UNIVERSIDAD DE CHILE
FACULTAD DE CIENCIAS FÍSICAS Y MATEMÁTICAS
DEPARTAMENTO DE INGENIERIA INDUSTRIAL

# DESARROLLO DE UN MODELO DE RECOMENDACIÓN DE COMPRA PARA CLIENTES DE UNA EMPRESA DE SEGUROS

**MEMORIA PARA OPTAR AL TITULO DE INGENIERO CIVIL INDUSTRIAL**

**MARCELA DEL PILAR MÉNDEZ CASTRO**

PROFESOR GUIA:
LUIS ABURTO LAFOURCADE

MIEMBROS DE LA COMISIÓN:
RICHARD WEBER HAAS
MANUEL REYES JARA

SANTIAGO DE CHILE
MARZO 2013

RESUMEN DE LA MEMORIA
PARA OPTAR AL TITULO DE
INGENIERO CIVIL INDUSTRIAL
POR: MARCELA MÉNDEZ CASTRO
FECHA: 22/03/2013
PROF. GUIA: SR. LUIS ABURTO L.

# DESARROLLO DE UN MODELO DE RECOMENDACIÓN DE COMPRA PARA CLIENTES DE UNA EMPRESA DE SEGUROS

El trabajo de título que a continuación se presenta, corresponde a la modelación de la probabilidad de aceptación de acciones de marketing directo sobre los clientes de la línea vida de la compañía Consorcio Nacional de Seguros, a quienes se les ofrece la contratación de seguros para automóviles a través del mecanismo de referirlos a los ejecutivos de call center de la compañía.

El proyecto se centra en el desarrollo de una metodología que, en consideración de la data e información disponible para los clientes de la línea vida (con y sin producto automóvil), mejore la efectividad de las campañas de venta cruzada de los seguros de automóvil sobre la cartera de clientes de los productos de vida, a través de la identificación y focalización de recursos sobre los clientes con mayor probabilidad de contratación.

El proceso de construcción del modelo, se basó en la identificación de los datos disponibles y su respectivo procesamiento y selección. Para la selección del modelo definitivo, se realizó pruebas reiterativas en las que se ajustaron arquitecturas y parámetros, y en base a la comparación de sus resultados se seleccionó el modelo final.

El resultado del modelo construido, permite identificar y rankear de manera periódica a los clientes que serán referidos al call center de la empresa, desde donde son contactados para ofrecerles la contratación del seguro de automóvil.

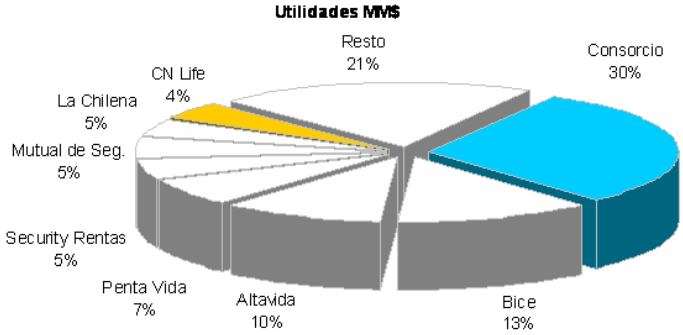
Como recomendación de iniciativas futuras, se plantean las siguientes alternativas: Construir un modelo de estimación de los intervalos de contratación, es decir, un modelo que cuantifique el tiempo que demora un cliente en contratar el seguro de automóvil dado que es un cliente de la línea vida, y/o construir un modelo del flujo inverso, es decir, que cuantifique la probabilidad de contratación de un producto de vida dado que el cliente tiene un seguro de automóvil.

**Tabla de Contenido**

## I. INTRODUCCIÓN

Consorcio Nacional de Seguros S.A. es una de las más importantes compañías de seguros individuales de Chile. Fundada en 1906 bajo el nombre de La Industrial, a lo largo de su historia ha experimentado una serie de fusiones con compañías menores hasta consolidarse en 1976 como Consorcio Nacional de Seguros (CNS). En 1986 fue adquirida por Bankers Trust N.Y.C., una de las más importantes instituciones de la banca mundial. Luego, se incorporaron como socios The Bank of Tokio y Chemical Bank N.Y. Finalmente. En el año 1997 se incorporan el grupo Banvida S.A. (Grupo Fernández León y José Antonio Garcés) y P&S (Grupo Hurtado Vicuña), posicionándose como socios mayoritarios con un 99,76% de la propiedad. Desde entonces, Consorcio ha liderado la industria de seguros tanto por el volumen total de ventas, activos, patrimonio, inversiones y cobertura geográfica. En la actualidad cuenta con más de 20 sucursales a lo largo del país, empleando a un total de cerca de 2000 personas.

Figura nº1: Utilidades del Mercado de Seguros FECU Junio 2007[1]




| Lugar | Compañía | Utilidades MM$ |
|---|---|---|
| 1° | **Consorcio** | **25,720** |
| 2° | Bice | 11,137 |
| 3° | Altavida | 8,988 |
| 4° | Penta Vida | 5,877 |
| 5° | Security Rentas | 4,423 |
| 6° | La Chilena | 4,192 |
| 7° | Mutual de Seg. | 4,004 |
| 8° | CN Life | 3,860 |
| | Resto | 17,755 |

La diversificación de sus actividades comerciales, ha conducido a Consorcio a convertirse en un conglomerado de 7 empresas independientes en términos jurídicos y contables, pero que comparten tanto el backoffice como la administración (RRHH, infraestructura), aprovechando de esta manera todas las sinergias y economías involucradas en la operación. Las empresas que en la actualidad componen Consorcio Financiero son: Consorcio Nacional de Seguros Vida, Consorcio Nacional de Seguros

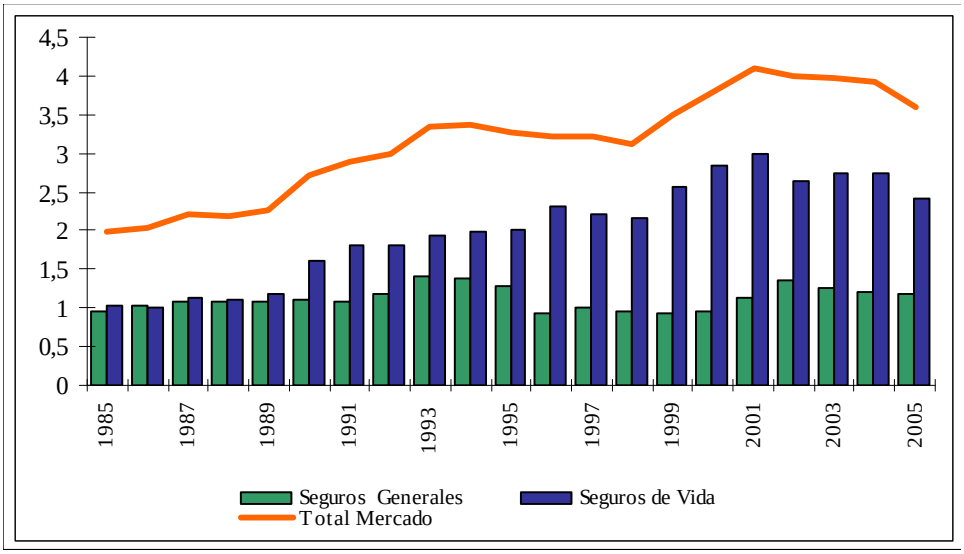
[1] Información consolidada a partir de datos reportados por las compañías aseguradoras a SVS.

Generales, Consorcio Créditos Hipotecarios, Consorcio Corredores de Bolsa, Compañía de Seguros CNLife, Consorcio AGF y Consorcio Administradora de Tarjetas de Crédito.

## I.II. EL MERCADO EN CHILE

El estudio realizado por LIMRA International "Actitudes de Consumidores y Comportamientos de Compra"[1] en 2005, indica que Chile es uno de los países latinoamericanos que más recursos per cápita[2] destina a la contratación de seguros y, que a pesar de esta tendencia (en dinámica variable desde al menos 1985), el mercado local aún no ha alcanzado su estado de madurez. En este sentido, el gran incentivo y desafío que hoy tienen las compañías de seguros dentro de este proceso de evolución en pleno desarrollo, lo constituye crecer y consolidarse a través de la participación de mercado.

Gráfico nº1: Evolución de la prima en Chile como % del PIB[3]




De la gráfica, se observa que tanto para las líneas de vida como generales, hoy el mercado se encuentra por debajo de los peak históricos alcanzados en el horizonte de tiempo comprendido desde 1985 a la fecha. En este sentido, las estrategias comerciales que predominan en el proceso de consolidación en participación de mercado de cada compañía, son básicamente 2:

1) Adquisición de nuevos clientes

2) Rentabilización de cartera cautiva

[1] "Actitudes de Consumidores y Comportamientos de Compra", Lori Chester (Senior Analyst), 2005
[2] Cerca de un 4% del PIB anual, seguido de Argentina con un índice cercano al 2%.
[3] Fuente: www.svs.cl

Es claro que antes de aumentar la participación de mercado vía incorporación de nuevos clientes, está el interesante desafío de rentabilizar de una manera más eficiente a los clientes actuales de las compañías. Existen 2 mecanismos básicos para abordar esta problemática:

a) Up Selling: bajo este foco, el objetivo es incrementar el valor asociado al par Cliente / Producto. En el mercado de los seguros esto es posible al, por ejemplo: incrementar los capitales asegurados en una póliza, adicionar coberturas contratadas a una póliza, etc.

b) Cross Selling: en este caso, el objetivo es incrementar la tenencia de productos asociados a un cliente, es decir, incrementar el factor N° de Pólizas / N° de Clientes. Esta situación, se produce cuando los clientes han encontrado en la oferta de la compañía, un mix que satisface más de una de sus necesidades de manera instantánea (contratación de más de 1 producto en t) o de manera progresiva en el tiempo (contratación de 1 producto en t y sucesivos en ti, con ti>t).

Ambos casos, pueden producirse tanto por acción espontánea de los clientes (al solicitar a su ejecutivo de venta la asesoría cuando el lo estime) como por resultado de la ejecución de acciones de marketing directo. Para este proyecto, una de las consecuencias directas de su aplicación es el aumento en los niveles de cross selling observados dentro de la cartera total de clientes vigentes de Consorcio Financiero, a través de la aplicación de acciones de marketing directo.

## I.III. LOS CLIENTES DE CONSORCIO

Hoy, Consorcio posee cerca de 140.000 clientes, distribuidos de manera heterogénea en sus líneas de negocio.

