

Presentación del Problema

Introducción

La empresa tiene un grupo de clientes que posee un producto de alta gama llamado *Paquete Premium*, son los clientes más valiosos.

Actualmente la empresa no hace campañas proactivas de retención de clientes, simplemente una vez que el cliente manifiesta que se quiere ir, reaccionan intentando retenerlo .

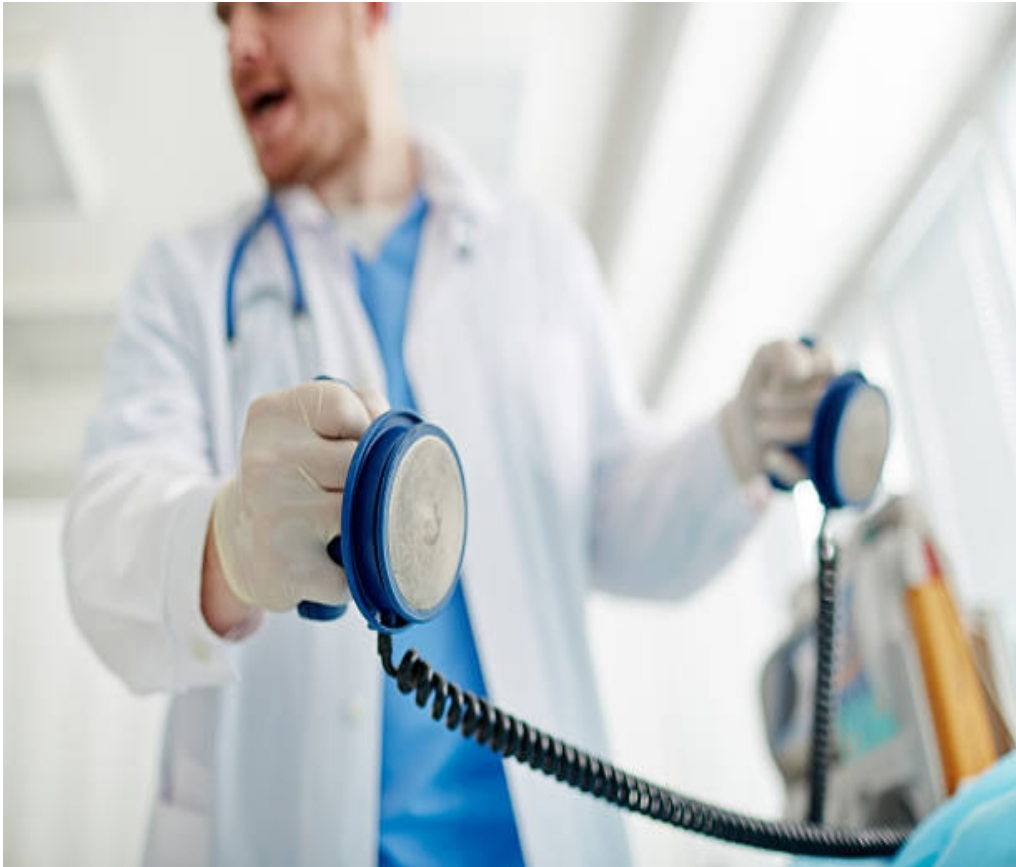
Objetivo

Se desea hacer un modelo predictivo para elegir a los clientes a los que se hará una campaña de marketing preventiva de retención, antes que manifiesten su voluntad de darse de baja.

