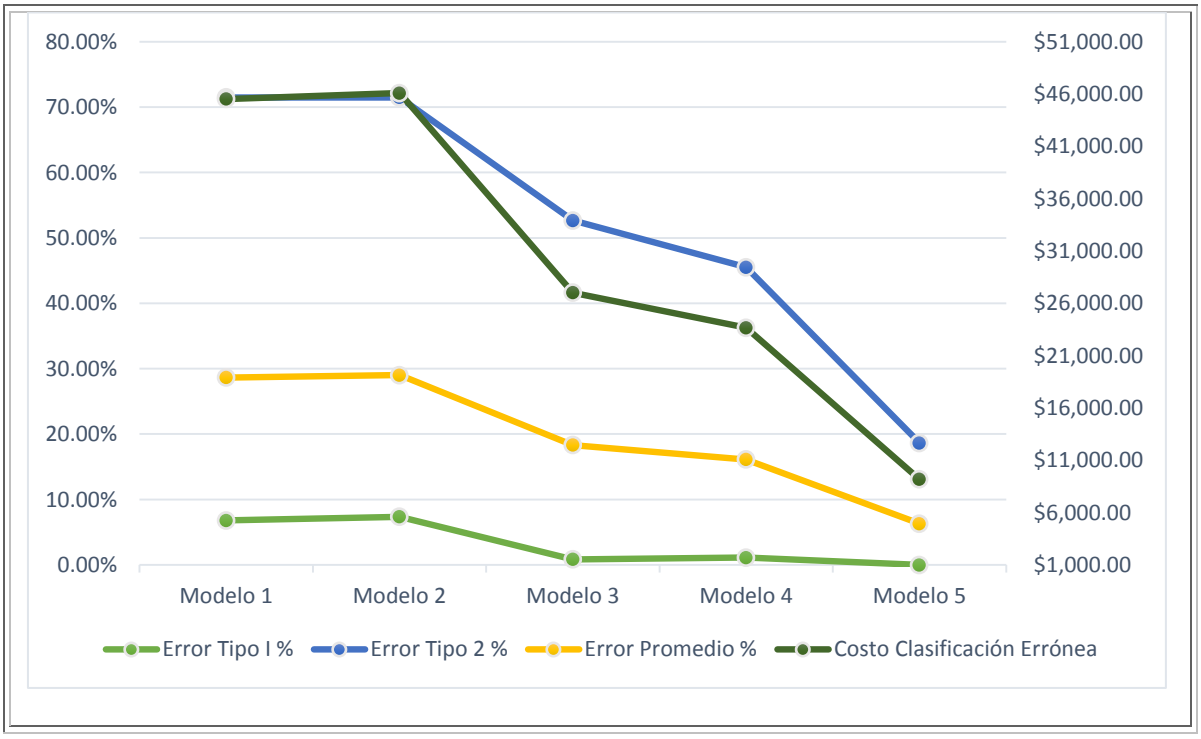
Si $X_2 \leq 34.5$, $X_{26}=1$, $X_4 \leq 67.5$, X_1=Quito y Cuenca, X_7=Empleado Público, Privado y actividades comercio mayor y menor, es desertor.

Si $X_2 \leq 21.5$, X_1=Quito y Cuenca, X_{17}=A, $X_{13}<934.5$, X_7=Empleado Público, Privado y actividades comercio mayor y menor, es desertor

Acorde a las reglas que se mencionaron, se puede observar cuáles son las variables más importantes para la predicción de la deserción de clientes en el modelo 5, corresponden a: ubicación geográfica (X_1), antigüedad del cliente (X_2), edad (X_4), actividad económica (X_7), ingreso neto (X_{13}), calificación central de riesgo (X_{17}) y si solo utiliza la T/C transacciones recurrentes (X_{26}).

La Figura 5.2.1 muestra el rendimiento de los 5 modelos, donde se puede observar que los modelos 1 y 2 no fueron muy significativos, dado los altos costos además de los porcentajes altos de error de una mala clasificación.

Figura 5.2.1 Desempeño de Predicción Árboles de Decisión



5.2.2 MODELO DE DESERCIÓN CON ADABOOST

En esta sección se construirá los 5 modelos empleando el algoritmo de Adaboost lo cual el objetivo es producir una mejor predicción que lo aplicado mediante árboles de decisión.

En la Tabla 5.2.2 muestra la calidad, precisión, desempeño y función de costo de los 5 modelos aplicados con Adaboost. Los modelos fueron realizados mediante la librería adabag en el Software R Versión 3.2.2 (en el Anexo IV se detalla los códigos utilizados para la construcción del modelo).

