Modelos de Churn Bancarios con R

Trabajo Final de la Maestría en Data Mining

Presentado por

Diego Ariel Oppenheim

Director

Martin Volpacchio

Fecha: 31/7/2017

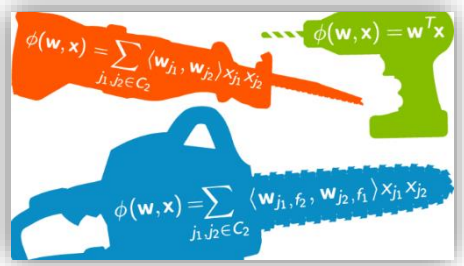
Objetivos del trabajo final

- Entender la situación actual de fuga de clientes del Banco
- Entender cuáles son las variables que mejor explican el fenómeno
- Desarrollar un modelo de Data Mining que permita predecir los clientes que se darán de baja dentro de los próximos dos meses
- Entender los motivos de bajas de clientes y definir acciones comerciales de retención
- Cuantificar el impacto de un programa de retención

Esquema General

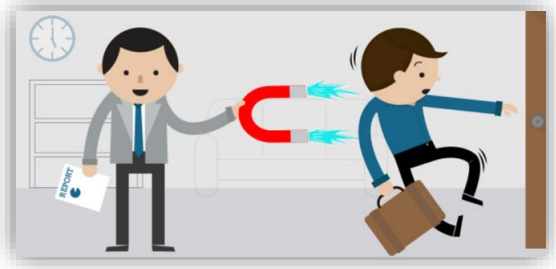
Fuentes de Información

Identificación de las principales variables que definen/caracterizan a los clientes que se fugarán



Clientes con mayor valor real y potencial para el Banco

Medidas de retención apropiadas para los clientes en riesgo de fuga



Cuantificar el incentivo, costos y ganancia esperada de un programa de retención

Estrategia post retención para aumentar la fidelización.

Introducción al concepto de Attrition

- Término utilizado para describir la pérdida del cliente

Tipos de Fuga

- **Total** → Cancela todos los acuerdos
- **Parcial** → Cancela algunos productos y para el resto utiliza la competencia
- **Escondida** → Inactividad por un largo periodo de tiempo

Motivos de fuga

- Precio
- Producto
- Servicio - Insatisfacción
- Mercado
- Organización
- Ofertas Competencia

Por qué retenerlos?

- Más rentable retener a un cliente valioso que captar uno nuevo
- Clientes a largo plazo tienden a consumir más
- Clientes a Largo Plazo resultan menos sensibles a actividades publicitarias de la competencia
- Nuevos clientes potencialmente riesgosos

Etapas del Trabajo - Concepto del KDD (*Knowledge Discovery in Databases*)

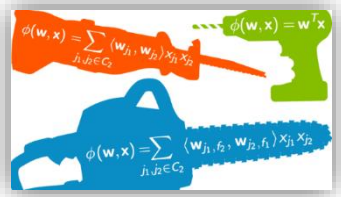
PROCESS CONSULTING

Integración

- Fuentes de información
- Análisis de datos
- IT
- Data Wharehouse
- Integración
- Multiplataforma
- Formatos diferentes
- Frecuencia de extracción

Pre-Procesamiento

- Valores ausentes
- Imputación de valores
- Outliers
- Duplicados, registros incoherentes
- Nuevas Variables
- Transformación de variable
- Feature Engineering



Data Mining

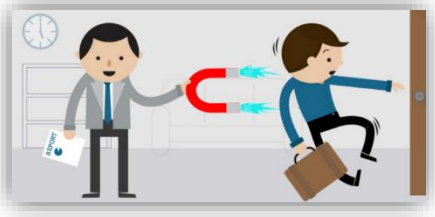
- Descubrimiento
- Patrones, secuencias
- Modelos de clasificación, predicción
- Confiabilidad
- Usabilidad
- Escalabilidad
- Robustez
- Estacionalidad de los datos

