Anales del Instituto de Actuarios Españoles, 4ª época, 23, 2017/1-19

BIG-DATA ANALYTICS IN INSURANCE

BIG-DATA ANALYTICS EN SEGUROS

Alemar E. Padilla-Barreto [1,2], Montserrat Guillen[1], Catalina Bolancé[1*]

Abstract

The big-data revolution has impacted the insurance industry more than expected, to become a paradigmatic example of what the new digital economy is. The large amount of data and predictive modeling in insurance represents a turning point and a golden opportunity to channel the theory of risk to the prediction of losses. The changes are radical and demand deep transformations at the organizational level. In this paper we present some reflections on what the incorporation of Analytics implies in an insurance company and we show its inherent complexity through a case of success.

Keywords: Big data**,** insurance, modelling, data analytics, lines of business, ROC curve

Resumen

La revolución del *big-data* ha impactado en el sector asegurador más de lo que se esperaba, hasta convertirse en un ejemplo paradigmático de lo que es la nueva economía digital. La gran cantidad de datos y la modelización predictiva en seguros representan un punto de inflexión y una oportunidad de oro para canalizar la teoría del riesgo hacia la predicción de las pérdidas. Los cambios son radicales y demandan transformaciones profundas a nivel organizacional. En este trabajo presentamos algunas reflexiones sobre lo que supone la incorporación del *Analytics* en una compañía de seguros y mostramos su inherente complejidad mediante un caso ya testado con éxito.

Palabras clave: Big-data, seguros, modelización, análisis de datos, líneas de negocio, curva ROC

Las autoras agradecen la ayuda recibida en el Proyecto ECO2016-76203-C2-2-P, ICREA Academia y AGAUR Programa de Doctorados Industriales.

[1] Dpto. Econometría, Estadística y Economía Aplicada, Riskcenter-IREA; Universitat de Barcelona, Av. Diagonal, 690, 08034 Barcelona.
[2] Zurich España, Vía Augusta, 200, 08021 Barcelona.
* Autor para la correspondencia (bolance@ub.edu)

DOI 10.26360/2017_1

1. Introducción

En la actualidad, la sociedad se adapta a un entorno en el que todo su dinamismo queda registrado en cuestión de segundos. Vivimos en una era en la que la mayoría de las fuentes de información están digitalizadas. Redes sociales, páginas web, *smartphones*, dispositivos telemáticos, entre otros, son los responsables de nutrir grandes sistemas de información. El cambio es constante, así como lo es la demanda de información. Por un lado, los individuos necesitan saber más, de ahí su urgencia por estar conectados y, por otro, las organizaciones necesitan sacar ventaja de la información disponible, lo cual supone descubrir aspectos - hasta ahora desconocidos - vinculados con el comportamiento de sus clientes, socios, riesgos, costes y operaciones, así como de la sociedad en general.

La transformación ha sido vigorosa y, sin lugar a dudas, es sinónimo de reinvención. Más allá de las distancias, los individuos, las empresas, los países y los continentes en su totalidad han tenido que adaptarse a los nuevos desafíos en la manera de entender y analizar los datos. Las posibilidades son innumerables y las ventajas se encuentran en manos de aquellas compañías capaces de adaptarse rápidamente. La era del big-data llegó para quedarse y el sector asegurador, como era de esperar, se ha visto afectado por esta vorágine en primera persona. Por ejemplo, la posibilidad de conseguir más información y, a su vez, más detallada de los asegurados plantea, entre otros, cambios en los modelos de tarificación.

La acumulación masiva de datos ha sido una práctica habitual en las compañías aseguradoras. Los ficheros contienen información individualizada y longitudinal para cada línea de negocio. El sistema de información es tan complejo, que actualmente los clientes pueden interactuar a través de distintos canales de mediación como: agentes y corredores, *callcenters*, redes sociales, internet y bancaseguros. En éste sentido, el aprovechamiento de los datos suele segmentarse y adaptarse a las necesidades y requerimientos de cada departamento. Aunque en muchos casos sea frecuente el uso de variables comunes, como por ejemplo la edad, el sexo, el tipo de póliza, la prima pagada y los tipos de descuentos, la realidad es que cada departamento incorpora variables específicas vinculadas a sus objetivos de negocio, suelen tener autonomía sobre los análisis que realizan y trabajan bajo una cultura de independencia interdepartamental.

Hasta ahora, los análisis a nivel interno habían estado orientados al informe (*reporting*), control y seguimiento de indicadores. Pocas empresas habían sido capaces de establecer - en la práctica- verdaderas sinergias entre departamentos. En general, el tiempo ha sido considerado como un recurso

limitado, donde lo urgente se ha antepuesto al cambio. Sin embargo, el reciente crecimiento digital ha desafiado la forma de entender el marco teórico asegurador, haciendo necesario que las entidades aseguradoras se redefinan a todo nivel.

