ESCUELA SUPERIOR POLITÉCNICA DEL LITORAL

Facultad de Ciencias Naturales y Matemáticas

Desarrollo de una metodología para detectar la fuga de clientes en un escenario no contractual.

PROYECTO INTEGRADOR

Previo la obtención del Título de:

INGENIERO EN ESTADÍSTICA INFORMÁTICA

Presentado por:

Nicolás Andrés Pazmiño Alarcón

GUAYAQUIL - ECUADOR

Año: 2018

DEDICATORIA

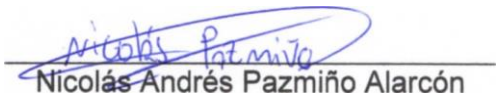
A mis padres y hermanos.

AGRADECIMIENTOS

Agradezco a los profesores de mi facultad por contribuir en mi proceso de formación profesional, especialmente a mi tutor Francisco Vera por su apoyo y a los directivos de la empresa que me dieron la apertura para desarrollar este proyecto.

DECLARACIÓN EXPRESA

“Los derechos de titularidad y explotación, me corresponde conforme al reglamento de propiedad intelectual de la institución; *Nicolás Andrés Pazmiño Alarcón* doy mi consentimiento para que la ESPOL realice la comunicación pública de la obra por cualquier medio con el fin de promover la consulta, difusión y uso público de la producción intelectual"



Nicolás Andrés Pazmiño Alarcón

EVALUADORES

PhD. Sandra García

PROFESOR DE LA MATERIA

PhD. Francisco Vera

PROFESOR TUTOR

RESUMEN

La retención de clientes es un asunto estratégico para las empresas, por lo tanto, estas deben implementar métodos para detectar posibles fugas de sus clientes; existen varios modelos para esto en escenarios contractuales, pero para escenarios no contractuales actualmente se pueden encontrar pocas alternativas debido a que no se observa con exactitud la fuga o abandono de un cliente debido a que no existe una ruptura de contrato.

Este proyecto tiene como objetivo proponer un método para detectar posibles fugas de clientes, esto se hizo mediante la estimación de la densidad del tiempo entre visitas de clientes de una manera no paramétrica utilizando kernels Gamma y en la estimación dando mayor importancia a las observaciones más recientes con el fin de detectar comportamientos anómalos del tiempo entre visitas de un cliente específico los cuales pueden significar una posible fuga del cliente. El método propuesto proporciona una manera de detectar posibles fugas de clientes de empresas cuya forma de hacer negocio se da en un escenario no contractual.

Palabras Clave: Estimación de densidad, kernels Gamma, fuga de clientes, escenario no contractual.

ABSTRACT

Client retention is an important topic for companies, therefore they should implement methods for detecting possible customer churn; several methods for this exists in contractual settings but few methods exists for non-contractual settings because in this type of setting it is not possible to observe the abandonment of a client given that there is no rupture of a contract, the objective of this project is to propose a method that can detect possible customer churn, this is done by density estimation of the time between visits of clients in a non-parametric way using Gamma kernels and by giving more relevance to recent observations in the estimation with the goal of detecting anomalies in the behavior of the time between visits for a specific client; anomalies that might be attributed to the churn or abandonment of the client. The proposed method provides a way of detecting possible customer churn for companies that do business in a non-contractual setting.

***Keywords**: Density estimation, Gamma kernels, Customer churn, non-contractual setting.*

ÍNDICE GENERAL

RESUMEN I

ABSTRACT II

ÍNDICE GENERAL III

ABREVIATURAS V

ÍNDICE DE FIGURAS VI

ÍNDICE DE TABLAS VII

CAPÍTULO 1 1

1. INTRODUCCION 1

1.1 Descripción del problema 1

1.2 Justificación del problema 2

1.3 Objetivos 3

1.3.1 Objetivo General 3

1.3.2 Objetivos Específicos 3

1.4 Marco teórico 3

1.4.1 Diagrama de Pareto 3

1.4.2 Estimación de densidad por kernels. 4

1.4.3 Método de bisección. 5

CAPÍTULO 2 7

2. METODOLOGIA 7

2.1 Creación de bases de datos. 7

2.2 Selección de clientes para seguimiento. 7

2.3 Elección de clientes a retener. 7

2.4 Derivación del método. 9

CAPÍTULO 3 20

3. RESULTADOS Y ANÁLISIS 20

3.1 Selección de la muestra.....20
3.2 Elección de parámetros 20
CAPÍTULO 4 25
4. CONCLUSIONES Y RECOMENDACIONES.....25
4.1. Conclusiones 25
4.2. Recomendaciones 25
BIBLIOGRAFÍA 26

ABREVIATURAS

ESPOL Escuela Superior Politécnica del Litoral

KDE Estimador de Densidad por Kernels

ÍNDICE DE FIGURAS

Gráfico 1.1. Diagrama de Pareto 4
Gráfico 2.1. Diagrama de Pareto de Ventas 8
Gráfico 2.2. Diagrama de Pareto de Visitas 8
Gráfico 2.3. Dispersión Real 9
Gráfico 2.4. Dispersión Logarítmica 9
Gráfico 2.5. Kernels Gaussianos 11
Gráfico 2.6. Kernels Gamma 13
Gráfico 2.7. Kernels Gamma Individuales 13
Gráfico 2.8. Frecuencia Visitas por Mes 14
Gráfico 2.9. Kernels Gamma Weight 17
Gráfico 2.10. Método de Bisección 18

ÍNDICE DE TABLAS

Tabla 3.1. Parámetros 21
Tabla 3.2. Cliente A 21
Tabla 3.3. Cliente B 21
Tabla 3.4. Cliente C 22
Tabla 3.5. Cliente D 22
Tabla 3.6. Cliente E 22
Tabla 3.7. Cliente F 22
Tabla 3.8. Cliente G 22
Tabla 3.9. Cliente H 22
Tabla 3.10. Cliente I 23
Tabla 3.11. Cliente J 23
Tabla 3.12. Cliente K 23

CAPÍTULO 1

1. INTRODUCCIÓN

En el presente proyecto se propone un método para la posible detección de fugas de clientes para empresas cuya manera de hacer negocio se da en un escenario no contractual, es decir, no existe un contrato de por medio entre el cliente y la empresa; esto dificulta la detección de la fuga de clientes ya que la empresa no observa con certeza cuando una persona deja de ser un cliente.

1.1 Descripción del problema

La retención de clientes es valiosa para las empresas y para esto se debe saber cuáles son los clientes que están próximos a fugarse, en un escenario no contractual como el que se da por ejemplo en el sector minorista la detección de la fuga de un cliente se puede complicar ya que no se puede observar con exactitud el abandono de este; retener o fidelizar clientes trae beneficios a las empresas; el costo de captar un nuevo cliente es significativamente mayor al costo de retener uno ya existente e incrementar la tasa de retención de clientes en un 5% puede incrementar las ganancias del negocio entre un 25% y 85% (Reichheld, 1990).