Tipos de retención de clientes

- Retención REACTIVA
- Retención PROACTIVA
- “Medicina” Preventiva

Retención Reactiva



El cliente manifiesta
su voluntad de irse.

generalmente, ya es
tarde

Retención Proactiva



El cliente aún no
manifiesta que quiere
irse, pero lo hará en ...
2 meses ...
con alta probabilidad

Intensive Care Unit

“Medicina” Preventiva



Se cuida la “salud” de cliente para evitar que enferme y muera

Los datos

- El ultimo día del mes, a las 23:59:59 se obtiene un snapshot del cliente
 - numero_de_cliente
 - foto_mes
 - detalles de la actividad del cliente
 - clase_ternaria
- Son 158 + 1 campos
- aprox 230k registros por snapshot, un reg x cliente
- Período 201801 a 202101

Los datos

- El dataset original está en el Repositorio de la Materia en la carpeta `datasetsOri`
- En la misma carpeta está el diccionario de datos

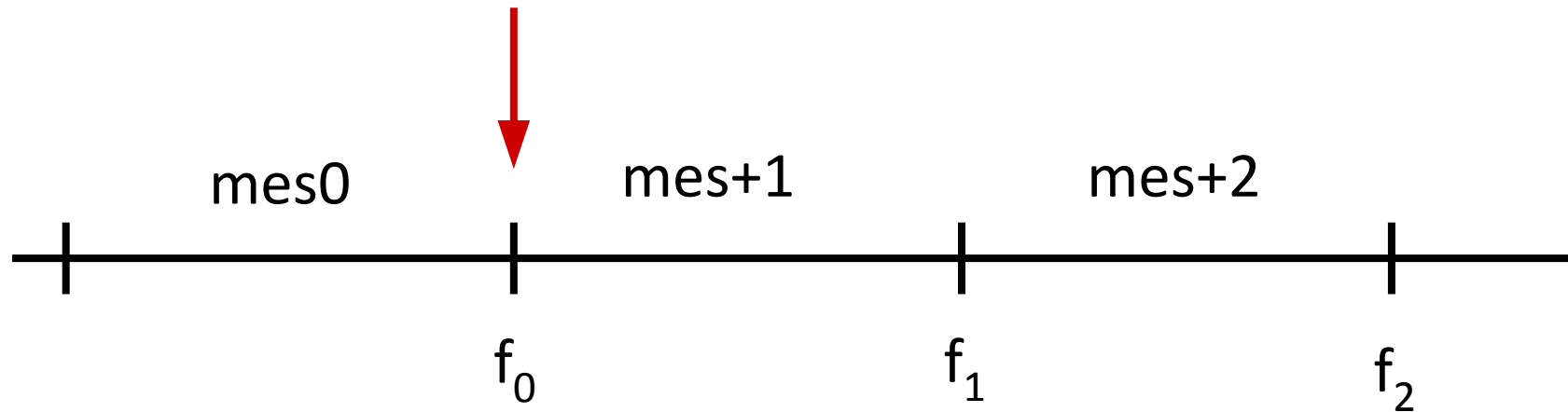
La clase

- Los datos del snapshot son “del pasado reciente”
datos que se conocen al último día del mes 23:59:59
- La clase se calcula mirando el futuro, los dos próximos meses, los dos snapshots siguientes.
Es el único campo del futuro.

la clase_ternaria

clase_ternaria	descripción
BAJA+1	se da de baja durante el mes+1
BAJA+2	se da de baja durante el mes+2
CONTINUA	sigue siendo cliente luego del mes+2

determinación de la Clase



A la fecha f_0 es cliente de paquete premium

- BAJA+1 se da de baja durante el mes+1, no aparece en f_1
- BAJA+2 se da de baja durante el mes+2, aparece en f_1 pero ya no está en f_2
- CONTINUA a la fecha f_2 sigue siendo cliente, aparece en f_0 , f_1 y f_2