La forma de analizar los datos ha cambiado y con ello una parte de la estructura interna del negocio. Internamente, el primer cambio ha sido a nivel cultural, es decir, entender que existe una necesidad no cubierta y hasta hace algunos años inexplorada, para la cual se justifica que se destinen parte de los recursos de la compañía y, entre otras acciones, se invierta tiempo y dinero en el diseño y ejecución de un plan estratégico basado en una cultura de datos – *data driven*- (ver McAfee y Brynjolfsson (2012) para reflexiones adicionales sobre lo que implica la revolución empresarial en torno al *big-data*). El segundo cambio ha sido actuar en consecuencia, siendo uno de los primeros pasos la actualización de los sistemas informáticos – nueva tecnología, formación de los equipos de trabajo, actualización de procesos, familiarización con el uso de datos disponibles -. De forma casi simultánea, también se ha producido la integración de la "ciencia de los datos" en los nuevos modelos que estudian los seguros (véase Guillen, 2016).

Todo esto ha derivado en la creación de equipos de trabajo, con perfiles interdisciplinares y especializados, capaces de explotar la gran cantidad de datos disponibles y, a su vez, aportar valor añadido al resto de áreas de la compañía y por ende al negocio. Los denominados equipos de *big-data analytics*, *analytics* o *advanced analytics* son cada vez más comunes dentro de las compañías de seguros. Su principal objetivo es la aplicación de técnicas analíticas avanzadas sobre grandes volúmenes de datos (Russom, 2011).

Nuestra principal contribución es la de dar a conocer la visión empresarial y técnica de lo que debería ser un Departamento de *Analytics* dentro de una compañía de seguros. Para ello, estableceremos algunas reflexiones en torno a éste tema y expondremos un caso de éxito basado en el análisis de la retención de clientes, donde se vincula la información de dos líneas de negocio de forma simultánea.

La estructura de éste artículo es la siguiente. En la Sección 2 nos centraremos en explicar la visión empresarial del Departamento de *Analytics*. En la Sección 3 presentaremos una aplicación de *big-data analytics* en seguros de no vida. Finalmente, en la sección 4 expondremos nuestras conclusiones y recomendaciones.

2. Visión empresarial del departamento de Analytics

En esta sección exponemos a grandes rasgos lo que consideramos deberían ser las responsabilidades de un Departamento de *Analytics*, su posición dentro de la estructura organizativa de la empresa, las ventajas de su creación y los retos que supone contar con un departamento de tal envergadura.

2.1 Responsabilidades

Entre las principales responsabilidades de un Departamento de *Analytics* se encuentran:

- Responder a las interrogantes de negocio planteadas a nivel directivo.
- Ofrecer soluciones de negocio rápidas a iniciativas basadas en datos.
- Identificar nuevas áreas de oportunidad.
- Dar soporte al resto de departamentos de la compañía.

2.2 Estructura organizacional

El Departamento de *Analytics* debe funcionar desde una perspectiva global del negocio, por ello debería depender directamente de la presidencia de la compañía, lo cual le proporcionaría autonomía sobre sus análisis y propuestas. La dependencia de otro departamento no está exenta de riesgos, pues su desarrollo quedaría supeditado a los criterios e intereses de un área concreta de la compañía en general, aun cuando esta última se encuentre alineada con las directrices generales de la empresa.

Por otra parte, el Departamento de *Analytics* necesita trabajar en paralelo con el departamento de tecnologías de la información (*Information Technology*) pues, en general, es este último el responsable de proporcionar las bases de datos necesarias para su posterior análisis, además de disponer de los recursos para su procesamiento inicial.

2.3 Roles analíticos

Harrington (2014) resalta la importancia del rol de las personas dentro de un programa de *analytics*. Dicho autor centra su énfasis en una gobernanza clara, el patrocinio necesario a nivel ejecutivo y el debido acceso la capacitación y los recursos necesarios. En éste sentido, establece las principales funciones organizativas – lo cual supone la creación de equipos

de trabajo específicos – para la creación y habilitación de modelos analíticos operativos, como lo son:

- Desarrollo de *Enterprise Analytics*
- Equipo de Arquitectura Técnica
- Equipo Empresarial / Funcional
- Equipo de gestión del cambio
- Equipo de la biblioteca de datos de *Analytics*

2.4 Ventajas

Entre las principales ventajas de la creación de un Departamento de *Analytics* se encuentran:

- La rapidez con la que se podrían abordar transacciones de negocio específicas.
- Generación automática y disponibilidad inmediata de diferentes tipos de informes.
- Monitorización en tiempo real de los indicadores de negocio.
- Seguimiento efectivo de resultados, derivados de implementaciones operativas.
- Implementación de reglas para la detección de fraudes en tiempo real.
- Mejor segmentación de los riesgos.
- Análisis de dependencias entre tipos de productos, características de los clientes, etc.
- Detección anticipada de abandonos.
- Ofrecer pólizas personalizadas, adaptadas a las características de los consumidores.
- Tarificación de productos con mayor precisión (ver Swedloff, 2014 para más detalles sobre sus implicaciones).