En escenarios no contractuales no se observa con certeza la fuga de un cliente, pues no existe ruptura de contrato y por tanto no existe una aviso por parte del cliente que abandonará la empresa, el único indicio que la empresa tiene de un posible abandono por parte del cliente es el tiempo desde el último contacto con este, para el caso del sector minorista sería el tiempo desde la última visita o compra del cliente en alguna sucursal, la ausencia prolongada desde su última visita a alguna sucursal puede indicar que este ya no regresará más debido a que decidió irse por la competencia o por alguna otra razón.

1.2 Justificación del problema

Las empresas cuyos negocios se dan en escenarios contractuales tienen la ventaja de que al suscribir un contrato con el cliente pueden observar con certeza cuándo un cliente abandona la empresa, pues para esto el cliente debe comunicar la ruptura del contrato que mantiene, esto le da a la empresa un historial claro de la cantidad de clientes activos o inactivos que tiene y en base a este historial puede analizar variables que definan un perfil o identifiquen a clientes que estén próximos o en peligro a fugarse, para esto pueden usar métodos como regresión logística, árboles de decisión, bosques aleatorios, etc.

Sin embargo, en las empresas cuyos negocios se dan en escenarios no contractuales este tipo de ventajas no están presentes, cabe señalar que en este escenario la empresa no observa con certeza el abandono del cliente debido a la falta de un contrato y no tiene la ventaja de conocer con exactitud cuántos clientes activos o inactivos tiene, por tanto, la empresa no puede utilizar las mismas técnicas que aplican las empresas en escenarios contractuales.

Consciente de la importancia de la fidelización de los clientes para la sostenibilidad de la empresa se requiere de un método de detección de fuga de clientes en el contexto de empresas que se manejan en escenarios no contractuales, el beneficio de este método estadístico radica en que la empresa podrá detectar clientes en riesgo de fuga y posteriormente sobre este conocimiento, diseñar estrategias para su retención.

En relación con lo teórico el método está fundamentado en técnicas de estadística no paramétrica lo cual es beneficioso ya que no se basa en supuestos obteniendo una herramienta más robusta. Metodológicamente su aplicación puede ser una herramienta valiosa para otras organizaciones con actividades no contractuales.

1.3 Objetivos

1.3.1 Objetivo General

Diseñar un método que permita la detección de comportamientos anómalos en el tiempo entre visitas de clientes en base a sus registros de frecuencia de compra en escenarios no contractuales.

1.3.2 Objetivos Específicos

- Discriminar los clientes con mejor perfil adquisitivo y de frecuencia de compra para la aplicación del método.
- Estimar la densidad de tiempo entre visitas de un cliente específico mediante técnicas de estadística no paramétrica.
- Incorporar en la estimación el efecto de los cambios en el comportamiento de visitas de clientes a través del tiempo.
- Establecer límites para alertas de comportamiento anómalo de clientes detectando posibles fugas o inactividades temporales del cliente.

1.4 Marco teórico

1.4.1 Diagrama de Pareto

Llamado así por Vilfrido Pareto, es un tipo de diagrama que contiene barras y una línea, donde los valores individuales son representados en orden descendiente por las barras, y el total acumulado es representado por la línea.

El eje vertical izquierdo es la frecuencia de ocurrencia, pero puede alternativamente representar costo u otra unidad de medida importante como dinero. El eje vertical derecho es el porcentaje acumulado del número total de ocurrencias, dólares totales u otra unidad de medida importante. Como los valores están en orden decreciente, la función acumulada es una función cóncava. El propósito del Pareto es resaltar los factores más importes de entre generalmente un grupo grande de factores.

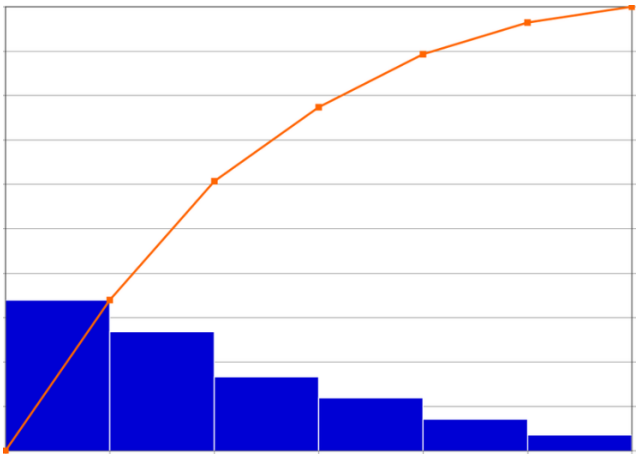


Gráfico 1.1. Diagrama de Pareto

Fuente: (Wikipedia)

1.4.2 Estimación de densidad por kernels.

Dada una muestra $X_1, \dots, X_n$ de una distribución $f(x)$ la cual se desea estimar, $f(x)$ es generalmente llamada la "verdadera" densidad, entonces el estimador de densidad por kernel (KDE) tiene la forma:

$$\hat{f}(x) = \sum_{i=1}^{n} w_i K_h(x - X_i) = \sum_{i=1}^{n} \frac{1}{h} w_i K\left(\frac{x - X_i}{h}\right) \qquad (1.1)$$

Donde $K_h(\cdot) = K(\cdot/h)/h$, K es la función kernel que satisface $\int K(u)\, du = 1, h > 0$ es el ancho de banda, parámetro que define que tan suavizada es la estimación de $f(x)$ y w_i es el peso que tiene la i-ésima observación los cuales son normalizados para que sumen a 1, es decir, $\sum_{i=1}^{n} w_i = 1$.

Si a todas las observaciones se les da el mismo peso, es decir, $\forall_i : w_i = \frac{1}{n}$, entonces el estimador de densidad por kernel (KDE) tiene la forma:

$$\hat{f}(x) = \frac{1}{nh}\sum_{i=1}^{n} K\left(\frac{x - X_i}{h}\right) \qquad (1.2)$$

1.4.3 Método de bisección.

El método de bisección es un método para hallar raíces en el cual de manera iterativa se divide en dos partes un intervalo y luego selecciona un subintervalo en el cual una raíz se debe encontrar para seguir procesando.

El método es aplicable para resolver numéricamente la ecuación $f(x) = 0$ para la variable real x, donde f es una función continúa definida en el intervalo $[a, b]$ y donde $f(a)$ y $f(b)$ tienen signos opuestos. En este caso se dice que a y b "encierran" a una raíz ya que por el teorema del valor intermedio la función continua f debe al menos contener una raíz en el intervalo (a, b).

En cada iteración el método divide el intervalo en dos calculando el punto medio $c = \frac{(a+b)}{2}$ del intervalo y el valor de la función $f(c)$ en ese punto. A no ser que c sea una raíz se tiene solo dos opciones: $f(a)$ y $f(c)$ tienen signos opuestos y "encierran" una raíz, o $f(c)$ y $f(b)$ tienen signos opuestos y encierran una raíz. El método selecciona el subintervalo que garantiza que "encierra" una raíz como nuevo intervalo para ser usado en la siguiente iteración. El proceso es continuado hasta que el intervalo es lo suficientemente pequeño.

