Métodos analíticos y modelos de predicción aplicados a la segmentación de clientes por valor

EXPERIENCIAS PRACTICAS

TAK NG

Definición del Customer Lifetime Value (CLV)

El CLV se puede definir como el valor presente de flujo de dinero al futuro, producto de la relación que una empresa tiene con el cliente

Valor del cliente (modelos probabilísticos)

$$CLV = \sum_{t=1}^{x} \frac{Valor\ esperado\ del\ Cliente\ en\ periodo\ t}{(1 + tasa\ de\ descuento)^{t-1}}$$

(Blattberg et al., 2009).

CLV = \sum_{t=1}^x\frac{Valor\ esperado\ del\ Cliente\ en\ periodo\ t}{(1+tasa\ de\ descuento)^{t-1}}

- El valor esperado puede calcularse desde una forma muy simple, o puede convertirse en un proyectos de varios meses
- El CLV se puede estimar a ***x*** períodos arbitrarios al futuro.
- Por lo general, cada período ***t*** es un mes
- La tasa de descuento se usa para convertir flujo de dinero futuro a valor presente

Modelo de anualidad

(Malthouse, 2013)

Imagina que una empresa vende un solo servicio. Pagas RD$ 200 al mes, por un período de un año y no se permite cancelar el servicio hasta cumplir un año.

Cada servicio tiene costo promedio para la empresa de RD$ 150 mensuales. La tasa de descuento mensual es de 0.01. Calcular el CLV a 12 meses

CLV = \sum_{t=1}^{T}\frac{valor\ esperado\ en\ periodo\ t}{(1+tasa\ descuento)^{t}} = valor\ esperado * \frac{1-(1+tasa\ descuento)^{-T}}{tasa\ descuento}

$$CLV = \sum_{t=1}^{T} \frac{valor\ esperado\ en\ periodo\ t}{(1 + tasa\ descuento)^{t}} = valor\ esperado * \frac{1 - (1 + tasa\ descuento)^{-T}}{tasa\ descuento}$$

$$CLV = (200 - 150) * \frac{1 - (1 + 0.01)^{-12}}{0.01} = 562.75$$

Modelo simple de retención

(Malthouse, 2013)

La misma empresa del ejemplo anterior tiene una probabilidad de retención mensual de 98%.

CLV = \sum_{t=1}^{T}\frac{valor\ esperado\ en\ periodo\ t * prob\ retencion^t }{(1+tasa\ descuento)^{t}}

$$CLV = \sum_{t=1}^{T} \frac{valor\ en\ periodo\ t * prob\ retencion^t}{(1 + tasa\ descuento)^t}$$

t	Valor esperado	Prob retención^t	CLV
1	$ 50.00	0.98	$ 48.51
2	$ 50.00	0.96	$ 47.07
3	$ 50.00	0.94	$ 45.68
4	$ 50.00	0.92	$ 44.32
5	$ 50.00	0.90	$ 43.00
6	$ 50.00	0.89	$ 41.73
7	$ 50.00	0.87	$ 40.49
8	$ 50.00	0.85	$ 39.28
9	$ 50.00	0.83	$ 38.12
10	$ 50.00	0.82	$ 36.98
11	$ 50.00	0.80	$ 35.89
12	$ 50.00	0.78	$ 34.82
			$ 495.89

Qué pasaría si la probabilidad de retención es diferente para cada mes?

Análisis de Supervivencia

Análisis de supervivencia es una rama de la estadística usada para estudiar la duración esperada hasta que algún evento ocurra (Survival Analysis, 2018).

Este es muy usado en medicina, donde se usa para analizar el momento transcurrido hasta que los pacientes de un estudio muere o tenga algún evento de interés.

Para el CLV, se puede usar análisis de supervivencia para estimar la probabilidad los clientes en seguir con un producto hasta que lo cancele

Conceptos importantes – Análisis de supervivencia

- Variable aleatoria: es una función que retorna un valor relacionado a un experimento aleatorio. En general, es difícil decir con exactitud cuándo ocurre un evento para un individuo en particular, pero se puede describir una distribución de valores.

  Por ejemplo, si tiras un dado, no se puede decir con exactitud qué número va a salir en ese momento. Sí es posible decir que la probabilidad que el número 1 tienen probabilidad de 1/6 en ocurrir.

- Tiempo de supervivencia: tiempo que dura hasta que un evento ocurra.