La clase

clase	mes ₀	mes ₁	mes ₂
BAJA+1	si	no	<i>no</i>
BAJA+2	si	si	no
CONTINUA	si	si	si

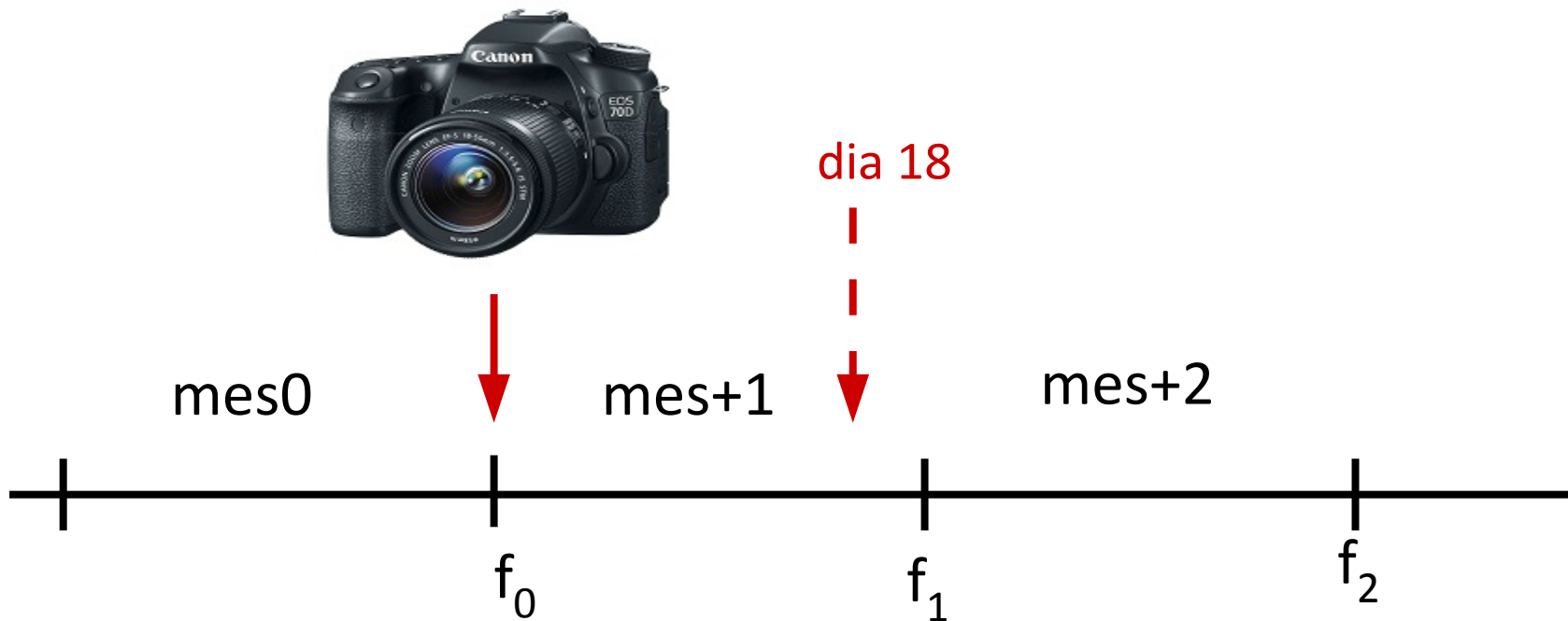
La materia tiene
una competencia
que representa el 60% de la nota

Que snapshots tienen clase ?

- Para calcular la clase debo conocer dos meses del futuro.
- El ultimo mes disponible es 202101, tiene la clase vacia, ya que no conozco el futuro
- El mes 202011 es el último mes con la clase completa { BAJA+1, BAJA+2, CONTINUA }