2.5 Fortalezas y debilidades

La incorporación de un Departamento de *Analytics*, favorece la adopción de nuevas estrategias de negocio basadas en datos. Así mismo, ayuda con el establecimiento de sinergias entre los distintos departamentos y facilita el intercambio de información entre equipos de trabajo. Además, beneficia la capacidad de reconocer el tipo de datos que le es útil a cada departamento, sus outputs complementan los análisis existentes y sirven como puente para la transición desde las pólizas clásicas a las pólizas telemáticas, sobre todo en los seguros de automóvil.

En contraposición, Russom (2011) menciona que los perfiles técnicos y habilidades inadecuadas, representan las principales barreras para el análisis de grandes volúmenes de datos. Así mismo, en caso de que el equipo de *analytics* no sea adecuadamente respaldado termina relegado y, por lo tanto, sus proyectos son vistos y acogidos por sus compañeros como acciones aisladas e injustificadas, en vez de relevantes para la compañía. Por otra parte, las posibles restricciones de acceso a los datos, junto a los retrasos que pudiesen generarse en la entrega de los mismos, debido a la dependencia de un proveedor de los datos (con frecuencia IT- *Information Technology*), le restan dinamismo a los proyectos y posteriores análisis del Departamento de *Analytics*.

3. Analytics en seguros de no vida

A continuación, presentaremos un ejemplo de la aplicación de big-data en seguros. El objetivo de este ejemplo es el estudio de la retención de clientes con pólizas contratadas simultáneamente en dos líneas de negocio distintas: seguro de autos y seguro de hogar. Para ello, se evalúa el ajuste de cuatro modelos predictivos en base a dos criterios diferentes. Este ejemplo es complementario al análisis realizado en Bolancé *et al.* (2016), donde los clientes exclusivamente habían contratado un seguro de auto o un seguro de hogar.

El software utilizado para el diseño de los modelos ha sido R y para el tratamiento de las bases de datos R y SAS.

3.1 Datos

Los datos han sido proporcionados por una importante compañía de seguros en España. Analizamos una muestra de 22.064 clientes, que se corresponden con una muestra aleatoria seleccionada entre aquellos clientes que habían decidido asegurar dos riesgos diferentes, concretamente auto y hogar, con la misma compañía. La información se obtiene a nivel cliente y póliza, para lo cual se tuvo en cuenta una base de datos con pólizas de auto, otra con pólizas de hogar y una con información sobre sus respectivas reclamaciones. Cada conjunto de datos contiene una variable "clave" a través de la cual se identifica a cada tipo de cliente particular. Por lo que la obtención de la base de datos utilizada en el proceso de modelización requirió la combinación de distintas fuentes de información, en este caso procedentes del mismo departamento de "*pricing*".

A efectos de este estudio sólo consideraremos clientes con una póliza de auto y una póliza de hogar, estos se corresponden con el 75,89% de los clientes que tienen contratadas pólizas en ambas líneas; el resto son clientes con más de una póliza contratada en una o en las dos líneas de negocio analizadas.

Finalmente, desde la perspectiva analítica hemos dividido la muestra en dos partes: 70% para entrenamiento (training set) y 30% para prueba (test set), de modo que podamos evaluar el ajuste de los modelos "a posteriori" o con datos que no han sido utilizados en la estimación o entrenamiento del mismo. Respecto a dicha división no existe una forma única que permita decidir cómo ha de realizarse la partición de los datos de cara a la creación de las bases de datos de entrenamiento y prueba. Sin embargo, lo que sí es cierto es que se han de tener en cuenta ciertas consideraciones relacionadas con la elección de ambas submuestras y la evaluación del modelo, para garantizar que los resultados sean generalizables y, a su vez, que el modelo sea extrapolable. En este sentido, existen factores importantes como el tamaño de la muestra y la dispersión de las submuestras de prueba y de entrenamiento, que no debería ser muy distinta (ver Myatt, 2007 y Dobbin y Simon, 2011). De todos modos, parece haber un consenso no explícito que ha hecho que la partición entrenamiento-prueba equivalente a 70%-30% sea la más utilizada en la práctica.

En las Tablas 1, 2 y 3 se describe la información utilizada en el ajuste y/o entrenamiento de los distintos modelos. La variable dependiente puede referirse a tres tipos de decisión por parte del cliente:

Si el cliente decide renovar o no su póliza de hogar, al margen de lo que haga con la póliza de autos.

