

# CARACTERIZACIÓN DEL PROCESO DE FUGA DE CLIENTES UTILIZANDO INFORMACIÓN TRANSACCIONAL

Carolina Segovia [csegovia@analytics.cl](mailto:csegovia@analytics.cl)

Luis Aburto [luaburto@analytics.cl](mailto:luaburto@analytics.cl)

Marcel Goic [mgoic@dii.uchile.cl](mailto:mgoic@dii.uchile.cl)

## Resumen

Para cualquier tipo de empresa es sumamente relevante detectar el momento en que un cliente se va de ella: es más barato retener a un cliente que captar uno nuevo. La metodología propuesta caracteriza el proceso de fuga de clientes utilizando sus datos transaccionales. La dinámica de la situación es modelada a través de cadenas de Markov, en donde cada estado representa un grupo de clientes con similares atributos Recency, Frequency y Monetary value (RFM). La metodología caracteriza el camino de los clientes hacia la fuga, estimando las probabilidades de fuga en cada período. En base a la información analizada, la metodología permitió detectar cómo los consumidores se mueven lentamente hacia la inactividad. Este conocimiento es importante para las empresas, porque les entrega una señal oportuna que pueden utilizar para definir acciones preventivas, enfocadas a retener a sus clientes más rentables.

**Palabras clave:** Fuga de clientes, RFM, Cadenas de Markov.

## Abstract

Customer churn detection is relevant for any company: the cost of acquiring a good new customer is far greater than the cost of retaining a good “old” one. The proposed methodology characterizes the churn process of customers using their transactional data. The dynamics of the situation are modeled with Markov chains, where each state represents a group of clients with similar Recency, Frequency and Monetary value attributes. The methodology characterizes the “path” that a customer leads to churn, besides estimating the churn probabilities in every period. Based on information analyzed, the methodology proposed detects how customers are moving slowly towards inactivity. This knowledge is valuable to managers because it gives them a timely signal they can use to define preventive actions to retain valuable customers.

**Key words:** Customer churn, RFM, Markov chains.

## 1. Introducción y definición del problema

El éxito de cualquier empresa depende de su capacidad para mantener a los buenos clientes, ya que a medida que aumenta la antigüedad de estos, también aumenta el beneficio que entregan a la empresa. Modelos que ayuden a conocer a los clientes son muy relevantes, ya que pueden ayudar a la empresa a focalizar sus políticas de marketing, concentrando sus esfuerzos en los clientes más rentables.

Marker (1998) estudió políticas de retención de clientes en una compañía de seguros, modelando la dinámica de la situación con cadenas de Markov, en donde cada estado representaba el número de períodos que el cliente permanecía activo. Más adelante, Pfeifer y Carraway (2000) modelaron con cadenas de Markov la relación de un cliente con una compañía de marketing directo, obteniéndose distintos valores de LTV para diferentes configuraciones RFM.

El presente trabajo caracterizará el proceso de fuga de clientes, modelando la dinámica de la situación a través de cadenas de Markov, con estados correspondientes a distintas clasificaciones RFM. La metodología se aplicará en el caso de un retail banking, el cual cuenta con 22.000 clientes, los cuales son empresas que lo contratan para que les proporcione el servicio de tarjetas de crédito en sus dependencias.

## 2. Desarrollo

Se utilizará la información transaccional de los clientes, de los meses de enero del 2003 a noviembre del 2004, específicamente las variables RFM. Para el análisis se ocupará una parte de la base de datos para entrenar los modelos, es decir, predecir con diferentes formas la matriz de probabilidades de transición y otra parte de la base para testeo, la cual se utilizará para elegir el mejor tipo de clasificación RFM. Tradicionalmente en problemas de data mining la partición utilizada es 80% de los datos para train y el 20% restante para test, esta será la participación a ocupar en este trabajo.

### Definición Criterio de Fuga

Será de suma importancia conversar con personas de la empresa, con el objetivo de conocer cuales son las variables relevantes para considerar a un cliente fugado. Además, dado las convenciones del modelo, si un cliente cae en el estado de fuga quedará ahí para siempre, por lo que para escoger un criterio de fuga hay que tener en cuenta que el porcentaje de clientes que vuelve sea pequeño.

