TESIS DE GRADO EN INGENIERIA INDUSTRIAL

INCREMENTO EN VENTA DE PLAZOS FIJOS UTILIZANDO MODELOS PREDICTIVOS DE RESPUESTA

AUTOR: MARTIN FILA

DIRECTORES: DR. OSCAR BENITEZ

DRA. PAOLA V. BRITOS

M. ING. CLAUDIO RANCAN

-2009-

Agradecimientos

Quiero agradecer en primer lugar a Patricia y Adrián, mis viejos, quienes sin duda son los máximos responsables (aún mas que yo mismo), de que pueda haber comenzado, transcurrido y terminado la carrera de Ingeniería Industrial tal como lo hice.

A la vez quiero dar mis respetos y agradecimientos al Dr. Oscar Benitez, director de esta tesis, quien me brindo su ayuda en forma totalmente desinteresada logrando que este trabajo fuese posible.

Por último agradecer a la Dra. Paola Britos quien me ayudó desde el primer momento y realizó correciones al trabajo, y al ingeniero Claudio Rancan quien finalmente aprobó esta tesis.

Resumen Ejecutivo

Los Bancos impulsan las ventas de sus productos mediante las llamadas Campañas Comerciales. Utilizando contactos telefónicos o personales a través de las sucursales, se ofrece a los clientes la toma de un Préstamo, la colocación de un Plazo Fijo o cualquier otro producto que se pretenda vender en el marco de la campaña.

Dada la gran diversidad de personas que son clientes de un mismo Banco, las condiciones económicas, sociales y financieras de cada uno de ellos son diferentes. Puede esperarse entonces que, en función de estas condiciones particulares y de sus necesidades, no todos tengan igual propensión a aceptar la oferta del producto que se trate.

Mediante este trabajo se creó un modelo predictivo de respuesta utilizando técnicas de Data Mining, el cuál permitió ordenar la cartera de clientes según la propensión estimada para cada uno de aceptar, en este caso, una oferta de colocación de un Plazo Fijo.

Contactando en primer lugar a aquellos clientes más propensos a aceptar la oferta, se incrementó el número de respuestas favorables y en consecuencia se lograron aumentos significativos en los volúmenes de venta de Plazos Fijos.

Executive Brief

Banks thrust their sales by organizing cross-selling campaigns to sell adittional products to their customers. Using telephone calls or direct contact, clients are offered loans, time deposits or any other product, depending on the campaign.

Due to the great diversity of people that can be bank clients, economic, social and financial conditions for each of them can be very different. This differences implie different needs and so, not every client can become attracted by the same products.

This proyect intended and accomplished to create a predictive response model, using Data Mining technics, wich allowed sorting the clients according to their expected attraction towards a particular product. In this case, the product was time deposits.

By reaching in the first place those clients on top of the list, the number of positive responses was increased and, as a result, a significant increase in time deposits sales was achieved.

Tabla de Contenidos

1 INTRODUCCION

Para impulsar los niveles de ventas, las empresas implementan estrategias de marketing cuyo objetivo es promocionar sus productos o servicios entre el público en general. Mediante esta acción, se busca que quien recibe el mensaje se sienta atraído por el producto y, en consecuencia, realice la compra.

Dos aspectos principales son los que caracterizan a estas acciones de promoción que, en los Bancos, se denominan Campañas Comerciales y cuyo objetivo es vender la mayor cantidad posible de productos o servicios.

El primero de ellos tiene que ver con el alcance, es decir, el número de personas a quien se quiere enviar el mensaje. A ese grupo se lo llama grupo objetivo (target-group).

El segundo aspecto esta relacionado con la calidad del grupo objetivo. Es decir, la respuesta obtenida. Por ejemplo, podemos comunicar a todas las mujeres del país que nuestra empresa ha sacado al mercado un nuevo modelo de taladro roto percutor y sin embargo el nivel de respuesta será lo suficientemente bajo como para considerar la campaña un fracaso. Si, en cambio, el grupo objetivo hubiesen sido hombres de entre 25 y 45 años, se puede afirmar casi con seguridad que se habrían logrado mejores resultados.

Ambos aspectos son determinantes en el éxito de las campañas y de su combinación surgen los niveles de venta finales. Según el alcance definido y cuanto se busque conocer el grupo objetivo, se estarán utilizando una u otra estrategia de marketing.

Si bien el alcance elegido dependerá del presupuesto disponible para las campañas y a su vez del mercado real que se estime para el producto o servicio, la calidad del grupo objetivo siempre es deseable y cuanto mejor sea, mejores resultados se obtendrán. No obstante esto, cabe aclarar que lograr dicho conocimiento tiene un costo asociado, el cuál, utilizando Data Mining, es lo suficientemente bajo como para ser una alternativa irrefutable.

Para lograr un alto nivel de respuesta, es necesario conocer el universo de clientes, estudiarlo, identificar cada individuo según sus patrones de comportamiento y luego determinar a qué grupo se quiere dirigir el esfuerzo de venta. A este procedimiento se lo llama segmentación.

Hay productos para los cuales la segmentación es más simple que para otros. Comprender el comportamiento de los clientes de un banco puede ser algo complejo pero a la vez la oportunidad es mayor ya que puede agregarse mucho valor.

Las publicidades tradicionales tal como las que pueden verse en la vía pública o en programas de televisión son publicidades masivas y se caracterizan por un gran alcance y una baja segmentación. Es decir, comunican a un gran número de personas aunque no identifican de manera individual a los receptores del mensaje.

Es fácil entender que, para lograr que una estrategia tenga posibilidades de ser exitosa, debe asegurar al menos uno de los dos aspectos que las caracterizan: un gran alcance o, ante un alcance reducido, una alta respuesta que asegure el nivel de ventas esperado.

Ninguno de estos dos casos se da en las estrategias comerciales del Banco en estudio.

En particular se tratará el caso de las campañas de Plazo Fijo donde el volumen de clientes alcanzado no supera el 1% del total de la cartera que, actualmente, ronda los 200.000 clientes. El problema reside en como se elige ese 1% (2.000 clientes)

Del total de clientes se realizan una serie de filtros que buscan asegurar en el grupo objetivo algunas características mínimas consideradas de importancia para lograr una buena respuesta en los clientes. Por ejemplo: “El cliente tiene que pertenecer al NSE A, B o C y ser mayor de 21 años”.

Al filtrar según estos criterios, se obtiene una base de 100.000 clientes que aún supera ampliamente el 1% definido como alcance. Por ende es necesario realizar una nueva selección sobre este grupo. ¿Cual es la estrategia utilizada para hacerlo? En forma aleatoria.

La situación actual es la siguiente: “Las tasas de éxito resultantes están por debajo de las esperadas”

Ante esta situación cabe preguntarse: ¿Es esta la mejor tasa de respuesta que puedo obtener? ¿Todos los clientes del Banco tienen la misma propensión a adquirir el producto? ¿Es posible saber algo acerca de la propensión de unos y otros a adquirir el producto ofrecido?

La respuesta para la primera pregunta es: Probablemente sí, mientras se siga utilizando la selección aleatoria de clientes.

La respuesta para la segunda pregunta es No. Cada persona tiene características personales y patrones de comportamiento observables que lo hacen más propenso a adquirir cierto producto o servicio. Por ejemplo: un hombre es más propenso a adquirir un taladro que una mujer.

Respecto de la tercera pregunta, la respuesta es Sí. Es posible hacerlo y mediante este trabajo se buscara mostrar en que forma es posible aumentar el éxito de las campañas comerciales simplemente mediante la selección inteligente del grupo objetivo. Es decir, eligiendo mejor a quien vender.

Para hacerlo se buscará asociar a cada cliente con la propensión que tiene de adquirir el producto ofrecido en determinada campaña. La selección ya no se hará mediante filtros o en forma aleatoria sino que se elegirá a quienes presenten una mayor probabilidad de adquirir el producto. Como resultado se espera aumentar la tasa de éxito de las campañas y por ende aumentar el beneficio económico percibido por el Banco.

2 EL NEGOCIO BANCARIO

2.1 Introducción

Un banco es una institución financiera cuya actividad económica se desarrolla en base a dos grandes tipos de operaciones: Operaciones *pasivas* y operaciones *activas*.

Las primeras, las *pasivas*, consisten en la captación de fondos. Por estos fondos que el banco toma prestado de sus clientes, paga a cambio una tasa de interés.

Por otro lado, mediante las llamadas operaciones *activas*, el Banco presta parte de ese dinero a otros clientes y recibe a cambio el pago de una tasa de interés más alta que la anterior. Se describen a continuación ambos tipos de operaciones.

2.2 Operaciones pasivas

Las operaciones pasivas son aquellas mediante las cuales el Banco capta (recibe) dinero de sus clientes: se trata de los depósitos de dinero.

Se pueden realizar depósitos de distintas características, los cuales se diferencian según la frecuencia con la que se permita retirar el dinero depositado, la posibilidad de gastar mas dinero del que se dispone y por todo lo cual el Banco paga distintas tasas de interés, según le convenga mas o menos el tipo de acuerdo que hizo con el cliente. Los tipos de depósitos que las personas utilizan con mayor frecuencia en los bancos son los siguientes:

- Cuentas corrientes
- Cajas de Ahorros
- Depósitos a Plazo Fijo

Las cuentas corrientes y las cajas de ahorro son “movilizables” en cualquier momento, es decir, el cliente puede retirar el dinero que tiene depositado cuando lo desee. Esto es una ventaja para quien realiza el depósito ya que le brinda flexibilidad para disponer del dinero. Sin embargo, como las extracciones son espontáneas, el

Banco no puede anticipar los volúmenes exactos de dinero de los que dispondrá en estas cuentas.

Dado que parte de los depósitos es a su vez entregado como préstamos a otros clientes mediante las operaciones activas, el Banco no dispone del 100% de los depósitos en sus cuentas.

Si los depositantes de cuentas corrientes y cajas de ahorro comenzaran a retirar el dinero de sus cuentas en forma inesperada, el Banco podría tener que negar ese derecho ya que, como se dijo antes, no tiene todo el dinero en su poder. En este caso podría haber consecuencias legales y hasta existe la posibilidad de quiebra.

Como existe un riesgo asociado a este tipo de depósitos, los intereses que el Banco paga a sus clientes por ellos son inferiores a los que paga por los depósitos a Plazo Fijo.

2.2.1 Productos Pasivos: Plazos Fijos

El Depósito a Plazo Fijo es un producto mediante el cual el cliente realiza un depósito de dinero en un Banco. Tienen fecha de inicio y fecha de fin. Durante ese período, el cliente no puede disponer del dinero. Una vez llegada la fecha de fin de contrato, el cliente obtiene el total del capital depositado al inicio más los intereses acumulados durante el período de contrato. En este caso, el Banco conoce exactamente por cuanto tiempo dispondrá del dinero y en que momento deberá devolverlo. Esto permite reducir el riesgo e incrementar el rendimiento económico que el Banco obtiene por él. Por estas razones, para los Bancos resulta un gran beneficio que los clientes elijan este producto y por ello pagan tasas de interés mayores que en el caso de cajas de ahorro o cuentas corrientes.

2.3 Operaciones activas

Las operaciones activas son aquellas mediante las cuales el banco entrega parte del dinero del que dispone a sus clientes y les cobra a cambio una tasa de interés. El dinero prestado es en parte capital del Banco y en parte corresponde al dinero depositado por los clientes.

Los productos activos más comunes mediante los cuáles un cliente puede solicitar dinero a un Banco son los Préstamos.

2.4 Margen de intermediación

Sabiendo que los bancos pagan una cantidad de dinero a las personas u organizaciones que depositan dinero (intereses de captación) y que cobran dinero por dar préstamos (intereses de colocación), cabe preguntarse de dónde obtiene un banco sus ganancias. La respuesta es que los de intereses de colocación, en la mayoría de los casos, son más altos que los intereses de captación; de manera que los bancos cobran más por dar recursos que lo que pagan por captarlos. Como resultado se obtiene un beneficio económico.

A la diferencia entre la tasa de interés de colocación y la de captación se lo denomina margen de intermediación y es la fuente principal de ingresos para un Banco.

3 CAMPAÑAS COMERCIALES

3.1 Introducción

Como se explicó anteriormente la actividad principal del Banco consiste en tomar dinero pagando a cambio cierta tasa de interés y luego prestar parte de ese dinero cobrando a cambio una tasa mayor. Tanto para prestar como para captar dinero, el Banco utiliza distintos productos ya sean pasivos o activos. Mediante dichos productos, un cliente puede efectuar un depósito (Caja de Ahorro, Cuenta corriente, Plazo Fijo) o pedir prestado dinero (Préstamos).

Si bien los clientes solicitan estos productos por iniciativa propia, el Banco encara acciones de promoción para impulsar sus ventas y de esa forma incrementar las operaciones activas y pasivas para luego, mediante el margen de intermediación, obtener mayores ganancias. Estas acciones tienen el nombre de Campañas Comerciales y tanto en la definición como en la posterior implementación, se involucra a distintas áreas del Banco entre las cuáles están las áreas de Marketing y de Inteligencia Comercial. Cuando las campañas están dirigidas a clientes existentes a los cuales se busca venderles productos adicionales a los que actualmente poseen, se las denomina campañas de cross-selling.

Los canales elegidos usualmente para contactarse con los clientes son vía e-mail, teléfono, a través de las sucursales o, en ocasiones, directamente a través de los ATMs (cajeros electrónicos) dependiendo del producto del que se trate.

3.2 Planificación y ejecución

Cuando se quiere salir al mercado con una campaña comercial, es necesario definir: el **producto** a vender, el **canal** mediante el cual se va a vender, el **período** durante el cual se desarolla la campaña y por último **a quién** se quiere vender (ver Ilustración 1).

Ilustración 1: Variables de control de las Campañas Comerciales

PRODUCTO OFRECIDO: Es el producto que se desea vender en la campaña; dependerá de las necesidades del Banco en cada momento.

Puede responder a una necesidad financiera, es decir, aumentar la captación o la colocación de fondos de cierto tipo (a plazo fijo, movilizable, etc) o a la fidelización de clientes mediante el ofrecimiento de productos que el mercado valora en ese momento.

Los atributos del producto (tasa de interés, costos de mantenimiento, disponibilidad de los fondos) se eligen en base a la rentabilidad esperada, en base a las necesidades del mercado y a su vez a través de benchmarking (análisis de la competencia)

CANALES DE VENTAS: Se trata del punto de contacto elegido para comunicarse con los clientes: sucursales, telefónico, cajeros automáticos, banca telefónica, o por

carta. En general se refuerza la acción de los canales de venta con folletos informativos que deben ser diseñados, impresos y distribuidos.

Para determinar que canal se utilizará en una campaña comercial se tienen en cuenta dos cuestiones. En primer lugar, la eficiencia de cada una de ellos en campañas pasadas. En promedio, algunos canales obtienen mejores resultados que otros. No obstante, es necesario tener en cuenta en segundo lugar la disponibilidad que presentan al momento de lanzar la campaña. Las sucursales, por ejemplo, están abocadas a otras tareas como pueden ser campañas lanzadas previamente, con lo cual habría que optar en ese caso por utilizar la Banca Telefónica o los cajeros automáticos.

PERIODO: El período es simplemente el espacio de tiempo durante el cual estará vigente la campaña. El contexto económico, político y social es un factor importante a considerar dado que la venta de un producto esta íntimamente relacionada con las necesidades de las personas en un momento dado.

GRUPO OBJETIVO: Son aquellos clientes a quienes se contactará para intentarles vender el producto. De su decisión depende el éxito o el fracaso de la campaña. Al igual que la resolución de un juicio depende del jurado elegido, las ventas resultantes dependen del grupo objetivo seleccionado.

3.3 Presupuesto

Realizar campañas comerciales tiene un costo para el Banco. El mismo depende de las características de la campaña pero en general los costos se dividen en costos de diseño, impresión y distribución (si se envían a domicilio) de los folletos que se entregan a los clientes como refuerzo del contacto directo. Si la campaña es telefónica y se realiza mediante un proveedor externo también es necesario pagar por dicho servicio.

El costo de horas hombre se considera un costo hundido ya que los sueldos fijos de los empleados se pagarán se realice o no la campaña. No obstante es necesario considerar la disponibilidad de horas hombre de cada canal (sucursal, call center interno, fuerza de ventas) las cuales también limitan el volumen de clientes a contactar.

El Banco asigna una parte del presupuesto a las Campañas Comerciales, el cuál limita tanto los costos como las horas hombre disponibles. Esto implica que no se puede contactar al 100% de los clientes. Una vez definido el presupuesto y en función de los costos, queda determinado la cantidad de contactos que se pueden realizar en cada campaña. Una campaña comercial contacta, en promedio, 3000 clientes.

3.4 Factores que influyen en su éxito

Para mejorar el éxito de una campaña comercial, se puede intervenir en cualquiera de los 4 puntos mencionados anteriormente: Producto (y sus atributos), Canal de venta (y pieza de marketing), Período (contexto) en el cual se lanza la campaña y por último grupo objetivo (grupo de clientes a contactar). Dicha relación se resume en la siguiente fórmula:

V(x)= f (Producto, Canal, Periodo, Grupo Objetivo) **V(x): Ventas**

Mejorando los atributos del producto (desde el punto de vista del cliente) puede incrementarse la demanda. Por ejemplo, ofreciendo una mejor tasa de interés. Eligiendo canales de contacto directos (sucursales) puede lograrse una mayor respuesta que si solo se envía una carta al domicilio del cliente. Lanzando la campaña en el momento adecuado, pueden aprovecharse situaciones particulares como, por ejemplo, la necesidad de préstamos hipotecarios dada la escasez de los mismos en el mercado.

Por último, el grupo objetivo se vuelve importante al existir una restricción en el alcance de la campaña y por ende, ser necesaria la selección de cierto grupo de clientes a contactar entre el total de la cartera. No todos los clientes responden de igual forma a las ofertas. A mayor respuesta, mayor volumen de ventas.

3.5 ¿Cómo se mide la efectividad de una campaña comercial?

Si bien las campañas son un medio mediante el cual se busca ayudar al Banco a lograr sus objetivos, es decir, lograr cierto nivel de ingresos través de la toma y préstamo de dinero, las mismas tienen objetivos propios.

El indicador más importante de la efectividad de una campaña comercial es la tasa de éxito (hit rate) que resulta simplemente de la división matemática entre el número de ventas efectivas y el número de clientes contactados, expresada en porcentaje.

El valor de este indicador al término de la campaña indicará en que medida se alcanzó el objetivo propuesto. El máximo valor posible corresponde al 100% que implicaría una situación ideal donde todos y cada uno de los clientes contactados adquieren el producto. Si bien es improbable alcanzar esa efectividad, en cada campaña se intenta obtener un valor lo más cercano posible a él.

El costo de contacto de cada cliente es significativamente inferior al ingreso que resulta de la venta efectiva. Si se gastara más dinero en convencer a un cliente que el esperado como ingreso al lograr la venta, no existiría posibilidad alguna de obtener beneficios, aún logrando el ideal de 100% en la tasa de éxito.

A continuación se muestra la ecuación para el cálculo de la tasa de éxito y el beneficio obtenido a raíz de la campaña.

$$E(\%) = V / C$$

$$B = V * Iu - C * Cu$$

E(%): Tasa de Éxito

V: Ventas (cantidad de clientes)

C: Clientes Contactados

Iu: Ingreso unitario

Cu: Costo unitario del contacto

B: Beneficio de la campaña

Igualando B = 0 es fácil ver que la tasa de éxito mínima debe ser superior a la relación entre el costo y el beneficio unitario, tal como muestra la siguiente ecuación:

Para B > 0	E > Cu / Iu

Aumentar las tasas de éxito es equivalente a aumentar las ventas y por ende a incrementar el beneficio obtenido por el Banco. La relación es directa.

No existen costos fijos a considerar dado que todo gasto depende del volumen de clientes que se contacta.

4 ESTRATEGIAS DE MARKETING

4.1 Marketing Masivo

El marketing masivo es una estrategia de venta mediante la cuál una empresa decide ignorar las diferencias entre los distintos segmentos de clientes (personas agrupadas según sus características) yendo detrás de todo el mercado con un único mensaje. La hipótesis en la cuál se basa esta estrategia es que alcanzando la mayor audiencia posible se maximiza la exposición al producto y por ende como resultado se obtendrá un mayor número de ventas. Un claro ejemplo son las publicidades a través de radio o televisión.

Sin embargo desde algún tiempo esta forma de llegar a los consumidores ya no logra ser tan efectiva como lo fue durante la mayor parte del siglo XX. Según una encuesta realizada por Yankelovich Partners Inc., el 65% de los consumidores se sienten constantemente bombardeados con demasiadas publicidades. A su vez, el 61% cree que el volumen de avisos esta fuera de control. El presidente de esta firma, J. Walker Smith, agrega: “El problema es que actualmente a los encargados de marketing no les preocupa la imprecisión, sin darse cuenta del impacto que eso tiene en aumentar la resistencia y reducir la productividad del marketing.[1]

[1] http://www.clickz.com/3344701 - Consumers Becoming Marketing-Resistant

En línea con lo anterior, en un artículo titulado “Cambiando de Marketing Masivo a Micro Marketing”, Ian Durrell Líder de Desarrollo de Negocios de Only Finance Ltd., expone: “Para satisfacer a los consumidores, las empresas deben pasar de transmitir mensajes masivos a un gran número de personas a alcanzar de forma individual a cada uno de ellos” [2]

4.2 Cambiando de estrategia

El comportamiento del mercado ha cambiado. Actualmente la diversidad en las ofertas es tan grande que frente a tantos avisos, las publicidades pierden impulso para tratar de instalarse en la mente de los consumidores. En consecuencia, solo una pequeña porción de los consumidores adquirirá el producto o servicio.

El marketing de segmentos no es una novedad. Se trata de identificar dentro del total de clientes potenciales, aquellos que cumplan con ciertas características (segmento) definidas según el producto para luego dirigir la publicidad a ese segmento en particular. Si bien el nivel de segmentación es mayor que en el marketing masivo, el marketing de segmentos también ha dejado de ser una estrategia novedosa y de gran utilidad.

[2] Shifting From Mass Marketing to Micro Marketing ttp://www.marketingprofs.com/5/durrell1.asp

En un artículo denominado “Uno por Uno”, Horacio Marchand (Licenciado en Adm. de Empresas y MBA en Austin, Texas) escribe: “Primero fue atender clientes masivamente y de forma anónima. Luego apareció la segmentación y se atendía a grupos específicos de clientes. Hoy se busca atender un cliente a la vez y generar conocimiento preciso de sus hábitos y patrones de consumo. El Segmento de Uno se está imponiendo.” El autor resalta en este párrafo la importancia de contactar a los clientes Uno a Uno, es decir, en forma individual. A esta forma de Marketing se la denomina “Marketing Directo”.

