UNIVERSIDAD SANTO TOMAS
PRIMER CLAUSTRO UNIVERSITARIO DE COLOMBIA

Facultad de Estadística
Trabajo de Grado
Enero 2016

EL MODELO DE UP-LIFTING APLICADO A UN SCORE DE RIESGO

UPLIFTING MODEL APPLIED TO A RISK SCORE

Carlos Abel Argüello Niño.[a]
caarguellon@gmail.com

Directora de tesis: Dra. Luz Mary Pinzón.[b]
luzpinzon@usantotomas.edu.co

RESUMEN

La minería de datos se ha utilizado ampliamente para optimizar el manejo de los clientes, con el fin de maximizar el retorno de la inversión. En particular, este trabajo trata del uso de los modelos de scoring en las campañas de comercialización de un producto.

Los modelos se desarrollan normalmente para identificar las características de los clientes que tienen más probabilidades de incurrir en un evento (caer en mora, comprar un producto, retirar un producto, etc.). Si bien estos modelos son útiles para identificar los clientes a los que se va a dirigir una campaña de marketing, esta campaña puede ser dirigida a clientes que ya han decidido que acción tomar con respecto al evento en cuestión (en este caso, compra de un producto), independientemente de si reciben o no la campaña (por ejemplo, correo electrónico, llamada).

Se propone la aplicación de una metodología para identificar a los clientes cuyas decisiones serán influenciados positivamente por campañas. La metodología propuesta es sencilla de implementar y se puede utilizar combinada con los algoritmos de aprendizaje supervisado más comúnmente utilizados. Esta metodología puede proporcionar al sector de telecomunicaciones una simple pero significativa mejora metodológica para optimizar sus acciones de marketing.

Palabras clave: Modelos de scoring, minería de datos, modelos predictivos, gestión de campañas de marketing, desarrollo de clientes, upselling y cross-selling, up-lifting, modelo con interacciones.

ABSTRACT

Data mining has been widely used to optimize the management of customers, in order to maximize the return on investment. In particular, this thesis is about the scoring models in the marketing of a product.

Models are usually developed to identify the characteristics of customers who have more possibilities to incur in an event (to be in arrears, buy a product, to withdraw a product, etc.). While these models are useful to identify clients that are going to run a marketing campaign, this campaign may be aimed at customers who have already decided what action to take regarding the event in question (in this case, buying a product), regardless of whether they receive the campaign (for example, email, phone).

It proposes the application of a methodology to identify customers whose decisions will be positively influenced by campaigns. The proposed methodology is simple to implement and can be used in combination with supervised learning algorithms most commonly used. This methodology can provide to the telecommunications sector a simple but significant methodological improvement to optimize their marketing.

[a]Estudiante de estadística Universidad Santo Tomás Bogotá
[b]Profesora de la facultad de estadística de la Universidad Santo Tomás Bogotá

Keywords: scoring models, data mining, predictive modeling, marketing campaign management, client development, upselling and cross-selling, up-lifting, model interactions.

Introducción

Entre las muchas aplicaciones de minería de datos y descubrimiento de conocimiento, las bases de datos provenientes del marketing son elementos claves para el desarrollo de métodos científicos que se aplican con frecuencia para reducir el riesgo en la toma de decisiones (Roberts, M.L . and Berger P.D y Kotler, P). En la última década, debido a la facilidad y capacidad de almacenar información, muchas industrias han implementado proyectos de almacenamiento de datos entre los que se destacan los referentes a los clientes. Esta práctica se ha convertido en requisito fundamental para maximizar el aprendizaje acerca de los comportamientos, hábitos y preferencias de estos, permitiendo a las empresas diseñar mejores ofertas, dirigir los mensajes de manera más precisa y utilizar los canales más adecuados para cada perfil de cliente (Fabris, P., 1988).

Dentro de las tareas más retadoras del marketing se encuentra estimar el retorno a la inversión de campañas publicitarias denominado ROI por sus siglas en inglés -Return Of investement-. Los publicistas han desarrollado métricas para estimarlo y con Internet estas mediciones se facilitan. Sin embargo, éstas principalmente miden tiempo de exposición y tiempo de interactividad con la marca, valores que posteriormente se comparan con el Top of mind.

Otra forma para estimar el ROI es mediante aplicaciones de minería de datos, donde se construyen modelos predictivos para determinar características de los clientes que respondieron positivamente a una campaña, es decir, que reaccionaron a la campaña bien sea comprando o participando según el mensaje. Luego para una campaña similar, el modelo puede ser usado para identificar segmentos que probablemente, responderán de manera positiva a la nueva campaña. Algunas ventajas de aplicar el modelo son evitar la recolección de información para evaluar la nueva campaña (Jorion, 2003) y canalizar de manera adecuada la comunicación (Almquist, 2001), dado que se pueden conocer los perfiles de clientes más interesantes para ésta.

Siguiendo la línea de minería de datos nos proponemos a hacer una aplicación a datos reales.

Justificación

Scoring se refiere al uso de conocimiento sobre el desempeño y características de préstamos en el pasado para pronosticar el desempeño de préstamos en el futuro. Así, cuando un analista de crédito valora el riesgo comparando mentalmente una solicitud de crédito en el presente con la experiencia que este mismo analista ha acumulado con otros clientes con solicitudes parecidas, está aplicando scoring, aunque sea un scoring implícito y subjetivo.

El scoring estadístico modela de manera individual el conocimiento cuantitativo del desempeño y características de los préstamos pasados, registrados en una base de datos, para pronosticar el desempeño de préstamos futuros. Esta técnica cuantifica el riesgo y tiene ventajas potenciales importantes.

Es una técnica consistente, porque la forma de evaluación del préstamo trata de la misma manera las solicitudes idénticas, por ejemplo, dos personas con las mismas características tendrán el mismo pronóstico de riesgo. Es explícita, se conoce y puede comunicarse el proceso exacto usado para pronosticar el riesgo (forma de evaluación). Permite evaluaciones y administración de riesgo refinadas. Puede probarse antes de usarlo , por ejemplo, una forma de calificación recién diseñada puede probarse para pronosticar el riesgo de los préstamos vigentes en la actualidad, usando solamente características conocidas al momento del desembolso. Este riesgo estimado puede compararse con el riesgo observado en la práctica hasta la fecha. Este procedimiento revela cómo habría funcionado la evaluación si hubiera estado en aplicación al momento de los desembolsos de los préstamos actualmente vigentes. El Scoring revela las relaciones

entre el riesgo y las características del prestatario, el préstamo, y el prestamista. Permite estimar el efecto del scoring en la rentabilidad, suponga que se conoce el costo neto de un préstamo "malo"que está aprobado y que también se conociera la utilidad neta de un préstamo "bueno". Dado el desempeño de la de calificación en la prueba histórica, el prestamista podría estimar el efecto directo sobre las utilidades de préstamos a diferentes tipos de clientes (Mark Schreiner).

Estas ventajas han hecho que esta técnica desarrollada inicialmente para el sistema financiero, sea aplicada a otros campos como en inteligencia de mercados. Por ejemplo, el Lead Scoring es una puntuación que se va asociando a cada uno de los contactos o leads en función de su comportamiento, es decir, una manera de dar una valoración numérica a las oportunidades de ventas, que se irán incrementando a medida que se vaya demostrando más interés en los contenidos web, e-mails, etc, sirve para conocer el grado de interés de nuestros prospectos o leads según las interacciones con nuestro ecosistema digital[1].

En Segmentación de campañas: dado el valor del cliente(score), se puede determinar la conveniencia de incluirlo en una campaña de recuperación o retención, de fidelización, venta de productos complementarios (up-selling) y adquisición de nuevos clientes.

En Análisis de ciclo de vida de un cliente: se puede predecir, mediante el cambio de valor(score) del cliente en el tiempo, en qué etapa de madurez está. Con esta información se podría determinar el momento de actuar y conocer la conveniencia de incluirlo en una campaña de venta[2]. Berson et al. (2000) define el çustomer churnçomo un término usado para denotar el movimiento de cliente de un proveedor a otro y çhurn management.[es] el término para describir el proceso que la empresa hace para retener a clientes rentables. La técnica de scoring también está siendo usada en el çhurn management".

La metodología de credit scoring no ha sido aplicada directamente para modelar el efecto de campañas publicitarias. Sin embargo, se conoce una aproximación de este modelos a partir de datos simulados (Lo,Victor. 2002). En este trabajo nos proponemos hacer una aplicación a un problema real para datos de una empresa de telecomunicaciones.

Objetivo general

Desarrollar un modelo de probabilidad de compra de un paquete de Voz-SMS asociado al efecto de la campaña publicitaria.

Objetivos específicos

- Construir un modelo que explique el evento compra de un paquete Voz-SMS a partir de las variables originales y artificiales correspondientes a las dimensiones: cliente, consumo, equipo, fidelidad, linea, PQR's, promociones, recargas, tráfico, valor, bonos, venta y otras.
- Desarrollar un método de optimización que explique la compra vs la campaña publicitaria.
- Encontrar segmentos propensos a la compra según el efecto de la campaña de marketing.

LOS DATOS

Se cuenta con 22 matrices de datos. Cada una corresponde a un mes a partir del Octubre del 2012 hasta Junio del 2014. Cada matriz tiene en filas los usuarios o líneas telefónicas y este número puede variar de mes a mes, entre 1'340.000 y 1'805.000, disminuir de acuerdo a las cancelaciones y/o suspensiones hechas durante el mes, o aumentar según el número de nuevos clientes y reactivaciones.

[1] http://www.markitude.com/que-es-y-para-que-sirve-el-lead-scoring/
[2] https://sites.google.com/site/jojooa/crm/definicion-de-score-de-cliente-que-es-el-score-de-cliente

El número de columnas es de 197 que corresponden a 174 variables continuas y 23 variables categóricas, que son las variables originales, registradas mensualmente por la compañía de telecomunicaciones. Estas variables están agrupadas en 14 dimensiones (Tablas 1 y 2):

# Variable	Dimensión	Contenido
1	IDENTIFICADOR	Fecha de corte de los datos (mensual)
2	IDENTIFICADOR	Número de celular prepago (MSISDN)
3	IDENTIFICADOR	Número de Identificación del cliente (Nro.DNI / Nro.RUC, etc) (p.e. DNI: 29240401)
4	IDENTIFICADOR	Número de identificación y número celular
5	BONOS	Total minutos de bono
6	BONOS	Total SMS de bono
7	BONOS	Total de MB de Datos de bonos y promociones
8	BONOS	Total consumo de minutos de bono y promociones
9	BONOS	Total consumo de SMS de bono y promociones
10	BONOS	Total consumo de MB de bono y promociones
11	BONOS	Cantidad de bonos entregados - Minutos
12	BONOS	Cantidad de bonos entregados - SMS
13	BONOS	Cantidad de bonos y promociones entregados - MB
14	BONOS	Consumo de $ promocionales y de bonos
15	BONOS	Total de $ otorgados por promociones
16	BONOS	Total de $ otorgados por bonos
17	BONOS	Promociones de $ entregadas
18	BONOS	Bonos de $ entregados
19	BONOS	$ gastados en promociones
20	CLIENTE	Tipo de la identificación del cliente, 1=cedula,2=nit,3=pasaporte,4=cedula Extranjeria
21	CLIENTE	Fecha de nacimiento del cliente
22	CLIENTE	Gênero del cliente (Masculino, Femenino)
23	CLIENTE	Ubigeo del Cliente (departamento)
24	CLIENTE	Ubigeo del Cliente (provincia)
25	CLIENTE	Ubigeo del Cliente (distrito)
26	CLIENTE	Agrupar DEPARMENTO_CD y PROVINCIA_CD (Lima, Lima Provincias, Norte, Sur y Centro)
27	CLIENTE	Total de Líneas Prepago que tiene un cliente (evaluar por Nro.Documento + Línea)
28	CLIENTE	Cantidad de líneas Prepago en estado Preactive (preactiva)(ver hoja de Negocio)
29	CLIENTE	Cantidad de líneas Prepago en estado ACTIVE (activa)(ver hoja de Negocio)
30	CLIENTE	Cantidad de líneas Prepago en Grace (periodo de gracia)(ver hoja de Negocio)
31	CLIENTE	Cantidad de líneas Prepago en estado Recycle(reciclada))(ver hoja de Negocio)
32	CLIENTE	Total de líneas Postpago
33	CLIENTE	Cantidad de líneas Postpago Activas
34	CLIENTE	Cantidad de líneas Postpago Suspendida
35	CONSUMO	Número de Conexiones de trafico de datos (cantidad de sesiones)
36	CONSUMO	Cantidad de Kb (subida/bajada) de Tráfico de datos (Incluye tráfico Blackberry)
37	CONSUMO	Número de intentos exitosos de inicio de sesión a APN blackberry.net.
38	CONSUMO	Cantidad de Kb (subida/bajada) de Tráfico cursado de Blackberrys
39	CONSUMO	Cantidad de Kb (subida/bajada) de Tráfico de datos (Incluye tráfico Blackberry) - Granel
40	EQUIPO	Marca de terminal según último tráfico
41	EQUIPO	Flag si el equipo fue adquirido en tienda o no
42	FIDELIDAD	Total puntos Club
43	FIDELIDAD	Total puntos canjeados Club
44	FIDELIDAD	Total de productos canjeados Club
45	FIDELIDAD	Cantidad de paquetes de minutos canjeados
46	FIDELIDAD	Cantidad de minutos canjeados
47	FIDELIDAD	Cantidad de paquetes de SMS canjeados
48	FIDELIDAD	Cantidad de SMS canjeados
49	FIDELIDAD	Cantidad de paquetes de Datos canjeados
50	FIDELIDAD	Cantidad de Kb canjeados

# Variable	Dimensión	Contenido
51	FIDELIDAD	Segmento definido por el área de Fidelidad
52	FIDELIDAD	Cantidad de recargas canjeados
53	FIDELIDAD	Cantidad de $ en recargas canjeados
54	FIDELIDAD	Cantidad de merchandasing canjeados
55	FIDELIDAD	Cantidad de equipos canjeados
56	FIDELIDAD	$ de los equipos canjeados
57	LINEA	Fecha de Activación de la línea
58	LINEA	Fecha de la última afiliación a números frecuentes
59	LINEA	Cantidad de bloqueos por desconocimiento por titularidad
60	LINEA	Cantidad de bloqueos de la línea por pérdida, robo o hurto
61	LINEA	Cantidad de cambios de opción de trios (numeros frecuentes - elegidos)
62	LINEA	Cantidad de cambio de trios (numeros frecuentes - elegidos)
63	LINEA	Cantidad de días desde que la línea fue activada (sysdate - fecha_activacion)
64	LINEA	Cantidad de días de una línea bloqueada por pérdida, robo o hurto
65	LINEA	Cantidad de días en estado bloqueado por desconocimiento de titularidad
66	LINEA	Cantidad de días de último estado (SYSDATE - FECHA_ULTIMO_ESTADO_DTTM)
67	LINEA	Cantidad de números frecuentes de SMS afiliados a la línea
68	LINEA	Cantidad de números frecuentes de VOZ afiliados a la línea
69	LINEA	Fecha de último cambio de estado
70	LINEA	Cantidad de veces que entró en estado Grace capa 1
71	LINEA	Cantidad de veces que entró en estado Grace capa 2
72	LINEA	Cantidad de veces que entró en estado Grace capa 3
73	LINEA	Cantidad de veces que entró en estado Grace capa 4
74	LINEA	Cantidad de veces que entró en estado Grace capa 5
75	LINEA	Plan Comercial asociado a la línea
76	LINEA	Cantidad de veces que entró en estado Recycle capa 1
77	LINEA	Cantidad de veces que entró en estado Recycle capa 2
78	LINEA	Fecha de última renovación
79	LINEA	Cantidad de veces que se REPONE un pack (línea/equipo)
80	LINEA	Cantidad de veces que se REPONE un pack (línea/equipo) por motivos de robo o pérdida
81	LINEA	Fecha de Última reposición
82	LINEA	Cantidad de veces que se RENUEVA un equipo o chip
83	LINEA	Cantidad de veces que se RENUEVA un equipo o chip por motivos de robo o pérdida
84	LINEA	Tipo de Numero de Frecuente. Opción Numeros frecuentes Actual (numeral)
85	LINEA	Estado de línea (Preactive, Active, Grace 1 al 5, Recycle zb21 zb22)
86	OTROS	Cantidad de visitas a CAC en el mes
87	OTROS	Cantidad de visitas a CAC en el mes
88	OTROS	Cantidad de pedidos en el mes referente a la línea
89	OTROS	Cantidad de quejas en el mes referente a la línea
90	OTROS	Cantidad de reclamos en el mes referente a la línea
91	OTROS	Número de SMS Salientes de Tipo Comercial en el mes (GSVA)
92	OTROS	Número de SMS Salientes de Tipo Informativo en el mes (GSVA)
93	OTROS	Cantidad de caídas de servicio en ubigeo de cliente
94	OTROS	Cobertura de servicio en ubigeo de cliente voz (tecnología mas alta)
95	OTROS	Cobertura de servicio en ubigeo de cliente datos (tecnología mas alta)
96	OTROS	Cantidad de números origen que llamaron
97	OTROS	Cantidad de números destino a los que llamó
98	OTROS	Cantidad de días de la última llamada
99	OTROS	Cantidad de días del último mensaje
100	OTROS	Número de días con llamada saliente en el mes