  Por ejemplo, en estudio de un grupo de pacientes con cáncer, sería el tiempo que dure estos hasta que mueran, luego de aplicar un tratamiento nuevo. El propósito sería medir si el nuevo tratamiento mejora el tiempo de supervivencia de los pacientes, cuando se compara a un tratamiento actual o con placebos

Conceptos importantes – Análisis de supervivencia

- T = variable aleatoria de tiempo de supervivencia
- t = valor específico de T
- S(t) = función supervivencia = Pr(T > t)

La función de supervivencia permite estimar la probabilidad de supervivencia

En general, no sabemos de antemano cuál es la distribución que describe la probabilidad de supervivencia. Se estima con observaciones

- $\hat{S}(t) = Estimacion\, de\, S(t)\, usando\, data\, historica\, o\, de\, experimentos$
- Censura (right censoring): pasa cuando el evento no ocurre dentro del tiempo de observación o de experimento

Ejemplo

time	status	x
9	1	Maintained
13	1	Maintained
13	0	Maintained
18	1	Maintained
23	1	Maintained
28	0	Maintained
31	1	Maintained
34	1	Maintained
45	0	Maintained
48	1	Maintained
161	0	Maintained

time	status	x
5	1	Nonmaintained
5	1	Nonmaintained
8	1	Nonmaintained
8	1	Nonmaintained
12	1	Nonmaintained
16	0	Nonmaintained
23	1	Nonmaintained
27	1	Nonmaintained
30	1	Nonmaintained
33	1	Nonmaintained
43	1	Nonmaintained
45	1	Nonmaintained

Fuente de la data: Rupert G. Miller (1997), *Survival Analysis*. John Wiley & Sons. ISBN: 0-471-25218-2

La data fue de un estudio sobre pacientes con leucemia. El experimento consistió en validar si dosis adicionales de quimioterapia mejoraba la supervivencia de los pacientes.

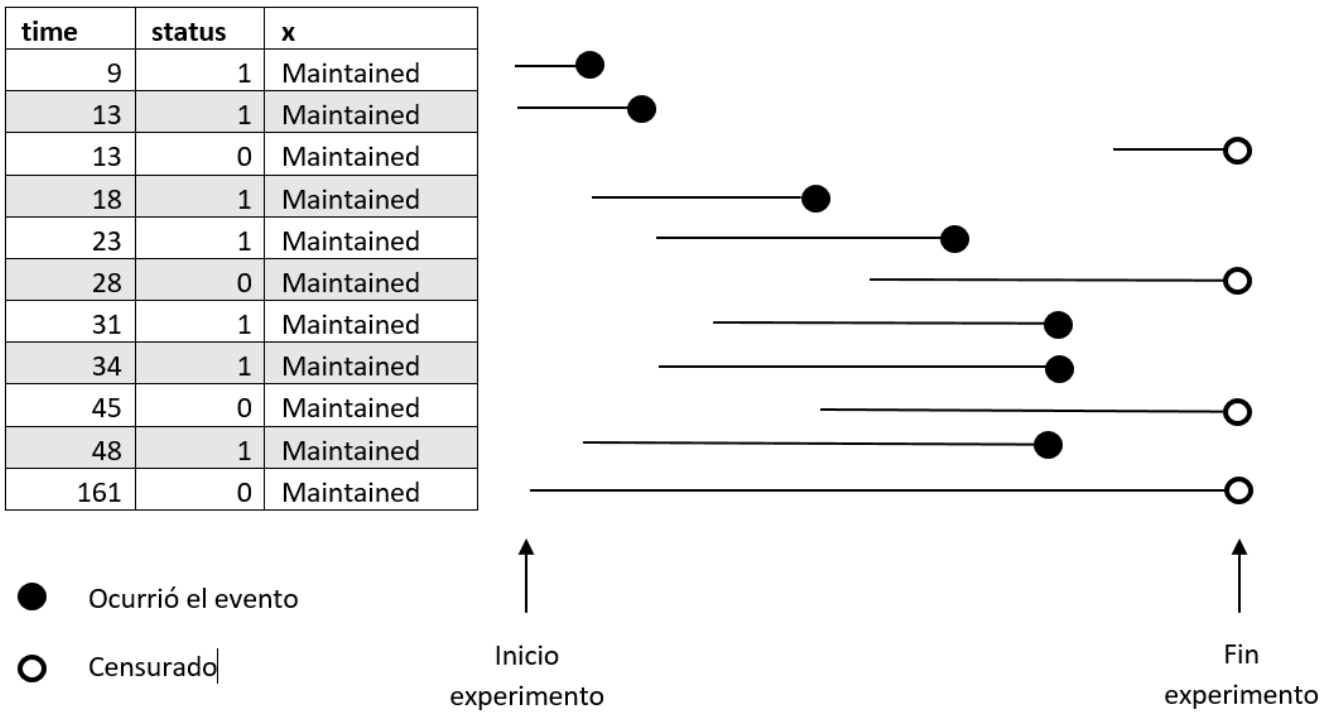
time: es tiempo de supervivencia en semanas