4.3 Marketing Directo

Se llama Marketing Directo a la estrategia de venta mediante la cual se contacta a los clientes en forma individual ya sea por teléfono, e-mail o en persona. Es un enfoque opuesto al Marketing Masivo. En lugar de enviar un único mensaje a toda la audiencia de potenciales clientes, se busca lograr contactos individuales con aquellos clientes que sean más propensos a aceptar la oferta. No obstante, conocer quienes son los clientes que con mayor probabilidad aceptarán la oferta implica conocer sus características y necesidades. Lo que se espera es lograr un mayor nivel de ventas sobre los clientes contactados.

4.4 La importancia del grupo objetivo

Las campañas comerciales utilizan marketing directo como estrategia de ventas. Esto puede hacerse mediante el envío de e-mails, en el caso de campañas vía Internet, realizando llamadas telefónicas o dirigiéndose a ellos vía correo postal.

En los casos mencionados, es necesario tener en cuenta dos aspectos: el costo de contactar a todos los clientes y el efecto negativo que puede provocar contactar a clientes que no quieren ser contactados. En el primer caso es necesario reducir los costos. Contactar a todos los clientes de la cartera (actualmente el Banco tiene 200.000 clientes) mediante correo postal o llamadas telefónicas son tan elevados que superan la ganancia esperada ya que solo una pequeña porción aceptará la oferta. En el segundo caso, es importante tener en cuenta que llamadas o folletos indeseados logran un efecto negativo que reduce el mercado potencial para futuras acciones de venta.

Además de los costos y el efecto negativo de contactos indeseados, las horas hombre disponibles, por ejemplo, en el caso de contactos telefónicos, es limitada.

En función de lo dicho anteriormente, se vuelve necesario seleccionar sub-grupos de clientes para cada campaña. Enfocar la campaña a aquellos clientes que son más propensos a adquirir el producto es un aspecto determinante para el éxito. La adecuada selección del grupo objetivo se vuelve, entonces, imprescindible.

5 EL PROBLEMA: SELECCIÓN DEL GRUPO OBJETIVO

5.1 Estrategias actuales de selección

Como se explicó anteriormente, que haya que elegir un sub-grupo de clientes a quien ofrecer los productos hace necesario definir los criterios en los cuales se basará dicha selección buscando al mismo tiempo que el grupo resultante tenga una mayor propensión a adquirir el producto. Dado lo observado en las estrategias actuales de selección del target de sus campanas de ventas, el Banco Itau no esta logrando cumplir con lo mencionado anteriormente.

Para poder entender en forma más detallada el problema, se hará una descripción del proceso de selección del grupo objetivo que tiene lugar en forma previa al lanzamiento de una campaña. La selección tiene 2 etapas.

5.1.1 Primera etapa: Filtrado

Al comienzo del proceso de selección, se cuenta con el total de clientes de la cartera, actualmente, dicho total alcanza los 200.000 individuos. El total de clientes se representa en la ilustración 2 donde A implica un cliente que, para cierto producto, canal y período, aceptaría la oferta comercial. Por otro lado, los N resultarían en contactos negativos.

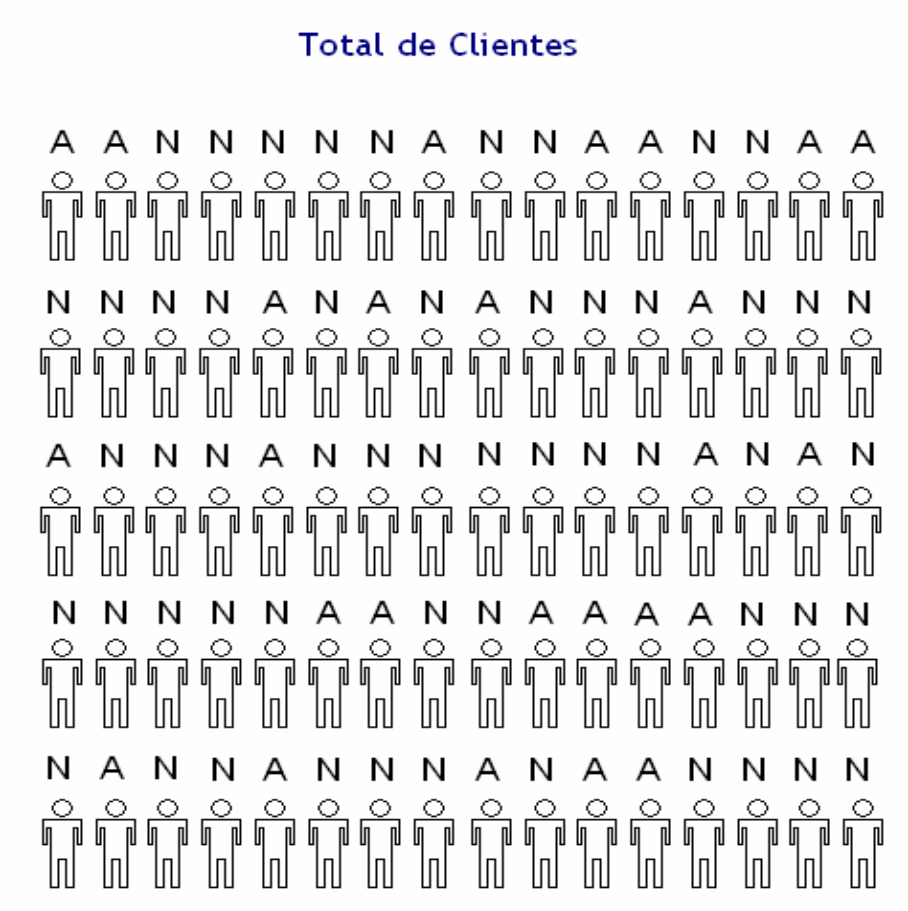


Ilustración 2: Total de Clientes (A: aceptará, N: No aceptará)

Para reducir el número de clientes a considerar para un potencial contacto, se realiza un primer filtro donde, se seleccionan aquellos clientes que cumplan las condiciones del tipo: “personas mayores a 30 años”, “NSE A o B”, “antigüedad mayor a dos años en el Banco”.

Habiendo concluido este primer paso, la base se ha reducido de 200.000 clientes a un total que, en general, se ubica en 100.000 clientes (ver Ilustración 3), siendo este un número aún elevado considerando que se contactarán aproximadamente 3.000.

Ilustración 3: Clientes seleccionados en primer filtro

El problema que se detecta en esta primera etapa es que los criterios de selección mencionados se definen en base a reglas preestablecidas y de las cuales no se ha comprobado que realmente estén relacionadas con la propensión de los clientes a aceptar el producto. Esto pone en duda si realmente, los individuos elegidos (pintados de verde en el grafico) realmente tienen mayores propensión a adquirir el producto que los dejados fuera (pintados de blanco en el grafico) Las mismas reglas se aplican en todos los casos.

5.1.2 Segunda Etapa: Selección aleatoria

La segunda parte consiste simplemente en tomar el grupo resultante del filtro inicial y realizar sobre dicho grupo una selección aleatoria. Es decir, partiendo de 100.000 clientes, se eligen 3000 individuos *al azar*.

Ilustración 4: Clientes seleccionados al azar

Estos individuos forman, finalmente, el grupo objetivo a contactar.

Se muestran de verde en la Ilustración 4, aquellos clientes que fueron elegidos y finalmente aceptaron la oferta, mientras que aquellos pintados de rojo fueron elegidos pero no aceptaron la oferta.

El problema aquí detectado es la selección aleatoria la cual no distingue entre las distintas características de los clientes y los considera igualmente atractivos. La variabilidad encontrada en una muestra de 100.000 clientes (1/2 del total de la cartera) es tan alta que tomarlos como semejantes demuestra una ineficiencia en la selección.

5.2 Conclusión de las estrategias actuales

En ambas etapas se encontraron puntos débiles en la estrategia de selección. En la primera etapa, se trato de la falta de fundamentos en los criterios utilizados, mientras que, en la segunda etapa, la selección aleatoria no agrega valor a la misma.

Si fuese posible refinar la selección de los clientes a contactar en función de criterios probados bajo rigor estadístico y aplicables al total de la cartera, sería posible distinguir aquellos clientes que tienen mayor propensión a adquirir el producto y así aumentar el éxito en las ventas de las campañas comerciales.

5.3 Tasa de éxito actuales

Dado que el número de clientes contactados es distinto en cada campaña, cuando se quiere determinar el éxito de una campaña se lo hace en términos de hit-rate y no de ventas absolutas. Es decir, el porcentaje de clientes que realizaron un depósito a Plazo Fijo sobre el total de clientes contactados.

A continuación se muestran las tasas históricas de éxito de campañas de venta de Banco Itau para Plazos Fijos[3].

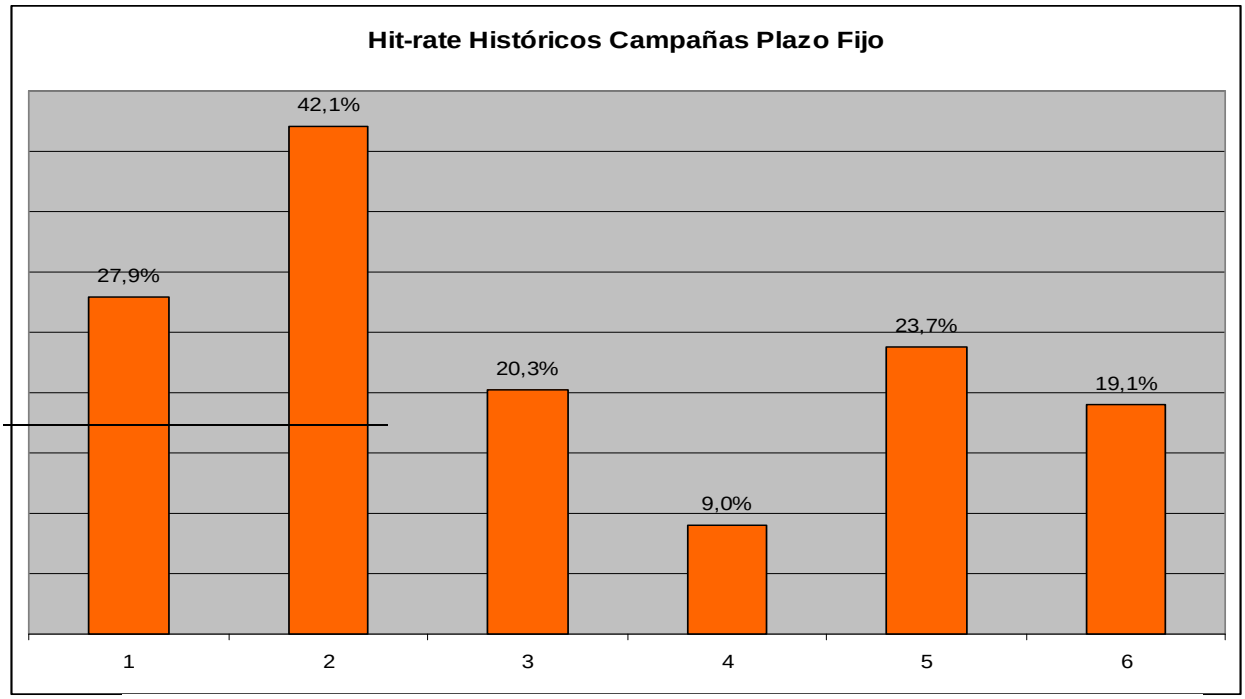


Gráfico 1: Tasas de éxito en campañas de Plazos Fijos

[3] Datos históricos Campañas Comerciales Plazos Fijos – Inteligencia Comercial Banco Itau

El hit-rate promedio es de 23.7%. Es decir que, en promedio, de cada 100 clientes contactados, casi 24 han aceptado la oferta. No obstante, a la hora de evaluar el impacto de la solución propuesta mas adelante, este número no será representativo dada la heterogeneidad entre las distintas campañas.

Como se explicó en el capítulo 3, “Campañas Comerciales”, existen 4 factores que determinan el grado de éxito de una campaña comercial: Producto ofrecido, período en que se desarrolla la campaña, canal y pieza de marketing, grupo objetivo contactado.

Las tasas de éxito observables en el gráfico anterior corresponden a campañas donde los productos ofrecidos en cada una no poseían las mismas características (distintas tasas de colocación), fueron ofrecidos en distintos momentos con lo cuál el contexto económico y social no fue el mismo, se utilizaron distintos canales influyendo en el éxito de cada contacto y por último las personas contactadas no fueron las mismas.

En vistas de esto, para medir el impacto obtenido mediante el cambio en una de estas 4 variables, será necesario realizar 2 pruebas simultáneas donde solo sea distinta la variable a monitorear. Las otras 3 permanecerán fijas. De esta forma se podrá asociar el cambio observado en el hit-rate con el cambio ejercido sobre la variable de control que en este caso será el Grupo Objetivo Contactado.

5.4 Objetivo y Alcance del Proyecto

El objetivo del trabajo es incrementar el hit-rate de las campañas comerciales mediante una mejor selección del grupo objetivo (variable de control). Si bien se hará para la venta del producto Plazo Fijo, el procedimiento es similar para Préstamos, Tarjetas de Crédito, etc. No se estudiará el impacto del canal de contacto ni se ahondará en la definición de los productos a ofrecer. Tampoco se estudiará el efecto del contexto en cada caso. Lo que se quiere lograr es que, una vez determinados el producto, el canal y el período poder determinar el grupo objetivo que mayor propensión tiene a aceptar la oferta.

Para medir el efecto logrado mediante la solución propuesta, se realizarán dos campañas simultáneas: una seleccionando el target group tal como se hace actualmente y la segunda según el procedimiento propuesto como solución.

5.5 Mejora esperada por la solución

Se espera poder aumentar el hit-rate en forma significativa. Lograr mejoras de, al menos, 15% entre ambas estrategias. Más adelante se calculará el impacto económico que implica lograr un aumento en dicho hit-rate o tasa de éxito.

6 DATA MINING

6.1 Introducción

6.1.1 ¿Qué es Data Mining?

Data Mining es el proceso que permite obtener información valiosa mediante la exploración de las grandes bases de datos, con el fin de poder responder preguntas clave de negocios. Durante ese proceso se utiliza un conjunto de técnicas de análisis y modelización que son utilizadas según la pregunta a resolver. La forma en que se logra esto es revelando relaciones implícitas, tendencias, patrones que estaban ocultos previamente para el ojo de los analistas.

6.1.2 Los orígenes del Data Mining

Las técnicas de Data Mining son el resultado de un largo proceso de investigación y desarrollo. Los componentes principales del Data Mining han evolucionado a lo largo de muchas décadas en áreas de investigación como la estadística, inteligencia artificial y aprendizaje computacional.

Redes neuronales, reglas inductivas y otros algoritmos, han sido utilizadas durante muchos años para el desarrollo de sistemas de reconocimiento de patrones, reconocimiento óptico de caracteres, aplicaciones científicas y hasta en las Bolsas de Comercio.

Históricamente solo los científicos y el mundo académico era capaz de contar con las grandes cantidades de datos requeridas para la aplicación de estas técnicas. Sin embargo, actualmente, estas técnicas se han acercado en forma radical al mundo de los negocios gracias al aumento generalizado en el almacenamiento de datos y la capacidad de procesamiento de las computadoras, quedando al alcance de la gran mayoría de las empresas.

El desarrollo de los softwares de aplicación ha hecho hincapié en aumentar la facilidad de utilización buscando aumentar el mercado potencial, a la vez que se desarrollaban métodos de análisis más eficientes.

Software con Interfases amigables, desarrollo tecnológico global y almacenamiento de datos, son los factores que hacen posible y práctica la utilización de Data Mining en una amplia variedad de rubros logrando obtener grandes beneficios por su aplicación.

6.2 Modelos de Respuesta

Una de las aplicaciones que ofrece el Data Mining, es la creación de los “Modelos Predictivos de Respuesta”. Al igual que utilizando regresión lineal pueden predecirse resultados futuros en función de una o más variables conocidas, los modelos de respuesta permiten anticipar el resultado futuro de acciones sobre clientes. En este caso, la acción es la oferta de un producto mediante contacto directo.

La forma que tienen estos modelos de predecir el resultado es encontrando, a partir de casos conocidos, cuales son las características comunes a aquellos clientes que han adquirido el producto. Aquellos atributos que resulten fuertemente relacionados con el resultado que se quiere predecir, serán considerados variables predictoras para el modelo.

6.3 La Solución: Modelos de Respuesta aplicados a Marketing Directo

Como se explicó en el punto anterior, los modelos de respuesta permiten conocer, en función de casos conocidos, la propensión que tiene cada cliente a aceptar una oferta. Aplicando el modelo a toda la cartera, puede establecerse quienes son aquellos clientes que con mayor probabilidad “comprarán” el producto, en este caso, Plazos Fijos. En consecuencia se logrará la selección de un grupo objetivo mediante técnicas de Data Mining con fundamentos estadísticos, procurando lograr el mayor nivel de ventas posible.

El modelo es la solución buscada para el problema mencionado en el capitulo 5. Es una herramienta complementaria al Marketing Directo que mejora la eficiencia de esta estrategia. A continuación se describe la estrategia propuesta de selección del grupo objetivo utilizando modelos de respuesta logrados con Data Mining.

6.3.1 Estrategia Propuesta

Lo que se logra mediante esta estrategia es asignar a cada cliente, mediante el modelo de respuesta, la probabilidad esperada de que, luego del contacto, realice un depósito de Plazo Fijo. El esquema que ilustra el procedimiento es el siguiente:

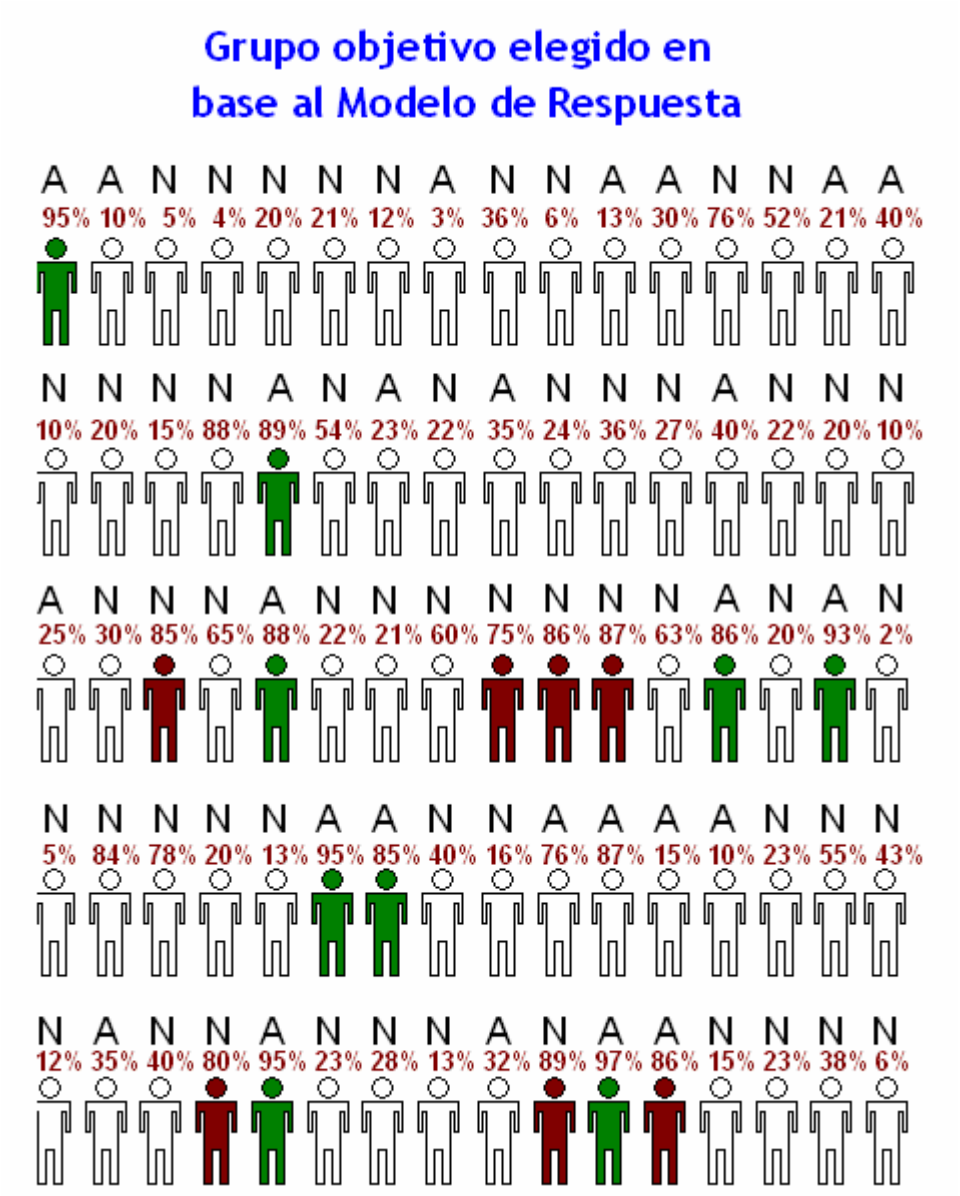


Ilustración 5: Clientes seleccionados con Modelo de Respuesta

Como se ve en la ilustración anterior, los clientes elegidos para el contacto son aquellos con las probabilidades de éxito más altas acorde al modelo.

Dada la incertidumbre inherente a las decisiones humanas, existirán clientes que rechazarán la oferta (Rojo).

6.4 Que se necesita para aplicar Data Mining

Para aplicar Data Mining en la resolución de cierto problema, es necesario asegurar 3 aspectos fundamentales de esta estructura: Software, Hardware y Datos.

6.4.1 Software

Los software de Data Mining pueden separarse en dos grupos. Por un lado están aquellos de licencia comercial que se venden, es decir, por los cuales es necesario pagar para poder utilizarlos. Por otro lado están los de licencia libre, cuya utilización es gratuita. Dentro de cada uno de estos dos grupos, existen distintas opciones de programas las cuales se diferencian por las diferentes herramientas que ofrecen para la resolución de diferentes tipos de problemas.

Se muestran a continuación los resultados de una encuesta realizada por KDnuggetsTM[4], asociación cuya labor consiste en la difusión de las novedades de todo aquello relacionado con Data Mining: noticias, programas, trabajos, cursos, etc.

El fin de la misma fue determinar la popularidad de cada una de las alternativas de software de Data Mining disponibles. La misma se realizó tanto para aquellos software de licencia comercial como para los de licencia gratuita.