1. Si el cliente decide renovar o no su póliza de autos, al margen de lo que haga con la póliza de hogar.
2. Si el cliente decide renovar tanto su póliza de autos como su póliza de hogar o, por el contrario, no renovar una o ninguna.

Entre las variables explicativas de la decisión de renovar existen tres grupos, las que se utilizan tanto para explicar la renovación de la póliza de hogar como de autos (Tabla 1), las que se utilizan para modelizar la renovación o no de la póliza de autos (Tabla 2) y las que se utilizan para modelizar la renovación o no de la póliza de hogar (Tabla 3). Para explicar la renovación conjunta de ambas pólizas se utilizan todas las variables descritas en las tablas 1, 2 y 3.

Tabla 1. Variables comunes utilizadas en cada modelo

Relacionadas con	Variables
Tomador	Sexo, Edad, Otras pólizas contratadas (1= Si, 0= No), Prima de la última renovación, Monto total de primas pagadas
Póliza	Antigüedad de la póliza, Diferencia de primas, Descuentos, Bonos, Tipo de pago (A=anual, S=semestral, T=trimestral), Tipo de mediador , Suplementos, Recargos, Ratio de cancelación por mediador

Fuente: Muestra de asegurados en el ramo de autos y hogar, 2015

Tabla 2. Variables relacionadas con el vehículo

Relacionadas con	Variables
Riesgo asegurado	Tipo de vehículo, Edad del primer conductor, Segundo conductor (Sí , No), Potencia, Peso/potencia, Número de asientos

Fuente: Muestra de asegurados en el ramo de autos, 2015

Tabla 3. Variables relacionadas con la vivienda

Relacionadas con	Variables
Riesgo asegurado	Tipo de vivienda, continente y contenido asegurado

Fuente: Muestra de asegurados en el ramo de hogar, 2015

3.2 Modelos

En ésta sección se describen brevemente los cuatro modelos de clasificación que se han utilizado en este estudio.

- Regresión logística
- Árboles de decisión condicionales (CTREE)
- Maquina vector soporte (SVM-*Support Vector Machine*)
- Redes neuronales (NN-*Neuronal Network*)

Es importante mencionar que el modelo de regresión logística, los árboles condicionales, las máquinas de vectores de soporte y las redes neuronales son utilizados en el contexto del aprendizaje supervisado. Posteriormente, el resultado numérico obtenido en cada caso es la probabilidad de renovación

de cada cliente basado en sus características personales, las particularidades de la póliza y las del objeto de riesgo. Es importante resaltar que además de modelizar la probabilidad de renovación de las pólizas de hogar y auto de forma independiente, también hemos querido presentar un modelo en el que se incluyen todas las variables tanto de hogar como de auto, cuya variable respuesta nos da información acerca de la renovación simultánea de ambas pólizas, es decir, obtendremos la propensión (probabilidad) de cada cliente a la renovación de sus dos pólizas durante el mismo periodo.

El modelo de regresión logística es un modelo lineal generalizado de respuesta binaria ampliamente utilizado a nivel actuarial (ver McCullagh y Nelder (1989) para más detalles sobre los modelos lineales generalizados y sus aplicaciones). Su ventaja reside en que es un modelo conocido, cuyos resultados son fáciles de entender y explicar. Así mismo, resulta bastante informativo dado que a-priori aporta pistas sobre la influencia de las variables en la modelización, lo cual es ideal al momento de decidir la configuración final del modelo. Al mismo tiempo sirve como referente, en el supuesto de que se deseen comparar resultados utilizando modelos más complejos. Dos de sus principales limitaciones son su forma funcional cerrada y la dificultad de extrapolar los resultados obtenidos en caso de que exista sobreajuste (*overfitting*).

Los árboles condicionales (CTREE-*Conditional Tree*) son un tipo especial de árbol de decisión en los que la selección de las variables se realiza en dos fases, primero se formula una hipótesis global de independencia en términos de m hipótesis parciales. Es decir, se evalúa si existe dependencia entre la variable respuesta y cada una de las variables explicativas, en caso de no poder rechazar la hipótesis nula de independencia planteada se detiene el proceso recursivo. En contraposición, si la hipótesis global de independencia es rechazada, el siguiente paso es medir el nivel de asociación entre la variable dependiente y cada una de las variables explicativas, lo cual permite generar nuevas divisiones del árbol de manera secuencial (ver Hothorn *et al.* (2006) para más detalles). Análogos a los árboles de decisión clásicos, los árboles condicionales se caracterizan por ser poderosas herramientas de clasificación y visualización, además de ser útiles en situaciones en las que el objetivo es agrupar segmentos de clientes, identificar características de un grupo, toma de decisiones de negocio, etc (ver por ejemplo Guelman *et al.* (2014) para más detalles sobre la implementación de árboles condicionales en la venta cruzada de pólizas de seguros). Otra de sus ventajas es su versatilidad en el caso de que existan relaciones no lineales y en el manejo de variables numéricas y categóricas de forma simultánea.