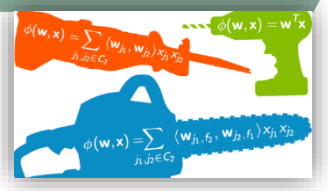
Validación

- Evaluación técnica y comecial
- Métricas de rendimiento
- Accuracy
- Recall
- AUC
- LIFT
- Curva Roc
- Matriz de confusión

Implementación

- Acción sobre el cliente (Evento)
- Comunicación
- Presentación de resultados
- Medición
- Plan de retención
- Servicio post venta





Imputación de valores faltantes - Nulos

- Descarte de registros
- Imputación de datos faltantes en base a árboles de decisión
- Conserva la distribución inicial de la variable

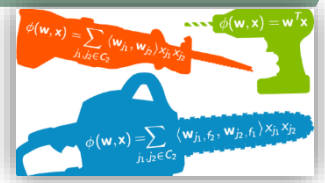
Data Set Original

Mbb	Edad	var_act
NA	3.5	1.4
NA	3	NA
4.7	3.2	1.3
4.6	3.1	1.5
NA	3.6	NA
5.4	NA	1.7
NA	3.4	1.4
4.8	3.4	1.5
4.4	2.9	1.4

Imputación datos faltantes

nuevo Data set imputado

Mbb	Edad	var_act
4.3	3.5	1.4
4.4	3	1.1
4.7	3.2	1.3
4.6	3.1	1.5
4.5	3.6	1.7
5.4	3.9	1.7
5.1	3.4	1.4
4.8	3.4	1.5
4.4	2.9	1.4



Balanceo de la muestra

- Problema de clases desbalanceadas
- Escasez de una de las clases (un conjunto de datos resulta significativo)
- Bajo % de fuga de la población de estudio

	Ventanas Temporales					
	Junio	Julio	Agosto			
Clientes	CHURN1	CHURN2	CHURN3	CHURN_TOTAL	Proporcion	Proporcion Churn
2.2M	0	0	0	0	97,6%	97,6%
9.2K	1	0	0	1	0,4%	2,4%
23K	0	1	0	1	1,0%	
22.7K	0	0	1	1	1,0%	
2.3M						

Se aplicó la técnica del sub-muestreo aleatorio que elimina al azar instancias de la clase mayoritaria

Foco en clientes Premium - **Alta participación en las ganancias**

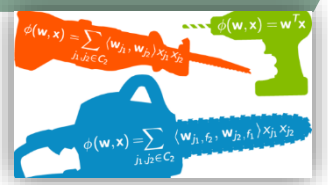
Muestra no Balanceada

Segmentos Select y Alta	Clientes	Proporción
Clientes NO fugados (0)	411.549	99,15%
Clientes Fugados (1)	3.546	0,85%
Total	415.095	100,00%

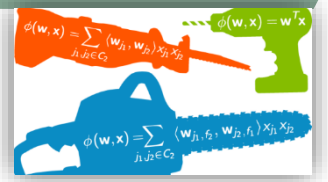
Muestra Balanceada

Segmentos Select y Alta	Clientes	Proporción
Clientes NO fugados (0)	25.000	87,58%
Clientes Fugados (1)	3.546	12,42%
Total	28.546	100,00%

Creación de nuevas variables - *Feature Engineering*



Derivadas de variables existente	Variables *Flag* (0-1)	Variables Ratio
• Discretización de la variable cuadrante (valor del cliente) • Discretizacion del mejor producto • Discretizacion de la zona de residencia • Discretizacion del estado civil	• Tenencia de tarjeta de crédito (últimos 3 y 6 meses) • Tenencia de tarjeta de débito (últimos 3 y 6 meses) • Cobro de sueldo hace 3 meses	• Variación de MBB (ingreso que deja al Banco) • Variación en la cantidad de productos bancarios • Variación de la cantidad de artículos consumidos • Variación del monto en consumo • Variación del monto total de las inversiones • Variación del saldo en cuenta corriente • Variación de la cantidad de extracciones por Home Banking **Variaciones a 3 y 6 meses**



Selección de variables

- Definir las variables más influyentes para los modelos
- Independencia entre variables, baja correlación con entre si y alta correlación con la variable target (fuga del cliente)