Los resultados fueron los siguientes[5]:

[4] Sitio Oficial de KDnuggets, http://www.kdnuggets.com

Como reconocimiento a su desempeño en la difusión de noticias relacionadas con el mundo del Data Mining, KDnuggets ha recibido varios premios siendo uno de ellos el "CRMsearch Editor's Choice Award" que es a su vez una reconocida fuente dedicada a la investigación en herramientas de decisión para el Manejo de la Relación con el Cliente.

[5] Resultados Encuesta KDnuggets:
http://www.kdnuggets.com/polls/2007/data_mining_software_tools.htm

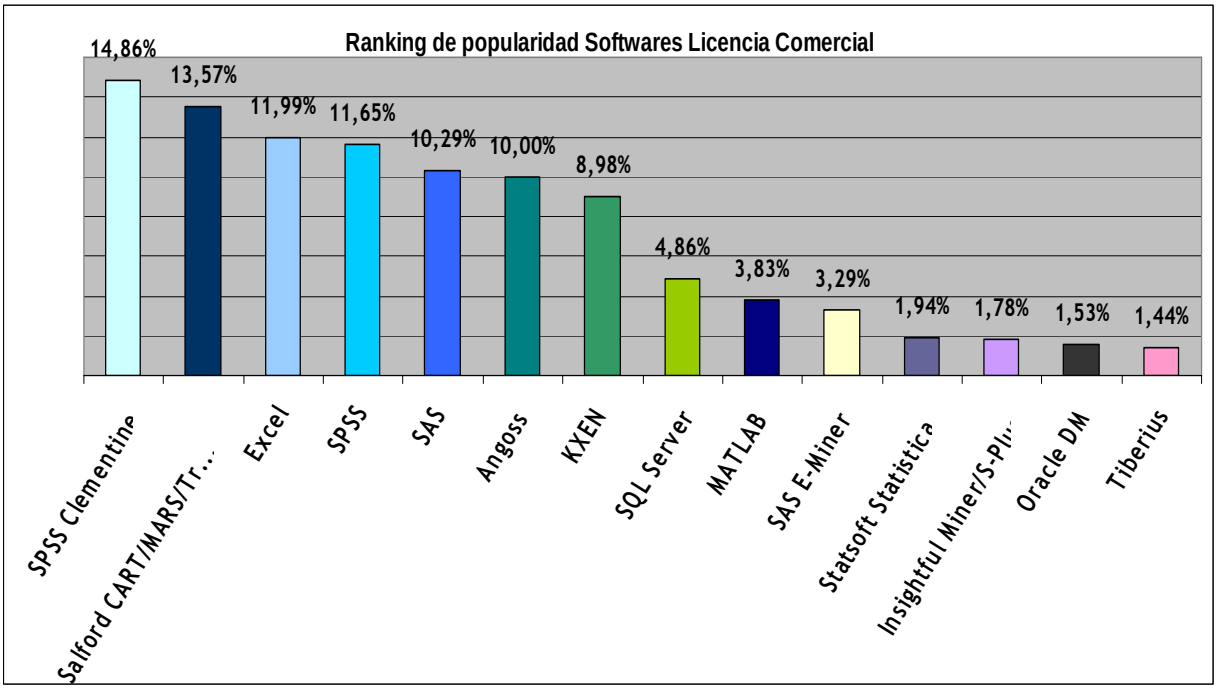


Gráfico 2: Resultados Encuesta sobre software Comerciales de Data Mining

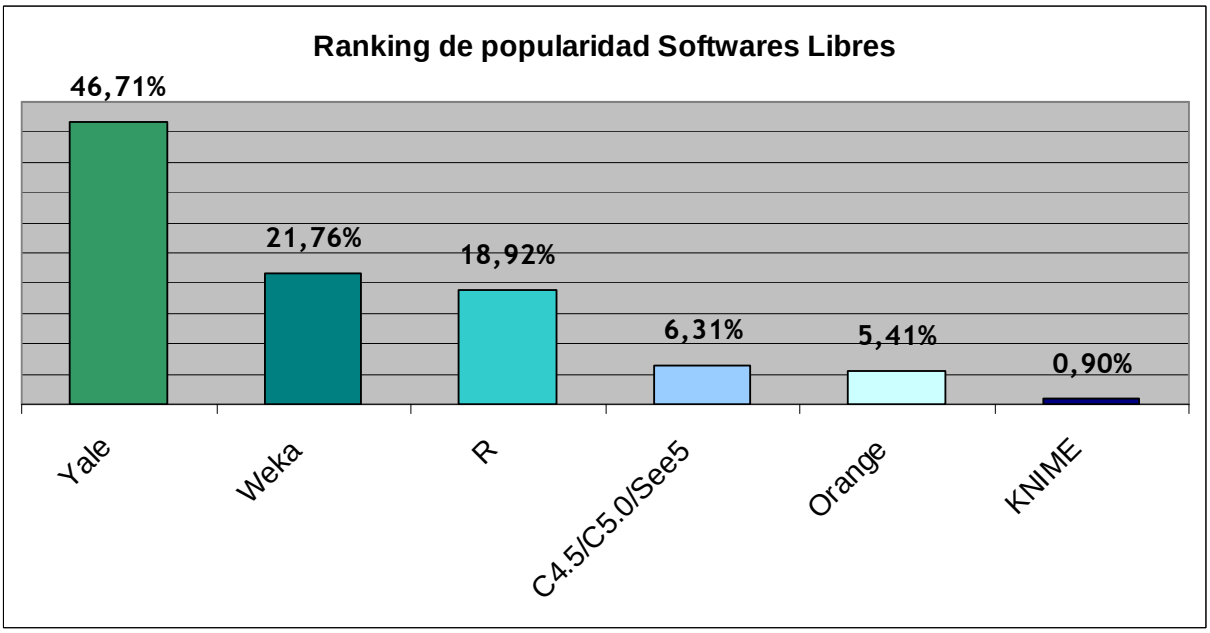


Gráfico 3: Resultados Encuesta sobre software Libres de Data Mining

Como se muestra en el Gráfico 2, el software comercial más popular es el SPSS Clementine con un 15% mientras que, tal como se ve en el gráfico 3, el software libre más popular según la encuesta es el Yale que obtuvo el 47% de los votos.

6.4.1.1 Funcionalidades del Software

A continuación se muestran dos tablas (Tabla 1 y Tabla 2)[6] que resumen las prestaciones ofrecidas por cada uno de los software incluídos en la encuesta de popularidad descripta en el punto 6.4.1., tanto comerciales como libres. En función del análisis deseado, se selecciona el software apropiado, teniendo en cuenta la posibilidad de necesitar otras funcionalidades en un futuro.

[6] The Data Mine: http://www.the-data-mine.com/bin/view/Software/DataMiningSoftware?sortcol=0;table=1;up=0#sorted_table

Software Comercial	Clasificación	Clusters	Regresión	Asociación	Text mining	Outliers	Visualización de Datos	Visualización de Patrones	Análisis de Redes Sociales	Análisis Web	Análisis de Redes Sociales
SPSS Clementine	x	x	x	x	x	x	x	x	x	x	x
Salford CART/MARS/TreeNet/RF	x	x	x	x	x	x	x	x	x	x	x
Excel	x	x	x	x		x	x	x			
SPSS	x	x	x	x	x	x	x	x	x	x	x
SAS	x	x	x	x		x	x			x	
Angoss	x	x	x	x			x	x	x	x	
KXEN	x	x	x	x	x	x	x	x	x	x	x
SQL Server	x	x	x	x		x	x				
MATLAB											
SAS E-Miner	x	x	x	x		x	x			x	
Statsoft Statistica	x	x	x	x	x	x	x	x	x	x	x
Insightful Miner/S-Plus	x	x	x	x		x	x	x	x		
Oracle DM	x	x	x	x	x	x	x	x	x	x	x
Tiberius	x					x	x	x			

Tabla 1: Funcionalidades de software Comerciales de Data Mining

Software Libre	Clasificación	Clusters	Regresión	Asociación	Text mining	Outliers	Visualización de Datos	Visualización de Patrones	Análisis de Redes Sociales	Análisis Web	Análisis de Redes Sociales
Yale	x	x		x		x	x	x			
Weka	x	x		x		x	x	x			
R	x	x	x	x	x	x	x	x	x	x	x
C4.5/C5.0/See5	x	x		x		x	x	x			
Orange	x	x	x	x		x	x	x			
KNIME	x	x		x		x	x	x			

Tabla 2: Funcionalidades de software Libres de Data Mining

6.4.2 Hardware

El hardware es la tecnología necesaria con la cual la empresa debe contar para poder utilizar los programas (software) de Data Mining. Es por eso que la capacidad de procesamiento necesaria de las computadoras dependerá del software elegido para la aplicación particular. Como se describió anteriormente, dicha elección se hará en función del objetivo buscado. Una vez seleccionado el programa, será necesario determinar el hardware. O de forma contraria, basándose en el equipamiento disponible, se optará entre aquellos programas que puedan funcionar con las computadoras disponibles en el momento.

En general, los proveedores de software de Data Mining, informan en su página Web cuales son los requerimientos de Hardware. En este proyecto, se optó por la utilización del programa Clementine.

Se detalla a continuación cuales son los requerimientos de hardware necesarios para poder operar con Clementine. Cabe aclarar que a su vez, el Sistema Operativo, también forma parte de los requerimientos y para soporte On Line se requiere un navegador de Internet.

Requerimientos:

Sistema Operativo:

- Microsoft Windows Vista® (Business and Enterprise) x32 or x64 Edition
- Microsoft Windows XP Professional® with Service Pack 2 x32 or x64 Edition
- Windows Server 2003® (Terminal Services)

Hardware:

- Procesador: Intel Pentium® o de clase similar o superior (para Windows 32-bit), familia de procesadores x64 (AMD 64 y EM64T, para Windows 64-bit)
- Monitor: Resolución 1024x768 o superior
- CD-ROM es necesario si la instalación se hace a partir de un CD
- Disco libre: 1 GB libre en Disco Duro
- Memoria RAM: 1 GB o superior, aunque se recomienda 2 GB o más
- Software: Microsoft Internet Explorer® 6.0 o superior para soporte online

6.4.3 Datos

Dado que los modelos predictivos se crean a partir de casos cuyo resultado ya es conocido, cuanto mayor sea el volumen de datos disponible, mayor la cantidad de

casos para estudiar. No existe una cantidad de registros determinada. No obstante, contar con datos acumulados durante al menos 2 años es un buen parámetro.

6.5 De los Datos al Modelo: El proceso del Data Mining

El proceso de Data Mining es un proceso estándar que se aplica en forma de pasos uno tras otro hasta alcanzar el objetivo. Comienza con los datos en su estado original y culmina con la creación del modelo.

Existen diversas metodologías propuestas para llevar a cabo un proceso de Data Mining, pero todas comparten la misma estructura y los mismos pasos a seguir. Se diferencian en la cantidad de componentes que consideran parte del proceso de Data Mining

El orden siempre debe respetarse si se quiere lograr un proceso ordenado y sin falencias aunque en ocasiones algunos pasos no son necesarios. Esto depende de la forma, calidad y naturaleza de los datos de que se dispone. La selección de variables y la creación del modelo siempre están presentes en un proyecto de Data Mining.

Los pasos se muestran en la ilustración 6 a continuación y se detallan en los capítulos siguientes:

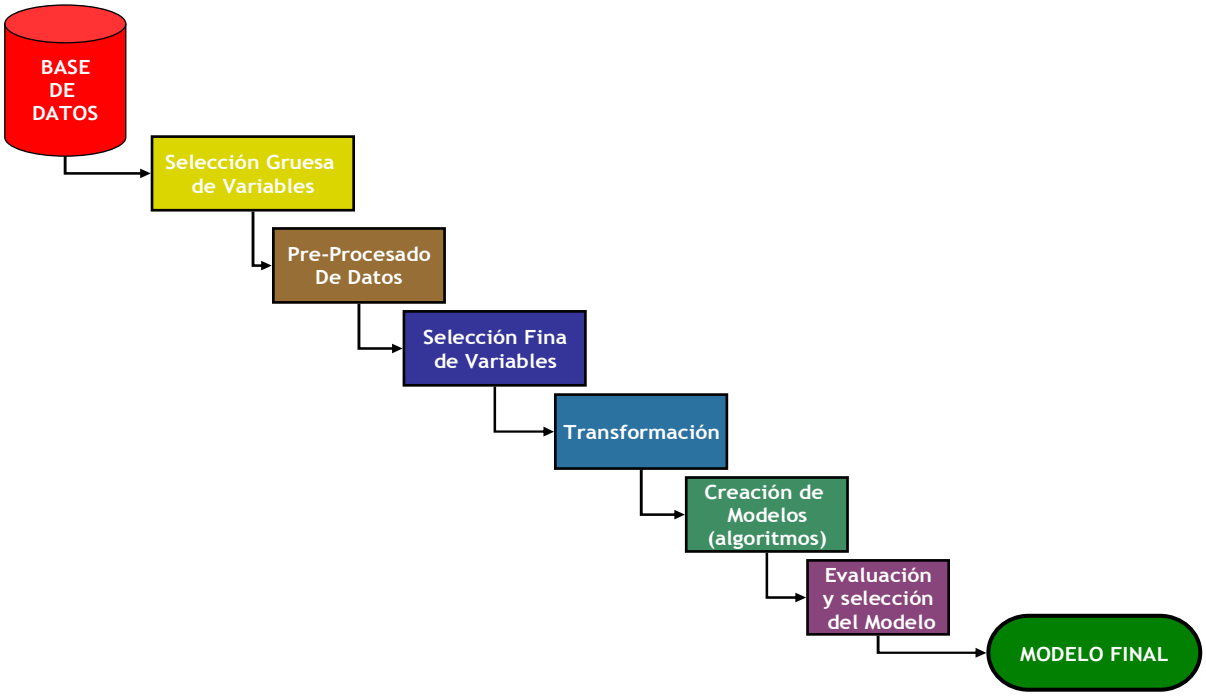


Ilustración 6: Proceso del Data Mining

6.5.1 Selección gruesa de Variables

Lo primero que se tiene al comienzo de un proyecto de Data Mining es una base de datos con una cantidad de variables desconocida dispersas en tablas que tampoco se conocen. La selección “gruesa” de variables consiste en determinar que tipo de variables se necesita recolectar en función del problema planteado.

Es necesario ser cuidadoso en esta primera selección de modo de no dejar afuera ninguna variable que pudiera ser de importancia para la creación del modelo. El resultado (output) de esta tarea debe ser un “set” de variables de número igual o inferior al original.

6.5.2 Pre-procesado de Datos

Cuanto mayor tamaño tiene una base de datos, mayores son las probabilidades de que existan errores en el registro de los mismos: “Ruidos”, valores faltantes o valores “atípicos" son situaciones comunes que se encuentran en las bases, dado su gran tamaño y las diferentes fuentes que impactan registros en ellas.

Es posible encontrarse con casilleros en blanco o valores que no se corresponden con el sentido común, como por ejemplo, edades negativas, ausencia del nombre, números de documento que incluyen letras, etc. Los datos en esa forma no son aptos para ser procesados.

En un proyecto que involucra Data Mining, la preparación de los datos es el conjunto de pasos que mas tiempo abarca (ver ilustración 7). En general insume el 60% del tiempo total del proyecto.

Gráfico 4: Distribución del tiempo en etapas del Data Mining

Para realizar el pre-procesado de Datos, existen distintas técnicas que se aplican en forma automática o manual*: Limpieza, integración y normalización.*

La limpieza de datos se utiliza para remover el ruido, reemplazar valores faltantes y corregir inconsistencias en los datos. En la integración se reúnen datos provenientes de distintas bases y se los almacena en único lugar (almacén de datos). Por último, la normalización de los datos consiste en reducir la escala de valores de una variable manteniendo la relación de magnitudes. Estas técnicas no son excluyentes entre sí sino que se complementan.

Se explican a continuación, en mayor detalle, las técnicas mencionadas.

6.5.2.1 Limpieza de Datos

Ciertos registros suelen estar vacíos (valores faltantes), contener “ruido” (error aleatorio de una variable o valores muy alejados de lo esperado), y muchas veces ser inconsistentes (datos mal cargados o registrados). Para hacer Data Mining, es necesario que las bases estén libres de estos errores, dado que la calidad final del modelo esta directamente relacionada con la calidad de los datos utilizados; “garbage in, garbage out”.

Existen varias técnicas para resolver estos inconvenientes.

6.5.2.1.1 Valores Faltantes

Los valores faltantes (no definidos) se tratan como valores no válidos. Para solucionar este tipo de errores se recurre a las técnicas que se muestran en la tabla a continuación:

Tipo de Error	Tecnica
Valores Faltantes	Ignorar el registro completo
	Completar con una constante por atributo
	Completar con la media del atributo
	Completar con el valor mas probable (predicción mediante regresión)

Tabla 3: Técnicas para corregir valores faltantes

La técnica de “completar con el valor mas probable”, es la más popular. En comparación con otros métodos, es la que utiliza la mayor información disponible para predecir los valores no definidos. Al considerar los valores de los demás atributos para estimar, por ejemplo, el “consumo anual con TC”, existe una mayor probabilidad de que se mantenga la relación entre ambas.

De las tres restantes, la más recomendada es la técnica de completar con la media del atributo aunque en ocasiones se decidirá ignorar el registro completo.

6.5.2.1.2 Ruido y outliers

En cualquier tipo de comunicación, el ruido es algo que hay que evitar, ya que ensucia el mensaje que se está transmitiendo. Cuando nos referimos a un modelo de predicción, podemos asumir que las variables independientes transmiten información sobre la variable dependiente. Dicho de otro modo, existen variables que comunican información que será útil para predecir la variable de interés.

Ahora bien, no siempre los modelos alcanzan un alto grado de precisión en sus predicciones debido a distintas razones. Por ejemplo, podría ser que las relaciones entre las variables independientes y la dependiente no sean lineales y la herramienta usada para construir el modelo sólo tenga en cuenta relaciones lineales.
Otra razón tiene que ver con el ruido en los datos. Se suele decir que los datos contienen ruido y esto interfiere con la creación de un buen modelo.
Se puede definir el ruido como una señal aleatoria que se superpone a la señal original y confunde al destinatario.

A continuación se muestra en un ejemplo primero una señal sin ruido y luego con ruido superpuesto.

Se tiene la siguiente relación entre dos variables:

$$y = x^2 - x + 2$$

Una tabla de datos que contenga esta relación tendrá dos variables X e Y. La variable X será la variable independiente e Y la dependiente. Se puede asumir que X transmite información acerca de Y y un buen modelo será capaz de usar la información de X para estimar Y.

Si se grafica esta relación puede verse que para cada valor de X existe solamente un valor de Y posible (ver gráfico 5):

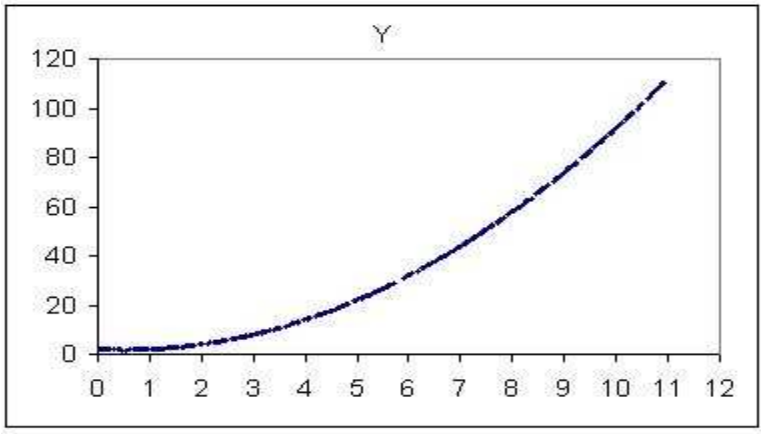


Gráfico 5: Función $y = x^2$

Los datos podrán contener varios valores de X iguales, por ejemplo, varias filas en donde X es 5, pero en cada una de estas filas, la variable Y tendrá el mismo valor: 22 (que es el resultado de la ecuación dada más arriba).

Supóngase ahora que los datos tienen la siguiente particularidad: para un mismo valor de X pueden existir varios valores de Y. O sea, ahora si existen varias filas en donde X es 5, no necesariamente el valor de Y será en todas ellas 22. Podría ser que en algunas sea 20, en otras 22 y en otras 30.

El gráfico de estos nuevos datos se verá así:

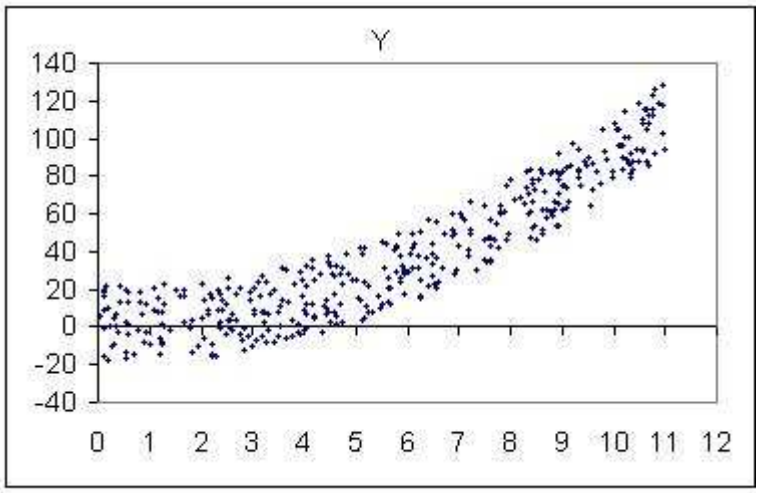


Gráfico 6: Función $y = x^2$ + ruido

Estos nuevos datos contienen ruido porque para una misma señal (en el ejemplo la señal es el valor de X) existen distintos valores que puede tomar la variable a predecir.
Un modelo construido con estos datos nunca será lo suficientemente preciso debido a que parte de la información contiene ruido. Es importante notar que no importa el tipo de herramienta usada para construir el modelo. Si los datos contienen ruido, el modelo no será perfecto, y el grado de precisión dependerá justamente del nivel de ruido.

La Tabla 4 a continuación resume las técnicas disponibles para evitar estos errores:

Tipo de Error	Tecnica
Ruido	Desviación típica de la media
	Clustering para identificar valores demasiado alejados (outliers)
	Formar grupos de valores y reasignarles un único valor por grupo

Tabla 4: Técnicas para corregir ruido

La técnica de la desviación típica de la media consiste en detectar los valores atípicos y extremos a partir del número de desviaciones típicas de la media. Por ejemplo, si tiene un campo con una media de 100 y una desviación típica de 10, se podría especificar 3,0 para indicar que cualquier valor inferior a 70 o superior a 130 debe tratarse como atípico.