El algoritmo recibe como entrada una función continua f, un intervalo $[a, b]$ y los valores de la función $f(a)$ y $f(b)$. Los valores de la función son de signo opuesto. En cada iteración se realizan los siguientes pasos:

1. Calcular c, el punto medio del intervalo, $c = \frac{(a+b)}{2}$.
2. Calcular el valor de la función en el punto medio, $f(c)$.

3. Si la convergencia es satisfactoria, esto es, si $c - a$ es suficientemente pequeña, o $|f(c)|$ es suficientemente pequeña, retornar c y terminar de iterar.
4. Examinar el signo de $f(c)$ y reemplazar $(a, f(a))$ o $(b, f(b))$ con $(c, f(c))$ de tal manera que exista una raíz dentro del nuevo intervalo.

CAPÍTULO 2

2. METODOLOGÍA

En este capítulo se presenta de manera detallada la metodología que se usó para el desarrollo del proyecto.

2.1 Creación de bases de datos.

Para la creación de la base de datos se utilizaron datos transaccionales facilitados por una empresa del sector minorista cuya matriz se encuentra en la ciudad de Guayaquil, por sigilo no se revela la identidad de la empresa. Estos datos transaccionales son únicamente de los clientes que poseen la tarjeta de membresía de la empresa.

2.2 Selección de clientes para seguimiento.

En el año 2012 la cantidad de clientes con tarjeta de membresía que realizó al menos una compra fue 1.477.151, este mismo año se registraron 272.392 nuevos clientes, estos clientes nuevos no fueron tomados en cuenta para el seguimiento quedando así 1.288.941 clientes.

Se tomaron los códigos de identificación de estos clientes y se hizo un seguimiento de la cantidad de visitas y cantidad de dólares que gastaron en la empresa durante el periodo de enero del 2012 hasta diciembre 2017, es decir, durante 6 años.

2.3 Elección de clientes a retener.

En última instancia el propósito del método propuesto es apoyar al proceso de retención y fidelización de clientes. Es importante definir el perfil de los clientes a los cuales se les aplicará el método.

Es razonable pensar que un perfil atractivo de cliente para retención y fidelización es aquel que realiza bastantes visitas y tiene consumos significativamente altos comparados con el resto, para poder discriminar estos clientes se utilizó como herramienta el diagrama de Pareto.

Se definió un diagrama de Pareto para las ventas (consumos) acumuladas durante los 6 años (enero 2012 a diciembre 2017) y otro para las visitas acumuladas durante los mismos 6 años. Se trabajó con los clientes que agregaban el 50% del consumo acumulado y con los clientes que agregaban el 50% de las visitas acumuladas, la cantidad de clientes para el Pareto de consumo acumulado es 146.735, aproximadamente el 11.38% de los clientes en seguimiento y la cantidad de clientes para el Pareto de visitas acumuladas es 156.382, que es aproximadamente el 12.13% de los clientes en seguimiento.

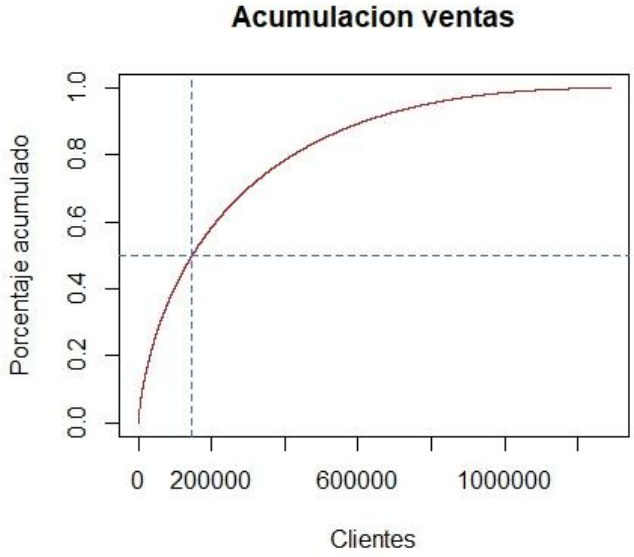


Gráfico 2.1. Diagrama de Pareto de Ventas
Fuente: El Autor

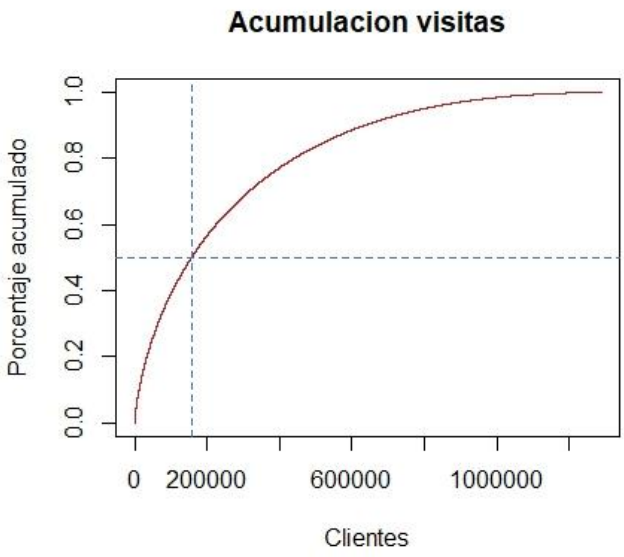


Gráfico 2.2. Diagrama de Pareto de Visitas
Fuente: El Autor

.

Se definió como grupo de clientes con perfil de interés a la intersección entre el conjunto de clientes del Pareto de consumo acumulado y el conjunto de clientes del Pareto de visitas acumuladas, el total de clientes en esta intersección es de 105.842, estos clientes son presentados con color azul en la gráfica de dispersión

que se muestra a continuación, en la escala real de los datos y en escala logarítmica para mejor apreciación.