Tabla n°1: Clientes por línea de Negocio Consorcio Financiero, Junio de 2007

| Línea de Negocio | N° de Clientes |
|---|---|
| Corredora de Bolsa | 4,070 |
| Crédito de Consumo Abierto | 6,587 |
| Crédito de Consumo Pensionados | 17,250 |
| Créditos Hipotecarios | 3,790 |
| Seguros Colectivos | 316 |
| Fondos Mutuos | 3,500 |
| **Seguros Generales** | **30,571** |
| Rentas Vitalicias | 45,893 |
| Tarjeta de Crédito | 9,087 |
| **Seguros de Vida** | **54,824** |
| Suma Directa | 175,888 |
| Rut Unitarios (Total Clientes) | 140,819 |

Línea Vida

Actualmente, Consorcio Nacional de Seguros Vida cuenta con aproximadamente 55.000 clientes vigentes en alguno de sus casi 20 productos[1]. Las opciones que ofrece Consorcio Vida a sus actuales y potenciales clientes son múltiples al momento de evaluar, no obstante es posible identificar 2 tipos de seguros de vida individual:

a) Seguros de Vida Tradicional: Los seguros tradicionales son aquellos que operan del modo en que fueron originalmente pactados, sin permitir que, en el transcurso del período de vigencia y/o según las necesidades del cliente, las condiciones puedan ser modificadas. No incluyen el concepto de ahorro y son seguros de riesgo, es decir, orientados exclusivamente a la protección del (los) asegurado(s) en una póliza. Entre ellos están: Temporales, Vida Entera y Dotales[2].

b) Seguros de Vida Flexibles: La rigidez de los productos de Vida Tradicional, dio origen a la creación de nuevos productos que combinaron Protección y Ahorro. En ellos, a diferencia de los planes tradicionales, se permite al asegurado variar la cantidad y los plazos de pagos de primas, además de aumentar o disminuir el monto asegurado en caso de fallecimiento. Están simultáneamente orientados al ahorro y la protección y se manejan generalmente a través de una cuenta individual donde se acreditan las primas y los intereses, y desde donde se descuenta el costo de seguro y los gastos operacionales. Finalmente, el asegurado puede disponer del monto de esta cuenta individual a la manera de un ahorro. Entre estos se cuentan: Vidactiva y Vidahorro100.

Línea Generales

Análogamente, Consorcio Nacional de Seguros Generales cuenta con más de 30.000 clientes vigentes en alguno de sus 2 productos principales: Seguros para automóviles y para bienes inmuebles. Consorcio Seguros Generales fue la primera empresa de seguros generales de Chile en comercializar productos con identidad propia (nombre comercial) y a través de ejecutivos de venta que forman parte de la compañía. Además de los productos FullCar (25,000 clientes) y FullHouse (12,000 clientes), Consorcio también ofrece el seguro obligatorio de accidentes personales (SOAP), necesario por ley para renovar el permiso de circulación de todo vehículo motorizado.

[1] Considerando los planes base, es decir, sin las variantes de ellos.
[2] Nombres comerciales de los distintos planes de seguros disponibles dentro de la oferta actual de Consorcio.

Cruce entre Líneas

El nivel de cross selling total dentro de la compañía es de 1.48 negocios por cliente, indicador construido de la siguiente forma:

Tabla n°2: Cross Selling Consorcio

| Factores | N° |
|---|---|
| N° Contratos Vigentes | 175,888 |
| N° Clientes Vigentes | 140,819 |
| **CS: Cross Selling** | **1.2490** |

Particularmente, dentro de las líneas Vida y Generales existen 4,866 clientes que actualmente tienen vigente al menos 1 producto de la línea vida y 1 de la línea generales. Estos clientes representan el 15,78% del total de clientes de generales y un 6,72% de los clientes de la línea vida.

## II. CONTEXTO DEL PROYECTO

### II.I. ORIGEN DEL PROYECTO

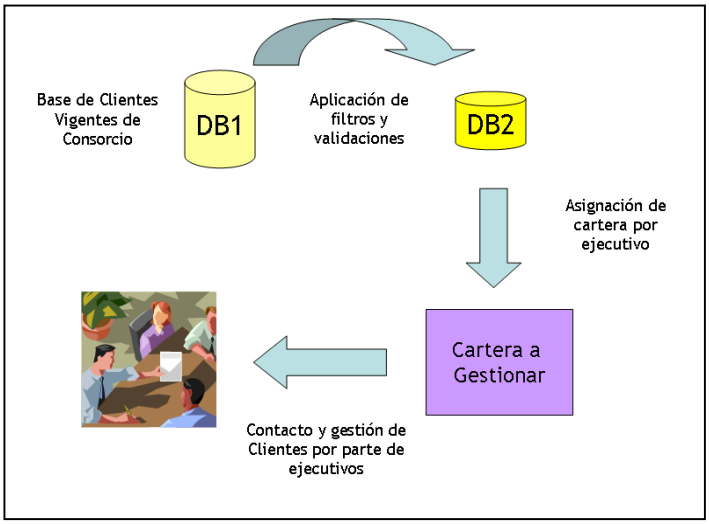
Como una estrategia de creación y consolidación interna de los conceptos de Campañas de Venta e Inteligencia de Negocios (en adelante BI), nace en Marzo de 2006 la subgerencia de Campañas y BI en Consorcio.

Entre sus objetivos centrales, el área de BI debe ser capaz de generar el conocimiento interno en relación a los clientes de consorcio (para todas las aristas relevantes desde el punto de vista comercial) y apoyar desde esta perspectiva, toda decisión en relación a ellos. En ese marco, estudios realizados por el área BI, han cuantificado los bajos niveles de cross selling entre las carteras de productos que actualmente componen la oferta de Consorcio. En adición a lo señalado en el ítem I.III, este es el punto de partida sobre el cual se gesta la iniciativa de desarrollar un modelo predictivo que permita, a través de la cuantificación de aceptación de oferta, determinar que clientes son los mejores prospectos para la aplicación de campañas de marketing directo. Para el desarrollo particular de este proyecto, la modelación busca cuantificar la probabilidad de aceptación de oferta de los clientes de la línea Vida frente a acciones de marketing directo que buscan promover la contratación de seguros automotrices (Línea Generales). No obstante, en el futuro, se contempla la realización de iniciativas similares sobre otros focos orientados igualmente en el aumento del cruce de las carteras de clientes.

## II.II. DEFINICIÓN DEL PROYECTO

El proyecto se define por la construcción de un modelo predictivo de compra del seguro de automóvil por parte de los clientes de la línea Vida de Consorcio frente a una acción de marketing directo. Las acciones de marketing directo de venta en Consorcio, se ejecutan a través de 2 mecanismos: como referidos a la fuerza de ventas, y como referidos a call center. El flujo asociado al proceso de referidos es el que en la figura adjunta se describe:

Figura nº2: Proceso tradicional de referidos




Uno de los focos centrales del proyecto de memoria es diseñar e implementar una metodología de inteligencia de negocios para la determinación eficiente del conjunto DB2, lo que en la práctica significa referir a los ejecutivos de venta y de call center mejores prospectos para campañas, es decir, clientes que cuantitativamente posean una mayor probabilidad[1] de compra del producto objeto de la acción. El desafío del modelo predictivo[2] es determinar tanto en volumen (cuantos) como composición (quienes) de la base de datos DB2.

## II.III. JUSTIFICACIÓN DEL PROYECTO

Hasta el momento de la realización del proyecto, todas las iniciativas comerciales de marketing directo que ha diseñado e implementado la subgerencia, se han basado (esencialmente) en la aplicación de filtros básicos y restricciones técnicas sobre los clientes de acuerdo a las características específicas asociadas a cada campaña. Sin

[1] En función del perfil de los clientes Consorcio.
[2] Complementado al juicio experto e interpretación de los resultados obtenidos.

embargo, existe la latente posibilidad de utilizar la data disponible para la cartera de clientes, para la modelación cuantitativa del problema y definir en base a ella una herramienta de selección.

Adicionalmente, una iniciativa que se ha repetido en al menos 4 oportunidades desde la creación de la subgerencia, es precisamente la que se desea modelar en este proyecto: ofrecer la contratación de seguros para automóviles a los clientes de la línea vida, a través de la gestión de call center. El origen de esta iniciativa encuentra su explicación en las siguientes consideraciones:

a) Un cliente de la línea vida es más rentable[1] que un cliente de la línea de seguros generales. En ese sentido, uno de lo objetivos que se desea lograr es fidelizar a los clientes de la línea Vida mediante la creación de barreras de salida.

b) Es el flujo que se desea activar desde el punto de vista comercial. La cartera de clientes de la línea vida es aproximadamente 2 veces mayor que la de generales, esto permite mayores posibilidades de aprendizaje en términos de implementar diversas fórmulas comerciales (sin agotar la data) que conduzcan a la identificación de configuraciones que en la práctica resulten más exitosas.

## II.IV. OBJETIVOS

Objetivo General

*"Diseñar una metodología para incrementar la efectividad de las campañas de marketing directo enfocadas en potenciar la venta cruzada de seguros de automóvil (línea seguros generales) sobre la cartera vigente de clientes de seguros de vida*"

Objetivos Específicos

- Construir un modelo predictivo que estime la probabilidad de cada cliente frente a la contratación de un seguro automotriz.
- Diseñar e implementar una metodología para la generación periódica de la variable predictiva "probabilidad de compra"[2] para el conjunto de clientes.
- Generar el conjunto de clientes con los mejores prospectos para la campaña comercial de marketing directo de venta de seguro automotriz.

## II.V. ALCANCES DEL PROYECTO

El objetivo central del proyecto es el diseño de una metodología que permita detectar a los clientes con mayor probabilidad de contratación, y en consecuencia, aquellos más proclives a aceptar una recomendación de compra realizada por la

[1] En términos de rentabilidad asociada a los productos. Una póliza de vida es en promedio aproximadamente 3 veces más valiosa que una de generales en términos de valor presente. Fuente: Gerencia Técnica, Consorcio Nacional de Seguros.
[2] Para los clientes vigentes de la línea vida en el producto fullcar.

empresa (acción comercial). En este contexto, es necesario definir la secuencia metodológica que permita implementar desde la construcción del conjunto de registros (clientes y variables), la selección de las variables relevantes y los resultados de la corrida del modelo seleccionado.

Por razones de tiempo, el proyecto no contempla la realización de pilotos o pruebas en ambiente real.

## II.VI. RESULTADOS ESPERADOS

a) Un modelo predictivo, basado en el stock histórico de datos de clientes, que cuantifique la probabilidad que tiene cada uno de ellos para la contratación de un seguro automotriz.

b) Implementación de una metodología analítica que permita la incorporación de mejoras continuas y calibración de los parámetros.

c) Mejorías observables en la tasa de respuesta de las campañas (en relación a las obtenidas hasta ahora).

d) Estimación a priori sobre el aporte de valor asociado a la realización de una campaña de venta cruzada.