Por otro lado, el análisis de clustering identifica los grupos homogéneos, y aquellos valores que no sean asignados a ningún grupo serán considerados outliers y eliminados. También se pueden tomar conjuntos de valores relativamente cercanos entre si y asignarles un único valor, de modo que se reduzca el ruido. Los valores eliminados pueden luego tratarse con las técnicas antes descriptas para valores faltantes.

6.5.2.1.3 Inconsistencias

Este tipo de errores es el que más tiempo lleva detectar y corregir dado que no existen formulas matemáticas que indiquen cuando un dato fue mal cargado.

La tabla 5 muestra las técnicas para eliminar las inconsistencias:

Tipo de Error	Tecnica
Inconsistencias	Detección y corrección manual
	Detección automática, corrección manual
	Rutinas de detección y corrección automáticas

Tabla 5: Técnicas para corregir inconsistencias

La detección y corrección manual implica realizar un “rastrillaje” de los registros observando los campos uno por uno para determinar si existe alguna inconsistencia en los valores. Por ejemplo: Nivel socio económico “R”, o edad -8, etc.

No obstante, pueden realizarse búsquedas automáticas donde por ejemplo, se pida detectar letras en campo numéricos, o niveles socioeconómicos distintos de A, B, C, D o E. Luego, o se borra el registro, o se consigue el valor real del cliente y se completa dicho campo.

El mejor caso es cuando se puede tanto detectar como corregir los valores en forma automática. Para el atributo “edad”, podría programarse una rutina donde se detecten edades negativas y se transformen en positivas, por ejemplo, si la edad es -42, la transforma en 42 asumiendo que el error esta en el signo.

6.5.2.2 Integración de Datos

Los datos de clientes generalmente se encuentran registrados en diferentes tablas cada una con diferentes tipos de información. Para poder utilizar toda esta información con Data Mining, es necesario juntarla en una única tabla. La Integración del esquema requiere responder la pregunta: ¿Como se relacionan las distintas fuentes de datos? Se explicará mediante un ejemplo:

Se quiere crear una tabla con el DNI como campo identificatorio de cada cliente.

De la tabla 1 (ver ilustración 8) puede asociarse directamente la cantidad de Plazos Fijos vigentes del cliente.

La tabla 2 (ver ilustración 8) contiene un dato de interés pero esta asociado al Id_cliente (identificación interna del Banco). Para poder llevar el dato de NSE de la tabla 2 a la tabla 1 es necesario encontrar la tabla que relaciona, a su vez, la variable DNI con la variable Id_Cliente. Esta tabla se muestra en la ilustración 8 como “Tabla Nexo”. La relación debe ser única.

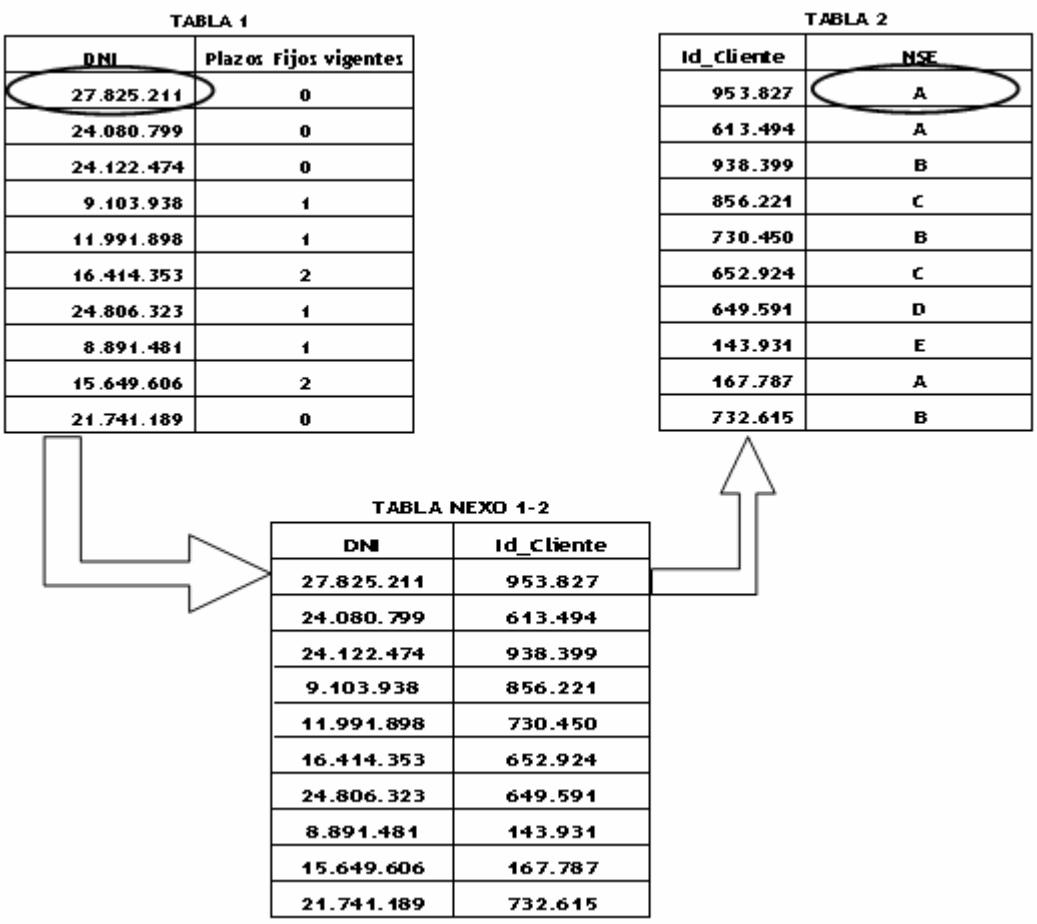


Ilustración 7: Ejemplo de integración de tablas

Como resultado se logra construir la tabla 3 (Ilustración 9):

TABLA 3

DNI	Plazos Fijos vigentes	NSE
27.825.211	0	A
24.080.799	0	A
24.122.474	0	B
9.103.938	1	C
11.991.898	1	B
16.414.353	2	C
24.806.323	1	D
8.891.481	1	E
15.649.606	2	A
21.741.189	0	B

Ilustración 8: Tabla integrada

La *Redundancia* es un aspecto a tener en cuenta cuando se Integran los datos. Evitar la redundancia implica evitar agregar 2 variables con alta correlación entre sí, en cuyo caso es recomendable eliminar una de las dos.

6.5.2.3 Normalización de Datos

Un atributo es normalizado transformando sus valores de modo que caigan dentro de un rango específico más pequeño, por ejemplo, entre 0 y 1. Esta técnica es de particular importancia cuando se utilizan algoritmos tipo redes neuronales o SVM. Estos métodos proveen mejores resultados si los datos fueron normalizados. En clasificación, la normalización de los valores de entrada hará más rápida la fase de aprendizaje y su principal ventaja es la de prevenir que los atributos que típicamente contienen valores mas grandes, dominen aquellos con valores menores.

Existen varios métodos para la normalización:

- Min-max: Dividir todos los valores por el valor absoluto más grande.
- Media cero: Hacer que los datos representen una normal de Media cero y Desvío 1
- Decimal: Correr el punto decimal mediante la división por 10,100,1000, etc

6.5.3 Selección Fina de Variables

Una vez realizada la selección “gruesa” de variables, mediante la cual se determina el mayor conjunto de datos requeridos para comenzar el análisis y luego aplicando a dicho conjunto las técnicas de pre-procesado (limpieza de datos, integración y normalización), se realiza la selección “fina” de variables.

El primer paso consiste en el prelavado. No hay que confundir con la limpieza de datos. En este paso se analizan los valores registrados para cada atributo y se ponen condiciones mínimas para asegurarse que las variables explicativas sean estadísticamente viables. Por ejemplo, si la variable “nacionalidad” en el 99% de los casos es “Argentino”, realmente no hace ningún aporte de información. Dicha variable queda entonces eliminada.

Luego del prelavado, se realiza la selección propiamente dicha. Esta selección no se hace en función de las variables que el analista elige en forma subjetiva (selección gruesa) sino que se hace en función de análisis objetivos que prueban la importancia de unas y otras en términos estadísticos. Dado que algunas de estas variables independientes resultan importantes para explicar la variable objetivo mientras que otras prueban ser indiferentes y se eliminan, el resultado de este paso será un “set” de variables menor que el original.

A continuación se explican en más detalle ambos pasos de la selección fina de variables: prelavado y métodos de selección.

6.5.3.1 Prelavado

Las condiciones impuestas como filtro en el prelavado dependen de cada caso aunque en general son similares.

Atributos que presentan demasiados valores perdidos (faltantes) o con escasa variabilidad son filtrados en este paso. Sirve como control en caso de no haberse realizada una adecuada limpieza de datos y a su vez agrega otros criterios para asegurar que el comportamiento de la variable sea estadísticamente viable. Como se mencionó antes, si una variable adquiere en el 99% de los casos el mismo valor, no agregará información útil a la predicción (variables categóricas). O, en el caso de variables numéricas, se determina el desvío estándar mínimo requerido (variabilidad mínima).

6.5.3.2 Métodos de Selección de Variables

Existen dos componentes importantes que distinguen los métodos de selección de variables: creación y evaluación.

Algunos métodos utilizan una medida para evaluar la importancia de cada variable y las variables se ordenan en función de dicho valor. Luego se eligen las primeras X cantidad de variables. Algunos programas asignan, además del número que mide la importancia, una etiqueta en función de dicho número que indica si la variable es: “importante”, “poco importante”, o “no importante”.

Por otro lado, existen ciertos métodos que determinan la importancia de un grupo de variables y no de cada variable individual. Así una variable que evaluada en forma individual pudo haber sido eliminada puede, en este segundo caso, pertenecer a un grupo que resultó adecuado y por ende ser considerada para la creación del modelo.

Más allá de poder analizar las variables en forma individual o su comportamiento grupal, los métodos de selección se dividen en dos grupos, según la naturaleza de las mediciones realizadas para evaluar a las variables, ya sea una a una o de a grupos. Se los llama modelos tipo “filtro” y modelos tipo “envuelto”.

A continuación se describen ambos tipos de métodos.

6.5.3.2.1 Métodos “Filtro”

El método tipo filtro, analiza el set de variables sin tener en cuenta su performance en el paso posterior de creación del modelo. No tiene en cuenta el algoritmo inductivo que se alimentará con ellas, así como tampoco la precisión final del modelo obtenido en ese paso.

Las elige según el poder de información contenido en ellas analizando su condición inherente para explicar la variable objetivo.

Existen 4 métodos de selección tipo filtro: Separación Bi-normal, Coeficientes de Correlación, F-Score, Algoritmo basado en entropía. Estos métodos requieren de menor tiempo de procesamiento que los de tipo envuelto y pueden, por ende, ser aplicados a grandes bases de datos que contienen un gran número de variables y registros.

No obstante la eficiencia operacional, la gran desventaja de los análisis tipo filtro es que una selección óptima de variables no es independiente de la calidad de la inducción realizada por el algoritmo en un paso posterior.

Dado que omite la performance de las variables de entrada al modelo, existen en forma alternativa métodos que sí consideran este factor. Son los modelos tipo envuelto que se describen a continuación.

6.5.3.2.2 Métodos tipo “Envuelto”

Los modelos tipo “envuelto” eligen las variables en función del resultado obtenido luego de probar la precisión del modelo obtenido mediante el algoritmo de generación. Por eso se dice que funciona como una caja negra donde las variables interactúan con el algoritmo, creando así distintos modelos según las variables consideradas. Midiendo la precisión de estos modelos se determina el grupo de variables que mejor funcionó para la predicción. Ese es el grupo finalmente seleccionado.

Teniendo en cuenta que es necesario probar la efectividad de cada grupo de variables combinada con cada algoritmo inductivo de prueba (en caso de querer probar con más de uno), este tipo de métodos demanda considerablemente mayor tiempo que los métodos tipo filtro.

6.5.3.2.3 “Filtro” vs. “Envuelto”

Las variables seleccionadas mediante el método tipo envuelto prueban dar origen, algoritmo de inducción mediante, a modelos más precisos.

Como se explicó anteriormente, esto se debe a que justamente esa es la medida que determina el mejor grupo de variables de entrada, maximizando la precisión del modelo.

Sin embargo, cuando el número de variables se vuelve grande, no es factible realizar este tipo de análisis sobre las variables de entrada. En estos casos, la selección debe hacerse mediante métodos tipo filtro. Son más rápidos y su precisión es igualmente aceptable.

6.5.3.2.4 Función de Evaluación de Variables

Para evaluar cada variable, los métodos tipo filtro utilizan una medida determinada. Las variables se ordenan en función de dicha medición y se eligen por último las primeras X. La naturaleza de dicho indicador puede ser de información, de distancia o de dependencia.

Existen 4 tipos de funciones de evaluación para los métodos filtro:

- Distancia (medida euclidiana de distancia)
- Información (Entropía, ganancia de información, etc)
- Dependencia (Coeficientes de correlación)
- Consistencia (sesgo mínimo de características)

Lo que caracteriza entonces a los métodos tipo filtro es, si evalúan variables en forma individual o grupal y en segundo lugar la función de evaluación elegida para medir las variables.

La metodología tipo envuelto, utiliza la precisión de la predicción final para evaluar la utilidad relativa de las variables de entrada del algoritmo.

Sin embargo, en la práctica, es necesaria una estrategia de búsqueda de grupos de prueba variables que sea robusta y eficiente.

Estrategias de búsqueda como selección “hacia delante” o “eliminación hacia atrás” permiten seleccionar grupos de prueba reduciendo el tiempo que llevaría probar todas las combinaciones posible de variables.

6.5.4 Creando el Modelo: Algoritmos de generación

Como resultado de la selección fina de variables se obtiene el set de variables que proveerá la información necesaria para entender el fenómeno que se quiere explicar o predecir. El próximo paso es la creación del modelo a partir de dichas variables de entrada.

El modelo es justamente una secuencia matemática de parámetros que una vez generado, se alimenta con los valores de las variables de entrada para cada individuo para predecir el valor de la variable objetivo (¿el cliente aceptará el producto?, ¿Qué enfermedad tiene el paciente?)

A continuación se describen los algoritmos mas importantes y frecuentes usados para la creación de Modelos Predictivos.

6.5.4.1 Árboles de decisión

Este tipo de algoritmos examina todos los campos de la base de datos para detectar el que proporciona la mejor clasificación o pronóstico dividiendo los datos en subgrupos. El proceso se aplica de forma recursiva, dividiendo los subgrupos en unidades cada vez más pequeñas hasta completar el árbol (según se definan determinados criterios de parada). Los campos objetivo y de entrada utilizados en la generación del árbol pueden ser intervalos numéricos o categóricos, dependiendo del algoritmo que se utilice. Si se usa un objetivo de rango, se genera un árbol de regresión; si se usa un objetivo categórico, se genera un árbol de clasificación.

6.5.4.1.1 C&R

El algoritmo de árbol de clasificación y regresión (C&R) genera un árbol de decisión que permite pronosticar o clasificar observaciones futuras. El método utiliza la partición reiterada para dividir los registros de entrenamiento en segmentos minimizando las impurezas en cada paso, donde un nodo se considera “puro” si el 100% de los casos del nodo corresponden a una categoría específica del campo objetivo que se quiere predecir. Los campos objetivo y predictor pueden ser de rango o categóricos. Todas las divisiones son binarias (sólo se crean dos subgrupos).

Los usos generales del análisis basado en árboles son:

Segmentación. Identifica personas con probabilidad de pertenecer a una determinada categoría.

Estratificación. Asigna casos en una o varias categorías, como grupos de alto, medio y bajo riesgo.

Reducción de datos y filtrado de variables. Selecciona un subconjunto útil de predictores de un gran conjunto de variables para usarlo en la creación de un modelo.

6.5.4.1.2 CHAID

El algoritmo CHAID genera árboles de decisión utilizando estadísticos de chi-cuadrado para identificar las divisiones óptimas. A diferencia de los algoritmos de C&R, CHAID puede generar árboles no binarios, lo que significa que algunas divisiones generarán más de dos ramas. Los campos objetivo y predictor pueden ser de rango o categóricos.

6.5.4.1.3 C5.0

El algoritmo C5.0 genera un árbol de decisión o un conjunto de reglas. El modelo divide la muestra basándose en el campo que ofrece la máxima ganancia de información en cada nivel. El campo objetivo debe ser categórico. Se permiten varias divisiones en más de dos subgrupos.

6.5.4.2 Redes

6.5.4.2.1 Redes Bayesianas

El nodo Red bayesiana permite crear un modelo de probabilidad combinando pruebas observadas y registradas con conocimiento del mundo real de "sentido común" para establecer la probabilidad de instancias utilizando atributos aparentemente no vinculados.

Las redes bayesianas se utilizan para realizar pronósticos en diferentes situaciones; algunos ejemplos son los siguientes:

- Selección de oportunidades de crédito con poco riesgo de fracaso.

- Estimación cuando se necesite reparar el equipo o piezas de recambio, en función de los datos de los sensores y los registros existentes.

• Solución de problemas de los clientes mediante herramientas de solución de problemas en línea.

• Diagnóstico y solución de problemas de redes de telefonía móvil en tiempo real.

• Evaluación de los riesgos potenciales y recompensas de proyectos de investigación y desarrollo para centrar los recursos en las mejores oportunidades.

Una red bayesiana es un modelo gráfico que muestra variables (que se suelen denominar nodos) en un conjunto de datos y las independencias probabilísticas o condicionales entre ellas. Las relaciones causales entre los nodos se pueden representar por una red bayesiana; sin embargo, los enlaces en la red (también denominados arcos) no representan necesariamente una relación directa de causa y efecto. Por ejemplo, una red bayesiana se puede utilizar para calcular la probabilidad de un paciente con una enfermedad concreta, con la presencia o no de algunos síntomas y otros datos relevantes, si las independencias probabilísticas entre síntomas y enfermedad son verdaderas, tal y como se muestra en el gráfico. Las redes son muy robustas en los puntos en los que falta información y realizan los mejores pronósticos posibles utilizando la información disponible.

Lauritzen y Spiegelhalter crearon un ejemplo común y básico de una red bayesiana en 1988. También se conoce como modelo "Asia" y es una versión simplificada de una red que se puede utilizar para diagnosticar a los nuevos pacientes de un médico; la dirección de los enlaces corresponde por lo general a la causalidad. Cada nodo representa una faceta que se puede relacionar con el estado de un paciente; por ejemplo, "fumador" indica que se trata de un fumador habitual y "VisitaAsia" muestra que recientemente ha viajado a Asia. Los enlaces entre los nodos muestran las relaciones probabilísticas; por ejemplo, fumar aumenta las posibilidades de que el paciente padezca bronquitis y cáncer de pulmón, mientras que la edad parece estar relacionada únicamente con la posibilidad de desarrollar cáncer de pulmón. De la misma forma, las anomalías detectadas en una radiografía de los pulmones pueden estar causadas por tuberculosis o cáncer, mientras que las posibilidades de que un paciente tenga dificultades respiratorias (disnea) aumentan si también padece bronquitis o cáncer de pulmón.

6.5.4.2.2 Redes Neuronales

Las redes neuronales son modelos simples del funcionamiento del sistema nervioso. Las unidades básicas son las neuronas, que generalmente se organizan en capas, como se muestra en la siguiente ilustración:

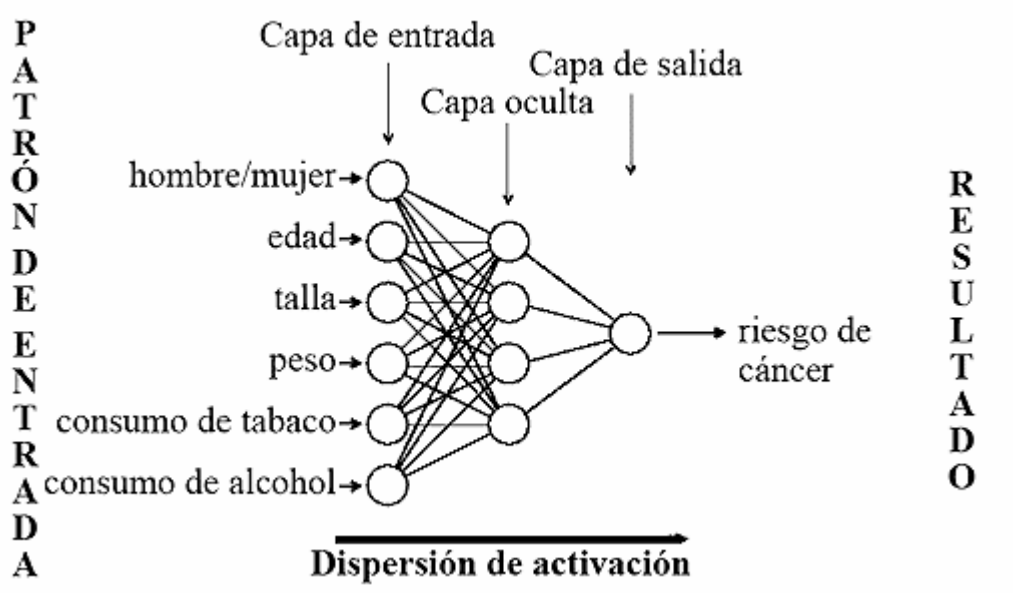


Ilustración 9: Estructura de una red neuronal

Una red neuronal, a menudo denominada perceptrón multicapa, es básicamente un modelo simplificado del modo en que el cerebro humano procesa la información. Funciona combinando en forma simultánea un número elevado de unidades simples de procesamiento interconectadas que parecen versiones abstractas de neuronas.

Las unidades de procesamiento se organizan en capas. Existen, generalmente, tres capas en una red neuronal: una capa de entrada, con unidades que representan los campos de entrada; una o varias capas ocultas; y una capa de salida, con unidades que representan los campos de salida. Las unidades se conectan con fuerzas de conexión variables (o ponderaciones). Los datos de entrada se presentan en la primera capa, y los valores se propagan desde cada neurona hasta cada neurona de la capa siguiente. Al final, se envía un resultado desde la capa de salida.

La red aprende examinando los registros individuales, generando un pronóstico para cada registro y realizando ajustes a las ponderaciones cuando realiza un pronóstico incorrecto. Este proceso se repite muchas veces y la red sigue mejorando sus pronósticos hasta haber alcanzado uno o varios criterios de parada.

Al principio, todas las ponderaciones son aleatorias y las respuestas que resultan de la red son, posiblemente, disparatadas. La red aprende a través del entrenamiento. Continuamente se presentan a la red ejemplos para los que se conoce el resultado, y las respuestas que proporciona se comparan con los resultados conocidos. La información procedente de esta comparación se pasa hacia atrás a través de la red, cambiando las ponderaciones gradualmente. A medida que progresa el entrenamiento, la red se va haciendo cada vez más precisa en la replicación de resultados conocidos. Una vez entrenada, la red se puede aplicar a casos futuros en los que se desconoce el resultado.

