

TRABAJO FIN DE MÁSTER

Título: Análisis de la retención y de la supervivencia en compañías aseguradoras: Un caso práctico.

Autor: Juan Guillermo Rusinque Campos

Tutor/res: Ramon Alemany

Curso: 2013 - 2014

Facultad de Economía y Empresa

Universidad de Barcelona

Trabajo Fin de Máster

Máster en Ciencias Actuariales y Financieras

**ANÁLISIS DE LA RETENCIÓN Y DE LA
SUPERVIVENCIA EN COMPAÑÍAS
ASEGURADORAS: UN CASO
PRÁCTICO**

Autor:

Juan Guillermo Rusinque Campos¹

Tutor/es:

Ramon Alemany

¹ E-mail: juanrus89@gmail.com

El contenido de este documento es de exclusiva responsabilidad del autor, quien declara que no ha incurrido en plagio y que la totalidad de referencias a otros autores han sido expresadas en el texto.

Resumen Ejecutivo

El sector asegurador a nivel europeo es un mercado muy competitivo y ésta característica se ve claramente en Autos que es uno de los ramos donde hay un mayor número de pólizas, una mayor cantidad de primas y una continua lucha por ganar más terreno dentro del sector atrayendo clientes de otras compañías por medio de diversas políticas de estrategia por la gran saturación que presenta el mercado.

De ahí viene la importancia de estudiar la retención porque es primordial para las empresas aseguradoras poder identificar tanto a sus mejores clientes como a los peores dado que sería muy perjudicial para una compañía tener el mayor porcentaje del sector si en su cartera sólo tiene a los clientes con peor rendimiento por una alta frecuencia o por un coste medio elevado. Así pues, el estudiar la retención es un primer paso en una estrategia de una compañía para identificar bien su cartera de clientes y cómo poder influir sobre ella para retener las pólizas con buen rendimiento y sanear la cartera de las pólizas que tengan una mala rentabilidad.

En el presente trabajo de investigación, el objetivo es hacer un análisis de la retención de pólizas mediante la aplicación de dos modelos y la comparación de los resultados obtenidos en cada uno de ellos. El primero consiste en analizar las variables que mayor influencia tienen en la decisión de renovar una póliza de Auto en un momento determinado mediante modelos lineales generalizados (LOGIT o PROBIT).

El segundo modelo se basa en un análisis de supervivencia sobre la duración que tiene una póliza dentro de una compañía y qué variables pueden hacer incrementar o disminuir esa duración. Para este último se utiliza el modelo de Riesgo Proporcional (Proportional Hazard model) propuesto por Cox.

Dada la tipología de los modelos se utilizaron dos bases de datos provenientes de una compañía aseguradora que opera en España. La primera base de datos consistía en una cartera de pólizas de un año determinado en el que la fecha de renovación o anulación de la póliza ya hubiese pasado. La segunda base de datos toma como base otra cartera de pólizas que renovaron en un período determinado y la evolución que ha tenido ésta durante aproximadamente cuatro años.

Para la aplicación de los modelos, algunas de las variables fueron modificadas con el fin de agrupar las categorías que resultaban ser no significativas con la categoría base. Para ello, se determinaba una categoría base o de referencia en aquellas variables que tuvieran categorías o en aquellas que, a criterio del autor, fuera necesario categorizar.

En cuanto a los resultados del trabajo, en el primer modelo se obtuvo que algunos factores como la cantidad de siniestros, tener franquicia contratada, algunas formas de pago o que la póliza sea suscrita en determinadas zonas geográficas del país; tenían una influencia negativa importante sobre la probabilidad de retener la póliza. Por otro lado, factores como el paquete de coberturas contratado o el sexo del tomador tenían una influencia positiva y significativa en la probabilidad de renovar la póliza.

En el segundo modelo se obtuvo que algunos de los factores que más influencia ejercían en la reducción del tiempo esperado o la probabilidad de supervivencia, tienen que ver con la forma de pago, la cantidad de siniestros o el hecho de tener franquicia contratada.

Mientras que los factores más importantes que hacían incrementar la probabilidad de supervivencia son el valor económico del vehículo o el paquete de coberturas contratado en la póliza.