Tabla 1: Lista de variables

# Variable	Dimensión	Contenido
101	OTROS	Número de días con llamada entrante en el mes
102	OTROS	Número de días con mensaje saliente en el mes
103	OTROS	Número de días con mensaje entrante en el mes
104	OTROS	Cantidad de servicios prepago
105	OTROS	Cantidad de servicios bam prepago
106	OTROS	Cantidad de servicios postpago
107	OTROS	Cantidad de servicios bam postpago
108	OTROS	Cantidad de servicios telefonía fija (TFI Pre y Post)
109	OTROS	Cantidad de servicios telefonía fija HFC
110	OTROS	Cantidad de servicios TV Cable HFC
111	OTROS	Cantidad de servicios Internet HFC
112	OTROS	Cantidad de servicios 2Play
113	OTROS	Cantidad de servicios TV Satelital
114	PQRs	Fecha del Tickler Más Reciente
115	PQRs	Estado del Tickler Más Reciente
116	PROMOCIONES	Total minutos de promociones
117	PROMOCIONES	Total SMS de promociones
118	PROMOCIONES	Cantidad de promociones entregados - Minutos
119	PROMOCIONES	Cantidad de promociones entregados - SMS
120	RECARGAS	Total $ de recargas al mes
121	RECARGAS	Total Cantidad de recargas al mes
122	RECARGAS	Valor de recargas Físicas
123	RECARGAS	Cantidad de recargas Físicas
124	RECARGAS	Valor de recargas Virtuales
125	RECARGAS	Cantidad de recargas Virtuales
126	RECARGAS	Fecha de la última recarga
127	RECARGAS	Días desde la ultima recarga (SYSDATE - RECARGA_DTTM)
128	RECARGAS	Canal por el que se realizó la recarga (USSD, WEB, IVR, POS,etc)
129	TRAFICO	Número de llamadas entrantes OFFNET - Movil
130	TRAFICO	Número de llamadas entrantes OFFNET - Fijos
131	TRAFICO	Número de llamadas entrantes ONNET - Fijo
132	TRAFICO	Número de llamadas entrantes ONNET - Móvil
133	TRAFICO	Número de llamadas Salientes al 123
134	TRAFICO	Numero de llamadas al 123 con duración de la conversación > 30 segundos
135	TRAFICO	Número de Llamadas a Números Frecuentes (elegidos)
136	TRAFICO	Número de Llamadas Facturables a Números Frecuentes (elegidos) - Granel
137	TRAFICO	Número de Llamadas Salientes OFFNET - Movistar
138	TRAFICO	Número de Llamadas Salientes OFFNET - Nextel
139	TRAFICO	Número de Llamadas Salientes OFFNET - Viettel
140	TRAFICO	Número de Llamadas Salientes OFFNET - LDI
141	TRAFICO	Número de Llamadas Salientes OFFNET - Fijos
142	TRAFICO	Número de Llamadas Salientes ONNET - Fijo
143	TRAFICO	Número de Llamadas Salientes ONNET - Móvil
144	TRAFICO	Cantidad de minutos entrantes OFFNET (segundos)
145	TRAFICO	Cantidad de minutos entrantes OFFNET - Fijos
146	TRAFICO	Cantidad de minutos entrantes ONNET - Fijo
147	TRAFICO	Cantidad de minutos entrantes ONNET - Móvil
148	TRAFICO	Cantidad de Minutos Salientes al 123
149	TRAFICO	Cantidad de Minutos Salientes al 123 con duración de la conversación > 30 segundos
150	TRAFICO	Cantidad de Minutos Salientes a Números Frecuentes (elegidos)

# Variable	Dimensión	Contenido
151	TRAFICO	Cantidad de Minutos Salientes Facturables a Números Frecuentes (elegidos) - Granel
152	TRAFICO	Cantidad de Minutos Salientes OFFNET - Movistar
153	TRAFICO	Cantidad de minutos Salientes OFFNET - Nextel
154	TRAFICO	Cantidad de minutos Salientes OFFNET - Viettel
155	TRAFICO	Cantidad de minutos Salientes OFFNET - Fijos
156	TRAFICO	Cantidad de minutos Salientes ONNET - Fijo
157	TRAFICO	Cantidad de minutos Salientes ONNET - Móvil
158	TRAFICO	Número de SMS Entrantes OFFNET
159	TRAFICO	Número de SMS Entrantes ONNET - Móvil
160	TRAFICO	Número de SMS Salientes OFFNET
161	TRAFICO	Número de SMS Salientes ONNNET - Móvil
162	TRAFICO	Número de Llamadas Salientes OFFNET - Granel
163	TRAFICO	Número de Llamadas Salientes OFFNET - Fijos - Granel
164	TRAFICO	Número de Llamadas Salientes ONNET - Fijo - Granel
165	TRAFICO	Numero de Llamadas Salientes ONNET - Móvil - Granel
166	TRAFICO	Cantidad de Minutos Salientes OFFNET - Granel
167	TRAFICO	Cantidad de minutos Salientes OFFNET - Fijos - Granel
168	TRAFICO	Cantidad de minutos Salientes OFFNET - LDI - Granel
169	TRAFICO	Cantidad de minutos Salientes ONNET - Fijo - Granel
170	TRAFICO	Cantidad de minutos Salientes ONNET - Móvil - Granel
171	TRAFICO	Número de SMS Salientes OFFNET -Granel
172	TRAFICO	Número de SMS Salientes ONNNET - Móvil - Granel
173	VALOR	Valor de Llamadas Salientesfacturables a Num Frecuentes
174	VALOR	Valor de Llamadas Salientes OFFNET - Movistar
175	VALOR	Valor de Llamadas Salientes OFFNET - Nextel
176	VALOR	Valor de Llamadas Salientes OFFNET - Viettel
177	VALOR	Valor de Llamadas Salientes OFFNET - Fijos
178	VALOR	Valor de Llamadas Salientes ONNET - Fijo
179	VALOR	Valor de Llamadas Salientes ONNET - Móvil
180	VALOR	Valor de Mensajes salientes OFFNET
181	VALOR	Valor de Mensajes Saliente ONNET - Móvil
182	VALOR	Valor de Kb (subida/bajada) de Tráfico de datos
183	VALOR	Valor de kb de Blackberry
184	VALOR	Valor de Llamadas Salientesfacturables a Num Frecuentes - Granel
185	VALOR	Valor de Llamadas Salientes OFFNET - Granel
186	VALOR	Valor de Llamadas Salientes OFFNET - Fijos - Granel
187	VALOR	Valor de Llamadas Salientes ONNET - Fijo - Granel
188	VALOR	Valor de Llamadas Salientes ONNET - Móvil - Granel
189	VALOR	Valor de Mensajes salientes OFFNET - Granel
190	VALOR	Valor de Mensajes Saliente ONNET - Móvil - Granel
191	VALOR	Valor de Kb (subida/bajada) de Tráfico de datos - Granel
192	VENTA	Código del PUNTO DE VENTA. (donde fue vendida la línea) (p.e. CAC, DAC, Cadena, Bodega)
193	VENTA	Campaña de Venta
194	VENTA	Fecha de venta de la línea
195	VENTA	Tipo de Chip (1,8,15,Pack) o los que hay
196	VENTA	Fecha de la última venta de una línea Prepago
197	VENTA	Fecha de la última venta de un Postpago

Tabla 2: Lista de variables

Variable Objetivo

Se define la variable objetivo del modelo, como la compra o no compra el paquete adicional del Voz-SMS, tomará valores 1 y 0, de la siguiente manera:

$$Variable\ Objetivo = \begin{cases} 1, \text{Compra un paquete adicional de Voz-SMS} \\ \\ 0, \text{No compra el paquete adicional de Voz-SMS} \end{cases}$$

Campaña

Se hace contacto con los clientes seleccionados de manera aleatoria y se les ofrece un paquete adicional de voz SMS. Este periodo de campaña dura un mes y es denominado offset porque el registro de la efectividad de la campaña se tiene solo hasta inicios del siguiente mes.

Se espera que compren el paquete y lo mantengan por lo menos un mes adicional al mes de la compra. Si compra el paquete, la activación se hace el primer día del siguiente mes. Mes denominado de activación. Si lo mantienen un mes adicional al mes de a compra, dicho mes se denomina de mantenimiento.

Construcción de la ABT (Analitic Base Table)

La base de datos sobre la que se genera el modelo se construye a partir de 12 bases llamadas ventanas de tiempo.

Una ventana de tiempo es una base de datos, formada por 10 meses consecutivos, distribuidos de la siguiente manera:

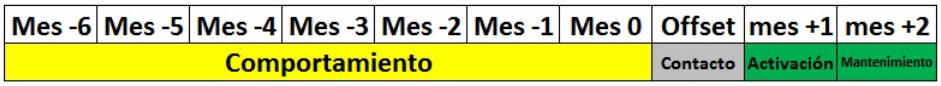


Figura 1: Ventana de tiempo

El mes 0 corresponde al final del periodo de evaluación del comportamiento, donde se aplicara el modelo en vez de hacer una selección aleatoria de los clientes.

Para la ventana 1, se tienen en cuenta los n individuos activos en el mes 0 y que hayan tenido por lo menos 4 meses de antigüedad consecutiva. Esto significa que el número de individuos en cada uno de los meses de la ventana, máximo pueden ser n.

A continuación, como ejemplo, la primera ventana de tiempo con sus respectivas fechas:

Mes	Ventana
oct-12	Mes -6
nov-12	Mes -5
dic-12	Mes -4
ene-13	Mes -3
feb-13	Mes -2
mar-13	Mes -1
abr-13	Mes 0
may-13	Offset
jun-13	mes +1
jul-13	mes +2

Figura 2: Ejemplo de una Ventana de tiempo

Las dimensiones de la base de datos de la ventana 1 son: máximo 7n líneas y 197 columnas correspondientes a las variables (véase: Lista de variables).

Las 12 ventanas de tiempo construidas con las 22 fechas disponibles, tienen la siguiente estructura y número de registros:

	Ventana 1	Ventana 2	Ventana 3	Ventana 4	Ventana 5	Ventana 6	Ventana 7	Ventana 8	Ventana 9	Ventana 10	Ventana 11	Ventana 12
oct-12	Mes -6											
nov-12	Mes -5	Mes -6										
dic-12	Mes -4	Mes -5	Mes -6									
ene-13	Mes -3	Mes -4	Mes -5	Mes -6								
feb-13	Mes -2	Mes -3	Mes -4	Mes -5	Mes -6							
mar-13	Mes -1	Mes -2	Mes -3	Mes -4	Mes -5	Mes -6						
abr-13	Mes 0	Mes -1	Mes -2	Mes -3	Mes -4	Mes -5	Mes -6					
may-13	Offset	Mes 0	Mes -1	Mes -2	Mes -3	Mes -4	Mes -5	Mes -6				
jun-13	mes +1	Offset	Mes 0	Mes -1	Mes -2	Mes -3	Mes -4	Mes -5	Mes -6			
jul-13	mes +2	mes +1	Offset	Mes 0	Mes -1	Mes -2	Mes -3	Mes -4	Mes -5	Mes -6		
ago-13		mes +2	mes +1	Offset	Mes 0	Mes -1	Mes -2	Mes -3	Mes -4	Mes -5	Mes -6	
sep-13			mes +2	mes +1	Offset	Mes 0	Mes -1	Mes -2	Mes -3	Mes -4	Mes -5	Mes -6
oct-13				mes +2	mes +1	Offset	Mes 0	Mes -1	Mes -2	Mes -3	Mes -4	Mes -5
nov-13					mes +2	mes +1	Offset	Mes 0	Mes -1	Mes -2	Mes -3	Mes -4
dic-13						mes +2	mes +1	Offset	Mes 0	Mes -1	Mes -2	Mes -3
ene-14							mes +2	mes +1	Offset	Mes 0	Mes -1	Mes -2
feb-14								mes +2	mes +1	Offset	Mes 0	Mes -1
mar-14									mes +2	mes +1	Offset	Mes 0
abr-14										mes +2	mes +1	Offset
may-14											mes +2	mes +1
jun-14												mes +2
7n=	**1.340.743**	**1.386.926**	**1.427.220**	**1.485.460**	**1.562.470**	**1.605.795**	**1.655.214**	**1.703.196**	**1.750.527**	**1.805.513**	**1.741.506**	**1.680.646**

Tabla 3: Estructura de las ventanas de tiempo

La ABT está formada por el apilamiento de las 12 ventanas anteriores y tiene 19'145.216 filas y 197 columnas.

Variables artificiales

Son variables que se construyen a partir de las 174 variables continuas de las base de datos, resumen la información a través del periodo de comportamiento para cada ventana. Estas variables ayudan a suavizar los valores de comportamiento de compra de los clientes, para alimentar el modelo con tendencias, medidas de dispersión, etc.

Las variables construidas van distinguidas por un sufijo; a continuación se presentan los sufijos utilizados que distinguen los cálculos incluidos en la tabla input del modelo:

- **Promedios con Sufijo "_M3":** Promedio de una variable en los meses M-1, M-2 y M-3, obte-

niendo 174*12 nuevas variables.

- **Promedios con Sufijo "_M6":** Promedio de una variable en los meses M-1, M-2, M-3, M-4, M-5 y M-6, obteniendo 174*12 nuevas variables.
- **Ratios con Sufijo "_R3":** Cociente entre el valor observado en el mes M0 y el promedio de la variable en los meses M-1, M-2 y M-3, obteniendo 174*12 nuevas variables.
- **Ratios con Sufijo "_R6":** Cociente entre el valor observado en el mes M0 y el promedio de la variable en los meses M-1, M-2, M-3, M-4, M-5 y M-6, obteniendo 174*12 nuevas variables.
- **IFR con Sufijo "_IF":** El Índice de fuerza relativa (RSI - Relative Strength Index) fue desarrollado por Welles Wilder en 1978 y refleja los cambios relativos entre los valores más altos y más bajos. Se trata de un oscilador, ya que los valores de este indicador varían entre 0 y 100. La fórmula que se utiliza para calcular el IFR es la siguiente:

$$RSI_t = 100 - \left(\frac{100}{1 + \frac{U_t}{D_t}} \right)$$

  Siendo: t : un punto concreto en el tiempo, específicamente el mes 7 o periodo de observación. U_t: el número de periodos de tiempo en que el valor **subió** con respecto al periodo inmediatamente anterior. D_t: el número de periodos de tiempo en que el valor **bajó** con respecto al periodo inmediatamente anterior.

  Este indicador se calcula con para cada variables en los meses M0 a M-6 (periodo de comportamiento), obteniendo 174*12 nuevas variables.
- **Desviación Estándar con Sufijo "_SD":** Corresponde a la desviación estándar de la variable de los meses M-1 a M-6 (seis meses de historia), obteniendo 174*12 nuevas variables.

Al finalizar el proceso descrito anteriormente, la ABT cuenta con **19'145.216 filas** y un total de **1.241 columnas**: 23 variables categóricas, 174 variables continuas originales y 1.044 variables continuas artificiales.

La base de datos

Después de construida la ABT, se requiere hacer un análisis previo que permita determinar la base sobre la cual se va a construir el modelo. Para ello se deben tener en cuenta:

- Exclusiones

  Dentro del contexto del negocio y del modelamiento. Algunas de estas exclusiones solamente se tienen en cuenta para el desarrollo del score, ya que cuando el modelo se aplique a las bases de datos generadas mensualmente, aún no se conoce el resultado del periodo de activación, por tanto, las exclusiones para dicho periodo no se pueden hacer.

  A continuación se presentan las exclusiones que se hacen a la ABT con el fin de generar la base de datos para el desarrollo del modelo:

  - Se excluyen los clientes que en el periodo de activación (PA) no se encuentran activos. (Exclusión 1)
  - Se excluye a los clientes que son Persona Jurídica y se deja a los que son Persona Natural. (Exclusión 2)
  - Se excluye a los contratos que tengan menos de 4 meses de antigüedad consecutiva, Mes-0, Mes-1,Mes-2,Mes-3 de cada ventana de tiempo. (Exclusión 3)

- Se excluye a aquellos clientes que se encuentran registrados en la lista antifraude. (Exclusión 4)

En la siguiente tabla se presentan los resultados de la aplicación de estas exclusiones:

Exclusión	Label	No Compradores	Compradores	Total	% Compradores	% No Compradores	% Columna
1	No activos en PA	1.355.673	41.928	1.397.601	3,00%	97,00%	7,30%
2	Persona Jurídica	222.431	1.568	223.999	0,70%	99,30%	1,17%
3	Antigüedad menor a 4 meses	1.603.374	4.825	1.608.198	0,30%	99,70%	8,40%
4	Lista antifraude	13.288	114	13.402	0,85%	99,15%	0,07%
	Lineas a calificar con el Score	15.810.098	93.833	15.903.931	0,59%	99,41%	83,07%
Total	Total	19.038.003	107.213	19.145.216	0,56%	99,44%	100,00%

Tabla 4: Exclusiones del modelo

Después de aplicar las exclusiones, se obtuvo el 83.07 %, de los registros de la ABT inicial, **15'903.931**, con una tasa de evento o de compra del 0.59 % (93.833).

Esta tasa de evento es muy pequeña, por lo que se debe ajustar la muestra con la técnica mostrada a continuación.

- Oversampling Dado que la tasa de compradores es muy pequeña (0.59 %), el modelo tendrá problemas para identificar el evento. Por esta razón se debe hacer un ajuste a la muestra de entrenamiento, técnica llamada **oversampling** (Scott, A.J. and Wild, C.J. 1986), que consiste en construir una muestra de entrenamiento con **312.777 filas**, los 93.833 compradores, y 218.944 no compradores seleccionados aleatoriamente; de esta forma la muestra de entrenamiento tiene un 30 % de compradores para generar el modelo.

Luego de aplicar las exclusiones y realizar el oversampling, la base de datos sobre la que se aplicará el modelo, cuenta con **312.777** filas y un total de **1.241** columnas: 23 variables categóricas, 174 variables continuas originales y 1.044 variables continuas artificiales.

Modelos de auto-aprendizaje

En algunas técnicas de minería de datos, se obtiene un modelo de conocimiento, que representa patrones de comportamiento observados en los valores de las variables del problema o relaciones de asociación entre dichas variables. También pueden usarse varias técnicas a la vez para generar distintos modelos, aunque generalmente cada técnica obliga a un preprocesado diferente de los datos; proceso denominado construcción del modelo. Posteriormente se realiza la interpretación y evaluación de datos, una vez obtenido el modelo. Para esto se debe proceder a su validación comprobando que las conclusiones que arroja son válidas y suficientemente satisfactorias. En el caso de haber obtenido varios modelos mediante el uso de distintas técnicas, se deben comparar los modelos en busca de aquel que se ajuste mejor al problema. Si ninguno de los modelos alcanza los resultados esperados, debe alterarse alguno de los pasos anteriores para generar nuevos modelos.