6.5.4.3 Regresión

6.5.4.3.1 Regresión Logística

La regresión logística, también denominada regresión nominal, es una técnica estadística para clasificar los registros a partir de los valores de los campos de entrada. Es análoga a la regresión lineal pero utiliza un campo objetivo categórico en lugar de uno numérico. Se admiten tanto los modelos binomiales (para objetivos con dos categorías discretas) como los multinomiales (para objetivos con más de dos categorías).

La regresión logística trabaja creando un conjunto de ecuaciones que relacionan los valores de los campos de entrada con las probabilidades asociadas a cada una de las categorías de los campos de salida. Una vez se ha generado el modelo, se puede utilizar para calcular las probabilidades de datos nuevos. Para cada registro, se calcula una probabilidad de pertenencia a cada categoría posible de salida. La categoría objetivo con la probabilidad más alta se asigna como el valor de salida pronosticado para cada registro.

Ejemplo: Un proveedor de servicios de telecomunicaciones está preocupado por el número de clientes que se están pasando a la competencia. Mediante los datos de uso de servicio puede crear un modelo binomial para pronosticar qué clientes tienen más probabilidad de contratar otro proveedor y personalizar las ofertas para retener el mayor número de clientes posible. Se utiliza un modelo binomial porque el objetivo tiene dos categorías diferentes (probabilidad de pasar a la competencia o no).

6.5.4.4 SVM (Support Vector Machines)

Este algoritmo, esta basado en fundamentos matemáticos relacionados con las distancias en el plano. La idea es ubicar a los distintos registros, cada uno de n dimensiones, como dos grupos de puntos separados en el hiper plano.

Gráficamente, solo es posible visualizarlo en el caso de n<=3.

Sin embargo, en términos matemáticos, el procedimiento es aplicable a registros de cualquier cantidad de variables.

El SVM construye, una vez alimentado con todos los registros (puntos n dimensionales), un hiperplano que separa el conjunto total en 2 sub grupos, de modo de maximizar la distancia entre ambos. La separación viene dada por la distancia mas chica entre un punto de un grupo y un punto del otro grupo. Cuanto mayor sea esa distancia, se supone que mejor caracterizado estará cada grupo y la clasificación de los casos no conocidos será mas precisa.

Se muestra en el gráfico 7 a continuación un caso de registros con 2 variables:

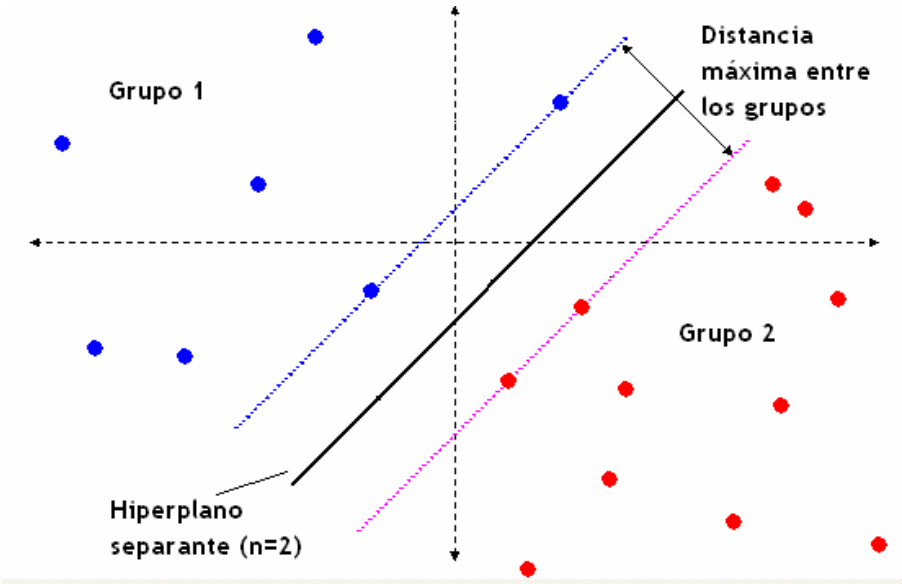


Gráfico 7: Registros clasificados mediante SVM

6.5.5 EVALUACION DE LOS MODELOS GENERADOS

Dado que se generan varios modelos a partir de los distintos algoritmos de generación, es necesario evaluarlos para determinar cual de todos es el que mejor funciona para la predicción.

Esto se hace “alimentándolos” con casos de resultado conocido y comparando la realidad contra lo predicho por el modelo.

Para conocer y comparar la calidad en la predicción de cada modelo, se utilizan 3 herramientas: Gráficos de Ganancia, Índice KS y tablas de Hosmer-Lemeshow.

6.5.5.1 Gráficos de Ganancia

El procedimiento para construir el gráfico de ganancia es el siguiente:

1. Una vez logrados los modelos, se selecciona un grupo de clientes cuyo resultado (variable objetivo) ya sea conocido. Se quiere comparar lo que dice el modelo contra lo que realmente ocurrió cuando se contacto al cliente en su momento (si aceptó o rechazó la oferta).

2. Se alimentan los modelos con los registros de prueba, manteniendo la variable objetivo (de resultado ya conocido) como incógnita.

3. El modelo asigna la probabilidad de éxito en la venta para cada cliente.

4. Se ordenan los clientes en función de dicha probabilidad, de mayor a menor.

5. En una columna contigua se indican con 1 o 0 si los casos fueron positivos o negativos en la realidad.

6. En la columna siguiente se representa el Acumulado de casos positivos real

7. Para finalizar, en las 2 últimas columnas, se obtienen los porcentajes de casos positivos en función del porcentaje contactado.

La tabla resultante es la siguiente.

Nº cliente	**Probabilidad**	Resultado	Acumulado	% de contacto	% de positivos
1	0,84	1	1	3%	8%
2	0,82	1	2	7%	17%
3	0,81	1	3	10%	25%
4	0,8	1	4	13%	33%
5	0,79	1	5	17%	42%
6	0,78	1	6	20%	50%
7	0,75	1	7	23%	58%
8	0,75	1	8	27%	67%
9	0,75	0	8	30%	67%
10	0,69	0	8	33%	67%
11	0,66	1	9	37%	75%
12	0,64	1	10	40%	83%
13	0,61	0	10	43%	83%
14	0,61	0	10	47%	83%
15	0,59	0	10	50%	83%
16	0,55	0	10	53%	83%
17	0,54	0	10	57%	83%
18	0,51	0	10	60%	83%
19	0,48	1	11	63%	92%
20	0,46	0	11	67%	92%
21	0,42	0	11	70%	92%
22	0,42	0	11	73%	92%
23	0,42	0	11	77%	92%
24	0,4	1	12	80%	100%
25	0,38	0	12	83%	100%
26	0,34	0	12	87%	100%
27	0,34	0	12	90%	100%
28	0,3	0	12	93%	100%
29	0,23	0	12	97%	100%
30	0,21	0	12	100%	100%

Tabla 6: Ejemplo de Tabla para construir un gráfico de ganancia

Descripción de los campos:

a. **Nº Cliente**: Identificación del cliente. A su vez indica cantidad acumulada de contactos.

b. **Probabilidad:** Asignada por el modelo y permite ordenar los clientes según su propensión a comprar el producto (Plazo Fijo)

c. **Resultado:** Valor real conocido de la variable objetivo. Es 1 si aceptó la oferta, 0 si la rechazó. (Un buen modelo debería acumular los 1 en los primeros contactos y los ceros al final)

d. **Acumulado:** Es el acumulado de casos positivos

e. **% de Contacto**: Nº cliente / Total de clientes (30)

f. **% Positivos**: Acumulado / Total de positivos (12)

El gráfico del ejemplo se ve así:

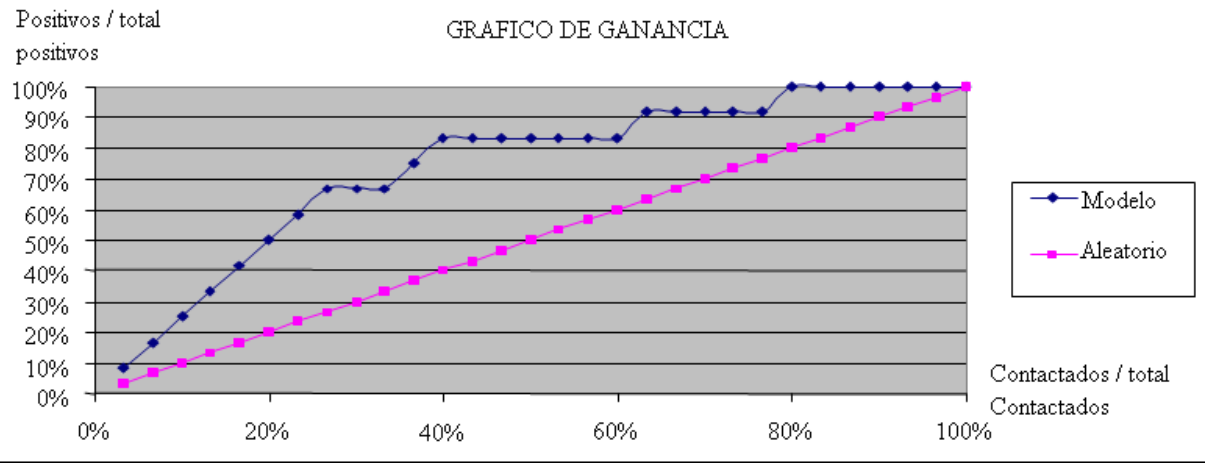


Gráfico 8: Gráfico de ganancia para tabla 6

Mientras que una selección aleatoria distribuye en forma homogénea a los aceptantes a medida que se realizan los contactos, el modelo acumula los casos más favorables en los primeros lugares.

Si se graficaran varios modelos a la vez, y para un mayor número de casos, el gráfico obtenido se vería como el siguiente:

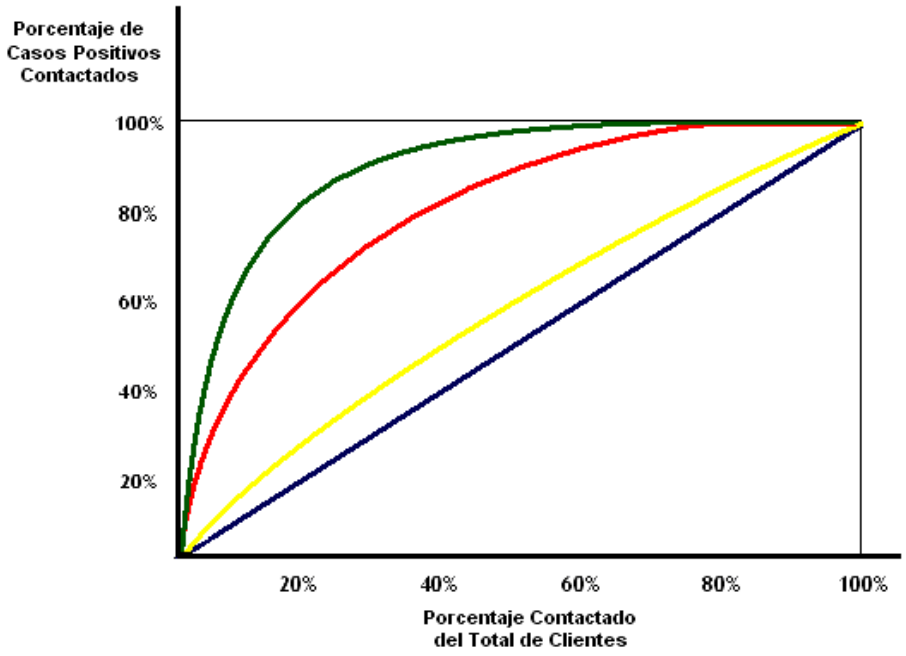


Gráfico 9: Gráficos de ganancia superpuestos para 3 modelos distintos

El grafico verde corresponde al modelo de mejor “performance”. Como se ve, contactando al 20% de la población, logra capturar el 80% del total de casos positivos.

Si en cambio no se utilizara un modelo y la selección fuese aleatoria, la curva sería la azul donde se reparten homogéneamente los casos positivos entre el total de la población. A medida que aumentan los clientes contactados, aumenta en igual proporción los contactos positivos.

El segundo mejor modelo es el rojo y en tercer lugar el amarillo.

Los gráficos de ganancia permiten ver cual es el valor agregado de los modelos.

6.5.5.2 Índice KS

Un método complementario al gráfico de ganancia es el índice KS. El gráfico muestra la separación porcentual entre la población que acepta la oferta (“Goods”) y

la que rechaza la oferta (“Bad”) ordenados por la probabilidad de aceptación asignada por el modelo. El valor máximo de separación, conocido como KS, indica la calidad del modelo.

El gráfico se construye de igual forma que el gráfico de Ganancia, pero graficando a su vez el porcentaje de los casos negativos totales contenidos en cada porción de la población.

A continuación se calcula el KS para el caso anterior (ver tabla 7) y se muestra el gráfico resultante (ver gráfico 10):

Nº cliente	Probabilidad (modelo)	Resultado Positivo (real)	Acumulado Positivos (acumulado real)	% de positivos	Resultado Negativo (real)	Acumulado Negativos (acumulado real)	% de negativos	% de contactados (del total de la muestra)	Indice KS (% pos - % neg)	Aleatorio
1	0,84	1	1	8%	0	0	0%	3%	8%	3%
2	0,82	1	2	17%	0	0	0%	7%	17%	7%
3	0,81	1	3	25%	0	0	0%	10%	25%	10%
4	0,8	1	4	33%	0	0	0%	13%	33%	13%
5	0,79	1	5	42%	0	0	0%	17%	42%	17%
6	0,78	1	6	50%	0	0	0%	20%	50%	20%
7	0,75	1	7	58%	0	0	0%	23%	58%	23%
8	0,75	1	8	67%	0	0	0%	27%	67%	27%
9	0,75	0	8	67%	1	1	6%	30%	61%	30%
10	0,69	0	8	67%	1	2	11%	33%	56%	33%
11	0,66	1	9	75%	0	2	11%	37%	64%	37%
12	0,64	1	10	83%	0	2	11%	40%	**72%**	40%
13	0,61	0	10	83%	1	3	17%	43%	67%	43%
14	0,61	0	10	83%	1	4	22%	47%	61%	47%
15	0,59	0	10	83%	1	5	28%	50%	56%	50%
16	0,55	0	10	83%	1	6	33%	53%	50%	53%
17	0,54	0	10	83%	1	7	39%	57%	44%	57%
18	0,51	0	10	83%	1	8	44%	60%	39%	60%
19	0,48	1	11	92%	0	8	44%	63%	47%	63%
20	0,46	0	11	92%	1	9	50%	67%	42%	67%
21	0,42	0	11	92%	1	10	56%	70%	36%	70%
22	0,42	0	11	92%	1	11	61%	73%	31%	73%
23	0,42	0	11	92%	1	12	67%	77%	25%	77%
24	0,4	1	12	100%	0	12	67%	80%	33%	80%
25	0,38	0	12	100%	1	13	72%	83%	28%	83%
26	0,34	0	12	100%	1	14	78%	87%	22%	87%
27	0,34	0	12	100%	1	15	83%	90%	17%	90%
28	0,3	0	12	100%	1	16	89%	93%	11%	93%
29	0,23	0	12	100%	1	17	94%	97%	6%	97%
30	0,21	0	12	100%	1	18	100%	100%	0%	100%

Tabla 7: Ejemplo de tabla para construir un gráfico de KS

En este caso, el índice KS (máxima separación) es del **72%** y se da cuando se ha contactado al 40% de la población.

Dado que en general un KS superior al 45% es un buen valor si se logra antes de alcanzar a la mitad de la población, se trata de un modelo de muy buena calidad.

El gráfico resultante se ve así:

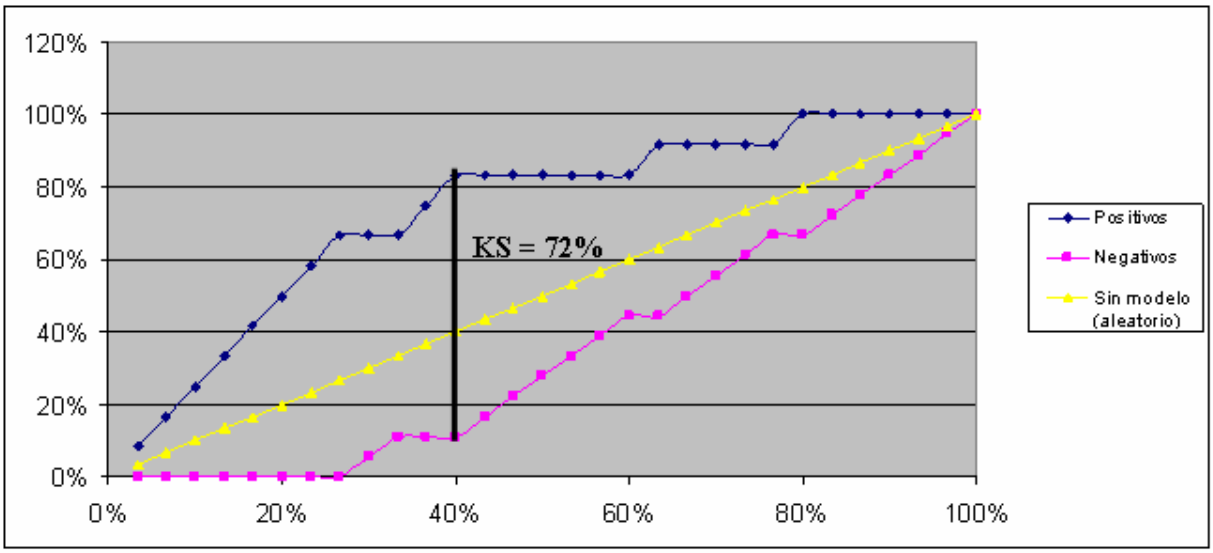


Gráfico 10: Gráfico KS de tabla 7

6.5.5.3 Tabla de Hosmer-Lemeshow

La tabla de Hosmer-Lemeshow recoge la información acerca de cómo actuó el modelo en comparación con la realidad. Se hace para cada decil de la muestra ordenada según las probabilidades asignadas por el modelo (ver tabla 8).

Decil	n Good (Real)	Tasa Good (Real)	n Good (Modelo)	Tasa Good (Modelo)	Error

10	44	52%	36	42%	18%
9	22	26%	18	21%	19%
8	21	25%	15	17%	30%
7	5	6%	13	15%	157%
6	4	5%	11	13%	172%
5	11	13%	10	12%	8%
4	6	7%	9	10%	47%
3	3	3%	8	9%	172%
2	2	3%	6	7%	191%
1	3	4%	5	6%	57%
Total	121		130		7%

Tabla 8: Tabla de Hosmer-Lemeshow

La columna “n Good Real” indica la cantidad de casos positivos que existen en el primer decil de la muestra ordenada por la probabilidad asignada por el modelo. La cantidad de casos positivos predicha por el modelo es la que se muestra en la columna “n Good Modelo”.

En este caso el modelo fue pesimista respecto de la realidad (pronosticó 36 cuando en realidad hubiesen sido 44). En este caso el error fue de (44 – 36) / 44 = 18%

Así en el decil de máxima probabilidad, el 10, hay 44 aceptaciones. Estas aceptaciones representan el 52% del total de aceptaciones (44/121). El modelo predice para este decil un total de 36 aceptaciones, que representan el 42% del total de aceptaciones predichas por el modelo (36/130).

Si observamos cómo se distribuyen las aceptaciones (“Good”) en el modelo vemos que son ordenadamente descendentes, tal como se espera de un modelo eficiente, con una única inversión en el decil 5 donde hay 11 aceptantes. El error global cometido por el modelo es del 7%.

6.6 Funcionalidades del Data Mining

Si bien el objetivo de un proyecto de Data Mining es único, como por ejemplo, “obtener un modelo de predicción de riesgo crediticio”, o “encontrar la combinación más apropiada de productos para realizar una oferta a los clientes”, es común que para lograrlo se deban resolver distintos tipos de problemas o tareas de análisis durante el transcurso del proyecto.

Estos problemas parciales aportan respuestas individuales que en su conjunto y luego de ser correctamente asociadas, permitirán obtener conclusiones más complejas que aquellas obtenidas de cada análisis por separado.

Las tareas de Data Mining pueden dividirse en Descriptivas y Predictivas. Según Berry y Linoff[7], las más importantes son las que se muestran en la siguiente ilustración:

[7] Berry and Linoff. 2008. Data Mining Techniques for marketing, sales and customer relationship

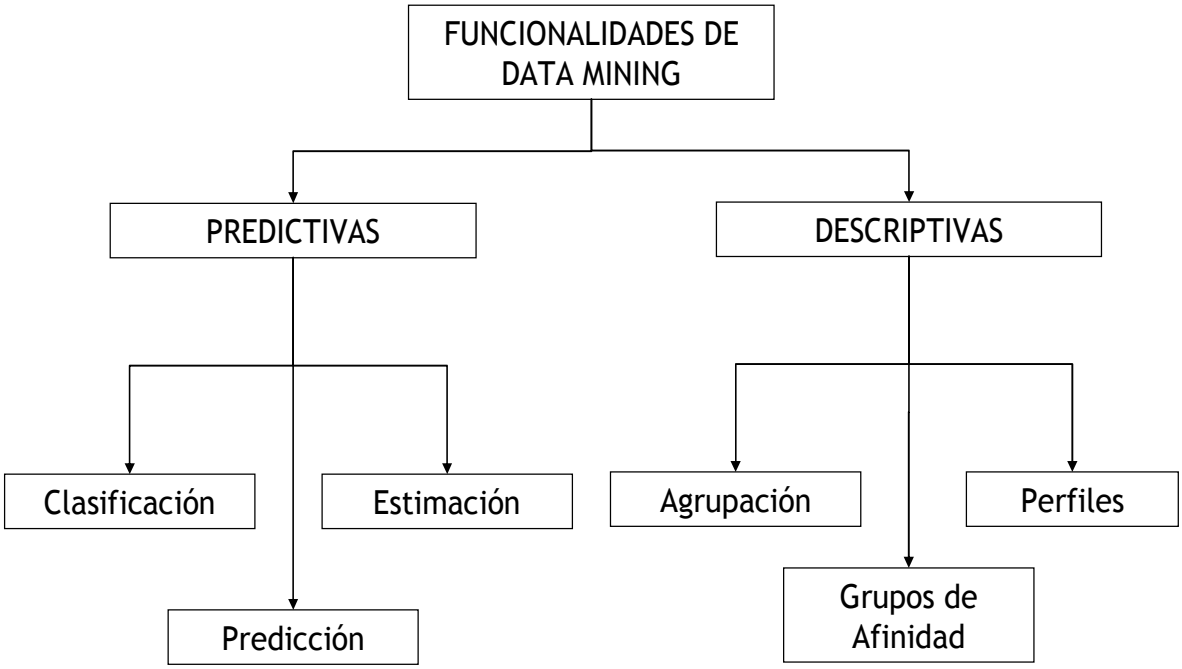


Ilustración 10: Clasificación de las funcionalidades del Data Mining

6.6.1 Predictivas

Las tareas predictivas son aquellas que buscan, a partir de casos conocidos, determinar las reglas de comportamiento de ciertas variables de modo tal de poder inferir el valor desconocido de la variable dependiente.

