Escuela de Negocios

Maestría en Dirección de Empresas

Tesis

La importancia de la explotación de datos para la correcta tarificación en las Compañías de Seguros

EMBA 2013 - Abril 2015

Karina Yabra
Tutor: Mariano Pérez

Sólo aquellos que han transitado
tormentas pueden sobrevivir huracanes.

Anónimo

AGRADECIMIENTOS

Quiero agradecerle a Leonardo, mi compañero de toda la vida, por haber estado a mi lado a lo largo de la maestría y haberse ocupado de nuestras hijas Victoria y Julieta cuando las horas de estudio, los innumerables trabajos prácticos, las lecturas y los tiempos de cursada no me permitían estar con ellos. Asimismo, quiero agradecerle profundamente por creer siempre en mí y por el apoyo incondicional que me brinda en cada proyecto que decido emprender.

Victoria y Julieta, ustedes son la luz que ilumina mi vida y hacen que cada día trate de ser la mejor versión de mí misma.

RESUMEN

Como su título lo indica, esta Tesis abordó la importancia de la explotación de datos para la correcta tarificación en las Compañías de Seguros. En tal sentido, en muchas oportunidades las Compañías de Seguros cuentan con un caudal abundante de información que no analizan, perdiéndose la posibilidad de encontrar patrones de interés que aporten conocimiento a la empresa.

En el caso particular del presente trabajo, se pretendió demostrar que mediante la aplicación de técnicas de minería de datos las aseguradoras podrían descubrir cuáles son los factores de riesgo que subyacen a su operatoria y, de esta manera, por un lado tarificar de manera más precisa y, por otro, definir y llevar adelante planes de acción para acotar las variables críticas logrando de este modo garantizar un resultado positivo para la Compañía de Seguros.

Son parte integrante de este trabajo el brindar un panorama sobre el seguro, algunos conceptos técnicos importantes de la industria aseguradora, se menciona luego la teoría relativa a la minería de datos para luego abordar un modelo matemático, en el cual se ha desarrollado la técnica de Árboles de Decisión, y donde se prueba la importancia de la explotación de datos para la tarificación en las Compañías de Seguros.

PALABRAS CLAVE

Explotación de datos, Tarificación, Compañías de Seguros, Minería de datos, Aseguradoras, Factores de Riesgo, Seguro, Árbol de Decisión.

Tabla de Contenidos

INTRODUCCIÓN

Tal como menciona Centeno et al. (2011), en los últimos años, se ha producido un gran crecimiento en nuestras capacidades de generar y colectar datos, debido básicamente al gran poder de procesamiento de las máquinas como al bajo costo de almacenamiento de los mismos.

Sin embargo, dentro de estas enormes masas de datos existe una gran cantidad de información oculta, de gran importancia estratégica, a la que no se puede acceder por las técnicas clásicas de recuperación de la información.

El descubrimiento de esta información oculta es posible gracias a la Minería de Datos (*Data Mining*), que entre otras sofisticadas técnicas aplica la inteligencia artificial para encontrar patrones y relaciones dentro de los datos permitiendo la creación de modelos, es decir, representaciones abstractas de la realidad. Pero es el descubrimiento del conocimiento (KDD, por sus siglas en inglés), que se encarga de la preparación de los datos y la interpretación de los resultados obtenidos, el que permite dar un significado a estos patrones encontrados.

Así el valor real de los datos reside en la información que se puede extraer de ellos, información que ayuda a tomar decisiones o mejorar nuestra comprensión de los fenómenos que nos rodean. Hoy, más que nunca, los métodos analíticos avanzados son el arma secreta de muchos negocios exitosos.

Utilizando métodos analíticos avanzados para la explotación de datos, los negocios incrementan sus ganancias, maximizan la eficiencia operativa, reducen costos y mejoran la satisfacción del cliente.

Este comportamiento no es ajeno a las compañías de seguros que cuentan con muchos datos de sus asegurados y potenciales asegurados y en muchos casos no saben qué hacer con ellos.

Frente a este hecho, esta tesis pretende describir cómo mediante la explotación de datos las compañías de seguros pueden identificar los factores de riesgo más significativos y, mediante ello, tarificar las distintas coberturas que comercializan de manera más precisa. Esto último, permitirá por un lado que los asegurados paguen una prima de seguro más acorde con su perfil de riesgo y por otro que las compañías de seguros puedan llevar adelante ciertos planes de acción para garantizar ganancias dentro de la organización.

Objetivo de la Tesis

Objetivo General e Hipótesis

El objetivo del presente trabajo de investigación es mostrar la factibilidad del uso de la Minería de Datos para tarificar las distintas coberturas que comercializan las Compañías de Seguros de manera más precisa identificando los factores de riesgo más significativos.

Específicamente, la hipótesis del presente trabajo es que mediante la aplicación de técnicas de minería de datos las aseguradoras pueden descubrir cuáles son los factores de riesgo que subyacen a su operatoria y, de esta manera, por un lado tarificar de manera más precisa y, por otro, definir y llevar adelante planes de acción para acotar las variables críticas logrando de este modo garantizar un resultado positivo para la Compañía de Seguros.

Objetivos Específicos

- Construir un modelo de minería de datos para una compañía de seguros líder en el país en comercialización de coberturas para equipos de tecnología móvil el cuál permita identificar las variables críticas que impactan en los siniestros a pagar,

- Mediante dicho modelo, estimar de manera precisa las primas que deberían abonar los distintos asegurados y determinar la brecha existente entre estos valores y los actualmente pagados,

- Identificar los cursos de acción que deberá llevar adelante la compañía de seguros para tener acotados los factores de riesgo que se hayan descubierto.

Alcance y limitaciones

El presente trabajo de investigación ha sido enfocado hacia el sector asegurador argentino; en particular se ha elegido a una importante aseguradora del país para construir el modelo de Minería de Datos.

Organización del Trabajo

El presente trabajo está organizado en nueve capítulos, los cuales son detallados a continuación:

En el capítulo 1 "El Seguro" se detalla qué es el seguro, las características que tienen los seguros y sus conceptos fundamentales.

En el capítulo 2 "La tarificación como elemento central de las compañías de seguro" se ofrece un panorama sobre los conceptos básicos asociados a la tarifación, y se mencionan algunos métodos usuales para tarifar.

En el capítulo 3 "Minería de Datos" se detalla qué es la minería de datos, los pasos que componen el proceso de extracción de conocimientos, las técnicas de minería de datos y cómo actualmente se aplica la minería de datos en compañías de seguros.

En el capítulo 4 "La empresa bajo estudio y su producto afectado" se presenta una descripción de la compañía de seguros que se va a analizar, la cual sirvió como base para construir el modelo para tarificar. En esta sección también se mencionan las características principales de los productos de seguros que fueron utilizados para construir el modelo.

El capítulo 5 "Preparación de los Datos" contiene el proceso de preparación de datos necesario para construir el modelo que se utilizará para tarificar. En él se muestra la selección del objetivo, el pre-proceso de datos y la transformación de los mismos así como un detalle de las bases de datos que se utilizaron en el análisis.

El capítulo 6 "Construcción del Modelo" como su nombre lo indica detalla las características del modelo que se desarrolló para llevar adelante el presente trabajo de investigación.

El capítulo 7 describe los resultados obtenidos del análisis de tarificación así como un análisis de sensibilidad en el que, una vez identificados los factores de riesgo, se examina el impacto que podría llegar a tener en los resultados la implementación de ciertos planes de acción.

Por último, en los siguientes dos capítulos se encuentran las conclusiones derivadas del presente trabajo, así como la información bibliográfica utilizada en el mismo.

Capítulo 1. EL SEGURO

Tal como detalla Osorio González, el seguro, ya sea bajo la forma de sociedades mutualistas o sociedades comerciales, ocupa parte importante de la vida económica moderna. Sin embargo, poco se conoce, en el ámbito general, de sus fundamentos teóricos y principios de funcionamiento.

La idea fundamental del Seguro reside en el hecho de que se crean asociaciones con la sola finalidad de cubrir conjuntamente pérdidas por sucesos determinados. En la actualidad la institución del seguro se funda en la mutualidad y la estadística. El estudio del seguro requiere de la estadística, pues el riesgo asumido debe tener cierta frecuencia con relación al conjunto de asegurados, respecto al cual es fundamental que el siniestro aparezca con la mayor certeza posible, aunque para el asegurado subsista la incertidumbre de la ocurrencia del siniestro. Dicho siniestro no puede ser frecuente pues se pierde la incertidumbre, ni raro porque no puede ser estudiado. La estadística solo puede ser eficaz para hechos que se repiten con regularidad, porque el azar observado en grandes masas, obedece a la Ley de los grandes números.

Definición del Seguro

El seguro es una actividad esencialmente económica, cuya finalidad es cubrir, mediante el concurso mutuo de todos los integrantes del mismo, la parte del costo social de la producción representada por la ocurrencia de siniestros individuales aleatorios, pero estadísticamente mensurables y predecibles para el conjunto.

Como institución, el seguro es un sistema de protección del hombre y de su patrimonio frente a diversos hechos que amenazan su integridad, su vida, su interés y su propiedad. Los hechos nocivos que causan pérdidas o daños son inciertos pero previsibles. El seguro garantiza el resarcimiento de un capital para reparar o cubrir la pérdida o daño que aparezca en cualquier momento, recibiendo como contraprestación un precio por adelantado por el servicio de protección que ofrece.

Es un hecho que el seguro, sea del tipo que sea, se basa en que cada asegurado del colectivo acuerde con la compañía pagar una cierta cantidad de dinero, conocida como prima, de manera que la suma de todas las primas recogidas se emplean para pagar los posibles siniestros ocasionados por el conjunto de los asegurados durante el período de tiempo pactado. De esta forma, el asegurado evita el desembolso de lo que constituirá la cuantía de sus siniestros, evidentemente desconocida, sustituyéndola por el pago de esa cantidad de dinero, la prima.

Lo deseable sería que el colectivo de asegurados que comparten sus gastos pagando una misma prima, sea lo más homogéneo posible en cuanto a sus riesgos se refiere, pues si no fuera así, los asegurados que generan más gastos a la compañía debería pagar primas mayores. En caso contrario, podría darse una selección desfavorable de los asegurados que ponen en peligro la solvencia de la empresa. Por ello, resulta esencial que la compañía

aseguradora sea capaz de clasificar a sus clientes del modo más homogéneo posible atendiendo al riesgo, de manera que los asegurados pertenecientes a un mismo grupo paguen idéntica prima.

Características del Seguro

Las características que tienen los seguros son las siguientes:

1. Económica: El seguro como institución económica, implica un conjunto de transferencia de valores, con arreglo a los siguientes factores:
    - Presencia de un conjunto de riesgos que, combinados entre sí, permiten compensar las pérdidas de unos cuantos con los aportes de la totalidad de los miembros del conjunto.
    - Existencia de un plan de seguro que garantice cierta continuidad.
    - Capacidad legal del individuo a recibir las prestaciones prometidas por el asegurador (vínculo legal entre asegurado y asegurador).
    - Onerosidad de las prestaciones.

2. Necesidad: El seguro cubre una necesidad del asegurado, reparación de un daño, satisfacción de una pérdida o de pago fortuito.

3. Mutualidad: Es la concurrencia de la comunidad amenazada por los riesgos, a la composición de las pérdidas ocurridas. El seguro es una especie de "fondo común" administrado por el asegurador en el que cada asegurado aporta una suma proporcional al riesgo que introduce.

4. Aleatoriedad: Los hechos que originan la prestación del asegurador deben ser, con respecto a cada individualidad, fortuitos y aleatorios, por lo menos al momento de su realización o de su conocimiento.

5. Tasabilidad en dinero: La pérdida probable ha de ser mensurable en dinero, en consecuencia, el seguro no responde de las consecuencias puramente morales o políticas de la pérdida.

6. Analogía de riesgos: Para determinar el volumen y valor de los riesgos, es necesario que estos tengan cierta homogeneidad, tanto cuantitativa como cualitativa. Esto significa que no hay seguros a la medida y al gusto del cliente. Para lograr la homogeneidad, la póliza juega un papel importante, pues, al fijar las condiciones de aseguramiento, minimiza las diferencias concretas de los riesgos y permite hacerlos más o menos similares.

Elementos Específicos

Son elementos específicos del seguro: el riesgo, la prima y la indemnización.

- El riesgo es el elemento aleatorio del seguro, definido como la incertidumbre de la pérdida económica
- La prima es el precio del seguro
- La indemnización es el pago efectuado por el asegurador cuando la pérdida se hace efectiva y exigible en los términos convenidos

Riesgo

En la terminología aseguradora, se emplea el concepto de riesgo para interpretar dos ideas diferentes: de una forma, riesgo como objeto asegurado; y de otra, riesgo como posible ocurrencia por azar de un acontecimiento que produce una necesidad económica y cuya aparición real o existencia se previene y garantiza en la póliza y obliga al asegurador a efectuar la prestación o indemnización que le corresponde.

Las características esenciales del riesgo, para ser objeto del seguro, son las siguientes:

a) Incierto y aleatorio
b) Posible
c) Concreto
d) Lícito
e) Fortuito
f) Contenido económico

Prima

Es el precio del contrato del seguro sin los impuestos. Llamada también prima comercial, es la prima pura más los recargos para la administración o gestión del seguro y un margen de utilidad para la empresa. Los recargos son los gastos de adquisición, formados básicamente por la comisión de intermediación que se paga al agente o corredor de seguro; los gastos de administración, que vienen a ser los gastos en que incurre el asegurador para el manejo de la cartera de seguros, como son sueldos y gastos generales de gestión; los recargos asignados a la utilidad razonable del asegurador, llamado también margen de beneficio.

Indemnización

El siniestro es la realización del riesgo. La indemnización es el pago que efectúa el asegurador al asegurado como consecuencia del siniestro. Sin embargo, en el lenguaje contable y estadístico de los seguros se suele utilizar la palabra siniestro como sinónimo de indemnización.

Tratándose de seguros de daños, el propósito del seguro es restituir al asegurado a su situación patrimonial inmediatamente anterior al momento del siniestro. De ahí que el seguro no pueda ser fuente de lucro para el asegurado.

La indemnización asume tres formas básicas:

a) Reparación del bien siniestrado

b) Reemplazo del bien siniestrado por otro similar

c) Entrega al asegurado de una suma de dinero equivalente a la pérdida sufrida.