status: 1 = murió; 0 = censurado

x: Maintained = dosis adicional de quimiterapia; Nonmaintained = no dosis adicional

Ilustración de censura

time	status	x
9	1	Maintained
13	1	Maintained
13	0	Maintained
18	1	Maintained
23	1	Maintained
28	0	Maintained
31	1	Maintained
34	1	Maintained
45	0	Maintained
48	1	Maintained
161	0	Maintained



En las censuras, el experimento termina pero no se observa el evento en los individuos

Las censuras se observan para las semanas 13, 28, 45 y 161

Note que no todos los individuos entraron al mismo tiempo al experimento

Estimador Kaplan–Meier

aml$x=Maintained

time	n.risk	n.event	n.event / n.risk	1 - (n.event / n.risk)	survival
9	11	1	0.091	0.909	0.909
13	10	1	0.100	0.900	0.818
18	8	1	0.125	0.875	0.716
23	7	1	0.143	0.857	0.614
31	5	1	0.200	0.800	0.491
34	4	1	0.250	0.750	0.368
48	2	1	0.500	0.500	0.184

aml$x=Nonmaintained

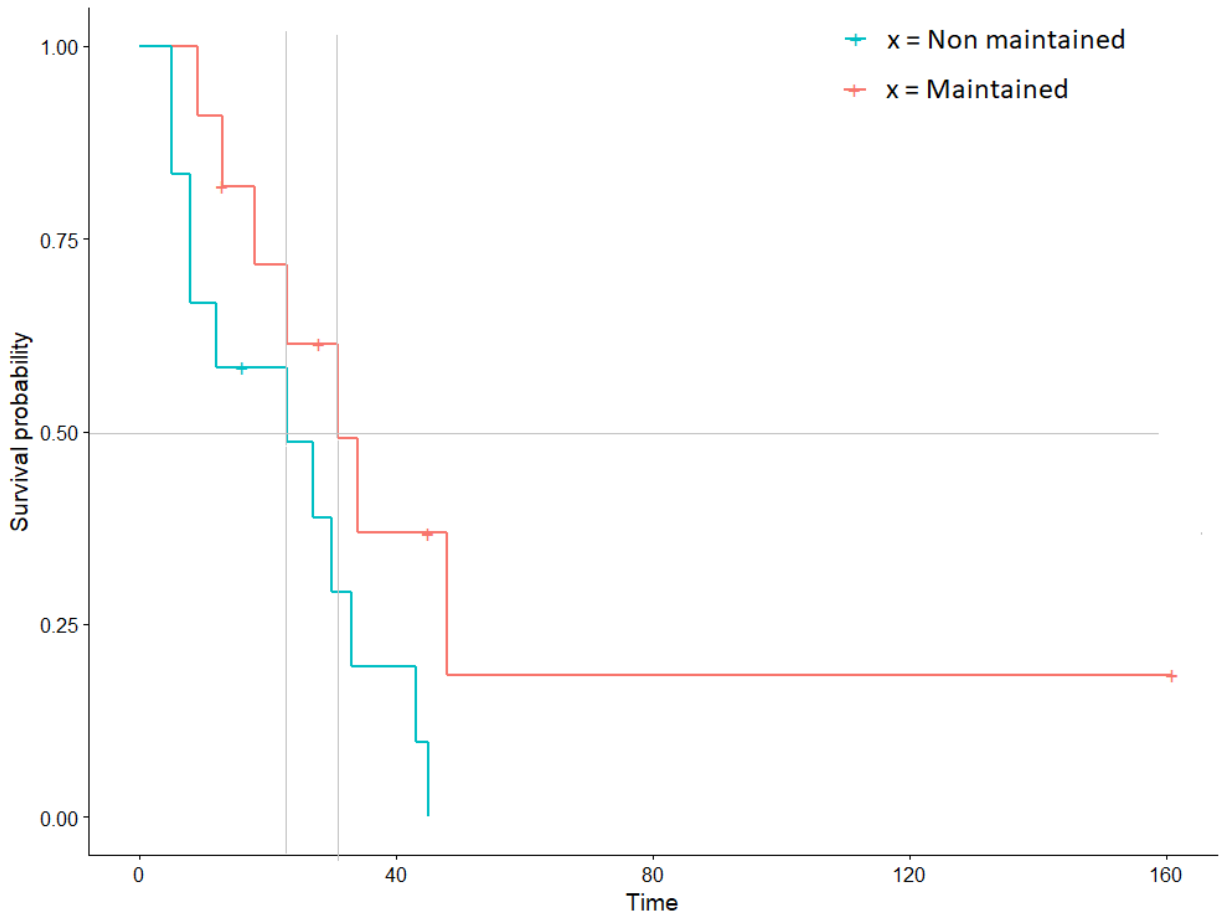
time	n.risk	n.event	n.event / n.risk	1 - (n.event / n.risk)	survival
5	12	2	0.167	0.833	0.833
8	10	2	0.200	0.800	0.667
12	8	1	0.125	0.875	0.583
23	6	1	0.167	0.833	0.486
27	5	1	0.200	0.800	0.389
30	4	1	0.250	0.750	0.292
33	3	1	0.333	0.667	0.194
43	2	1	0.500	0.500	0.097
45	1	1	1	0	0

