EGADE Business School
Tecnológico de Monterrey

Aplicaciones de analítica de datos a los negocios II

PROF: JUAN C. BUSTAMANTE

JUCBUSTAM@TEC.MX

Normas para la conexión síncrona:

1. La clases tiene un back-up garantizado (Grabación disponible en la nube de Zoom).
2. Ingresar a la clase con la cámara del equipo de computo encendida.
3. La **cámara deberá permanecer encendida a lo largo de la clase**.
4. Al ingresar a la clase deben silenciar el micrófono del equipo de computo.
5. Levantar la mano es una opción cuando se quiere preguntar algo durante la sesión de clase, pero les recomiendo que mejor hagamos uso intensivo del chat del canal general para hacer preguntas.
6. En caso de necesitar hacer una pregunta, puede interrumpir la clase sin problema, activando el micrófono de vuestro equipo de computo, luego de la pregunta desactívelo nuevamente.
7. Para una buena clase online es indispensable **el debate, así que foméntelo!!!**.
8. Toda la información se gestiona en CANVAS LMS.

EGADE Business School
Tecnológico de Monterrey

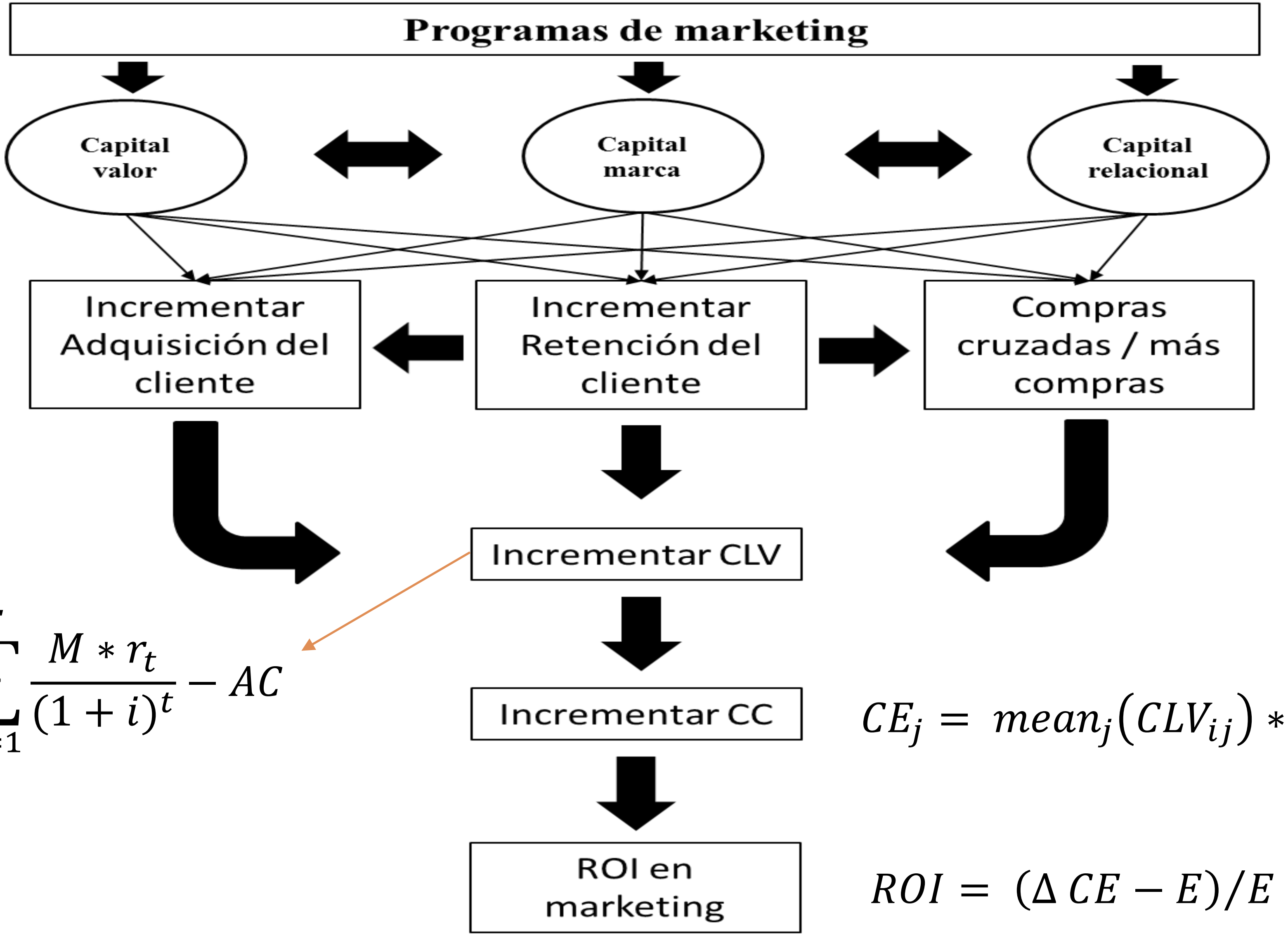
Cronograma de trabajo:

Sesiones	Contenidos	Actividad		Fecha
1	Información general del curso	Utility of classification algorithms		Martes 18/04
2	Algoritmo de regresión logística	Ejecutar script	**Solución caso:** Retention modelling at Scholastic Travel Company (A) and (B)	Martes 25/04
3	Algoritmo Naïve Bayes	Ejecutar script		Martes 02/05
4	Algoritmo k-nearest-neighbors (KNN)	Ejecutar script		Martes 09/05
5	Algoritmo Support vector machine	Ejecutar script		Martes 16/05
6	Algoritmo Decision Trees	Ejecutar script		Martes 23/05
7	Algoritmo Random Forest	Ejecutar script		Martes 30/05
8	Modelo RFM	Ejecutar script	**Solución caso**: CD Now	Martes 06/06
9	Modelo valor de vida del cliente (I)	Ejecutar script		Martes 13/06
10	Modelo valor de vida del cliente (II)	Ejecutar script		Martes 20/06
11	Análisis de series de tiempo	Ejecutar script		Martes 27/06
12	**Proyecto final**	Presentación en equipos		Martes 04/07
	Evaluación final			

1

〈 Valor de vida del cliente (CLV)

(1) Modelo virtuoso del ROI en marketing



$$\boldsymbol{CLV} = \sum_{t=1}^{T} \frac{M * r_t}{(1+i)^t} - AC$$

$$CE_j = mean_j(CLV_{ij}) * POP$$

$$ROI = (\Delta\, CE - E)/E$$

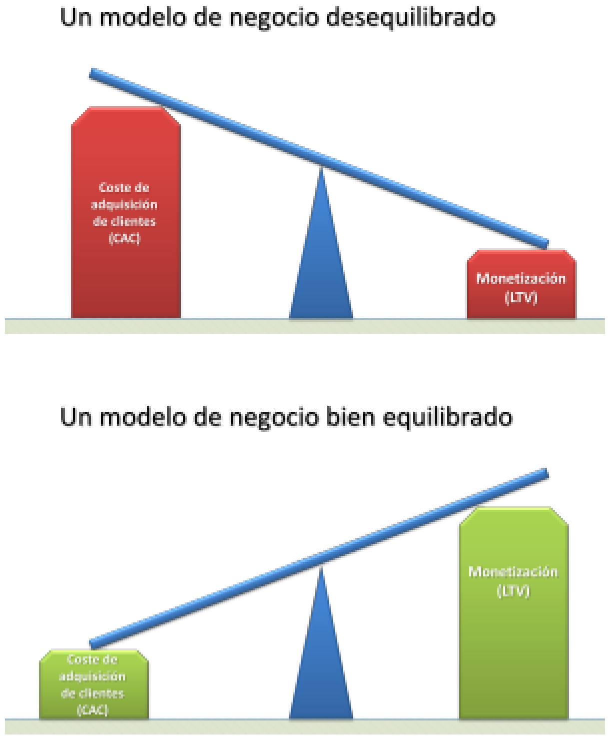
(2) Construcción del modelo virtuoso del ROI en marketing