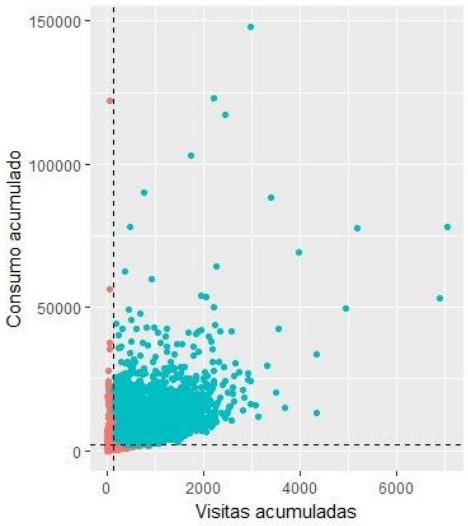


Gráfico 2.3. Dispersión Real

Fuente: El Autor

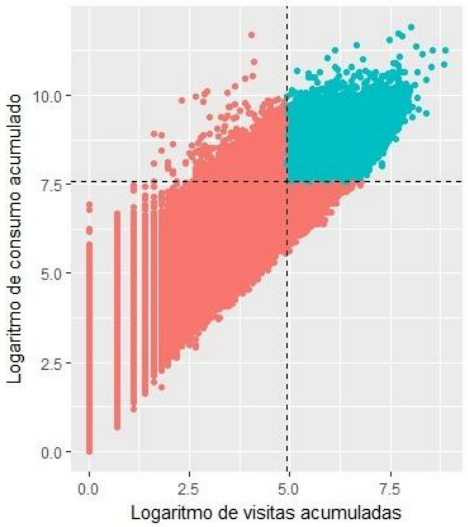


Gráfico 2.4. Dispersión Logarítmica

Fuente: El Autor

El método se aplicará sobre este grupo de clientes.

2.4 Derivación del método.

Como ya se planteó anteriormente las empresas que hacen negocios en escenarios no contractuales tienen la desventaja de no poder observar con certeza el abandono o fuga de un cliente de la empresa, es por esto que para este método se decidió principalmente utilizar el tiempo entre visitas de un cliente específico como variable para detectar posibles fugas de clientes, esto es, estimar la densidad del tiempo entre visitas de una manera no paramétrica mediante kernels, luego, una vez estimada la densidad obtener un límite superior para el tiempo entre visitas definiendo un percentil, por ejemplo, 95, 99 o 99.9.

Esto se hace para detectar comportamientos anómalos de los tiempos entre visitas del cliente, si se detecta una anomalía, es decir, si el cliente se ha demorado significativamente más de lo que acostumbra en realizar otra visita esto puede ser un indicio de fuga o abandono del cliente de la empresa.

Para la estimación de la densidad del tiempo entre visitas la elección del kernel es importante dada la naturaleza da la densidad a estimar, dado que estamos estimando tiempos la variable está acotada teóricamente; los tiempos pueden ser mayores o iguales a cero $[0, \infty)$.

Generalmente para este tipo de estimaciones se utilizan kernels simétricos, por ejemplo, kernels gaussianos, el problema con utilizar kernels de este tipo para la estimación de tiempos entre visitas se da en la frontera o vecindad en cero, es decir, para tiempos entre arribos muy pequeños kernels simétricos que se encuentran cerca de la frontera asignan peso fuera del soporte de la densidad, que en este caso es $[0, \infty)$.

Para ilustrar esto nos enfocaremos en un cliente especifico al cual llamaremos cliente “XYZ”, este cliente durante el periodo de seguimiento (6 años) ha realizado 413 visitas (412 observaciones de tiempos entre visitas). En el siguiente *Grafico 2.5* se presenta la estimación de la densidad de tiempo entre visitas mediante kernels gaussianos (simétricos), donde cada barra de color negro representa una observación y la línea roja puntada es la estimación de la densidad.

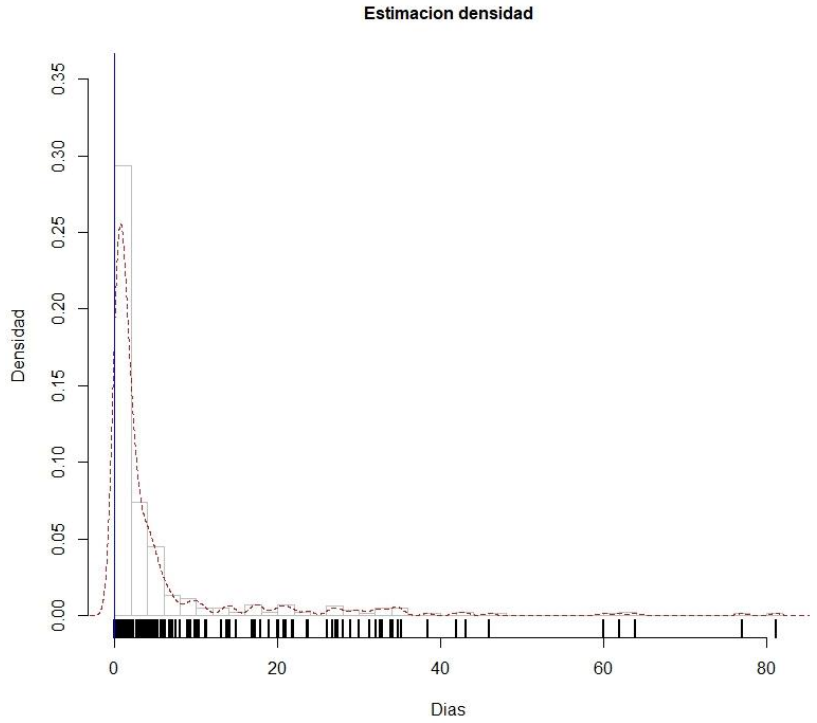


Gráfico 2.5. Kernels Gaussianos

Fuente: El Autor

En el *Gráfico 2.5.* claramente se puede evidenciar como kernels simétricos que se encuentran cerca de la frontera asignan peso fuera del soporte de la densidad.

Para estimar la densidad de este tipo de variables como el tiempo entre visitas mediante kernels simétricos lo adecuado sería hacer un ajuste para evitar este problema, varias alternativas existen para esto. La manera más natural de lidiar con este problema de la cota inferior en cero es usar el enfoque general de estimación mediante kernels simétricos y luego rescatarlos localmente en relación con la cantidad de masa del kernel local que se encuentra dentro del soporte de la variable, esto se conoce como el método de renormalización (Jones, 1993), otro método para hacer este ajuste es el método de reflexión y en esencia lo que hace es reflexionar la data en la frontera o cota, estos métodos son discutidos en (Silverman, 1986) y en (Cwik, 1993).

Una manera de abordar más naturalmente este problema dado que el soporte de la variable tiempo entre visitas es el conjunto $[0, \infty)$, es utilizar kernels asimétricos,

específicamente kernels gamma. En (Chen.S, 2000) se propone utilizar la distribución de probabilidad Gamma:

$$G_{a,b}(t) = \frac{t^{a-1}e^{-t/b}}{\Gamma(a)b^a}, \qquad a, b, t > 0 \quad (2.1)$$

Para reemplazar los kernels simétricos tradicionales para estimar la densidad de variables aleatorias continuas no negativas. El estimador de densidad de kernels gamma en (Chen.S, 2000), con un ancho de banda $h > 0$, es definido como:

$$\hat{f}_C(x) = \frac{1}{n}\sum_{i=1}^{n} G_{x/h+1,h}(X_i) = \frac{1}{n}\sum_{i=1}^{n} \frac{{X_i}^{x/h}e^{-X_i/h}}{\Gamma(x/h+1)h^{x/h+1}} \quad (2.2)$$