Algunos Conceptos Técnicos dentro del Seguro

Frecuencia Siniestral (*Frequency*)

La Frecuencia Siniestral es la probabilidad de que ocurra un siniestro en un determinado período de tiempo. Suele referirse a la cantidad de siniestros que ocurrieron en un período de tiempo en relación al número de asegurados expuestos a riesgo en ese mismo período.

Severidad o Intensidad (*Severity*)

La Severidad o la intensidad es el valor esperado del costo de los siniestros.

Muchas veces dicha severidad está determinada por la Suma Asegurada contratada. En otros casos, la severidad depende del valor de reparación o valor de reemplazo del objeto siniestrado.

Prima Pura (*Risk Premium*)

La Prima Pura es la porción del precio final que paga el cliente que se destina al pago de los siniestros.

La Prima Pura se calcula como el producto de la frecuencia siniestral y la severidad.

Representa la unidad más simple y básica del concepto de prima, por cuanto significa el coste real del riesgo asumido por el asegurador, sin tener en cuenta sus gastos de gestión.

Prima Ganada o Devengada (*Earned Premium*)

La prima ganada o devengada por una compañía de seguros en un período de tiempo es la porción de la prima que la compañía de seguro ha ganado o dicho de otro modo le pertenece a ésta. Dicha prima ganada se calcula en base al ratio del tiempo transcurrido.

Ratio Siniestral (*Loss Ratio*)

El ratio Siniestral o Loss Ratio es el ratio entre el monto de los siniestros incurridos en un período, esto es, los siniestros pagados y reservados y la prima ganada en el período.

Por ejemplo, si una compañía de seguros abona $80 de siniestros por cada $150 de prima luego el ratio siniestral es del 53%.

Cabe aclarar que un 1% de caída o de incremento en el Loss Ratio puede hacer que la compañía de seguros gane o pierda mucho dinero por lo que es muy importante monitorear este indicador.

Descomposición del precio que abona el cliente

El precio que abona un cliente a la compañía de seguros se llama Premio. Dicho premio se descompone de la siguiente forma:

$$Premio = Prima\ de\ Tarifa + Impuestos$$

Siendo:

$$Prima\ de\ Tarifa = Prima\ Pura + Comisiones + Gastos + Margen\ de\ Beneficio$$

Capítulo 2. LA TARIFICACIÓN COMO ELEMENTO CENTRAL DE LAS COMPAÑÍAS DE SEGUROS

Tal como desarrolla Lancheros (2011), el precio de un bien o servicio será aquel que permita cubrir los costos y provea un margen de utilidad a quien lo ofrece y sea aceptado por quien lo demanda. Para la mayoría de los bienes y servicios, el precio se puede establecer de manera directa, dado que el costo de producción es conocido, sin embargo para los seguros esto último no es cierto.

Los seguros pueden ser definidos como un contrato entre el tomador de la póliza y la compañía de seguros, mediante el cual la compañía se obliga, a cambio de una suma de dinero (prima), al pago a un tercero (beneficiario) de una cantidad de dinero que usualmente tiene un límite (valor asegurado), si determinados eventos ocurren dentro de un período específico de tiempo. Dado que la ocurrencia de los eventos que dan origen a un pago por parte de la compañía es incierta, el costo final de una póliza no puede ser conocido en el momento de la venta, lo cual hace que el proceso de establecer el precio que debe pagar el tomador sea más complejo que establecer el precio para otro tipo de productos. En este proceso deben tenerse en cuenta aspectos técnicos, regulatorios, económicos, etc.

Algunos principios fundamentales que deben tenerse en cuenta a la hora de calcular las tarifas son los siguientes:

- Las tarifas deben ser suficientes para cubrir los costos de las reclamaciones más los gastos y proveer un margen de utilidad para las compañías de seguros.

- Las tarifas deben estar directamente relacionadas con el riesgo, esto es a mayor riesgo mayor tarifa.

- Las tarifas deben ser el producto de la utilización de información estadística que cumpla exigencias de homogeneidad y representatividad.

Para poder cumplir con estos principios es necesario tener en cuenta la naturaleza de los diferentes tipos de seguros y los diferentes riesgos asociados a éstos.

En la tarificación moderna, son fundamentales las leyes de los grandes números. Estas leyes establecen que a medida que el número de observaciones independientes de una población aumenta, la media de la muestra se acerca cada vez más a la media de la población, lo cual permite que por medio de la recolección, compilación y análisis de volúmenes grandes de datos, los actuarios puedan desarrollar interpretaciones probabilísticas más precisas sobre las pérdidas de las aseguradoras.

Conceptos básicos asociados a la Tarificación

Primas

La prima es el monto que el tomador paga por la cobertura del seguro. Las primas pueden ser emitidas, devengadas o ganadas, no devengadas o en vigencia.

Las primas emitidas corresponden a las primas asociadas con pólizas que fueron emitidas en un período específico de tiempo. Las primas devengadas son aquellas que la compañía ha ganado durante un período determinado, pues ya ha cubierto el riesgo. Las primas no devengadas corresponden a la porción de la prima que aún no ha sido ganada por la aseguradora. Las primas en vigencia corresponden al valor de las primas emitidas, para todas las pólizas que se encuentran vigentes en un momento específico de tiempo.

Expuestos

Un expuesto es la unidad básica de riesgo que mide la exposición a una pérdida.

De manera similar a las primas, los expuestos pueden clasificarse en expuestos emitidos, devengados, no devengados o en vigencia.

Los expuestos emitidos corresponden al número de expuestos asociados con pólizas emitidas en un período específico de tiempo. Los expuestos devengados corresponden a los expuestos asociados a la porción de las pólizas que ya transcurrió. Los expuestos no devengados corresponden a la diferencia entre los expuestos emitidos y los expuestos devengados y los expuestos en vigencia corresponden al número de expuestos asociados a las pólizas que se encuentran vigentes en una fecha determinada.

Cuando las pólizas tengan una vigencia menor a la vigencia anual, los expuestos deben ser ajustados consecuentemente. En el caso de que la información no esté desagregada, sino que por ejemplo se tengan los expuestos mensuales, usualmente se asume que las pólizas se emiten a mitad de período.

Siniestros y Reclamaciones

Un siniestro es la materialización de un riesgo cubierto por la póliza. Una reclamación es la solicitud de pago que hace un asegurado o beneficiario tras la ocurrencia de un siniestro.

Debido a que usualmente los siniestros no son reportados inmediatamente, es necesario mantener una reserva para los siniestros que ya ocurrieron pero aún no han sido reportados y para aquellos que habiendo sido reportados no han sido suficientemente reservados. Esta reserva es conocida como reserva de IBNR (Incurred but not reported). Las técnicas para el cálculo de esta reserva quedan fuera del alcance del presente trabajo.

Pérdidas

Las pérdidas corresponden al valor que las compañías deben cancelar por concepto de siniestros.

Valor Asegurado

El valor asegurado corresponde al monto máximo pagadero en caso de siniestro, previamente estipulado en las condiciones de la póliza.

Gastos

Las compañías incurren en gastos para el ajuste y la liquidación de siniestros, así como para la emisión y mantenimiento de las pólizas. Estos gastos pueden ser fijos (no dependen del monto de las primas) o variables. Usualmente suele usarse información histórica para su estimación.

Gastos de Ajuste de Siniestros

Los gastos de ajuste de siniestros son aquellos en los que debe incurrir la compañía para el ajuste y la liquidación de los siniestros (Costos del departamento de indemnizaciones, honorarios jurídicos, etc.). Pueden dividirse en gastos asignables (se pueden asignar directamente a un siniestro) o gastos no asignables.

Gastos de Suscripción

Los gastos de suscripción son gastos en que debe incurrir la compañía para la emisión y mantenimiento de las pólizas.

Gastos de Adquisición

Son los gastos en que incurre la compañía para la emisión de las pólizas. El principal rubro dentro de los gastos de adquisición son las comisiones a intermediarios.

Impuestos y Licencias

Son los impuestos que abonan los asegurados dentro del premio y los impuestos que pagan las compañías de seguros.

Gastos Generales

Estos son gastos que aunque no hacen parte del negocio, son necesarios para el funcionamiento de las compañías, tales como salarios, servicios públicos, alquiler de la oficina, etc.

Margen de Riesgo y Utilidad

Dado que las compañías deben mantener requerimientos de capital para respaldar los riesgos que asumen, los accionistas esperan recibir una utilidad.

La utilidad en el negocio de seguros proviene de dos fuentes: el resultado operacional (ingresos menos egresos) y los ingresos generados por la inversión de las reservas.

Adicionalmente, las pérdidas pueden ser mayores a las pérdidas esperadas y por esta razón es necesario establecer un margen de riesgo para esas desviaciones adversas. Usualmente este margen se escoge de manera que la probabilidad de que las pérdidas superen su valor esperado sea pequeña.

La ecuación fundamental del seguro

Para un producto cualquiera fuera de seguros, la ecuación que describe el precio sería la siguiente:

$$Precio = Costo + Utilidad$$

En la actividad aseguradora, los costos provienen de las pérdidas, los gastos asociados a las reclamaciones, los gastos de adquisición y mantenimiento de las pólizas y otros gastos que aunque no hacen parte de la actividad aseguradora son necesarios para el funcionamiento de la compañía. La utilidad de las compañías corresponde a la diferencia entre los ingresos y los egresos provenientes de la suscripción de pólizas más el producto por inversiones. Dado que la prima es el precio que se paga por un seguro, en virtud de la ecuación anterior, tendríamos la siguiente relación:

$$Prima = Pérdidas + Gastos\ de\ Ajuste\ de\ Siniestros + Gastos\ de\ Suscripción + Utilidad$$

El propósito de la tarificación consiste en asegurar que la anterior ecuación, llamada usualmente "ecuación fundamental del seguro", esté balanceada. Esto es, que las primas sean suficientes para cubrir todos los costos asociados con la transferencia del riesgo y se obtenga un cierto margen de utilidad.

Para lograr dicho objetivo, es necesario que la tarificación sea prospectiva, esto es, que refleje las condiciones del momento en que las tarifas se van a aplicar, y que la ecuación esté balanceada a nivel agregado y a nivel individual.

En el proceso de tarificación usualmente tiende a utilizarse información histórica para estimar los costos futuros que serán utilizados en el cálculo de las primas.

Sin embargo, cuando se usa la experiencia histórica deben hacerse ajustes, a fin de que esta experiencia sea representativa de la experiencia futura, dado que existen diversos factores que afectan los componentes de la ecuación fundamental del seguro, tales como

cambios operacionales, presiones inflacionarias, cambios en el portafolio y cambios regulatorios.

El equilibrio debe darse a nivel agregado y también a nivel de segmento y a nivel individual. El equilibrio a nivel agregado se refiere a que el total de primas debe ser suficiente para cubrir los costos y alcanzar el margen de utilidad esperado, sin que las primas sean excesivas, pues esto traería desventajas competitivas.

El equilibrio a nivel de segmento y a nivel individual está relacionado con la suficiencia y la equidad de las tasas, lo que implica que la tarifa debe ser proporcional al riesgo. El balance a nivel de segmento y a nivel individual es muy importante, pues en el caso de que las tarifas no sean equitativas podría presentarse un fenómeno conocido como selección adversa, que consiste en que los buenos riesgos se retiran de la compañía por considerar que las tarifas son excesivas y los malos riesgos se quedan, lo cual lleva a una espiral de aumentos de tarifas y costos que puede volverse insostenible si no se corrige la situación.

Con el fin de evitar este problema, las compañías utilizan diferentes variables, conocidas como variables de clasificación, para segmentar los riesgos de la mejor manera posible.

Información de Pólizas o Certificados

La base de pólizas o certificados brinda información sobre las características de los riesgos y sobre las exposiciones para las diferentes coberturas. Los campos más usuales dentro de la base de pólizas son los siguientes:

- Identificador de la póliza
- Identificador del riesgo

  Si varios riesgos son cubiertos por una misma póliza, es común asociar un identificador a cada uno de ellos.
- Fechas Relevantes

  Cada registro individual contiene las fechas de emisión y terminación de las pólizas o coberturas de las pólizas. En el caso de que se creen registros ante cambios durante la vigencia de la póliza (por ejemplo un cambio en el deducible), se registra la fecha del cambio.
- Primas: corresponde a las primas emitidas asociadas a cada registro
- Exposición: corresponde a la exposición asociada a cada registro.
- Valor asegurado: corresponde al valor asegurado asociado a cada registro.
- Características del riesgo

Aquí se encuentran las características asociadas a la suscripción y la tarificación, las cuales dependen del ramo.

Información de siniestros y reclamaciones

Usualmente, además de la base de pólizas, se tiene una base de datos de siniestros que captura la información correspondiente a los siniestros asociados a las diferentes pólizas y los diferentes riesgos. Por lo general cada registro corresponde a una transacción (pago o cambio en la reserva de siniestros denunciados). Los campos corresponden a fechas y otros datos importantes con respecto a las reclamaciones.

Esta información se registra por cobertura.

Algunos de los campos típicos en una base de siniestros son los siguientes:

- Identificador de la Póliza
- Identificador del riesgo

  Por lo general, la base de datos de reclamaciones contiene un identificador que permite asociar la reclamación con el riesgo que la originó.
- Identificador de la reclamación
- Identificador del reclamante
- Fechas relevantes asociadas al siniestro

  Es importante registrar la fecha de ocurrencia del siniestro, la fecha de reporte del siniestro, la fecha de reserva y la fecha en que se pagó el siniestro o fecha en que se presentó un cambio en la reserva de siniestros denunciados. También suele registrarse la fecha en que cambió el estatus de la reclamación.
- Estatus de la reclamación

  Se registra si la reclamación está abierta o si ya está cerrada (esto es, si ha sido pagada). En algunas ocasiones es necesario reabrir reclamaciones que ya habían sido cerradas y es útil tener un registro de estas reclamaciones.
- Número de reclamaciones

  Este campo identifica el número de reclamaciones por cobertura, asociadas con la ocurrencia del siniestro.
- Valor pagado

  Este campo registra el valor pagado. Si el seguro es susceptible a pérdidas catastróficas (terremotos, inundaciones, etc.), es conveniente registrarlos de forma separada.