Desde el punto de vista de los resultados esperados asociados a la aplicación de la metodología, se espera que el modelo desarrollado constituya una herramienta de apoyo y soporte de las decisiones comerciales asociadas a la realización de campañas, lo cual en la práctica se traduce en:

a) Mejor utilización de los recursos humanos y financieros: al referir mejores prospectos para la gestión de venta.

b) Mayor rentabilización de los clientes para la compañía: ya que incentiva la contratación de productos adicionales en los clientes de la cartera vigente (tiende a incrementar los niveles de cross selling en la cartera).

# III. ENFOQUE METODOLÓGICO

## III.I. INTRODUCCIÓN

El problema general de la predicción de probabilidad de compra se ha resuelto básicamente mediante dos enfoques:

1) Modelos de Clasificación: en general, a través de la modelación con árboles de decisión. Estos modelos se caracterizan por detectar reglas explicativas y de interacción entre las variables para la cuantificación de la variable predictiva. En este caso, como resultado de la aplicación del modelo se caracterizan y

descubren una serie de reglas de interacción entre las variables, que permiten la cuantificación de la variable predictiva.

2) Modelos de Regresión: en los que la variable a predecir puede ser explicada por un conjunto de variables independientes. Estos modelos, generalmente, corresponden a la formalización de una teoría que explica las relaciones subyacentes entre la variable en estudio y el conjunto de variables independientes. En este caso, se utiliza una "función" que es la que se aplica sobre los datos para determinar el valor de la variable predictiva.

Este proyecto tiene su enfoque primario en el desarrollo del modelo basado en árboles de decisión, sin embargo, se contempla la realización de modelos alternativos para la comparación de los resultados obtenidos.

Comparación de enfoques

En el paper "Comparing Adaptive and tradicional techniques for direct marketing"[1], los autores realizaron la comparación de la precisión de los modelos más utilizados para la resolución de problemas de marketing directo. Se compararon Redes neuronales, Árboles y Regresión Logística. Todas técnicas frecuentemente utilizadas para abordar problemas de marketing directo y orientados en la definición del conjunto óptimo de clientes sobre quienes se realizaran las acciones definidas. Los modelos fueron comparados en términos de precisión, interpretabilidad, transparencia y tiempo involucrado en la ejecución del modelo.
Para el estudio, se preparó y preprocesó un conjunto predefinido de variables clásicas de problemas de marketing directo. El resultado fue un conjunto de 135.000 registros (correspondientes a representantes de empresas) asociados a 74 variables más una binaria correspondiente a si respondió o no en acciones pasadas. Las variables independientes se agruparon en 4 grupos: Demográficas (ramo, número de empleados, etc.), Operacionales (cantidad de productos comprados en el pasado, etc.), Situacionales (Productos en que se manifestó interés) y Personales (sexo, edad, cargo en la empresa, etc.). Como es habitual para la aplicación de modelos, la data se particionó en conjuntos de training y test. El conjunto training fue la entrada para la construcción de los modelos mientras que el conjunto test fue usado para la selección del mejor modelo de la técnica usada.
Adicionalmente, se reservó un conjunto de validación al cual se aplicó cada uno de los mejores modelos, de manera de observar diferencias entre los modelos más allá de los resultados obtenidos sobre el conjunto test.

Resultados observados

Desde el punto de vista de la interpretabilidad, es claro que para todas las metodologías desarrolladas es posible realizar interpretaciones de los resultados obtenidos. La regresión logística provee de una lista con variables más relevantes para la predicción además de la importancia relativa entre ellas dentro del modelo (ponderadores). En tanto que CHAID, provee una estructura de árbol de directa

[1] Eiven, Euverman, Peelen, Slisser& Wesseling, 2005.

interpretación y reglas de interacción claras para la predicción en base a las variables, no es posible determinar ni la interacción entre las variables ni su importancia relativa. En términos de tiempos de ejecución sólo se consideraron los tiempos propios de la ejecución computacional, sin contar el tiempo de preparación y diseño del modelo. La cuantificación es la siguiente:

Tabla n°3: Tiempo de ejecución metodología de modelación

| Modelo | Magnitud de Tiempo Ejecución |
|---|---|
| CHAID | Minutos |
| Neural Net | Horas |
| Log Reg | Minutos |

Finalmente, en consideración de:

1) Que es necesario que los resultados sean interpretados por diversos usuarios y/o clientes de Consorcio
2) La experiencia y comprensión del área relativa a la técnica de árboles de decisión[1]
3) Al ajuste que ha demostrado la técnica frente a la naturaleza de las variables disponibles

se desarrollará un modelo fundado en la metodología de Árboles de Decisión. No obstante, se desarrollarán modelos alternativos basados en técnicas distintas, que permitan hacer comparaciones sobre la efectividad de los modelos.

En consideración tanto de los resultados empíricos expuestos como la experiencia y cercanía del área en relación a las metodologías de data mining, se diseño la metodología para abordar el problema de modelación objeto de este proyecto.

## III.II. METODOLOGÍA

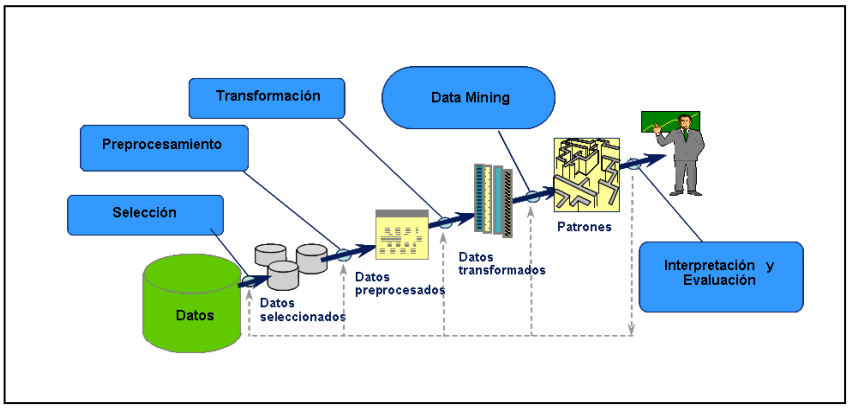
En consideración de que el proyecto busca constituir una herramienta de apoyo a las decisiones comerciales periódicas de la subgerencia de campañas de venta y BI, se diseño una metodología estructurada[2] de forma simple de modo que sea posible capacitar a más de un usuario (Ingeniero de Inteligencia de Negocios) tanto para la generación como para calibraciones futuras de modelo.
Se construirán modelos, con diferentes arquitecturas, en base a la metodología de árboles de decisión, los que se compararan en términos de precisión en la predicción. Los modelos se desarrollarán en software estadístico SPSS y, bajo el prisma del desarrollo de la metodología KDD.

[1] La subgerencia desarrolla y gestiona desde hace aprox. 1 año los modelos de fuga de cliente, basados en árboles de decisión.
[2] En el marco del desarrollo del proyecto se documentará la secuencia de ejecución del modelo.

## III.II.I. ETAPAS DE LA MODELACIÓN

Como marco metodológico base en el desarrollo de la modelación del problema central, se consideró las acciones definidas en un proceso KDD. El proceso KDD fue definido por Fayyad[1] como “un proceso no trivial de identificar patrones, previamente desconocidos, válidos, nuevos, potencialmente útiles y comprensibles dentro de los datos”.

Figura n°3: Proceso KDD




El proceso KDD se define en sus especificaciones dentro de cada problema particular, en función de las características únicas que lo determinan.
El problema de descubrir y extraer el conocimiento implícito en una base de datos, involucra una secuencia de pasos que van desde la manipulación de los datos para su preprocesamiento y transformación, hasta la investigación e inferencia sobre ellos.

Básicamente, la secuencia de trabajo se resume en:

- Integración y recopilación de data: Consiste en la obtención de información sobre el dominio de data disponible (características, atributos, fuentes de datos, etc.). En función de la información recopilada se decide cual será el marco de datos que se utilizará para la construcción del conjunto de datos que será el input inicial del proceso de modelación.

- Selección, limpieza y transformación: Identificación y cuantificación de deficiencias en la calidad de datos como: outliers, missing values.

- Data mining: Etapa cuyo objetivo central es la modelación del problema y la obtención de resultados de predicción / estimación.

- Evaluación e interpretación: Es necesario interpretar y dotar de coherencia a los resultados obtenidos. Es claro que la interpretación se enriquece gracias al feedback de etapas anteriores de la modelación, pero es en la etapa de

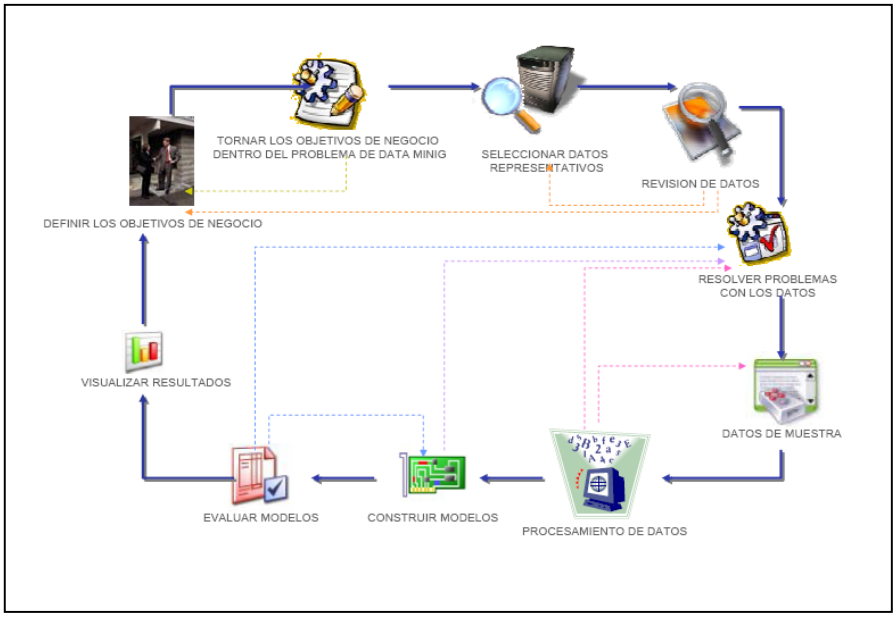
[1] “From data mining to knowledge discovery in databases”; Usama Fayyad, Gregory Piatetski-Shaphiro & Padhraic Smith; 1996.

consolidación de resultados que la interpretabilidad se convierte en el factor fundamental de validación del proceso completo.

- Difusión y uso: Consiste en el implementación y uso práctico de los resultados.