Este estimador elimina el sesgo en la frontera dado que está definido en la recta real no negativa. En (Yongho Jeon, 2013) se propone el siguiente estimador con diferente parametrización al de (Chen.S, 2000):

$$\hat{f}_G(x) = \frac{1}{n}\sum_{i=1}^{n} G_{X_i/h+1,h}(x) = \frac{1}{n}\sum_{i=1}^{n} \frac{x^{X_i/h}e^{-x/h}}{\Gamma(X_i/h+1)h^{X_i/h+1}} \quad (2.3)$$

Ahora, estimando la densidad de tiempo entre visitas del cliente “XYZ” mediante (2.3) con $h = 1$ se elimina el problema de asignar pesos fuera del soporte de la densidad como se puede observar en el *Gráfico 2.6.*

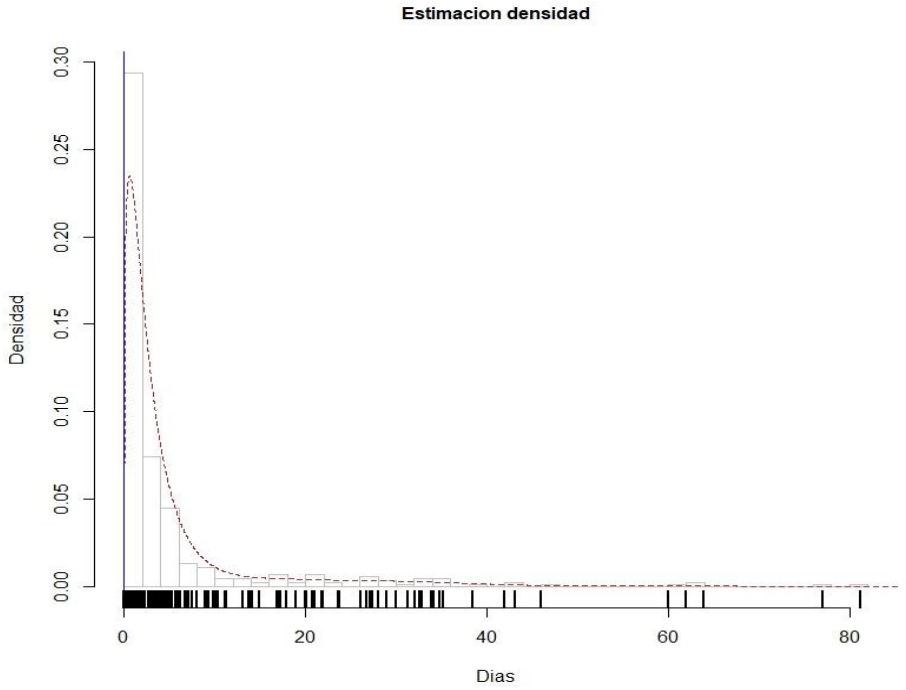


Gráfico 2.6. Kernels Gamma

Fuente: El Autor

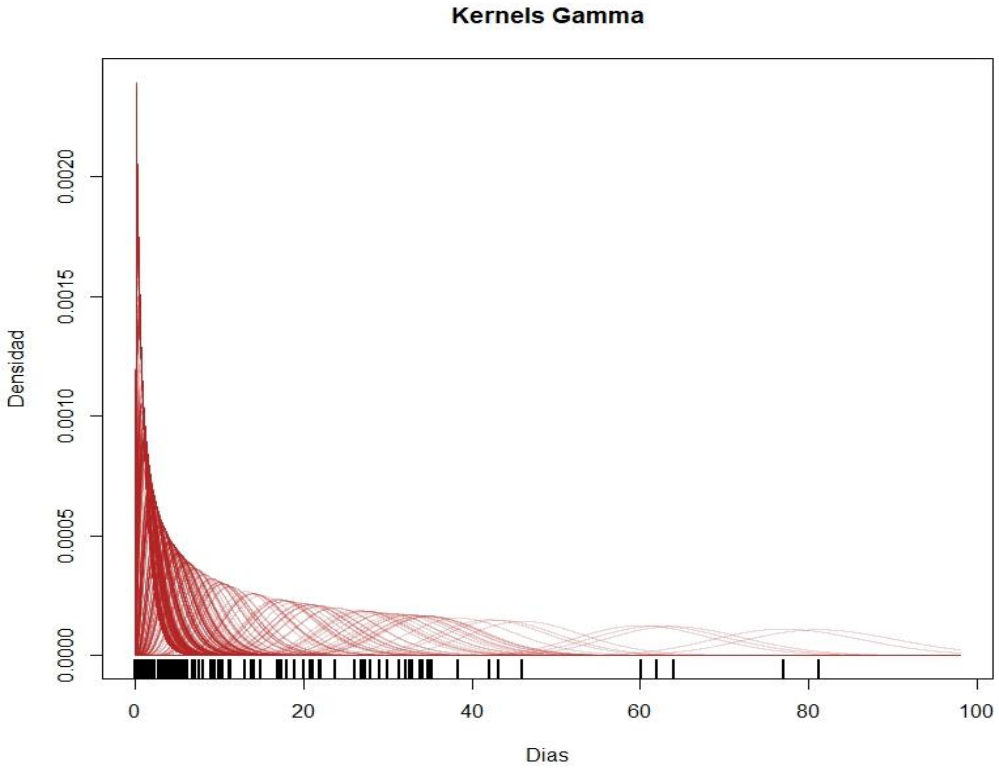


Gráfico 2.7. Kernels Gamma Individuales

Fuente: El Autor

Todavía queda un ajuste por hacer, el comportamiento de visitas de este cliente ha sido analizado durante 6 años es probable que la frecuencia con la que realiza visitas a la empresa haya variado durante esos años, una forma de verificar esto es hacerlo de manera gráfica.