$$\hat{S}(t) = \prod_{i:\ t_i \leq t} \left(1 - \frac{n.event_i}{n.risk_i}\right)$$

$n.event_i$: cantidad de eventos ocurridos en el tiempo t_i

$n.risk_i$: cantidad de individuos en riesgo en el tiempo t_i

Curva de supervivencia



En el ejemplo de leucemia, podemos graficar el resultado de la función de supervivencia estimada con Kaplan-Meier

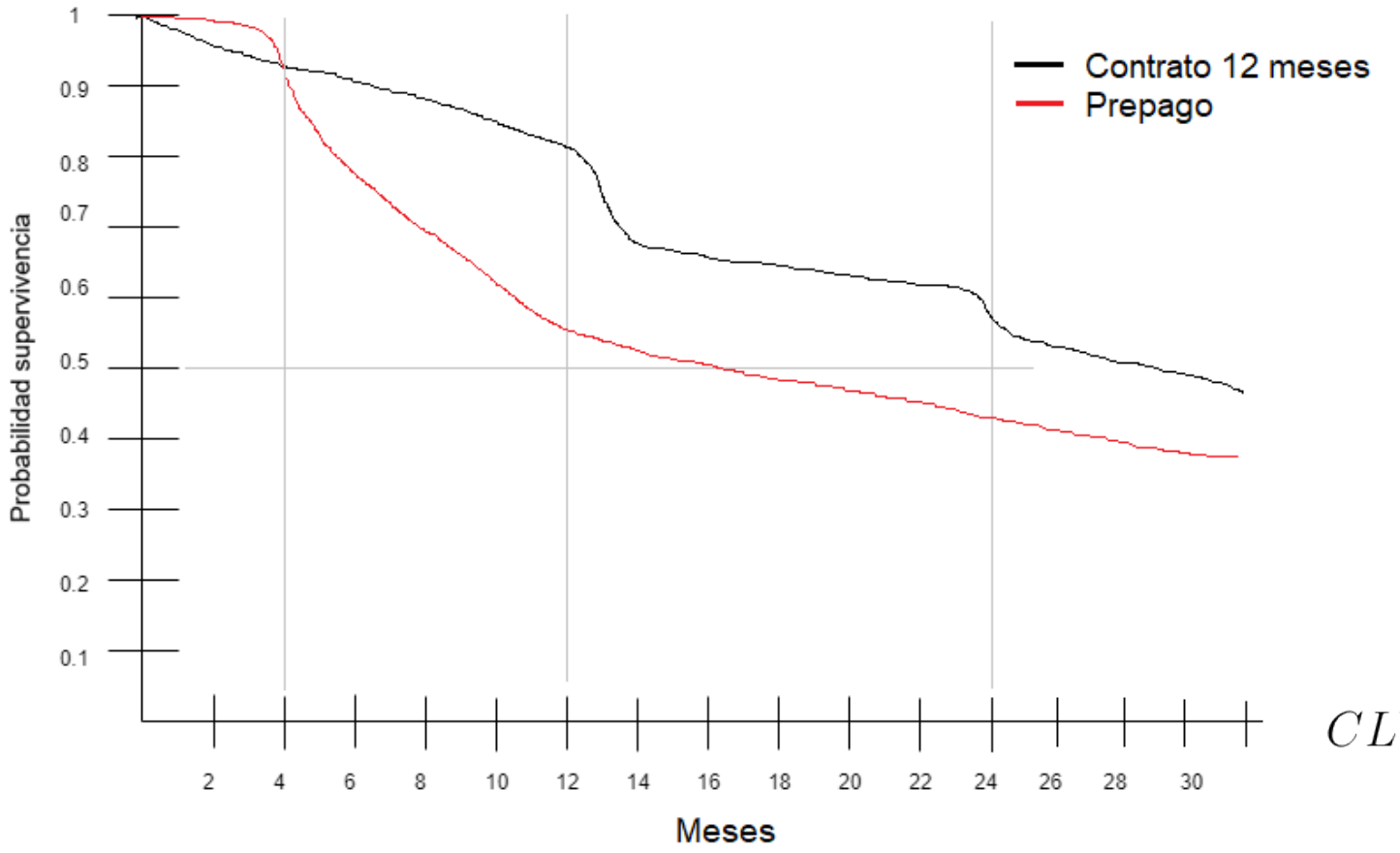
Noten que el 50% de los pacientes sin dosis adicional del tratamiento sobreviven hasta 22 semanas.