- Reserva de Siniestros Denunciados

  Aquí se incluye la reserva de siniestros denunciados, o cambios en la misma.

- Gastos Asignables al siniestro

  Se registran los gastos que se puedan asignar directamente al siniestro (por ejemplo honorarios jurídicos), en caso de que aplique.

- Salvamentos y recobros

  Se registran los salvamentos y recobros asociados con la reclamación.

- Características de la Reclamación

  Aquí se puede registrar información adicional sobre la reclamación. Por ejemplo, en coberturas de automóviles se podría registrar el tipo de accidente.

Agrupamiento de los datos

Un paso fundamental en la tarificación es el agrupamiento de los datos. Es recomendable recolectar los datos de la manera más detallada posible, con el fin de poder generar varios análisis. Los métodos más comunes de agrupamiento de datos son año calendario, año de ocurrencia, año póliza y año de reporte aunque también pueden utilizarse otros períodos como trimestres o semestres.

Año Calendario

En este caso se consideran todas las primas y siniestros que ocurrieron durante el año calendario, sin tener en cuenta la fecha de emisión de la póliza, la fecha de accidente o la fecha de reporte. Las primas y los expuestos devengados corresponden a todas las primas y expuestos devengados durante el período. Las reclamaciones pagadas corresponden a todos los siniestros pagados durante ese año sin importar la fecha de ocurrencia o de reporte. Las reclamaciones incurridas corresponden a las reclamaciones pagadas en el año más el cambio en las reservas de siniestros avisados. Al final del año calendario, las reclamaciones incurridas son fijas.

Una ventaja de este tipo de agrupamiento es que los datos están disponibles de manera rápida cuando se acaba el año. La principal desventaja es que no hay una correspondencia en el tiempo entre las primas y los pagos. Las primas corresponden a pólizas emitidas durante el año anterior o durante el año actual, mientras que los pagos pueden proceder de pólizas emitidas varios años atrás. Esta agrupación es útil cuando los siniestros son reportados y pagados rápidamente.

Año de Ocurrencia

En este tipo de agrupamiento las primas y los expuestos devengados se calculan de la misma manera que en el año calendario. La diferencia está en que las pérdidas que se tienen en cuenta son las que ocurrieron durante ese año (o período). Este tipo de agrupamiento permite una mejor correspondencia entre las primas y las pérdidas ya que las primas devengadas durante ese año, se comparan con las pérdidas ocasionadas por siniestros ocurridos durante ese año o período. Debido a que al analizar el período las pérdidas totales no se conocen, se hace necesaria su estimación.

Año póliza

En este caso se tienen en cuenta todas las primas y pérdidas relacionadas con las pólizas suscritas durante el año sin importar cuando se pagaron o reportaron las pérdidas. Las primas devengadas y los expuestos devengados corresponden a todas las primas devengadas y los expuestos devengados de cada una de las pólizas suscritas durante el año. En este caso, ni las primas devengadas, ni los expuestos devengados, ni las pérdidas, están completas al finalizar el año póliza.

Este tipo de agrupación presenta la mejor correspondencia entre primas y pérdidas, pues se están comparando tanto las primas devengadas, como los siniestros incurridos de las pólizas suscritas en un determinado año.

Año de Reporte

Este método es similar al año calendario, salvo que las pérdidas se agregan de acuerdo al año en que fueron reportadas. Este tipo de agrupación se utiliza en ramos tales como Cumplimiento y Responsabilidad Civil.

Información externa

Cuando se quiere tarificar una nueva línea de negocio, o cuando el volumen de información de una compañía no es suficiente, surge la necesidad de utilizar información externa. Esta información externa puede provenir de datos de la industria (que pueden ser recolectados por los supervisores u otras asociaciones), datos de los competidores, datos de otros sectores relacionados con los seguros, entre otros. En la medida de lo posible, se debe buscar información que cumpla con los mismos estándares de calidad que se piden para la información interna.

Métodos de Tarificación

Los métodos más empleados en el cálculo de las primas son el método de prima pura y el método de Loss Ratio. Para utilizar cualquiera de ellos es necesario estimar y proyectar las pérdidas del período de experiencia (período observado), que son indicativas de las pérdidas en el período de proyección (período en el cual las tarifas estarán vigentes).

Estimación y proyección de las pérdidas

Desarrollo de las pérdidas

Al momento de la tarificación es usual que sólo se conozca el valor de las reclamaciones que han sido pagadas y el valor reservado para aquellas reclamaciones que han sido reportadas a la compañía, correspondientes a un cierto período.

Para obtener el valor total de las pérdidas correspondientes a un cierto período o pérdidas últimas, es necesario estimar a partir de estos valores conocidos, el valor de las pérdidas que aún no han sido avisadas o que no fueron suficientemente reservadas, (IBNR).

Para estimar el IBNR y las pérdidas últimas, existen metodologías como la de *Chain Ladder*, la de Bornhuetter Fergusson, o métodos estocásticos. Pero, como se ha dicho anteriormente, la explicación de estos métodos de cálculo de IBNR se encuentra fuera del alcance de esta tesis.

Gastos de Ajustes de Siniestros

Como se mencionó previamente, además del valor de los siniestros, también es necesario tener en cuenta los gastos en que la compañía incurre para ajustarlos y liquidarlos. Los gastos de Ajuste de Siniestros se dividen en gastos asignables y gastos no asignables. Los gastos asignables son aquellos que se pueden asignar a un siniestro particular. Dentro de este tipo de gastos se encuentran los honorarios jurídicos correspondientes a un caso particular, o el valor que se paga a un ajustador por un caso específico. Por su parte, los gastos no asignables, como su nombre lo indica, no están relacionados con un siniestro en particular (por ejemplo los salarios del departamento de indemnizaciones).

En tarificación es común agregar las pérdidas con los gastos asignables, para calcular la prima.

Deducibles, Límites y Coaseguros

Estos tres conceptos, limitan la pérdida de la compañía y por lo tanto afectan el cálculo de la prima pura. Una de las razones que tiene en cuenta la compañía a la hora de fijar un deducible es la razón de eliminación de pérdida (*Loss Elimination Ratio*), que busca cuantificar la fracción de las pérdidas esperadas que se elimina al colocar un deducible. Es importante tener en cuenta que el establecimiento de deducibles y límites en el desarrollo de los productos, permite que los costos del seguro sean menores, y que una mayor cantidad de personas puedan acceder a ellos. Asimismo permiten reducir la frecuencia y el riesgo moral.

Credibilidad

En muchas ocasiones no se cuenta con una cantidad suficiente de datos, para producir tarifas estables y adecuadas, y es necesario utilizar información adicional, que está relacionada con la información observada. En este caso, se utiliza la teoría de credibilidad, que permite combinar la información observada, con información externa, para obtener una mejor estimación de las tarifas.

En términos generales, la teoría de la credibilidad genera una estimación asignándole un peso Z a la experiencia observada, y un peso (1-Z) a la experiencia relacionada, donde Z es un factor que varía entre cero y uno. Existen diferentes metodologías para calcular Z como el método de credibilidad clásica, o la credibilidad de Bühlmann.

Este punto se encuentra fuera del alcance de esta Tesis.

Otros Aspectos

Para que las pérdidas históricas puedan ser un buen indicador de las pérdidas esperadas durante el período en que van a estar vigentes es necesario realizar varios ajustes. Los principales ajustes tienen que ver con el tratamiento de siniestros extraordinarios, siniestros catastróficos, cambios en los beneficios, e inflación.

Método de la Prima Pura

Según la definición detallada previamente, la prima pura corresponde al costo promedio por expuesto, esto es:

$$Prima\ Pura = Frecuencia * Severidad$$

En el método de prima pura, se estiman las pérdidas totales o pérdidas últimas y se calcula el número de expuestos correspondiente al período de experiencia.

Posteriormente, las pérdidas totales son proyectadas teniendo en cuenta aspectos tales como la inflación, para que sean representativas de las que se observarán en el período en el cual las tarifas estarán vigentes. Por último la prima pura es cargada con un factor de gastos, así como un margen por riesgo y utilidad.

Método de Loss Ratio

El objetivo del método Loss Ratio o razón de pérdida es determinar en qué porcentaje debe ser aumentada o disminuida la tasa actual, para alcanzar la utilidad esperada. Para ello se emplea un factor de ajuste, que se calcula como:

$$Factor\ de\ Ajuste = \frac{Razón\ de\ pérdida\ esperada\ efectiva}{Razón\ de\ pérdida\ permisible} - 1$$

Donde la razón de pérdida efectiva corresponde a:

$$Razón\ de\ Pérdida\ Efectiva = \frac{Pérdidas\ esperadas\ en\ el\ período\ de\ proyección}{Pérdidas\ devengadas\ en\ las\ tasas\ actuales}$$

Y la razón de pérdida permisible se define como:

$$Razón\ de\ pérdida\ permisible = 1 - \%\ Gastos - \%\ Riesgo\ y\ Utilidad$$

Comparación del método de Prima Pura y el método de Loss Ratio

Los resultados obtenidos con el método de prima pura y con el método de Loss Ratio son equivalentes, siempre y cuando los datos y las hipótesis sean consistentes para ambas aproximaciones. Entonces surgen las siguientes preguntas: ¿En qué casos es mejor utilizar el método de prima pura? ¿En qué casos es preferible utilizar el método de Loss Ratio?

Para responder estas preguntas es necesario tener en cuenta las siguientes consideraciones:

- Dado que mediante el método de Loss Ratio se estima un porcentaje de cambio para las tasas actuales, esta metodología no se puede utilizar para un nuevo producto o una nueva compañía. A pesar de que no exista información histórica, el método de prima pura se puede aplicar estimando la prima pura esperada, a partir de datos externos, juicios actuariales, entre otros.

- En la metodología de Loss Ratio es indispensable ajustar las primas del período de experiencia, en términos de las primas actuales. Por lo tanto, si ha habido múltiples cambios de tarifa que dificulten la actualización de las primas, es recomendable utilizar el método de prima pura.

- Si la información de expuestos es difícil de obtener o si los expuestos no están claramente identificados, se recomienda el uso del método de Loss Ratio pues no requiere estos datos.

Tarificación por clases

Como se mencionó previamente, además de alcanzar la suficiencia en el agregado, también es importante alcanzar la suficiencia por clases de riesgo y a nivel individual.

Dado que usualmente no existe la información suficiente para tarificar los riesgos a nivel individual, las compañías utilizan el método de tarificación por clases, que consiste en agrupar riesgos con pérdidas potenciales similares y estimar tasas diferentes para cada clase.

Para poder agrupar los riesgos es necesario establecer qué variables los segmentan de manera efectiva en grupos con pérdidas potenciales similares. Las variables que se utilizan

son llamadas variables de clasificación. Los diferentes valores o rangos de las variables de clasificación, se conocen como categorías o niveles de la variable.

La clasificación adecuada de los riesgos, mediante variables de clasificación, permite establecer tarifas equitativas y evitar la selección adversa, que resulta en un espiral de tarifas cada vez más altas y una menor rentabilidad para la compañía. De la misma manera, identificar características de segmentación importantes, que no están siendo utilizadas por los competidores, permite atraer clientes con bajo riesgo y, en el largo plazo, permite aceptar de manera rentable una mayor cantidad de negocios.

Métodos Univariados

En el caso de los métodos univariados, se calculan las relatividades de los diferentes niveles de las variables de clasificación con respecto a la clase base. Para ello se puede utilizar métodos como el método de prima pura, o el método de Loss Ratio, los cuales fueron discutidos previamente.

Métodos Multivariados

Dado que la tarificación utilizando métodos univariados de clasificación puede presentar distorsiones, debido principalmente a que no siempre se tienen en cuenta las relaciones entre las diferentes variables de tarificación, en los últimos años se ha extendido el uso de métodos multivariados, tales como los modelos lineales generalizados.

El uso de métodos multivariados como los modelos lineales generalizados, permite modelar todas las variables a la vez, teniendo en cuenta las interacciones (esto es que el efecto de una variable varíe de acuerdo con los niveles de otra), así como la relación entre los expuestos de cada variable. También permite asumir una distribución para la frecuencia y para la severidad, establecer mediante criterios estadísticos cuáles son las variables más relevantes y realizar diagnósticos, los cuales sirven para evaluar qué tan apropiado es el modelo que se está asumiendo, entre otras ventajas.

Algunos de los métodos multivariados que se utilizan en tarificación son los modelos lineales generalizados (MLG), el análisis factorial y las redes neuronales.

Capítulo 3. MINERÍA DE DATOS

Definición

La minería de datos es un conjunto de herramientas y técnicas de análisis de datos que por medio de la identificación de patrones extrae información interesante, novedosa y potencialmente útil de grandes bases de datos que puede ser utilizada como soporte para la toma de decisiones.

Para descubrir conocimiento de la información se pueden utilizar varias formas de análisis por medio de las cuales se puede llegar a identificar patrones y reglas en los datos para luego crear escenarios, esta información se puede representar por medio de modelos matemáticos sobre datos históricos y con esto se crea un modelo de minería de datos. Después de haber creado un modelo de minería de datos, se puede examinar nueva información a través del modelo evaluando si se apega a los patrones o reglas definidos.

Datos, información y conocimiento

Datos

Los datos son en esencia números o textos que pueden ser procesados en una computadora. En la actualidad, las organizaciones acumulan grandes cantidades de datos en distintos formatos y en distintas bases de datos, entre las que se incluyen datos operacionales o transaccionales en las que se almacenan costos, ventas, inventarios, contabilidad, etc.