De todos modos las tareas que llevan a resolver uno y otro caso pertenecen a la rama de la predicción.

A continuación se describen las 3 tareas predictivas más importantes.

6.6.1.1 Clasificación

La Clasificación consiste en examinar las características de cierto objeto y en base a ellas asignarlo a una de un conjunto de clases predefinidas. Los objetos a ser clasificados son generalmente representados por registros (filas) en una base de datos donde las columnas contienen los valores de los distintos atributos del objeto.

Mediante el “acto” de clasificación se agrega una última columna con el código de clase.

Las variables que se predicen (clase a la que pertenece cada objeto) mediante este tipo de análisis son de tipo categóricas, es decir, no numéricas y sin orden establecido.

Ejemplos de aplicación:

- Clasificando solicitantes de Tarjetas de Crédito en Riesgo “Bajo”, “Medio” o “Alto”
- Identificando si el cliente “Comprará X”, “No Comprará X”
- Determinando si el cliente debe recibir “Tratamiento A”, “Tratamiento B”, “Tratamiento C”

Dado que las clases en las cuales se ubican a unos y otros registros son definidas por el usuario, a este tipo de problema se lo define como aprendizaje supervisado.

6.6.1.2 Estimación

La Estimación, al igual que la Clasificación, busca predecir el valor de una variable asociada a la persona u objeto en estudio. La diferencia es que en lugar de predecir variables categóricas, predice variables numéricas continuas. Dado el conjunto de atributos (x1, x2, x3, x4), la Estimación devuelve el valor numérico de cierta variable continua: “Peso”, “Deuda en Tarjeta de Crédito”, etc.

En la práctica, la estimación es utilizada para realizar tareas que típicamente serían de clasificación. Por ejemplo: Un banco quiere clasificar crediticiamente a sus clientes según sean de “Alto Riesgo”, “Medio Riesgo”, “Alto Riesgo”. Para ello podría utilizar un análisis de clasificación y asignar a cada cliente a una de las clases. Sin embargo un enfoque similar sería asignar a cada cliente un valor que represente el

nivel de riesgo crediticio, por ejemplo, un valor entre 0 y 1. La tarea de clasificación ahora debería determinar una “probabilidad umbral” a partir de la cual se considere que el cliente pasa de una categoría a la otra.

Este enfoque tiene la ventaja de que los objetos individuales (clientes en este caso), pueden ser rankeados de acuerdo al valor de riesgo asignado. Permite priorizar con mayor detalle.

Ejemplos de aplicación:

- Estimar el gasto anual en tarjeta de Crédito de cada Cliente
- Determinar el Valor Presente de un cliente para la empresa
- Estimar la propensión de los empleados a dejar la empresa

6.6.1.3 Predicción

La Predicción es similar a la Clasificación o la Estimación en el sentido que buscan conocer el valor de una variable desconocida en función de casos conocidos. La diferencia es que los registros son clasificados en base a algún comportamiento futuro estimado o valor futuro estimado. En una tarea de predicción la única forma de determinar el éxito del modelo es “esperar y ver”. La razón primaria para tratar esta tarea de forma separada a la Clasificación o a la Estimación es que en la Predicción existen aspectos adicionales relacionados con la relación temporal de las variables de entrada y la variable objetivo.

Cualquiera de las técnicas utilizadas para la clasificación o la estimación pueden ser adaptadas para utilizar en análisis de predicción mediante ejemplos de entrenamiento donde el valor de la variable a ser predicha es conocido junto con datos históricos para esos ejemplos. Los datos históricos son usados para construir un modelo que explique los valores observados actuales. Luego aplicando dicho modelo a los datos actuales se obtiene como resultado un comportamiento futuro estimado.

Ejemplos de aplicación:

- Probabilidad de ejercer el voto a favor de cierto candidato
- Propensión a adquirir cierta enfermedad

6.6.2 Descriptivas

En esta categoría se encuentran aquellas tareas cuyo objetivo es describir y entender los datos pero solo aquellos que son conocidos. No buscan encontrar los valores de una variable objetivo.

6.6.2.1 Grupos de Afinidad

El objetivo de los Grupos de Afinidad es el de determinar “que cosas van juntas”. El caso mas representativo es el de determinar que productos se adquieren juntos en una compra de supermercado. Los canales de venta de productos masivos utilizan este conocimiento para colocar uno al lado del otro en la góndola a aquellos productos que en general se llevan juntos.

La misma estrategia se sigue para armar paquetes de turismo o para amar paquetes atractivos de productos o servicios.

Ejemplos de aplicación:

- Encontrar aquellos productos que en general son adquiridos juntos
- Países que son visitados juntos en viajes a Europa

6.6.2.2 Perfiles (Profiling)

El análisis de Perfiles el proceso mediante el cual se busca describir las características de un grupo de clientes, identificando los atributos más destacados de dicho grupo, es decir, las variables que más lo representan.

Dichas variables pueden ser de tipo demográfico, situación económica, tipo de consumos, etc.

Se trata de responder preguntas como:

- ¿Cuáles son los atributos demográficos de mis mejores clientes?
- ¿Qué productos son los que consumen con más frecuencia mis clientes?
- ¿Porqué adquieren los productos de mi empresa?
- ¿Dónde se encuentran aquellos no clientes que son similares a mis clientes actuales?

6.6.2.3 Agrupación (Clustering)

La Agrupación es la tarea de segmentar una población heterogénea en sub-grupos más homogéneos (clusters).

Lo que difiere a la Agrupación de la Clasificación, es que no se basa en clases predefinidas por el usuario. Es decir, los objetos se ordenan solos en base a la “similitud” que presenten entre sí.

El usuario es luego el responsable de determinar el significado de cada agrupación, si existiese, y asignárselo al grupo.

Realizar este análisis sobre síntomas de pacientes enfermos podría indicar distintas enfermedades. Realizarlo sobre clientes podría indicar diferentes segmentos de mercado. En general este tipo de análisis es el paso previo a la aplicación de otras funcionalidades del Data Mining.

Por ejemplo: En vez de contestar la pregunta ¿Qué clientes respondieron mejor a las ofertas?, primero realizar una agrupación y luego analizar la respuesta de cada segmento por separado.

Ejemplos de aplicación:

- Identificar segmentos de clientes
- Identificar países con idiosincrasia similar

6.6.3 Otras funcionalidades

6.6.3.1 Text Mining (Minería de Texto)

Este tipo de análisis permite trabajar con datos no estructurados como pueden ser campos de texto que se completan por los clientes en forma libre. Es una aplicación muy utilizada para detectar posibles amenazas e incrementar la eficiencia de los sistemas de seguridad.

Se utiliza con varias finalidades, por ejemplo:

- Comprender en detalle las preferencias del cliente mediante el análisis de los campos de notas de las aplicaciones de los centros de llamadas
- Descubrir temas comunes y conceptos importantes en las respuestas a las preguntas abiertas de las encuestas
- Predecir qué tipos de fraude y abuso pueden ocurrir y dónde, gracias al análisis de la información de textos como la que ofrecen los campos de notas y los correos electrónicos.

- Proteger la seguridad pública de un modo más eficaz utilizando el análisis predictivo de textos para mejorar los modelos de posibles amenazas por parte de individuos o de grupos.

En 2007, la división Europol contra el Crimen desarrolló un sistema de análisis para rastrear al crimen organizado a nivel global. Este sistema hecho en base a técnicas de text mining, llevó a Europol a experimentar el progreso más significativo en la lucha contra el crimen a nivel internacional.[8]

6.6.3.2 Outliers (puntos extremos)

Los llamados “Outliers”, cuya traducción al español es “anómalos”, son aquellos casos en los cuales no se cumplen las leyes válidas para el resto de los registros.

Cuando se analizan bases, en general se ordenan en filas los registros (Ej. Clientes), mientras que en columnas se completan los datos para cada variable de ese cliente (Ej. Nombre, Edad, Saldo en Cuenta corriente, etc).

En data mining, cada registro (o cliente), se piensa como un punto n-dimensional, cuyas dimensiones toman los valores de cada una de las n variables.

[8] http://en.wikipedia.org/wiki/Text_mining#cite_note-2#cite_note-2

Para detectar los puntos anómalos, se detectan aquellos puntos cuya distancia a sus puntos vecinos son mayores. No es posible visualizar esta distancia en el espacio ya que en general los puntos tienen más de 3 dimensiones pero es matemáticamente posible saber cual es la distancia al punto vecino más próximo. Aquellos que estén más lejos serán eliminados de la muestra dado que no se consideran representativos.

Este tipo de análisis incrementa el nivel de precisión del modelo de Data Mining final.

6.6.3.3 Análisis de Secuencias

Los algoritmos de análisis de secuencias (Sequence analysis algorithms), detectan secuencias que se repiten en la información como por ejemplo cierta cadena de genes en el genoma de un ser humano. Año a año se descubren mas secuencias de genes y proteínas. Comparando secuencias conocidas contra las nuevas secuencias cuya funcionalidad se desconoce, es posible ir descubriendo que función cumplen.

Otra aplicación frecuente es la de evitar ataques informáticos mediante la detección de secuencias de caracteres que se repiten en este tipo de códigos.

6.6.3.4 Análisis Web

El campo de acción de este tipo de análisis son los negocios on-line. Es decir, es de utilidad siempre y cuando se tenga una página Web en la que ingresen usuarios y naveguen por ella. El indicador mas frecuente con el que se cuenta en estos casos es el número de visitas al sitio por día. Sin embargo, conocer solamente el volumen de personas que ingresan a la página no es suficiente para entender el perfil de los consumidores y las características personales, o los patrones de navegación que presentan aquellos que resultaron compradores del producto o servicio.

Estos módulos se ocupan de objetivos vitales de los negocios online como puede ser el aumento de tasas de adhesión mediante capacidades de análisis predictivo que van desde la segmentación automatizada de visitantes a la inteligencia para el marketing de un motor de búsqueda preceptivo.

Permite conocer “la posibilidad de adhesión por visitante individual” que es un indicador de negocios más útil que “número de visitas”.

6.6.3.5 Análisis de Redes Sociales

Una red social se define como una estructura de individuos que están relacionados (directa o indirectamente) a través de intereses comunes como por ejemplo amistad o confianza.

El análisis de las redes sociales es el estudio que permite entender su estructura y comportamiento. Su aplicación va desde marketing de productos hasta motores de búsqueda. Recientemente ha habido un rápido aumento del interés en este tipo de análisis dentro de la comunidad del Data Mining

Un ejemplo de aplicación fue durante el caso Enron donde se utilizó para analizar las relaciones entre los distintos miembros y así detectar a los involucrados en la estafa.

7 DESARROLLO DEL MODELO

Hasta aquí ya se ha hecho una introducción al negocio bancario y se ha explicado la importancia que tiene, dentro de este, la colocación de Plazos Fijos. A su vez se explicó como el Banco favorece las ventas de dicho producto mediante las campañas comerciales. Luego se explicó que el grupo objetivo se determina prácticamente en forma aleatoria, lo cual responde a una estrategia de marketing masivo por la pobre y casi ausente segmentación de los clientes. Esto trajo la posibilidad de incorporar una herramienta novedosa para ser más inteligentes en dicha selección: Modelos de Respuesta con Data Mining. Se explicó como un Modelo de Respuesta puede ayudar a incrementar en forma sustancial las tasas de éxito de las campañas comerciales, indicando con fundamentos científicos, a que clientes vender.

Se hizo a continuación un repaso general por el mundo del Data Mining, explicando todos los pasos del proceso que da origen a los modelos y algunas de las aplicaciones conocidas hasta ahora para esta herramienta.

Resta ahora explicar, teniendo en cuenta los requisitos detallados en el punto 6.6 y siguiendo el proceso descripto en el capítulo 6.3, de que manera se obtuvo el modelo que permitió incrementar las tasas de éxito en la colocación de Plazos Fijos en forma significativa para el Banco.

7.1 Requisitos

Tal como lo explica el punto 6.6, para comenzar con las tareas de Data Mining hubo que asegurar que se contaba con Software, Hardware, Datos.

7.1.1 Software

El Software utilizado elegido fue el SPSS Clementine.

La elección se realizo principalmente en base a la gran popularidad que tiene el programa entre los consumidores (ver grafico 2).

Este software, además de ser el más popular según la encuesta presentada, se ubica entre los que ofrecen mayor cantidad de funcionalidades (ver Tabla 1 de funcionalidades)

7.1.2 Hardware

No fueron necesarias inversiones en computadoras. En función de los requisitos para el software elegido, se determinó que tanto la capacidad de procesamiento instalada como la memoria en disco eran suficientes para soportar la instalación y utilización del mismo.

7.1.3 Datos

Por razones de seguridad un Banco registra y realiza back-ups automáticos de todas las operaciones que realizan sus clientes (consumos con tarjeta de crédito, tasa del préstamo solicitado, cantidad de Plazos Fijos colocados en los últimos 5 años). Se guardan por un período no menor a 2 años.

Las variables sociodemográficas (edad, sexo, lugar de residencia) son registradas al momento del alta de la cuenta del cliente.

7.2 Desarrollo de la solución

7.2.1 Selección gruesa de variables

Se eligieron 111 variables (ver Tabla 9) del total disponible en las bases del banco entre las cuales se encuentra la variable objetivo,”Tenencia de Plazo Fijo” (SI / NO). Por su naturaleza se las puede agrupar en: Cantidades de Producto (25), Tenencia de Productos (25), Consumos y Características de los productos Bancarios (50), sociodemográficas (5), identificación (1), otras (5).

Nombre del atributo	Descripción	Tipo
clinum	codigo identificatorio del cliente	**Identificación**
fecha_alta	fecha de ingreso como cliente	**Otras**
OPERACIÓN_USD		
CANAL_USD		
mes		
ORIGEN		
NSE	Nivel Socio Económico (A,B,etc)	**Socio Demográficas**
sexo	Genero sexual (M o F)	
Region	Region de residencia	
Estado Civil	Casado, Soltero, etc	
edad	Edad	
Cant_PP	Cantidad Actual de Prestamos Personales	**CANTIDAD**
Cant_P_ACORD	Cantidad Actual de Prestamos Pre Acordados	
Cant_PH	Cantidad Actual de Prestamos Hipotecarios+	
Cant_REF_PP	Cantidad Actual de Prestamos Personales Refinanciados	
Cant_CC	Cantidad de Cuentas Corrientes	
Cant_CA	Cantidad de Cajas de Ahorro	
Cant_DA	Cantidad de Débitos Automáticos Asociados a sus cuentas	
Cant_CJ		
Cant_SA		
Cant_SV		
Cant_MC	Cantidad de Tarjetas Master Card	
Cant_MC_TIT	Cantidad de Tarjetas Titulares Master Card	
Cant_MC_ADI	Cantidad de Tarjetas Adicionals Master Card	
Cant_VI	Cantidad de Tarjetas Visa	
Cant_VI_TIT	Cantidad de Tarjetas Titulares Visa	
Cant_VI_ADI	Cantidad de Tarjetas Adicionales Visa	
Cant_SH	Cantidad de Seguros de Hogar	
Cant_SAU	Cantidad de Seguros Automotor	
Cant_OTROS_ATM		
Cant_TD	Cantidad de Tarjetas de Débito	
Cant_PMC		
Cant_ATM		
Cant_Caja	Cantidad de Cajas de Seguridad	
Cant_IP		
Cant_HB		

Tabla 9: Variables seleccionadas por filtro “grueso”

Flag_PP	Tenencia Actual de Prestamos Personales	**TENENCIA (variables binarias SI o NO)**
Flag_P_ACORD	Tenencia Actual de Prestamos Pre Acordados	
Flag_PH	Tenencia Actual de Prestamos Hipotecarios+	
Flag_REF_PP	Tenencia Actual de Prestamos Personales Refinanciados	
Flag_CC	Tenencia de Cuentas Corrientes	
Flag_CA	Tenencia de Cajas de Ahorro	
Flag_DA	Tenencia de Débitos Automáticos Asociados a sus cuentas	
Flag_CJ		
Flag_SA		
Flag_SV		
Flag_MC	Tenencia de Tarjetas Master Card	
Flag_MC_TIT	Tenencia de Tarjetas Titulares Master Card	
Flag_MC_ADI	Tenencia de Tarjetas Adicionals Master Card	
Flag_VI	Tenencia de Tarjetas Visa	
Flag_VI_TIT	Tenencia de Tarjetas Titulares Visa	
Flag_VI_ADI	Tenencia de Tarjetas Adicionales Visa	
Flag_SH	Tenencia de Seguros de Hogar	
Flag_SAU	Tenencia de Seguros Automotor	
Flag_otros_ATM		
Flag_TD	Tenencia de Tarjetas de Débito	
Flag_PMC		
Flag_ATM		
Flag_Caja	Tenencia de Cajas de Seguridad	
Flag_IP		
Flag_HB		
CA_VIGENTE		**PRODUCTOS Y CONSUMO**
CA_SUM_CRED		
CA_SUM_DEB		
SDO_CA_SUM_ARP_USD		
SDO_CA_PROM_ARP		
SDO_CA_PROM_USD		
SDO_CA_PROM_EUR		
SDO_CC_MAX_MES		
SDO_CC_SUM_CRED		
SDO_CC_PROM_ARP_DEU		
SDO_CC_FIN_MES		
HABERES		
JUBILACIONES		
PF_SUM_CAP_ARP_USD		
PF_CAP_PROM_ARP		
PF_CAP_PROM_USD		
PF_CAP_PROM_EUR		
PMO_PER_SDO_VENCER		
PMO_PER_SDO_VENCIDO		
PMO_PRE_SDO_VENCER		
PMO_PRE_SDO_VENCIDO		
PMO_RE_SDO_VENCER		
PMO_RE_SDO_VENCIDO		
SEGAU		
SEGHOG		
TC_CONS_ARP		
TC_CONS_USD		
TC_SUM_CONS_ARP_USD		
TC_SUM_SALD_ARP_USD		
TC_SALD_ARP		
TC_SALD_USD		
TC_CANT_ADIC		
TC_LIMITE_V		
TC_LIMITE_M		
TC_FINAN_V		
TC_FINAN_M		
TC_DNV_V		
TC_DNV_M		
TC_COMPRAS_ARP		
TC_COMPRAS_USD		
TC_COMPRAS_ARP_USD		
SDO_CC_PROM_ARP		
TD_CANT		
TD_MONTO		
TC_Limite_Total		
TC_Finan_Total		
TC_Dnv_Total		
PP_Monto_Total		
PR_Monto_Total		
RE_Monto_Total		

Tabla 9 (cont.): Variables seleccionadas por filtro “grueso”

7.2.2 Pre-Procesado de Datos

7.2.2.1 Limpieza de Datos

Por razones legales y de seguridad, el Banco debe asegurar la calidad de todos los datos que se registran tanto en forma automática como manual, por lo que en general son limpios y no contienen errores.

El Banco Central realiza auditorías frecuentes para asegurar que esto se cumpla y penaliza con fuertes multas los incumplimientos detectados en este sentido.

De todas formas, se relevó el proceso de carga manual de Datos en el cual se asientan los valores de las variables sociodemográficas para los nuevos clientes. Dado que existe control cruzado al momento de la carga (uno carga el otro controla) se aceptó en forma previa la calidad de todos los datos presentes en las Bases.

No obstante, en el pre-lavado, se filtraron aquellas variables que por contener valores faltantes, ruido, o valores atípicos, no eran factibles de utilización en la generación del modelo.

Cabe aclarar que el caso de los Bancos es un caso especial ya que la correcta operación del negocio esta basada casi por completo en el flujo de información. Es por esto que la etapa de limpieza puede ser, como ocurrió en el presente trabajo, omitida. Esto no significa tomar datos de mala calidad sino, en vez de corregir los errores que pudieran haber, eliminar las variables y registros inviables mediante el prelavado.

7.2.2.2 Integración de Datos

En el Banco se identifica a los clientes según un código alfanumérico de 6 caracteres al que se lo conoce como *clinum***.** No existen dos clientes con el mismo clinum.

Esta fue la variable que permitió la integración de todas las tablas.

7.2.2.3 Normalización de Datos

No se llevo a cabo una normalización de los Datos. Como se explicó anteriormente, no siempre se aplican indefectiblemente todos los pasos sino que en función de las necesidades se determina cuales realizar. En este caso las variables no mostraron diferencias significativas a priori en las magnitudes relativas como para esperar la que unas predominen sobre otras.

7.2.3 Selección Fina de Variables

7.2.3.1 Prelavado

Una vez realizada la integración y consolidación de todas aquellas variables elegidas en la selección gruesa en una única tabla, se procedió al proceso de prelavado.

Mediante este proceso lo que se buscó es asegurar que las variables que iban a ser analizadas cumplieran con determinadas condiciones mínimas de aceptación, tal que los resultados obtenidos en forma posterior tuviesen significancia estadística.

En este paso, se deben definir los criterios de filtro (cribado) para las variables y registros tal que se eliminen aquellas y aquellos que no aportan información útil. Las opciones de filtro se basan en los valores observados de la variable en análisis, sin contemplar la eficacia predictiva de la misma. Los registros o variables cribadas se excluyen del proceso.

A continuación se muestra como se realiza el prelavado utilizando el software.

Se pueden elegir entre 0 y 5 criterios de filtro o cribado (ver Ilustración 12). Es posible modificar los valores umbral que el programa recomienda por default.