Disponibilidad del snapshot

Por este motivo los BAJA+1
dan pérdida



Dataset

Última
foto

Futuro

202011

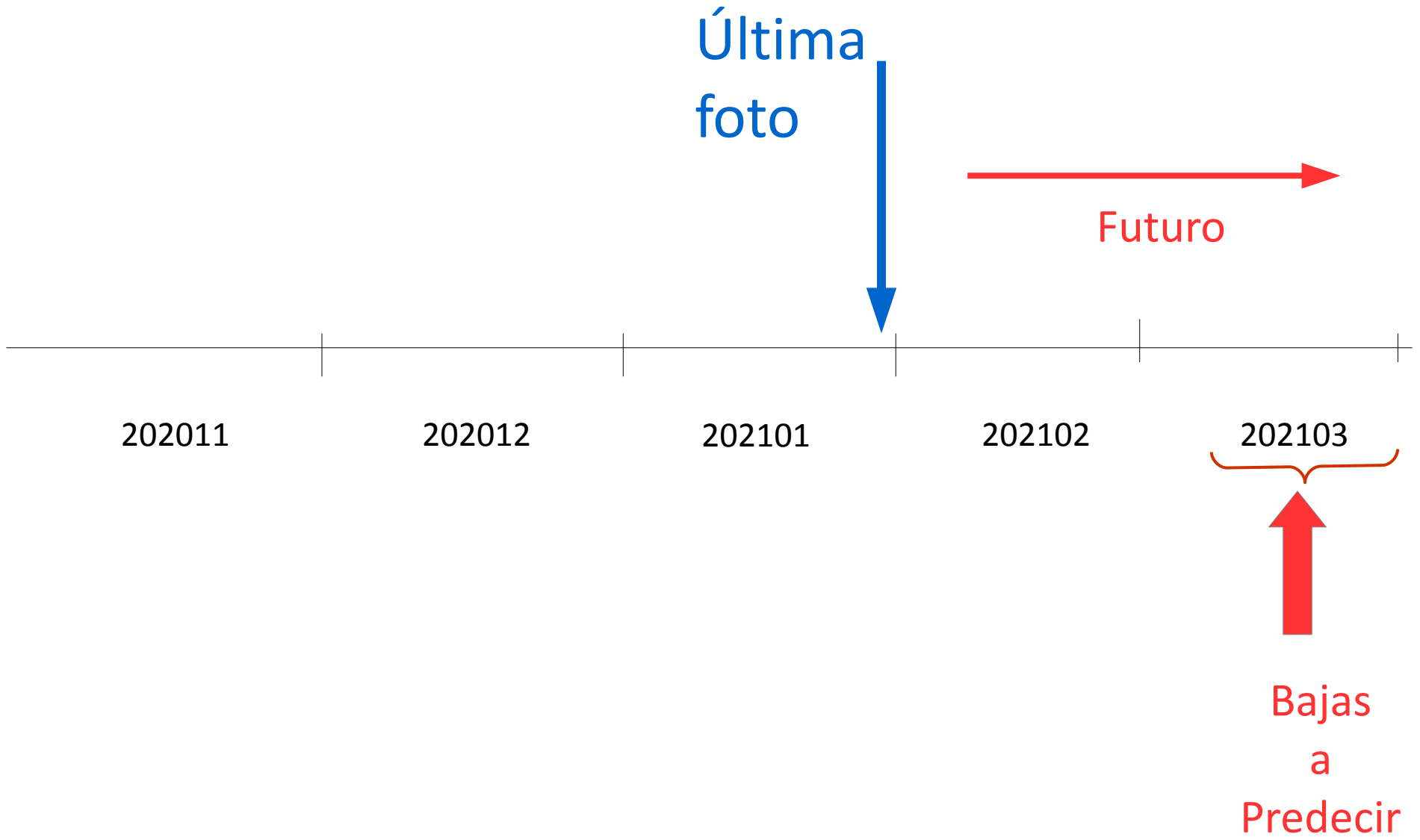
202012

202101

202102

202103

Bajas
a
Predecir



Decir
“Las BAJA+2 de enero”

es lo mismo que

“los clientes que tienen paquete
premium al 31-enero y se van a dar de
baja *durante* marzo”

Campaña de Marketing de Retención Proactiva

Campaña de retención

Una vez que se conocen los datos del snapshot del mes₀, lo que sucede el día 18 del mes₁, se quiere hacer una campaña de marketing de retención proactiva a clientes que el modelo prediga tienen alta probabilidad de darse de baja *durante* el mes₂

(La campaña se hace solo a algunos clientes, a los *muy enfermos*, definitivamente no se hace a todos los clientes)

Campaña de retención

- La campaña consiste de un estímulo, que tiene un costo de \$ 1,250
- En experimentos pasados se ha encontrado que de las personas que se iban a ir en el mes₂ y reciben el estímulo, el 50% decide quedarse y el otro 50% se va durante el mes₂
- Un cliente que se queda, deja una ganancia futura de \$ 100,000