Tomando en cuenta lo anterior se decidió utilizar un criterio de fuga mixto, en el cual si un cliente llegaba a tener un R/F (que tan atrasado o adelantado está un cliente) mayor o igual a 20 se consideraba fugado, pero si el cliente no cumplía con ese criterio, pero llegaba a completar una inactividad de 120 días, también se consideraba fugado. Con este criterio se fuga el 6,5% de los clientes en cada período, y un 5% de estos vuelve a realizar transacciones al período siguiente.

### Definición de experimentos

Cada experimento será una cadena de Markov, en cada uno se variarán los parámetros que se pueden ver en la Figura 1.

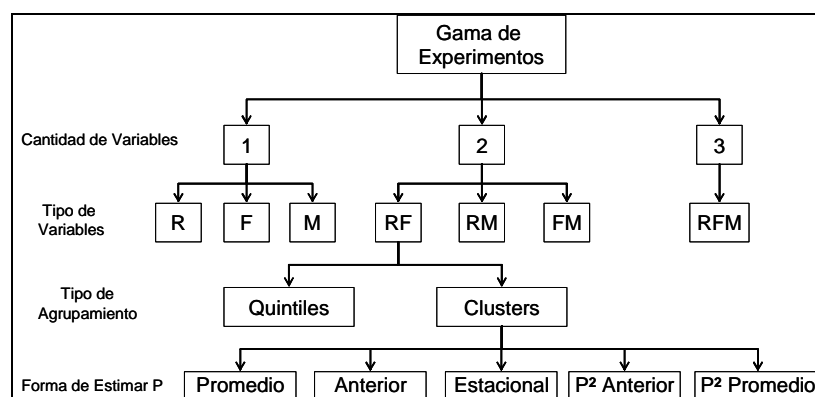


Figura 1: Experimentos a realizar

**a. Tipos de Variables:** Se harán experimentos con una sola variable a la vez (R, F o M), con dos variables (RF, RM y FM) y por último con las tres variables (RFM).

**b. Tipo de Agrupamiento:** Los estados se construirán a partir de agrupamientos de los clientes, con respecto a sus variables RFM. Así habrá dos tipos de agrupamiento:

1. **Quintiles:** Agrupamiento de clientes, en donde se crean por cada variable RFM cinco segmentos de clientes, en donde cada uno con la misma cantidad de clientes. Esto se logra ordenando de menor a mayor los clientes por cada variable, luego se dividen en cinco los clientes, donde el quintil uno representa a los clientes con mejores valores del atributo y el quintil cinco los con peores.
2. **Clusters:** Esta clasificación se realiza minimizando la varianza dentro de cada grupo y maximizando la varianza entre grupos. Para esto se utilizará el algoritmo k medias, el que se aplicará para cada una de las clasificaciones (R, RF, RM, etc.). En cada caso se deberá determinar el número óptimo de clusters, lo que se hará analizando la contribución marginal, que significa incluir un cluster más (observando los valores que entrega la distancia euclidiana).

**c. Estimación de probabilidades de transición:** Si se desea saber cuantas personas hay en cada estado en el período t se ocuparán dos tipos de estimaciones. A continuación se muestran las estimaciones de un período, es decir, las que estiman el paso de t-1 a t:

Anterior: Será igual a la probabilidad de transición del período anterior.

$$P'_{ij}(t-1, t) = T_{ij}(t-2, t-1) \quad (1)$$

donde:

$P'_{ij}(t-1, t)$ : Estimación de la probabilidad de pasar del estado i al j, desde el período (t-1) al t.

$T_{ij}(t-2, t-1)$ : Probabilidad empírica de pasar del estado i al j, desde el período (t-2) al (t-1).