De esta manera, variables como la forma de pago o el número de siniestros tienen el mismo efecto en ambos modelos (reduciendo tanto la probabilidad de renovar como la duración esperada de una póliza). En cambio, los clientes con paquetes de cobertura más completos o con vehículos con mayor valor económico son los que más probabilidad tienen de renovar o de durar más tiempo dentro de la compañía.

Tabla de contenido

Resumen Ejecutivo.....	5
1. Introducción	9
2. Metodología	11
3. Datos	14
3.1. Modelo de Retención	14
3.2. Modelo de supervivencia	15
4. Resultados	17
4.1. Modelo de Retención	17
4.2. Modelo de Supervivencia.....	20
5. Conclusiones	26
6. Bibliografía	29
Anexo 1	30
Anexo 2	30
Anexo 3	31
Anexo 4	32

1. Introducción

El mercado asegurador es un mercado maduro y muy competitivo en el que las entidades aseguradoras están en una continua lucha por retener a sus propios clientes y por ganar terreno dentro del sector atrayendo a los clientes de las otras entidades ofreciendo mejores precios y servicios.

El hecho de poder retener una póliza o un cliente genera muchos más beneficios que el hecho de tener un nuevo cliente. Esto es consecuencia de que se llevarán a cabo diferentes acciones sobre el nuevo cliente con el fin de captarlo (por ejemplo mayores descuentos o más coberturas) lo que afectará a la prima de la póliza haciéndola más baja que la prima del antiguo cliente. Esto implica que se obtendrán menos ingresos si se pierde un cliente y se obtiene uno nuevo, respecto a los ingresos derivados de haber podido renovar la póliza del primer cliente.

A pesar de la importancia que tiene la retención para las empresas, la literatura que existe es muy reciente teniendo en cuenta que muchas compañías aseguradoras comenzaron su actividad a finales del siglo XIX. Aun así, diferentes autores han estudiado sobre los beneficios que se obtendrían con la retención de clientes. Una leve disminución en la tasa de abandono impulsaría los beneficios de un determinado banco entre un 25% y un 85% (Reichheld y Sasser, 1990). Larivière y Van den Poel (2004) demostraron que dentro de la cartera de clientes de una entidad financiera belga, la mayor proporción de clientes tenía el producto financiero más básico (cuenta de ahorros) y que éstos eran los que tenían la tasa de retención más baja. Pero el impacto que tenía la compra de cualquier otro producto financiero dentro de la misma compañía era de incrementar esa tasa de retención. Esto muestra como la implementación de diferentes estrategias de mercado, como la venta cruzada, puede ayudar a mejorar la retención de clientes dentro de una empresa.

Por otro lado, el análisis de la retención también presenta una serie de ventajas para la compañía. Por medio del análisis se pueden detectar segmentos en la cartera en función de las características de los clientes y desarrollar acciones de marketing con el objetivo de mejorar la satisfacción con la empresa y sus productos de aquellos segmentos donde interese tener unas tasas de retención más altas que en otros.

En cuanto al análisis de supervivencia una de sus posibles utilidades es saber el tiempo que dura una póliza dentro de una cartera y ver en función de una serie de variables como puede incrementar o disminuirse ese tiempo. Brockett et al. (2008) realizaron un análisis de supervivencia a una cartera de clientes con pólizas en varios ramos y mostraron que la influencia de compañías externas junto con el número de siniestros y el cambio de dirección del tomador eran los tres factores más relevantes (la influencia de compañías externas la que más) cuando un cliente quería cancelar todas sus pólizas. Esto demuestra la alta competitividad que hay dentro del sector asegurador y que el análisis de supervivencia es una herramienta que ayuda a encontrar posibles causas para que un cliente cancele su póliza de seguros.

Larivière y Van den Poel (2005) también demostraron otra utilidad del estudio de la supervivencia a partir del momento en que un cliente emite una queja a la compañía y el momento en que la abandona o compra un nuevo producto. Se observó que los clientes que se quejan tienen una mayor probabilidad de quedarse en la compañía que los que no

lo hacen. Incluso hay una mayor probabilidad de que estos clientes adquieran otro producto. También hallaron que los factores que más influyen en el comportamiento del cliente al tener una incidencia con la compañía son tener una solución rápida al problema y una compensación que en este caso será financiera dado la tipología de clientes que se habían estudiado.