Ganancia de la campaña

$$\text{Ganancia} = 0.5 * \$100,000 * \text{aciertos} \\ - \$1,250 * \text{envios}$$

envios = cantidad de estímulos

aciertos = clientes que reciben el estímulo y se iban a ir en mes₂

Ganancia de la campaña

$$\text{Ganancia} = \$50,000 * \text{aciertos} - \$1,250 * \text{envios}$$

$$\text{envios} = \{ \text{BAJA}_1, \text{BAJA}_2, \text{CONTINUA} \}$$

$$\text{aciertos} = \{ \text{BAJA}_2 \}$$

Ganancia de la campaña

$$\begin{aligned} \text{Ganancia} = & \$ 50,000 * \text{BAJA}_2 \\ & - \$ 1,250 * (\text{BAJA}_1 + \text{BAJA}_2 + \text{CONTINUA}) \end{aligned}$$

$$\begin{aligned} \text{Ganancia} = & \$ 48,750 * \text{BAJA}_2 \\ & - \$ 1,250 * (\text{BAJA}_1 + \text{CONTINUA}) \end{aligned}$$

Evaluación

¿Cómo se evalúa la lista de clientes que entrega un alumno ?

El profesor, que conoce el futuro, sabe para cada `numero_de_cliente` cual es la clase { BAJA+1, BAJA+2, CONTINUA} , y suma la ganancia de cada registro.

Ganancia de cada registro entregado

CLASE REAL (futuro)	Ganancia
BAJA+1	-1,250
BAJA+2	48,750
CONTINUA	-1,250

$$\text{Ganancia} = 48,750 * \text{'BAJA+2'} - 1,250 * (\text{'BAJA+1'} + \text{'CONTINUA'})$$

Pero un modelo devuelve probabilidades !

¿Pero, qué lista se debe entregar ?

Un modelo predictivo
asignará a cada cliente de 202101
una probabilidad de darse de baja *durante* 202103.

¿Cómo elijo cuales probabilidades son las que me
sirven ?

(a continuación derivación matemática)

Función ganancia

$$\text{ganancia} = 48750 * \text{'BAJA+2'} - 1250 * (\text{'BAJA+1'} + \text{'CONTINUA'})$$

$$\text{ganancia} \geq 0 \text{ \textcolor{green}{sii}}$$

$$48750 * \text{'BAJA+2'} - 1250 * (\text{'BAJA+1'} + \text{'CONTINUA'}) \geq 0$$

$$\text{dado que 'BAJA+2'} \geq 0, \text{'BAJA+1'} \geq 0 \text{ y 'CONTINUA'} \geq 0$$

$$\text{\textcolor{green}{sii}}$$

$$48750 * \text{'BAJA+2'} \geq 1250 * (\text{'BAJA+1'} + \text{'CONTINUA'})$$

Función ganancia

sii

$$48750 * 'BAJA+2' \geq 1250 * ('BAJA+1' + 'CONTINUA')$$

sii sumo $1250 * 'BAJA+2'$ de ambos lados de la igualdad

$$48750 * 'BAJA+2' + 1250 * 'BAJA+2' \geq \\ 1250 * ('BAJA+1' + 'CONTINUA') + 1250 * 'BAJA+2'$$

sii

$$50000 * 'BAJA+2' \geq 1250 * ('BAJA+2' + 'BAJA+1' + 'CONTINUA')$$

Función ganancia

sii

$$50000 * \text{'BAJA+2'} \geq 1250 * (\text{'BAJA+1'} + \text{'BAJA+2'} + \text{'CONTINUA'})$$

sii

$$\text{'BAJA+2'} / (\text{'BAJA+1'} + \text{'BAJA+2'} + \text{'CONTINUA'}) \geq 1250 / 50000$$

sii

$$\text{'BAJA+2'} / (\text{'BAJA+1'} + \text{'BAJA+2'} + \text{'CONTINUA'}) \geq 0.025$$