Las máquinas de vectores de soporte (SVM-*Support Vector Machine*) representan un método de predicción alternativo y comúnmente utilizado en problemas de clasificación, análisis de regresión y detección de valores extremos (Boser *et al.*, 1992). Su principal característica es que los datos son mapeados en un espacio de dimensión superior, en donde las clases de la variable respuesta se separan mediante un hiperplano de división óptimo (ver Suykens *et al.*, 1999, Meyer y Wien, 2001, Meyer *et al.*, 2012 y Hornik *et al.*, 2006). Algunas de las ventajas de éste método es que el uso de funciones núcleo (*kernel*) simplifica la selección del borde de separación entre clases, permitiendo que esta tenga una forma no lineal, y los resultados son fácilmente extrapolables a otros datos distintos a los de la muestra. Por otra parte, entre sus principales limitaciones destacan la velocidad y el tiempo de ejecución de los algoritmos durante las fases de entrenamiento y prueba.

Las redes neuronales son métodos complejos de procesamiento de información, inspirados en asociaciones neuronales biológicas (véase Hastie, 1998). Usualmente son vistas como especies de "cajas negras", dado que los resultados son poco intuitivos y, además, aportan poca información acerca de la influencia de las variables explicativas sobre la variable respuesta, dicho análisis deberá realizarse a-posteriori. Entre sus principales ventajas destaca su habilidad para detectar relaciones complejas no lineales entre las variables explicativas y la variable respuesta y su capacidad para aprender de tales relaciones. En este sentido, los pesos dentro de la red se ajustan de forma gradual durante la fase de entrenamiento con el objetivo de reducir las diferencias entre el valor real y el valor predicho de la variable respuesta (para más detalles ver Tu (1996) quien expone las ventajas y desventajas entre el uso de redes neuronales y regresión logística en predicciones médicas). Así mismo, el proceso de modelización no requiere de la especificación de ningún modelo a-priori, a diferencia de otras técnicas de modelización no lineal (para más detalles ver Livingstone *et al.*, 1997). Entre sus debilidades más importantes se sitúan la sensibilidad de los resultados respecto al ajuste de los parámetros que controlan el algoritmo; su poca versatilidad en el caso de que se presenten correlaciones extremas entre las variables explicativas utilizadas y, finalmente, su propensión al sobreajuste.

Medidas de para evaluar la capacidad predictiva de los modelos

Los métodos utilizados estiman la probabilidad de clase p_i, es decir, la probabilidad de que la i-ésima póliza sea retenida, lo cual permite definir las clases predichas al comparar p_i con diferentes puntos de corte de clasificación $t \in [0, 1]$. Cada una de estas comparaciones produce una

matriz de confusión, a través de la cual es posible determinar la proporción de verdaderos positivos (TP), falsos positivos (FP), verdaderos negativos (TN) y falsos negativos (FN) para cada modelo.

Con el objetivo de evaluar el acierto de los modelos en la predicción de la deserción, se han escogido dos criterios diferentes. El primer criterio (C1) está basado en la máxima sensibilidad y especificidad, mientras que el segundo criterio (C2) representa el área bajo la curva ROC (AUC-*Area Under Curve*). Así,

$$C1 = max(sensibilidad + especificidad) \\ = max\,(TP/(TP + FN) + TN/(TN + FP))$$

$$C2 = max(AUC)$$

3.3 Resultados

A continuación, en las Figuras 1 y 2 se muestran los resultados de las curvas ROC asociadas a cada uno de los modelos propuestos y para cada una de las líneas de negocio -hogar y auto- analizadas. La Figura 1 se corresponde con las curvas de ROC de los modelos para el ajuste de la retención en el seguro de hogar. En dicha figura se observan diversos cruces entre curvas y en general un mejor ajuste para los árboles condicionales y la máquina de vectores de soporte. En la Figura 2, asociada a la retención en el seguro de autos, las curvas ROC de los distintos modelos presentan un comportamiento más homogéneo, excepto al inicio de las curvas ROC donde se aprecia que los resultados del árbol condicional son un poco mejores.

El hecho de que las curvas se crucen implica que no hay un modelo que domine o mejore completamente al resto. Es decir, en función de la especificidad deseada el modelo óptimo podría variar.

Hasta aquí la propensión de los clientes a renovar es vista desde la perspectiva de hogar y desde la perspectiva de auto por separado. Sin embargo, la Figura 3 muestra los resultados de la propensión a la renovación de ambas pólizas (hogar y auto) de los mismos clientes en función de todas sus características, es decir, los modelos han sido definidos de manera que se tuviesen en cuenta todas las variables de hogar y de auto de forma simultánea. En esta figura podemos observar múltiples intersecciones, y aunque no existe una curva que domine sobre el resto, los resultados de la regresión logística, CTREE y SVM son mejores que los obtenidos con la red

neuronal. Por cada intersección tendremos diferentes niveles de falsas alarmas donde un clasificador supera a las otras.