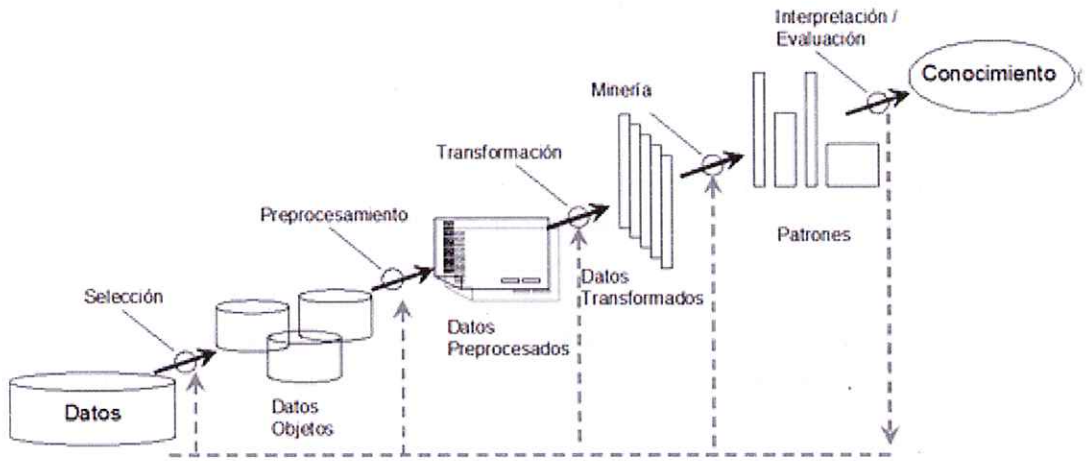
Información

Los patrones, asociaciones o relaciones entre los datos proporcionan información, por ejemplo el análisis de transacciones de un punto de venta brinda información sobre qué cantidad de productos se han vendido y durante cuánto tiempo.

Conocimiento

La información puede ser convertida en conocimiento partiendo de patrones históricos.

Pasos que componen el proceso de extracción de conocimiento de bases de datos (KDD)



Así, los pasos que componen al proceso KDD son cinco: selección del objetivo, pre-proceso de datos, transformación, minado de datos e interpretación de los resultados.

La selección del objetivo tiene como finalidad estudiar el problema y decidir cuál es la meta del proyecto. Una vez definido el problema, se identifican las fuentes de datos internas o externas y se selecciona el subconjunto de datos necesarios para la aplicación de un algoritmo de minería de datos.

El pre-proceso de datos consiste en estudiar los datos seleccionados para entender el significado de los atributos y para detectar errores de integración, por ejemplo, datos repetidos con distinto nombre o datos que significan lo mismo en diferente formato.

Una vez que se tienen los datos pre-procesados, se procede a la transformación final de los mismos, esto con el fin de que se ajusten al formato de entrada del algoritmo seleccionado.

El siguiente paso es el minado de datos propiamente dicho. Aquí se aplican los diferentes algoritmos de análisis a los datos ya transformados. La finalidad en esta etapa es encontrar patrones útiles e interesantes en los datos.

Por último, se procede a interpretar y evaluar los resultados obtenidos en la etapa de minado de datos. Aquí, el usuario debe valorar los resultados conseguidos y, de ser necesario, aplicar una y otra vez los algoritmos de Data Mining hasta encontrar información útil y valiosa. Esto último hace que el proceso KDD sea un proceso iterativo y de búsqueda continua, en donde el conocimiento y la intuición del usuario juegan un papel fundamental en el proceso.

Protocolo de un proyecto de Minería de Datos

Un proyecto de minería de datos tiene varias fases necesarias que son, esencialmente:

- Comprensión: del negocio y del problema que se quiere resolver
- Determinación, obtención y limpieza: de los datos necesarios
- Creación de modelos matemáticos
- Validación, comunicación: de los resultados obtenidos
- Integración: si procede, de los resultados en un sistema transaccional o similar

La relación entre todas estas fases sólo es lineal sobre el papel. En realidad, es mucho más compleja y esconde toda una jerarquía de sub-fases. A través de la experiencia acumulada en proyectos de minería de datos se han ido desarrollando metodologías que permiten gestionar esta complejidad de una manera más o menos uniforme.

Técnicas de Minería de Datos

En los últimos años han existido muchos avances en las investigaciones y desarrollos relacionados con la minería de datos, como resultado, se han desarrollado diversas técnicas y sistemas relativos al data mining. Diferentes esquemas de clasificación pueden ser usados para categorizar métodos y sistemas de minado de datos, como el tipo de base de datos a estudiar (relacional, orientada a objetos, multimedia, etc.), el tipo de conocimiento que se quiere extraer (reglas de asociación, reglas de clasificación, clustering, etc.), así como las técnicas que serán aplicadas en el proceso (basadas en patrones, teoría estadística, teoría matemática, enfoques integradores, etc.).

En la práctica, los métodos de data mining más utilizados caen dentro de la categoría de tipo de conocimiento a extraer. Las técnicas de minado de datos pertenecientes a esta categoría buscan hacer predicción y/o descripción de un fenómeno determinado.

Varios algoritmos y técnicas como Clasificación, Clustering, Regresión, Inteligencia Artificial, Redes Neuronales, Reglas de Asociación, Árboles de Decisión, Algoritmos genéticos, etc. son utilizados para descubrir conocimientos de las bases de datos.

Reglas de Asociación

Mediante el minado de reglas de asociación se pueden encontrar interesantes relaciones de asociación o correlación en los datos. Dada la gran cantidad de datos que continuamente se recolectan y almacenan, muchas industrias se han interesado por encontrar reglas de asociación en sus bases de datos. El descubrimiento de interesantes relaciones de asociación en grandes cantidades de registros transaccionales, puede ayudar en diversos procesos de toma de decisiones relacionados con el negocio.

Una regla de asociación es un criterio que implica ciertas relaciones de asociación entre distintos objetos de una base de datos, tales como "ocurren juntos" o "uno implica al otro". Matemáticamente se representa como una implicación de la forma A ➔ B, en donde A y B

representan conjuntos de atributos con intersección vacía ($A \cap B = \emptyset$), de tal forma que la regla se presenta en un conjunto de transacciones D con una confianza del $\alpha\%$.

Clasificación y Predicción

La clasificación y la predicción son dos formas de análisis de datos que pueden ser usadas para extraer modelos que describen importantes clases de datos o predicen valores futuros.

En la clasificación de datos se desarrolla una descripción o modelo para cada una de las clases presentes en la base de datos. Existen muchos métodos de clasificación tales como los árboles de decisión, los métodos estadísticos, las redes neuronales, y los conjuntos difusos, entre otros.

La predicción puede ser vista como la construcción y uso de modelos para evaluar las clases de una muestra sin clasificaciones, o para evaluar el valor, o rango de valores, que un atributo debería de tener para una muestra determinada.

Agrupamiento o *Clustering*

El *Clustering* identifica grupos de datos que son "similares". La similitud puede medirse mediante funciones de distancia especificadas por los usuarios o por expertos.

Cuando se utiliza la técnica de *clustering*, se obtiene un diagrama en el cual se muestra cómo los datos caen dentro de distintos grupos (*clusters*). En el caso más simple, se asocia a cada dato un *cluster*, dibujando los datos o instancias en un diagrama de dos dimensiones.

Algunos algoritmos de *clustering* permiten a una instancia pertenecer a uno a más *clusters*, como resultado, el diagrama en dos dimensiones muestra cómo se solapan los subconjuntos de datos (como un Diagrama de Venn).

Otros algoritmos asocian instancias a *clusters* de manera probabilística, así para cada instancia, existe una probabilidad asociada o un "grado de pertenencia" con el cual se asigna a un determinado *cluster*.

Árboles de Decisión

Un árbol de decisión es un modelo de predicción cuyo objetivo principal es el aprendizaje inductivo a partir de observaciones y construcciones lógicas. Son muy similares a los sistemas de predicción basados en reglas, que sirven para representar y categorizar una serie de condiciones que suceden de forma sucesiva para la solución de un problema.

Un árbol gráficamente se representa por un conjunto de nodos, ramas y hojas. El nodo principal o raíz es el atributo a partir del cual se inicia el proceso de clasificación; los nodos internos corresponden a cada una de las preguntas acerca del atributo en particular del problema. Cada posible respuesta a los cuestionamientos se representa mediante un nodo hijo. Las ramas que salen de cada uno de estos nodos se encuentran etiquetadas con los

posibles valores del atributo. Los nodos finales o nodos hoja corresponden a una decisión, la cual coincide con una de las variables clase del problema a resolver.

Este modelo se construye a partir de la descripción narrativa de un problema, ya que provee una visión gráfica de la toma de decisión, especificando las variables que son evaluadas, las acciones que deben ser tomadas y el orden en el que la toma de decisión será efectuada. Cada vez que se ejecuta este tipo de modelo, sólo un camino será seguido dependiendo del valor actual de la variable evaluada. Los valores que pueden tomar las variables para este tipo de modelos pueden ser discretos o continuos.

Contenido de un Árbol de Decisión

Nodos de Decisión

Un nodo de decisión representa un punto en el que se debe tomar una decisión. De estos nodos salen ramas que son las posibles decisiones a tomar y se representan con un cuadrado (□).

Las ramas que nacen de un nodo de decisión representan las alternativas. Un nodo de decisión podrá tener tantas ramas como alternativas existan.

Un Árbol de Decisión puede contener varios momentos de decisión. Asimismo, pueden existir varios nodos de decisión consecutivos.

Nodos de Probabilidad o Acontecimiento

Un nodo de probabilidad o acontecimiento (estado de la naturaleza) representa el momento en que se produce un evento aleatorio. De estos nodos salen ramas que representan posibles resultados para eventos inciertos en los que no se tiene control sobre su resultado. Se representan con un círculo (○).

Las ramas que nacen de un nodo de probabilidad o acontecimiento muestran los distintos comportamientos que puede exhibir una variable no controlable o aleatoria.

Al igual que ocurre con los nodos de decisión, un Árbol puede contener varios nodos de probabilidad de forma sucesiva.

Ramas

Las Ramas se esquematizan con el símbolo ⟶ y representan los distintos caminos que se pueden emprender cuando se toma una decisión o bien ocurre un evento aleatorio.

Construcción de un Árbol de Decisión

En un árbol de decisión la secuencia de eventos se desarrolla de izquierda a derecha. Las probabilidades se indican en las ramas de estado de la naturaleza. Son probabilidades condicionales de eventos que ya sucedieron. Los valores en dinero en los extremos de las ramas dependen de decisiones tomadas y estados de la naturaleza previamente ocurridos. La decisión que resulta de un análisis de árbol no es una definitiva sino condicionada a la ocurrencia de eventos que sucedan en la próxima decisión.

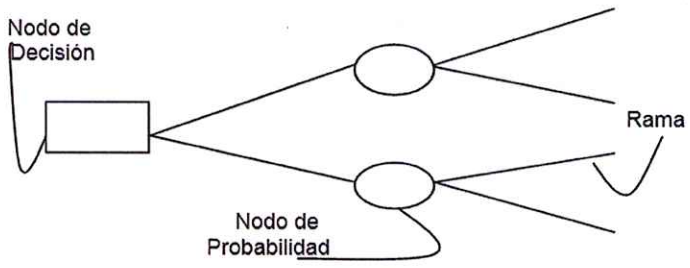
Un Árbol de Decisión se evalúa de atrás hacia adelante, teniendo en cuenta la influencia de las decisiones o eventos aleatorios. En el caso de los nodos de decisión se evalúan las mejores alternativas. En el caso de los eventos aleatorios se indican las probabilidades asociadas a cada evento.

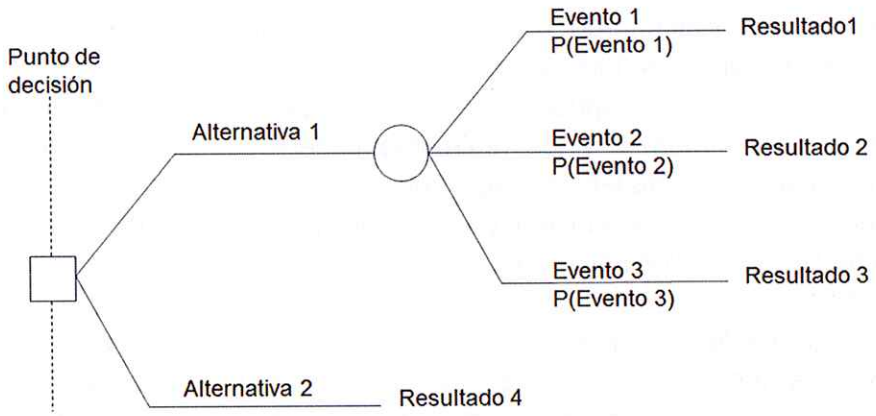
Pasos a seguir en la construcción de un Árbol de Decisión

1. Definir el problema
2. Dibujar el Árbol de Decisión
3. Asignarle probabilidades a los diversos estados de la naturaleza
4. Estimar los resultados de cada una de las posibles combinaciones de alternativas y estados
5. Resolver el problema calculando los valores monetarios esperados de cada nodo. Para ello, se procede desde las ramas finales hacia el origen del árbol, calculando los valores esperados en los nodos de incertidumbre y para cada decisión. Se siguen estas reglas hasta que se llega al nodo raíz del árbol.

Estructura de un Árbol de Decisión

Los siguientes gráficos muestran la estructura de un Árbol de Decisión:





Regresión Lineal Múltiple

Es la más utilizada para formar relaciones entre datos.

En estadística la regresión lineal o ajuste lineal es un método matemático que modela la relación entre una variable dependiente Y, las variables independientes X_i y un término aleatorio ε. Este modelo puede ser expresado como:

$$Y_t = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots \ldots + \beta_P X_P + \varepsilon$$

Donde:

Y_t: Variable dependiente, explicada o regresando.

$X_1, X_2, \ldots, X_p$: Variables explicativas, independientes o regresores.

$\beta_0, \beta_1, \ldots \beta_p$: Parámetros, que miden la influencia que las variables explicativas tienen sobre el regresando.

β_0 es la intersección o término "constante", las β_i ($i > 0$) son los parámetros respectivos a cada variable independiente, y "p" es el número de parámetros independientes a tener en cuenta en la regresión. La regresión lineal puede ser contrastada con la regresión no lineal.

Requisitos y limitaciones

Hay ciertos requerimientos necesarios para poder utilizar la técnica de regresión múltiple:

- *Linealidad*: La variable a explicar depende linealmente de las variables explicativas. Si la respuesta no aparenta ser lineal, debemos introducir en el modelo componentes no lineales. Otro tipo de respuesta no lineal es la interacción. Para ello, se ha de incluir en

el modelo términos de interacción, que equivalen a introducir nuevas variables explicativas que en realidad son el producto de dos o más de las independientes.