Función ganancia

sii

$$'BAJA+2' / ('BAJA+1' + 'BAJA+2' + 'CONTINUA') \geq 0.025$$

Si BAJA+2 son los positivos, entonces

$$'BAJA+2' / ('BAJA+1' + 'BAJA+2' + 'CONTINUA') = \text{prob(POS)}$$

o sea

$$\text{ganancia} \geq 0 \quad \text{sii} \quad \text{prob(POS)} \geq 0.025$$

Función ganancia

Conclusión, la campaña es rentable si incluyo a todos los clientes que su probabilidad real de BAJA+2 sea mayor o igual a 0.025 (2,5%)

Notar que este punto de corte depende solamente del problema y para nada del dataset.

No depende de la distribución de las clases en el dataset.

Ejemplo scoring $\text{prob}(\text{BAJA}+2) > 0.025$

MODELO $\text{prob}(\text{BAJA}+2)$	Acción Enviar	Realidad	Ganancia
0.010	NO	CONTINUA	
0.200	SI	BAJA+1	-1250
0.850	SI	BAJA+2	+48750
0.030	SI	CONTINUA	-1250
0.019	NO	BAJA+2	
0.005	NO	BAJA+1	
0.026	SI	BAJA+1	-1250
0.001	NO	CONTINUA	
0.099	SI	BAJA+2	+48750

Observaciones

Sobre su entrega

Una excelente entrega en estos datos sería:

CLASE REAL	registros	ganancia unitaria	ganancia total
CONTINUA	5,403	\$ -1,250	\$ - 6,753,750
BAJA+1	267	\$ -1,250	\$ - 333,750
BAJA+2	614	\$ 48,750	\$ 29,932,500
TOTAL	6,284	-----	\$ 22,845,000

Observar

Esta excelente entrega tan solo le acertó a 614 BAJA+2 del total de 6,284 de registros o sea, apenas acertó al 9.8%

Sin embargo, generó una ganancia > \$ 22.8M
esto se debe a la asimetría en los pesos que se asignan a los positivos y a los negativos
\$48,750 vs \$1,250

Consideraciones

Consideraciones

La empresa jamás ha hecho una campaña proactiva de retención de clientes (solo hace campañas de cross selling).

Su trabajo como científico de datos es pasarle al departamento de marketing el día 20 de cada mes la lista de clientes que a su entender se darán de baja durante el mes siguiente (BAJA+2)

Consideraciones

Como marketing no cree en usted, los primeros tres meses no enviará el estímulo a los clientes, sino que verificará la ganancia que habría tenido la campaña de retención.

Al cuarto mes, marketing dividirá al azar la lista que usted le pase en dos mitades.

A una mitad la llamará *grupo de control* y no hará nada.

A la otra mitad le enviará el estímulo.

A los n meses medirá el porcentaje de clientes que se dieron de baja de cada uno de los grupos.

Esto se llama A/B testing

Consideraciones

En un futuro lejano, marketing se sofisticará, empezará a probar con distintos estímulos, de distinto costo, registrará dicha información,

y le pedirá a usted realizar un modelo predictivo que para cada tipo de estímulo calcule la probabilidad que ese cliente que va a ser un BAJA+2 permanezca en la empresa si recibe *ese* estímulo.

Consideraciones

En un futuro realmente muy lejano, marketing dará un salto intelectual y comenzará a preguntarse :
como debería tratar rentablemente a cada uno de mis clientes de forma que la probabilidad de BAJA+2 siempre esté por debajo de cierto umbral, y así evitar que lleguen a la Unidad de Terapia Intensiva.

Para concluir

A pesar que son todos clientes de paquete premium, existen diferencias entre ellos, y algunos dejan mas ganancia a la empresa que otros, el \$ 100,000 es un promedio.

Una vez que se consolide en la Ciencia de Datos usted deberá sofisticar su modelo para que considere ganancias distintas por cliente.

FIN