Para la construcción del modelo y su validación, se deben generar dos muestras. En este caso caso, la muestra que fue seleccionada para la construcción del modelo integra las ventanas de Octubre 2013, Noviembre de 2013 y Enero 2014, Marzo 2014. La validación se hará sobre todas las ventanas disponibles.

	Ventana 1	Ventana 2	Ventana 3	Ventana 4	Ventana 5	Ventana 6	Ventana 7	Ventana 8	Ventana 9	Ventana 10	Ventana 11	Ventana 12
oct-12	Mes -6											
nov-12	Mes -5	Mes -6										
dic-12	Mes -4	Mes -5	Mes -6									
ene-13	Mes -3	Mes -4	Mes -5	Mes -6								
feb-13	Mes -2	Mes -3	Mes -4	Mes -5	Mes -6							
mar-13	Mes -1	Mes -2	Mes -3	Mes -4	Mes -5	Mes -6						
abr-13	Mes 0	Mes -1	Mes -2	Mes -3	Mes -4	Mes -5	Mes -6					
may-13	Offset	Mes 0	Mes -1	Mes -2	Mes -3	Mes -4	Mes -5	Mes -6				
jun-13	mes +1	Offset	Mes 0	Mes -1	Mes -2	Mes -3	Mes -4	Mes -5	Mes -6			
jul-13	mes +2	mes +1	Offset	Mes 0	Mes -1	Mes -2	Mes -3	Mes -4	Mes -5	Mes -6		
ago-13		mes +2	mes +1	Offset	Mes 0	Mes -1	Mes -2	Mes -3	Mes -4	Mes -5	Mes -6	
sep-13			mes +2	mes +1	Offset	Mes 0	Mes -1	Mes -2	Mes -3	Mes -4	Mes -5	Mes -6
oct-13				mes +2	mes +1	Offset	Mes 0	Mes -1	Mes -2	Mes -3	Mes -4	Mes -5
nov-13					mes +2	mes +1	Offset	Mes 0	Mes -1	Mes -2	Mes -3	Mes -4
dic-13						mes +2	mes +1	Offset	Mes 0	Mes -1	Mes -2	Mes -3
ene-14							mes +2	mes +1	Offset	Mes 0	Mes -1	Mes -2
feb-14								mes +2	mes +1	Offset	Mes 0	Mes -1
mar-14									mes +2	mes +1	Offset	Mes 0
abr-14										mes +2	mes +1	Offset
may-14											mes +2	mes +1
jun-14												mes +2
Sin restricciones y con Oversampling	22.173	22.937	23.603	24.567	25.840	26.557	26.443	27.210	28.950	28.845	28.801	26.850

Tabla 5: Estructura de las ventanas de tiempo

La base de datos que se tiene al final de este proceso, y que será utilizada para la construcción del modelo, tiene **109.348** filas y **1.241** columnas.

LA METODOLOGÍA

Posterior a la construcción de la base de datos, que en este punto cuenta con **312.777 filas** y un total de **1.241 columnas**, se procede a la implementación de la metodología con la que se dará alcance a los objetivos planteados, esta cuenta con cuatro etapas fundamentales:

- **VARIABLES:** En esta etapa se construye la variable objetivo y, con base en esta, se realiza la selección inicial de variables, que serán el input para la siguiente etapa.
- **MODELAMIENTO:** En esta etapa se construye el algoritmo con el que se modela la variable objetivo a partir de las variables input.
- **VALIDACIÓN DEL MODELO:** En esta etapa se analiza el modelo construido con el fin de darle el mejor uso a sus resultados.
- **SCORE Vs CAMPAÑAS: UP-LIFTING:** Esta etapa corresponde al objetivo general del presente trabajo, es otra técnica de optimización del score.

A continuación se detalla cada una de las etapas que componen la metodología, se dan a conocer los resultados obtenidos y su análisis.

Etapa 1: VARIABLES

Variable objetivo del modelo: Compra de un paquete adicional de Voz-SMS

Recordemos la definición de la variable objetivo del modelo: compra o no compra el paquete adicional del Voz-SMS, tomará valores 1 y 0, de la siguiente manera:

$$Variable\ Objetivo = \begin{cases} 1, \text{Compra un paquete adicional de Voz-SMS} \\ \\ 0, \text{No compra el paquete adicional de Voz-SMS} \end{cases}$$

SELECCIÓN INICIAL DE VARIABLES

La selección inicial de variables predictivas es un paso muy importante ya que tenemos que asegurar que el modelo sea "parsimonioso", es decir, que tenga el mayor poder predictivo posible con el menor numero de variables. El objetivo de este paso es seleccionar una serie de variables que son las de mayor potencial para ser consideradas *a priori* como variables independientes en el modelo final.

Esta selección se puede realizar con diferentes metodologías, como análisis bivariado, análisis discriminante, criterio de experto, arboles de decisión; todas estas son útiles para evaluar las variables antes de decidir cuáles se deben introducir al modelo.

En el presente trabajo, la selección se realiza con el análisis bivariado. Para este análisis es necesario categorizar las variables continuas, luego, analizar el poder predictivo de estas categorías con respecto a la variable objetivo del modelo por medio del Weigth of evidence (WOE), y posteriormente construir el indicador denominado Information Value (IV).

A continuación se describe el proceso para categorizar las variables.

- Weigth of evidence (WOE)

  El WOE mide la fuerza predictiva de cada una de las categorías de las variables en la discriminación entre buenos y malos, es equivalente a la probabilidad que una persona en cada categoría compre o no compre.

  La forma de calcular el WOE, en cada categoría i de la variable, es la siguiente:

$$WOE_i = ln\left(\frac{Distr\ No\ Compradores_i}{Distr\ Compradores_i}\right) \times 100$$

  donde:

$$Distr\ No\ Compradores_i = \frac{No\ compradores\ en\ la\ categoria\ i}{No\ compradores\ en\ la\ Variable}$$

$$Distr\ Compradores_i = \frac{Compradores\ en\ la\ categoria\ i}{Compradores\ en\ la\ Variable}$$

  La multiplicación por 100 se hace para facilitar la lectura de los valores. Los números negativos implican que el atributo particular está prediciendo una mayor proporción de no compradores que de compradores.

- Information Value (IV)

  Una vez se ha calculado el WOE para cada categoría de la variable, se determina el poder predictivo total de la variable, es decir, la capacidad de discriminar compradores de no compradores, para esto se calcula el IV, que es la suma ponderada de los WOE's en cada variable.

$$IV = \sum_{i=1}^{k}(Distr\ No\ Compradores_i - Distr\ Compradores_i) * WOE_i$$

  Siendo k = número de categorías de la variable.

  Con base en esta metodología, una regla general con respecto al Information Value de cada variable (Siddiqui, 2006) es la siguiente:

- Menos de 0.02: No predictiva
- 0.02 a 0.1: Poder predictivo débil
- 01 a 0.3: Poder predictivo medio
- 0.3 0.5: Poder predictivo fuerte
- mas de 0.5: Poder predictivo fuerte

Variables con Information Value mayores a 0.5 pueden ser sobrepredictoras, así que deben ser revisadas con precaución, para determinar si se incluyen o no en el modelo.

Ejemplo del Análisis Bivariado para la variable **M3-Cantidad de números destino a los que llamó durante el mes**, esta variable corresponde al promedio de números marcados en los meses M-0, M-1 y M-2.

M3-Cantidad de números destino a los que llamó durante el mes	Clientes	%	Compradores	% Compradores	WOE	Information Value (IV)
Categoria 01: low-60.2	27773	25%	5372	19%	63%	9%
Categoria 02: 60.2-132.5	37844	35%	11194	30%	7%	0%
Categoria 03: 132.5-304.5	35317	32%	13517	38%	-32%	3%
Categoria 04: 304.5-high	8414	8%	3970	47%	-68%	4%
Total	**109348**	**100%**	**34053**	**31%**		**16%**

Tabla 6: Tabla WOE

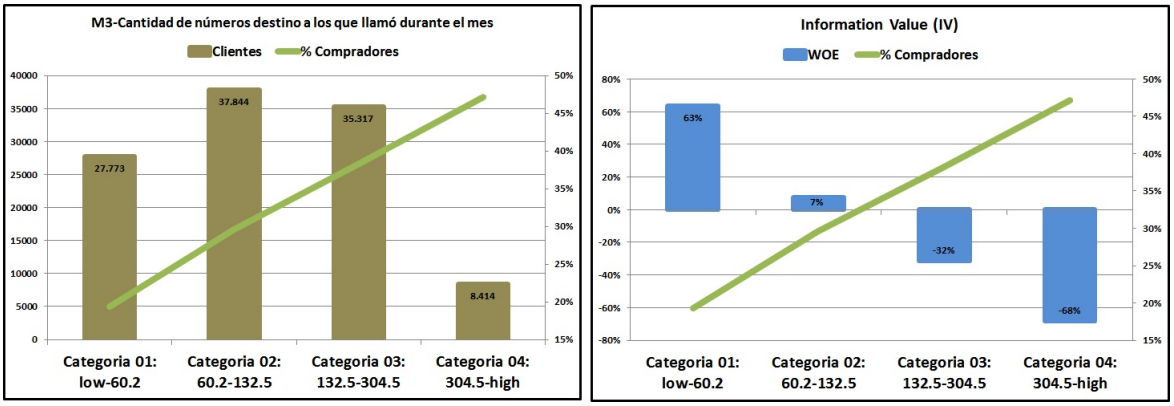


Figura 3: Análisis Gráfico WOE

Como se observa en el ejemplo anterior, la categoría con mayor tasa de compradores es la número 4, mas de 304.5 minutos en promedio en los últimos tres meses (M3), lo ideal al categorizar cada variable, y hacer el análisis del WOE, es lograr un ordenamiento ascendente o descendente por la **tasa de compra**, con esto se asegura la capacidad de discriminación de cada variable y se logra una mejor interpretación de los resultados. Algo igualmente importante a tener en cuenta es que las categorías no tengan una cantidad muy pequeña de clientes (menores a 0.5 %), ya que cuando el modelo este calificando a los usuarios las categorías muy pequeñas pueden desaparecer, con esto se perdería el ordenamiento logrado en la construcción del modelo; tampoco categorías muy grandes (mayores a 95 %) ya que estas categorías atrapan todo el evento modelado y no se logra una discriminación adecuada.

Sobre la variable del ejemplo anterior, se puede observar que cumple con los criterios de selección, y tiene un information value de 16.3 %, es decir una poder predictivo medio, con lo que se determina que va a ser pre-seleccionada para el modelo.

De esta manera se evalúan todas las variables de la base de datos, y de acuerdo con el criterio de escogencia mencionado anteriormente, se selecciona el conjunto de variables descrito a continuación.

Variables Pre-Seleccionadas

Una vez que ha sido calculado el IV para cada variable, se seleccionan las que tengan un mejor desempeño en el indicador, entre 0.02 y 0.5, a continuación la primera selección de variables que cuenta con 55 variables.

N°	Label	IV	Origen
1	C3_Flag cantidad de minutos adicionales salientes OFFNET del plan a Movistar	0,49	Original
2	Cantidad de números origen que llamarona a una línea	0,41	Original
3	Porcentaje de consumo de soles incluidos en el plan. Para planes actuales	0,27	Original
4	Tipo Tope de consumo (1=T.Abierto/ 2=Cero/ 3=Automático/ 4=Variable, etc)	0,27	Original
5	Número de llamadas entrantes ONNET	0,24	Original
6	Cantidad de números destino a los que llamó	0,24	Original
7	Número de visitas al CAC	0,12	Original
8	Cantidad de minutos salientes ONNET a RPC Claro	0,11	Original
9	Número de SMS entrantes ONNET	0,11	Original
10	C6_Flag cantidad de minutos adicionales salientes OFFNET del plan a Movistar	0,09	Original
11	Número de días con mensaje entrante en el mes	0,08	Original
12	C6_Flag cantidad de Kb adicionales (subida/ bajada) de tráfico GPRS	0,02	Original
13	R6_Cantidad de minutos adicionales Salientes ONNET	0,49	Construida
14	M6_Número de visitas al CAC	0,46	Construida
15	M2_Cantidad de minutos adicionales Salientes ONNET	0,46	Construida
16	M6_Cantidad de minutos Entrantes OFFNET - Nextel	0,45	Construida
17	R3_Cantidad de minutos adicionales Salientes ONNET	0,45	Construida
18	R3_Número de llamadas Adicionales Salientes OFFNET al plan a Movistar	0,42	Construida
19	R6_ Cantidad de Kb incluidos en el plan	0,41	Construida
20	R6_Valor de llamadas Adicionales Salientes ONNET	0,41	Construida
21	M6_Cantidad de minutos Incluidos ONNET	0,39	Construida
22	M3_Cantidad de minutos Incluidos ONNET	0,39	Construida
23	M6_ Cantidad de días en que consume todos sus soles incluidos	0,39	Construida
24	M3_ Cantidad de Kb incluidos en el plan	0,37	Construida
25	M6_Cantidad de numeros frecuentes de VOZ asociados a la línea	0,36	Construida
26	IF_Cantidad de minutos incluidos ONNET	0,36	Construida
27	M6_Porcentaje de consumo de soles incluidos en el plan. Para planes actuales	0,34	Construida
28	M6_Cantidad de números destino a los que llamó	0,33	Construida
29	R3_Porcentaje de consumo de soles incluidos en el plan. Para planes actuales	0,33	Construida
30	M3_Valor de recargas por canal Recargas de Bodegas	0,29	Construida
31	R6_Porcentaje de consumo de soles incluidos en el plan. Para planes actuales	0,27	Construida
32	M6_Número de SMS Entrantes ONNET	0,26	Construida
33	M2_ Cantidad de días en que consume todos sus soles incluidos	0,22	Construida
34	R6_Número de llamadas Adicionales Salientes ONNET	0,22	Construida
35	M2_Número de días con mensajes entrantes en el mes	0,22	Construida
36	R3_Número de llamadas Adicionales Salientes ONNET	0,19	Construida
37	R6_Número de llamadas Adicionales Salientes OFFNET al plan a Movistar	0,19	Construida
38	M3_Proporción de mensajes salientes sobre el total e mensajes	0,18	Construida
39	R3_ Cantidad de Kb incluidos en el plan	0,18	Construida
40	R3_Cantidad de minutos Incluidos ONNET	0,18	Construida
41	M3_Cantidad de números destino a los que llamó	0,16	Construida
42	M3_Cantidad de días en que consume todos sus MMS incluidos	0,12	Construida
43	M3_ Cantidad de días en que consume todos sus soles incluidos	0,09	Construida
44	M3_Proporción de uso de Kb en GPRS	0,08	Construida
45	M3_Valor cargo fijo mensual de paquetes	0,07	Construida
46	M6_Valor de mensajes descargas	0,07	Construida
47	R3_Valor de llamadas Adicionales Salientes ONNET	0,06	Construida
48	M3_Porcentaje de consumo de soles incluidos en el plan. Para planes actuales	0,06	Construida
49	M2_Número de llamadas Adicionales Salientes ONNET	0,06	Construida
50	M2_ Cantidad de días en que consume todos sus SMS incluidos	0,06	Construida
51	M2_Valor de llamadas Adicionales Salientes ONNET	0,06	Construida
52	M6_Valor promedio recarga ventanilla	0,05	Construida
53	M2_Cantidad de números destino a los que llamó	0,05	Construida
54	R6_Cantidad de minutos incluidos ONNET	0,04	Construida
55	M2_Porcentaje de consumo de soles incluidos en el plan. Para planes actuales	0,02	Construida

Tabla 7: Selección inicial de variables

En este conjunto de variables se pueden observar aspectos interesantes:

- Se tienen 43 variables construidas, lo que indica que estas tienen mejor poder de predicción que las variables originales, de las cuales solo se tienen 12 en esta pre-selección.
- De las variables construidas, 28 son de promedios (M2, M3, M6), esto debido a la estabilidad que se logra con el suavizamiento característico de los promedios en general.
- 14 variables son ratios (R3, R6), lo que indica que el comportamiento de un cliente en el ultimo mes tiene un peso alto en la determinación de la intención de compra.
- 1 variable es IF, como se mencionó anteriormente, es una medida de la tendencia de la variable en los 6 meses del periodo de comportamiento.

Luego de esta pre-selección de variables, pasaremos al ajuste de un modelo para variable respuesta dicotómica, este proceso se describe a continuación.

Etapa 2: MODELAMIENTO

Como se mencionó anteriormente, la variable objetivo del modelo es dicotómica, donde el 0 indica la no ocurrencia del evento y el 1 indica que el evento ha ocurrido; para estos modelos se utilizan comúnmente tres técnicas de modelamiento: Arboles de decisión, Regresión Logística y Redes Neuronales; en este caso se utilizó la Regresión Logística,que es una de las herramientas más utilizadas para la estadística aplicada y análisis de datos discretos, básicamente, hay cuatro razones para ello:

1. Tradición, es el modelo mas utilizado para los casos de variable respuesta binaria.
2. Además de la aproximación heurística descrita a continuación, el registro de la cantidad $\frac{p}{(1-p)}$ juega un papel importante en el análisis de tablas de contingencia (los "log odds"). La clasificación es como tener una tabla de contingencia con dos columnas (clases) y un número infinito de filas (valores de x). Con una tabla de contingencia finito, podemos estimar los log odds para cada fila empíricamente, con sólo tomar los recuentos en la tabla. Con una infinidad de filas, necesitamos algún tipo de esquema de interpolación; regresión logística es la interpolación lineal para los log-odds.
3. Está estrechamente relacionado con la distribución de la "familia exponencial", donde la probabilidad de algún vector v es proporcional a $e^{\beta_0} + \sum_{j=1}^{m} f_j(v)\beta_j$.. Si uno de los componentes de v es binaria, y las funciones f_j son toda la función identidad, entonces obtenemos una regresión logística. Las familias exponenciales surgen en muchos contextos en teoría estadística, así que hay un montón de problemas que se pueden solucionar con la regresión logística.
4. Generalmente funciona bien como un clasificador. Y sus resultados son relativamente sencillos de interpretar, tanto para evaluar el modelo como para explicarlo al usuario final.