El 50% del grupo con tratamiento adicional sobrevive hasta 30 semanas

Recuerden que este resultado es una muestra pequeña, y este no necesariamente es clínicamente significativo

También la forma de escoger la muestra es importante. Por ejemplo, la edad de los pacientes puede influir en la sobrevivencia. Si un grupo es joven y otro es de edad avanzada, puede haber un sesgo

Estimación de probabilidad de retención (ejemplo de telco)



En Contratos de 12 meses, hay más cancelaciones cerca de cada 12 meses, que es el fin de contrato

En Prepago, hay cancelaciones masivas en el mes 4, pues se inactiva un SIM luego de ausencia de recarga y uso. Los primeros 12 meses son críticos, y luego las cancelaciones se suavizan

$$CLV = \sum_{t=1}^{T} \frac{valor\ en\ periodo\ t * prob\ retencion^t}{(1 + tasa\ descuento)^t}$$

Cálculo valor en período t

$$Valor_t = Ingresos_t - Costos_t$$

Ingresos

En general, los principales ingresos están registrado en los sistemas de facturación de la empresa, y están bien segregados. De otra forma, no se cobra al cliente.

Costos

Muchos costos se registran de forma manual, y existen un histórico limitado

No todos los costos se pueden fácilmente distribuir entre los clientes.

Muchas veces, los costos aparecen como un campo agregado en los estados de las empresas, pero es difícil segregarlos por cliente

Ejemplos de ingresos y costos

Industria	Ingreso	Costo relacionados al producto del cliente
Telecomunicaciones	Facturación mensual, recargas (planes prepago), ingreso por interconexión, cargos por mora	Planes de lealtad, descuento por suscripción, descuentos de temporada, instalaciones, costos de aparatos, servicio al cliente (presencial y call center), costo interconexión, mora, mercadeo, comisiones de venta, fraudes, canal de pago
Bancos	Intereses por financiamiento en préstamos y tarjetas, ingresos por servicios, cargos por mora	Programas de lealtad, provisiones, encaje legal, mora, mercadeo, servicio al cliente (presencial y call center), comisiones de venta, fraudes, canal de pago
Seguros	Prima de seguro privado, prima de seguro familiar de salud	Reclamos (se debe de segmentar los reclamos), fraudes, atención al cliente, corredores de seguro, comisión por venta, campañas de prevención, red de prestadores
Retail. **Es posible que no se pueda identicar una parte de los clientes**	Venta de mercancías	Costo de productos (se necesita entender bien). Por ejemplo, los vegetales y productos refrigerados tienen un costos de manejo más alto que los enlatados. Velocidad de rotación de productos, robos, productos dañados Planes de lealtad, ofertas de temporada, servicio al cliente, mercadeo, canal de pago