Para el cálculo de la función de pérdida se consideraron los mismos valores que se utilizaron para Árboles de Decisión.

Tabla 5.2.2 Rendimiento de Predicción mediante Árboles de Decisión

	Modelo 1	Modelo 2	Modelo 3	Modelo 4	Modelo 5
e_f = Error Tipo I %	7.83%	7.64%	1.03%	0.19%	0.47%
e_s =Error Tipo 2 %	66.73%	72.39%	47.17%	10.79%	43.33%
Error Promedio %	27.72%	29.51%	16.60%	3.77%	14.94%
Loss= Costo	$ 43,081.58	$ 47,042.58	$ 24,408.00	$ 5,442.24	$ 21,844.29
ACC Precisión (Accuracy) %	72.28%	70.49%	83.40%	96.23%	85.06%

Se puede observar en la Tabla 5.2.2, que el primer tipo de error del *Modelo 1* es 7.83% y el segundo tipo de Error es 66.73%, por otro lado, se puede observar que los costos por una mala clasificación de los clientes dan como resultado $43,081.58.

En el *Modelo 2*, el Error Tipo I y Error Tipo II son 7.64% y 72.39% respectivamente, el promedio del error es 29.51% y el costo por una mala clasificación dio $47,042.58.

Continuando con el *Modelo 3* el error promedio es 16.60% y los costos por una mala clasificación en la predicción son $24,408, con una precisión del 83.40%.

Generalmente hablando, cuando el error es menor que el 25%, el modelo es lo suficientemente poderoso para ser utilizado en una aplicación comercial [10]. El siguiente cuadro muestra la Matriz de Confusión del Modelo 3.

Matriz de Confusión Modelo 3		Predicción	
		No Desertor *(Negativo)*	**Desertor** *(Positivo)*
Valor Real	**No Desertor** *(Negativo)*	1062 *Precisión Negativa*	11 *Falso Positivo*
	Desertor *(Positivo)*	258 *Falso Negativo*	289 *Precision Positiva*

Por otro lado, en el *Modelo 4*, el % del error promedio dio como resultado 3.77 lo cual es menor que los demás modelos, siendo el mejor modelo, hablando en términos de Error Promedio, Precisión y Costos de Mala Clasificación, la precisión del modelo fue del 96.23% y los costos de una mala clasificación son $5,442.24. A continuación se muestra la Matriz de Confusión del Modelo 4.

Matriz de Confusión Modelo 4		Predicción	
		No Desertor *(Negativo)*	**Desertor** *(Positivo)*
Valor Real	**No Desertor** *(Negativo)*	1071 *Precisión Negativa*	2 *Falso Positivo*
	Desertor *(Positivo)*	59 *Falso Negativo*	488 *Precision Positiva*

A continuación, se menciona las variables que el modelo 4 mediante la aplicación del algoritmo de Adaboost consideró como importantes para predecir la deserción de clientes.

Var	Descripción	Importancia
X2	Antigüedad(meses)	67.38
X21	Monto Promedio consumo diferido	21.71
X24	Total transacciones facturadas	5.36
X7	Actividad económica (categórica)	1.97
X1	Ubicación geográfica (categórica)	1.08
X22	Cantidad veces sobregirado	0.89
X13	Ingreso neto	0.74
X25	Total transacciones masivas	0.36
X6	Estado civil (categórica)	0.29
X16	Monto deuda sistema financiero	0.21
X3	Canal radicador (categórica)	-
X4	Edad	-
X5	Género (categórica)	-
X8	Segmento (categórica)	-
X9	Dummy, si tiene cuenta corriente es 1 sino 0	-
X10	Dummy, si tiene cuenta de ahorro es 1 sino 0	-
X11	Dummy, si tiene depósito a plazo es 1 sino 0	-
X12	Dummy, si tiene préstamo es 1 sino 0	-
X15	Categoría tarjeta Visa (categórica)	-
X17	Calificación central riesgo	-
X18	Total monto consumo anual	-
X19	Monto Promedio consumo	-
X20	Monto Promedio consumo rotativo	-
X23	Monto Promedio sobregiros	-
X26	Dummy, si solo utiliza la T/C pagos recurrentes es 1 sino 0	-