Modelamiento de probabilidades condicionales

En nuestro caso de estudio, la variable objetivo es discreta. En particular, hay muchas situaciones con este tipo de variable: Llueve o no en una zona determinada, un individuo es portador o no de un virus, un préstamo será pagado, o no será pagado; una persona tendrá una enfermedad cardíaca en los próximos cinco años, o no lo hará. Además de la variables respuesta binaria, se tiene un conjunto de variables de entrada, las cuales pueden ser continuas y/o discretas. Cómo podríamos modelar y analizar esos datos?

se puede tratar de llegar a una regla construida empíricamente la salida binaria de las variables de entrada. Esto se llama clasificación, y es un tema importante por su gran utilización en diferentes contextos de

negocio. Sin embargo, esta forma de clasificación es bastante subjetiva, especialmente porque no hay una regla perfecta, ya que la técnica no podrá tener en cuenta el efecto exacto de cada una de las variables de entrada. Se pueden construir y perfeccionar una serie de reglas de clasificación, que a menudo será útil, pero aun no explicará exactamente a la variable binaria. En definitiva, queremos probabilidades, lo que significa que tenemos que ajustar un modelo estocástico.

El escenario ideal, de hecho, sería tener la distribución condicional de la respuesta Y, teniendo en cuenta las variables de entrada, $Pr(Y|X)$. Esto nos diría que tan precisas son nuestras predicciones. Si nuestro modelo dice que hay una posibilidad de 51 % de llueva, sería menos exacto que si hubiera dicho que había una posibilidad del 99 % de la nieve (aunque incluso una probabilidad del 99 % no es una cosa segura).

Vamos a escoger una de las clases y la llamaremos "1ƶ a la otra "0". (No importa cuál es cuál Entonces Y se convierte en una variable indicadora, y puede decirse de que $Pr(Y = 1) = E(Y)$. Del mismo modo, $Pr(Y = 1|X = x) = E(Y|X = x)$. En una frase: la probabilidad condicional es la esperanza condicional de la variable indicadora. Esto nos ayuda porque a estas alturas sabemos todo acerca de la estimación de las esperanzas condicionales. Lo más fácil para nosotros para hacer en este punto sería la de elegir una nuestro la función de regresión para la variable de indicador; esta será una estimación de la función de probabilidad condicional.

Supongamos que $Pr(Y = 1|X = x) = p(x; \Theta)$, para alguna función parametrizada por Θ, además, se supone que las observaciones son independientes el una de la otra. La función de la probabilidad (condicional) es:

$$\prod_{i=1}^{n} Pr(Y = y_i|X = x_i) = \prod_{i=1}^{n} p(x_i; \Theta)^{y_i}(1 - p(x_i; \Theta)^{1-y_i})$$

Recordemos que en una secuencia de ensayos Bernoulli $y1, ..., yn$, donde hay una probabilidad constante de éxito p, la probabilidad es:

$$\prod_{i=1}^{n} p^{y_i}(1 - p)^{1-y_i}$$

Como se mencionó anteriormente, esta probabilidad se maximiza cuando $p = \hat{p} = n^{-1} \sum_{i=1}^{n} y_i$. Si cada ensayo tenía su propio p_i probabilidad de éxito, esta probabilidad se convierte en:

$$\prod_{i=1}^{n} p_i^{y_i}(1 - p_i)^{1-y_i}$$

Sin algunas restricciones, la estimación del modelo por máxima verosimilitud no funciona; conseguiríamos $p_i = 1$ cuando $y_i = 1, p_i = 0$ cuando $y_i = 0$, y no se llegaría nada. Si por el contrario se supone que el p_i no son sólo números arbitrarios sino que están unidos entre sí, esas limitaciones dan estimaciones de los parámetros no triviales, y vamos a generalizar. En el tipo de modelo que estamos hablando, la restricción, $p_i = p(xi; \Theta)$, nos dice que p_i debe ser el mismo cada vez que x_i es el mismo, y si p es una función continua, entonces valores similares de x_i debe conducir a valores similares de p_i. Suponiendo que p es conocido, la probabilidad es una función de Θ, y podemos estimar Θ mediante la maximización de la probabilidad.

Regresión Logística

Para resumir: tenemos una variable de salida binaria Y, y queremos modelar la probabilidad condicional $Pr(Y = 1|X = x)$ como una función de x; los parámetros desconocidos en la función deben ser estimados por máxima verosimilitud. Por ahora, la primera inquietud seria ¿cómo podemos utilizar la regresión lineal para resolver esto?.

1 La idea más obvia es asumir que $p(x)$ una función lineal de x. Cada incremento de x haría sumar o restar una cantidad a la probabilidad. El problema conceptual aquí es que p toma valores entre 0 y 1, y las funciones lineales no tienen limites. Por otra parte, en muchas situaciones encontramos empíricamente rendimientos decrecientes": el cambio en p requiere un cambio mayor en x cuando p es ya grande (o pequeña) que cuando p es cerca de 1/2. Los modelos lineales no pueden hacer esto.

2 La siguiente idea más obvia es asumir que $p(x)$ una función lineal de x, por lo que el cambio de una variable de entrada se multiplica la probabilidad por una cantidad fija. El problema es que los logaritmos tienen límites en una sola dirección, y funciones lineales no lo son.

3 Por último, la modificación más fácil de $ln(P)$ que tiene una gama ilimitada es la transformación logística (o logit), $ln\frac{P}{1-p}$. Podemos hacer de esto una función lineal de x sin temor a obtener resultados absurdos. (Por supuesto, los resultados todavía podrían ser malos, pero es una buena aproximación.)

Esta última alternativa es la **regresión logística**.

Formalmente, el modelo de regresión logística modelo es:

$$ln\frac{p(x)}{1-p(x)} = \beta_0 + x * \beta$$

Resolviendo para p, esto da:

$$p(x; b; w) = \frac{e^{\beta_0 + x*\beta}}{1 - e^{\beta_0 + x*\beta}} = \frac{1}{e^{-(\beta_0 + x*\beta)}}$$

Observe que esta expresión es mucho más fácil de entender en términos de la probabilidad.

Para minimizar la tasa de errores de clasificación, debemos predecir $Y = 1$ cuando $p?0,5$ y $Y = 0$ cuando $p < 0,5$. Esto significa predecir 1 siempre que $\beta_0 + x * \beta$ no es negativo, y 0 en caso contrario. Así, la regresión logística nos da un clasificador lineal. La expresión que discrimina adecuadamente las dos clases predichas es la solución de $\beta_0 + x * \beta = 0$, que es un punto si x es unidimensional, una línea si se trata de dos dimensiones, etc.

La regresión logística nos da el límite entre las clases, por esto es un buen clasificador. Se hace más robusto, da predicciones más detalladas, y se puede ajustar de una manera diferente; pero esas predicciones robustas podrían estar equivocadas. El uso de regresión logística para predecir las probabilidades con una variable objetivo dicotómica es una opción de modelado, al igual que la regresión lineal es una opción de modelado para predecir variables cuantitativas.

En ninguno de los casos es la adecuación del modelo está garantizada. Empezamos planteando el modelo, para conseguir algo con qué trabajar, y terminamos (si sabemos lo que estamos haciendo) comprobando si realmente lo hace coincidir con los datos, o si tiene defectos sistemáticos.

Función de probabilidad de regresión logística

Debido a la regresión logística predijo probabilidades, en lugar de clases, podemos ajustarlo utilizando verosimilitud. Para cada punto de datos, tenemos un vector de características, x_i, y una clase observada, y_i. La probabilidad de que la clase estaba bien p, si $y_i = 1$, o $1-p$, si $y_i = 0$. La probabilidad es entonces:

$$L(\beta_0, \beta) = \prod_{i=1}^{n} p(x_i)^{y_i}(1 - p(x_i))^{1-y_i}$$

Se podría sustituir en la ecuación real de p, pero las cosas van a ser más claras en un momento si no se hace. Las probabilidades giran en torno a las siguientes sumas:

$$\begin{aligned} l(\beta_0, \beta) &= \sum_{i=1}^{n} y_i * ln(p(x_i)) + (1 - y_i) * ln(1 - p(x_i)) \\ &= \sum_{i=1}^{n} ln(1 - p(x_i)) + \sum_{i=1}^{n} y_i * ln(\frac{p(x_i)}{1 - p(x_i)}) \\ &= \sum_{i=1}^{n} ln(1 - p(x_i)) + \sum_{i=1}^{n} y_i * (\beta_0 + x_i * \beta) \\ &= \sum_{i=1}^{n} -ln(1 - e^{\beta_0 + x_i * \beta}) + \sum_{i=1}^{n} y_i * (\beta_0 + x_i * \beta) \end{aligned}$$

donde finalmente utilizaremos la ecuación:

$$ln\frac{p(x_i)}{1 - p(x_i)} = \beta_0 + x_i * \beta$$

.

Métodos de selección automática

Existen varios métodos para construir el modelo de regresión, es decir, para seleccionar de entre todas las variables que introducimos en el modelo, cuáles son las que necesitamos para explicarlo. El modelo de regresión se puede construir utilizando las siguientes técnicas:

- Hacia adelante
  1. Se inicia con un modelo vacío (sólo el intercepto).
  2. Se ajusta un modelo y se calcula el $p - valor$ de incluir cada variable por separado.
  3. Se selecciona el modelo con la más significativa.
  4. Se ajusta un modelo con la(s) variable(s) seleccionada(s) y se calcula el $p - valor$ de añadir cada variable no seleccionada por separado.
  5. Se selecciona el modelo con la más significativa.
  6. Se repite 4 y 5 hasta que no queden variables significativas para incluir.
- Hacia atrás
  1. Se inicia con un modelo con TODAS las variables candidatas.
  2. Se eliminan, una a una, cada variable y se calcula la pérdida de ajuste al eliminar.
  3. Se selecciona para eliminar la menos significativa;
  4. Se repite 2 y 3 hasta que todas las variables incluidas sean significativas y no pueda eliminarse ninguna sin que se pierda ajuste.
- Stepwise
  - Se combinan los métodos adelante y atrás.
  - Puede empezarse por el modelo vacío o por el completo, pero en cada paso se exploran las variables incluidas, por si deben salir y las no seleccionadas, por si deben entrar.

- No todos los métodos llegan a la misma solución necesariamente.

La forma de determinar si un modelo mejora o no, en cada iteración de los métodos anteriores, es a través de contrastes de hipótesis. En los pasos en los que se prueba introducir una nueva variable, se realizan contrastes condicionales de razón de verosimilitud con el modelo del paso anterior como hipótesis nula y cada uno de los nuevos como hipótesis alternativa en cada contraste. Aquellos contrastes cuyo $p - valor$ sea inferior al nivel de significación requerido determinarán las variables susceptibles de ser introducidas en el modelo en ese paso. Entre todas, se elegirá la que más mejore el modelo, en el sentido de reducir más la varianza, con un $p - valor$ adecuado. En los pasos en los que se prueba eliminar una variable, los contrastes tienen como hipótesis nula los modelos resultantes de eliminar cada una de las variables y como hipótesis alternativa el modelo que se tenía del paso anterior.

En el presente trabajo utilizaremos como método de selección el **Stepwise**.

Ajuste del modelo de regresión logística

Con el conjunto de 55 variables resultante del proceso de selección inicial (ver: Variables Pre-Seleccionadas), ajustamos un modelo logístico, recordemos que estas variables fueron categorizadas con la ayuda del WOE (Weigth of evidence) expuesto anteriormente.

Poder de clasificación

El primer indicador del poder predictivo del modelo es la proporción compradores y no compradores clasificados correctamente.

Classification Table
Data Role=TRAIN Target Variable=target_VSMS

Objetivo	Salida	N	%
0	0	172039	71,1208
0	1	8836	3,6528
1	0	11779	4,8694
1	1	49243	20,357

% total de casos clasificados correctamente
91,48

Tabla 8: Tabla de Clasificación

Como se observa en la tabla anterior, el evento compra definido como 1 en la variable objetivo, es clasificado por el modelo como comprador en un 20.34 %, el no comprador definido por la variable objetivo es clasificado correctamente en el 71.12 % de los casos, lo que nos da un porcentaje total de 91.47 % de clasificación correcta, siendo este porcentaje alto para los modelos construidos en categorías de consumo.

Selección de variables

A continuación, se muestran las variables seleccionadas por el modelo, dada la estructura de las variables, las estimaciones del modelo logístico con método de selección **Stepwise** se hacen para cada una de las categorías de las mismas:

Nombre de la variable	Categoria	DF	Estimate	Error	Chi-Square	Pr > ChiSq
Intercepto		1	0,2797	0,0319	77,0	<.0001
Número de visitasal CAC	01:low-0.5,	1	-0,2640	0,0217	147,9	<.0001
Número de visitasal CAC	02:0.5-7.5	1	-0,0389	0,0236	2,7	0,1
R3-Porcentaje de consumo de soles incluidos en el plan. Para Planes actuales	01:low-0.845285, __	1	-0,1823	0,0190	92,4	<.0001
M3-Valor Cargo fijo mensualde paquetes	01:low-0.9833333	1	-0,1178	0,0275	18,4	<.0001
M3-Valor Cargo fijo mensualde paquetes	02:0.9833333-59.918433,	1	-0,2164	0,0178	148,0	<.0001
M2-Número deLlamadas Adicionales SalientesONNET	01:low-0.25,	1	-0,0721	0,0203	12,6	0,0
M2-Número deLlamadas Adicionales SalientesONNET	02:0.25-2.25	1	-0,0436	0,0217	4,0	0,0
Número de díascon mensaje entrante en elmes	01:low-13.5	1	-0,2248	0,0344	42,8	<.0001
Número de díascon mensaje entrante en elmes	02:13.5-25.5,	1	-0,0509	0,0213	5,7	0,0
Número de díascon mensaje entrante en elmes	03:25.5-30.5	1	0,0677	0,0229	8,7	0,0
M3-Cantidad de númerosdestino a los que llamó	01:low-60.166667	1	-0,1430	0,0291	24,1	<.0001
M3-Cantidad de númerosdestino a los que llamó	02:60.166667-132.5,	1	-0,1103	0,0208	28,3	<.0001
M3-Cantidad de númerosdestino a los que llamó	03:132.5-304.5	1	0,0687	0,0208	10,9	0,0
Cantidad de númerosorigen que llamaron a una línea	01:low-34.5	1	-0,1159	0,0259	20,0	<.0001
Cantidad de númerosorigen que llamaron a una línea	02:34.5-123.5,	1	-0,0548	0,0159	11,8	0,0
Cantidad de númerosorigen que llamaron a una línea	03:123.5-344.5	1	0,0291	0,0165	3,1	0,1
C6-Flag Cantidadde Kb adicionales (subida/bajada)de Tráfico GPRS	01:low-0.5,	1	0,0679	0,0179	14,3	0,0
C6-Flag Cantidadde Kb adicionales (subida/bajada)de Tráfico GPRS	02:0.5-1.5	1	0,0336	0,0209	2,6	0,1
C6-Flag Cantidad de Minutos Adicionales Salientes OFFNET del Plan a Movistar	01:low-0.5,	1	-0,2476	0,0182	185,7	<.0001
C6-Flag Cantidad de Minutos Adicionales Salientes OFFNET del Plan a Movistar	02:0.5-1.5	1	-0,0435	0,0185	5,5	0,0
C6-Flag Cantidad de Minutos Adicionales Salientes OFFNET del Plan a Movistar	03:1.5-3.5	1	0,0341	0,0189	3,3	0,1
RFM Recargas	01:low-2782.5,	1	-0,3503	0,0172	413,4	<.0001
RFM Recargas	02:2782.5-43328	1	0,0078	0,0182	0,2	0,7

Figura 4: Tabla de Parametros

La anterior tabla de parámetros muestra las variables que entraron en el modelo, las categorías de dichas variables así como sus ponderadores y significancias. En esta tabla es posible evaluar la lógica de las variables así como su poder predictivo. En ese orden de ideas, un signo positivo representa mayor probabilidad de compra y en caso contrario (signo negativo) representa menor probabilidad de compra. Tal como se muestra notar en dicha tabla, todas las variables son significativas al 95 % de confianza.

La lectura de la anterior tabla de parámetros es la siguiente:

- Variables: Corresponde a las variables predictoras del el modelo y el intercepto.
- Categorías: Corresponde a las categorías o agrupamientos de las variables, obtenidos anteriormente por medio del WOE.
- DF: Esta columna muestra los grados de libertad que corresponden al parámetro. Cada parámetro estimado en el modelo requiere de un DF y define la distribución de chi-cuadrado para probar si el coeficiente de regresión individual es cero.
- Estimate: Estas son las estimaciones de regresión para los parámetros del modelo.
- Error estándar: Estos son los errores estándar de las estimaciones de los coeficientes de regresión.
- Chi-Cuadrado y `Pr>chisq`: Estas son las estadísticas de prueba y los valores de p, respectivamente, probando la hipótesis nula de que el coeficiente de regresión de cada predictor individual (categoría) es cero. La prueba estadística de Chi-Cuadrado es la proporción al cuadrado de la estimación del error estándar del respectivo predictor. El valor de Chi-Cuadrado sigue una distribución central de Chi-cuadrado con grados de libertad dados por DF, que se utiliza para poner a prueba contra la hipótesis alternativa de que la estimación no es igual a cero.