Para el caso de este proyecto se utilizará una versión extendida del proceso KDD, que se caracteriza por explicitar la estrecha relación, interacción y coherencia que debe existir entre el proyecto de data mining y la naturaleza y objetivos del negocio.

Figura n°4: Extensión del Proceso KDD




## III.II.I.I. Objetivos de Negocio

Dentro de la operación de Consorcio Nacional de Seguros, uno de los objetivos centrales desde el punto de vista de este proyecto es la ordenación y selección eficiente de los clientes que serán objeto de las acciones de marketing directo, a fin de obtener mayores utilidades netas como resultado de la implementación de ellas. Este objetivo tiene implícitas consideraciones que trascienden al ámbito de la aplicación de acciones, y se relacionan tanto al uso eficiente de los recursos disponibles (humanos, financieros, infraestructura, etc.) como a una mayor rentabilización de los clientes vigentes. La idea táctica de incrementar la tenencia de productos de los clientes es el core de este proyecto, que busca determinar a quienes y cuando ofrecerles la contratación de un producto adicional en la compañía.

## III.II.I.II. Incorporar objetivos de Negocio dentro del Proyecto de Data Mining

Una vez identificados los objetivos del negocio, se establece la interacción e intercambio entre el modelador y el cliente de proyecto de manera que el modelador capture la esencia del problema y detecte las oportunidades tanto de modelación como enriquecimiento del proyecto. En este caso en particular, el modelador esta inserto dentro de la operación de Consorcio, lo que presenta ventajas y dificultades: es posible acceder al cliente en forma directa y presentar tanto avances como inquietudes, pero restringe la posibilidad de tomar distancia para reenfocar desde distintas perspectivas. Sin embargo, la participación activa de diferentes roles dentro de la empresa, asegura la revisión continua de los supuestos y criterios objetivos utilizados en la realización del proyecto.

### III.II.I.III. Selección de Datos Representativos

Visual Time, la actual base operacional de Consorcio, se encuentra en funcionamiento desde 2004, año en que se ejecutó un proceso migratorio de los datos alojados hasta ese instante en el sistema Ingress, respondiendo a una necesidad creciente de mejorar tanto el funcionamiento como las capacidades de almacenaje de una base de datos en sostenido desarrollo, debido principalmente al crecimiento de tanto de productos como líneas de negocios. En él, se almacenan y relacionan hoy los datos correspondientes a los clientes y negocios para las líneas vida y generales.
Las principales tablas que componen el modelo de datos, se clasifican básicamente en:

*Tablas de Clientes*: Donde se almacenan todos los atributos correspondientes a los Clientes de Consorcio Vida y Generales, donde se entiende por Cliente a toda persona (natural o jurídica) que tuvo o tiene alguna relación con la compañía. En ellas se almacenan los datos de: Clientes, Intermediarios, Proveedores, Empleados, etc.

*Tablas de Negocios*: Donde se almacenan los atributos correspondientes a los negocios y las condiciones contractuales particulares en que se establece la relación entre la compañía y el cliente.

Los atributos de cada tabla se especifican en la sección anexos.

### III.II.I.IV. Selección de datos alineados a objetivos del proyecto

Para la selección de datos que se ajusten a los objetivos del estudio, se consideró 2 condiciones:

*Datos a partir de 2004*: Se considerarán exclusivamente Clientes / Pólizas cuya vigencia sea posterior a Enero de 2004, fecha a partir de la cual todos los nuevos negocios se almacenaron en el nuevo sistema operacional Visual Time. Básicamente, las principales razones que motivan esta decisión son: el universo de negocios que se concentran a partir de esa fecha congrega a más del 60% de los negocios actualmente vigentes y la

incertidumbre de la eventual perdida de datos en el proceso migratorio entre sistemas. Esta premisa, reduce el universo de más de 54.000 clientes vigentes de vida a 37.403.

*Flujo de contratación*: Dentro del universo total de datos correspondientes a los clientes vigentes de la cartera Consorcio, se considerará para la modelación el Conjunto I (Inicial) definido como:

**I = (VA) U (V)**

Donde:

a) Conjunto VA: Todos[1] los clientes naturales Cj que actualmente tienen contratado un seguro de vida (V) y un seguro automotriz (A). El conjunto A, se explica por las consideraciones comerciales que definen el proyecto: es el flujo de contratación que se desea activar y/o potenciar como resultado de la aplicación del modelo predictivo. Dadas las condiciones antes definidas, el conjunto VA contiene 1.201 clientes, los que representan un 3.2% de la base de clientes vida vigentes a partir de enero de 2004.

b) Conjunto V: Todos los clientes que actualmente tienen contratado sólo seguros de vida, los que suman un total de 36.202, conteniendo al 96.8% del total de los clientes de la línea vida.

Tal como está definido, el conjunto I tiene como foco central la caracterización del perfil de los clientes que cumplen con tenencia de ambos productos (seguro de vida y seguro automóvil). El hecho de excluir de I el conjunto de clientes que contrataron el seguro automotriz producto de la aplicación de una campaña de referidos a la gestión de call center, tiene como objetivo independizar al modelo del factor humano: resulta metodológicamente complejo incorporar este factor a la modelación, y adicionalmente, el foco central del proyecto se basa en el perfil de compra, por lo que la ganancia en términos reales de introducir factores adicionales no se justifica.

No obstante, el conjunto C definido como los clientes que contrataron un seguro automotriz producto de campañas implementadas por la subgerencia de campañas de venta y BI, será incluido en el conjunto de testeo del modelo, es decir, se utilizará como conjunto de control para la validación de la predicción del modelo.

### III.II.I.V. Descripción de las variables utilizadas

Las variables disponibles para la modelación, se pueden agrupar básicamente en 3 grupos, en función de su naturaleza: socio demográficas, de comportamiento y de entorno.

[1] Todos menos los clientes que contrataron seguro automotriz producto de una campaña de referidos. Este conjunto será utilizado posteriormente para medir la efectividad de la predicción del modelo.

Variables socio demográficas

Variables correspondientes a características personales propias de los clientes (edad, género y estado civil).

Tabla n°4: Definición de las variables socio demográficas

| Nombre | Descripción |
|---|---|
| Fecha de nacimiento | Fecha (dd/mm/aaaa) de los clientes dueños de las pólizas |
| Género | Género del titular de la póliza |
| Estado Civil | Estado civil del cliente titular de la póliza |

Variables de comportamiento

Son las variables que capturan el comportamiento de los clientes hacia la institución, principalmente es la tenencia o cancelación de otros productos ya sea de vida o generales6, fechas de inicio/termino de la relación contractual, tenencia de coberturas extras asociadas a alguno productos analizados, vía y frecuencia de pago e interacción con la entidad aseguradora.

Tabla n° 5: Definición de las variables de comportamiento del cliente

| Nombre | Descripción |
|---|---|
| Fecha Inicio/ Término | Fecha de inicio de la relación contractual y fecha estipulada (o real) de término de la misma. |
| Coberturas | Tipo de coberturas extras asociadas a cada uno de los productos. |
| Porcentaje de pagos al día | Porcentaje de pagos al día, de acuerdo a la antigüedad, frecuencia de pago y número de pagos de cada póliza |
| Medio de pago | Vía de pago elegida por el cliente para realizar los pagos correspondientes a su póliza. |
| Frecuencia de pago | Frecuencia de tiempo elegida por el cliente para realizar los pagos de las pólizas a su haber. |
| PeopleSoft | Software CRM con el que interactúa el cliente de forma presencial, telefónica o por internet. Se registran todas las interacciones de los clientes en este sistema; la variable usada es si existió interacción (“YES”) o no (“NO”) por parte del cliente con el sistema, en cualquiera de sus vías. |
| Otro producto vigente | Esta variable se refiere a la posesión efectiva de algún otro producto dentro de la empresa que esté en estado VIGENTE. Esta variable incluye a los productos Generales o de Vida, para todo el horizonte de tiempo analizado. Para un mejor análisis de la variable posteriormente, ésta se desglosa en productos de Vida o Generales con un horizonte de tiempo de seis meses (caso que se consideró representativo del estado actual) y para todo el rango de tiempo analizado. |
| Otro producto terminado | Esta variable se refiere a la posesión efectiva de algún otro producto dentro de la empresa que esté en estado TERMINADO. Esta variable incluye a los productos Generales o de Vida, para todo el horizonte de tiempo analizado. Para un mejor análisis de la variable posteriormente se desglosa en productos de Vida o Generales con un horizonte de tiempo de seis meses (caso que se consideró representativo del estado actual) y para todo el rango de tiempo analizado. |

Variables de entorno

Variables asociadas tanto al producto como a la venta del producto y al intermediario que efectuó la venta del mismo.

Tabla n° 6: Definición de las variables de entorno

| Nombre | Descripción |
|---|---|
| Número de póliza | Identificador primario, la predicción se hace en base a la póliza y no al cliente, debido a que la unidad de negocio con que opera la institución aseguradora son las pólizas. |
| Rut | Identificador secundario, es el número de cédula de identidad del cliente, el cual posee un número determinado de pólizas. Se usa principalmente para establecer la relación del cliente con otros productos. |
| Antigüedad | Antigüedad del cliente dueño de la póliza, se expresa en meses debido a que la unidad de tiempo con que trabaja la institución aseguradora son los meses. |
| Agente | Es el símil del agente de cuenta de una institución bancaria para las compañías aseguradoras. La variable usada es el cambio ("YES") o mantención ("NO") del agente a lo largo del tiempo, no importando el número de cambios de agente. |

## Caracterización del conjunto I

a) Estadísticas descriptivas sobre el conjunto I: Las siguientes son las distribuciones porcentuales % comparativas para los conjuntos VA y V, en cada una de las principales variables descriptivas:

Edad

| Rango de Edad | Clientes con FullCar | Clientes sin FullCar |
|---|---|---|
| Rango = 1, Edad <=31 | 18.03% | 20.31% |
| Rango = 2, 32<=Edad<=36 | 25.79% | 23.91% |
| Rango = 3, 37<=Edad<=41 | 20.96% | 18.49% |
| Rango = 4, 42<=Edad<=48 | 14.68% | 18.36% |
| Rango = 5, Edad>=48 | 20.54% | 18.93% |

Para la variable edad, los puntos de corte para cada intervalo fueron construidos utilizando la una funcionalidad específica de la herramienta SPSS, que permite a través de la aplicación de un criterio predeterminado como input, segmentar y categorizar las variables numéricas y así pasar de un conjunto con múltiples valores a un espacio más reducido.