Figura 1. Hogar – curvas ROC

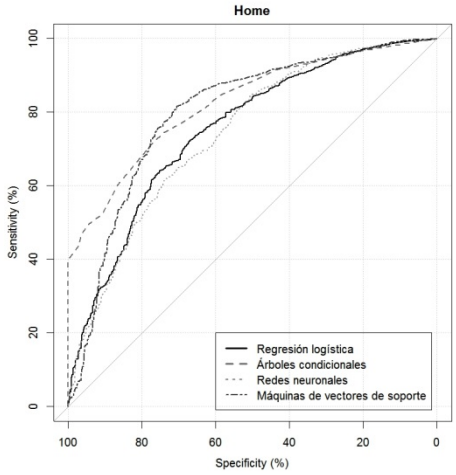


Figura 2. Auto – curvas ROC

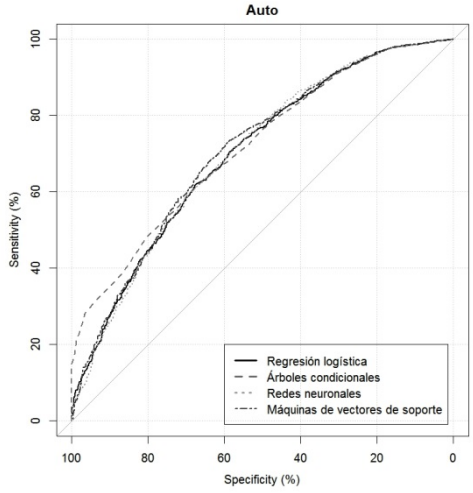
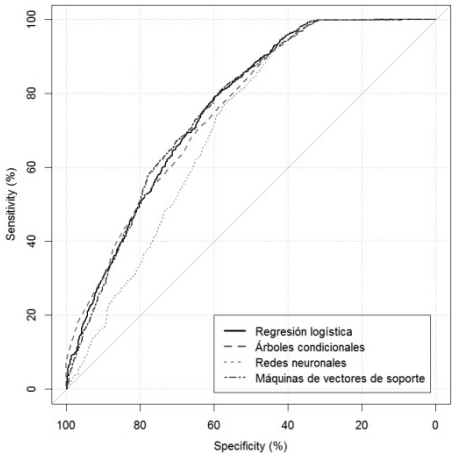


Figura 3. Hogar y Auto – curvas ROC



Las Tablas 4, 5 y 6 muestran los resultados de C1 y C2 utilizando el conjunto de datos de prueba. En general, el valor de los criterios varía según la línea de negocio. En el caso de hogar el árbol condicional es el que tiene un mejor ajuste, mientras que en auto el SVM y CTREE son los mejores. Cuando la información de ambas líneas de negocio es fusionada en un solo modelo los resultados de los modelos que pudieron ajustarse son muy parecidos, excepto para la red neuronal.

Tabla 4. Criterios de desempeño para cada modelo en los datos de prueba de Hogar

		Modelos			
		Regresión logística	**Árbol condicional**	**Red neuronal**	**Máquina de vectores de soporte**
Criterios	C1	1.39	1.49	1.35	1.21
	C2 (%)	75	83	74	80
Umbral óptimo	C1	0.94	0.92	0.91	0.97

Fuente: Cálculos propios

Tabla 5. Criterios de desempeño para cada modelo en los datos de prueba de Auto

		Modelos			
		Regresión logística	**Árbol condicional**	**Red neuronal**	**Máquina de vectores de soporte**
Criterio	C1	1.29	1.29	1.29	1.32
	C2 (%)	70	72	70	71
Umbral Óptimo	C1	0.89	[0.89-.090]	0.89	0.97

Fuente: Cálculos propios

Tabla 6. Criterios de desempeño para cada modelo en los datos de prueba de Hogar y Auto

		Modelos			
		Regresión logística	**Árbol condicional**	**Red neuronal**	**Máquina de vectores de soporte**
Criterio	C1	1.39	1.36	1.37	1.38
	C2 (%)	76	76	71	76
Umbral óptimo	C1	0.87	0.78	0.83	0.96

Fuente: Cálculos propios

Teniendo en cuenta el criterio C1 y considerando el modelo que mejor ajusta en cada caso, hemos construido las matrices de confusión[3] utilizando el "Umbral óptimo" que aparece en la última fila de las tablas 4, 5 y 6. Dicho "Umbral óptimo" equivale al valor de la probabilidad a partir del cual se predice que el individuo renovará su póliza. El objetivo es el de visualizar desde otra perspectiva el poder predictivo de cada modelo. A partir de los resultados de la Tabla 7 se observa que el porcentaje de aciertos sobre la renovación o no del seguro de hogar es del 71%. Por su parte, en la Tabla 8 se observa un porcentaje de acierto en la renovación de los seguros de auto es del 66%. Mientras que para el caso conjunto (hogar y auto) obtenemos un 75% de casos bien clasificados.