Se excluyeron las variables con un coeficiente de correlación mayor o igual a 0.8

Ingreso Promedio y Suma del Ingreso tienen alta correlación positiva → Se excluye la que menor correlación tiene con la variable target

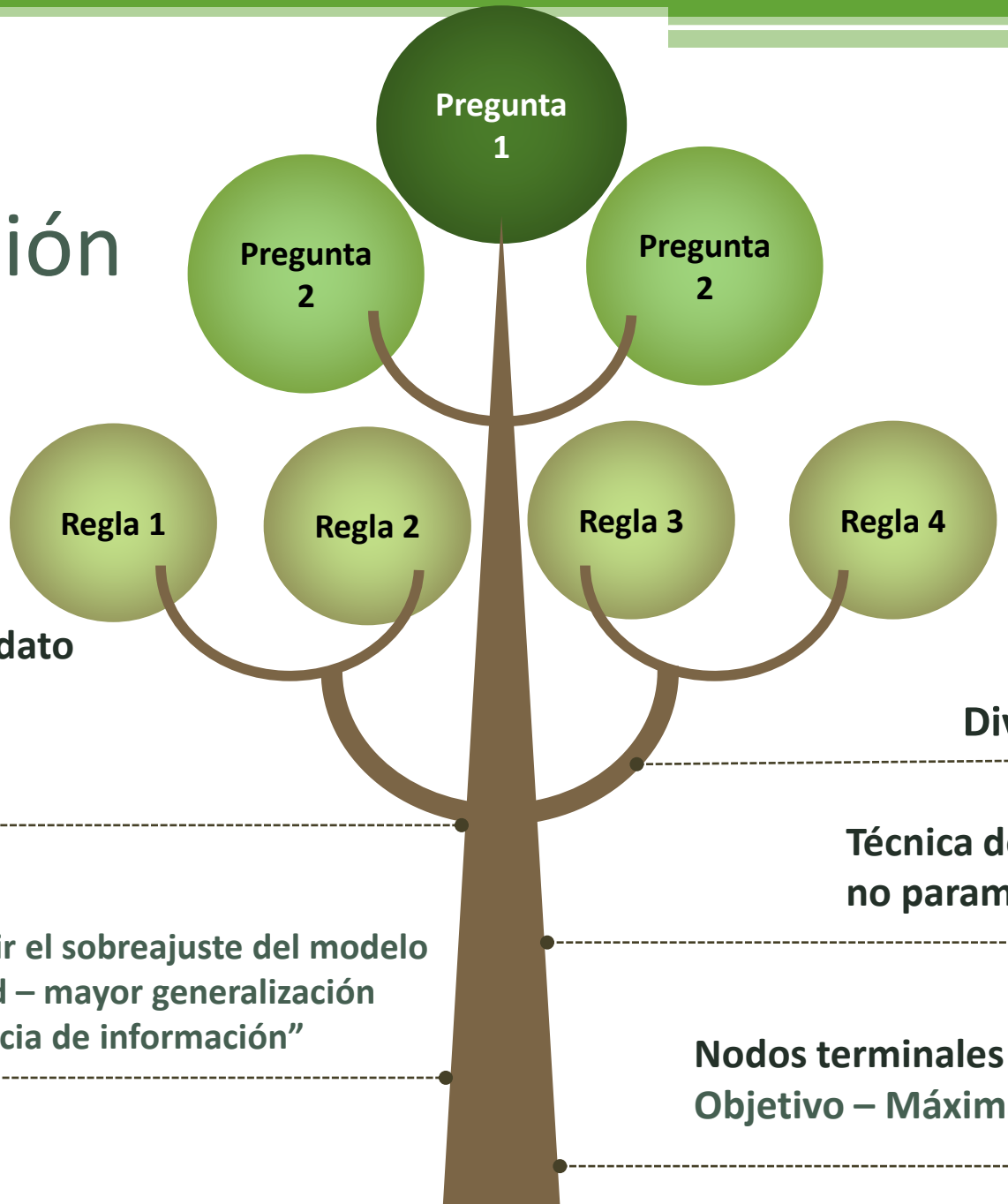
Lapso de compra y Attrition tienen alta correlación negativa → Se mantiene la variable