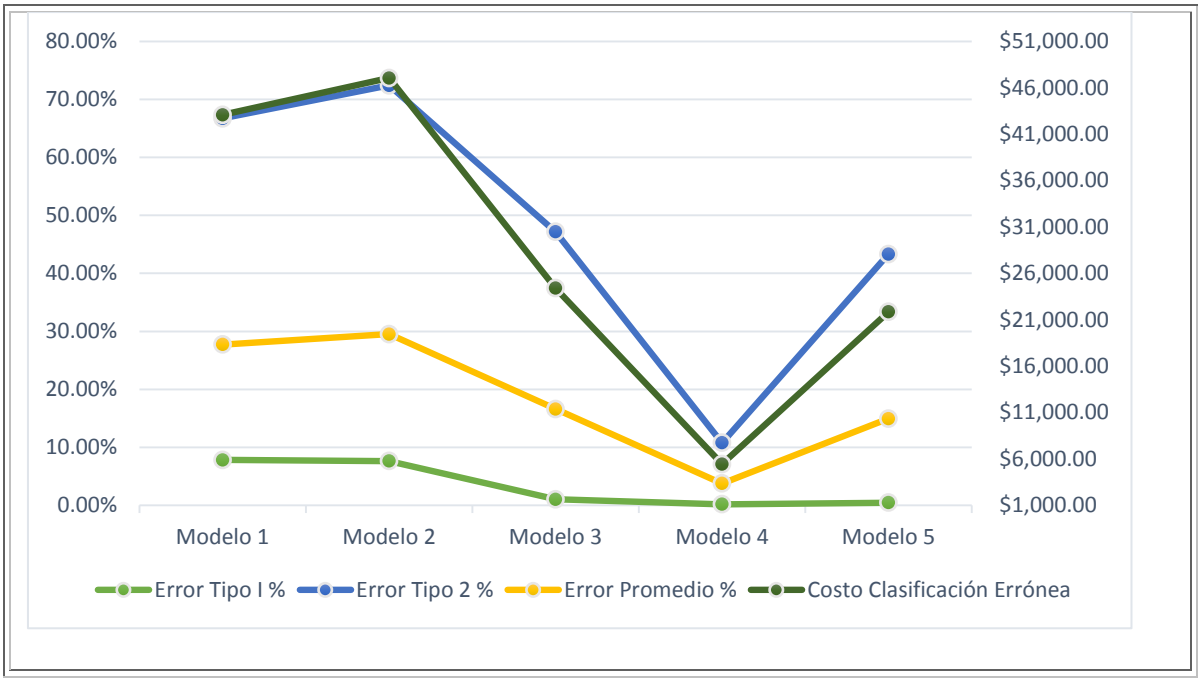
Se puede observar que las variables más importantes que considera el Modelo 4 son antigüedad del cliente, monto promedio consumos diferidos, total transacciones facturadas, actividad económica, ubicación geográfica, cantidad de veces sobregirado, ingresos netos, total transacciones masivas, estado civil y monto deuda sistema financiero.

Finalmente, el modelo 5 el promedio del error es mayor que el modelo 4 siendo 14.94%, el costo de pérdida es $21,844 y la precisión del modelo fue del 85.06%. En este modelo el algoritmo de Adaboost la única que consideró importante fue Antigüedad del cliente.

Matriz de Confusión Modelo 5		Predicción	
		No Desertor *(Negativo)*	**Desertor** *(Positivo)*
Valor Real	**No Desertor** *(Negativo)*	1073 *Precisión Negativa*	0 *Falso Positivo*
	Desertor *(Positivo)*	102 *Falso Negativo*	445 *Precision Positiva*

La Figura 5.2.2 se muestra el rendimiento de los 5 modelos, donde se puede observar que el modelo 4 fue el mejor modelo que dio Adaboost.

Figura 5.2.2 Desempeño de Predicción - Adaboost



5.3 MODELO Y DISCUSIÓN

En la sección anterior se mostró el rendimiento de los 5 modelos aplicando Árboles de Decisión y Adaboost, el modelo 4 resultó mejor desde la perspectiva en cuanto a precisión y costos por mala clasificación lo cual los resultados del modelo son consistentes.

Aplicando Árboles de Decisión se determinó que el mejor modelo es el 5 con una precisión del 93.70%, costo por mala clasificación de $9,180 y un error promedio del 6.30%.

De todos los modelos del Adaboost, el modelo 4 tuvo un promedio de error de predicción del 3.77% que fue el error más bajo de Adaboost y que son datos de prueba y no de entrenamiento, lo cual este resulta ser el mejor modelo y objetivamente hablando refleja las variables predictoras que ayudará al Banco a detectar los clientes potencialmente a desertar.

CONCLUSIONES

Las entidades financieras requieren contar con herramientas que le permitan predecir la deserción de clientes y así decidir sobre a quiénes destinar sus esfuerzos de retención de clientes. En esta tesis ha sido posible estudiar las variables predictoras que permitan determinar la deserción del cliente.