- *Normalidad y equidistribución de los residuos*: Se llaman residuos las diferencias entre los valores calculados por el modelo y los realmente observados en la variable dependiente. Para tener un buen modelo de regresión no es suficiente con que los residuos sean pequeños. La validez del modelo requiere que los mismos se distribuyan de modo normal y con la misma dispersión para cada combinación de valores de las variables independientes.
- *Número de variables independientes*: uno podría llegar a estar tentado a incluir en el modelo cualquier variable que esté en una base de datos con la esperanza de que cuantas más varialbes se incluyan mejor será el modelo. Sin embargo, en este último caso se corre el riesgo de cometer error de tipo I. Otro problema es que si se espera ajustar unas pocas observaciones usando muchas variables, muy probablemente se consiga una aproximación muy artificial y además muy sensible a los valores observados.
- *Colinealidad*: si dos variables independientes están estrechamente relacionadas y ambas son incluidas en el modelo, muy posiblemente ninguna de las dos sea considerada significativa, aunque si hubiésemos incluido sólo una de ellas, sí.
- *Observaciones anómalas*: Se debe poner especial cuidado en identificarlas (y descartarlas si procede) pues tienen gran influencia en el resultado. Hay veces que son producto de pequeños errores en la entrada de datos, pero de grandes consecuencias en el análisis.

La Minería de Datos en las Compañías de Seguros

Tal como detalla López (2006), la minería de datos (data mining), puede definirse como el proceso de seleccionar, explotar y dar forma a grandes cantidades de datos para descubrir patrones antes desconocidos.

La minería de datos mejora los modelos existentes detectando variables más relevantes que las utilizadas hasta el momento, identificando términos de interacción y detectando relaciones no lineales.

La minería de datos puede ayudar a las compañías de seguros en prácticas empresariales como:

- Cálculo de primas
- Captación de nuevos clientes
- Fidelización de clientes
- Desarrollo de nuevas líneas de productos
- Creación de informes de riesgos geográficos
- Detección de reclamaciones fraudulentos
- Administración de campañas sofisticadas
- Estimación de las provisiones para siniestros pendientes

- Apoyo a los reguladores para comprender las tarifas y modelos de la empresa
- Coordinación de los departamentos de marketing y actuarial

Esta tesis profundiza sobre el primer punto, es decir sobre el cálculo de las primas como se detalla seguidamente.

Cálculo de Primas

Identificación de factores de riesgo que predicen beneficios, siniestros y pérdidas

La pregunta clave para la fijación de tarifas es la siguiente: ¿Cuáles son los factores de riesgo a tener en cuenta para predecir la probabilidad de que ocurra un siniestro y el importe al que pueden ascender las indemnizaciones?

Pese a que muchos de los factores de riesgo que afectan a las primas son obvios, pueden existir relaciones no intuitivas y sutiles entre las variables que son difíciles o imposibles de identificar sin aplicar análisis más complejos.

Mejora de la precisión predictiva: segmentar bases de datos

Para mejorar la precisión en las predicciones, las bases de datos deben segmentarse en grupos homogéneos. A continuación, los datos de cada grupo podrán explorarse, analizarse y modelarse. Dependiendo del análisis que se quiera hacer, la segmentación puede realizarse utilizando variables asociadas a factores de riesgo, beneficio o comportamiento.

Dado que estos segmentos a menudo revelan contrastes marcados, podrán interpretarse más fácilmente. Como resultado, los actuarios pueden predecir con mayor exactitud la probabilidad de un siniestro y la indemnización a la que éste puede ascender.

Modelos de *data mining* como los árboles de decisión y las redes neuronales pueden predecir el riesgo de forma más precisa que los modelos actuariales actuales. De esta forma, las compañías de seguros pueden establecer las tasas de primas de forma más precisa lo que resulta en modelos de pricing más exactos y, en consecuencia, permite que las compañías de seguros se encuentren en una posición más competitiva.

Capítulo 4. LA EMPRESA BAJO ESTUDIO Y SU PRODUCTO AFECTADO

En este capítulo se muestra un detalle de la compañía de seguros en estudio, así como la problemática a la que se enfrenta actualmente: la identificación de los riesgos más significativos.

La empresa bajo estudio es una compañía de Seguros y Servicios de primer nivel de Argentina que opera asimismo en Europa, Norteamérica, el Caribe y Sudamérica

Dicha aseguradora cuenta con más de 60 años de experiencia global. La casa matriz, es una empresa de Fortune 500, con sede en la ciudad de Nueva York, y cotiza en la Bolsa de Nueva York.

Esta compañía de seguros comercializa los siguientes seguros:

- Programas de Extensión de Garantía
- Seguros de Mercado Financiero
- Seguros de Protección Móvil

Producto bajo análisis

A los efectos de este estudio, se ha procedido a analizar la cartera de Seguros de Protección Móvil antes descripto.

Este seguro cubre a los equipos celulares ante los siguientes riesgos:

- Falla mecánica
- Daño Accidental
- Robo y Hurto
- Falla de Software

El principal cliente de esta compañía es uno de los mayores operadores de telefonía móvil en Argentina.

En este seguro, ante un siniestro, la compañía de seguros puede reparar o remplazar el equipo siniestrado según el daño haya sido total o parcial. Del mismo modo, si el siniestro es parcial pero no se consiguen las partes se procede a remplazar el equipo por otro.

La compañía de seguros puede remplazar el equipo siniestrado utilizando equipos nuevos o equipos reacondicionados. En este último caso, el costo de estos celulares es menor.

Asimismo, en caso de robo o de destrucción total se le cobra al asegurado un deducible del 50%. Este deducible no opera en el caso de siniestros de falla mecánica o en el caso de siniestro por daño accidental cuando el reemplazo es producto de que no se consiga las partes necesarias para reparar el equipo siniestrado.

Capítulo 5. PREPARACIÓN DE LOS DATOS

Proceso de extracción del conocimiento

Como se ha detallado previamente, los pasos que componen al proceso de extracción de conocimiento en bases de datos (KDD) son cinco:

1. Selección del objetivo
2. Pre-proceso de datos
3. Transformación de datos
4. Minado de datos
5. Interpretación de resultados

Los primeros tres pasos están orientados básicamente a la preparación de los datos, mientras que los últimos dos, corresponden tanto al minado de datos, propiamente dicho, como a la interpretación de los resultados arrojados por los algoritmos de minado.

Selección del Objetivo

Problema Observado

La selección del objetivo tiene como finalidad estudiar el problema y decidir cuál es la meta del proyecto. En este caso, el problema observado es identificar los distintos factores de riesgo asociados a una cartera de asegurados específica para tarificar de manera más precisa.

Meta del Estudio

Con base en lo anterior, se ha fijado una meta para el estudio que consiste en desarrollar un modelo que permita identificar los factores de riesgo y calcular primas más precisas y, con ello, en caso de corresponder, aplicar planes de acción y garantizar el resultado a futuro de la compañía de seguros.

Fuentes de Datos

A los efectos de efectuar el presente trabajo se procedió a analizar por un lado la base de expuestos de la cartera de asegurados del Seguro de Protección Móvil y por otro la base de siniestros de este producto para luego relacionarlas y, obtener de esta forma, cuáles deberían ser las tarifas para cada segmento de asegurados.

Asimismo, a fin de calcular las tarifas para cada segmento se utilizó el método de la Prima Pura y el método del Loss Ratio.

Base de Expuestos

Los campos de dicha base son los siguientes:

- Dealer
  - ARCO: Protección Empresas y profesionales – Cartera propia
  - ARDF: Programa anti-daño
  - ARGC: Protección Empresas y profesionales – Cartera Adquirida
  - ARGI: Protección Individuos – Cartera Adquirida
  - ARIN: Protección Individuos – Cartera propia
  - ARPS: Programa Post-pago
- Número de certificado
- Fecha de inicio de vigencia
- DNI
- Nombre del Asegurado
- Código Postal
- Teléfono
- Gama de celular
  - Base
  - Baja
  - Media
  - Alta
  - Low Premium
  - Premium
  - Iphone
- Marca del Celular
- Modelo del Celular
- Número de Serie
- Status (certificado activo o cancelado)

Actualmente la base de expuestos de esta cartera es de 680,000 asegurados.

No obstante lo anterior, en el presente trabajo se analizó de forma mensual la base de expuestos de los últimos dos años (desde Enero 2013 a Diciembre de 2014).

En Noviembre de 2013 la compañía de seguros le adquirió una cartera de asegurados del programa de Protección Móvil a otra aseguradora. Debido a lo anterior, en el presente trabajo se decidió agrupar a los asegurados por Cartera Propia y Cartera Adquirida según los *Dealers* de cada una de dichas aseguradoras.

La razón de lo anterior obedece a que la compañía de seguros vendedora de la cartera no necesariamente tenía los mismos criterios de suscripción que la compradora por lo que unos y otros asegurados pueden llegar a comportarse de forma distinta lo que se traduce en posibles diferentes riesgos.

Asimismo, se agrupó a la base de expuestos según la gama de celular asegurado.

Base de Siniestros

La base de siniestros tiene los siguientes campos

- Dealer
- Número de certificado
- Número de Siniestro
- Solución inicial
  - Reparación
  - Reemplazo
- Solución definitiva
  - Reparación
  - Reemplazo
- Tipo de Reemplazo
  - Celular Nuevo
  - Celular Reacondicionado
- Gama de celular
  - Base
  - Baja
  - Media
  - Alta
  - Low Premium
  - Premium
  - Iphone
- Fecha de ocurrencia del siniestro
- Fecha de pago del siniestro
- Costo del siniestro
- Precio del equipo para el cliente
- Deducible
- Tipo de siniestro
  - Robo
  - Daño
  - Falla Mecánica
  - Falla de Software
- Centro de Servicio
- Status de siniestro
  - Denegado
  - No Denegado

Desde el año 2013 se han registrado aproximadamente 185.000 siniestros que son los que se han considerado para abordar el presente trabajo.

Al igual que con la base de expuestos, la base de siniestros se la agrupó por cartera propia o cartera adquirida y por gama de celular.

Asimismo, se analizaron los siniestros desde inicios del 2013 hasta Diciembre de 2014 según el criterio de Fecha de Ocurrencia de manera mensual.

Pre-proceso de Datos

El pre-proceso de datos consiste en estudiar los datos seleccionados para entender el significado de los atributos, detectar errores en la información, estandarizar datos, hacer agrupaciones, etc.

En este caso el pre-proceso consistió básicamente en la detección de errores, estandarización, y el agrupamiento de datos.

Como se mencionó previamente, las dos bases de datos se agruparon por Cartera Propia o Cartera Adquirida y por Gama de Celular Asegurado.

Tal como se comentó en el punto anterior, la razón del agrupamiento por Cartera Propia o Cartera Adquirida obedece a que no necesariamente los criterios de suscripción entre una y otra cartera son similares por lo que los asegurados de una y otra cartera pueden comportarse de manera disímil.

Por su parte, la razón del agrupamiento por Gama de Celular radica en que los riesgos son distintos entre celulares más económicos y los más caros. Por ejemplo, un celular más costoso es más factible de ser robado que uno celular más básico. Asimismo, la agrupación por Gama también obedece al que el costo de los siniestros según las gamas puede es muy distinto (tanto en lo que hace al costo de reparación como al costo de reemplazo en caso de siniestro).

En el caso de la base de siniestros, se agruparon los siniestros por año de ocurrencia y no por fecha efectivamente de pago pues, tal como se estableció previamente, este criterio permite calcular de una manera más fehaciente cuál es la frecuencia siniestral.

Por otro lado, se agruparon los siniestros por tipo de denuncia según éstas sean de Robo, Falla Mecánica, Daño Accidental y Falla de Software y se analizó si la reclamación terminó en reparación o reemplazo y en este último caso si el reemplazo se resolvió con un equipo nuevo o con un equipo reacondicionado.

Transformación de los Datos

En base a lo descripto anteriormente, es decir a la corrección de errores en las bases y a la estandarización de ciertos atributos se han transformado los datos que alimentan el modelo que permite identificar los factores de riesgo y calcular las primas de manera más certera.

A los efectos del cálculo de la prima se utilizó el Método de la Prima Pura y el Método del Loss Ratio ambos antes descriptos.

Minado de Datos

El minado de datos es la etapa principal del proceso de extracción de conocimiento en bases de datos. Durante esta fase, se aplican algoritmos de extracción de conocimiento en datos que ya han sido preparados y transformados en etapas anteriores.

En el caso del presente trabajo, se ha construido un modelo en la herramienta Microsoft Excel cuyo objetivo es calcular, en base a técnicas actuariales y al concepto de Árboles de Decisión cuáles deberían ser las primas que deberían abonar los asegurados en cada segmento. Este análisis también se corroboró con el Método del Loss Ratio.

Asimismo, a fin de calcular todo el universo de posibilidades se desarrolló un programa en Visual Basic por Application (VBA) para Microsoft Excel tal como se detalla en las siguientes secciones.

Interpretación de Resultados

El modelo creado, asimismo, permite identificar los casos en los que la siniestralidad se ha disparado debido a un factor de riesgo específico permitiendo definir planes de acción para revertir tales situaciones.

Capítulo 6. CONSTRUCCIÓN DEL MODELO

Determinación de la Prima Pura

Variables relevantes para el cálculo de la Prima Pura

Frecuencia Siniestral:

Tal como se detalló previamente, las variables que resultan relevantes para el cálculo de la Prima Pura son las siguientes:

- Probabilidad de Robo
- Probabilidad de Daño Accidental
- Probabilidad de Falla Mecánica
- Probabilidad de Falla de Software

Los cuatro variables anteriores determinan la Frecuencia Siniestral antes explicada.

Dichas frecuencias varían según la gama del celular y si existe o no un deducible asociado. Asimismo, y tal como se estableció anteriormente, la frecuencia siniestral también puede variar dependiendo de si los asegurados pertenecen a la cartera propia o a la cartera adquirida.

A fin de calcular cada una de estas frecuencias se relacionó para cada gama de celular y si dicha gama pertenece a la cartera propia o a la cartera adquirida la cantidad de casos que tuvieron siniestros en un período determinado dividido por la cantidad de expuestos que había en ese mismo período de análisis.

Las frecuencias antes descriptas se han calculado mensualmente desde Enero de 2013 a Diciembre de 2014 a fin de analizar la evolución de estas variables y definir cuál debería ser la más representativa como valor para predecir la frecuencia a futuro.

A los efectos del cómputo de estas frecuencias se efectuó un programa en VBA que, en base a lo antes explicado, relaciona la base de siniestros y la base de expuestos y calcula cada una de estas variables.