[3] La matriz de confusión es una tabla de frecuencias cruzadas que compara los valores reales de la variable con los valores predichos, y determina en cuantos casos el modelo se equivoca tanto en un sentido (renovación) como en otro (no renovación).

Con el objetivo de analizar si la capacidad predictiva de los modelos es homogénea en función de grupos de asegurados, hemos calculado los porcentajes de aciertos para hombres y mujeres. Los resultados no muestran diferencias entre dichos porcentajes en los modelos para el seguro de autos y en los modelos para ambos ramos. Por el contrario, el porcentaje de aciertos es mayor para los hombres en los modelos para el seguro de hogar, en la práctica, este último resultado implica que la compañía podría considerar que las mujeres se comportan de un modo distinto al esperado en un mayor número de casos.

Tabla 7. Matriz de confusión del modelo óptimo (CTREE) en el seguro de hogar. Umbral óptimo c=0.92

		Predicho	
		No renueva	Renueva
Real	No renueva	361	103
	Renueva	1789	4375

Fuente: Diseño propio

Tabla 8. Matriz de confusión del modelo óptimo (SVM) en el seguro de autos. Umbral óptimo c =0.97

		Predicho	
		No renueva	Renueva
Real	No renueva	431	226
	Renueva	2019	3952

Fuente: Diseño propio

Tabla 9: Matriz de confusión del modelo óptimo (regresión logística) en ambos ramos simultáneamente. Umbral óptimo c=0.87

		Predicho	
		No renueva	Renueva
Real	No renueva	577	380
	Renueva	1233	4438

Fuente: Diseño propio

Esta aplicación, es un ejemplo claro de cómo a nivel empresarial es posible sacar partido de la información disponible y a los *outputs* generados por el Departamento de *Analytics*. Identificar la propensión simultánea de un cliente a la renovación o no de sus pólizas es un indicador que aporta valor a la compañía. Por una parte el Departamento de cliente, sería capaz de tomar

acciones preventivas ante posibles anulaciones o por el contrario cuidar de aquellos clientes buenos que tienen más de una póliza contratada con la compañía. Luego, a nivel de tarificación sería posible la calibración de los modelos predictivos en función de una o más variables que sirvan como indicadores de la propensión a la renovación de los clientes para cada cartera. Además las acciones del Departamento de tarificación y el Departamento de cliente estarían alineadas en el sentido de que ambos tendrían una misma estrategia para el grupo de riesgo identificado. Por su parte, a nivel directivo, las decisiones estratégicas de negocio tendrían en cuenta un colectivo específico de clientes a los que se les debe tratar de forma diferenciada y por ende el negocio en sí mismo sería capaz de reorientarse de forma más asertiva.

4. Conclusiones y recomendaciones

La transformación requiere de un cambio de paradigma. El seguimiento de indicadores clásico, el informe (*reporting*) y la toma final de decisiones basada en la intuición no son suficientes. Cada vez es más popular la frase "los datos hablan por sí solos", y el éxito está en manos de las compañías que decidan evolucionar y "escucharlos".

Todos los empleados de la compañía deben entender los motivos y los beneficios de la creación de un Departamento de *Analytics*. La aceptación favorece la colaboración y con ello resulta más fácil establecer sinergias entre el resto de departamentos y el equipo de *Analytics*. El fin último es aportar valor al negocio, así que cuanto mayor sea la disposición a compartir los detalles operativos de cada área, mayor será el input que recibirá el Departamento de *Analytics* y mayores los beneficios globales.

Los cambios que supone la creación de un equipo de *Analytics* dentro de una compañía de seguros vs. la rapidez con la que se espera tener resultados, se encuentran desfasados temporalmente, es decir, resulta casi imposible tener un Departamento de *Analytics* si no se dispone de la tecnología, los recursos humanos y la experiencia necesaria. De ahí que sea necesario que los equipos de trabajo dispongan del perfil técnico específico y los líderes se encuentren familiarizados con el sector asegurador, además de tener conocimientos sólidos en *big-data analytics*. En este sentido, una alternativa inicial es la de contar con asesores expertos externos, para, en un primer momento, liderar los proyectos, capacitar a los equipos internos y realizar la transferencia del conocimiento necesario.

Por otra parte, y como comentamos al inicio, cada departamento suele disponer de ficheros de datos específicos para las actividades de análisis e informes usuales. En éste sentido, urge la democratización de la información. La capacidad de las compañías de centralizar el acceso a los datos en un único lugar, de modo que cualquier departamento tenga acceso transversal a los datos de otros departamentos, supone un verdadero reto, donde el primer beneficiado sería el Departamento de *Analytics*.