En el capítulo 3 se hizo un análisis de supervivencia mediante los estimadores de Kaplan – Meier y Regresión de Cox, donde se pudo determinar las variables explicativas que determinan las posibles causas de la deserción de clientes. Sin embargo, el % error promedio fue de 45.19% en los datos de prueba lo cual resulta ser muy alto. Por consiguiente, se propone combinar árboles de decisión mediante Adaboost, de tal forma que incremente el poder de predicción de deserción.

En el capítulo 5 se aplica el algoritmo de Adaboost, donde se evalúan 5 modelos, donde el modelo 1 es construido en base a las variables predictoras que dio como resultado el estimador de Kaplan – Meier, el modelo 2 se construyó en base a las variables predictoras aplicando Regresión de Cox, por otro lado el modelo 3 se construyó con variables de comportamiento transaccional, externas e ingresos netos, así mismo el modelo 4 contiene todas las variables que comprenden las de tipo transaccional, externas y demográficas y finalmente el modelo 5 consiste solo en las variables de tipo demográficas y externas.

La elección de aplicar Adaboost favorece la precisión y nos da como mejor modelo el 4, lo cual es más robusto en cuanto a mayor precisión en la predicción, porcentaje de error más bajo y menor costos por errores de mala clasificación. En este modelo se utilizaron el mayor número de variables lo cual permitió predecir de forma más completa el comportamiento de los clientes al momento de desertar. Esto quiere decir que el proceso de selección de atributos es fundamental en la construcción de modelos, sobretodo en la precisión del modelo y menor porcentaje de error.

Así mismo este modelo proporciona reglas que son acordes al negocio del banco y fáciles de entender, de tal forma que resulta una guía para los directivos del Banco en las estrategias de marketing para la retención y fidelización del cliente.

La aplicación de una metodología ha sido exitosa. Primero porque ha sido posible construir un modelo mediante Adaboost de tal forma que nos permite caracterizar las variables predictoras en la deserción. Segundo porque fue posible estimar modelos propuestos sobre una base de datos que contiene la transaccionalidad de los clientes, comprobando la aplicabilidad de estas metodologías. Finalmente, porque los resultados obtenidos con los modelos propuestos son competitivos respecto al enfoque tradicional que realizan en la actualidad.

RECOMENDACIONES

Con el propósito de facilitar el análisis y mejorar la calidad de predicción de deserción de clientes, es necesario realizar las siguientes recomendaciones:

- Recopilar y almacenar la información relacionada con transacciones y disponer de información actualizada en cuanto a demografía del cliente, esto con el propósito de generar un modelo de deserción de clientes. Cabe recalcar que para esto se requiere una estructura de base de datos adecuada para evitar problemas de extracción de información.

- Se recomienda, puesto que el monitoreo se debe realizar periódicamente, realizar una implementación automática de reportes de monitoreo sugeridos, de tal forma que se pueda determinar cuáles son los clientes propensos a desertar y así realizar esfuerzos necesarios para su retención.

- Para la implementación automática de reportes de monitoreo, es necesario implementar un modelo que permita predecir la deserción de clientes mediante aplicación de Adaboost, de tal forma que el Banco se pueda anticipar lo suficiente para no sufrir una deserción de clientes y así mismo mejorar la efectividad de campañas de fidelización.

- Promover el uso del software libre R, en la construcción de modelos y generación de nuevas metodologías en la institución financiera.

- Estudiar las posibles ventajas que implicaría la utilización de modelos de deserción en el fortalecimiento de las relaciones entre la entidad y el cliente.

BIBLIOGRAFÍA

[1]: De la página web: www.telegrafo.com.ec/economia/item/tarjetahabientes-deben-en-promedio-85732-al-mes.html.

[2]: European Journal of Operational Research, Volume 197, Issue 1, Pearson Educación S.A., 16 Agosto 2009, Páginas 402–411,**Modeling churn using customer lifetime value,** de Nicolas Glady, Bart Baesens y Christophe Croux.

[3]: Del libro: Dirección de Marketing, Edición del Milenio, de Kotler Philip, Cámara Dionicio, Grande Ildefonso y Cruz Ignacio, Pearson Educación S.A., 2000, Págs. 52 y 55. **Libro de Consulta: Fundamentos de Marketing, 13a. Edición,** de Stanton William, Etzel Michael y Walker Bruce, Mc Graw Hill, 2004.

[4] Customer churns prediction using improved balanced random forests, Yaya Xie, Xiu Li, E.W.T. Ngai, Weiyun Ying.