Modelos

Árboles de decisión

Bagging – Boosting – Random Forest

Árbol de decisión



- Robusto a Outliers
- Aplicable a cualquier tipo de dato
- Sencilla interpretación
- Reutilizable

Overfitting

Poda para prevenir el sobreajuste del modelo
Mayor simplicidad – mayor generalización
Técnica de "ganancia de información"

Divide y vencerás

Técnica de clasificación no paramétrica

Nodos terminales

Objetivo – Máxima pureza

Modelos: *Bagging*

- Muestreo selectivo – Aumento de la performance
- Creación de diferentes modelos usando muestras aleatorias con reemplazo y luego combina y ensambla los resultados mejorando la predicción

Modelos: *Boosting*

- Cada nuevo clasificador presta mayor atención a los datos clasificados erróneamente por los clasificadores anteriores
- Combinación de resultados para obtener un clasificador con mayor poder de predicción

Modelos: *Random Forest*

- Cada árbol contiene una muestra aleatoria de observaciones y es construido con variables seleccionadas de forma aleatoria

Métricas de Rendimiento

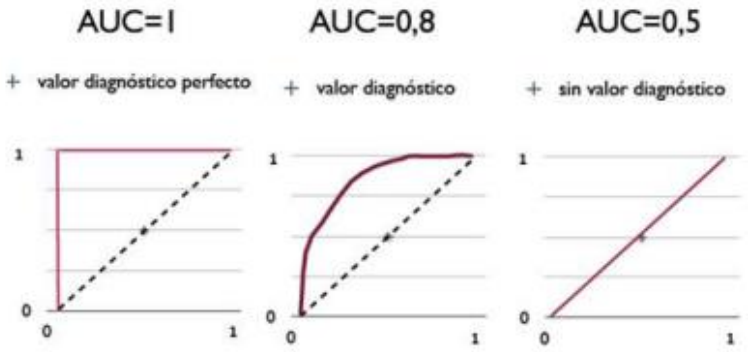
- Matriz de Confusión, muestra los resultados de los modelos de clasificación binaria

		Clase Actual	
		NO CHURN	CHURN
Clase Predicha	NO CHURN	TP	FN
	CHURN	FP	TN

Sensitivity	→ TP / (TP+FN)
Specificity	→ TN / (FP+TN)
Positive Predictive Value	→ TP / (TP+FP)
Negative Predictive Value	→ TN / (FN+TN)
Accuracy	→ (TP+TN) / (TP+FP+FN+TN)

AUC - Área bajo la curva ROC - LIFT

- Comparación entre las predicciones correctas e incorrectas del clasificador
- LIFT: Indica la cantidad de veces que mejora la predicción del modelo con respecto al azar



Principales variables seleccionadas con los modelos

CLUSTER 0	1?
VARIABLES	
ANTIGUEDAD	
AVG_CANTIDAD	x
AVG_CONSUMO_ACT	
AVG_HOME_BANKING_TRX	
AVG_SALDO_ACT	x
EDAD	
LAPSO_COMPRA	
LIM_CREDITO	x
MBB	
SUM_MBB	
MOV90_CTA	
RIESGO_SCORE_VERAZ	
SCORE_PROMEDIO	
USO_SALDO_PROPIO_TC	
VAR_CONSUMO	
VAR_CROSS_SELLING	
VAR_HOME_BANKING_TRX	
VAR_INGRESO_12	
VAR_MBB	
VAR_SUM_MBB	x
VAR_MOV90_CTA	
VAR_SALDO	
VAR_SCORE_PROMEDIO	

CLUSTER 1	0?
VARIABLES	
ANTIGUEDAD	
AVG_CAJERO_PPIO_MONTO_ACT	x
AVG_CONSUMO_ACT	
AVG_HOME_BANKING_TRX	
CROSS_SELLING	x
EDAD	
LAPSO_COMPRA	
MBB	
MOV90_CTA	
RIESGO_SCORE_VERAZ	
SCORE_PROMEDIO	
SUM_MBB	
USO_SALDO_PROPIO_TC	
VAR_CANTIDAD	x
VAR_CONSUMO	
VAR_CROSS_SELLING	
VAR_HOME_BANKING_TRX	
VAR_INGRESO_12	
VAR_MBB	
VAR_MOV90_CTA	
VAR_SALDO	
VAR_SCORE_PROMEDIO	