Cuando una compañía disponga de más de un equipo de *Analytics*, por ejemplo, porque cuenta con diferentes sedes en diversas partes del mundo, la relación debe ser absolutamente cercana. Aunque resulte evidente, la realidad es que no lo es. En éste sentido, han de compartir información, reciclar ideas e intercambiar casos de éxito y fracaso. Se trata de poder trabajar alineados, re-aprovechar la experiencia existente, facilitar la estandarización de procesos y la creación de modelos corporativos fácilmente reproducibles.

Por último, el caso de estudio presentado resume algunos de los beneficios que le aportaría al negocio implementar *Big Data Analytics* dentro de sus procesos. Por una parte, sería posible implementar métodos predictivos alternativos a los convencionales con el objetivo de complementar y mejorar la modelización realizada con los métodos clásicos. Así mismo, éste tipo de análisis serviría como punto de partida para el análisis de las dependencias entre diferentes grupos de riesgo, teniendo en cuenta información de más de un departamento de forma simultánea. Tal y como se ha evidenciado, la combinación de información del cliente a nivel global permite mejorar la capacidad predictiva de los modelos de retención. Por lo tanto, teniendo en cuenta que podrían incorporarse más líneas de negocio al análisis y con ello más información, es muy recomendable que los modelos utilizados dentro de las aseguradoras puedan escalar a otro nivel de uso más general que el de los informes específicos para cada ramo. Establecer en las entidades un equipo exclusivamente dedicado y capacitado para llevar a cabo análisis de tal envergadura, sin lugar a dudas, marcaría la diferencia entre el pasado y el futuro del sector asegurador.

Referencias

Bolancé, C., Guillen, M. y A. E. Padilla-Barreto (2016). Predicting detection in non-life motor and home insurance. *Lectures on Modeling and Simulation* 2, 107-120.

Boser, B. E., Guyon, I. M. y V. N. Vapnik (1992). A training algorithm for optimal margin classifiers. *Proceedings of the fifth annual workshop on Computational Learning Theory*, 144-152.

Dobbin, K. K. y R. M. Simon (2011). Optimally splitting cases for training and testing high dimensional classifiers. *BMC Medical Genomics* 4, 31, 1-8.

Guelman, L., Guillen, M. y A. M. Pérez-Marín (2014). A survey of personalized treatment models for pricing strategies in insurance. *Insurance: Mathematics and Economics*, 58, 68-76.

Guillen, M. (2016). Big data en seguros. *Índice: revista de estadística y sociedad* 67, 28-30.

Harrington, E. (2014). Building an analytics team for your organization part I. http://iianalytics.com/research/building-an-analytics-team-for-your-organization-part-i (7 de julio de 2017).

Hastie, T. (1998). Neural network. En *Encyclopedia of Biostatistics*, P. Armitage y T. Colton (eds.), John Wiley & Sons, Ltd, New York, 2986-2989.

Hornik, K., Meyer, D. y A. Karatzoglou (2006). Support vector machines in R. *Journal of Statistical Software* 15(9), 1-28.

Hothorn, T., Hornik, K. y A. Zeileis (2006). Unbiased recursive partitioning: A conditional inference framework. *Journal of Computational and Graphical Statistics* 15(3), 651-674.

Livingstone, D. J., Manallack, D. T. y I. V. Tetko (1997). Data modelling with neural networks: advantages and limitations. *Journal of computer-aided molecular design* 11(2), 135-142.

Myatt, G. J. (2007). *Making sense of data: a practical guide to exploratory data analysis and data mining*. John Wiley & Sons. New York.

McAfee, A. y E. Brynjolfsson (2012). Big data: the management revolution. *Harvard Business Review* 90(10), 60-68.

McCullagh, P. y J. A. Nelder (1989). *Generalized linear models*, 37. CRC press.

Meyer, D. y F. T. Wien (2001). Support vector machines. *R News* 1(3), 23-26.

Meyer, D., Dimitriadou, E., Hornik, L., Weingessel, A., Leisch, F. y C. C. Chang (2012). Package e1071: Misc functions of the department of statistics (e1071), TU Wien. *R package version*, 1-6.

Russom, P. (2011). Big data analytics. *TDWI Best practices report*, Fourth quarter 19, 40.

Suykens, J. A. y J. Vandewalle (1999). Least squares support vector machine classifiers. *Neural Processing Letters* 9(3), 293-300.

Swedloff, R. (2014). Risk classification's big data (r) evolution. *Connecticut Insurance Law Journal* 21, 339-374

Tu, J. V. (1996). Advantages and disadvantages of using artificial neural networks versus logistic regression for predicting medical outcomes. *Journal of clinical epidemiology* 49(11), 1225-1231.