[5] Credit card churn forecasting by logistic regression and decision tree, Guangli Nie, Wei Rowe, Lingling Zhang, Yingjie Tian, Yong Shi

[6] Fuente: http://en.wikipedia.org/wiki/Kaplan%E2%80%93Meier_estimator

[7] Fuente: http://daynebatten.com/2015/02/customer-churn-cox-regression/

[8] Fuente: http://en.wikipedia.org/wiki/Proportional_hazards_model

[9] Fuente: www.r-statistics.com/2013/07/creating-good-looking-survival-curves-the-ggsurv-function/

[10] Credit card churn forecasting by logistic regression and decision tree, Guangli Nie, Wei Rowe, Lingling Zhang, Yingjie Tian, Yong Shi

[11] Customer churns prediction using improved balanced random forests, Yaya Xie, Xiu Li, E.W.T. Ngai, Weiyun Ying.

[12] Model of Customer Churn Prediction on Support Vector Machine, XIA Guo, JIN Wei-dong.

[13] Customer churn prediction by hybrid neural networks, Chih-Fong Tsai, Yu-Hsin Lu

[14] Del libro: Understanding Machine Learning from Theory to algorithms, First Edition , Shai Shalev – Shwartz and Shai Ben - David, Cambridge University Press, 2014, Pages 101 to 112.

[15] De la página web:
http://www.cse.buffalo.edu/~jcorso/t/CSE555/files/lecture_boosting.pdf
[16] De la página web: https://en.wikipedia.org/wiki/Confusion_matrix
[17] De la página web:
http://oldemarrodriguez.com/yahoo_site_admin/assets/docs/Presentaci%C3%B3n_-_KNN.20085205.pdf

ANEXO I

VARIABLES UTILIZADAS

VARIABLES DEMOGRÁFICAS

- ✓ **Ubicación geográfica:** Ubicación geográfica del tarjetahabiente que es utilizada por la institución financiera.
- ✓ **antiguedad_cliente:** antigüedad del cliente en meses.
- ✓ **canal_radicador:** indica si el cliente obtuvo la tarjeta por Preventa o Mercado Natural, se entiende por Mercado Natural cuando el cliente se acerca a la entidad a solicitar la tarjeta de crédito, mientras que Preventa el cliente obtiene la tarjeta de crédito mediante campaña realizada con el Banco y otorga la cliente.
- ✓ **edad:** Edad del cliente
- ✓ **género:** Género del cliente
- ✓ **estado_civil:** Estado civil
- ✓ **actividad_economica:** Actividad económica a la que se dedica
- ✓ **ingreso_neto:** corresponde a los ingresos del Tarjetahabiente
- ✓ **ingreso_mensual:** corresponde a los ingresos del Tarjetahabiente + cónyuge
- ✓ **Segmento:** clasificación del segmento del Tarjeta habiente de acuerdo a sus ingresos netos.
    - o **Básico:** Ingresos menor a $500
    - o **Estándar:** Ingresos de $500 a $1,499
    - o **Medio:** Ingresos de $1,500 a $2,999
    - o **Medio Alto:** Ingresos de $3,000 a $4,999
    - o **Alto:** Ingresos superior o igual a $5,000

- ✓ **ind_ctacte:** indica si el cliente tiene o no una cuenta corriente en el banco
- ✓ **ind_ctaaho:** indica si el cliente tiene o no una cuenta de ahorro en el banco
- ✓ **ind_deppla:** indica si el cliente tiene o no una cuenta de depósito a plazo en el banco
- ✓ **monto_total_cartera:** monto de deuda total en el banco

VARIABLES EXTERNAS

- ✓ **calificación:** calificación correspondiente a la Central de Riesgo
- ✓ **monto_riesgo_sistema:** monto del riesgo total en el sistema (es decir el monto total de deuda que el cliente registra en todas las entidades financieras)

VARIABLES DE COMPORTAMIENTO

- ✓ **descripcion_tarjeta:** descripción de la tarjeta de crédito, si es Platinum, Oro, etc.
    - Descripción de Tarjeta se refiere a las diferentes categorías que la entidad tiene en su producto Tarjeta de Crédito en la Marca Visa. Cada categoría tiene diferentes beneficios y target. Cabe recalcar que las tarjetas en la categoría de afinidad o marcas compartidas son alianzas estratégicas que realiza el Banco con establecimientos, como por ejemplo Visa Mi Comisariato, Visa De Prati, etc. Mientras que las corporativas es la tarjeta exclusivamente diseñada para respaldar y distinguir a los ejecutivos de su empresa.
- ✓ **saldo_consumido:** saldo total consumido (rotativo + diferido) en el tiempo t
- ✓ **saldo_rotativo:** saldo total consumido en rotativo en t