Estimación de ingresos y costos en períodos futuros

- Ingresos y costos registrados para períodos futuros en los sistemas operacionales. Por ejemplo, tabla de amortización de préstamos, cuotas de seguros, etc.
- Promedio ingresos y costos de últimos N meses

CLV en empresas con múltiples productos

Es muy probable que un cliente tenga más de un producto con una empresa. En este caso, se debe de calcular el Product Lifetime Value para cada producto, y sumar los valores para obtener el CLV del cliente

Las características de los productos pueden ser muy diferentes. Por ejemplo, una tarjeta de crédito vs certificado de depósito

$$PLV = \sum_{t=1}^{T} \frac{valor\ en\ periodo\ t * prob\ retencion^t}{(1 + tasa\ descuento)^t}$$

$$CLV = \sum_{p=1}^{N} PLV_p\ del\ cliente$$

Modelos de Data Mining

Se puede usar modelos de data mining, donde no se asume sobre el valor de los períodos t ni la probabilidad de retención

$$PLV = f(x) + error$$ (Malthouse, 2013)

Donde

PLV - es la variable dependiente que queremos predecir

f(x) - es la función del modelo de data mining

x - son variables independientes disponibles antes del período de predicción

error - se asume como errores independientes y homoscedástico entre las distintas observaciones

Modelos de Data Mining

La elección correcta de variables independientes es crucial para que el modelo dé un resultado aceptable.

Modelos de data mining populares son regresión, random forest, support vector machine, entre otras.

No vamos a abundar en estos modelos en esta charla

Modelos Probabilísticos vs Data Mining

Modelos probabilíticos

Pros

- Son más fáciles de calcular
- Son más transparentes en el cálculo del CLV

Cons

- Las suposiciones de los modelos probabilíticos pueden no ser correctas
- No hay herramientas estándares en la literatura para validar el resultado del modelo

Modelos Data Mining

Pros

- Se toma menos suposiciones
- Existen herramientas estándares para validar y comparar los resultados de diferentes modelos de Data Mining

Cons

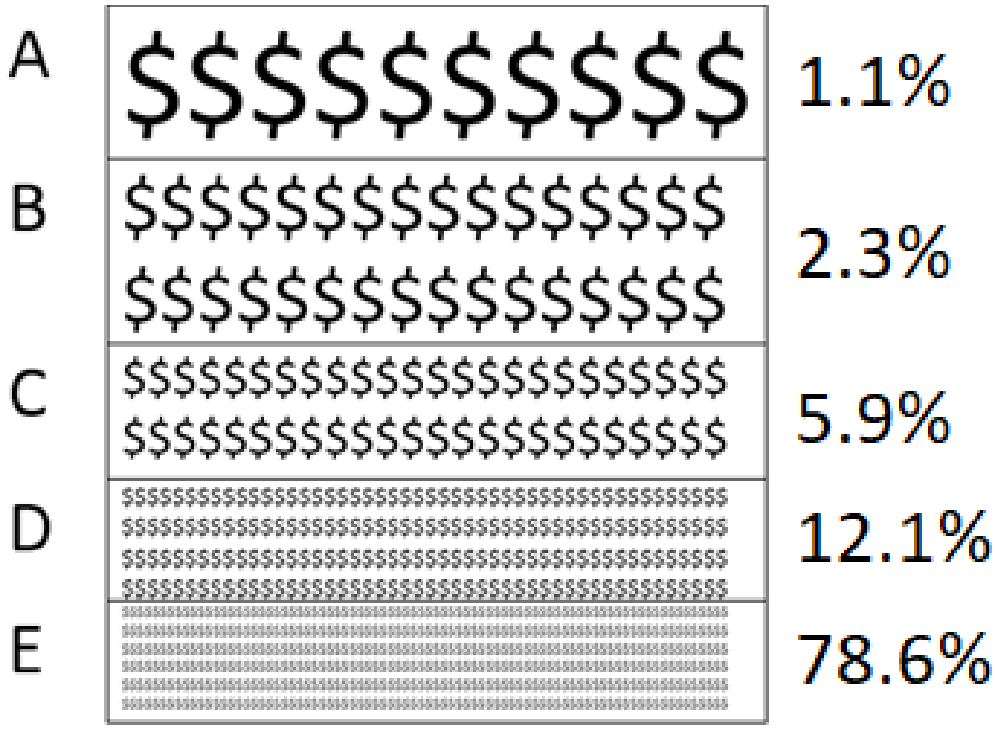
- El cálculo para obtener el resultado NO es transparente, y no es fácil de explicar a negocio
- Hay que trabajar más en la parte de feature engineering
- Puede tomar más tiempo de implementación