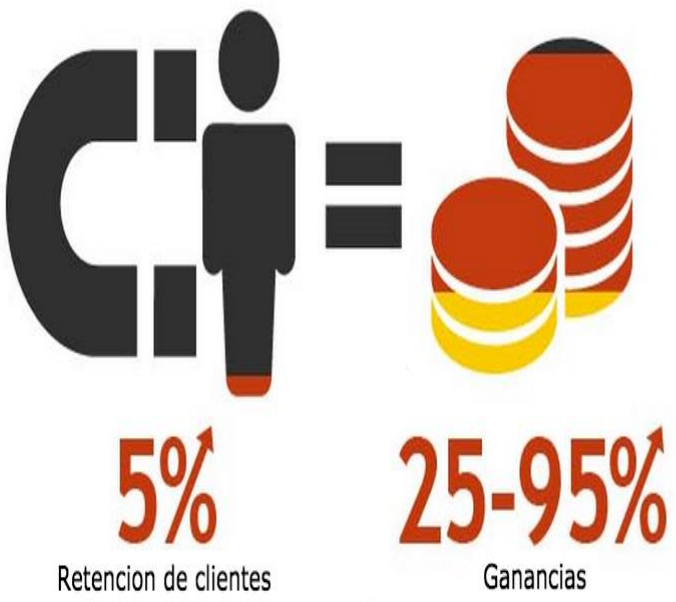
Adquisición de clientes.

Cross Selling

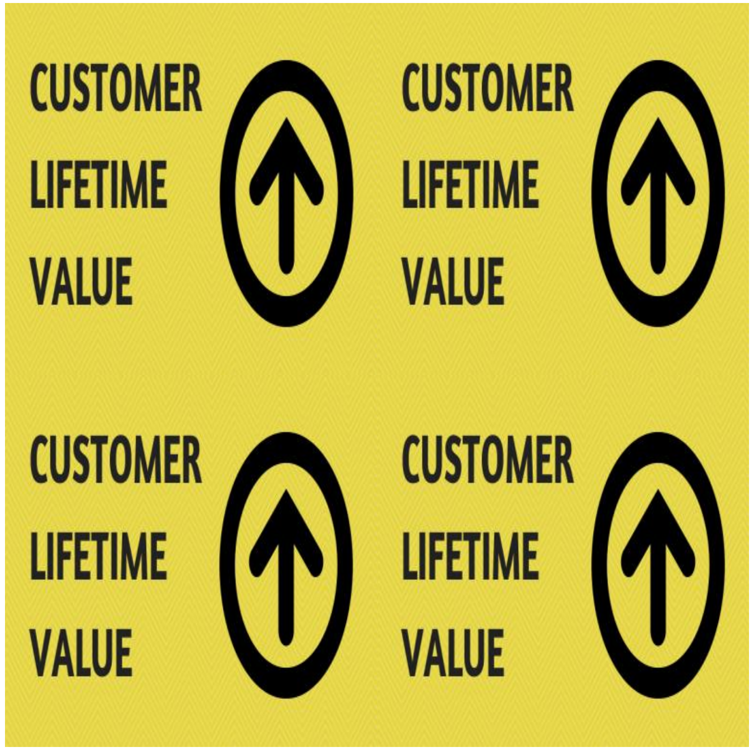
Up Selling

Retención de clientes.





(2) Construcción del modelo virtuoso del ROI en marketing (II).



Valor actual de los futuros beneficios aportados por el **cliente** durante su relación con la empresa. Se trata, por tanto, de valorar a cada cliente en función de su **aportación al margen neto con sus compras**, la frecuencia con la que realiza dichas compras, la fidelidad prevista representada por la **tasa de retención** y los costos generados en la gestión de la relación con el consumidor.

$$\boldsymbol{CLV} = \sum_{t=1}^{T} \frac{M * r_t}{(1+i)^t} - AC$$

(3) Cómo obtener el CLV.

1. Lo primero que se tiene que calcular es la tasa de descuento a partir del tipo de interés previsto.

   La tasa de descuento se calcula con la fórmula: $D = (1+i)^{\wedge n}$

2. Se tiene que calcular los ingresos de los clientes en los 5 años.

   a. Clientes: número de clientes de cada año. Los del año 1 son un dato de arranque. A partir del año 2 se calculan en función de la tasa de retención y referencia.

   b. Tasa de retención: porcentaje de clientes que continúan comprando de un año a otro.

   c. Ventas anuales promedio: Se calcula multiplicando el precio promedio por unidad por el promedio de unidades vendidas por cliente

   d. Ingresos totales: Se calcula multiplicando el número de clientes de cada año por las ventas anuales promedio.