- ✓ **saldo_diferido:** saldo total consumido en diferido en t
- ✓ **cupo:** cupo asignado en la tarjeta de crédito en t
- ✓ **saldo_disponible:** saldo disponible en la T/C en t
- ✓ **ind_consumo_local:** indica si realizó consumos locales en t
- ✓ **ind_consumo_exterior:** indica si realizó consumos en el exterior en t
- ✓ **cantidad_trx_carga_masiva:** número de débitos recurrentes realizados en la T/C, ej. (Direct Tv, TV Cable, Luz, etc.) en t
- ✓ **monto_carga_masiva:** monto total de débitos recurrentes realizados en T/C, ejm (Direct Tv, TV Cable, Luz, etc.) en t
- ✓ **cantidad_trx_pagos:** # de pagos realizados en t a la T/C
- ✓ **monto_pagos:** total de pagos realizados en t a la T/C
- ✓ **cantidad_trx_fact:** cantidad de transacciones realizas en t
- ✓ **monto_facturado:** monto en $ facturado del mes en t

ANEXO II

MATRIZ CORRELACIÓN

Var	Descripción	Mínimo	Máximo
X_1	Ubicación geográfica (categórica)	1	6
X_2	Antigüedad(meses)	12	281
X_3	Canal radicador (categórica)	0	1
X_4	Edad	18	113
X_5	Género (categórica)	1	2
X_6	Estado civil (categórica)	1	5
X_7	Actividad económica (categórica)	1	6
X_8	Segmento (categórica)	1	5
X_9	Dummy, si tiene cuenta corriente es 1 sino 0	0	1
X_{10}	Dummy, si tiene cuenta de ahorro es 1 sino 0	0	1
X_{11}	Dummy, si tiene depósito a plazo es 1 sino 0	0	1
X_{12}	Dummy, si tiene préstamo es 1 sino 0	0	1
X_{13}	Ingreso neto	0	150,000
X_{14}	Ingreso mensual	0	204,000
X_{15}	Categoría tarjeta Visa (categórica)	1	7
X_{16}	Monto deuda sistema financiero	2.01	8,114,040
X_{17}	Calificación central riesgo	1	5
X_{18}	Total monto consumo anual	0	244,378
X_{19}	Monto Promedio consumo	0	20,365
X_{20}	Monto Promedio consumo rotativo	0	19,231
X_{21}	Monto Promedio consumo diferido	0	3,438
X_{22}	Cantidad veces sobregirado	0	12
X_{23}	Monto Promedio sobregiros	0	1,374
X_{24}	Total transacciones facturadas	0	647
X_{25}	Total transacciones masivas	0	82
X_{26}	Dummy, si solo utiliza la T/C pagos recurrentes es 1 sino 0	0	1