Modelos de Migración (Malthouse, 2013)

- En industrias como tiendas online, supermercados, aerolíneas, y tiendas al detalle, el cliente no transacciona de forma regular, y no se captura en los sistemas operacionales cuándo un cliente termina su relación con la empresa. Es más, se dan casos que no se puede identificar al cliente
- Los modelos probabilísticos y de Data Mining no se ajustan muy bien a estos modelos de negocios.
- Los modelos de migración ven al cliente en dos estados: compran o no compran
- No vamos a abundar en modelos de migración

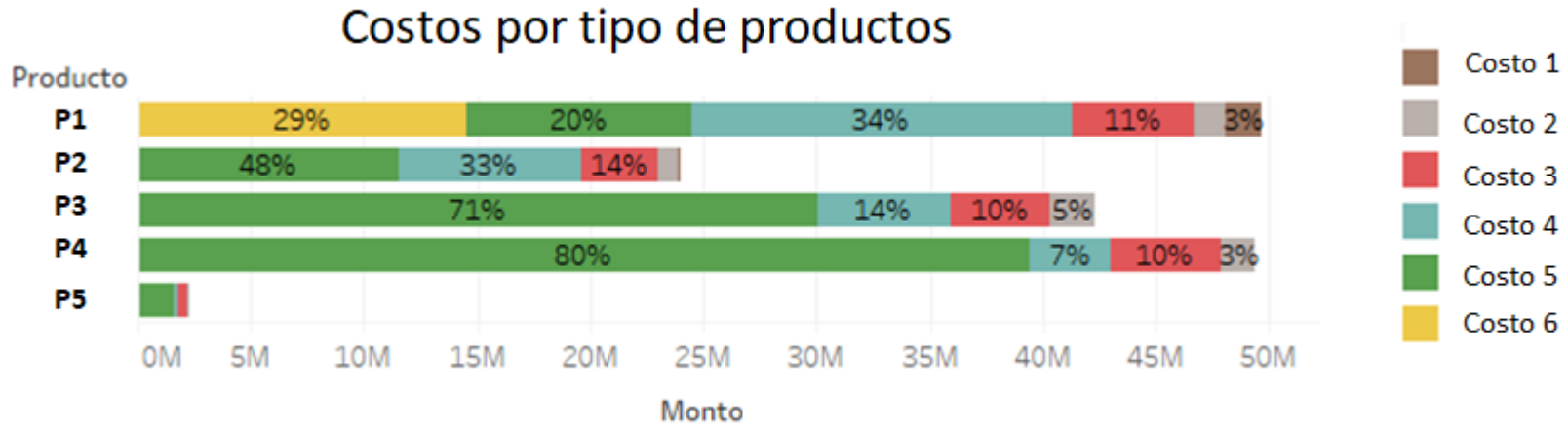
Usos del Customer Lifetime Value

Segmentación de clientes en quintiles de valor



- Cada segmento tiene el 20% del valor de la empresa
- El 3.4% de los clientes concentran el 40% del valor (A y B)
- El 78.6% de los clientes (segmento E) tienen apenas el 20% del valor de la empresa
- Regla del pareto (20% - 80%)
- Permite clasificar los clientes de forma muy sencilla

Entender estructura de ingresos y costos por productos y subproductos



Como el CLV necesita estimar los ingresos y costos, estos mismos se pueden usar para entender los ingresos y costos de los productos

Entender cómo la venta cruzada agrega valor al cliente

Cantidad de productos	Cantidad de clientes	%
1	8,200	82%
2	1,200	12%
3	500	5%
4	100	1%
Total Clientes	10,000	100%

Cantidad de productos	A	B	C	D	E	
1	0.02%	1.00%	3.20%	10.78%	85.00%	100.00%
2	1.00%	2.00%	8.00%	19.00%	70.00%	100.00%
3	5.00%	4.00%	11.00%	20.00%	60.00%	100.00%
4	10.00%	15.00%	20.00%	22.00%	33.00%	100.00%