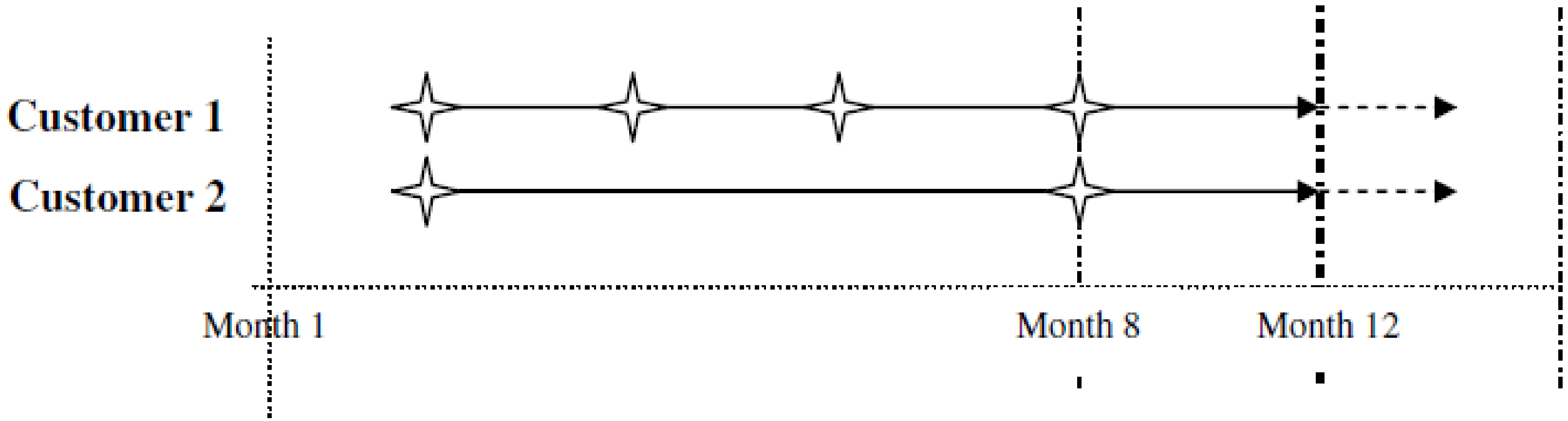
3. Se calculan los costos.

   a. % de costos: Porcentaje de costos que implica la venta y/o la operación del negocio. Es este caso se estiman en un 5%.

   b. Se obtiene multiplicando los ingresos totales por el porcentaje de costos.

4. Se calculan el beneficio y el valor de vida del cliente.

(3) Cómo obtener el CLV (Individual approach)