Variables de control		X1	X2	X3	X4	X5	X6	X7	X8	X9	X10	X11	X12	X13	X14	X15	X16	X17	X18	X19	X20	X21	X22	X23	X24	X25	X26
CHURNER	X1	1.000	-.021	-,046**	.010	-.018	,037**	,093**	,053**	.021	-,042**	.019	,061**	,057**	,058**	.006	,047**	-.005	.001	.001	.000	.008	-.004	-.009	-.006	-.004	-,062**
	X2	-.021	1.000	-,375**	,185**	.011	.014	-,053**	.020	,034*	-.024	.001	.002	.027	,066**	-.010	.018	-.003	,132**	,134**	,131**	,047**	.004	.019	.024	-,031*	,040**
	X3	-,046**	-,375**	1.000	,040**	-,032*	.022	-.016	-,086**	-,167**	-,149**	-,120**	-,096**	-,055**	-,057**	-,030*	-.007	-,032*	-,029*	-,029*	-,030*	.004	-,144**	-.017	-,111**	,104**	,074**
	X4	.010	,185**	,040**	1.000	-,033*	,329**	,117**	,104**	.022	-,072**	,069**	,037**	,079**	,068**	,099**	,030*	-.026	,091**	,092**	,086**	,067**	-,122**	.014	-.023	,028*	,077**
	X5	-.018	.011	-,032*	-,033*	1.000	.017	.008	-,125**	.020	,087**	.007	-.014	-,099**	-,046**	-.025	-.021	.006	.004	.004	.001	.023	.002	-.012	-.024	-,087**	,027*
	X6	,037**	.014	.022	,329**	.017	1.000	,091**	,091**	,048**	-,032*	,052**	,027*	,051**	,050**	.018	,050**	.023	,053**	,053**	,051**	,028*	-.015	.027	.012	.022	-.018
	X7	,093**	-,053**	-.016	,117**	.008	,091**	1.000	,144**	,119**	-.009	,061**	,064**	,128**	,101**	,030*	.021	.024	.017	.019	.009	,087**	-,064**	-.006	-.002	-.007	-.020
	X8	,053**	.020	-,086**	,104**	-,125**	,091**	,144**	1.000	,205**	-.007	,089**	,162**	,533**	,441**	,145**	,109**	-.013	,132**	,133**	,132**	,030*	-,038**	,093**	,141**	,067**	-,034*
	X9	.021	,034*	-,167**	.022	.020	,048**	,119**	,205**	1.000	,116**	,137**	,305**	,154**	,146**	,068**	,030*	-,035*	,037**	,035*	,034*	.010	-.019	.023	,091**	-,062**	-.007
	X10	-,042**	-.024	-,149**	-,072**	,087**	-,032*	-.009	-.007	,116**	1.000	,148**	,127**	-.011	-.004	.001	-.025	.016	-,086**	-,088**	-,087**	-.026	.009	-,033*	.015	-,147**	,042**
	X11	.019	.001	-,120**	,069**	.007	,052**	,061**	,089**	,137**	,148**	1.000	,133**	,095**	,084**	.017	.008	-.026	-.011	-.011	-.011	.000	-.020	.019	,064**	-,027*	.002
	X12	,061**	.002	-,096**	,037**	-.014	,027*	,064**	,162**	,305**	,127**	,133**	1.000	,150**	,136**	.012	,028*	-.014	,027*	.026	,027*	-.008	.010	-.002	,028*	-,055**	-.006
	X13	,057**	.027	-,055**	,079**	-,099**	,051**	,128**	,533**	,154**	-.011	,095**	,150**	1.000	,779**	,130**	,139**	-.010	,071**	,070**	,070**	.015	-,041**	,036**	,075**	,036**	.016
	X14	,058**	,066**	-,057**	,068**	-,046**	,050**	,101**	,441**	,146**	-.004	,084**	,136**	,779**	1.000	,126**	,144**	-.012	,077**	,076**	,076**	.013	-,034*	,032*	,081**	,043**	.011
	X15	.006	-.010	-,030*	,099**	-.025	.018	,030*	,145**	,068**	.001	.017	.012	,130**	,126**	1.000	,069**	-,040**	,146**	,149**	,150**	.020	-,084**	,078**	,080**	,028*	,041**
	X16	,047**	.018	-.007	,030*	-.021	,050**	.021	,109**	,030*	-.025	.008	,028*	,139**	,144**	,069**	1.000	-.011	,097**	,096**	,097**	.011	-.016	.018	,081**	.020	.002
	X17	-.005	-.003	-,032*	-.026	.006	.023	.024	-.013	-,035*	.016	-.026	-.014	-.010	-.012	-,040**	-.011	1.000	,064**	,067**	,068**	.005	,173**	.019	,036**	-,033*	-,064**
	X18	.001	,132**	-,029*	,091**	.004	,053**	.017	,132**	,037**	-,086**	-.011	,027*	,071**	,077**	,146**	,097**	,064**	1.000	,996**	,990**	,237**	,317**	,348**	,511**	,135**	-,266**
	X19	.001	,134**	-,029*	,092**	.004	,053**	.019	,133**	,035*	-,088**	-.011	.026	,070**	,076**	,149**	,096**	,067**	,996**	1.000	,993**	,240**	,314**	,350**	,509**	,132**	-,269**
	X20	.000	,131**	-,030*	,086**	.001	,051**	.009	,132**	,034*	-,087**	-.011	,027*	,070**	,076**	,150**	,097**	,068**	,990**	,993**	1.000	,127**	,325**	,349**	,500**	,130**	-,263**
	X21	.008	,047**	.004	,067**	.023	,028*	,087**	,030*	.010	-.026	.000	-.008	.015	.013	.020	.011	.005	,237**	,240**	,127**	1.000	-,034*	,072**	,174**	,043**	-,097**
	X22	-.004	.004	-,144**	-,122**	.002	-.015	-,064**	-,038**	-.019	.009	-.020	.010	-,041**	-,034*	-,084**	-.016	,173**	,317**	,314**	,325**	-,034*	1.000	,129**	,170**	,036**	-,197**
	X23	-.009	.019	-.017	.014	-.012	.027	-.006	,093**	.023	-,033*	.019	-.002	,036**	,032*	,078**	.018	.019	,348**	,350**	,349**	,072**	,129**	1.000	,294**	,050**	-,102**
	X24	-.006	.024	-,111**	-.023	-.024	.012	-.002	,141**	,091**	.015	,064**	,028*	,075**	,081**	,080**	,081**	,036**	,511**	,509**	,500**	,174**	,170**	,294**	1.000	,288**	-,336**
	X25	-.004	-,031*	,104**	,028*	-,087**	.022	-.007	,067**	-,062**	-,147**	-,027*	-,055**	,036**	,043**	,028*	.020	-,033*	,135**	,132**	,130**	,043**	,036**	,050**	,288**	1.000	-,263**
	X26	-,062**	,040**	,074**	,077**	,027*	-.018	-.020	-,034*	-.007	,042**	.002	-.006	.016	.011	,041**	.002	-,064**	-,266**	-,269**	-,263**	-,097**	-,197**	-,102**	-,336**	-,263**	1.000