## Género de los Clientes

| Sexo Clientes | Clientes con Fullcar | Clientes sin Fullcar |
|---|---|---|
| 1 (Mujer) | 42.35% | 46.22% |
| 2 (Hombre) | 57.65% | 53.77% |

## Estado civil

| Estado Civil Clientes | Clientes con Fullcar | Clientes sin Fullcar |
|---|---|---|
| 1 (Casado) | 66.04% | 63.88% |
| 2 (Soltero) | 29.77% | 32.20% |
| 3 (Viudo) | 1.47% | 1.00% |
| 4 (Divorciado) | 1.68% | 2.13% |
| 5 (No Informado) | 1.05% | 0.79% |

## Tenencia total de pólizas de la línea vida

| N Total Pólizas Vida | Clientes con Fullcar | Clientes sin Fullcar |
|---|---|---|
| 1 | 69.39% | 82.93% |
| 2 | 23.06% | 13.70% |
| 3 | 5.45% | 2.50% |
| 4 | 2.10% | 0.87% |

## Medio de pago de las pólizas de vida

| Medio de Pago (Vida) | Clientes con Fullcar | Clientes sin Fullcar |
|---|---|---|
| Aviso de cobranza | 3.56% | 4.19% |
| Cuponera | 0.84% | 2.63% |
| Descuento por Planilla | 0.63% | 0.24% |
| PAC | 87.63% | 85.80% |
| Transbank/Diners | 7.34% | 7.13% |

## Prima total mensual (considera todas las pólizas de vida)

| Prima Total Mensual (UF) | Clientes con Fullcar | Clientes sin Fullcar |
|---|---|---|
| 1 (<=1.75) | 29.98% | 33.25% |
| 2 (1.751 - 2.7) | 31.45% | 33.15% |
| 3 (2.71-14.996) | 36.69% | 32.68% |
| 4 (>=14.997) | 1.89% | 0.92% |

## Ciudad de residencia

| Residencia con Fullcar | % | Residencia sin Fullcar | % |
|---|---|---|---|
| 43 (Santiago) | 49.90% | 43 (Santiago) | 47.54% |
| 11 (Concepción) | 7.55% | 11 (Concepción) | 6.98% |
| 52 (Viña del Mar) | 5.03% | 52 (Viña del Mar) | 4.66% |
| 5 (Calama) | 3.98% | 34 (Punta Arenas) | 4.08% |
| 34 (Punta Arenas) | 2.94% | 49 (Valparaíso) | 3.47% |
| 49 (Valparaíso) | 2.94% | 44 (Talca) | 3.34% |
| 9 (Chillán) | 2.73% | 32 (Puerto Montt) | 2.40% |
| 44 (Talca) | 2.73% | 45 (Temuco) | 2.40% |
| 36 (Rancagua) | 2.52% | 9 (Chillán) | 2.29% |
| 45 (Temuco) | 2.52% | 26 (osorno) | 1.98% |
| 20 (Linares) | 2.31% | 2 (Antofagasta) | 1.87% |
| 32 (Puerto Montt) | 2.10% | 19 (La Serena) | 1.73% |

b) Consideraciones de desarrollo: Dado que la cartera de vida congrega más de 1 producto con diferentes características (principalmente en cuanto a la naturaleza del negocio), para esta etapa se consideró el análisis por separación de cartera, las cuales se distribuyen de la siguiente manera:

Tabla n°7: Número de clientes por producto

| Producto | N° de Clientes |
|---|---|
| Accidentes Personales | 875 |
| APV | 8,508 |
| Contigo en Viaje | 769 |
| Dotal | 145 |
| Gastos Médicos | 628 |
| Opción Mayor | 266 |
| Protección Activa | 1,758 |
| SEF | 10,586 |
| Vida Entera | 365 |
| Vida Preferente | 1,430 |
| **Vidactiva** | **7,141** |
| Vidahorro 100 | 4,353 |
| Otros | 579 |
| Total | 37,403 |

El conjunto sobre el cual se desarrollaran los modelos, es la cartera del producto Vidactiva, tanto por el volumen, como por la antigüedad de la cartera (la más antigua del mix de productos de la línea vida). En esta línea, se descarta el producto SEF, que si bien cuenta con un importante número de clientes, es una cartera que actualmente se encuentra en proceso de decrecimiento sostenido debido a que el producto ya no se comercializa. La metodología presentada debe en el futuro ser replicada sobre el resto de los clientes Consorcio, de acuerdo se detecten oportunidades desde el punto de vista comercial.

## Caracterización de los clientes de Vidactiva

Para la etapa preliminar de caracterización, se construyó a través de la técnica disponible en SPSS Conglomerados en 2 fases, una segmentación que permita identificar los principales grupos y sus características dentro del conjunto sobre el cual posteriormente se realizará la predicción. Las siguientes son las principales carácterísticas de los conglomerados detectados y los parámetros utilizados en la segmentación:

Tabla n°8: Parámetros de la segmentación

| Medida | Parámetro utilizado |
|---|---|
| Medida de distancia | Log Verosimilitud |
| N° conglomerados | Máximo 5 |
| Criterio de Conglomeración | Bayesiano de Schwarz (BIC) |

En tanto que las variables utilizadas fueron:

Tabla n°9: Variables de la segmentación

| Variables |
|---|
| Prima promedio mensual Vidactiva (cat) |
| Edad (cat) |
| Ciudad (cat) |
| N Total productos vida |
| Medio de pago (1° póliza) |
| Estado Civil (cat) |
| Sexo (cat) |

Como resultado de esta etapa de detección de subgrupos relevantes dentro de la cartera del producto Vidactiva, se obtuvo lo siguiente:

**Distribución de conglomerados**

| | | N | % de combinados | % del total |
|---|---|---|---|---|
| Conglomerado | 1 | 2438 | 36.6% | 36.5% |
| | 2 | 2207 | 33.1% | 33.1% |
| | 3 | 2025 | 30.4% | 30.3% |
| | Combinados | 6670 | 100.0% | 99.9% |
| Casos excluidos | | 6 | | .1% |
| Total | | 6676 | | 100.0% |

Donde cada uno de los conglomerados se caracteriza principalmente por:

Prima prom mens Vidactiva (Categorizada)

| | | <= 1.750 | | 1.751 - 2.700 | | 2.701 - 14.996 | | 14.997+ | |
|---|---|---|---|---|---|---|---|---|---|
| | | Frecuencia | Porcentaje | Frecuencia | Porcentaje | Frecuencia | Porcentaje | Frecuencia | Porcentaje |
| Conglomerado | 1 | 1084 | 49.3% | 1321 | 59.9% | 0 | .0% | 33 | 50.0% |
| | 2 | 31 | 1.4% | 160 | 7.3% | 1986 | 90.3% | 30 | 45.5% |
| | 3 | 1084 | 49.3% | 724 | 32.8% | 214 | 9.7% | 3 | 4.5% |
| | Combinados | 2199 | 100.0% | 2205 | 100.0% | 2200 | 100.0% | 66 | 100.0% |

edad (Categorizada)

| | | <= 31 | | 32 - 36 | | 37 - 41 | | 42 - 48 | | 49+ | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| | | Frecuencia | Porcentaje | Frecuencia | Porcentaje | Frecuencia | Porcentaje | Frecuencia | Porcentaje | Frecuencia | Porcentaje |
| Conglomerado | 1 | 350 | 26.0% | 734 | 45.8% | 636 | 51.1% | 439 | 36.4% | 279 | 22.0% |
| | 2 | 62 | 4.6% | 280 | 17.5% | 321 | 25.8% | 621 | 51.4% | 923 | 72.6% |
| | 3 | 932 | 69.3% | 590 | 36.8% | 287 | 23.1% | 147 | 12.2% | 69 | 5.4% |
| | Combinados | 1344 | 100.0% | 1604 | 100.0% | 1244 | 100.0% | 1207 | 100.0% | 1271 | 100.0% |

MEDIOPAGO_Nombre

| | | Aviso de cobranza | | Cuponera | | Descuento por Planilla | | PAC | | Transbank/Diners | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| | | Frecuencia | Porcentaje | Frecuencia | Porcentaje | Frecuencia | Porcentaje | Frecuencia | Porcentaje | Frecuencia | Porcentaje |
| Conglomerado | 1 | 114 | 41.3% | 62 | 37.1% | 7 | 38.9% | 2094 | 36.5% | 161 | 33.8% |
| | 2 | 78 | 28.3% | 56 | 33.5% | 5 | 27.8% | 1869 | 32.6% | 199 | 41.7% |
| | 3 | 84 | 30.4% | 49 | 29.3% | 6 | 33.3% | 1769 | 30.9% | 117 | 24.5% |
| | Combinados | 276 | 100.0% | 167 | 100.0% | 18 | 100.0% | 5732 | 100.0% | 477 | 100.0% |

CLIENTES_Estado Civil

| | | 1 | | 2 | | 3 | | 4 | | 5 | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| | | Frecuencia | Porcentaje | Frecuencia | Porcentaje | Frecuencia | Porcentaje | Frecuencia | Porcentaje | Frecuencia | Porcentaje |
| Conglomerado | 1 | 2358 | 55.2% | 0 | .0% | 17 | 24.6% | 32 | 22.9% | 31 | 64.6% |
| | 2 | 1917 | 44.8% | 167 | 7.8% | 46 | 66.7% | 60 | 42.9% | 17 | 35.4% |
| | 3 | 0 | .0% | 1971 | 92.2% | 6 | 8.7% | 48 | 34.3% | 0 | .0% |
| | Combinados | 4275 | 100.0% | 2138 | 100.0% | 69 | 100.0% | 140 | 100.0% | 48 | 100.0% |

CLIENTES_Sexo

| | | 1 | | 2 | |
|---|---|---|---|---|---|
| | | Frecuencia | Porcentaje | Frecuencia | Porcentaje |
| Conglomerado | 1 | 1179 | 38.5% | 1259 | 34.9% |
| | 2 | 768 | 25.1% | 1439 | 39.9% |
| | 3 | 1117 | 36.5% | 908 | 25.2% |
| | Combinados | 3064 | 100.0% | 3606 | 100.0% |

En resumen y considerando la diferencia % como característica principal de los grupos, se observa:

- ✓ Conglomerado 1: Hombres, casados, que utilizan preferentemente el medio de pago PAC (pago automático de cuentas), tienen entre 32 y 36 años y pagan una prima mensual en el rango UF 1.751 y UF 2.7.
- ✓ Conglomerado 2: Hombres, casados, que utilizan el medio de pago PAC, tienen más de 49 años y pagan una prima mensual entre UF 2.701 y UF 14.996.
- ✓ Conglomerado 3: Mujeres, solteras, que tienen hasta 31 años, que utilizan medio de pago PAC y que pagan una prima mensual hasta UF 1.751.