1. Lo primero es calcular sí un cliente está activo. P (Activo).

$$\text{P (Activo)} = \left(\frac{T}{N}\right)^{n}$$

2. Estimar el CLV:

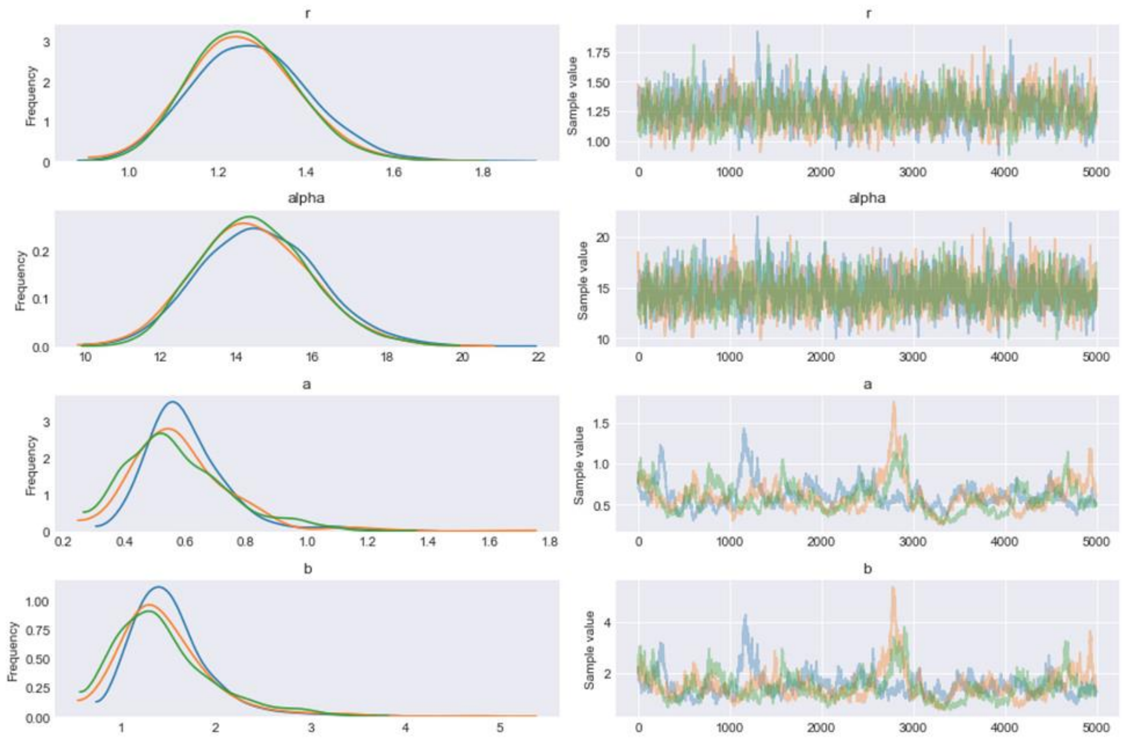
$$CLV_{it} = \sum_{n=t+1}^{t+x} P_{in}\, x\, \frac{MCP}{(1+d)^n} - \sum_{n=1}^{x} M_{in} x \left(\frac{1}{1+d}\right)^{n-1} - A$$

2

Modelos avanzados

El Modelo de Pareto/NBD.

El modelo de Pareto/NBD, también conocido como SMC, analiza el comportamiento del consumidor en base a su consumo reiterado, en un contexto no contractual. Del análisis se desprenden la probabilidad condicionada a su histórico de compras, de que un cliente permanezca activo, así como el número de transacciones esperadas para un cliente aleatorio.



El Modelo de Pareto/NBD (Supuestos básicos)

(1) Mientras está activo, el número de transacciones realizadas por un cliente en un periodo de tiempo t, se distribuye como una Poisson con media λt.

(2) Por lo tanto, la probabilidad de observar x transacciones en un intervalo de tiempo (0, t] se muestra como

$$P(X(t) = x|\lambda) = \frac{(\lambda t)^x e^{-\lambda t}}{x!}, \qquad x = 0, 1, 2 \ldots$$

Esto es equivalente a suponer que el intervalo de tiempo entre cada transacción se distribuye como una exponencial con tasa de transacción λ,

$$f\left(t_j - t_{j-1}|\lambda\right) = \lambda e^{-\lambda(t_j - t_{j-1})}, \qquad t_j > t_{j-1} > 0,$$

donde t_j es el momento de la compra j.

El Modelo de Pareto/NBD (Supuestos básicos)

(3) La heterogeneidad en la tasa de transacción de los clientes λ, sigue una distribución Gamma con forma (*shape parameter*) r y escala (*scale parameter*) α.

$$g(\lambda|r,\alpha) = \frac{\alpha^r \lambda^{r-1} e^{-\lambda\alpha}}{\Gamma(r)}$$

(4) Cada cliente tiene un ciclo de vida indeterminado (*lifetime*) de longitud τ, tras el que se considerará inactivo. El momento en el que el cliente se transforma en inactivo se distribuye como una exponencial con tasa de abandono μ.

$$f(\Gamma|\mu) = \mu e^{-\mu\Gamma}$$

El Modelo de Pareto/NBD (Supuestos básicos)

(5) La heterogeneidad en las tasas de abandono de los clientes sigue una distribución Gamma con forma s y escala β.

$$g(\mu\,|s) = \frac{\beta^s \mu^{s-1} e^{-\mu\beta}}{\Gamma(r)}$$

(6) La tasa de transacción λ y de abandono μ, varían independientemente de los consumidores.