Severidad Siniestral:

A los efectos del cálculo de la Severidad Siniestral, para cada cobertura, gama y si el asegurado pertenece a la cartera propia o a la cartera adquirida se modelaron las siguientes variables que inciden en la severidad esperada:

- La proporción de celulares que se reparan versus los que se reemplazan

  A los efectos de este cálculo se analizó la base de siniestros desde Enero de 2013 a Diciembre de 2014 y, en base a cada tipo de siniestro (robo, daño accidental, falla mecánica y falla de software), gama de celular y tipo de cartera (adquirida o propia), se computó cuántos de estos celulares derivaron en reparación y cuantos derivaron en reemplazo y se calcularon luego la proporciones respectivas.

  Este análisis se efectuó de manera mensual desde Enero de 2013 a Diciembre de 2014 a fin de analizar la evolución de estas variables y definir cuál debería ser la más representativa como valor para predecir el porcentaje de reparación y reemplazo a futuro.

  A fin de poder computar cada una de estas variables se corrió un programa en VBA que contempló todo el universo de posibilidades y arrojó el resultado de estas variables.

- Dentro de los que equipos que se reemplazan, la proporción de los celulares que se reemplazan por celulares nuevos versus los que se reemplazan por celulares reacondicionados

  Aquí también se analizó de manera mensual la base de siniestros desde Enero de 2013 a Diciembre de 2014 y, en base a cada tipo de siniestro, gama de celular y tipo de cartera, se computó dentro de los equipos que se reemplazan cuántos fueron con celulares nuevos y cuántos con celulares reacondicionados.

  En base a los resultados obtenidos, se definieron cuáles deberían ser los valores más representativos para predecir el porcentaje de reemplazos por nuevos y por reacondicionados a futuro.

  Al igual que en los casos anteriores, en esta oportunidad se corrió un programa en VBA para abarcar todo el universo de posibilidades.

- La proporción de los celulares que se reemplazan por destrucción total versus los que se reemplazan por falta de alguno de sus componentes en el mercado

  En el último tiempo se ha detectado que muchos celulares siniestrados terminaban en reemplazo no debido a destrucción total sino que por el contexto económico y político que transita actualmente el país en el que, debido al cierre de las importaciones, no existen en muchas ocasiones en Argentina las partes necesarias para efectuar la reparación. En este último caso, y a pedido del Tomador de la póliza –el *carrier*–, la Aseguradora no cobra el deducible del 50% por lo que, a priori, tanto la frecuencia como el valor de los siniestros se han disparado en el último tiempo.

- Debido a lo anterior, otra de las variables que se modelaron en el presente análisis fue si el reemplazo se produjo por Falta de Partes o si fue un Reemplazo por Destrucción Total.

Al igual que en los casos previos, se analizó de manera mensual la base de siniestros desde Enero de 2013 a Diciembre de 2014 y, en base a cada tipo de siniestro, gama de celular y tipo de cartera, se computó dentro de los equipos que se reemplazan cuántos fueron por destrucción total y cuántos por falta de partes.

En base a los resultados obtenidos, se definieron cuáles deberían ser los valores más representativos para predecir el porcentaje de reemplazos por destrucción total y por falta de partes.

Al igual que en los casos anteriores, a los efectos del análisis se corrió un programa en VBA para abarcar todo el universo de posibilidades.

- Monto del siniestro promedio

Para cada tipo de siniestro, gama y cartera (propia o adquirida) se determinó el monto de siniestro promedio de:

  - Reparación
  - Reemplazo por equipo nuevo por destrucción total
  - Reemplazo por equipo reacondicionado por destrucción total
  - Reemplazo por equipo nuevo por falta de partes
  - Reemplazo por equipo reacondicionado por falta de partes

A los efectos de su cómputo se analizó de manera mensual la base de siniestros desde Enero de 2013 a Diciembre de 2014 y, en base a los resultados arrojados, se determinó la evolución de los mismos y se definieron los valores más representativos para predecir los montos de los siniestros promedios a futuro.

Cabe aclarar que, dado el contexto inflacionario que vive la Argentina, resulta imprescindible hacer este análisis de manera periódica para, eventualmente, corregir las primas que se les cobran a los asegurados.

Al igual que en los casos anteriores se corrió un programa en VBA para abarcar todas las posibilidades antes descriptas.

- Monto del deducible promedio

Al igual que en el caso del siniestro promedio, para cada tipo de siniestro, gama y cartera se determinó el monto del deducible promedio de:

  - Reparación
  - Reemplazo por equipo nuevo por destrucción total
  - Reemplazo por equipo reacondicionado por destrucción total
  - Reemplazo por equipo nuevo por falta de partes
  - Reemplazo por equipo reacondicionado por falta de partes

A los efectos de su cómputo se analizó de manera mensual la base de siniestros desde Enero de 2013 a Diciembre de 2014 y, en base a los resultados arrojados, se determinó la evolución de los deducibles y se definieron los valores más representativos para predecir los montos de los deducibles promedios a futuro.

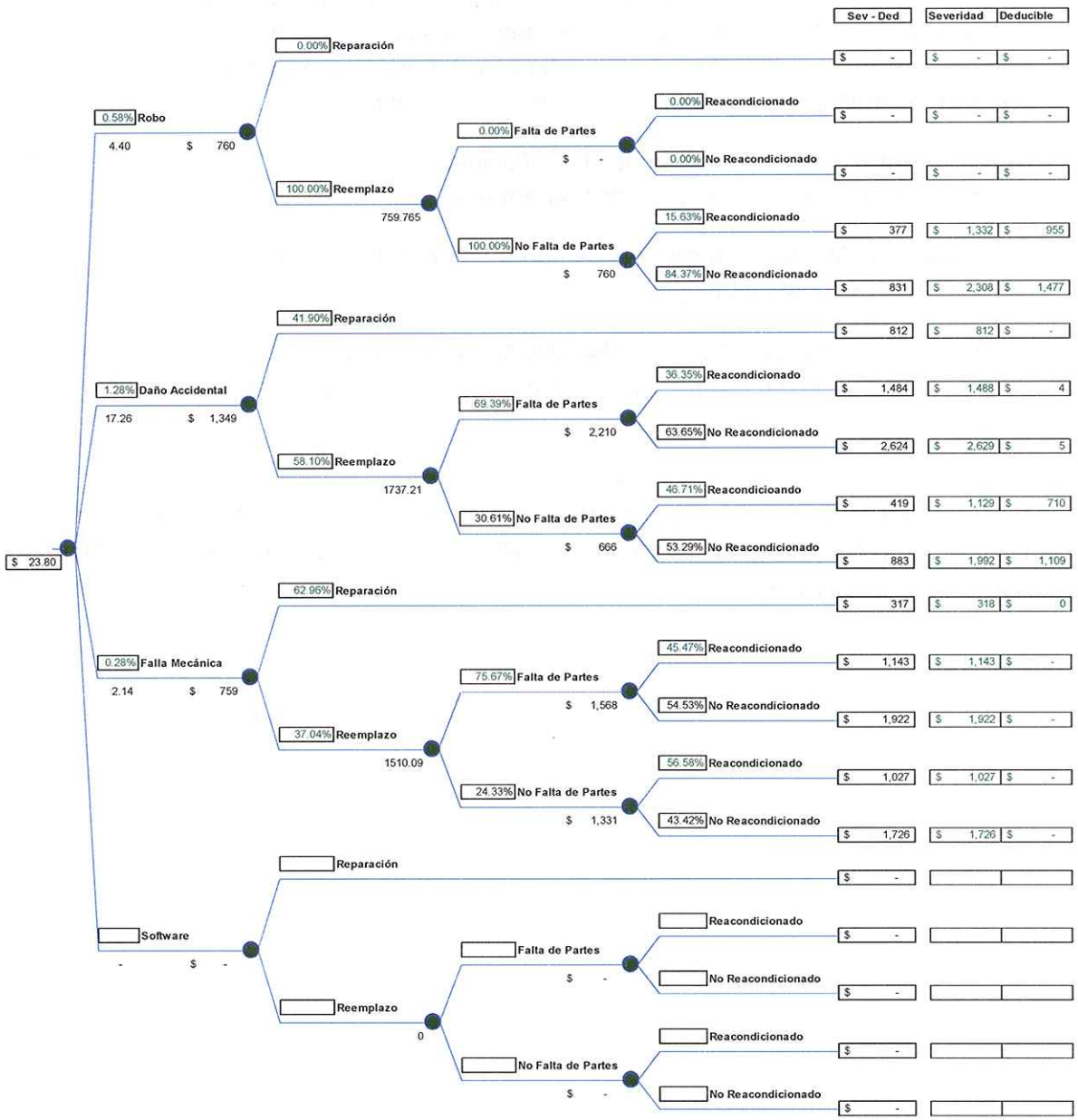
Cabe aclarar que, dado el contexto inflacionario que vive la Argentina, resulta imprescindible hacer este análisis de manera mensual.

Al igual que en los casos anteriores se corrió un programa en VBA para abarcar todas las posibilidades antes descriptas.

Todos los conceptos antes detallados determinan para cada gama, tipo de siniestro y tipo de cartera la Severidad del Programa tal como se explicita en la subsección siguiente.

Esquema para el cálculo de la Prima Pura

A fin de calcular la Prima Pura, y siguiendo la estructura de un Árbol de Decisión y los cálculos de la frecuencia y severidad antes explicados el esquema que se utilizó para calcular este concepto fue el siguiente:



La combinación de todas estas variables determina la Prima Pura o Costo Esperado del Siniestro.

Cabe aclarar que en el armado de la estructura previamente descripta se ha tenido en cuenta los pasos a seguir en la construcción de Árboles de Decisión desarrollados en el Marco Teórico de la presente Tesis.

Es decir, luego de definido el problema, en este caso obtener la mejor estimación del costo esperado de un siniestro o prima pura, se procedió a dibujar el árbol y, en base a información histórica, se le asignó una probabilidad condicional de ocurrencia a los diversos estados de la naturaleza posibles (probabilidad de robo, daño accidental, reparación, reemplazo, reemplazo por equipos reacondicionados, etc.).

Asimismo, y a fin de estimar los resultados de cada una de las posibles combinaciones de alternativas y resolver el problema calculando los valores monetarios esperados de cada nodo, tal como se mencionó previamente, se procedió a considerar los valores de severidades de reparación, reemplazo por equipos nuevos, reemplazo por equipos reacondicionados, etc. y sus deducibles asociados.

Finalmente, para obtener el costo esperado o prima pura se procede desde las ramas finales hacia el origen del árbol, calculando los valores esperados en los nodos de probabilidad hasta llegar al nodo raíz del árbol.

Agrupamiento por Segmentos

A fin de calcular los distintos valores de Prima Pura, y tal como lo describiéramos previamente, se hicieron las siguientes agrupaciones:

- Cartera Propia – Gama Base
- Cartera Propia – Gama Baja
- Cartera Propia – Gama Medium
- Cartera Propia – Gama Alta
- Cartera Propia – Gama Low Premium
- Cartera Propia – Gama Premium
- Cartera Propia – Gama Iphone
- Cartera Propia – Total de Gamas

- Cartera Adquirida – Gama Base
- Cartera Adquirida – Gama Baja
- Cartera Adquirida – Gama Medium
- Cartera Adquirida – Gama Alta
- Cartera Adquirida – Gama Low Premium
- Cartera Adquirida – Gama Premium
- Cartera Adquirida – Gama Iphone
- Cartera Adquirida – Total de Gamas

- Cartera Total – Total de Gamas

De lo anterior, se desprende que el modelo construido arroja 17 árboles de decisión distintos que permiten calcular la prima pura para cada segmento.

Determinación de las Variables relevantes para el cálculo de la Prima Pura

Tal como se explicó previamente, a fin de calcular para cada Gama de Celular y si es Cartera Propia o Cartera Adquirida cada una de las variables antes detalladas en el Árbol de Decisión se analizó el comportamiento de cada una de dichas variables tomando en consideración la evolución de éstas de forma mensual durante los últimos dos años y luego se seleccionó el valor que resultaba ser la mejor estimación del comportamiento a futuro.

Tal como se detalló previamente, a fin de calcular la evolución de todas estas variables se corrió un programa desarrollado en VBA para Microsoft Excel.

Determinación de la Prima de Tarifa y Premio

Tal como se describiera previamente, una vez obtenida la prima pura para cada segmento se le adicionó el nivel de comisiones, los impuestos, los gastos y el margen de beneficio de la compañía para arribar así a la prima de tarifa y premio.

Comparación entre el Premio obtenido y el realmente cobrado

Una vez efectuado el análisis se compararon los valores antes calculados con los actualmente cobrados para identificar GAPs entre uno y otro.

Análisis "What If"

Finalmente, y a fin de entender los factores de riesgo más relevantes, se sensibilizaron ciertas variables para ver qué sucedía con la siniestralidad total del programa si se efectuaban ciertas medidas o planes de Acción.

En tal sentido, los planes de Acción que se modelaron fueron:

- ¿Qué pasaría si se agregara al programa un deducible de reparación del 30% y un deducible por falta de partes correspondiente al 30% del valor que hubiese salido la reparación?
- ¿Qué pasaría si se hacen acuerdos con los talleres para aumentar el nivel de reparabilidad a un cierto porcentaje. Por ejemplo, el 60%?
- ¿Qué pasaría si para ciertas gamas se aumentara la prima en un "x"%?
- ¿Qué pasaría si se aumenta el nivel de los celulares reacondicionados para utilizarlos en siniestros de remplazos en desmedro de los celulares nuevos?

Creación de la Herramienta de Software a Utilizar

Tal como se describiera previamente, con el fin de obtener los valores correspondientes a la Prima Pura y demás datos necesarios para llegar al precio final a pagar por el cliente, se construyó una herramienta en MS Excel que, a grandes rasgos, contiene lo siguiente:

- Una primer hoja que contiene el árbol de decisión tal y como se ha presentado previamente en este mismo capítulo. Asimismo, en esta primera hoja, se pueden encontrar el resto de las variables requeridas para el cálculo del precio final (variables tales como impuestos, comisiones, gastos, margen de utilidad, etc.).