III.II.I.VI. Revisión de la data y resolución de problemas[1]

Preprocesamiento y transformación de la data

La manera en que los datos se encuentran en una base de datos, generalmente no representa de forma limpia o eficiente la información disponible. Para sortear esta dificultad es necesario preprocesar los datos y corregir errores e inconsistencias detectadas dentro de la base.
Para la realización de este proceso, se tomará como base las mejores prácticas de la minería de datos.

Etapa de preprocesamiento

El objetivo central del preprocesamiento de las variables es minimizar o eliminar las inconsistencias dentro de la base de datos utilizada para la modelación. Los principales focos del preprocesamiento son:

- Missing values: Dado que la predicción se realizará sobre los clientes de la línea vida y, se tendrá como entrada las variables disponibles para esa línea de negocios, no se tienen valores perdidos para las variables relacionadas a los clientes (como sexo, edad) ya que son variables fundamentales en la evaluación e ingreso de los negocios de esa línea. Sólo se encontraron valores perdidos en ciudad (51) los cuales fueron completados por la moda de la variable (Santiago).

- Cantidad de variables utilizadas en el modelo: En consideración del limitado número de variables disponibles para la modelación, no se excluirán a priori ninguna de ellas. De manera posterior, se evaluará la importancia relativa de las mismas dentro del modelo resultante.

Etapa de transformación

Entre las transformaciones aplicadas a la data, se encuentran:

1) Edad del cliente: Variable numérica a partir de la fecha de nacimiento del cliente (dd/mm/aaaa) se construyó la variable edad del cliente de la siguiente forma:

*Edad (años)= 2007 – Año (DBirth_Date)*

[1] Bibliografía Principal: "The Data Ware House ETL Toolkit"; Ralph Kimball & Joe Caserta.

2) Antigüedad de la póliza Vida (primera): Variable numérica que determina la antigüedad del cliente en la línea vida, a partir de la fecha de inicio de vigencia de la póliza más antigua vigente:

*Antigüedad (meses)= Redondear((DateHoy – DDate_Orig)/365.25*12,0)*

3) Número total de pólizas vida (N Total vida): Tenencia total de pólizas vida hoy para cada cliente Cj:

*N Total vida (Cj)= Cuenta (Npolicy)*

4) Prima total mensual: Variable numérica que adiciona todas las primas en productos vida que el cliente Cj desembolsa de manera mensual

*Prima total mensual (Cj) = Suma (prima i), donde i = 1,...,n con n= número de pólizas vigentes*

5) Intervalos óptimos para la variable Prima total mensual: En la herramienta SPSS, se selecciona la intervalación óptima para una variable. En este caso, se seleccionó que los intervalos fueran homogéneos en términos de casos, con lo que se obtuvo la siguiente agrupación:

Tabla n°10: Intervalos de primas

| Prima Total Mensual (UF) | Intervalo |
|---|---|
| prima <=1.75 | 1 |
| 1.751 <=prima<= 2.7 | 2 |
| 2.71 <=prima<= 14.996 | 3 |
| prima >=14.997 | 4 |

6) Coberturas adicionales en la póliza: A cada tipo de cobertura adicional (distinta a fallecimiento) se asocia una variable binaria Yes / No, de acuerdo a si está (Yes) o no (No) asociada a la póliza del cliente.

7) Interacciones con CRM Peoplesoft: Variable binaria Yes / No, asociada a si el cliente (rut) tiene o no registro de interacciones con la plataforma Peoplesoft.

### III.II.I.VII. Técnica de Modelación

Se modeló el problema con 3 técnicas: Árboles de decisión (como técnica central) y (como técnicas comparativas) Naive Bayes y Support Vector Machines, variando de manera sistemática (grillas) su arquitectura. La idea de recorrer e implementar variaciones de una misma técnica de modelación, busca principalmente:

⇒ Seleccionar aquel que resulte más apropiado para la predicción (en términos de error, ajuste, performance, interpretabilidad) para la predicción

⇒ Extraer la mayor cantidad de información de cada modelo desarrollado a fin de aprender de cada uno de ellos.

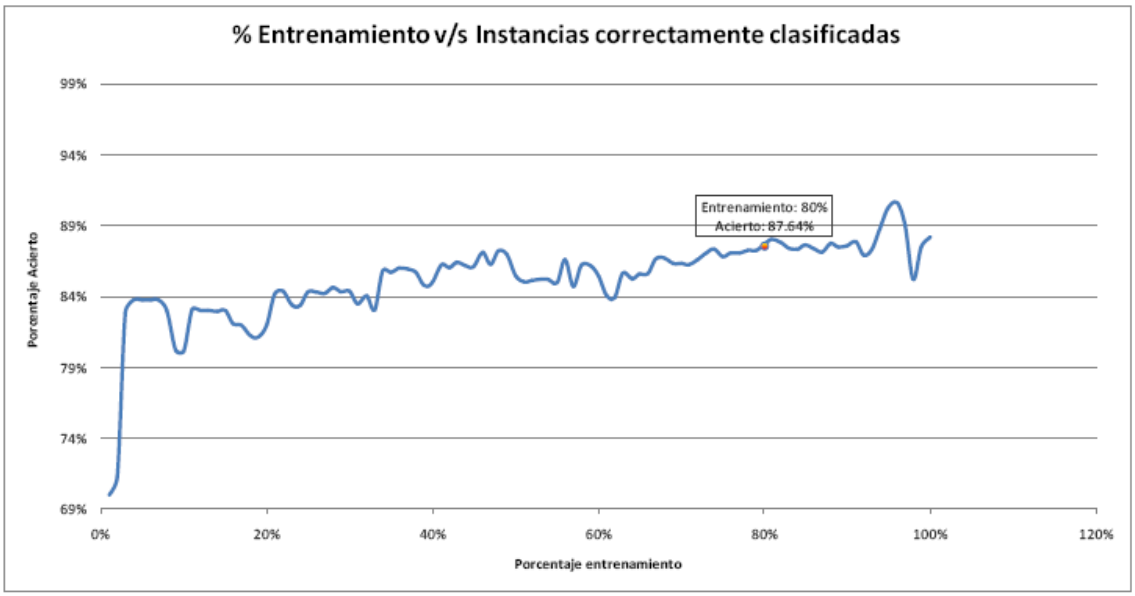
Adicionalmente, se debe considerar los costos y complejidad tanto del preprocesamiento como de la solución final y se seleccionará aquella que reporte la mejor combinación costo / beneficio a los roles usuarios del proyecto final.

Porcentaje de entrenamiento y validación

> "En muchos casos, la función que define la curva de precisión puede adecuarse bien al comportamiento del algoritmo y del conjunto de datos, permitiendo obtener los parámetros que la describe" (Ricardo Blanco Vega (2006), "Extracción y contextualización de reglas comprensibles a partir de modelos de "caja negra").

Sobre la base de la idea anterior, y la evidencia de la literatura, se definió las muestras como 80% training y 20% test. En la gráfica adjunta, de % de entrenamiento v/s instancias correctamente clasificadas, se puede visualizar la estabilidad que la curva alcanza en torno al 80% de training.

Gráfico n° 2: Entrenamiento v/s clasificación de instancias




Instancias por hojas

Se utilizará un número variable de instancias por hoja, ligado a la búsqueda de la mejor arquitectura.

Arquitectura eficiente

Dado que para cada modelo, se requiere del input de arquitecturas preestablecidas por el modelador, el problema de determinación de la arquitectura eficiente será abordado de acuerdo al recorrido de Grillas. Esta técnica consiste en la conformación de una "grilla" de puntos definidos por configuraciones factibles de arquitecturas para cada uno de los modelos. La determinación de la arquitectura que mejor modela el problema, se obtiene como resultado del recorrido de la grilla y la comparación de los resultados de modelación que de cada configuración probada se desprenden. Una vez que constituidos los modelos con la respectiva arquitectura, se determinará cual de los modelos provee de una mejor solución para el problema central.

Sobre ajuste en los modelos

Con el objetivo de evitar el sobre ajuste del modelo, se utilizó cross validation en la corrida en SPSS de cada arquitectura.

III.II.I.VIII. Interpretación y evaluación de resultados

Resultados del modelo

Como se mencionó anteriormente, la distribución utilizada para training/test fue 80%/20%. A partir de la salida del modelo, se construyen indicadores que permiten cuantificar el grado de ajuste y/o precisión de la predicción del modelo. El conjunto de indicadores[1] que permitirán comparar tanto la efectividad de la predicción como el ajuste del modelo al problema, son los siguientes:

$$Sensibilidad = \frac{\text{M(Clientes compra)/R(Clientes compra)}}{\text{N°Real Clientes compra}}$$

$$Especificidad = \frac{\text{M(Clientes No compra)/R(Clientes No compra)}}{\text{N°Real Clientes No compra}}$$

$$Exactitud = \frac{\text{Sensibilidad} + \text{Especificidad}}{2}$$

$$\text{Precisión} = \frac{\text{M(Clientes compra)} + \text{R(Clientes compra)}}{\text{M (Clientes compra)}}$$

[1] Fuente: Han, J., Lamber, M. "Data mining: Concepts and techniques".

Para los datos de la modelación a través de árboles de decisión, se obtuvieron las siguientes mediciones para cada indicador:

Tabla n°11: Valorización de indicadores de ajuste

| | Indicadores de ajuste de los modelos | | | |
|---|---|---|---|---|
| Modelo | % Exactitud | % Sensibilidad | % Especificidad | % Precisión |
| 1 | 62 | 56 | 68 | 67 |
| 2 | 63 | 55 | 70 | 67 |
| 3 | 65 | 55 | 75 | 61 |
| 4 | 68 | 60 | 75 | 69 |
| 5 | 59 | 52 | 65 | 67 |
| 6 | 71 | 74 | 67 | 69 |
| 7 | 67 | 55 | 79 | 69 |
| 8 | 65 | 61 | 69 | 65 |
| 9 | 67 | 56 | 77 | 64 |
| 10 | 67 | 54 | 80 | 64 |

En función de la mejor combinación de indicadores, se estableció que el modelo 6, caracterizado por el vector de indicadores (Ex, S, Es, P)= (70, 61, 77, 75) es la mejor combinación definida por la siguiente arquitectura subyacente:

**Resumen del modelo**

| | | |
|---|---|---|
| Especificaciones | Método de crecimiento | CRT |
| | Variable dependiente | TieneFC |
| | Variables independientes | edad_rec, Primamenscateg, Mediopago_rec, Ciudad_rec, NVidactiva, CLIENTES_Sexo, CLIENTES_EstadoCivil |
| | Validación | Validación cruzada |
| | Máxima profundidad de árbol | 8 |
| | Mínimo de casos en un nodo filial | 10 |
| | Mínimo de casos en un nodo parental | 5 |
| Resultados | Variables independientes incluidas | CLIENTES_Sexo, CLIENTES_EstadoCivil, Primamenscateg, edad_rec, Mediopago_rec, NVidactiva, Ciudad_rec |
| | Número de nodos | 41 |
| | Número de nodos terminales | 21 |
| | Profundidad | 8 |

Para la elección del modelo que mejor se ajusta al problema, se priorizó el nivel del indicador “% Precisión” ya que es el que mejor permite cuantificar los casos que posteriormente se convertirán en contratación del producto, y en ese sentido, permite cuantificar a priori el valor aportado por la campaña a realizar.