Promedio: Será igual al promedio de las transiciones anteriores.

$$P'_{ij}(t-1, t) = \text{Promedio} \{T_{ij}(1,2), T_{ij}(2,3), \dots, T_{ij}(t-2, t-1)\} \quad (2)$$

Estacional: Será igual a la transición de los mismos meses pero del año anterior.

$$P'_{ij}(\text{mes } (y-1)_{\text{año } x}, \text{mes } y_{\text{año } x}) = T_{ij}(\text{mes } (y-1)_{\text{año } x-1}, \text{mes } y_{\text{año } x-1}) \quad (3)$$

A continuación se muestran las estimaciones de dos períodos, las que estiman el paso de t-2 a t:

Anterior P<sup>2</sup>: El paso del período t-2 al t, será igual al cuadrado del paso del período (t-3) al (t-2).

$$P'_{ij}(t-2, t) = (T_{ij}(t-3, t-2))^2 \quad (4)$$

Promedio P<sup>2</sup>: Será igual al promedio de todos los pasos de un período anteriores, elevado al cuadrado.

$$P'_{ij}(t-2, t) = \text{Promedio} \{T_{ij}(1,2), T_{ij}(2,3), \dots, T_{ij}(t-3, t-2)\}^2 \quad (5)$$

**d. Errores de predicción:** Se utilizarán las siguientes medidas de error:

Error absoluto: Porcentaje de equivocaciones del modelo, con respecto al máximo de equivocaciones posibles. El máximo error absoluto que puede tener un experimento es 100%.

$$\text{error absoluto} = \frac{\sum_i \sum_j \text{ABS} \left( (N'_{ij}(t, t + \Delta t)) - (N_{ij}(t, t + \Delta t)) \right)}{N(t)} \quad (6)$$

donde:

$N'_{ij}(t, t + \Delta t)$ : Número de clientes que el modelo dijo que pasarían del estado i al j, desde el período t al (t + Δt).

$N_{ij}(t, t + \Delta t)$ : Número de clientes que de verdad pasaron del estado i al j, desde el período t al (t + Δt),  $N(t)$ : Número total de clientes que había en el período t y  $ABS(x - y)$ : Corresponde al módulo de la resta entre x e y.

**Ponderado por número de transiciones:** Este error es similar al ponderado, pero se mide con respecto al número de transiciones ocurridas, haciéndose cargo del problema de tener experimentos con distinto número de estados.

$$ponderado\ por\ transiciones = \sum_i \sum_j \frac{ABS(N'_{ij}(t, t + \Delta t) - N_{ij}(t, t + \Delta t))}{(N_{ij}(t, t + \Delta t))} \cdot \frac{1}{número\ de\ transiciones} \quad (7)$$

El máximo de este error no es 100%. Esta complicación no es importante debido a que este valor toma sentido al compararlo entre distintas clasificaciones.

**Ponderado por distancias:** Se toma el error absoluto y se pondera por la distancia existente entre dos estados. Esta distancia corresponde a la distancia euclidiana entre los centros de los dos clusters.

$$ponderado\ por\ distancias = \frac{\sum_i \sum_j ABS(N'_{ij}(t, t + \Delta t) - N_{ij}(t, t + \Delta t))}{N(t)} \cdot distancia(i, j) \quad (8)$$

Este error tampoco posee 100% como máximo, pudiendo alcanzar también valores mucho mayores.

La efectividad de cada clasificación se medirá con el error absoluto y la comparación entre clasificaciones se hará mirando los valores de los tres errores. La primera parte se realizará con los datos de train y la segunda con los de test.

## Mejor estimación de P

Para las transiciones totales, en la mayoría resultó mejor el método Anterior. En las de fuga, el mejor método fue el Estacional. Esto se aprecia en la Tabla 1.

**Tabla 1:** Resumen mejores métodos de estimación de P

Clasificación		Total	Fuga
R	Quintiles	P <sup>2</sup> Anterior	Estacional
	Clusters	Anterior	Estacional
F	Quintiles	Anterior	Estacional
	Clusters	Anterior	Estacional
M	Quintiles	Anterior	Estacional
	Clusters	Anterior	Estacional
RF	Quintiles	P <sup>2</sup> Anterior	Estacional
	Clusters	Anterior	Estacional
RM	Quintiles	P <sup>2</sup> Anterior	Estacional
	Clusters	Anterior	Estacional
FM	Quintiles	Anterior	Estacional
	Clusters	Anterior	Estacional
RFM	Quintiles	P <sup>2</sup> Anterior	Estacional
	Clusters	Anterior	Estacional

## Mejor Clasificación

Para comparar los rendimientos de todas las clasificaciones se medirán los tres tipos de error, en la base de test. La Tabla 2 muestra los valores del indicador promedio + desviación de todas las clasificaciones.