Análisis de multicolinealidad

A continuación se presentan los diagnósticos de colinealidad que se realizaron sobre las variables definitivas del modelo que se presentaron en la tabla anterior. En la siguiente tabla se pueden observar la tolerancia y la inflación de la varianza. La tolerancia es el coeficiente de determinación R^2 de una variable explicativa sobre todas las demás; la inflación de la varianza es el inverso multiplicativo de la tolerancia. A fin de entender cómo detectar colinealidad (en este caso multicolinealidad) de debe tener en cuenta

que una tolerancia de menos de 0.20 y / o un VIF (inflación de la varianza) de 5 o 10 y por encima indica un problema de multicolinealidad. De acuerdo a estos dos indicadores, no se presenta ninguno de estos dos casos.

Parameter	Label	Tolerance	Variance Inflation
Intercept	Intercept	.	0.00000
CPPOS01CLI005	Número de visitas al CAC	0.97605	1.02454
CPPOS01CON008_R3	R3-Porcentaje de consumo de soles incluidos en el plan. Para Planes actuales	0.87909	1.13754
CPPOS01FAC011_M3	M3-Valor Cargo fijo mensual de paquetes	0.84278	1.18654
CPPOS01TRA005_M2	M2-Número de Llamadas Adicionales Salientes ONNET	0.67483	1.48185
CPPOS01TRA023	Número de días con mensaje entrante en el mes	0.62289	1.60541
CPPOS01TRA056_M3	M3-Cantidad de números destino a los que llamó	0.65881	1.51789
CPPOS01TRA057	Cantidad de números origen que llamaron a una línea	0.58480	1.70999
CPPOS02CON002_C6	C6-Flag Cantidad de Kb adicionales (subida/bajada) de Tráfico GPRS	0.90226	1.10832
CPPOS02TRA004_C6	C6-Flag Cantidad de Minutos Adicionales Salientes OFFNET del Plan a Movistar	0.64179	1.55813
CPPOS02TRA042	RFM Recargas	0.84263	1.18675

Tabla 9: Tolerancia e Inflación de la Varianza

Tendencia de las variables significativas

Se observa una relación directa entre las categorías de la variable y la tasa de compra, en todas las variables incluidas en el modelo.

En la variable "Numero de visitas al CAC", a medida que aumenta la cantidad de visitas, la tasa de compradores es mayor. En efecto se observa que cero visitas tiene 27.66 % de target mientras que 7 o mas visitas tiene un 48.25 % de target.

Ahora bien, la variable "Numero de días con mensaje entrante en el mesïndica que a mayor cantidad de días, es más alta la tasa de compradores. De hecho, si tiene mensajes entrantes todos el mes (30 o mas), su probabilidad de compra de paquetes de VSMS es de 49.11 %. En la variable Ç6-Flag cantidad de minutos adicionales salientes OFFNET del plan a Movistar"la mayor probabilidad de compra de paquetes Ontop-Voz y SMS se encuentra en los clientes con con 3 o mas flag en 6 meses, esto significa que los clientes que usan en 3 o mas meses, de 6 observados, minutos adicionales a Movistar, tienen la mayor probabilidad de compra.

De la misma manera se leen las tasas de compra de las demás variables incluidas en el modelo, que se pueden ver en la siguiente tabla.

	Categoría	Label categoría	Tasa de Compra	% Compradores	% Total
Número de visitasal CAC	1	0	27,66%	58,65%	66,02%
	2	Entre 1 y 6	36,06%	33,42%	28,86%
	3	7 y mayores	48,25%	7,93%	5,12%
R3-Porcentaje de consumo de soles incluidos en el plan. Para Planes actuales	1	Menor a 0.845285	19,29%	14,51%	23,43%
	2	0.845285 y mayores	34,77%	85,49%	76,57%
M3-Valor Cargo fijo mensualde paquetes	1	Menor a 0.9833333	19,25%	6,01%	9,73%
	2	Entre 0.9833333 y 59.918433	31,77%	86,55%	84,85%
	3	03:59.918433 y mayores	42,68%	7,44%	5,43%
M2-Número deLlamadas Adicionales SalientesONNET	1	0	25,03%	58,59%	72,90%
	2	Entre 0.5 y 2	38,10%	13,38%	10,93%
	3	2.5 y mayores	54,01%	28,04%	16,17%
Número de díascon mensaje entrante en elmes	1	Menor a 13	17,82%	11,87%	20,73%
	2	Entre 13 y 25	29,46%	41,93%	44,33%
	3	Entre 25 y 30	39,84%	38,22%	29,88%
	4	30 y mayores	49,11%	7,98%	5,06%
M3-Cantidad de númerosdestino a los que llamó	1	Menor a 60.166667	19,34%	15,78%	25,40%
	2	Entre 60.166667 y 132.5	29,58%	32,87%	34,61%
	3	Entre 132.5 y 304.5	38,27%	39,69%	32,30%
	4	04:304.5 y mayores	47,18%	11,66%	7,70%
Cantidad de númerosorigen que llamaron a una línea	1	Menor a 34	17,69%	11,37%	20,02%
	2	Entre 34 y 123	28,93%	37,21%	40,05%
	3	Entre 123 y 344	38,71%	42,45%	34,14%
	4	344 y mayores	48,33%	8,98%	5,78%
C6-Flag Cantidadde Kb adicionales (subida/bajada)de Tráfico GPRS	1	0	28,24%	59,41%	65,52%
	2	1	33,79%	18,68%	17,21%
	3	Mayor a	39,54%	21,92%	17,26%
C6-Flag Cantidad de Minutos Adicionales Salientes OFFNET del Plan a Movistar	1	0	22,63%	41,65%	57,33%
	2	1	32,71%	18,20%	17,32%
	3	Entre 1 y 3	43,47%	21,28%	15,24%
	4	Mayores a 3	58,13%	18,88%	10,11%
RFM Recargas	1	Menor a 2782	22,95%	36,58%	49,65%
	2	Entre 2782 y 43328	32,79%	27,83%	26,43%
	3	43328 y mayores	46,33%	35,59%	23,92%

Tabla 10: Tendencia - Variables del modelo

Etapa 3: Validación del modelo

Para analizar el funcionamiento del modelo, se aplicó a toda la base de datos de 2013-04 a 2014-03, sin oversampling. De tal manera que se pueda evaluar su efectividad en la proporción real del evento modelado: **Compra de un paquete adicional de Voz-SMS**.

Para esta validación, se utilizaron los tres indicadores que se describen a continuación, que fueron aplicados en cada una de las ventanas de tiempo.

- **KS**

  El estadístico de Kolmogorov y Smirnov (KS) es una medida de separación y consiste en medir cuán distintas son las funciones de de distribución acumulada de No compradores y Compradores

para cada decil del score; busca la mayor diferencia entre las distribuciones acumuladas, en valor absoluto:

$$KS = Max|Dist\ Acum\ No\ Compradores - Dist\ Acum\ Compradores|$$

Mientras mayor sea la máxima distancia entre estas distribuciones, mejor discrimina el modelo. En la construcción, se tiene un KS de 36.47 %, y se mantiene estable en los meses de validación, especialmente a partir de agosto de 2013, periodo en el cual se mejoró la calidad de la información en la compañía.

- **Gini**

  El estadístico de Gini indica el numero de compradores detectados sobre el total de compradores de la muestra.

$$Gini = \frac{Compradores\ detectados}{Compradores\ Totales}$$

  Este estadístico también es estable en el tiempo, entre agosto de 2013 y marzo de 2014 está alrededor del 47.66 % de se obtuvo en la construcción del modelo.

- **ROC**

  Este indicador compara la sensibilidad (clasificación correcta de compradores) frente a 1-Especificidad (Clasificación correcta de no compradores) del modelo evaluado:

$$Sensibilidad = VP/(VP + FN)$$

$$Especificidad = VN/(FP + VN)$$

  Donde:

  VP = Verdaderos positivos

  FN = Falsos Negativos

  FP = Falsos positivos

  VN = Verdaderos Negativos

  Como se puede ver en la tabla de resultados, se mantiene cercano al 73.95 % de la construcción a partir de agosto de 2013.

La siguiente tabla muestra los resultados de los indicadores para todas las ventanas de tiempo.

Base	Población	KS	GINI	ROC
Total	**6.844.569**	**36,47%**	**47,66%**	**73,95%**
abr-13	1.340.743	15,75%	22,36%	61,35%
may-13	1.386.926	16,06%	21,55%	60,97%
jun-13	1.427.220	14,20%	19,16%	59,74%
jul-13	1.485.460	19,10%	25,65%	62,99%
ago-13	1.562.470	38,20%	49,75%	75,03%
sep-13	1.605.795	37,34%	49,10%	74,69%
oct-13	1.655.214	37,72%	49,50%	74,88%
nov-13	1.703.196	37,45%	48,21%	74,24%
dic-13	1.750.527	34,20%	45,57%	72,91%
ene-14	1.805.513	35,33%	46,49%	73,38%
feb-14	1.741.506	35,22%	45,88%	73,08%
mar-14	1.680.646	33,08%	43,88%	72,01%

Tabla 11: Resumen de Validación

En general, para todos los periodos de validación se observa un buen desempeño y estabilidad en los tres estadísticos calculados.

Análisis del score del modelo

Se define el score para cada individuo como la probabilidad de compra de un paquete adicional de Voz-SMS, multiplicado por 1000. Este score se divide en deciles, con el fin de identificar los grupos con mayor probabilidad de compra, analizar las distribuciones e identificar el punto del score donde se tiene la mayor discriminación entre compradores y no compradores: KS.

Analizando las distribuciones acumuladas, observamos que:

- El noveno decil tiene el mayor KS; entendido como la diferencia entre % acumulado de compradores y el % acumulado de no compradores:

RANGO	No Compradores	% No Compradores	% No compradores Acumulado
10	669.439	9,83%	9,83%
9	674.118	9,90%	19,73%
8	673.402	9,89%	29,62%
7	694.880	10,20%	39,82%
6	680.327	9,99%	49,81%
5	682.991	10,03%	59,84%
4	685.161	10,06%	69,90%
3	680.945	10,00%	79,90%
2	662.949	9,73%	89,63%
1	706.304	10,37%	100,00%
Total	**6.810.516**	**100,00%**	.

RANGO	Compradores	% Compradores	% compradores Acumulado
10	14.717	43,22%	43,22%
9	4.420	12,98%	56,20%
8	3.210	9,43%	65,62%
7	2.799	8,22%	73,84%
6	2.212	6,50%	80,34%
5	1.869	5,49%	85,83%
4	1.605	4,71%	90,54%
3	1.338	3,93%	94,47%
2	1.134	3,33%	97,80%
1	749	2,20%	100,00%
Total	**34.053**	**100,00%**	.

KS
33,39%
36,47%
36,01%
34,03%
30,53%
25,99%
20,64%
14,58%
8,17%
0,00%
36,47%

Tabla 12: Tabla KS del modelo

Gráficamente, el KS se puede representar como la distancia máxima entre las curvas acumuladas de compradores y no compradores (linea color naranja):

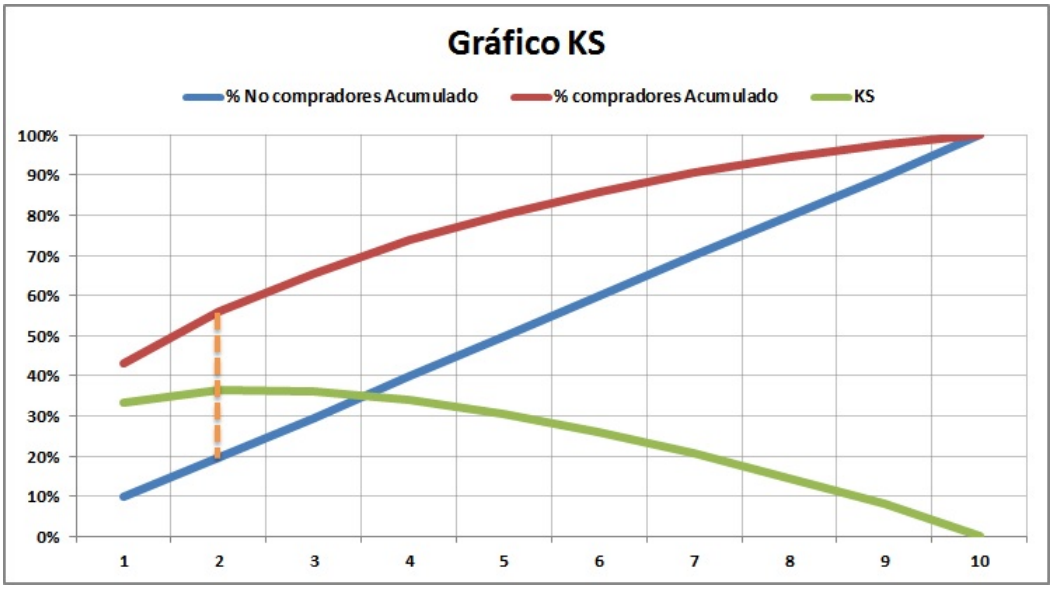


Figura 5: Gráfico KS del modelo

- Los deciles están en relación directa con la tasa de compra, lo que indica que en los deciles superiores del score están los clientes con mayor tasa de compra; esta es una propiedad deseable en la construcción de modelos de score, porque facilita su utilización y lectura.

RANGO	Total	Compradores	Tasa Compra
10	684.156	14.717	2,15%
9	678.538	4.420	0,65%
8	676.612	3.210	0,47%
7	697.679	2.799	0,40%
6	682.539	2.212	0,32%
5	684.860	1.869	0,27%
4	686.766	1.605	0,23%
3	682.283	1.338	0,20%
2	664.083	1.134	0,17%
1	707.053	749	0,11%
Total	**6.844.569**	**34.053**	**0,50%**

Tasa de compra

Tabla 13: Ordenamiento por tasa de compra

- **Lift del modelo** Es una medida de la mejora del modelo con respecto a la tasa de compra general y se calcula para cada decil del score.

RANGO	Total	Compradores	Tasa Compra	Lift
10	684.156	14.717	2,15%	4,30
9	678.538	4.420	0,65%	1,30
8	676.612	3.210	0,47%	0,94
7	697.679	2.799	0,40%	0,80
6	682.539	2.212	0,32%	0,64
5	684.860	1.869	0,27%	0,54
4	686.766	1.605	0,23%	0,46
3	682.283	1.338	0,20%	0,40
2	664.083	1.134	0,17%	0,34
1	707.053	749	0,11%	0,22
Total	**6.844.569**	**34.053**	**0,50%**	**1,00**

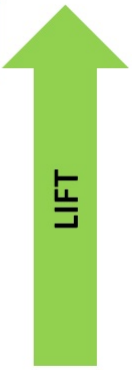


Tabla 14: Lift de Score

Por ejemplo, para el decil superior, contactar a los 684.156 clientes, captura 4.3 veces mas compradores que si lo hiciera aleatoriamente en toda la base de datos.

- Al tener los porcentajes de población total y de compradores, se pueden hacer análisis muy importantes sobre el score predictivo construido en el presente trabajo:

RANGO	Total	Total Acumulado	%Total	%Total Acum	Compradores	Compradores	% Compradores	% compradores Acumulado
10	684.156	684.156	10,00%	10,00%	14.717	14.717	43,22%	43,22%
9	678.538	1.362.694	9,91%	19,91%	4.420	19.137	12,98%	56,20%
8	676.612	2.039.306	9,89%	29,79%	3.210	22.347	9,43%	65,62%
7	697.679	2.736.985	10,19%	39,99%	2.799	25.146	8,22%	73,84%
6	682.539	3.419.524	9,97%	49,96%	2.212	27.358	6,50%	80,34%
5	684.860	4.104.384	10,01%	59,97%	1.869	29.227	5,49%	85,83%
4	686.766	4.791.150	10,03%	70,00%	1.605	30.832	4,71%	90,54%
3	682.283	5.473.433	9,97%	79,97%	1.338	32.170	3,93%	94,47%
2	664.083	6.137.516	9,70%	89,67%	1.134	33.304	3,33%	97,80%
1	707.053	6.844.569	10,33%	100,00%	749	34.053	2,20%	100,00%
Total	**6.844.569**	.	**100,00%**	.	**34.053**	.	**100,00%**	.

Tabla 15: Poblaciones y Tasas de compra por deciles

En el noveno decil del score, la población acumulada es de 19.91 %:

- En el caso extremo, si contactamos a todas la población (6'844.569 usuarios), esperamos una efectividad de compra del producto del 0.05 % (34.053 compradores).
- Si contactamos a todas las personas del decil superior, que corresponde al 10 % (684.156 usuarios), esperamos atrapar el 43.22 % de los compradores (14.717)
- Si contactamos a las personas de los dos deciles superiores, estamos alcanzando al 19.91 % de la población (1'362.694 usuarios)y esperamos captar el 56.20 % de los compradores (19.137). En este caso, el incremento en la tasa de compra es importante, confirmando el punto de corte sugerido por el cálculo del KS.
- Si consideramos contactar a las personas que se encuentran en los 3 deciles superiores, es decir el 29.79 % de la población (2'039.306 usuarios), se esperaría captar el 65.62 % de los compradores (22.347).

Para decidir si se contactan a las personas de los dos mayores deciles como lo sugiere el KS, o de los tres mayores como se analizó, es necesario tener en cuenta variables propias del negocio, como presupuesto, capacidad operativa, etc.

El score es una potente herramienta que permite enfocar y optimizar el contacto de clientes, siempre enfocado a mejorar el ROI de las campañas comerciales.