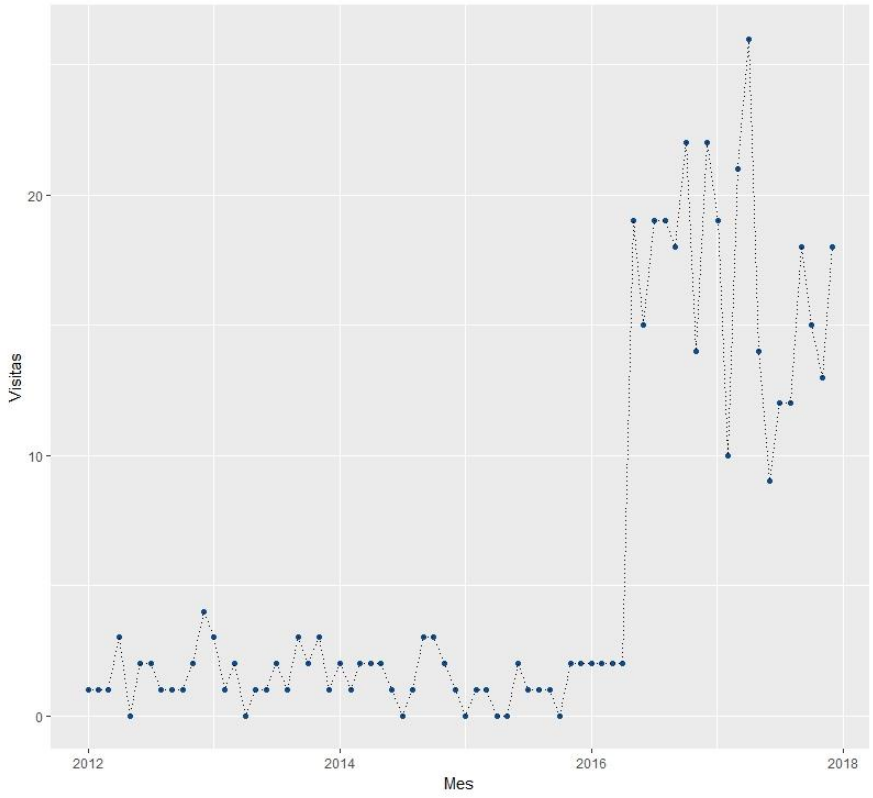


Gráfico 2.8. Frecuencia Visitas por Mes

Fuente: El Autor

Cada punto en la *Gráfico 2.8.* representa el número de visitas que realizó el cliente "XYZ" en un mes especifico, como se puede observar, los primeros años el cliente en promedio venia unas 2 veces por mes, pero luego a partir del 2016 su frecuencia de visita promedio ascendió a unas 15 visitas al mes, esto implica que en los últimos años sus tiempos entre visitas se redujeron significativamente y es importante tomar en cuenta esto al momento de estimar la densidad de tiempo entre visitas para poder detectar comportamientos anómalos.

La forma de incorporar estos cambios del comportamiento más reciente del tiempo entre visitas del cliente es darle mayor peso o mayor relevancia a las ultimas

observaciones al momento de hacer la estimación mediante kernels gamma, es decir no darles el mismo peso $\frac{1}{n}$ a las observaciones.

Se define el peso para la i – ésima observación de la siguiente manera:

$$w_i = \alpha(1-\alpha)^{T-i} \quad (2.4)$$

Donde,

$$i = 1,2,\dots,T$$
$$0 < \alpha < 1$$

El parámetro α se encarga de la relevancia que se les da a las observaciones más recientes, mientras mayor sea α mayor relevancia se le da a las ultimas observaciones y por ende menor relevancia a las observaciones más antiguas. T es el número de observaciones.

Se puede mostrar que $\{w_i\}$ es una progresión geométrica, con razón:

$$r = \frac{\alpha(1-\alpha)^{T-i}}{\alpha(1-\alpha)^{T-(i-1)}} = \frac{\alpha\left[(1-\alpha)^{T}(1-\alpha)^{-i}\right]}{\alpha[(1-\alpha)^{T}(1-\alpha)^{-i}(1-\alpha)]}$$

$$r = \frac{1}{1-\alpha} \quad (2.5)$$

Para que la densidad estimada integre a 1, la suma de los pesos debe ser 1 también, es decir, $\sum_{i=1}^{T} w_i = 1$. La suma hasta el n-ésimo termino de $\{w_i\}$ esta dada por:

$$S = \alpha(1-\alpha)^{T-1}\left(\frac{1-\left(\frac{1}{1-\alpha}\right)^{n}}{1-\left(\frac{1}{1-\alpha}\right)}\right)$$

$$S = (1-\alpha)^{T-n} - (1-\alpha)^{T} \quad (2.6)$$

Para que $\sum_{i=1}^{T} w_i = 1$ se cumpla tendría que suceder $\boxed{S = 1 \ \wedge n = T}$, por tanto, tendríamos:

$$S = (1-\alpha)^{T-n} - (1-\alpha)^{T} \ \Rightarrow \ 1 = 1 - (1-\alpha)^{T}$$

$$(1-\alpha)^{T} = 0$$

Dado que $0 < \alpha < 1$ esto significa que $\boxed{T \to \infty \Rightarrow \sum_{i=1}^{T} w_i \to 1}$.

Entonces lo que se puede hacer es aproximar $\sum_{i=1}^{T} w_i$ a 1.

$$S \approx 1 \Rightarrow S + \epsilon = 1, \epsilon > 0$$

$$1 - \epsilon = 1 - (1-\alpha)^{T}$$

$$(1-\alpha)^{T} = \epsilon \quad (2.7)$$

Finalmente tenemos,

$$T = \frac{\ln(\epsilon)}{\ln(1-\alpha)} \quad (2.8)$$

Donde $\epsilon > 0$ es el error, esto es, $\boxed{\epsilon \to 0 \Rightarrow \sum_{i=1}^{T} w_i \to 1}$.

Ahora, basándose en la definición (1.1) e incorporando (2.4) y (2.3) se define el siguiente estimador:

$$\hat{f}_{GW}(x) = \sum_{i=1}^{n} w_i \cdot G_{X_i/h+1,h}(x) = \sum_{i=1}^{n} \alpha(1-\alpha)^{T-i} \frac{x^{X_i/h} e^{-x/h}}{\Gamma(X_i/h+1) h^{X_i/h+1}} \quad (2.9)$$

Usando (2.9) para estimar la densidad del tiempo entre visitas del cliente “XYZ” con $\alpha = 0.1$, $T = 412$ y $h = 1$ se puede observar gráficamente como al incorporar los pesos w_i la estimación de la densidad ahora es diferente ya que toma en cuenta los cambios en la frecuencia de visitas del cliente a través del tiempo.