La clasificación escogida es RM cluster, debido a que presenta el mejor rendimiento para los errores absoluto y ponderado por distancias (en fuga) y en el ponderado por transiciones posee un valor casi igual

a la clasificación que obtienen el mejor resultado. Además posee también resultados totales aceptables, siendo incluso la mejor en el error ponderado por transiciones.

**Tabla 2:** Comparación de errores entre clasificaciones

Clasificación		Absoluto		Ponderado por transiciones		Ponderado por distancias	
		Total	Fuga	Total	Fuga	Total	Fuga
R	Quintiles	11.78%	2.43%	26.29%	42.35%	519.74%	489.51%
	Clusters	3.90%	1.85%	33.34%	<b>18.91%</b>	<b>178.47%</b>	278.63%
F	Quintiles	4.33%	2.30%	52.38%	38.89%	221.41%	492.80%
	Clusters	2.73%	2.81%	34.35%	42.35%	201.95%	615.54%
M	Quintiles	3.18%	2.27%	47.90%	37.91%	189.23%	498.12%
	Clusters	<b>2.50%</b>	2.43%	35.17%	64.68%	188.42%	534.11%
RF	Quintiles	15.80%	2.65%	1644.56%	53.28%	600.20%	535.65%
	Clusters	3.95%	1.90%	26.78%	19.20%	222.40%	577.96%
RM	Quintiles	16.05%	3.08%	1572.90%	82.11%	618.01%	635.08%
	Clusters	3.75%	<b>1.20%</b>	<b>18.85%</b>	19.86%	195.55%	<b>246.57%</b>
FM	Quintiles	7.62%	2.91%	20.86%	54.20%	312.72%	624.71%
	Clusters	2.63%	2.72%	39.39%	37.84%	201.39%	580.19%
RFM	Quintiles	23.25%	3.42%	347.02%	63.33%	781.87%	712.43%
	Clusters	3.65%	2.20%	22.73%	29.79%	183.07%	320.40%

### 3. Análisis de resultados

Se obtuvieron cinco segmentos, las características de los cuales se presentan en la Tabla 3.

**Tabla 3:** Caracterización de Segmentos

# Cluster	Nombre	R	F	M (UF)	Trans	Antigüedad	Max Inactividad	R/F
1	Leales	1.30	2.80	1.40	671.84	386.59	17.94	0.65
2	Estables	13.68	11.18	3.20	70.315	367.26	46.67	2.53
3	En tránsito a fugarse	34.68	15.65	2.71	40.729	344.24	54.08	4.38
4	En riesgo	64.62	19.59	2.84	26.723	333.41	57.08	6.00
5	Con alta probabilidad de fugarse	99.64	21.41	2.22	18.528	324.76	56.37	7.66

En el primer cluster están los clientes con mejor R y así sucesivamente hasta llegar al 5 que es el con peor. La frecuencia está directamente relacionada con el R. Los clientes leales poseen menor monto promedio por transacción, el grupo 2 es el con mejor y el resto son similares. Los clientes del primer grupo realizan muchas más transacciones que los otros grupos.

La antigüedad no es una variable que discrimine entre los diferentes grupos. Los mejores grupos poseen mejor valor del atributo Max Inactividad, lo mismo sucede con el R/F.

En la tabla 4 se presentan las características de los fugados.

**Tabla 4:** Características de fugados

	R	F	M (UF)	Trans	Antigüedad	Inactividad	R/F
Prom	108.24	18.26	2.31	66.96	347.49	45.26	42.27
Desvest	8.48	6.16	0.35	32.52	125.11	15.51	24.26
Coef. De variación	0.08	0.34	0.15	0.49	0.36	0.34	0.57

Los fugados poseen un R promedio mayor que 100, un F de 18 y un M de 2,31 UF, se encuentran mejor representados por sus características R y M, dado los valores del coeficiente de variación pequeños. El coeficiente de variación es alto para las variables transacciones y R/F, se concluye que estas variables no son características de los fugados, ya que aunque el R/F sirve para establecer el criterio de fuga, los clientes fugados pueden tener cualquier valor de R/F siempre que sea mayor que 20, por lo tanto los valores pueden ser muy dispersos.