Pero los efectos que tienen las variables sobre las que se hacen los estudios sobre retención o supervivencia no son constantes en el tiempo. Por esta razón es que se han venido desarrollando modelos que recojan el impacto de las variables para cada momento determinado del tiempo. Guillen et al. (2011) demostraron que las predicciones de perder un cliente pueden variar a lo largo del tiempo y que deberían ser reajustadas para una estimación más precisa de la supervivencia de un cliente dentro de la compañía.

Por las anteriores razones y los estudios realizados por diversos autores es que el objetivo de este trabajo es comparar dos formas de analizar la retención. Por un lado, el análisis se realizará para un momento determinado del tiempo mediante un modelo de regresión logística (LOGIT) o un modelo de regresión PROBIT, mientras que por el otro lado se pretende realizar un análisis de supervivencia a lo largo de un tiempo determinado mediante el modelo de riesgo proporcional (o Proportional Hazard model). De esta manera, se pretenderá llegar a conclusiones sobre la influencia de distintas variables tanto en un modelo como en el otro, hallar las posibles similitudes o diferencias y contrastar con lo que otros autores hayan demostrado sobre la permanencia o duración de un cliente en una compañía aseguradora.

Para cumplir con este objetivo el trabajo se estructurará de la siguiente manera: en primera instancia, se hará una breve introducción a los modelos de regresión (tanto LOGIT como PROBIT) y al modelo de Proportional Hazard con el fin de facilitar al lector la interpretación de los resultados que se van a obtener. En una segunda parte, se hará un análisis descriptivo de las variables de las dos bases de datos proporcionadas por la compañía aseguradora con las que se trabajarán y que ayudarán a establecer una categoría de referencia en cada uno de los modelos. Estas categorías de referencia irán agrupando a aquellas categorías que resulten ser no significativas mientras se trabaja con los modelos. En la tercera parte, se analizarán los resultados obtenidos en los dos modelos para que en la cuarta y última parte se recojan las conclusiones derivadas de la realización del trabajo.

2. Metodología

Todo el desarrollo del trabajo se basará en los modelos lineales generalizados (*Generalized linear models*, GLM) y en el modelo de supervivencia de Cox. De los modelos GLM se utilizará el modelo de regresión logística o LOGIT y el modelo de regresión PROBIT porque son los que generalmente se aplican cuando la variable dependiente es dicotómica y discreta (Renovar sí o no). El modelo de supervivencia de Cox se usa normalmente para estimar el tiempo entre que un individuo nace (el tomador contrata una póliza) y el momento en que muere (en este trabajo será que el tomador cancele la póliza o que llegue hasta la fecha final del estudio).

Para el modelo LOGIT se toma una variable aleatoria Y cuyos valores sean 1 o 0. Tomará el valor de 1 cuando el suceso ocurre y el valor 0 si se da lo contrario. En este estudio la renovación será el suceso y por lo tanto si el tomador renueva será un 1 y si no lo hace será un 0. Para explicar a esta variable aleatoria Y , se tomará un conjunto de variables explicativas X para las cuáles, a través del modelo, se estimará un vector de parámetros β . El conjunto de variables explicativas del modelo pueden ser tanto continuas como discretas.²

La probabilidad de que ocurra el suceso (que el tomador renueve) viene determinada entonces por la siguiente ecuación:

$$\pi = P(Y = 1) = \frac{e^{x_i'\beta}}{1 + e^{x_i'\beta}}$$

Y la probabilidad de que no renueve será entonces:

$$1 - \pi = P(Y = 0) = \frac{1}{1 + e^{x_i'\beta}}$$

El modelo PROBIT tiene un planteamiento similar pero si en el LOGIT la función de enlace canónica es la logística, en el PROBIT es la función de distribución de la normal estándar. Por lo tanto, la probabilidad de que renueve y que no renueve viene determinada por:

$$\pi = P(Y = 1) = \Phi^{-1}(x_i'\beta)$$

$$1 - \pi = P(Y = 0) = 1 - \Phi^{-1}(x_i'\beta)$$

La estimación de los parámetros del modelo LOGIT se realiza mediante el criterio de la máxima verosimilitud que se desarrolla a partir de expresar en logaritmos la probabilidad conjunta de observaciones Y_i independientes e igualmente distribuidas como una Bernoulli:

$$L(\beta) = \prod_{i=1}^n \left[\frac{e^{x_i'\beta}}{1 + e^{x_i'\beta}} \right]^{y_i} \left[1 - \frac{e^{x_i'\beta}}{1 + e^{x_i'\beta}} \right]^{1-y_i}$$

² En el **Anexo 1** se encontrará más información sobre el modelo LOGIT

$$\ln\{L(\beta)\} = \sum_{i=1}^n \left[y_i \ln \left\{ \frac{e^{x_i' \beta}}{1 + e^{x_i' \beta}} \right\} + (1 - y_i) \ln \left\{ 1 - \frac{e^{x_i' \beta}}{1 + e^{x_i' \beta}} \right\} \right]$$

Este modelo ha sido utilizado por varios autores para el estudio de la retención como Brockett et al. (2008) que basándose en covariables conocidas quieren determinar la probabilidad de que un hogar que inicialmente tiene más de una póliza, las cancele todas simultáneamente. La misma metodología fue aplicada para predecir la probabilidad de que una póliza de Contenido, Hogar o Auto sea cancelada dentro de un período de tres meses condicionada a la información recogida en el año previo (Guillen et al., 2003).

Otros autores tratan de comparar al modelo de regresión logística con el método de árboles de decisión y el método de las redes neuronales artificiales, todo esto dentro del ámbito del *Data Mining*. Au et al. (2003) quisieron demostrar cómo usar las técnicas del *Data Mining* y el software para ajustar y comparar diferentes modelos de predicción de reducción de clientes (Customer attrition) usando datos de una de las mayores empresas proveedoras de servicios de telecomunicaciones en Estados Unidos.

Un análisis similar hicieron Smith et al. (2000) para demostrar el potencial del *Data Mining* para lograr crecimiento en el mercado y beneficios. Ellos intentan analizar tanto la retención como el nivel de prima basándose en el conocimiento de los clientes que probablemente renovarán, su nivel de riesgo y su sensibilidad al incremento de precios; con el fin de establecer las primas que permitan retener y adquirir nuevos clientes.

En cuanto al análisis de supervivencia, este trabajo se basará en el modelo de regresión de Cox o modelo de Riesgo proporcional (Proportional Hazard) desarrollado desde 1972. Este modelo permite la incorporación de variables dentro del estudio de la supervivencia. De la misma manera que se realizó con los modelos LOGIT y PROBIT, se tomarán una serie de variables X para las que se estimará un vector de parámetros β . Estas variables son características particulares de cada individuo que harán incrementar o disminuir la probabilidad de que ocurra el suceso. La forma general de la fórmula del modelo de Riesgo proporcional es:

$$\lambda(t|X) = \lambda_0(t) * c(X'\beta)$$

Donde $\lambda_0(t)$ es la función “baseline” o función de riesgo base, y $c(X'\beta)$ es una función que se aplica sobre las variables explicativas y cuyos valores finales serán no negativos. Es una forma de estimar la distribución del tiempo que hay entre un momento determinado y el momento en que ocurre un suceso. En este caso, el tiempo entre la fecha inicial del contrato de la póliza (sea una renovación o una nueva póliza) y el momento en que se anula. En el modelo de Cox, la función antes mencionada está determinada en la forma:

$$c(X'\beta) = e^{X'\beta}$$

Por lo tanto la fórmula general quedaría de la siguiente manera

$$\lambda(t|X) = \lambda_0(t) * \exp(X'\beta)$$

Así pues, ésta fórmula cumpliría con la propiedad de que las funciones de riesgo de dos individuos diferentes sean proporcionales y que ésta relación sea constante e independiente del tiempo cuando las variables seleccionadas son fijadas en el momento inicial.