Además, tanto el modelo de Pareto/NBD como el modelo BG/NBD sólo requerirán dos datos del histórico de compras de cada cliente: la fecha de su última transacción, parámetro que llamaremos *recency*, y el número de transacciones hechas en un tiempo determinado, definido como *frecuency* (frecuencia de compra).

El Modelo de Pareto/NBD (Notación)

La notación será $(X = x, t_x, T)$ donde x es el número de transacciones observadas en el periodo $(0, T]$ y $t_x (0 < t_x \leq T)$ es el tiempo de la última transacción.

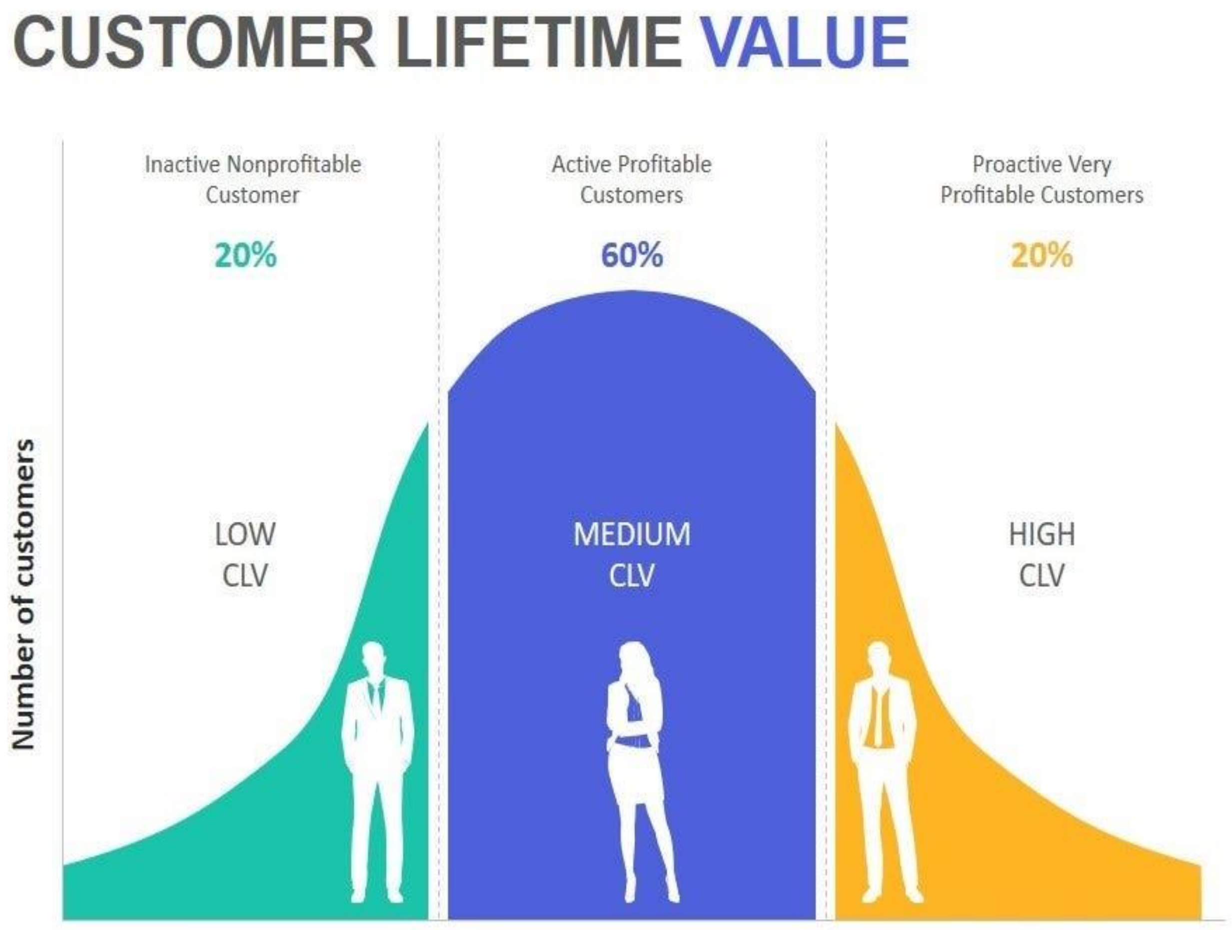
Usando estas dos estadísticas agregadas clave, con el modelo SMC podemos extraer conclusiones económicas para un número significativo de KPIs (*Key Performance Indicators*) tales como:

- $E[X(t)]$ como las compras esperadas en un periodo de tiempo de longitud t para lo que es necesario calcular el volumen de transacciones esperadas para la base de clientes a lo largo del tiempo.
- $P(X(t) = x)$ la probabilidad de observar x transacciones en un periodo de tiempo de longitud t.
- $E[Y(t)|X = x, t_x, T]$ es el número esperado de transacciones en el periodo $(T, T + t]$ para un cliente con comportamiento determinado (x, t_x, T).