La Figura 2 muestra las probabilidades de transición. La probabilidad de quedarse en el estado de Fuga es 1, puesto que si una persona cae ahí, quedará fugada para siempre. El estado más estable es el 1, los demás son solo de paso. El 5 es el que tiene mayor probabilidad de fuga y el 1 es el que posee la menor.

$$P = \begin{bmatrix} 0,86 & 0,09 & 0,03 & 0,02 & 0,00 & 0,01 \\ 0,36 & 0,32 & 0,14 & 0,14 & 0,01 & 0,03 \\ 0,18 & 0,24 & 0,15 & 0,04 & 0,29 & 0,09 \\ 0,09 & 0,16 & 0,12 & 0,04 & 0,14 & 0,44 \\ 0,04 & 0,10 & 0,09 & 0,03 & 0,00 & 0,74 \\ 0,00 & 0,00 & 0,00 & 0,00 & 0,00 & 1,00 \end{bmatrix}$$

**Figura 2:** Matriz de probabilidades de transición

Con el objetivo de conocer el camino de los clientes hacia la fuga se hizo un análisis de dos períodos (Tabla 5). Cada celda representa el porcentaje de clientes que se va a fuga, luego de haber realizado cualquier tipo de transición. La notación NA significa que no existió la primera transición. El 0.00% significa que si existió la primera transición, pero que luego 0% de esos se fueron a fuga.

**Tabla 5:** Análisis dos períodos fuga

Estados Iniciales	Estados Intermedios				
	1	2	3	4	5
1	2.18%	6.44%	20.95%	53.55%	NA
2	0.81%	1.18%	4.09%	49.73%	63.76%
3	0.18%	0.51%	1.38%	12.75%	73.96%
4	0.00%	0.00%	0.66%	9.06%	77.63%
5	0.00%	0.00%	0.00%	11.53%	NA

Un alto porcentaje de los clientes que hacen transiciones radicales (como del 1 al 4), al siguiente estado se fugan. La mayoría de los clientes que transita al estado 5 al siguiente período se va a fuga. Además se observa que la mayoría de estas probabilidades son pequeñas, lo que indica que hay clientes no se van rápidamente a la fuga, pudiendo realizar sobre estos acciones enfocadas a retenerlos.

#### 4. Conclusiones

Se generó un procedimiento replicable para predecir fuga de clientes. Para poder utilizarlo y obtener buenos resultados basta con establecer un buen criterio de fuga que sea aplicable al caso específico.

Se pudo obtener la matriz de probabilidades de transición, dentro de lo cual se encuentran las probabilidades de fuga.

El camino a la fuga de los clientes se puede dividir en dos grandes tipos, el de aquellos que se van rápidamente a la fuga y el de los que se van acercando lentamente a ella. De esta manera, se pueden detectar a los clientes antes que se vayan, pudiendo así realizar acciones enfocadas a retenerlos.

#### 5. Referencias

- [1] Briones D. 2002. Modelo de predicción de fuga de clientes para Banco Estado. Tesis Ingeniería Civil Industrial. Universidad de Chile.
- [2] Fayyad, U., Piatestky-Shapiro, G., Smyth, P. 1996. From data mining to knowledge discovery in databases. American association for artificial intelligence 0738-4602, 37-54.
- [3] GrosPELLIER D. 2002. Modelo para estimar el valor de vida de los clientes de un banco. Tesis Ingeniería Civil Industrial. Universidad de Chile.
- [4] Marker, J. 1998. Studying Policy Retention Rates Using Markov Chains. Casualty actuarial society v85.

- [5] Pfeifer, P., Carraway, R. 2000. Modeling Customer Relationships as Markov Chains. *Journal of interactive marketing* v14, no. 2:43-55.
- [6] Reicheld, F y Sasser, W. 1990. Zero defections: Quality comes to service. *Harvard Business Review*, V: Septiembre – Octubre 1990. Pp: 250-258.
- [7] Ross, S. 1996. *Stochastic Processes*. University of California, Berkley, segunda edición. Pp: 163-230.
- [8] Two Crows Corporation. 1999. *Introduction to Data Mining and Knowledge Discovery*, tercera edición <http://www.twocrows.com>
- [9] Weber, R. 2004. *Apuntes IN60E: Aplicación de minería de datos en la empresa*.