Etapa 4: METODOLOGÍA PROPUESTA: Up-Lifting

Hasta este punto, hemos ajustado un modelo de Scoring de riesgo, que ha sido ampliamente analizado en la literatura y se usa en muchos casos de negocio. Hsta el momento el score construido nos permite tener el siguiente lift de 131.25 en los tres deciles superiores:

RANGO	Total	Compradores	Tasa Compra	Lift
10	2.039.306	22.347	65,62%	**131,25**
9				
8				
7	4.805.263	11.706	34,38%	68,75
6				
5				
4				
3				
2				
1				
Total	**6.844.569**	**34.053**	**100,00%**	**1,00**

Tabla 16: Lift en lor tres deciles superiores

Entendiendo el contexto de nuestro problema, la compañía va a llamar o a envíar mensajes de texto a los 2'039.306 clientes de los tres deciles superiores, para ofrecerles un paquete adicional de voz-SMS, sin embargo, si el cliente necesita minutos o mensajes adicionales, puede estar contemplando la adquisición del paquete antes de recibir la campaña, con lo que la campaña seria innecesaria, incluso puede surtir el efecto contrario: que el cliente se moleste y desista de su compra. Paralelo a esto, hay clientes que no pensaban comprar el paquete adicional, pero al ser contactados con el mensaje de texto o la llamada, decidirán comprarlo, son estos los clientes a los que se deben dirigir las campañas de mercadeo.

Ahora bien, el objetivo es refinar la decisión que se tomó con el modelo, utilizando el efecto de una campaña anterior, que tenía un mensaje similar al que se usara ahora. para esta metodología se hace una simulación con interacción de la variable campaña, que toma valor de 1 si recibió la campaña y 0 si no la recibió.

Objetivo

Partiendo de la hipótesis de que la respuesta a la campaña de marketing hace la diferencia en el evento compra o no compra del paquete adicional, se propone construir un modelo posterior, incluyendo la variable de campaña $T_i \in \{1, 0\}$ donde (1) indica que recibió la campaña, y (0) que no la recibió.

El principal objetivo de la metodología es la maximización de la diferencia entre el valor esperado de la probabilidad de compra dado que recibió campaña $T_i = 1$ y el valor esperado dado que no la recibió $T_i = 0$, es decir,

$$Maximizar \sum_{n=1} \{E(y_i|x_i; Recibio\ Campana) - E(y_i|x_i; No\ Recibio\ Campana)\}$$

Donde y_i es la variable objetivo del modelo, $y_i = 1$ indica la compra del paquete, y $y_i = 0$ indica la no compra del paquete.

Esta maximización se hace con el fin de identificar a los clientes (o potenciales clientes) cuyas respuestas serán influenciadas de manera positiva por la campaña.

Metodología Up-Lifting

Para lograr la maximización, es necesario simular dos probabilidades: $E(y_i|x_i; Recibio\ Campana)$ y $E(y_i|x_i; No\ Recibio\ Campana)$, para todos los individuos de la población.

Si Y_i es una variable binaria que indica si el cliente i responde a una campaña, tenemos en cuenta el siguiente conjunto de variables: X_i la matriz de variables predictoras, T_i, donde $T_i = 1$ si i recibió la campaña, y $T_i = 0$ si i no recibió la campaña, y $X_i * T_i$, la interacción entre cada una de las variables predictoras y la variable de campaña. Podemos modelar la tasa de respuesta mediante una regresión logística de la siguiente forma:

$$P_i = E(y_i|x_i) = \frac{1}{1 + exp^{-(\alpha + \beta' X_i + \gamma' T_i + \delta' X_i T_i)}}$$

donde $\alpha, \beta, \gamma, \delta$ son parámetros a estimar.

Nótese que α es el intercepto, β es un vector de parámetros de los principales efectos de las variables independientes, γ denota el efecto principal del tratamiento, y δ los efectos adicionales de las variables predictoras en interacción con la campaña.

Por lo general se aplicará algún procedimiento de reducción de variables, que este caso se realizó en la construcción del score. A partir de las estimaciones de los parámetros obtenidas con la regresión logística se calcula:

$$P_i(x_i|Tratamiento) - P_i(x_i|Control) = \frac{1}{1 + exp^{-(\alpha + \beta' X_i + \gamma' T_i + \delta' X_i T_i)}} - \frac{1}{1 + exp^{-(\alpha + \beta' X_i)}}$$

Esta es la probabilidad para cada individuo simulando que tuvo campaña y que no tuvo campaña. Aquellos clientes para los que la diferencia sea positiva serán seleccionados para la próxima campaña de mercadeo, porque la probabilidad de compra dado que recibió la campaña es mayor que la probabilidad de compra dado que no recibió la campaña.

Integración del up-lifting con el modelo de scoring

A partir del up-lifting descrito anteriormente, y del modelo ya ajustado, se construyó la matriz de segmentación de clientes que nos permite visualizar los siguientes grupos de clientes:

- **A:** Clientes que tienen una alta probabilidad de compra, presentan una respuesta positiva a la campaña, a estos clientes se les realizará campaña.
- **B:** Clientes que, aunque tienen una alta probabilidad de compra, presentan una respuesta negativa a la campaña, no se les realizará campaña pues su intención de compra puede deteriorarse.
- **C:** Clientes que tienen una alta probabilidad de compra, pero no presentan una respuesta definida positiva o negativa a la campaña, es decir, la diferencia $(T = 1) - (T = 0) \cong 0$, se les realizará campaña solo si los recursos los permiten.

- **D:** Clientes que tienen una baja probabilidad de compra, no es recomendable hacerles campaña.

La lectura de la matriz puede verse de la siguiente forma:

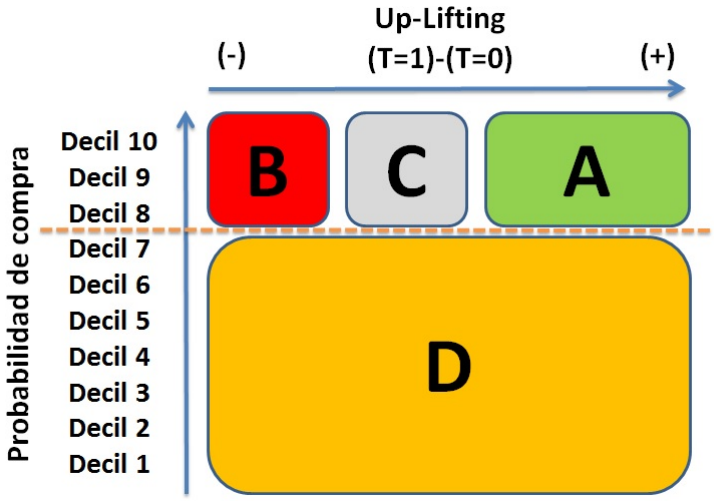


Figura 6: Matriz Uplifting

En la figura anterior, se muestra la forma general de la matriz de UpLifting. Sobre el eje Y tenemos los deciles del score de construido anteriormente, por tanto hacia la parte superior de la matriz se encontrarán los clientes con mayor probabilidad de compra. En el eje X, se encuentra el valor de UpLifting, los valores a la izquierda representan un efecto de la campaña cero o incluso negativo. Hacia el lado derecho se encuentran los clientes susceptibles a la influencia de la campaña. El valor del Up-lifting se dividió en cuartiles, y se tomará como segmento A a los dos cuartiles mas altos (derecha), como segmento B al cuartil mas bajo (izquierda), y como segmento C al que esta en medio de los dos anteriores.

Como una mejor forma de analizar la matriz, se presentan las siguientes proporciones de la tabla, en donde se identifican los porcentajes de población en cada uno de los cuadrantes sugeridos para la estrategia.

Total Población

Score	Up-Lifting 1	2	3	4
10				
9	1.026.216	423.506	576.830	
8				
7				
6				
5				
4	4.818.016			
3				
2				
1				

% de Población

Score	Up-Lifting 1	2	3	4
10				
9	15,0%	6,2%	8,4%	
8				
7				
6				
5				
4	70,4%			
3				
2				
1				

Tabla 17: Proporciones de población - Matriz Uplifting

Como se puede ver en la tabla, el modelo de up-lifting sugiere un segmento mucho mas pequeño que el que inicialmente se plantea con el score, 8.4 % vs 29.6 %, esto es importante porque reduce la cantidad de usuarios a contactar, por ende reduce los costos de las campañas realizadas.

En la siguiente tabla, para cada segmento, se visualiza la tasa de compra del score calculado, que para los tres deciles superiores es del 64.7 %. Esta medida será la base de partida para las estrategias dirigidas a dichos segmentos.

Total Compradores

	Up-Lifting			
Score	1	2	3	4
10				
9	12.189	5.700	4.157	
8				
7				
6				
5				
4	12.006			
3				
2				
1				

% de Compradores

	Up-Lifting			
Score	1	2	3	4
10				
9	35,8%	16,7%	12,2%	
8				
7				
6				
5				
4	35,3%			
3				
2				
1				

Tabla 18: Proporciones de Compradores - Matriz Uplifting

El hallazgo mas relevante de este proceso es que con el score construido teníamos que contactar al 29.6 % de la población para encontrar al 64.7 % de los compradores, pero como el Up-lifting nos muestra que al segmento B no se le debe contactar porque comprarán sin hacerles campaña, y al segmento C tampoco se le debe contactar porque la campaña no tendrá ningún efecto, solo tendríamos que hacer campaña al 8.4 % del segmento A para atrapar al mismo 64.7 % de los compradores que propuso el score.

Conclusiones

Se ha presentado un nuevo enfoque para el modelado de la respuesta en minería de datos. Esto supone una mejoría a las metodologías actuales, ya que aborda directamente el objetivo de maximizar el Up-Lifting. En concreto, la metodología propuesta, sin embargo, identifica los clientes (o potenciales clientes) de tal manera que la diferencia incremental entre las respuestas de tratamiento y de control se maximiza. Esto es particularmente importante para las campañas de desarrollo de clientes (upselling y cross-selling).

La metodología propuesta es simple y se puede aplicar fácilmente a técnicas de modelado no lineal para aprendizaje supervisado como la regresión logística, los árboles de decisión y las redes neuronales, que son las técnicas más comúnmente utilizadas.

Por otra parte, con respecto a la necesidad planteada en principio, resuelta por el modelo, podemos decir que:

- El modelo de probabilidad de compra de paquetes ONTOP VOZ - SMS para el segmento Postpago presenta estadísticos de discriminación altos (ejemplo: ROC de 73.95 %), que muestran que su uso podrá ser un diferenciador importante para llegar a tener estrategias eficaces de upsell de clientes.
- El ordenamiento en términos de tasas de compra es bastante satisfactorio, y podrá aportar al diseño de las estrategias de negocio necesarias.

- Se presenta estabilidad en el modelo, vista a través del tiempo, especialmente en los periodos más recientes, lo cual fue comprobado en los periodos de desarrollo y de validación del score.

- Se recomienda hacer un backtesting en un periodo de seis meses o un año a fin de garantizar que la estabilidad y el poder de discriminación se mantiene. Esto se debe realizar debido a que el objetivo del modelo es generar cambios en el comportamiento de los clientes con respecto a la compra de paquetes adicionales de VOZ-SMS, lo cual cambiará el mercado y las condiciones del mismo.

Trabajo futuro

Esperamos que con este documento se abra una nueva línea de investigación y por lo tanto se beneficien las técnicas de minería de datos en los diferentes sectores de la industria.

Hemos considerado las siguientes áreas de profundización, para ser trabajadas en el futuro o en otros proyectos donde se aplique la metodología propuesta:

- Mas de un tipo de campaña: Cuando se presentan múltiples ofertas, diferentes canales, diferentes mensajes, etc., la metodología propuesta se puede aplicar fácilmente. Sin embargo, se generarán más variables de interacción con el modelo, esto complicará la simulación de las probabilidades. Por esto será necesario diseñas otra manera de integrar las variables simuladas y el calculo de la diferencia a maximizar.

- Otras formas de estimación: Se pueden aplicar otros procedimientos similares, por ejemplo, el ajuste de modelos con campaña y sin campaña de manera separada, pueden ser estudiados de manera rigurosa, incluso, con simulaciones empíricas.

- Otras formas de validación: Se pueden utilizar diferentes configuraciones para la validación, por ejemplo, realizar simulaciones para examinar la sensibilidad del procedimiento ante diversos factores tales como la categorización y la variabilidad de los datos en el tiempo.

Agradecimientos

Quiero agradecer a mi tutora de tesis, doctora Luz Mary Pinzón, por compartir conmigo su amplio conocimiento y experiencia, por la oportunidad de trabajar a su lado, por la paciencia que tuvo y por su dedicación, ya que sin ella este proyecto no hubiera sido posible.

Agradezco a mi familia, en especial a mi madre, por el apoyo y admiración, por sus hermosas palabras en los momentos buenos y malos de mi formación profesional.

A todos los colegas y compañeros, de quienes he aprendido mucho sobre esta hermosa profesión.

A todos, muchas gracias.

Referencias

[1] Lo, V. S. Y., *The true lift model - a novel data mining approach to response modeling in database marketing*, SIGKDD Explorations, 4(2):78-86, (2002)

[2] Scott, Alastair , *Fitting Logistic Regression Models in Case-Control Studies with Complex Sampling*, R. L. Chambers and C. J. Skinner, (2003)

[3] Teresa Costa Cor, *BONDAD DE AJUSTE Y ELECCIÓN DEL PUNTO DE CORTE EN REGRESIÓN LOGÍSTICA BASADA EN DISTANCIAS. APLICACIÓN AL PROBLEMA DE CREDIT SCORING.*,Anales del Instituto de Actuarios Españoles, 3ª época, 18, 2012/19-40, (2012)

[4] Naeem Siddiqi, *Credit Risk scorecards. Developing and implementing intelligent credit scoring*, John Wiley & Sons Inc., (2006)

[5] Raymond Anderson, *The Credit Scoring Toolkit - Theory and Practice for Retail Credit Risk Management and Decision Automation*, OXFORD, (2007)

[6] Goran Kraljevi, *Modeling Data Mining Applications for Prediction of Prepaid Churn in Telecommunication Services*, ATKAFF 51(3), 275?283, (2010)

[7] Philippe Jorion, *Financial Risk Manager Handbook*, John Wiley & Sons, (2003)

[8] Kotler P., *Marketing Management, 8th edition.*, Prentice-Hall, chapter 24, (1994)

[9] Peppers, D. and Rogers, M. *The One-to-One Fieldbook.*, Doubleday, (1999).

[10] Fabris, P. *Advanced navigation: Marketing secrets from the financial sector show how data mining charts a profitable course to customer management.*, CIO magazine, 11, No.15, p.50-55., (1998).

[11] Almquist, E. and Wyner G. *Boost your marketing ROI with experimental design.*, Harvard Business Review, Oct, p.135-141, (2001).

[12] Blattberg, R.C., Getz, G., and Thomas J.S. *Customer Equity: Building and Managing Relationships As Valuable Assets.*, Harvard Business School Press, (2001).

[13] Schreiner, Mark *Ventajas y Desventajas del Scoring Estadístico para las Microfinanzas.*, Center for Social Development Washington University in St. Louis., (2002).

[14] Berson, A., Smith, S., and Thearling, K *Building data mining applications for CRM.*, New York, NY: McGraw-Hill, (2000).

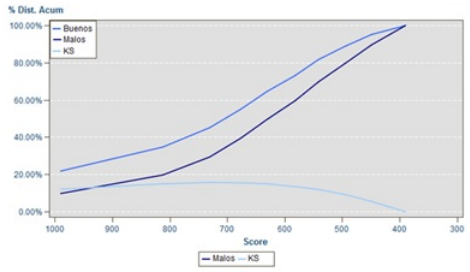
Anexo 1 - Tablas de validación

A continuación se presentan las tablas de validación de todos los meses de la base de datos, desde 2013-04 hasta 2014-04. La lectura de estas tablas se hace tal como se describió en la etapa 3: **Validación del modelo**.