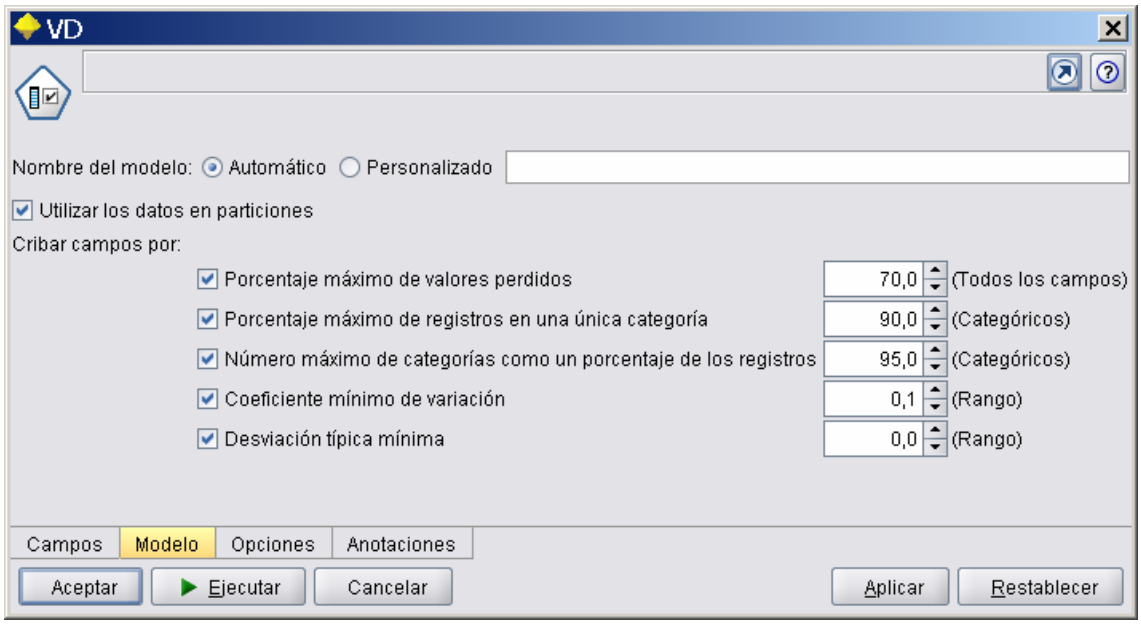


Ilustración 11: Selección de criterios y valores umbral en Prelavado

Descripción de los criterios:

• **Porcentaje máximo de valores perdidos**: Criba campos con demasiados valores perdidos, expresados como un porcentaje del número total de registros. Los campos con un alto porcentaje de valores perdidos proporcionan poca información predictiva.

• **Porcentaje máximo de registros en una categoría única**: Criba campos con demasiados registros dentro de la misma categoría en relación con el número total de registros. Por ejemplo, si el 95% de los clientes de la base de datos conduce el mismo tipo de coche, no sería útil incluir esta información para distinguir a un cliente de otro. Cualquier campo que exceda el máximo especificado se criba. Esta opción sólo se aplica a campos categóricos.

• **Número máximo de categorías como un porcentaje de registros:** Criba campos con demasiadas categorías en relación con el número total de registros. Si un porcentaje elevado de las categorías contiene sólo un único caso, puede que el campo sea de uso limitado. Por ejemplo, si cada cliente lleva un sombrero diferente, será improbable que esta información sirva a la hora de modelar patrones de comportamiento. Esta opción sólo se aplica a campos categóricos.

• **Coeficiente mínimo de variación:** Criba campos con un coeficiente de varianza menor o igual que el mínimo especificado. Esta medida es el cociente de la desviación típica del predictor dividida por la media del predictor. Si este valor es cercano a cero, no habrá mucha variabilidad en los valores de la variable. Esta opción sólo se aplica a campos de rango numérico.

• **Desviación típica mínima**: Criba campos con desviación típica menor o igual que el mínimo especificado. Esta opción sólo se aplica a campos de rango numérico.

7.2.3.2 Selección de Variables

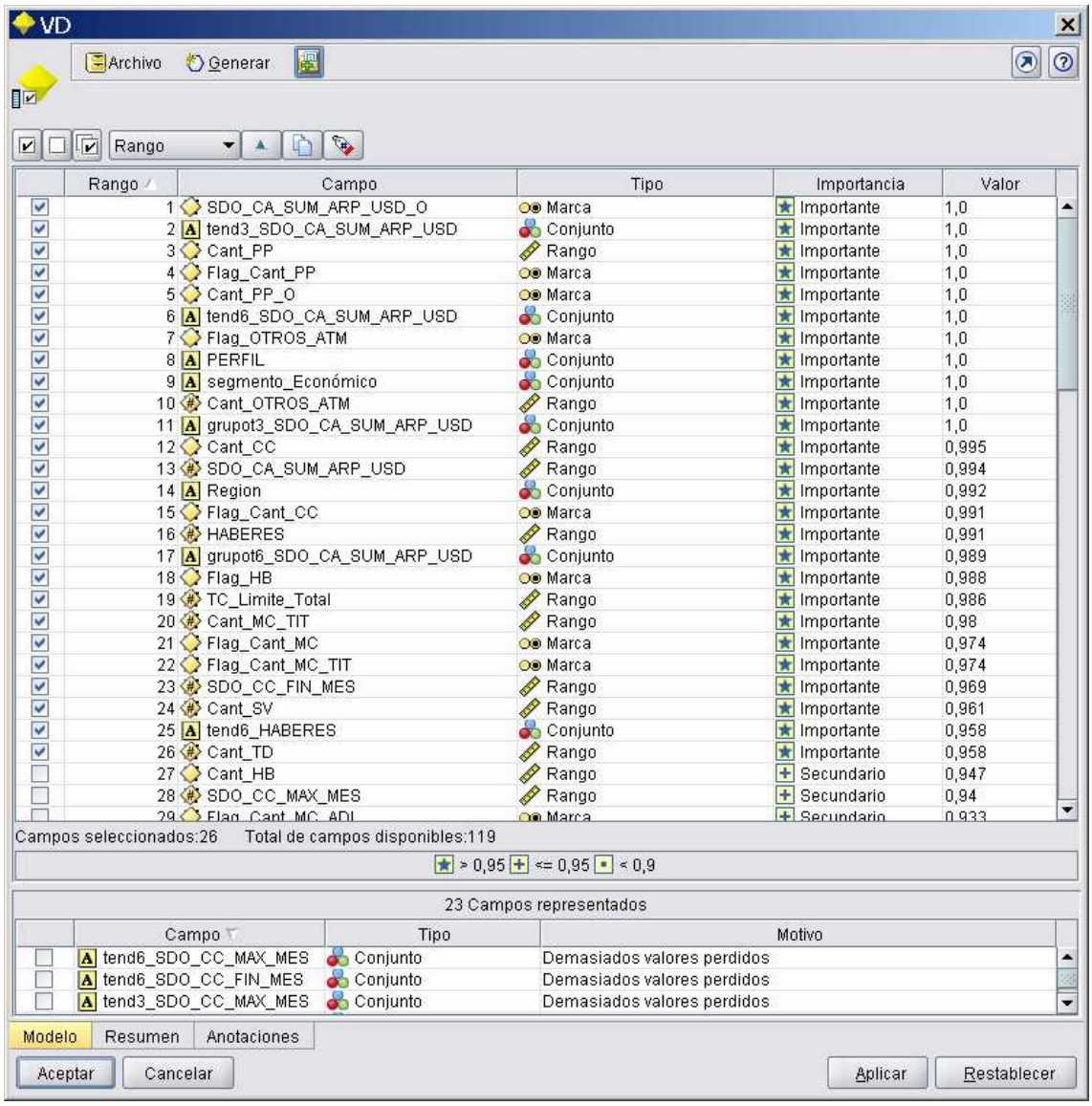
Habiendo establecido los criterios de cribado para el prelavado, se realiza sobre los registros y variables que no fueron excluídas, el filtrado de estas últimas pero ahora en función de su calidad como predictoras. Para hacerlo se utilizó un modelo *tipo filtro*, es decir, se asigna a cada variable valores de importancia predictiva sin importar como interactúan con el algoritmo de predicción que se utilizará en forma posterior para la creación del modelo.

A continuación se muestra la pantalla del software en la cual se muestran los resultados

En la sección sección inferior se listan las variables excluídas por no cumplir con alguno de los criterios del prelavado.

Al mismo tiempo en la parte superior, las variables quedan ordenadas según su importancia para la predicción de la variable objetivo.

La naturaleza de la medida utilizada es de Dependencia, es decir, involucra coeficientes de correlación entre variables.



Rango	Campo	Tipo	Importancia	Valor
1	SDO_CA_SUM_ARP_USD_O	Marca	Importante	1,0
2	tend3_SDO_CA_SUM_ARP_USD	Conjunto	Importante	1,0
3	Cant_PP	Rango	Importante	1,0
4	Flag_Cant_PP	Marca	Importante	1,0
5	Cant_PP_O	Marca	Importante	1,0
6	tend6_SDO_CA_SUM_ARP_USD	Conjunto	Importante	1,0
7	Flag_OTROS_ATM	Marca	Importante	1,0
8	PERFIL	Conjunto	Importante	1,0
9	segmento_Económico	Conjunto	Importante	1,0
10	Cant_OTROS_ATM	Rango	Importante	1,0
11	grupot3_SDO_CA_SUM_ARP_USD	Conjunto	Importante	1,0
12	Cant_CC	Rango	Importante	0,995
13	SDO_CA_SUM_ARP_USD	Rango	Importante	0,994
14	Region	Conjunto	Importante	0,992
15	Flag_Cant_CC	Marca	Importante	0,991
16	HABERES	Rango	Importante	0,991
17	grupot6_SDO_CA_SUM_ARP_USD	Conjunto	Importante	0,989
18	Flag_HB	Marca	Importante	0,988
19	TC_Limite_Total	Rango	Importante	0,986
20	Cant_MC_TIT	Rango	Importante	0,98
21	Flag_Cant_MC	Marca	Importante	0,974
22	Flag_Cant_MC_TIT	Marca	Importante	0,974
23	SDO_CC_FIN_MES	Rango	Importante	0,969
24	Cant_SV	Rango	Importante	0,961
25	tend6_HABERES	Conjunto	Importante	0,958
26	Cant_TD	Rango	Importante	0,958
27	Cant_HB	Rango	Secundario	0,947
28	SDO_CC_MAX_MES	Rango	Secundario	0,94
29	Flag_Cant_MC_ADI	Marca	Secundario	0,933

Campo	Tipo	Motivo
tend6_SDO_CC_MAX_MES	Conjunto	Demasiados valores perdidos
tend6_SDO_CC_FIN_MES	Conjunto	Demasiados valores perdidos
tend3_SDO_CC_MAX_MES	Conjunto	Demasiados valores perdidos

Ilustración 12: Filtro de variables y asignación de importancia para la predicción

Como se muestra en la ilustración anterior, las variables se ordenaron según la medida elegida para determinar su importancia.

En lugar de seleccionar las primeras X variables, se optó por mantener todas aquellas catalogadas como “Importante” por el software. Lógicamente esta etiqueta esta directamente relacionada con el valor de importancia obtenido, el cual es superior a 0,95 para todos los casos.

7.2.4 Creación del Modelo mediante Algoritmos de Generación

Una vez seleccionadas las variables mas relevantes para la predicción, se procede con la generación del modelo predictivo.

Los algoritmos probados fueron:

- Redes Neuronales
- C5
- CHAID
- Combinación (C5 + Red Neuronal + CHAID)

Tomando como input las variables elegidas en el paso 7.2.3, se indicó al programa que generara los modelos predictivos para cada caso.

A continuación se procedió con la evaluación de cada modelo en forma individual y por último la combinación de ellos. Para hacerlo se determinó al azar una muestra de comprobación sobre la cual se corrieron los 4 modelos y se observaron los resultados.

7.2.5 Combinación de modelos parciales para obtener el modelo final

Como se mencionó anteriormente, se combinaron 3 algoritmos para obtener pronósticos más precisos de los que pueden conseguirse de los modelos individuales. Al combinar predicciones de varios modelos, pueden evitarse las limitaciones en modelos individuales que dan como resultado una precisión global superior. Los modelos combinados de esta forma suelen ejecutarse tan bien como el mejor de los modelos individuales y, en ocasiones, mejor.

El método utilizado es la votación, que funciona cuadrando el número de veces que cada posible valor pronosticado se elige y seleccionando el valor con mayor valor total. Por ejemplo, si tres de los cinco modelos pronostican sí y los otras dos pronostican no, sí gana por 3 a 2. A dicho registro se le asigna la mayor probabilidad individual que hayan arrojado los 3 modelos que votaron sí.

7.2.6 Evaluación y selección del Modelo Predictivo

A continuación se muestran los resultados de la evaluación para cada uno de los 4 modelos generados. Las herramientas utilizadas para evaluarlos uno respecto de otro fueron:

- Gráfico de ganancia
- Índice KS
- Tabla de Hosmer-Lemeshow

Para todos los casos, la evaluación consta de alimentar el modelo con una muestra de n registros con resultado conocido y obtener la probabilidad de aceptación de cada cliente que entrega el modelo. Luego se ordenan los registros según dicha probabilidad y se determina cuanto más exitosos hubieran resultado los contactos si se hubiese seguido lo indicado por el modelo.

7.2.7 Modelo de tipo C5

A continuación se muestra el gráfico de ganancia obtenido para este modelo:

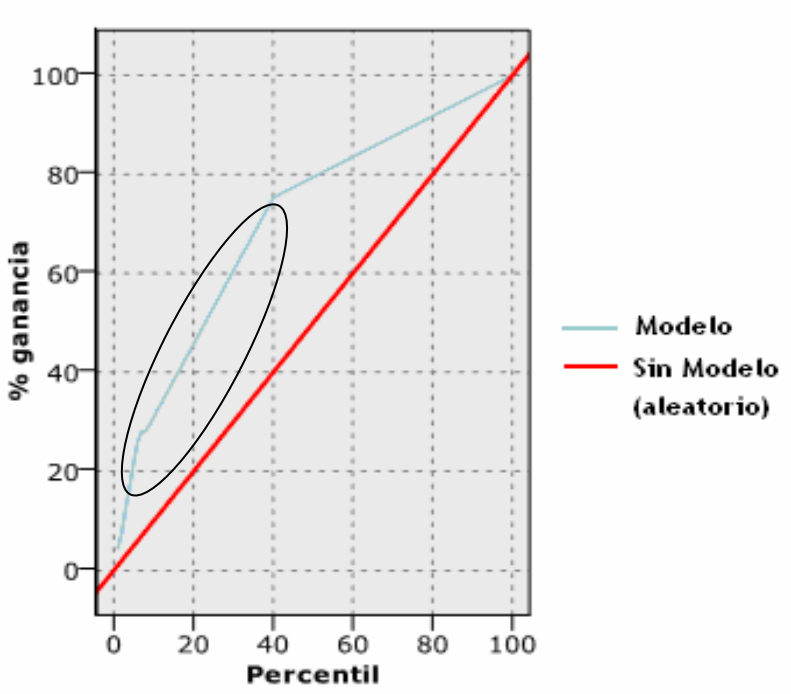


Gráfico 11: Gráfico de Ganancia para el Modelo obtenido mediante C5

Lo que se observa en este modelo (ver Gráfico 11) es un crecimiento lineal, casi paralelo a la efectividad sin modelo. La ganancia es algo baja dado que para los primeros deciles la curva del Modelo se despega “lentamente” de la del Modelo, sobre todo a partir del primer quiebre que se da en el percentil 10 aproximadamente.

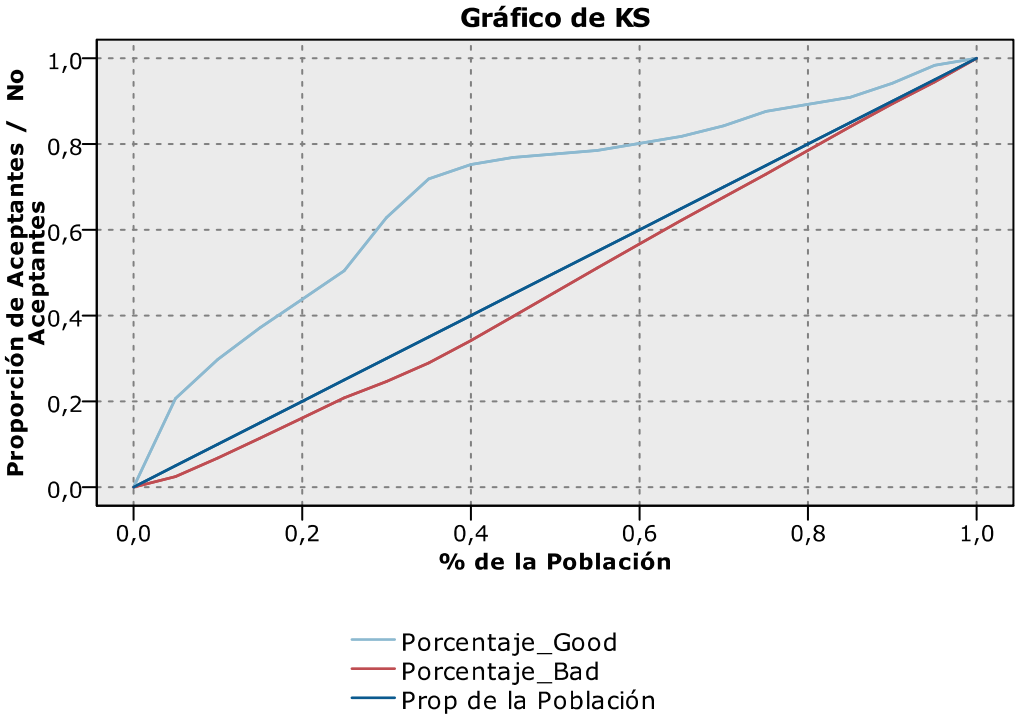


Gráfico 12: Gráfico de KS para el Modelo obtenido mediante C5

El máximo KS alcanzado por el modelo es de 42 que esta por debajo de 45, resultando bajo. Además la separación entre “Good” y “Bad” a lo largo de la muestra es pobre (ver gráfico 12).

Decil	n Good Real	Tasa Good Real	n Good Modelo	Tasa Good Modelo	Error
10	36	42%	41	49%	15%

9	17	20%	13	16%	22%
8	23	27%	13	16%	43%
7	15	18%	13	15%	13%
6	3	4%	5	5%	51%
5	3	4%	3	4%	15%
4	5	6%	3	4%	32%
3	6	7%	3	4%	43%
2	6	7%	3	4%	43%
1	7	8%	3	4%	51%
Total	121		102		15%

Tabla 10: Tabla Hosmer-Lemeshow para el Modelo obtenido mediante C5

Se observa en la tabla anterior una tendencia creciente de la “Tasa Good” en los deciles de menor propensión (4 a 1). Esto no es un buen indicio dado que la proporción de aceptantes debe ser decreciente a medida que decrece la propensión asignada por el modelo. El decil 4, categorizado por el modelo como más propenso a aceptar el producto que el decil 1, resultó en la realidad ser menos propenso (6% contra 8%).

El error global del modelo es del 15% por defecto, es decir, fue pesimista al determinar los aceptantes (121 reales contra 102 pronosticados).

7.2.8 Modelo de tipo CHAID

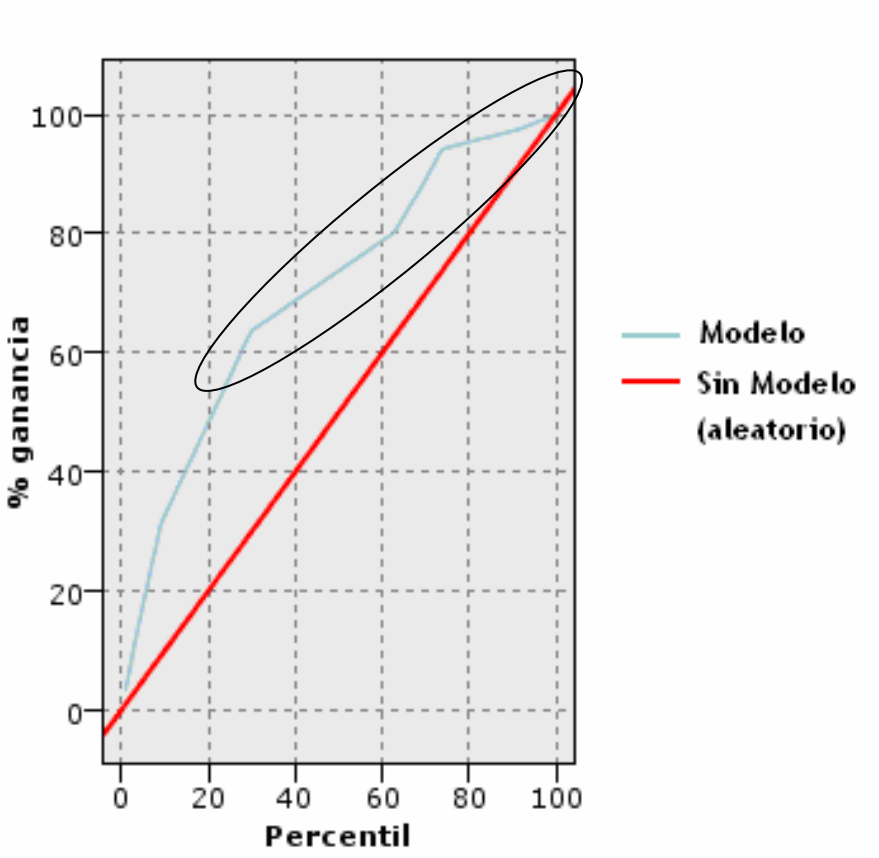


Gráfico 13 Gráfico de Ganancia para el Modelo obtenido mediante CHAID

En el gráfico anterior pueden observarse 2 cosas. Primero, la ganancia es baja, dado que para un 40% de la muestra no se logra contactar al 70% de los aceptantes. Por otro lado, se observa inestabilidad en la curva del modelo (Esta

curva representa la evolución del porcentaje de aceptantes a medida que crece el porcentaje de contactados)

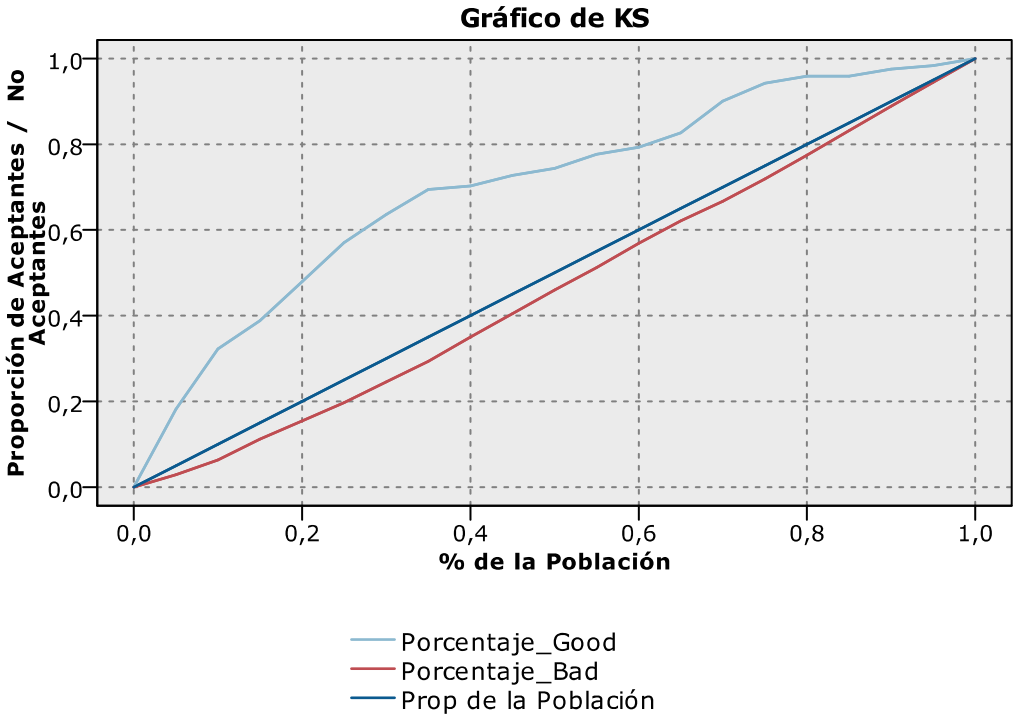


Gráfico 14 Gráfico de KS para el Modelo obtenido mediante CHAID

El valor de KS es 40 (gráfico 14), aún mas bajo que para el modelo generado con algoritmo del tipo C5.