Comportamientos diferenciados para cada uno de los *grupos de clientes seleccionados (Clusters)*

Variables:

- Cantidad de compras (transacciones) en los últimos 3 meses
- Saldo en cuenta de los últimos 3 meses.
- Variación de la suma de MBB (Ingreso que deja el cliente al Banco)
- Cantidad de extracciones en cajeros
- Tenencia (en número) de productos bancarios
- Cantidad de compras en los últimos 3 meses

Evaluación de los modelos

VALIDACIÓN CON DATA FUTURA (1 AÑO)

Clases altamente desbalanceadas para ambos clusters para la variable Target (Attrition)

	Churn	No Churn	% Churn
Cluster 0	3.897	254.787	1,53%
Cluster 1	725	227.513	0,32%
Total	4.622	482.300	0,96%

Validación de los modelos con datos a Abril 2016

Modelos Random Forest y Boosting

AUC
96%-97%

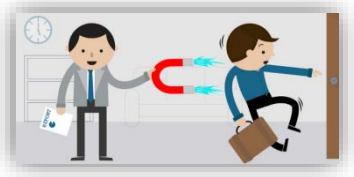
Accuracy
95%-99%

TN, *True Negative*(Especificidad)
90% (para el Cluster 0 el modelo encontró a 3715 de los 3879 clientes fugados)

Falsos Positivos
60%-77%
El modelo Boosting clasificó correctamente al **23%-42%** de los clientes que predijo como Churners

LIFT - Primeros 20.000 clientes
11.23
Más probable a abandonar el Banco a un cliente seleccionado al azar

Programa de Retención → + LIFT + Rentabilidad + Éxito del programa



N: Número total de clientes, 486.922

Pr: Porcentaje de clientes a impactar con el programa de retención (Percentil 8), 8%

Pe: Son los verdaderos positivos que el modelo detectará, es el porcentaje de clientes impactados por el programa que efectivamente serán fugas del Banco, (tener en cuenta que el *Attrition* de la cartera para los *Cluster* 0 y 1 es de 0.96%), 10.22%

e: Tasa de éxito del programa, se asumen 3 escenarios posibles, [10%, 30%, 50%]

C: Costo del incentivo ofrecido por el Banco, 50 USD

Co: Costo de contactar al cliente, 5 USD

CLV: Valor del cliente para el Banco en caso en que sea retenido, 4000 USD

V: Valor del programa de retención (ganancia/pérdida)

*Aquellos clientes impactados por el programa fueron los de **mayor valor comercial** para el Banco
*Mayor probabilidad de fuga
*Mayor LIFT
*Se busca reducir la tasa de los Falsos Positivos (clientes que permanecen en el Banco habiendo recibido un incentivo)

$V=486.922*8%*(10.22%*10%*($4.000-$50-$5)+ 10.22%*(1-10%)*(-$5)+(1-10.22%)*(-$50-$5))

Resultado e Implementación del Programa de Retención

		Tasa de Éxito del Programa de Retención					
		10%	15%	20%	25%	30%	35%
Customer Lifetime Value (USD)	$4.000	$ -371.595	$ 414.350	$ 1.200.294	$ 1.986.238	$ 2.772.183	$ 3.558.127
	$4.500	$ -172.621	$ 712.810	$ 1.598.241	$ 2.483.672	$ 3.369.103	$ 4.254.534
	$5.000	$ 26.352	$ 1.011.269	$ 1.996.187	$ 2.981.105	$ 3.966.022	$ 4.950.940
	$5.500	$ 225.325	$ 1.309.729	$ 2.394.134	$ 3.478.538	$ 4.562.942	$ 5.647.347
	$6.000	$ 424.298	$ 1.608.189	$ 2.792.080	$ 3.975.971	$ 5.159.862	$ 6.343.753
	$6.500	$ 623.272	$ 1.906.649	$ 3.190.027	$ 4.473.404	$ 5.756.782	$ 7.040.159
	$7.000	$ 822.245	$ 2.205.109	$ 3.587.973	$ 4.970.837	$ 6.353.702	$ 7.736.566
	$7.500	$ 1.021.218	$ 2.503.569	$ 3.985.920	$ 5.468.271	$ 6.950.621	$ 8.432.972