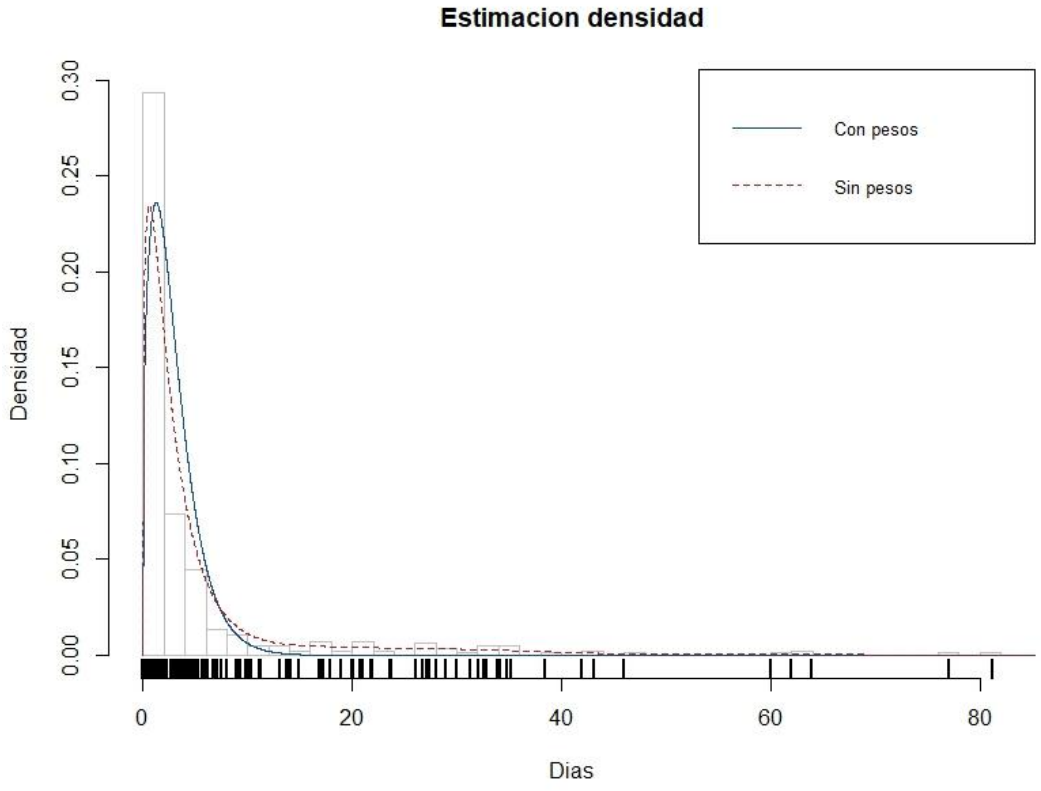


Gráfico 2.9. Kernels Gamma Weight

Fuente: El Autor

Con (2.7) podemos obtener que tan cercano a 1 es $\sum_{i=1}^{T} w_i$, reemplazando α y T tenemos que $\epsilon = (1-0.1)^{412} = 1.40576x10^{-19}$; un numero muy cercano a 0, lo que implica que $\sum_{i=1}^{T} w_i = 1 - 1.40576x10^{-19}$, lo cual es muy cercano a 1.

Lo que queda es obtener el límite para la alerta de comportamientos anómalos del tiempo entre visitas del cliente “XYZ”.

Sea $\hat{F}_{GW}$ la acumulada de $\hat{f}_{GW}$ y sea $p \in (0,1)$, sabemos que $\hat{F}_{GW}(x)$ es continua $\forall_x$, el objetivo es encontrar x que satisfaga $\hat{F}_{GW}(x) = p$, es decir, queremos encontrar las raíces de la siguiente ecuación:

$$\hat{F}_{GW}(x) - p = 0 \quad (2.10)$$

Sabemos que $\hat{F}_{GW}$ por ser una distribución acumulada es monótona creciente, por tanto, (2.10) tendrá una sola raíz. Las condiciones son adecuadas para utilizar el método de bisección o de búsqueda binaria para resolver (2.10). Para este caso se fija $p = 0.9999$ y se encuentra la raíz de (2.10) mediante el método de bisección.

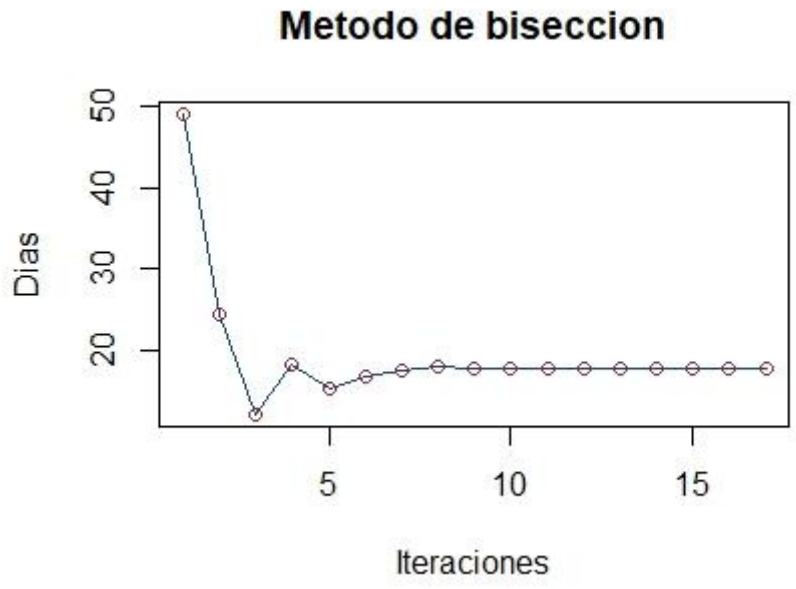


Gráfico 2.10. Método de Bisección
Fuente: El Autor

Como se puede observar en el *Gráfico 2.10.* la solución encontrada es $x = 17.749$ con una tolerancia al error absoluto $\delta = 0.001$, se toma $\lceil x \rceil = 18$ como solución entera, esto quiere decir que la alerta se activará si la última visita del cliente fue hace 18 días o más.

Cabe recalcar que al no incorporar los pesos w_i el límite para la alerta de comportamientos anómalos hubiera sido 96 días lo cual no reflejaba el cambio en el comportamiento de visitas del cliente “XYZ” a través del tiempo.

CAPÍTULO 3

3. RESULTADOS Y ANÁLISIS

En esta sección se presentan los principales resultados de la aplicación del método a una muestra aleatoria de clientes.

3.1 Selección de la muestra

Se seleccionó una muestra aleatoria de 500 clientes del grupo de 105.842 clientes escogidos con perfil atractivo para fidelización definido en el capítulo anterior, se realizó un filtro para trabajar solo con los clientes de esta muestra que hayan visitado al menos una vez alguna sucursal en los últimos 6 meses del 2017; una vez realizado el filtro quedaron 462 clientes.

3.2 Elección de parámetros

La elección del parámetro de ancho banda h sigue siendo un problema abierto en las estimaciones de densidades mediante kernels, existen propuestas para la selección automática de este parámetro pero de manera general la elección del parámetro se reduce a probar varios valores para h y corroborar cual funciona mejor dentro del contexto del problema; para este caso se hizo eso, se probaron varios valores para h y gráficamente se escogió cual daba un mejor ajuste, luego de esto se escogió $h = 1$, para los parámetros α y p se probaron los valores $(0.2, 0.15, 0.1)$ y $(0.99,\ 0.999,\ 0.9999, 0.99999)$ respectivamente, se aplicó el método a los 462 clientes de la muestra y se estimaron individualmente sus densidades de tiempo entre visitas usando la información de enero del 2012 a diciembre del 2017 (6 años) y luego se contrastó cuantas alertas se activarían en el periodo de prueba del primer día de enero del 2018 al primer día de marzo del 2018 (2 meses) para estos 462 clientes, en la siguiente tabla se muestran los resultados.