Tabla de Validación:
Base=sasdlSAS_POSTPAGO_ABT_VSMS_201304 / Exclusión=Exclusion_VSMS / BGI=TARGET_Ontop_VSMS / Score=score1000_VSMS

RANGO	Min Score	Max Score	Dif	Total	%Total	%Total Acum	Malos	Tasa Malos	%Malos Acum	Buenos	Tasa Buenos	%Buenos Acum	KS	Tasa Buenos Acum	Odds Acum	Odds
10	813	990	178	133,997	9.99%	9.99%	129,614	96.73%	9.81%	4,383	3.27%	21.88%	12.06%	3.27%	29.57	29.57
9	731	813	82	133,903	9.99%	19.98%	131,321	98.07%	19.76%	2,582	1.93%	34.77%	15.01%	2.60%	37.46	50.86
8	677	731	54	130,029	9.70%	29.68%	127,940	98.39%	29.44%	2,089	1.61%	45.20%	15.75%	2.28%	42.95	61.25
7	630	677	47	134,229	10.01%	39.69%	132,261	98.53%	39.46%	1,968	1.47%	55.02%	15.56%	2.07%	47.28	67.21
6	583	630	48	138,105	10.30%	49.99%	136,149	98.58%	49.77%	1,956	1.42%	64.78%	15.02%	1.94%	50.65	69.61
5	542	583	41	131,386	9.80%	59.79%	129,715	98.73%	59.59%	1,671	1.27%	73.12%	13.54%	1.83%	53.72	77.63
4	495	542	46	136,261	10.16%	69.95%	134,533	98.73%	69.78%	1,728	1.27%	81.75%	11.97%	1.75%	56.27	77.86
3	449	495	46	132,427	9.88%	79.83%	130,994	98.92%	79.69%	1,433	1.08%	88.90%	9.21%	1.66%	59.10	91.41
2	391	449	58	132,716	9.90%	89.73%	131,451	99.05%	89.65%	1,265	0.95%	95.22%	5.57%	1.59%	62.07	103.91
1	344	390	47	137,690	10.27%	100.00%	136,732	99.30%	100.00%	958	0.70%	100.00%	0.00%	1.49%	65.93	142.73
Total	-	-	-	1,340,743	100.00%	-	1,320,710	98.51%	-	20,033	1.49%	-	15.75%	-	-	65.93

KS	GINI	AR	ROC
15.75%	22.36%	22.70%	61.35%



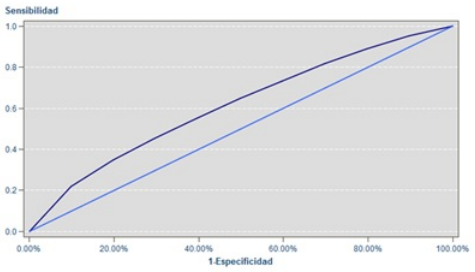
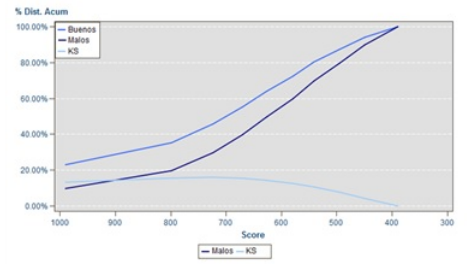


Tabla 19: Tabla de Validacion - Abril 2013 (Validación)

Tabla de Validación:
Base=sasdlSAS_POSTPAGO_ABT_VSMS_201305 / Exclusión=Exclusion_VSMS / BGI=TARGET_Ontop_VSMS / Score=score1000_VSMS

RANGO	Min Score	Max Score	Dif	Total	%Total	%Total Acum	Malos	Tasa Malos	%Malos Acum	Buenos	Tasa Buenos	%Buenos Acum	KS	Tasa Buenos Acum	Odds Acum	Odds
10	799	990	191	138,526	9.99%	9.99%	132,880	95.92%	9.75%	5,646	4.08%	23.00%	13.25%	4.08%	23.54	23.54
9	724	799	75	138,174	9.96%	19.95%	135,165	97.82%	19.67%	3,009	2.18%	35.26%	15.59%	3.13%	30.97	44.92
8	671	724	53	139,318	10.05%	30.00%	136,739	98.15%	29.71%	2,579	1.85%	45.77%	16.06%	2.70%	36.03	53.02
7	627	671	44	138,742	10.00%	40.00%	136,424	98.33%	39.73%	2,318	1.67%	55.21%	15.49%	2.44%	39.94	58.85
6	579	627	48	137,063	9.88%	49.88%	134,920	98.44%	49.63%	2,143	1.56%	63.94%	14.31%	2.27%	43.08	62.96
5	541	579	38	140,193	10.11%	59.99%	138,131	98.53%	59.77%	2,062	1.47%	72.34%	12.57%	2.13%	45.86	66.99
4	495	541	46	138,799	10.01%	70.00%	136,819	98.57%	69.81%	1,980	1.43%	80.41%	10.60%	2.03%	48.19	69.10
3	449	495	45	135,818	9.79%	79.79%	134,110	98.74%	79.65%	1,708	1.26%	87.37%	7.71%	1.94%	50.60	78.52
2	390	449	59	141,014	10.17%	89.96%	139,365	98.83%	89.88%	1,649	1.17%	94.08%	4.20%	1.85%	53.02	84.51
1	344	390	46	139,279	10.04%	100.00%	137,827	98.96%	100.00%	1,452	1.04%	100.00%	0.00%	1.77%	55.50	94.92
Total	-	-	-	1,386,926	100.00%	-	1,362,380	98.23%	-	24,546	1.77%	-	16.06%	-	-	55.50

KS	GINI	AR	ROC
16.06%	21.55%	21.94%	60.97%



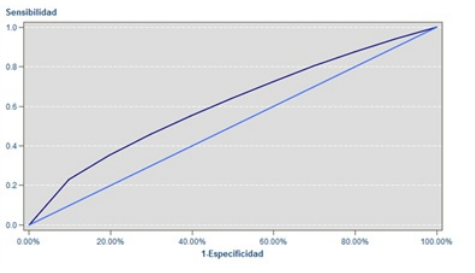
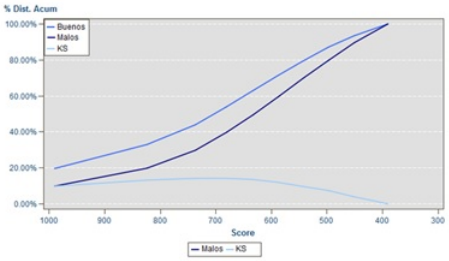


Tabla 20: Tabla de Validacion - Mayo 2013 (Validación)

Tabla de Validación:
Base=sasdl.SAS_POSTPAGO_ABT_VSMS_201306 / Exclusión=Exclusion_VSMS / BGI=TARGET_Ontop_VSMS / Score=score1000_VSMS

RANGO	Min Score	Max Score	Dif	Total	%Total	%Total Acum	Malos	Tasa Malos	%Malos Acum	Buenos	Tasa Buenos	%Buenos Acum	KS	Tasa Buenos Acum	Odds Acum	Odds
10	824	990	166	142,719	10.00%	10.00%	138,066	96.74%	9.84%	4,653	3.26%	19.63%	9.80%	3.26%	29.67	29.67
9	737	824	87	142,717	10.00%	20.00%	139,560	97.79%	19.78%	3,157	2.21%	32.95%	13.17%	2.74%	35.55	44.21
8	681	737	56	142,729	10.00%	30.00%	140,125	98.18%	29.76%	2,604	1.82%	43.94%	14.18%	2.43%	40.11	53.81
7	633	681	48	142,314	9.97%	39.97%	139,946	98.34%	39.74%	2,368	1.66%	53.93%	14.20%	2.24%	43.63	59.10
6	588	633	45	138,962	9.74%	49.71%	136,814	98.45%	49.48%	2,148	1.55%	63.00%	13.51%	2.10%	46.52	63.69
5	543	588	45	142,627	9.99%	59.70%	140,630	98.60%	59.50%	1,997	1.40%	71.42%	11.92%	1.99%	49.34	70.42
4	497	543	46	146,496	10.26%	69.97%	144,596	98.70%	69.81%	1,900	1.30%	79.44%	9.63%	1.89%	52.04	76.10
3	451	497	46	142,248	9.97%	79.93%	140,412	98.71%	79.81%	1,836	1.29%	87.19%	7.38%	1.81%	54.21	76.48
2	391	451	59	139,145	9.75%	89.68%	137,649	98.92%	89.62%	1,496	1.08%	93.50%	3.88%	1.73%	56.76	92.01
1	344	390	47	147,263	10.32%	100.00%	145,722	98.95%	100.00%	1,541	1.05%	100.00%	0.00%	1.66%	59.22	94.56
Total	.	.	.	1,427,220	100.00%	.	1,403,520	98.34%	.	23,700	1.66%	.	14.20%	.	.	59.22

KS	GINI	AR	ROC
14.20%	19.16%	19.49%	59.74%



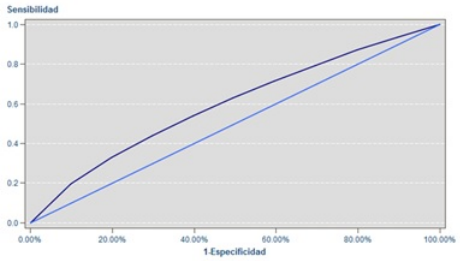
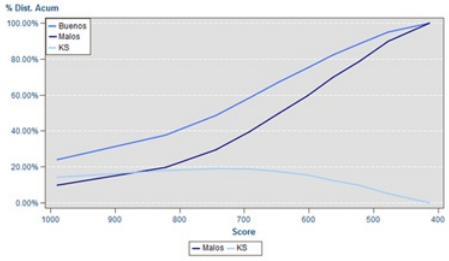


Tabla 21: Tabla de Validacion - Junio 2013 (Validación)

Tabla de Validación:
Base=sasdl.SAS_POSTPAGO_ABT_VSMS_201307 / Exclusión=Exclusion_VSMS / BGI=TARGET_Ontop_VSMS / Score=score1000_VSMS

RANGO	Min Score	Max Score	Dif	Total	%Total	%Total Acum	Malos	Tasa Malos	%Malos Acum	Buenos	Tasa Buenos	%Buenos Acum	KS	Tasa Buenos Acum	Odds Acum	Odds
10	822	990	168	148,525	10.00%	10.00%	143,877	96.87%	9.81%	4,648	3.13%	24.03%	14.21%	3.13%	30.95	30.96
9	744	822	78	147,542	9.93%	19.93%	144,899	98.21%	19.70%	2,643	1.79%	37.69%	17.99%	2.46%	39.61	54.82
8	692	744	52	146,154	9.84%	29.77%	144,040	98.55%	29.52%	2,114	1.45%	48.62%	19.10%	2.13%	46.02	68.14
7	647	692	45	149,694	10.08%	39.85%	147,793	98.73%	39.60%	1,901	1.27%	58.45%	18.85%	1.91%	51.35	77.75
6	602	647	45	150,807	10.15%	50.00%	149,117	98.88%	49.77%	1,690	1.12%	67.18%	17.41%	1.75%	56.15	88.24
5	563	602	39	145,771	9.81%	59.81%	144,238	98.95%	59.61%	1,533	1.05%	75.11%	15.50%	1.64%	60.15	94.09
4	522	563	41	151,310	10.19%	70.00%	149,924	99.08%	69.84%	1,386	0.92%	82.27%	12.44%	1.53%	64.33	108.17
3	477	522	45	130,257	8.77%	78.77%	129,062	99.08%	78.64%	1,195	0.92%	88.45%	9.81%	1.46%	67.38	108.00
2	414	477	63	166,506	11.21%	89.98%	165,232	99.23%	89.91%	1,274	0.77%	95.04%	5.13%	1.38%	71.70	129.70
1	344	414	70	148,894	10.02%	100.00%	147,934	99.36%	100.00%	960	0.64%	100.00%	0.00%	1.30%	75.79	134.10
Total	.	.	.	1,485,460	100.00%	.	1,466,116	98.70%	.	19,344	1.30%	.	19.10%	.	.	75.79

KS	GINI	AR	ROC
19.10%	25.65%	25.99%	62.99%



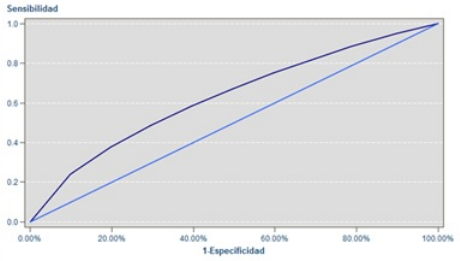


Tabla 22: Tabla de Validacion - Julio 2013 (Validación)

Tabla de Validación:
Base=sasdl.SAS_POSTPAGO_ABT_VSMS_201308 / Exclusión=Exclusion_VSMS / BGI=TARGET_Ontop_VSMS / Score=score1000_VSMS

RANGO	Min Score	Max Score	Dif	Total	%Total	%Total Acum	Malos	Tasa Malos	%Malos Acum	Buenos	Tasa Buenos	%Buenos Acum	KS	Tasa Buenos Acum	Odds Acum	Odds
10	825	990	165	156,244	10.00%	10.00%	152,242	97.44%	9.80%	4,002	2.56%	42.81%	33.00%	2.56%	38.04	38.04
9	747	825	78	156,065	9.99%	19.99%	154,703	99.13%	19.76%	1,362	0.87%	57.38%	37.61%	1.72%	57.22	113.59
8	689	747	58	156,331	10.01%	29.99%	155,341	99.37%	29.76%	990	0.63%	67.96%	38.20%	1.36%	72.76	156.91
7	641	689	49	156,313	10.00%	40.00%	155,598	99.54%	39.78%	715	0.46%	75.61%	35.83%	1.13%	87.41	217.62
6	596	641	45	149,597	9.57%	49.57%	149,008	99.61%	49.38%	589	0.39%	81.91%	32.54%	0.99%	100.14	252.98
5	550	596	46	159,218	10.19%	59.76%	158,717	99.69%	59.60%	501	0.31%	87.27%	27.67%	0.87%	113.45	316.80
4	506	550	45	157,856	10.10%	69.87%	157,450	99.74%	69.73%	406	0.26%	91.61%	21.88%	0.78%	126.45	387.81
3	464	506	41	157,187	10.06%	79.93%	156,837	99.78%	79.83%	350	0.22%	95.36%	15.53%	0.71%	139.08	448.11
2	394	464	70	157,135	10.06%	89.98%	156,853	99.82%	89.93%	282	0.18%	98.37%	8.44%	0.65%	151.87	556.22
1	344	394	50	156,524	10.02%	100.00%	156,372	99.90%	100.00%	152	0.10%	100.00%	0.00%	0.60%	166.13	1,028.76
Total	.	.	.	1,562,470	100.00%	.	1,553,121	99.40%	.	9,349	0.60%	.	38.20%	.	.	166.13

KS	GINI	AR	ROC
38.20%	49.75%	50.05%	75.03%

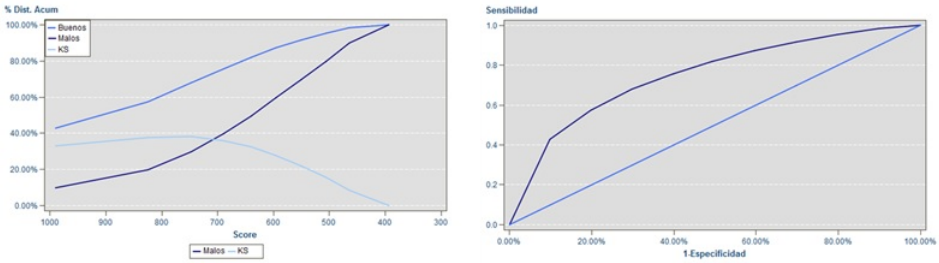


Tabla 23: Tabla de Validacion - Agosto 2013 (Validación)

Tabla de Validación:
Base=sasdl.SAS_POSTPAGO_ABT_VSMS_201309 / Exclusión=Exclusion_VSMS / BGI=TARGET_Ontop_VSMS / Score=score1000_VSMS

RANGO	Min Score	Max Score	Dif	Total	%Total	%Total Acum	Malos	Tasa Malos	%Malos Acum	Buenos	Tasa Buenos	%Buenos Acum	KS	Tasa Buenos Acum	Odds Acum	Odds
10	824	990	166	159,952	9.96%	9.96%	156,192	97.65%	9.78%	3,760	2.35%	42.55%	32.77%	2.35%	41.54	41.54
9	747	824	77	161,190	10.04%	20.00%	159,905	99.20%	19.79%	1,285	0.80%	57.09%	37.30%	1.57%	62.66	124.44
8	690	747	57	160,516	10.00%	29.99%	159,629	99.45%	29.79%	887	0.55%	67.13%	37.34%	1.23%	80.20	179.97
7	641	690	49	160,254	9.98%	39.97%	159,536	99.55%	39.78%	718	0.45%	75.25%	35.47%	1.04%	95.53	222.19
6	596	641	45	160,976	10.02%	50.00%	160,402	99.64%	49.82%	574	0.36%	81.75%	31.92%	0.90%	110.14	279.45
5	552	596	44	160,589	10.00%	60.00%	160,134	99.72%	59.85%	455	0.28%	86.90%	27.04%	0.80%	124.47	351.94
4	511	552	41	152,338	9.49%	69.49%	151,955	99.75%	69.37%	383	0.25%	91.23%	21.86%	0.72%	137.40	396.75
3	471	511	40	168,541	10.50%	79.98%	168,204	99.80%	79.90%	337	0.20%	95.04%	15.14%	0.65%	151.92	499.12
2	408	471	63	158,343	9.86%	89.84%	158,065	99.82%	89.80%	278	0.18%	98.19%	8.39%	0.60%	165.27	568.58
1	344	408	64	163,096	10.16%	100.00%	162,936	99.90%	100.00%	160	0.10%	100.00%	0.00%	0.55%	180.71	1,018.35
Total	.	.	.	1,605,795	100.00%	.	1,596,958	99.45%	.	8,837	0.55%	.	37.34%	.	.	180.71

KS	GINI	AR	ROC
37.34%	49.10%	49.37%	74.69%

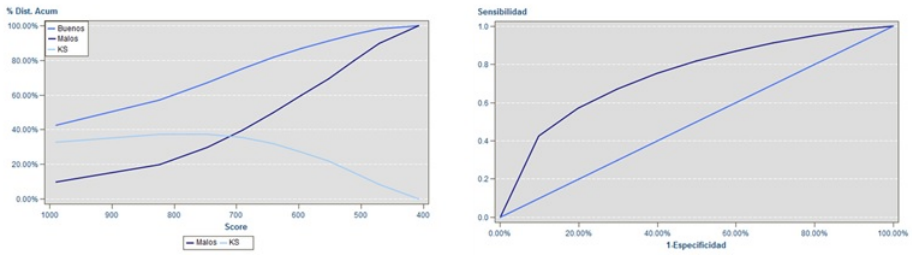
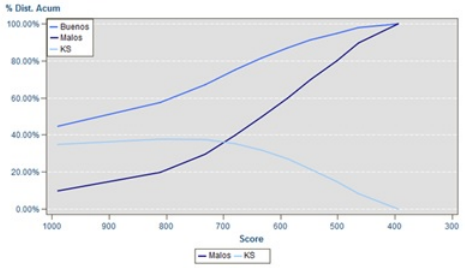


Tabla 24: Tabla de Validacion - Septiembre 2013 (Validación)

Tabla de Validación:
Base=sasdl.SAS_POSTPAGO_ABT_VSMS_201310 / Exclusión=Exclusion_VSMS / BGI=TARGET_Ontop_VSMS / Score=score1000_VSMS