- Para alimentar cada una de las celdas que componen el árbol de decisión se generó una hoja por cada una de las siguientes variables: Frecuencia, Porcentaje de Distribución, Severidad y Deducible, donde dependiendo de su naturaleza, se van

completando con información que se encuentra en la base de expuestos y/o en la base de siniestros.

Luego, y en línea con lo expresado a lo largo de esta tesis, se cuenta con una base de expuestos y con una base de siniestros que son las que proporcionan toda la información requerida para abordar el modelo desarrollado.

- Tal lo dicho, una de las hojas de la herramienta de cálculo contiene un resumen segmentado por *dealer* y gama de celular, donde se totaliza, mes por mes del período estudiado, la cantidad de expuestos (personas aseguradas) correspondiente. Este resumen se obtiene a partir de una base completa de expuestos para este producto, con información histórica desde Enero de 2013 a Diciembre de 2014.

  Los campos más significativos de esta base histórica se han mencionado en el capítulo previo.

- Finalmente, en otra de las hojas de la herramienta de cálculo se encuentra la Base de Siniestros, donde se puede ver el detalle de todos los siniestros ocurridos en el período bajo estudio (incluyendo tanto siniestros aprobados como denegados).

  Los campos de dicha base de siniestros se han detallado en el capítulo previo.

Con la ayuda del lenguaje VBA utilizable en Excel, más la batería de herramientas propias que este software pone a disposición (macros, tablas dinámicas y fórmulas por doquier), se logró estudiar la información contenida en las bases de siniestros y expuestos permitiendo obtener las primas y premios que mejor representan a cada uno de los segmentos bajo estudio.

Capítulo 7. RESULTADOS

En el presente capítulo se detallan los resultados que surgen del análisis efectuado.

En tal sentido, en esta sección se presentan las primas puras, primas de tarifa y premio de cada segmento y se lo compara con el nivel actual que abonan los clientes y se identifican los GAPs resultantes.

Asimismo, en este capítulo se muestra cuál es el Loss Ratio de cada segmento y se lo compara con el Loss Ratio Permisible. Este último surge de la cotización original que se efectuó cuando se creó este programa.

Finalmente, en este capítulo se detallan algunas iniciativas que podrían revertir ciertos resultados y se cuantifica el impacto de las mismas. Para ello, se tomó a la cartera total y se cuantificó el impacto en el Loss Ratio.

Primas Teóricas versus Primas realmente cobradas en cada Segmento

A continuación, se presentan para cada segmento antes descripto la prima de tarifa que surgen del análisis efectuado y se lo compara con los niveles de prima de tarifa actuales:

			RESULTADOS SEGÚN ANÁLISIS EFECTUADO					
Cartera	Gama	Expuestos Actuales	Prima Pura	Prima de Tarifa	Premio	Prima de Tarifa Actual	GAP en %
	Base	549	$ 0.1	$ 0.1	$ 0.2	$ 19.5	15572%
	Low	7,647	$ 1.3	$ 2.8	$ 3.5	$ 22.0	674%
	Mid	25,169	$ 4.7	$ 10.2	$ 12.6	$ 28.1	176%
Propia	High	11,400	$ 9.3	$ 20.0	$ 24.8	$ 41.7	108%
	Low Premium	117,424	$ 17.3	$ 37.1	$ 45.9	$ 46.5	25%
	Premium	163,217	$ 45.5	$ 97.8	$ 121.0	$ 62.3	-36%
	Iphone	-	$ -	$ -	$ -	$ -	0%
Propia	**Total**	**325,406**	**$ 27.9**	**$ 60.0**	**$ 74.2**	**$ 52.2**	**-13%**
	Base	1,207	$ 0.2	$ 0.5	$ 0.6	$ 17.5	3724%
	Low	44,414	$ 3.1	$ 6.7	$ 8.3	$ 18.6	176%
	Mid	71,372	$ 4.8	$ 10.2	$ 12.6	$ 23.1	125%
Adquirida	High	43,300	$ 7.6	$ 16.4	$ 20.3	$ 37.7	129%
	Low Premium	-	$ -	$ -	$ -	$ -	0%
	Premium	194,149	$ 34.4	$ 73.9	$ 91.4	$ 56.6	-23%
	Iphone	782	$ 0.8	$ 1.8	$ 2.2	$ 97.9	5308%
Adquirida	**Total**	**355,224**	**$ 21.2**	**$ 45.6**	**$ 56.4**	**$ 42.7**	**-6%**
Total	**Total**	**680,630**	**$ 23.8**	**$ 51.2**	**$ 63.3**	**$ 47.3**	**-8%**

Tal como surge del cuadro anterior, donde para completar cada una de las filas se ha tomado 1 a 1 la información resultante de los 17 árboles de decisión mencionados en el capítulo "Construcción del Modelo", en el caso de las gamas más bajas la prima cobrada al asegurado es superior a la que surge del análisis. Incluso, en algunos casos la dispersión entre una y otra es muy alta.

De forma contraria, en el caso de la gama Premium (donde se concentra la mayor cantidad de los celulares expuestos) la prima que abona el cliente es inferior a la que se le debería cobrar.

Tal como el lector puede anticipar, en el caso antes descripto, los celulares de gamas bajas subsidian a los celulares de gamas altas. Esta situación se da muchas veces en seguros llamados colectivos en la que existe un grupo de asegurados bajo un mismo Tomador.

Este análisis permite identificar que las primas a nivel de cada segmento no están bien tarificadas.

En el agregado, la Prima de tarifa promedio que abona el cliente se encuentra un 8% por debajo del valor que se debería abonar. ($47,3 versus $51,2).

Esta brecha entre ambas primas se torna crítica pues en el agregado el programa está al límite de comenzar a arrojar pérdida pues la prima que abonan los asegurados casi no alcanza para cubrir los siniestros, pagar las comisiones y pagar los gastos de la compañía.

Como conclusión de este análisis, la compañía de seguros sujeta a estudio podría corregir las tarifas para cada gama de celulares buscando ser más competitiva en las gamas más bajas (bajando las primas) e incrementando las tarifas para las gamas altas a fin de que éstas sean suficientes para hacer frente a los compromisos de la aseguradora.

Comparación entre el Loss Ratio Observado y el Loss Ratio Permisible en cada Segmento

El siguiente cuadro detalla para cada segmento el Loss Ratio observado y lo compara con el Loss Ratio permisible:

Cartera	Gama	Expuestos Actuales	Loss Ratio Actual	Loss Ratio Permisible	GAP en %
	Base	549	0.3%	46.5%	-99.4%
	Low	7,647	6.0%	46.5%	-87.1%
	Mid	25,169	16.8%	46.5%	-63.8%
Propia	High	11,400	22.3%	46.5%	-52.0%
	Low Premium	117,424	37.1%	46.5%	-20.2%
	Premium	163,217	73.0%	46.5%	57.0%
	Iphone	-	0.0%	46.5%	-100.0%
Propia	**Total**	**325,406**	**53.4%**	**46.5%**	**14.9%**
	Base	1,207	1.2%	46.5%	-97.4%
	Low	44,414	16.9%	46.5%	-63.7%
	Mid	71,372	20.6%	46.5%	-55.7%
Adquirida	High	43,300	20.3%	46.5%	-56.4%
	Low Premium	-	0.0%	46.5%	-100.0%
	Premium	194,149	60.8%	46.5%	30.7%
	Iphone	782	0.9%	46.5%	-98.2%
Adquirida	**Total**	**355,224**	**49.6%**	**46.5%**	**6.7%**
Total	**Total**	**680,630**	**50.4%**	**46.5%**	**8.3%**

Al igual que como puede observarse en el análisis de la Prima Pura, en el caso del análisis del Loss Ratio las gamas bajas tienen un nivel de Loss Ratio por debajo del Loss Ratio Permisible. Sin embargo, este no es caso para la gama Premium en donde el Loss Ratio actual es bastante superior al nivel permitido.

Otra vez aquí, queda corroborado que las gamas bajas subsidian a las gamas superiores.

En el agregado el Loss Ratio global del programa se encuentra un 8.3% por encima que el Loss Ratio Permisible (50,4% versus 46,5%)

Tal como mencionara previamente, en muchas ocasiones un 1% de brecha en el Loss Ratio real versus el permisible puede arrojar una ganancia o una perdida considerable a un programa de seguros.

Análisis What If

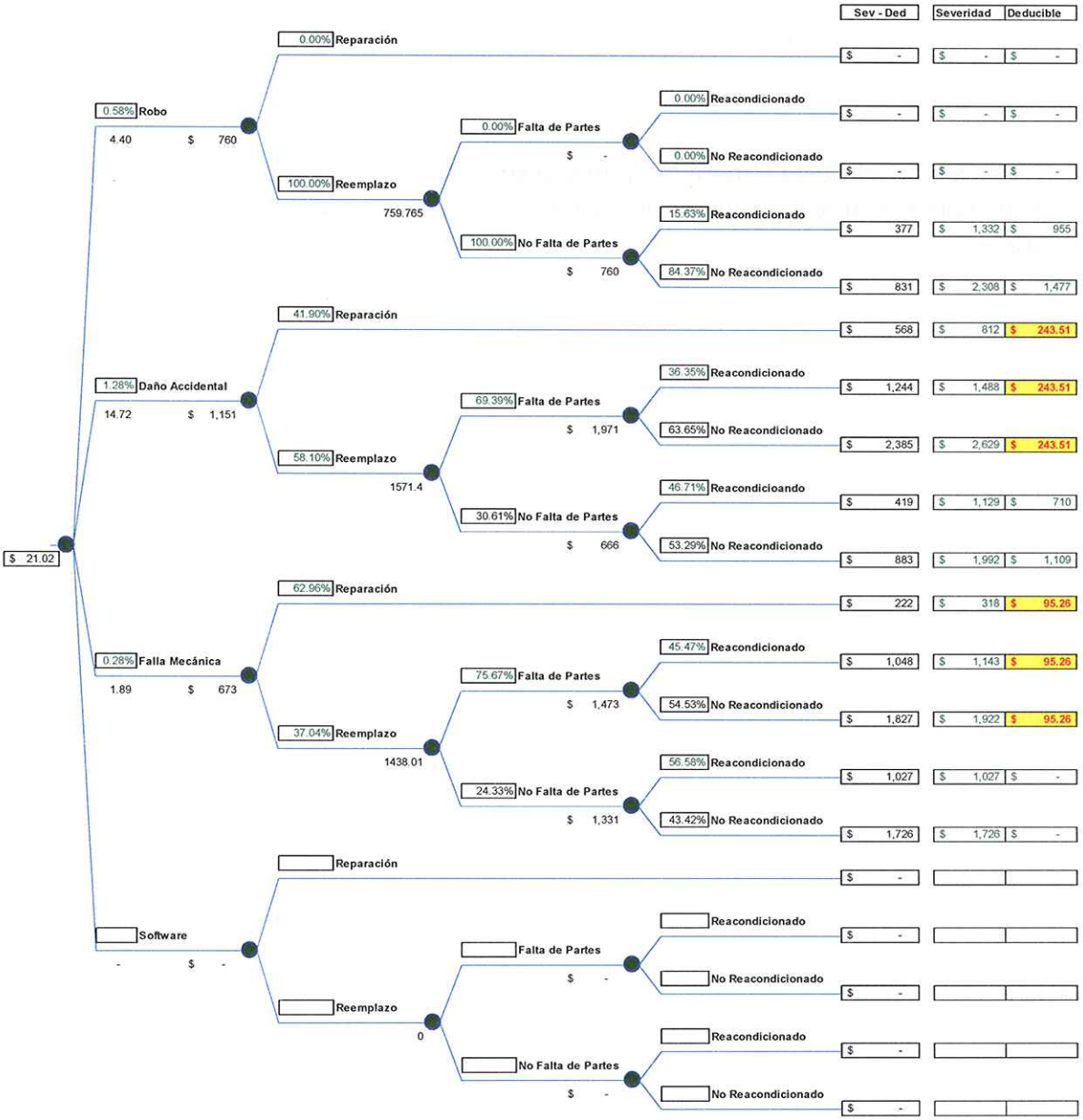
En esta sección se presentan algunas iniciativas para revertir el resultado del Programa y se detalla cuál sería el impacto en el resultado del Loss Ratio.

Introducción de un deducible de reparación del 30% y un deducible por falta de partes correspondiente al 30% del valor que hubiese costado la reparación

En el caso de que se introduzca este deducible el Loss Ratio del programa total se reduce de 50,4% a 44,5% según el siguiente detalle:

$$Loss\ Ratio = \frac{Costo\ Esperado\ Promedio}{Prima\ de\ Tarifa\ Promedio} = \frac{\$21{,}02}{\$47{,}26} = 44{,}5\%$$

A fin de calcular el Costo Esperado Promedio se ajustaron las celdas de Deducible del Árbol de Decisión (celdas sombreadas en amarillo) para reflejar el 30% antes descripto. El Árbol de Decisión en este caso es el siguiente:



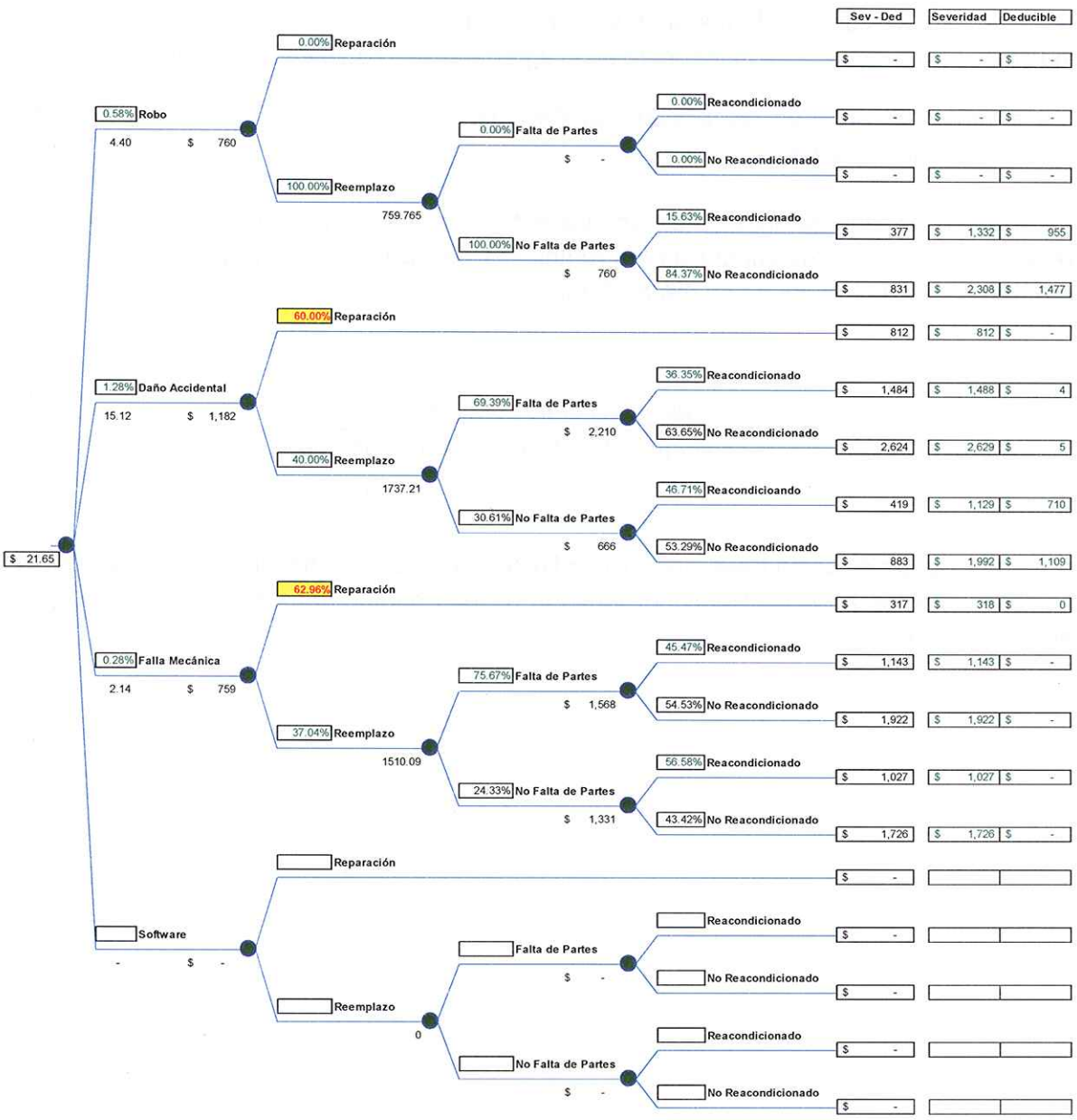
Esta medida permitiría estar a nivel global por debajo del Loss Ratio Permisible lo que garantizaría una ganancia para la Compañía de Seguros.

Incremento del nivel de reparabilidad hasta alcanzar un nivel del 60% o superior

En el caso de que la compañía de seguros trabaje con los talleres -vía incentivos- para incrementar el nivel de reparabilidad de los celulares siniestrados el Loss Ratio del programa total se reduciría de 50,4% a 45,8% según el siguiente detalle:

$$Loss\ Ratio = \frac{Costo\ Esperado\ Promedio}{Prima\ de\ Tarifa\ Promedio} = \frac{\$21{,}65}{\$47{,}26} = 45{,}8\%$$

A fin de calcular el Costo Esperado Promedio se ajustó en el Árbol de Decisión las celdas correspondientes al % de Reparación (celdas sombreadas en amarillo) según el siguiente detalle:



Incremento de la prima en ciertas gamas de celulares

En el caso de que se introduzca un aumento de las primas en ciertas gamas de celulares el Loss Ratio del programa total se reduce de 50,4% a 44,4% según el siguiente detalle:

$$Loss\ Ratio = \frac{Costo\ Esperado\ Promedio}{Prima\ de\ Tarifa\ Promedio} = \frac{\$23{,}80}{\$53{,}64} = 44{,}4\%$$

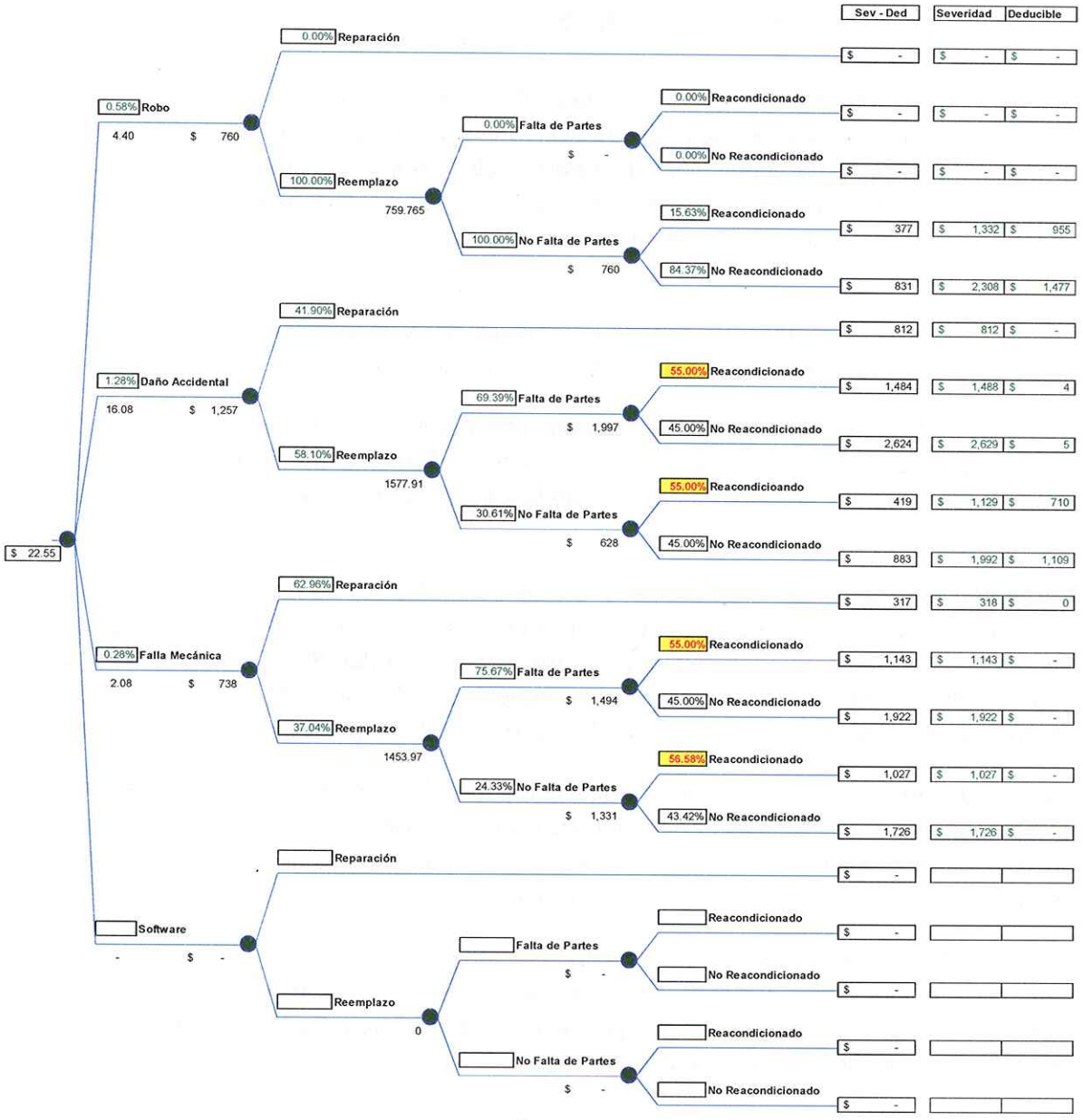
En este caso se aumentó la prima de la Gama Premium tal que sea suficiente para cubrir los costos esperados. Este incremento en el agregado se tradujo en un aumento del 13,5%.

Incremento del nivel de remplazos por Celulares Reacondicionados hasta alcanzar un nivel mínimo del 55%

En el caso de que se introduzca un aumento en el nivel de remplazos de celulares reacondicionados en desmedro de los remplazos por celulares nuevos el Loss Ratio del programa total se reduce de 50,4% a 47,7% según el siguiente detalle:

$$Loss\ Ratio = \frac{Costo\ Esperado\ Promedio}{Prima\ de\ Tarifa\ Promedio} = \frac{\$22{,}55}{\$47{,}26} = 47{,}7\%$$

A fin de calcular el Costo Esperado Promedio se ajustó en el Árbol de Decisión las celdas correspondientes al % de Reacondicionados (celdas sombreadas en amarillo) según el siguiente detalle:


Sev - Ded
Severidad
Deducible
$ 22.55
0.58% Robo
4.40 $ 760
0.00% Reparación
$ -
$ - $ -
100.00% Reemplazo
759.765
0.00% Falta de Partes
$ -
0.00% Reacondicionado
$ -
$ - $ -
0.00% No Reacondicionado
$ -
$ - $ -
100.00% No Falta de Partes
$ 760
15.63% Reacondicionado
$ 377
$ 1,332 $ 955
84.37% No Reacondicionado
$ 831
$ 2,308 $ 1,477
1.28% Daño Accidental
16.08 $ 1,257
41.90% Reparación
$ 812
$ 812 $ -
58.10% Reemplazo
1577.91
69.39% Falta de Partes
$ 1,997
55.00% Reacondicionado
$ 1,484
$ 1,488 $ 4
45.00% No Reacondicionado
$ 2,624
$ 2,629 $ 5
30.61% No Falta de Partes
$ 628
55.00% Reacondicioando
$ 419
$ 1,129 $ 710
45.00% No Reacondicionado
$ 883
$ 1,992 $ 1,109
0.28% Falla Mecánica
2.08 $ 738
62.96% Reparación
$ 317
$ 318 $ 0
37.04% Reemplazo
1453.97
75.67% Falta de Partes
$ 1,494
55.00% Reacondicionado
$ 1,143
$ 1,143 $ -
45.00% No Reacondicionado
$ 1,922
$ 1,922 $ -
24.33% No Falta de Partes
$ 1,331
56.58% Reacondicionado
$ 1,027
$ 1,027 $ -
43.42% No Reacondicionado
$ 1,726
$ 1,726 $ -
Software
- $ -
Reparación
$ -
Reemplazo
0
Falta de Partes
$ -
Reacondicionado
$ -
No Reacondicionado
$ -
No Falta de Partes
$ -
Reacondicionado
$ -
No Reacondicionado
$ -

Capítulo 8. CONCLUSIONES

Actualmente la información no es escasa sino, por el contrario, abundante. En este contexto, se hace imprescindible poder seleccionar la información relevante criteriosamente y analizarla para tomar decisiones exitosas. Este comportamiento no es ajeno a las compañías de seguros que cuentan con muchos datos de sus asegurados presentes y pasados así como la historia de los siniestros a su cargo y en muchas ocasiones no saben qué hacer con todos ellos.

Frente a este hecho, esta Tesis abordó la temática de la importancia de la explotación de datos para la correcta tarificación en las Compañías de Seguros.

Como se estableció a lo largo de la presente Tesis, el objeto que persigue un sistema de Tarificación es la obtención de primas equitativas para cada grupo homogéneo de asegurados teniendo siempre en cuenta que la solvencia de la compañía de seguros debe estar garantizada.

En tal sentido, para mejorar la precisión en las estimaciones de las primas, las bases de datos de las Compañías de Seguros deben segmentarse en grupos homogéneos y, así, los datos de cada grupo pueden explorarse, analizarse y modelarse. Como resultado de lo anterior, las compañías de seguros pueden predecir con mayor exactitud la probabilidad de ocurrencia de un siniestro y el monto de la indemnización a que ésta puede ascender.

Asimismo, y en línea con lo anterior, a fin de elaborar las tarifas, y respetando el principio de equidad, las aseguradoras deben identificar los factores de riesgo más significativos, es decir, los que explican de una mejor manera el comportamiento de la siniestralidad como variable endógena de una cartera de asegurados.

Si bien muchos de los factores de riesgo que afectan a las primas son obvios, pueden existir relaciones no intuitivas y sutiles entre las variables que son difíciles o imposibles de identificar sin aplicar análisis más complejos. Aquí también resulta primordial la explotación de datos para encontrar estas relaciones.

Esta Tesis ha demostrado que mediante la aplicación de técnicas de minería de datos las compañías de seguros pueden descubrir cuáles son los factores de riesgo que subyacen a su operatoria y, de esta forma, calcular primas más precisas para sus asegurados. Tal como se ha visto en el capítulo de resultados, habiendo aplicado minería de datos se ha podido calcular primas más certeras que las actualmente cobradas en la compañía, lo que permite obtener múltiples beneficios para el producto sujeto a análisis. Entre ellos, vale destacar los siguientes:

- Obtener ventajas competitivas frente a otras compañías de seguros,
- Mejorar su rentabilidad,

- Permitir orientar las campañas de marketing a los riesgos más sanos en detrimento de los riesgos más adversos,

- Definir y llevar adelante planes de acción para acotar sus variables críticas y así no poner en compromiso su operatoria.

Capítulo 9. REFERENCIAS

Centeno, Doffourt, García, Gómez, González, Granado, Loyo, Pérez & Pérez. (2011) Minería de Datos – El arte de sacar conocimiento de grandes volúmenes de datos

Gustavo Alexi Osorio González. Manual Básico del Seguro,

Diana Carolina Lancheros. (2011) Tarificación: elemento central de la actividad aseguradora – Aplicación a los Seguros Generales

Carmen López. (2006) Análisis avanzados de grandes volúmenes de datos en el sector seguros

S. Hameetha Begum (2013). Data Mining Tools and Trends – An Overview

Galit Shmueli. Nitin R. Patel. Peter C Bruce. (2005) Data Mining in Excel: Lecture Notes and Cases

Bharati M. Ramageri. Data Mining Techniques and applications

Ruxandra Petre. Data Mining Solutions for the Business Environment

Max Bramer (2007), *Principles of Data Mining,* Londres, Reino Unido, Springer.

Neftalí de Jesús Calderón Méndez. (2006) Minería de Datos: Una herramienta para la toma de decisiones

Carlos Bousoño Calzón, Antonio Heras Martínez, Piedad Tolmos Rodríguez Piñero. Factores de Riesgo y cálculo de primas mediante técnicas de aprendizaje

Hebe Alicia Cadaval, Árboles de Decisión – Otra herramienta del modelo general