Matriz de confusión

Una matriz de confusión es una herramienta de visualización comúnmente usada en el ámbito de inteligencia artificial. Cada columna de la matriz representa las instancias pertenecientes a una clase pronosticada mientras que las filas representan a las instancias pertenecientes a la clase real. La matriz de confusión permite visualizar fácilmente los porcentajes de acierto y desacierto en la predicción de cada una de las clases. La siguiente es la matriz de confusión asociada al modelo de la mejor arquitectura:

Tabla n°12: Matriz de confusión

| Condición Real / Predicción | No contrata | Contrata |
|---|---|---|
| No contrata | 350 | 129 |
| Contrata | 165 | 370 |

En base a la matriz, es posible determinar el acierto del modelo en cada clase:

$$ClaseContrata = \frac{(370)}{(370+129)} = 74.14\%$$

$$ClaseNoContrata = \frac{(350)}{(350+165)} = 67.96\%$$

Y el acierto global del modelo:

$$AciertoClases = \frac{(350+370)}{(350+165+129+370)} = 71\%$$

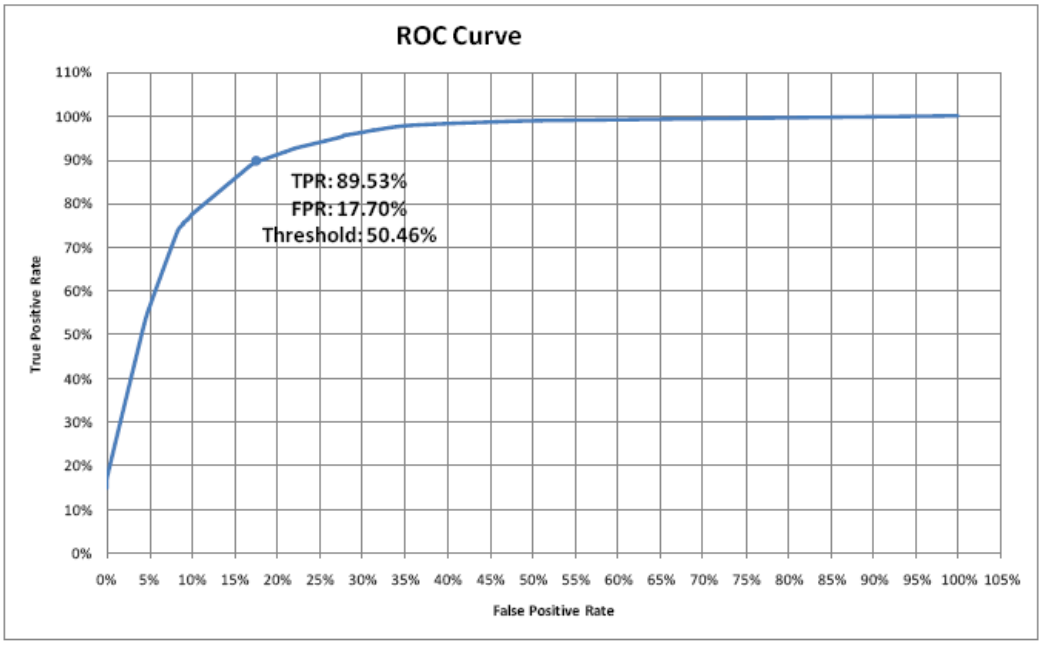
Análisis de sensibilidad sobre el modelo de árboles

Curva ROC

El análisis de sensibilidad se realizó mediante las curvas ROC (Receiver Operating Characteristic). Estas curvas, se utilizan frecuentemente para problemas de 2 clases, y permiten seleccionar el subconjunto de posibilidades de clasificación que tiene un comportamiento óptimo para todas las posibilidades de costo. En este caso, se definió como doblemente costo predecir erróneamente un caso que en la práctica correspondía a una contratación y el modelo lo clasifico como no compra.

El análisis ROC se relaciona de forma directa con el análisis de costo/beneficio en el diagnóstico de la toma de decisiones. En el modelo presentado las predicciones se hacen en base a un 50% de probabilidad, esto quiere decir que sí una instancia tiene probabilidad mayor o igual al 50% de CONTRATAR el seguro de automóvil, el algoritmo clasifica la instancia como SI COMPRA, lo mismo ocurre en el caso contrario. Al variar el porcentaje con que se predice si una instancia pertenece a una clase u otra obtenemos los tradeoffs en la predicción. La adjunta es la curva ROC obtenida para la contratación del seguro de automóvil, sobre la cual se justifica el uso de 50% de probabilidad como clasificación positiva de compra. Los índices utilizados en el análisis son el True Positive Rate (TPR) y False Positive Rate (FPR).

Gráfico n°3: Curva ROC




Considerando predicción de compra positiva para cualquier probabilidad mayor o igual a 50% (esp. 50.46%) se llega a los resultados presentados en la figura anterior.

Cabe destacar que el óptimo se obtiene para un Threshold de 49.83% obteniendo un TPR de 89.54% y un FPR de 17.7%.

Variables relevantes dentro del modelo

Se realizo un ranking de los atributos en base al ReliefF Ranking (disponible en el software estadístico WEKA). Este algoritmo, se funda en la técnica **nearest neighbour** (del vecino más cercano) asignando un peso relativo a cada atributo. El peso de cada atributo modifica en función de la habilidad para distinguir entre los valores de la variable clase (compra, no compra). El ranking así construido muestra en qué grado influye cada una de las variables utilizadas (de forma independiente) sobre la intención de compra o no compra de cada cliente. Los resultados son los siguientes:

Tabla n°13: Ranking de variables

| Peso | Atributo |
|---|---|
| 9.95% | Prima (cat) |
| 7.48% | N° Pólizas |
| 6.33% | Antigüedad |
| 4.40% | Edad |
| 3.20% | Estado Civil |
| 3.05% | Sexo |
| 2.95% | Vía de Pago |
| 1.77% | Porcentaje Pagos |
| 1.61% | Otro Vigente Vida |
| 1.03% | Otro Vigente General |
| 1.03% | PeopleSoft |

Comparación con otros modelos

La técnica de árboles de decisión junto a Support Vector Machines (SVM) y Naive Bayes son frecuentemente utilizados en la resolución de este tipo de problemas[1] de dos clases. Con el fin de realizar una comparación válida, se desarrollaron modelos[2] para los algoritmos alternativos en consideración de la misma base de matriz de costos y composición training / test. Los resultados, en base a TPR/FPR para la clase compra, se presentan a continuación:

Comparación algoritmos predicción compra

| | CRT | Naive Bayes | SVM |
|---|---|---|---|
| TPR | 89.53% | 91.07% | 90.11% |
| FPR | 17.70% | 32.37% | 25.97% |

[1] Katharina Morik and Hanna Kopcke (2004), "Analysing custumer churn in Insurance data"

[2] Modelos desarrollados en el software estadístico WEKA.

Comparación algoritmos clasificación total

| | CRT | Naive Bayes | SVM |
|---|---|---|---|
| Correctamente clasificadas | 71.00% | 62.32% | 77.28% |
| Incorrectamente clasificadas | 29.00% | 37.68% | 22.72% |

El algoritmo utilizado en el modelo es notoriamente superior en el porcentaje de clasificación total, respecto del modelo Naive Bayes. Mientras que es un 6% inferior a SVM. Para la clase compra, el TPR es aproximadamente un 1% inferior al resto de los modelos, pero se logra un FPR 8% menor que el algoritmo más cercano (SVM).

### III.II.I.IX. Output de la modelación

La tabla adjunta es una gráfica de presentación para el resultado final producto del output del modelo de árboles. El objetivo de gestión detrás de la entrega del output es focalizar los esfuerzos comerciales (traducidos en la aplicación de campañas de venta telefónica) sobre aquellos clientes que presentan una mayor probabilidad de contratación del producto.

Tabla n°14: Output del modelo, probabilidad de contratación

| Cliente | Predicción de Probabilidad de compra |
|---|---|
| Cliente 1 | 10.35% |
| Cliente 2 | 54.21% |
| Cliente 3 | 92.65% |
| … | |
| … | |
| Cliente k | 57.11% |
| … | |

## IV. CONCLUSIONES

El objetivo inicial de este trabajo de titulo, fue la construcción de un modelo de propensión de compra para clientes de la línea vida de Consorcio sobre la línea generales. La consecución de este objetivo se a bordo de acuerdo al enfoque KDD de minería de datos, dentro del cual, se consideró el desarrollo de diferentes alternativas de técnicas de modelación a fin de comparar los resultados obtenidos. El modelo final, correspondiente al modelo basado en árboles de decisión, fue elegido tanto por el ajuste de sus resultados como por decisión de la empresa.

Respecto de los datos utilizados en el modelo final, aunque los atributos de entrada para la modelación fueron todos los que se tenía a disposición, de la selección e importancia relativa de atributos dentro del modelo se obtuvieron las siguientes variables relevantes que caracterizan la propensión de compra:

- Nivel de prima mensual que pagan los clientes
- N° de pólizas contratadas
- Antigüedad en la compañía
- Edad del cliente
- Estado civil del cliente
- Sexo del cliente
- Medio de pago elegido por el cliente

El modelo final para la predicción es un árbol de decisión, construido en el software estadístico SPSS, que se caracteriza por la siguiente arquitectura:

- 8 niveles de profundidad
- 41 nodos
- Mínimo elementos en nodo parental: 10
- Mínimo elementos en nodo filial: 5

Este modelo, se caracteriza por el siguiente vector de indicadores (Exactitud, Sensibilidad, Especificidad, Precisión)= (70%, 61%, 77%, 75%).

Si bien se obtienen mejores resultados globales de predicción para el modelo alternativo SVM, se opta por el modelo de árboles desarrollado en SPSS por la facilidad y familiaridad de uso que tiene el área de BI de la empresa tanto con la herramienta como con la técnica.

Con el output obtenido de la modelación, se rankeo el conjunto de clientes enviado a la gestión de call center de la empresa. En función de este ranking (desde el cliente más propenso al menos propenso) los ejecutivos de call center focalizarán su gestión privilegiando en términos de tiempo y dedicación.