Decil	n Good Real	Tasa Good Real	n Good Modelo	Tasa Good Modelo	Error
10	39	46%	41	48%	5%
9	19	22%	22	26%	17%
8	19	22%	21	24%	9%
7	8	10%	5	6%	40%
6	5	6%	5	6%	2%
5	6	7%	5	6%	18%
4	13	15%	4	4%	73%
3	7	8%	2	2%	76%
2	2	2%	1	2%	33%

1	3	4%	1	1%	61%
Total	121		106		12%

Tabla 11: Tabla Hosmer-Lemeshow para el Modelo obtenido mediante CHAID

Se observa una inversión de la tendencia decreciente para la "Tasa Good Real" en el decil 4 (ver Tabla 11). El modelo muestra un error global del 12% por defecto.

7.2.9 Modelo Red Neuronal

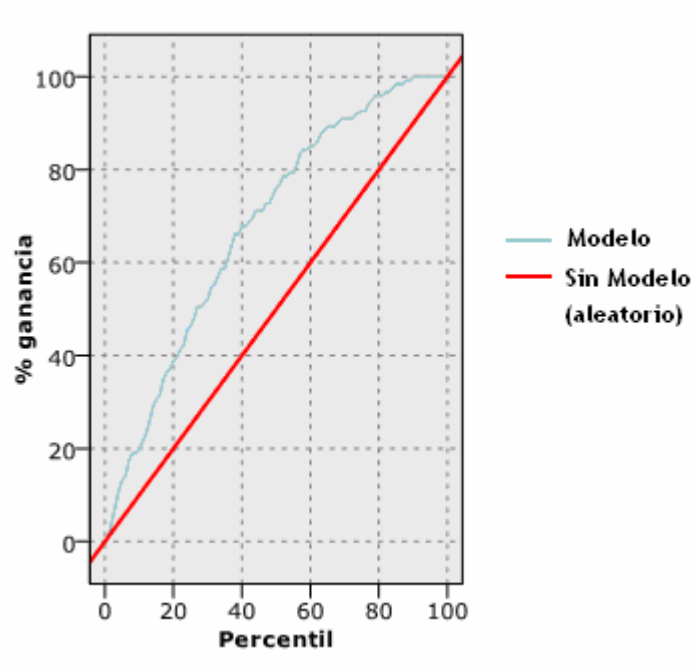


Gráfico 15: Gráfico de Ganancia para el Modelo obtenido mediante Red Neuronal

Si bien no hay inestabilidad en la evolución, existe una baja acumulación de aceptantes en los primeros deciles (deciles más propensos según el modelo). Cuando se alcanza al 40% de la población, no se logra contactar al 70% de los aceptantes (ver gráfico 15). Esta ganancia se considera baja.

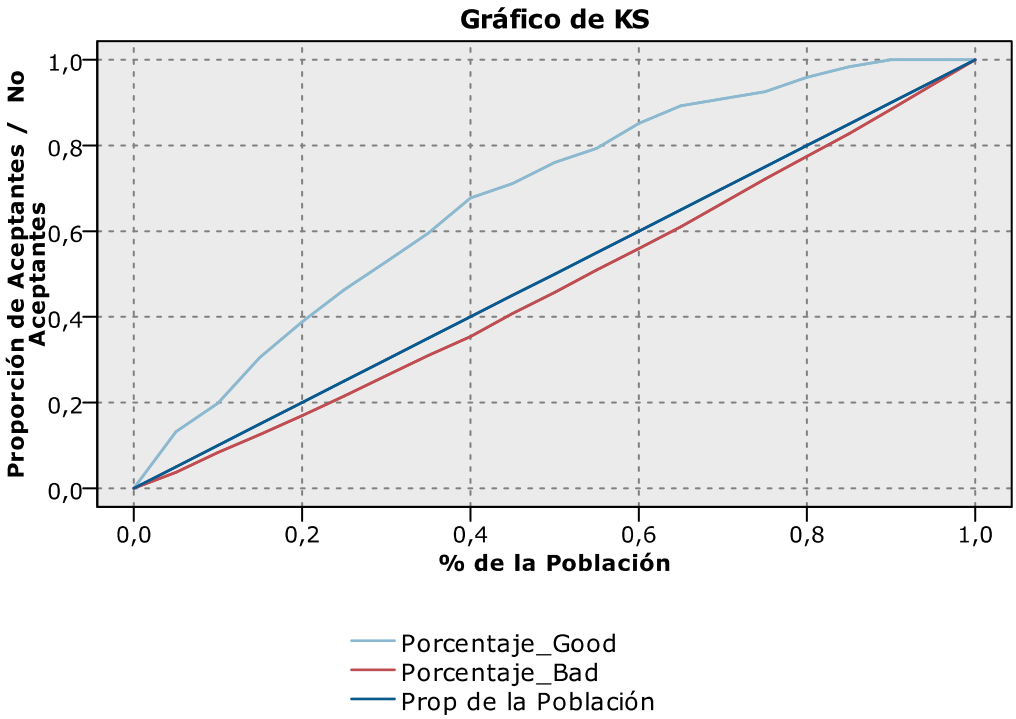


Gráfico 16: Gráfico de KS para el Modelo obtenido mediante algoritmo tipo Red Neuronal

El gráfico anterior muestra una buena separación entre “Good” (aceptantes) y “Bad” (no aceptantes) a lo largo de la curva aunque el KS máximo del modelo es de 32, el cual es bajo.

Decil	n Good Real	Tasa Good Real	n Good Modelo	Tasa Good Modelo	Error
10	24	28%	35	41%	44%

9	23	27%	33	39%	44%
8	17	20%	32	37%	87%
7	18	21%	30	36%	68%
6	10	12%	28	33%	184%
5	11	13%	27	32%	147%
4	7	8%	17	20%	142%
3	6	7%	10	11%	60%
2	5	6%	7	8%	39%
1	0	0%	6	7%	
Total	121		224		85%

Tabla 12: Tabla Hosmer-Lemeshow para el Modelo obtenido mediante algoritmo tipo Red Neuronal

La tabla 12 muestra los casos ordenados, casi sin inversiones, pero con alta tasa de error (85%).

7.2.10 Combinación (C5 + Red Neuronal + CHAID)

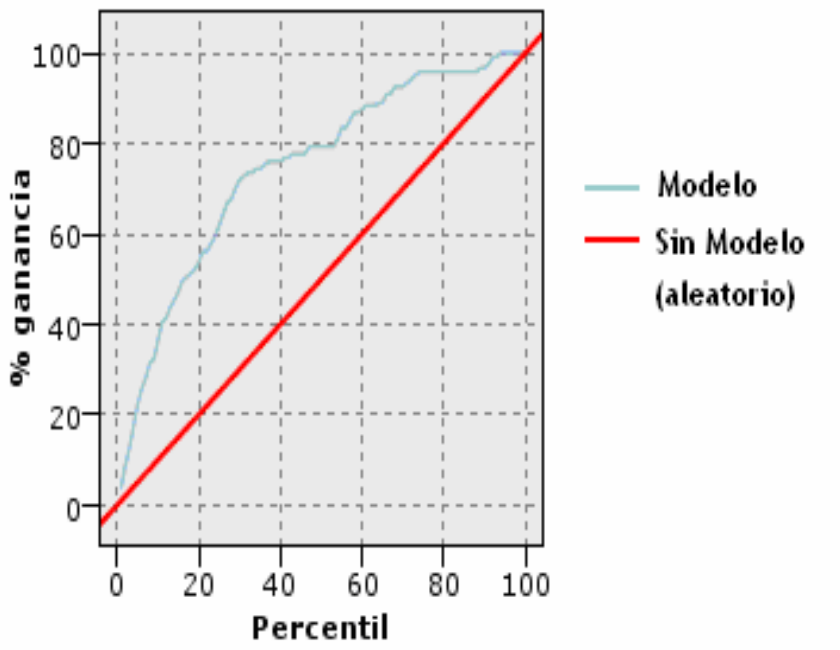


Gráfico 17 Gráfico de Ganancia para Modelo combinado obtenido mediante C5 + Red neuronal + CHAID

El gráfico anterior de Ganancia del modelo combinado final para la muestra de comprobación indica que el modelo agrupa cerca del 80% de los clientes que aceptaron la oferta dentro del 40% de población con mayor probabilidad asignada. Además la ganancia es buena a lo largo de toda la curva ya que la curva del modelo se despega rápidamente logrando contactar casi el 60% de los aceptantes en el 20% de la población (ver gráfico 17).

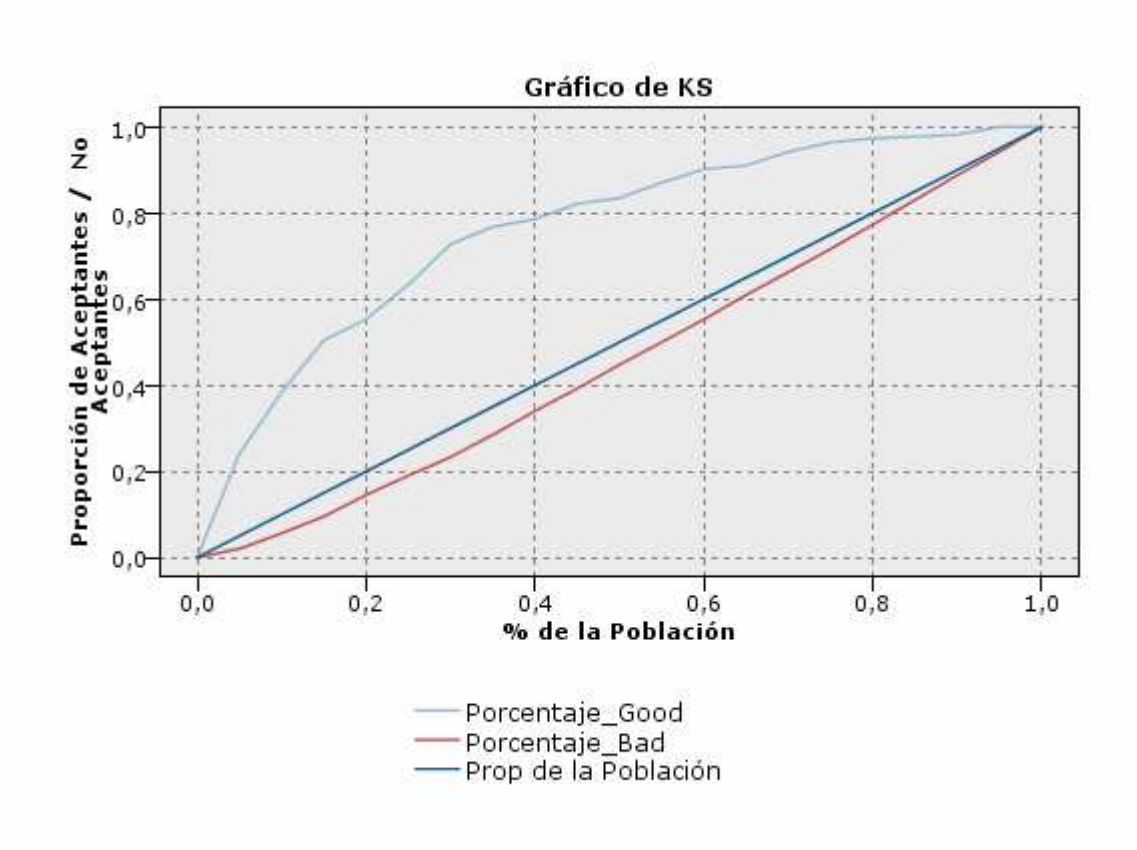


Gráfico 18 Gráfico de KS para Modelo combinado obtenido mediante C5 + Red neuronal + CHAID

Como muestra el gráfico anterior el KS es de 48 y la separación es buena a lo largo de toda la curva.

Decil	n Good Real	Tasa Good Real	n Good Modelo	Tasa Good Modelo	Error
10	44	52%	36	42%	18%
9	22	26%	18	21%	18%

8	16	19%	15	17%	6%
7	12	14%	13	15%	8%
6	7	8%	11	13%	57%
5	6	7%	10	12%	67%
4	5	6%	9	10%	80%
3	3	4%	8	9%	167%
2	3	4%	6	7%	100%
1	3	4%	5	6%	67%
Total	121		130		7%

Tabla 13: Tabla Hosmer Lemeshow para Modelo combinado obtenido mediante C5 + Red neuronal + CHAID

La tabla anterior muestra que para el decil de mayor probabilidad asignada por el modelo, se encontraron 52% de casos positivos, aún mas que el 42 % pronosticado por el modelo. El error es del 18% por defecto, es decir, el modelo fue pesimista en su estimación. El error por defecto se prefiere al error por exceso.

Si se observa la tendencia de la “Tasa Good Real” se ve que a medida que desciende la probabilidad asignada por el modelo, los deciles contienen una proporción cada vez menor de aceptantes, tal como se espera de un modelo eficiente. No hay inversión.

8 IMPLEMENTACION Y RESULTADOS

8.1 Prueba del Modelo en caso Real

El modelo seleccionado fue el Combinado (C5 + Red Neuronal + CHAID), ya que en las 3 pruebas resultó superior a sus competidores. Mediante el gráfico de ganancia se comprobó que se acumulaba el 80% de aceptantes en el primer 40% de población contactada, creciendo en forma exponencial. El índice KS obtenido mediante el segundo gráfico fue de 48, lo cual, según se explicó, da idea de un buen modelo. Por último, la tabla H-L mostró que no existían inversión en la cantidad de aceptantes a medida que se descendía en los deciles.

Una vez obtenido el modelo se lo puso a prueba. La forma de determinar el impacto de la implementación del modelo para la selección de clientes es utilizándolo en una campaña real.

Se seleccionaron 2 grupos de clientes. El primero, el grupo testigo, se seleccionó según la vieja estrategia.

El segundo se eligió alimentando el modelo con la tabla de clientes y luego de obtener las probabilidades de aceptación de cada cliente, se eligieron aquellos más propensos.

Luego de realizar las ofertas comerciales a todos los clientes, se midió la tasa de éxito de ambos grupos, es decir, el porcentaje de ventas logrado sobre el total de clientes contactados en cada estrategia.

Los resultados se muestran en la siguiente tabla:

Campaña Plazo Fijo					
SIN MODELO (testigo)			CON MODELO		
Contactados	Vendidos	Tasa de Éxito	Contactados	Vendidos	Tasa de Éxito
2000	***350***	***18%***	***2000***	***720***	***36%***

Tabla 14: Comparación entre ambas estrategias de selección de clientes

8.2 Impacto económico logrado

8.2.1 Ingresos Generados

En 2008 se dieron de alta por campañas comerciales un total de **2000 Plazos Fijos[9].**

A continuación se muestra la tabla[10] que contiene los parámetros necesarios para calcular el rendimiento esperado de un plazo fijo (la ganancia resultante por la venta)

Tasa pagada que se paga al cliente	11,70%	
Plazo Promedio de colocación (meses)	**4**	
Tasa cobrada a la Mesa de Dinero	15,15%	
Spread Anual	3,45%	(tasa cobrada - tasa pagada)
Spread Mensual	**0,2875%**	
Monto Promedio colocado en Plazos Fijos	**35000**	
Beneficio por PF vendido	**$ 403**	(Monto Promedio * Spread Mensual * Plazo Promedio de Colocación)

Tabla 15: Cálculo de Beneficio por PF vendido

Esto significa que cada Plazo Fijo adicional que se vende representa para el Banco un ingreso de $403 (ver tabla 15).

[9] Informe de Performance 2008 – Campañas Comerciales Banco Itau

[10] Datos de Inteligencia Comercial – Banco Itau

Para calcular el volumen esperado de ventas en 2009, se supondrá que durante 2009 la cantidad de clientes contactados, el contexto, los canales utilizados y la atractividad del producto son similares a los del 2008.

Dado que la relación entre tasas es de 2, como se asumen igual cantidad de contactos en el 2009, las ventas de plazos fijos se calculan como:

Volumen de Ventas 2009 (con modelo) = 2000 * 36% / 18% = 4000 Plazos Fijos

Comparando ambos escenarios, sin modelo y con modelo se observa que, gracias a la implementación del modelo, se logra un incremento en las ganancias del Banco de $ 805.000 anuales (ver tabla 16).

	Volumen	Ganancia Generada
Plazos Fijos dados de alta 2008 sin modelo	2000	$ 805.000
Plazos Fijos dados de alta 2009 con modelo	4000	$ 1.610.000
Incremento en la Ganancia del Banco	**2000**	**$ 805.000**

Tabla 16: Incremento en Ganancias logradas

8.2.2 Costos incurridos

Los costos adicionales en los que se incurrió para lograr la solución tienen tres componentes: Licencia del Software, Hardware y Mano de Obra. A continuación se detalla cada uno de ellos

8.2.2.1 Software

Es en realidad una inversión. La licencia para la adquisición del software tiene un valor de $50.000.

Calculando una vida útil de 5 años y realizando la división simple entre el monto pagado y los años de utilización se concluye que:

Costo anual del software = $60.000 / 5 años = **$12.000 / año**

8.2.2.2 Hardware

Si bien no fue necesaria la instalación de equipos adicionales, se considera que en algún momento será necesaria la compra de una nueva computadora para reemplazar la asignada a los proyectos de Data Mining. Se supone una vida útil de 3 años A precios actuales, un equipo como el descripto según los requerimientos del Software tiene un costo de:

Costo Anual del Hardware (PC) = $ 4.000 / 3 años = $ 1.300 / año

8.2.2.3 Mano de Obra

El tiempo de dedicación necesario para lograr el modelo predictivo y luego acompañar en todas las etapas de implementación, control, evaluación y actualización es de **4 meses efectivos.**

Actualmente el sueldo percibido por este profesional es de $10.000 pero al Banco le cuesta **$12.000** debido a las cargas sociales.

Dado que trabaja solo 4 meses al año abocado a los modelos de Plazos Fijos, se tiene que:

Costo Anual Mano de Obra = $12.000 / mes * 4 meses = $ 48.000 / año

8.2.3 Impacto económico total

Como se explicó anteriormente:

Ingresos adicionales generados = **$ 805.000/año**

Costos adicionales incurridos = $12.000 (Software) + $1.300 (Hardware) + $48.000 (Mano de Obra) = **$ 61.300**

BENEFICIO LOGRADO = $ 805.000 – $ 61.300 = $ 743.000 / año

9 CONCLUSION Y FUTURAS LINEAS DE INVESTIGACION

Se probó que la tasa de éxito utilizando el modelo supera ampliamente la obtenida mediante la vieja estrategia de supuestos y selección aleatoria.

La tasa de éxito se multiplicó por un facto de 2, logrando aumentar las ventas en igual proporción.

Quedando demostrada la eficiencia de esta herramienta mediante el aumento en las ventas de Plazos Fijos, se logró demostrar el poder del Data Mining, el cual puede ser aplicado a muchos otros casos de negocios dentro de la Empresa.

Actualmente se están desarrollando nuevos modelos para los productos Préstamos y Tarjetas de Crédito.

Dado que la metodología a seguir para lograr el modelo es similar para cada proyecto de Data Mining, el procedimiento descripto en el presente trabajo puede ser utilizado para la resolución de los nuevos proyectos a encarar.

Queda demostrado que la “experiencia” de un individuo en cierto campo puede ser reemplazada, y con mejores resultados, por datos históricos, los cuales mediante un análisis apropiado, aportan conocimiento con validez estadística. Un análisis científico riguroso como el aportado por esta técnica logra desentrañar relaciones ocultas hasta para el ojo de los “expertos” y permite tomar decisiones más acertadas que impactan directamente en los ingresos percibidos por la empresa.

Bibliografía

Noticias

- **¿Para qué sirve el "business intelligence"?**

http://www.infobaeprofesional.com/notas/80821-Para-que-sirve-el-business-intelligence.html&cookie

- **Business intelligence o el fin de las empresas "tontas"**

http://tecnologia.infobaeprofesional.com/notas/68301-Business-intelligence-o-el-fin-de-las-empresas-tontas.html?cookie

- **Los beneficios de la inteligencia de negocios**

http://tecnologia.infobaeprofesional.com/notas/51865-Los-beneficios-de-la-inteligencia-de-negocios.html

- **La inteligencia de datos no está madura en la Argentina**

http://tecnologia.infobaeprofesional.com/notas/43472-La-inteligencia-de-datos-no-esta-madura-en-la-Argentina.html

- **Research and Markets: Applied Data Mining for Business and Industry, 2nd Edition**

http://www.tradingmarkets.com/.site/news/Stock%20News/2270751/

- **Biotech companies testify on data mining**

http://www.milforddailynews.com/news/x148103593/Biotech-companies-testify-on-data-mining

Papers

- **Note on Data Mining and BI in Banking Sector**

http://www.iimahd.ernet.in/publications/data/Note%20on%20Data%20Mining%20&%20BI%20in%20Banking%20Sector.pdf

- **C. Olivia Rud. Data Warehousing for Data Mining: A Case Study. Páginas 119-125**
- **Chapman et al., 1999. CRISP-DM 1.0 Step-by-step data mining guide**
- **Ramin Mikaili et al., 2000. Data Mining: An implementation reference guide**
- **Xindong Wu et al., 2008. Top 10 algorithms in data mining**

Libros

- **Alex Berson et al., 1999. Building Data Mining Applications for CRM**
- **Berry and Linoff, 2008. Data Mining Techniques for marketing, sales and customer relationship**