RANGO	Min Score	Max Score	Dif	Total	%Total	%Total Acum	Malos	Tasa Malos	%Malos Acum	Buenos	Tasa Buenos	%Buenos Acum	KS	Tasa Buenos Acum	Odds Acum	Odds
10	811	990	179	165,519	10.00%	10.00%	161,682	97.68%	9.82%	3,837	2.32%	44.71%	34.89%	2.32%	42.14	42.14
9	732	811	79	165,522	10.00%	20.00%	164,422	99.34%	19.80%	1,100	0.66%	57.53%	37.72%	1.49%	66.05	149.48
8	680	732	52	162,602	9.82%	29.82%	161,776	99.49%	29.63%	826	0.51%	67.15%	37.52%	1.17%	84.66	195.86
7	633	680	47	168,041	10.15%	39.98%	167,357	99.59%	39.79%	684	0.41%	75.12%	35.33%	0.97%	101.63	244.67
6	588	633	45	165,117	9.98%	49.95%	164,563	99.66%	49.79%	554	0.34%	81.58%	31.79%	0.85%	117.10	297.05
5	548	588	40	166,070	10.03%	59.98%	165,605	99.72%	59.84%	465	0.28%	87.00%	27.15%	0.75%	131.99	356.14
4	502	548	46	165,328	9.99%	69.97%	164,955	99.77%	69.86%	373	0.23%	91.34%	21.48%	0.68%	146.75	442.24
3	464	502	38	165,847	10.02%	79.99%	165,551	99.82%	79.92%	296	0.18%	94.79%	14.88%	0.61%	161.76	559.29
2	394	464	69	160,993	9.73%	89.72%	160,720	99.83%	89.68%	273	0.17%	97.97%	8.30%	0.57%	175.62	588.72
1	344	394	51	170,175	10.28%	100.00%	170,001	99.90%	100.00%	174	0.10%	100.00%	0.00%	0.52%	191.87	977.02
Total	.	.	.	1,655,214	100.00%	.	1,646,632	99.48%	.	8,582	0.52%	.	37.72%	.	.	191.87

KS	GINI	AR	ROC
37.72%	49.50%	49.76%	74.88%



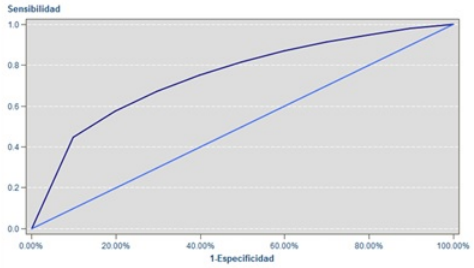
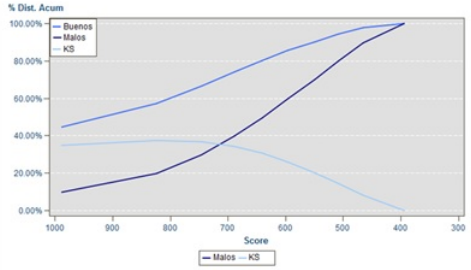


Tabla 25: Tabla de Validacion - Octubre 2013 (Construcción)

Tabla de Validación:
Base=sasdl.SAS_POSTPAGO_ABT_VSMS_201311 / Exclusión=Exclusion_VSMS / BGI=TARGET_Ontop_VSMS / Score=score1000_VSMS

RANGO	Min Score	Max Score	Dif	Total	%Total	%Total Acum	Malos	Tasa Malos	%Malos Acum	Buenos	Tasa Buenos	%Buenos Acum	KS	Tasa Buenos Acum	Odds Acum	Odds
10	824	988	164	170,287	10.00%	10.00%	166,012	97.49%	9.80%	4,275	2.51%	44.65%	34.85%	2.51%	38.83	38.83
9	746	824	78	170,217	9.99%	19.99%	169,013	99.29%	19.78%	1,204	0.71%	57.23%	37.45%	1.61%	61.15	140.38
8	689	746	57	169,174	9.93%	29.92%	168,286	99.48%	29.72%	888	0.52%	66.50%	36.79%	1.25%	79.05	189.51
7	641	689	49	171,122	10.05%	39.97%	170,398	99.58%	39.78%	724	0.42%	74.07%	34.29%	1.04%	95.01	235.36
6	596	641	45	165,058	9.69%	49.66%	164,470	99.64%	49.49%	588	0.36%	80.21%	30.72%	0.91%	109.15	279.71
5	550	596	46	175,334	10.29%	59.96%	174,808	99.70%	59.81%	526	0.30%	85.70%	25.89%	0.80%	123.46	332.34
4	509	550	42	170,664	10.02%	69.98%	170,245	99.75%	69.86%	419	0.25%	90.08%	20.21%	0.72%	137.20	406.31
3	465	509	43	170,661	10.02%	80.00%	170,249	99.76%	79.92%	412	0.24%	94.38%	14.46%	0.66%	149.79	413.23
2	394	465	71	166,082	9.75%	89.75%	165,753	99.80%	89.70%	329	0.20%	97.82%	8.11%	0.61%	162.23	503.81
1	344	394	51	174,597	10.25%	100.00%	174,388	99.88%	100.00%	209	0.12%	100.00%	0.00%	0.56%	176.90	834.39
Total	.	.	.	1,703,196	100.00%	.	1,693,622	99.44%	.	9,574	0.56%	.	37.45%	.	.	176.90

KS	GINI	AR	ROC
37.45%	48.21%	48.48%	74.24%



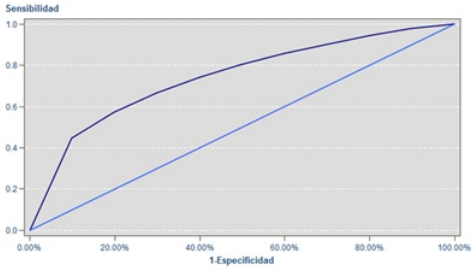


Tabla 26: Tabla de Validacion - Noviembre 2013 (Construcción)

Tabla de Validación:
Base=sasdl.SAS_POSTPAGO_ABT_VSMS_201312 / Exclusión=Exclusion_VSMS / BGI=TARGET_Ontop_VSMS / Score=score1000_VSMS

RANGO	Min Score	Max Score	Dif	Total	%Total	%Total Acum	Malos	Tasa Malos	%Malos Acum	Buenos	Tasa Buenos	%Buenos Acum	KS	Tasa Buenos Acum	Odds Acum	Odds
10	805	989	184	174,461	9.97%	9.97%	170,489	97.72%	9.79%	3,972	2.28%	40.51%	30.72%	2.28%	42.92	42.92
9	731	805	74	175,619	10.03%	20.00%	174,313	99.26%	19.81%	1,306	0.74%	53.84%	34.03%	1.51%	65.33	133.47
8	680	731	51	174,625	9.98%	29.97%	173,630	99.43%	29.78%	995	0.57%	63.98%	34.20%	1.20%	82.65	174.50
7	634	680	46	175,335	10.02%	39.99%	174,520	99.54%	39.81%	815	0.46%	72.30%	32.49%	1.01%	97.76	214.14
6	592	634	42	174,519	9.97%	49.96%	173,842	99.61%	49.80%	677	0.39%	79.20%	29.41%	0.89%	111.63	256.78
5	549	592	43	175,622	10.03%	59.99%	175,059	99.68%	59.85%	563	0.32%	84.94%	25.09%	0.79%	125.10	310.94
4	502	549	46	175,168	10.01%	70.00%	174,653	99.71%	69.89%	515	0.29%	90.20%	20.31%	0.72%	137.57	339.13
3	460	502	42	172,924	9.88%	79.88%	172,478	99.74%	79.79%	446	0.26%	94.75%	14.95%	0.66%	149.53	386.72
2	394	460	66	172,347	9.85%	89.72%	172,044	99.82%	89.68%	303	0.18%	97.84%	8.16%	0.61%	162.74	567.80
1	344	394	50	179,907	10.28%	100.00%	179,695	99.88%	100.00%	212	0.12%	100.00%	0.00%	0.56%	177.55	847.62
Total	.	.	.	1,750,527	100.00%	.	1,740,723	99.44%	.	9,804	0.56%	.	34.20%	.	.	177.55

KS	GINI	AR	ROC
34.20%	45.57%	45.82%	72.91%

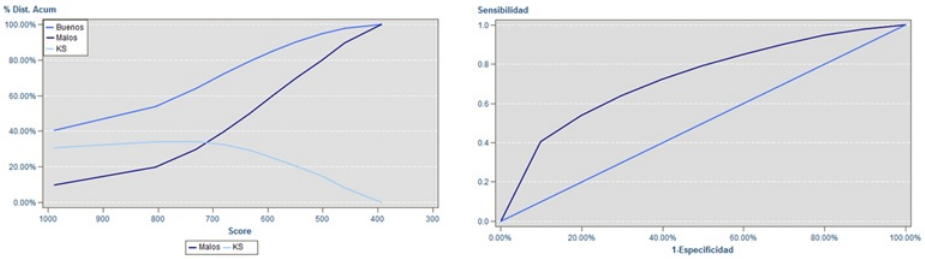


Tabla 27: Tabla de Validacion - Diciembre 2013 (Validación)

Tabla de Validación:
Base=sasdl.SAS_POSTPAGO_ABT_VSMS_201401 / Exclusión=Exclusion_VSMS / BGI=TARGET_Ontop_VSMS / Score=score1000_VSMS

RANGO	Min Score	Max Score	Dif	Total	%Total	%Total Acum	Malos	Tasa Malos	%Malos Acum	Buenos	Tasa Buenos	%Buenos Acum	KS	Tasa Buenos Acum	Odds Acum	Odds
10	828	989	162	180,550	10.00%	10.00%	176,170	97.57%	9.81%	4,380	2.43%	41.36%	31.54%	2.43%	40.22	40.22
9	750	828	78	180,534	10.00%	20.00%	179,077	99.19%	19.79%	1,457	0.81%	55.12%	35.33%	1.62%	60.86	122.91
8	692	750	58	178,032	9.86%	29.86%	177,015	99.43%	29.65%	1,017	0.57%	64.72%	35.07%	1.27%	77.66	174.06
7	644	692	49	182,362	10.10%	39.96%	181,507	99.53%	39.77%	855	0.47%	72.80%	33.03%	1.07%	92.59	212.29
6	597	644	47	181,262	10.04%	50.00%	180,515	99.59%	49.82%	747	0.41%	79.85%	30.03%	0.94%	105.76	241.65
5	552	597	45	179,952	9.97%	59.97%	179,358	99.67%	59.82%	594	0.33%	85.46%	25.64%	0.84%	118.63	301.95
4	510	552	42	180,632	10.00%	69.97%	180,105	99.71%	69.85%	527	0.29%	90.43%	20.58%	0.76%	130.91	341.76
3	466	510	45	180,681	10.01%	79.98%	180,243	99.76%	79.89%	438	0.24%	94.57%	14.68%	0.69%	143.18	411.51
2	394	465	72	180,059	9.97%	89.95%	179,701	99.80%	89.90%	358	0.20%	97.95%	8.05%	0.64%	155.57	501.96
1	344	394	50	181,449	10.05%	100.00%	181,232	99.88%	100.00%	217	0.12%	100.00%	0.00%	0.59%	169.49	835.17
Total	.	.	.	1,805,513	100.00%	.	1,794,923	99.41%	.	10,590	0.59%	.	35.33%	.	.	169.49

KS	GINI	AR	ROC
35.33%	46.49%	46.77%	73.38%

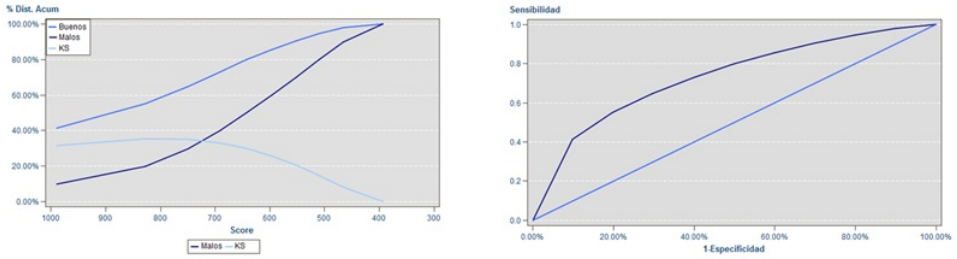
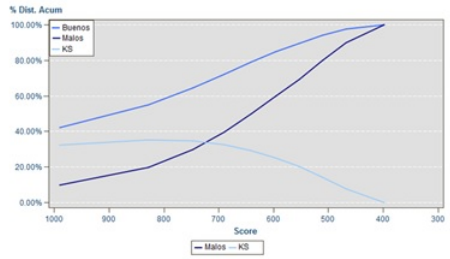


Tabla 28: Tabla de Validacion - Enero 2014 (Construcción)

Tabla de Validación:
Base=sasdl.SAS_POSTPAGO_ABT_VSMS_201402 / Exclusión=Exclusion_VSMS / BGI=TARGET_Ontop_VSMS / Score=score1000_VSMS

RANGO	Min Score	Max Score	Dif	Total	%Total	%Total Acum	Malos	Tasa Malos	%Malos Acum	Buenos	Tasa Buenos	%Buenos Acum	KS	Tasa Buenos Acum	Odds Acum	Odds
10	828	990	162	174,096	10.00%	10.00%	169,817	97.54%	9.81%	4,279	2.46%	42.15%	32.34%	2.46%	39.69	39.69
9	747	828	81	173,546	9.97%	19.96%	172,243	99.25%	19.76%	1,303	0.75%	54.98%	35.22%	1.61%	61.28	132.19
8	690	747	58	174,660	10.03%	29.99%	173,696	99.45%	29.79%	964	0.55%	64.47%	34.68%	1.25%	78.79	180.18
7	641	690	48	172,222	9.89%	39.88%	171,438	99.54%	39.69%	784	0.46%	72.20%	32.50%	1.06%	93.75	218.67
6	596	641	45	176,176	10.12%	50.00%	175,483	99.61%	49.83%	693	0.39%	79.02%	29.19%	0.92%	107.53	253.22
5	553	596	43	174,204	10.00%	60.00%	173,609	99.66%	59.85%	595	0.34%	84.88%	25.03%	0.82%	120.25	291.78
4	511	553	42	163,621	9.40%	69.40%	163,147	99.71%	69.28%	474	0.29%	89.55%	20.27%	0.75%	131.92	344.19
3	468	511	43	183,677	10.55%	79.94%	183,216	99.75%	79.86%	461	0.25%	94.09%	14.23%	0.69%	144.74	397.43
2	399	468	69	175,086	10.05%	90.00%	174,727	99.79%	89.95%	359	0.21%	97.63%	7.67%	0.63%	157.12	486.71
1	344	398	55	174,218	10.00%	100.00%	173,977	99.86%	100.00%	241	0.14%	100.00%	0.00%	0.58%	170.53	721.90
Total	.	.	.	1,741,506	100.00%	.	1,731,353	99.42%	.	10,153	0.58%	.	35.22%	.	.	170.53

KS	GINI	AR	ROC
35.22%	45.88%	46.15%	73.08%



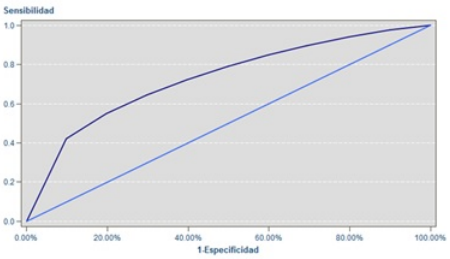
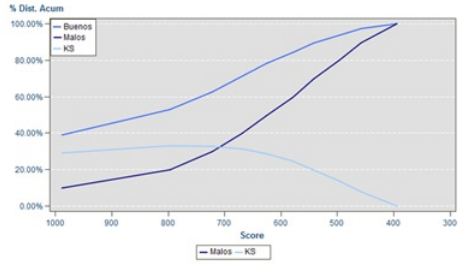


Tabla 29: Tabla de Validacion - Febrero 2014 (Construcción)

Tabla de Validación:
Base=sasdl.SAS_POSTPAGO_ABT_VSMS_201403 / Exclusión=Exclusion_VSMS / BGI=TARGET_Ontop_VSMS / Score=score1000_VSMS

RANGO	Min Score	Max Score	Dif	Total	%Total	%Total Acum	Malos	Tasa Malos	%Malos Acum	Buenos	Tasa Buenos	%Buenos Acum	KS	Tasa Buenos Acum	Odds Acum	Odds
10	797	988	191	167,435	9.96%	9.96%	165,367	98.76%	9.87%	2,068	1.24%	38.97%	29.10%	1.24%	79.97	79.96
9	722	797	75	168,288	10.01%	19.98%	167,546	99.56%	19.87%	742	0.44%	52.95%	33.08%	0.84%	118.47	225.80
8	669	722	53	168,103	10.00%	29.98%	167,591	99.70%	29.87%	512	0.30%	62.60%	32.72%	0.66%	150.66	327.33
7	625	669	44	168,337	10.02%	39.99%	167,879	99.73%	39.90%	458	0.27%	71.23%	31.33%	0.56%	176.82	366.55
6	579	625	46	164,548	9.79%	49.79%	164,172	99.77%	49.69%	376	0.23%	78.31%	28.62%	0.50%	200.33	436.63
5	542	579	38	164,751	9.80%	59.59%	164,439	99.81%	59.51%	312	0.19%	84.19%	24.68%	0.45%	223.14	527.05
4	497	542	45	172,872	10.29%	69.87%	172,593	99.84%	69.81%	279	0.16%	89.45%	19.64%	0.40%	246.38	618.61
3	458	497	39	168,398	10.02%	79.89%	168,176	99.87%	79.85%	222	0.13%	93.63%	13.78%	0.37%	269.22	757.55
2	394	458	63	162,916	9.69%	89.59%	162,719	99.88%	89.56%	197	0.12%	97.34%	7.78%	0.34%	290.45	825.98
1	344	394	51	174,998	10.41%	100.00%	174,857	99.92%	100.00%	141	0.08%	100.00%	0.00%	0.32%	315.69	1,240.12
Total	.	.	.	1,680,646	100.00%	.	1,675,339	99.68%	.	5,307	0.32%	.	33.08%	.	.	315.68

KS	GINI	AR	ROC
33.08%	43.88%	44.02%	72.01%



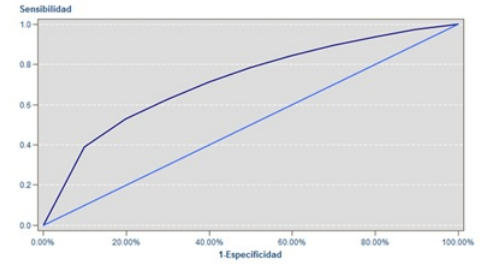


Tabla 30: Tabla de Validacion - Marzo 2014 (Validación)