La mayoría de los clientes tiene 1 producto con la empresa (82%). Hay mucha oportunidad de venta cruzada

De los clientes con 1 producto, el 85% de estos pertenecen al segmento E (generan poco valor)

A medida que los clientes adquieren más productos, el valor de estos incrementan

Medición del valor de los clientes en el tiempo

		Octubre 2018						
		A	**B**	**C**	**D**	**E**	**No es cliente**	**Cliente nuevo**
Julio 2018	**A**	**80%**	7%	5%	3%	2%	2%	1%
	B	2%	**85%**	4%	2%	2%	1%	2%
	C	1%	2%	**90%**	3%	2%	2%	1%
	D	0.05%	1%	2%	**92%**	2%	4%	1%
	E	0%	1%	2%	3%	**76%**	8%	10%

Los clientes del segmento A están perdiendo valor rápido. Incluso, se van de la empresa

Los clientes de los segmentos D y E tienden a seguir siendo clientes de bajo valor.

En el segmento E, el 8% de los clientes se fueron, y se agregó un 10% de clientes a ese segmento. Probablemente, la empresa está haciendo mucho esfuerzo para traer nuevos clientes de bajo valor, y estos se van en poco tiempo. Solo hay objetivos de venta de nuevo clientes?

Hay problema de lealtad en clientes

Experiencias prácticas en la implementación de CLV

Pasos para implementar un CLV

1) Entender cómo negocios va a usar el CLV

2) Entender las diferentes líneas de productos de la empresa, y decidir cuáles entran a formar parte del CLV

3) Entender los ingresos y costos de cada producto

4) Obtener acceso a data relacionada a los productos: clientes, ingresos, costos, snapshot mensuales, etc. Haver análisis preliminar

Pasos para implementar un CLV

5) Hacer un borrador de los modelos y las fórmulas usadas para calcular el CLV

6) Presentar el borrador a los gerentes de productos y otros departamentos, y obtener su feedback

7) Implementar el CLV. Durante la implementación, tu entendimiento de negocio y de la data va a cambiar, y vas a ajustar tu modelo inicial

8) Presentar el CLV a la gerencia, en lenguaje de negocio

Dificultades en implementar un CLV

Apoyo gerencial

Muchas veces, la gerencia no sabe decirte directamente que preguntas tienen o entienden bien que es un CLV

Es nuestra responsabilidad presentar resultados en lenguaje de negocio, respondiendo preguntas claves de negocio.

Vas a necesitar apoyo gerencial para obtener tiempo de recursos claves

Difultades en implementar un CLV

Dificultades de negocio

Entender las peculiaridades de cada producto, especialmente en los ingresos y costos

Dificultades en implementar un CLV

Dificultades técnicas

-Tiempo escaso de personal clave de procesos y de TI

-Falta de documentación en los sistemas operacionales

-Diversidad de sistemas, con ID de clientes diferentes (primary keys)

-Histórico no disponible o muy difícil de obtener

-Data sucia

-Gran volumen de data (cientos de gigabyte a terabytes de data). Se necesitan servidor potente, buen software, saber como eficientizar los queries y los procesos

-Logicas complejas que no se pueden expresar en query. Hay que programar

Dificultades en implementar un CLV

Dificultades técnicas (cont.)

-Organización del código, se necesita experiencia de software para crear codigo mantenible

-Dificultad para actualizar data, especialmente si hay muchas fuentes manuales

-Falta de recursos para la implementacion. En mi experiencia, serás el único recurso disponible y no siempre 100%

-Grabar los datos claves de los cálculos, con nombres de campos entendibles, de forma historica. Para fines de auditoria, campanas, análisis detallado de ingresos, costos, comparar históricamente al cliente, etc

Preguntas

?

Referencias

Blattberg, R. and Deighton, J. (1996). Manage marketing by the customer equity test. Harvard Business Review, July-August:136–144

Malthouse, Edward (2013). Segmentation and Lifetime Value Models Using SAS. SAS Institute

Anualidad. https://es.wikipedia.org/wiki/Anualidad . Wikipedia . Accesado 2018-10-24

Análisis de supervivencia. https://en.wikipedia.org/wiki/Survival_analysis . Wikipedia. Accesado 2018-10-24

Kleinbaum, David and Klein, Mitchel (2012). Survival Analysis. A Self Learning Text. Springer