El Modelo BG/NBD:

Este modelo está basado en el Modelo de Pareto/NBD con la única diferencia de cómo y cuándo determinar si un cliente está inactivo. Básicamente es un modelo estocástico de Comportamiento del comprador para entornos no contractuales de tiempo discreto.

Por tanto, el modelo BG/BB requiere la misma información que el modelo Pareto/NBD, pero como modela discretamente oportunidades de transacción, esta información se puede condensar en una matriz recency-frequency.



El Modelo BG/NBD (Supuestos básicos):

(1) El número de transacciones realizadas por un cliente mientras está activo en un periodo de tiempo t, se distribuye como una Poisson con media λt.
Por lo tanto, la probabilidad de observar x transacciones en un intervalo de tiempo (0, t] se muestra como

$$P(X(t) = x|\lambda) = \frac{(\lambda t)^x e^{-\lambda t}}{x!}, \qquad x = 0, 1, 2 \ldots$$

Esto es equivalente a suponer que el intervalo de tiempo entre cada transacción se distribuye como una exponencial con tasa de transacción λ, por ejemplo,

$$f(t_j - t_{j-1}|\lambda) = \lambda e^{-\lambda(t_j - t_{j-1})}, \qquad t_j > t_{j-1} > 0,$$

donde t_j es el momento de la compra j.

(2) La heterogeneidad en la tasa de transacción de los clientes, λ, sigue una distribución gamma con forma r y escala α.

$$f(\lambda|r, \alpha) = \frac{\alpha^r \lambda^{r-1} e^{-\lambda\alpha}}{\Gamma(r)}, \qquad \lambda > 0$$

El Modelo BG/NBD (Supuestos básicos):

(3) Después de una transacción cualquiera, un cliente se convierte en inactivo con probabilidad p. Por lo tanto, el punto en el que un cliente abandona se distribuye a lo largo de varias transacciones de acuerdo a una distribución geométrica ("desplazada") cuya función de probabilidad será

$$P(cliente\ inactivo\ inmediatamente\ después\ de\ la\ compra\ j) \quad = p(1-p)^{j-1},$$

$$j = 1, 2, 3 \ldots$$

(4) La heterogeneidad en p sigue una distribución de densidad beta de la forma

$$f(\,p|\,a, b) = \frac{p^{a-1}(1-p)^{b-1}}{B(a,b)}\ , \qquad 0 \leq p \leq 1$$

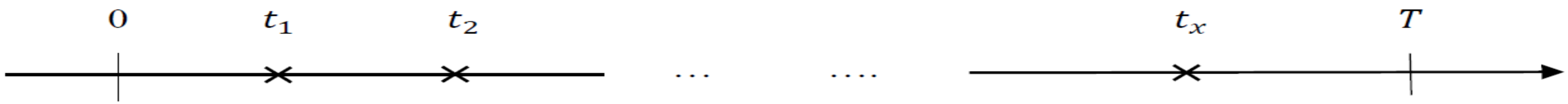
donde $B(a, b)$ es la función Beta que puede expresarse en términos de funciones Gamma:

$$B(a, b) = \frac{\Gamma(a) + \Gamma(b)}{\Gamma(a + b)}$$

(5) La tasa de transacción λ, y la probabilidad de abandono p varía con independencia de los clientes.

Desarrollo de un modelo BG/NBD (Supuestos básicos):

Consideremos un cliente que ha realizado x transacciones en el periodo $(0, T]$ ocurridas en diferentes momentos $t_1, t_2, \dots, t_x$:



Derivaremos la función de probabilidad a nivel individual de la siguiente manera:

- La probabilidad de que la primera transacción ocurra en el momento t_1 tiene una probabilidad exponencial estándar igual a $\lambda e^{-\lambda t_1}$.
- La probabilidad de que la segunda transacción ocurra en el momento t_2 tiene una probabilidad exponencial estándar igual a $(1-p)\lambda e^{-\lambda(t_2 - t_1)}$.