GANANCIA ADICIONAL ante una mejora del 1% en el LIFT

		Tasa de Éxito del Programa de Retención					
		10%	15%	20%	25%	30%	35%
Customer Lifetime Value (USD)	$4.000	$ 17.709	$ 25.568	$ 33.428	$ 41.287	$ 49.146	$ 57.006
	$4.500	$ 19.698	$ 28.553	$ 37.407	$ 46.261	$ 55.116	$ 63.970
	$5.000	$ 21.688	$ 31.537	$ 41.386	$ 51.236	$ 61.085	$ 70.934
	$5.500	$ 23.678	$ 34.522	$ 45.366	$ 56.210	$ 67.054	$ 77.898
	$6.000	$ 25.668	$ 37.506	$ 49.345	$ 61.184	$ 73.023	$ 84.862
	$6.500	$ 27.657	$ 40.491	$ 53.325	$ 66.159	$ 78.992	$ 91.826
	$7.000	$ 29.647	$ 43.476	$ 57.304	$ 71.133	$ 84.962	$ 98.790
	$7.500	$ 31.637	$ 46.460	$ 61.284	$ 76.107	$ 90.931	$ 105.754

SUPUESTOS CONSIDERADOS:

1. Distribución constante del CLV
2. Tasa de éxito del 20%
3. Ganancia media de 2MM USD (en 3 meses)

ÉXITO DEL PLAN DE RETENCION ANTE:

- Clientes con gran valor para el Banco
- Alto riesgo de fuga para los próximos meses
- Correcta comunicación, canal e incentivo con el cliente
- Revisión y mantenimiento de los modelos predictivos

Conclusiones

- Se trata de un problema del desbalanceo de clases (existe una clase predominante y una minoritaria, fuga de clientes). Relación 99.5% a 0.5%
- Balanceo de casos mediante sub-muestreo, eliminación de registros de clase dominante.
- Problema de la sub especificación del modelo: No contar con todas las variables para poder explicar el fenómeno (existencia de limitaciones y supuestos)
- Ranking de clientes en base a la probabilidad de fuga
- Acciones comerciales de retención, necesidad de bajar el CHURN del 12% anual
- Cambios rápidos en las condiciones económicas del país
- Oportunidad de mejora (+ retención - adquisición)

Conclusiones

- Aquellos clientes que cobraban en el Banco hace 6 meses y que redujeron la cantidad de productos en un 30% y redujeron en mas de un 20% la cantidad de artículos comprados, tienen el triple de probabilidad de abandonar el Banco (X3)
- Dentro del grupo de clientes que no acreditaba sueldo en el Banco, aquellos que disminuyeron el saldo en vista en un 25%, no consumieron en mas de 2 meses y disminuyeron en un 30% la cantidad de movimientos voluntarios totales, tienen el quíntuple de probabilidad de fuga (X5).

Trabajo Futuro

- Inclusión de nuevas variables
- Revisión de los algoritmos utilizados
- Optimización de parámetros de los modelos
- Implementación de nuevos algoritmos. Ejemplo, *gradient boosting* (XGBOOST)

  A diferencia del algoritmo Boosting aplicado en el trabajo, el XGBOOST construye arboles de manera secuencial añadiendo a cada iteración el árbol que mejor compensa los errores de los arboles previos. Esto se realiza árbol a árbol.

- *Ensembles* de modelos para explorar diferentes hipótesis sobre los datos
- Predicción de fuga mes a mes (Función de Hazard)
- Afinar supuestos de *Customer Life Time Value*

Dudas, preguntas?

Muchas gracias!

Diego Ariel Oppenheim