**. La correlación es significativa en el nivel 0,01

*. La correlación es significativa en el nivel 0,05

ANEXO III

Librerías R Versión 3.2.2

rpart

rpart.plot

caret

grid

mvtnorm

stats4

modeltools

base

zoo

sandwich

strucchange

gee

geepack

Matrix

CompQuadForm

e1071

gskat

survival

mvtnorm

coin

multcomp

TH.data

Party

Código R

Modelo 2

Cargar Base:

```
modelo2 <- read.csv('C:/Users /BaseModelo.csv')
```

Visualizar data.frame:

```
str(modelo2)
```

Dividir conjunto de datos por muestreo aleatorio simple para obtener el conjunto de datos de entrenamiento y validación,

```
set.seed(2)
ind=sample(2,nrow(modelo2),replace=TRUE,prob=c(0.7,0.3))
trainset2=modelo2[ind==1,]
testset2=modelo2[ind==2,]
```

Construcción árbol de clasificación:

```
modelo2.rp=rpart(CHURNER~.,data=trainset2)
```

Devuelve el nodo:

```
modelo2.rp
```

Visualizar los parámetros delo modelo, y analizar mediante el valor de CP analizar si es necesario podar el árbol

```
printcp(modelo2.rp)
```

Figura del valor CP del modelo 2

```
plotcp(modelo2.rp)
```

Muestra el resumen del modelo construido: variables de importancia y las reglas del árbol de decisión

```
summary(modelo2.rp)
```

Figura del árbol, donde se visualiza mejor las reglas

```
tree <- rpart(CHURNER ~ ., data=trainset2)
old.par <- par(mfrow=c(2,2))
# put 4 figures on one page
prp(tree, main="default prp\n(type = 0, extra = 0)")
prp(tree, main="type = 4, extra = 6", type=4, extra=6, faclen=0)
# faclen=0 to print full factor names
cols <- ifelse(tree$frame$yval == 1, "darkred", "green4")
# if survived
prp(tree, main="assorted arguments",
extra=106, # display prob of survival and percent of obs
nn=TRUE, # display the node numbers
```

```
fallen.leaves=TRUE, # put the leaves on the bottom of the page
branch=.5, # change angle of branch lines
faclen=0, # do not abbreviate factor levels
trace=1, # print the automatically calculated cex
shadow.col="gray", # shadows under the leaves
branch.lty=3, # draw branches using dotted lines
split.cex=1.2, # make the split text larger than the node text
split.prefix="is ", # put "is " before split text
split.suffix="?", # put "?" after split text
col=cols, border.col=cols, # green if survived
split.box.col="lightgray", # lightgray split boxes (default is white)
split.border.col="darkgray", # darkgray border on split boxes
split.round=.5) # round the split box corners a tad
# the old way for comparison
plot(tree, uniform=TRUE, compress=TRUE, branch=.2)
text(tree, use.n=TRUE, cex=.6, xpd=NA) # cex is a guess, depends on your window size
title("plot.rpart for comparison", cex=.6)
par(old.par)
```

Predicción de Desertores

```
predictions2=predict(modelo2.rp, testset2, type="class")
```

Muestra la tabla de Predicción

```
predictions2
```

Predicción Mediante Matriz de Confusión

```
confusionMatrix(table(predictions2,testset2$CHURN))
```

ANEXO IV

Librerías R Versión 3.2.2

rpart

mlbench

caret

lattice

ggplot2

adabag

Código R

Modelo 2

Cargar Base:

```
modelo2  <- read.csv('C:/Users/BaseModelo.csv')
```

Visualizar data.frame:

```
str(modelo2)
```

Dividir conjunto de datos por muestreo aleatorio simple para obtener el conjunto de datos de entrenamiento y validación,

```
set.seed(2)
ind=sample(2,nrow(modelo2 ),replace=TRUE,prob=c(0.7,0.3))
trainset=modelo2 [ind==1,]
testset=modelo2 [ind==2,]
```

Variables de importancia que considera el modelo

```
churn.boost= boosting(CHURNER~.,data=trainset,mfinal=10,coeflearn="Freund",
boos=FALSE, control=rpart.control(maxdepth=3))
```

Predicción de Desertores mediante Matriz de Confusión

```
churn.boost.pred$confusion
churn.boost.pred$error
```