Esto continúa para las siguientes transacciones hasta que:

- La probabilidad de que la transacción x ocurra en el momento t_x es la probabilidad de permanecer activo en el momento t_{x-1} que sería igual a

$$(1-p)\lambda e^{-\lambda(t_x - t_{x-1})}.$$

- Por último, la probabilidad de no observar ninguna compra en $(t_x, T]$ es la probabilidad de que el cliente permanezca inactivo en t_x más la probabilidad de que permanezca activo pero que no haya hecho compras en ese intervalo, que es igual a

$$p + (1-p)e^{-\lambda(T - t_x)}$$

Desarrollo de un modelo BG/NBD (Supuestos básicos):

Derivación de P (X (t)=x).

Considerando las variables aleatorias $X(t)$ como el número de transacciones que ocurren en un periodo de tiempo de longitud t (empezando en el tiempo 0), y T_x como el tiempo de la transacción x, tenemos que $X(t) \geq x \Leftrightarrow T_x \leq t$.

Derivando una expresión como $P(X(t) = x)$, subrayaremos la relación existente entre el número de eventos y los periodos entre dichos eventos.
Entonces,

$$P(X(t) = x) = P(cliente\ activo\ después\ de\ la\ compra\ x)\ .\ P(T_x \leq t\ y\ T_{x+1} > t) + \delta_{x>0}$$
$$.P(cliente\ inactivo\ después\ de\ la\ compra\ x)\ .\ P(T_x \leq t)$$

Bajo el supuesto de que el tiempo entre las transacciones se distribuye como una exponencial $P(T_x \leq t\ y\ T_{x+1} > t)$ que es simplemente la probabilidad de Poisson en la que $X(t) = x$ y $P(T_x \leq t)$. Por tanto,

$$P(X(t) = x|\lambda, p) = (1-p)^x \frac{(\lambda t)^x e^{-\lambda t}}{x!} + \delta_{x>0}\, p(1-p)^{x-1} \left[1 - e^{-\lambda t} \sum_{j=0}^{x-1} \frac{(\lambda t)^j}{j!}\right]$$

Desarrollo de un modelo BG/NBD (Supuestos básicos):

Todas las expresiones desarrolladas debajo están condicionadas a la tasa de transacción λ, y la probabilidad de abandono p, ambas son no observadas.

Para derivar las expresiones equivalentes para un cliente elegido aleatoriamente, tomamos los resultados esperados a nivel individual por encima de las distribuciones combinadas para λ y p, como vemos en (1) y (2). Esto nos da los siguientes resultados.

Tomando el valor esperado de (3) por encima de la distribución de λ y p se obtiene la siguiente expresión de la función de probabilidad para un cliente aleatorio con histórico de compras $(X = x, t_x, T)$:

$$\begin{aligned} &\mathrm{P}(\text{alive at } n+1 \mid \alpha, \beta, \gamma, \delta, x, t_x, n) \\ &\quad = \frac{\mathrm{B}(\alpha+x, \beta+n-x)}{\mathrm{B}(\alpha, \beta)} \frac{\mathrm{B}(\gamma, \delta+n+1)}{\mathrm{B}(\gamma, \delta)} \\ &\qquad \cdot \mathrm{L}(\alpha, \beta, \gamma, \delta \mid x, t_x, n)^{-1}. \end{aligned}$$

Los cuatro parámetros del modelo BG/NBD ($\alpha, \beta, \gamma, \delta$) pueden estimarse por el método de máxima verosimilitud de la siguiente manera.

Bibliotecas

Pandas	Manipulación y análisis de datos
NumPy	Creación de vectores y matrices Colección de funciones matemáticas
matplotlib	Generación de gráficos
seaborn	Biblioteca de visualización basada en Matplotlib
scikit learn	Biblioteca de aprendizaje automático
LIFETIMES	Biblioteca probabilística para estimar “alive” o “die”

Dataset Retail
e-commerce