Tabla 3.1. Parámetros

		p			
		0.99	0.999	0.9999	0.99999
α	0.2	108	44	26	15
	0.15	93	33	22	13
	0.1	77	29	17	11

Como se puede observar en la tabla anterior si fijamos un α y se aumenta p se verifica que el un numero de alertas disminuye lo cual era de esperarse ya que la exigencia para que un tiempo entre visitas sea considerado anómalo aumenta, por tanto, menos alertas se activaron. Un patrón similar se puede observar si se fija p y se disminuye α; el número de alertas disminuye, esto se da ya que al disminuir α se esta disminuyendo la relevancia a las observaciones más recientes y aumentado la relevancia a las observaciones más antiguas.

Con esto, se escoge $\alpha = 0.1$ y $p = 0.99999$, ahora se analiza el comportamiento de cada uno de los 11 clientes que tuvieron comportamientos anómalos en sus tiempos entre visitas para observar si en realidad se fugaron, tuvieron un periodo de inactividad temporal o fueron solo falsas alarmas. Se define como inactividad temporal a la ausencia de un cliente en un periodo mayor o igual a 4 meses.

Tabla 3.2. Cliente A

	FECHA_HORA	tiempo
1	11/16/2017	#N/A
2	1/15/2018	60
3	7/22/2018	188
4	8/2/2018	10
Limite		56

Tabla 3.3. Cliente B

	FECHA_HORA	tiempo
1	7/25/2017	#N/A
2	3/16/2018	233
3	8/2/2018	139
Limite		108

Tabla 3.4. Cliente C

	FECHA_HORA	tiempo
1	10/18/2017	#N/A
2	5/9/2018	203
3	8/2/2018	84
Limite		77

Tabla 3.5. Cliente D

	FECHA_HORA	tiempo
1	9/18/2017	#N/A
2	6/1/2018	256
3	8/2/2018	62
Limite		93

Tabla 3.6. Cliente E

	FECHA_HORA	tiempo
1	9/21/2017	#N/A
2	2/9/2018	141
3	2/12/2018	3
4	4/9/2018	56
5	4/16/2018	7
6	4/18/2018	2
7	8/2/2018	105
Limite		100

Tabla 3.7. Cliente F

	FECHA_HORA	tiempo
1	10/16/2017	#N/A
2	4/18/2018	185
3	5/9/2018	21
4	5/15/2018	6
5	6/20/2018	36
6	7/3/2018	13
7	7/16/2018	13
8	8/2/2018	16
Limite		90

Tabla 3.8. Cliente G

	FECHA_HORA	tiempo
1	10/28/2017	#N/A
2	1/14/2018	77
3	1/15/2018	2
4	1/30/2018	15
5	2/3/2018	3
6	3/25/2018	51
7	3/30/2018	4
8	5/11/2018	43
10	5/25/2018	14
11	5/27/2018	2
12	7/15/2018	48
13	8/2/2018	18
Limite		72

\

Tabla 3.9. Cliente H

	FECHA_HORA	tiempo
1	12/8/2017	#N/A
2	3/2/2018	84
3	3/5/2018	3
4	4/18/2018	44
5	5/20/2018	32
7	5/30/2018	10
8	6/12/2018	13
9	7/4/2018	22
10	8/2/2018	28
Limite		54

Tabla 3.10. Cliente I

	FECHA_HORA	tiempo
1	9/2/2017	#N/A
2	8/2/2018	333
Limite		120

Tabla 3.11. Cliente J

	FECHA_HORA	tiempo
1	12/31/2017	#N/A
2	1/7/2018	7
3	8/2/2018	207
Limite		49

Tabla 3.12. Cliente K

	FECHA_HORA	tiempo	FECHA_HORA	tiempo
1	12/27/2017	#N/A	4/30/2018	19
2	2/19/2018	54	5/18/2018	18
3	3/16/2018	25	5/26/2018	8
4	3/18/2018	2	5/28/2018	2
5	3/22/2018	4	6/18/2018	21
6	3/30/2018	8	6/29/2018	11
7	4/2/2018	3	6/30/2018	1
8	4/7/2018	5	7/2/2018	2
10	4/9/2018	2	7/5/2018	3
11	4/11/2018	2	7/7/2018	2
			7/8/2018	1
			7/9/2018	1
			7/17/2018	8
			7/23/2018	6
			7/26/2018	3
			8/2/2018	6
			Limite	51

De las tablas presentadas se puede observar resaltado con color rojo los tiempos entre visitas que sobrepasan los 120 días (4 meses) ya que se definió a los periodos mayores a este como periodo de inactividad temporal o fuga parcial del cliente. Los clientes A, B, C, D, E, F, I y J fueron correctamente identificados como fugados o fugados parcialmente, los clientes G, H y K fueron falsas alarmas ya que ninguno de sus tiempos entre visitas excedió los 120 días.

Esto nos deja con 8 clientes de 11 que fueron correctamente identificados como fugas parciales o totales, esto nos da aproximadamente un 73% de correcta detección.

CAPÍTULO 4

4. CONCLUSIONES Y RECOMENDACIONES

4.1. Conclusiones

- Se logró discriminar los clientes con mejor perfil adquisitivo y de frecuencia de compra mediante el uso de diagramas de pareto, obteniendo un total de 105.842 clientes con perfil atractivo para retención, en un periodo de 6 años de seguimiento.
- Se consiguió estimar la densidad del tiempo entre visitas de un cliente especifico mediante técnicas de estadística no paramétrica utilizando kernels.
- Se eliminó el problema de asignación de pesos fuera del soporte de la variable utilizando kernels asimétricos gamma.
- Se logró incorporar el efecto de los cambios en el comportamiento de visitas del cliente aplicando pesos w_i que daban mayor relevancia a las observaciones más recientes acorde con el valor que se le daba al parámetro α.
- Se obtuvo un 73% de correcta detección de fuga total o parcial de los clientes cuyos días de ausencia excedían sus límites de tiempo entre visitas.

4.2. Recomendaciones

Como recomendación para mejorar la estimación de los tiempos entre visitas como alternativa se podrían utilizar anchos de banda variable, el ancho de banda puede ser pequeño donde existen muchas observaciones y grande en donde existen pocas observaciones.

BIBLIOGRAFÍA

Chen,S. (2000). Probability Density Function Estimation Using Gamma Kernels. *Annals of the Institute of Statistical Mathematics*, 471-480.

Cwik, J. A. (1993). Data-dependent bandwidth choice for a grade density. *Statistics & Probability Letters 16*, 397–405.

Jones, M. (1993). Simple boundary correction for kernel density estimation. *Statistics and computing*, 135-146.

Reichheld, F. F. (Septiembre de 1990). *Zero Defections: Quality Comes to Services.* Harvard Bussines. Accedido el 11 de junio, 2018, desde https://hbr.org/1990/09/zero-defections-quality-comes-to-services

Silverman, B. W. (1986). Density Estimation for Statistics and Data Analysis. School of Mathematics University of Bath.

Yongho Jeon, J. H. (2013). A gamma kernel density estimation for insurance loss data. *Insurance: Mathematics and Economics 53*, 569-579.