Para la mantención del modelo desarrollado, se recomienda incorporar los resultados de las campañas comerciales que la compañía realice, a través tanto de la evaluación de la correlación entre los resultados del modelo y la conducta real de los

clientes (probabilidad del cliente v/s contratación real) como la incorporación de los mismos para ajustar la arquitectura del modelo.

Como iniciativas futuras, que complementen la efectividad del modelo y fortalezcan su uso comercial, se plantean las siguientes alternativas:

- Construir un modelo de estimación de los intervalos de contratación, es decir, un modelo que cuantifique el tiempo que demora un cliente en contratar el seguro de automóvil dado que es un cliente de la línea vida.

- Construir un modelo de flujo inverso, es decir, un modelo que cuantifique la probabilidad de contratación de un producto de vida

## V. REFERENCIAS BIBLIOGRÁFICAS Y ELECTRÓNICAS

### Referencias Documentales: Libros, Publicaciones y Memorias de título


ABRAHAM, M. y LODISH, L. 1993. “An implemented system for improving promotion productivity using store scanner data”. Information Resource Inc., the Wharton School, University of Pennsylvania.

ACUÑA, E. y RODRIGUEZ, A. 2004. “The treatment of missing values and its effect in the classifier accuracy”. University of Puerto Rico at Mayaguez, Puerto Rico.

ART, MARY. 2007. “Using data to drive sales: database marketing company practices”. LIMRA.

BOX, G. E. P, JENKINS, G. M., y REINSEL, G.C. 1994. “Time Series Analysis: Forecasting and Control”, Englewood Cliffs, NJ: Prentice Hall, Inc.

CARVAJAL, JP. 2002. “Desarrollo de un modelo de propensión de compra para una empresa de comunicaciones”.

DIAZ, A. y PAULO, A. 2002. “Introducción de tecnologías de inteligencia de negocios al proceso de ventas del área residencial de Telefónica CTC Chile”.

EIBEN, AE., EUVERMAN, W., PEELEN, E., SLISSER, F., WESSELING, J. 1996. “Comparing Adaptive and traditional techniques for direct marketing”. 4th European Congress on Intelligent Techniques and Soft Computing, Verlag Mainz.

GUYON, I. y ELISSEEFF, A. 2003. “An Introduction to Variable and Feature Selection”. MIT Press. Cambridge, MA, USA.

HAN, J., LAMBER, M. 2001, “Data mining: Concepts and techniques”. Morgan Kaufman Publishers, San Francisco.

KIMBALL, R. y CASERTA, J. 2004. “The Data Warehouse ETL Toolkit: Practical Techniques for Extracting, Cleaning, Conforming and Delivering Data”. John Wiley & Sons. 416p.

RUIZ, M. 2004. “Identificación de clientes en una empresa de telecomunicaciones con alta probabilidad de realizar venta cruzada utilizando técnicas de data mining”.

**Otros**

Apuntes del curso IN60E, Profesor Richard Weber, semestre Otoño 2005.
Apuntes del curso IN72O, Profesor Richard Weber, semestre Primavera 2005.
Página de la Superintendecia de Valores y Seguros: www.svs.cl
Página de Consorcio Nacional de Seguros: www.consorcio.cl

## VI. ANEXOS

## ANEXO A: DEFINICIÓN DE SEGURO[1]

Conceptos Básicos

Un seguro, es un contrato mediante el cual el asegurado, a través del pago de una prima al asegurador, adquiere el derecho a recibir de este una prestación que le indemnice ante daños o hechos que puedan producirse dentro de los límites determinados en el contrato.

Uno de los pilares fundamentales del la actividad aseguradora es el mutualismo. El seguro es un sistema en que un grupo de individuos suman aportes para la formación de un fondo único, cuya finalidad de suplir, en determinado momento, necesidades eventuales de algunos de sus miembros afectados por un acontecimiento imprevisto.

El principio básico de funcionamiento de un seguro es que para las personas resulta razonable, frente a la posibilidad de una pérdida o daño de gran magnitud, optar por la certeza de una pérdida menor a cambio de un pago convenido.

Los elementos comunes a todo seguro son:

a) El riesgo: Es una amenaza de pérdida o deterioro que afecta a bienes determinados o a derechos específicos, la vida, la salud o la integridad física e intelectual de un individuo. La noción de riesgo está ligada a un bien, derecho o condición física o intelectual y representa la posibilidad de que algo dañe este bien, derecho o persona.
b) La cosa asegurable: Es el elemento sujeto a riesgo. Pueden ser: cosas corporales (bienes materiales) o incorporales (derechos), la vida o integridad física de una persona.
c) La prima: Es el precio del seguro y lo que el asegurado debe pagar por asegurar contra el riesgo a la cosa asegurable.
d) La póliza: Es el documento físico que ostenta las condiciones del contrato de seguro. La ley vigente establece que este documento debe contener: Datos de asegurado y asegurador (nombres, domicilios, etc.), declaración de la calidad que toma el contratante (asegurado o contratante para 3° personas), designación clara del valor y naturaleza de los bienes asegurados, los riesgos que el asegurador asume (incendio, muerte, enfermedad, robo, etc.), instante de tiempo en que comienza y termina el riesgo para el asegurador, prima (tiempo, lugar y forma de pago), fecha de celebración del contrato, etc.
e) La indemnización: Es la suma de dinero que el asegurador se obliga a pagar al asegurado en caso de la ocurrencia de un siniestro y según las condiciones señaladas en la póliza. Representa la compensación por la pérdida sufrida y nunca puede ser para el asegurado oportunidad de ganancia.

[1] Basado en “Manual Autoinstructivo de Productos Consorcio”, Gerencia de Recursos Humanos Consorcio Financiero, 2006.

Clasificación de los Seguros

Los seguros y las compañías que los comercializan, se clasifican en 2 grandes grupos:

a) Grupo I – Seguros Generales: Son aquellos seguros que cubren riesgos de incendio, marítimos, de transporte terrestre y demás, que aseguran la reparación de daños causados por acontecimientos que puedan o no ocurrir.
b) Grupo II – Seguros de Vida: Cubren riesgos de vida u otros que aseguren al tenedor de la póliza un capital, una póliza saldada, o una renta para sí o su(s) beneficiario(s). En caso de muerte del asegurado, el capital es a favor de un (o unos) beneficiarios(s).

En Chile, por disposiciones legales, ninguna compañía aseguradora puede realizar en forma conjunta el negocio de seguros del Grupo I y II bajo una misma razón social. Es por esto que las empresas aseguradoras ofrecen productos de ambos grupos a través de empresas con razones sociales diferentes.

## ANEXO B: TABLAS DEL MODELOS DE DATOS

Las tablas que serán fuente central para este estudio y sus respectivos atributos principales son:

1) Tabla POLICY: Atributos de la Póliza

| POLICY | |
|---|---|
| DDATE_ORIGI | Fecha de efecto (inicio de vigencia) original de la póliza/certificado |
| DEXPIRDAT | Fecha de vencimiento (fin de vigencia) de la póliza |
| DNULLDATE | Fecha de anulación de la póliza/certificado |
| NAGENCY | Código de la agencia--Valores posibles según tabla 5555 |
| NBRANCH | Código del ramo comercial.--Valores posibles según tabla 10 |
| NCAPITAL | Monto de capital asegurado de la póliza/certificado |
| NINTERMED | Código del intermedario principal de la póliza |
| NOFFICE | Código de la Sucursal. Valores según Table9 |
| NOFFICEAGEN | Código de la oficina--Valores posibles según tabla 5556 |
| NPAYFREQ | Frecuencia de pago de la prima.--Valores únicos según tabla 36 |
| NPOLICY | Número identificativo de la póliza/ cotización/ solicitud |
| NPREMIUM | Monto de prima anual de la póliza |
| NPRODUCT | Código del producto |
| NPROPONUM | Número de la propuesta o cotización que da origen a la póliza |
| SCERTYPE | Tipo de registro--Valores únicos-- 1- Propuesta-- 2- póliza-- etc. |
| SCLIENT | Código que identifica al cliente titular de los recibos de la póliza o certificado |
| SSTATUS_POL | Estado de la póliza/certificado.--Valores únicos según tabla 181 |

2) Tabla CLIENT: Atributos de Clientes

| CLIENT | |
|---|---|
| DBIRTHDAT | Fecha de nacimiento del cliente |
| SLASTNAME | Apellido paterno del cliente.Sólo se indica para personas naturales |
| SCLIENAME | Nombre del cliente |
| SFIRSTNAME | Nombre(s) del cliente. Sólo se indica para personas naturales |
| SCLIENT | Código que identifica al cliente |
| NCIVILSTA | Estado civil del cliente.--Valores únicos según tabla 14 |
| NOFFICE | Código de la sucursal. Valores posibles según table 9 |
| SSEXCLIEN | Sexo del cliente. Valores únicos según tabla 18 |

3) Tabla ADDRESS: Atributos de direcciones

| ADDRESS | |
|---|---|
| NBRANCH | Código del ramo comercial.--Valores posibles según tabla 10 |
| NOFFICE | Código de la sucursal--Valores posibles según table 9 |
| NPROVINCE | Código de la regisn--Valores posibles según tabla de regiones (Province) |
| NPOLICY | Número identificativo de la póliza/ cotización/ propuesta |
| SDEPARTMENT | Número del departamento |
| NPRODUCT | Código del producto |
| DNULLDATE | Fecha de anulación del registro |
| NLOCAL | Código de la ciudad.--Valores posibles según tabla de ciudades (tab_locat) |
| NFLOOR | Número de piso |
| SZONE | Ciudad, localidad, provincia |
| SCLIENT | Código que identifica al cliente |
| SCERTYPE | Tipo de registro--Valores únicos-- 1- Propuesta-- 2- póliza-- etc. |

4) Tabla COVER: Atributos de Coberturas de las Pólizas

| COVER | |
|---|---|
| DCOMPDATE | Fecha del computador en que se crea o actualiza el registro |
| DNULLDATE | Fecha de anulación del registro |
| NBRANCH | Código del ramo comercial.--Valores posibles según tabla 10 |
| NCAPITAL | Monto de capital asegurado de la cobertura |
| NCAPITALI | Monto de capital asegurado inicial de la cobertura |
| NCOVER | Código de la cobertura |
| NCURRENCY | Código de la moneda.--Valores reservados según tabla 11 |
| NPOLICY | Número identificativo de la póliza/ cotización/ solicitud |
| NPREMIUM | Monto de prima anual de la cobertura |
| NPRODUCT | Código del producto |
| NROLE | Figura con la que actúa el cliente en la póliza.--Valores únicos según tabla 12 |
| SCERTYPE | Tipo de registro--Valores únicos-- 1- Propuesta-- 2- Póliza-- etc. |
| SCLIENT | Código que identifica al cliente |