$$\frac{\lambda(t|X_1)}{\lambda(t|X_2)} = \frac{\lambda_0(t) * c(X_1'\beta)}{\lambda_0(t) * c(X_2'\beta)} = \frac{c(X_1'\beta)}{c(X_2'\beta)}$$

A esta relación se le llama hazard ratio o ratio de riesgo y es ésta la razón por la que el modelo recibe el nombre de Proportional Hazard. El modelo de Cox normalmente se ajusta con métodos de máxima verosimilitud para poder estimar el vector de parámetros β . Para estimar la probabilidad de que un individuo fallezca (no renueve) en el momento t se deben conocer las funciones de riesgo de todos los individuos que aún permanecen vivos (pólizas en vigor) en el momento inicial o momento cero:

$$\lambda(t|X_i) = \lambda_0(t) * c(X_i'\beta)$$

La probabilidad de que el individuo i no renueve, dentro de todo el conjunto de individuos que tienen pólizas en vigor, es la proporción que tiene la función de riesgo del individuo i dentro de la suma total de funciones de riesgo de todos los individuos en riesgo.

$$\Pr(X = x_i|t) = \frac{\lambda(t|X_i)}{\sum_{i=1}^n \lambda(t|X_i)}$$

Por ejemplo, si hay cinco individuos vivos (pólizas en vigor), la probabilidad de que el cuarto individuo fallezca (no renueve) sería:

$$\Pr(X = x_4|t) = \frac{\lambda(t|X_4)}{\lambda(t|X_1) + \lambda(t|X_2) + \lambda(t|X_3) + \lambda(t|X_4) + \lambda(t|X_5)}$$

Y dado que todas las funciones de riesgo tienen la misma función de riesgo base $\lambda_0(t)$, la probabilidad se puede reescribir de la siguiente manera:

$$\Pr(X = x_4) = \frac{\exp(X_4'\beta)}{\exp(X_1'\beta) + \exp(X_2'\beta) + \exp(X_3'\beta) + \exp(X_4'\beta) + \exp(X_5'\beta)}$$

En el **Anexo 2** se encontrará más información sobre el modelo de riesgo proporcional.

3. Datos

Para el desarrollo de este trabajo se disponen de datos provenientes de una compañía aseguradora que opera en España. Los datos consisten en dos bases de datos, una para el modelo de retención aplicando el LOGIT y el PROBIT mientras que en la otra se aplicará el modelo Proportional Hazard.

3.1. Modelo de Retención

En la primera base de datos se dispone de una muestra de pólizas del ramo Auto entre las que se encuentran pólizas que ya estaban en cartera y su renovación se hacía a lo largo del año 20XX junto con aquellas nuevas pólizas que entraron en la compañía durante ese mismo año. Las variables presentes en la base de datos y que se utilizarán en el modelo son las que se listan en la siguiente tabla:

Nombre de la Variable	Descripción
VIGOR	Si la póliza está en vigor o no
Forma de pago	Forma de pago: A, B o C
Sexo	Género: Varón o Mujer
Age	Edad del conductor
Antig	Antigüedad de póliza
Valor	Valor del vehículo
Franq	Franquicia de la póliza
Region	Comunidad autónoma a la que pertenece
EdadVeh	Antigüedad del vehículo
Mes	Mes de suscripción de la póliza
Canal	Canal de distribución
New_Ren	Cartera o nueva producción
Cobertura	Paquete de coberturas contratado
Siniestros	Número de siniestros ocurridos
Varabsprima	Incremento de prima
IncrePor	Incremento porcentual de prima

Tabla 1: Variables del modelo de Retención. Fuente: Elaboración propia.

Las variables *Valor*, *Franq*, *Region*, *Cobertura*, *varabsprima* e *IncrePor* son variables creadas o modificadas por el autor a partir de las variables originales de la base de datos. En la variable *Valor* se han creado cuatro categorías en función del valor del vehículo (*0-10k*, *10k-20k*, *20k-50k*, *>50k*). La variable *Franq* tiene tres categorías en función de si la póliza no tiene franquicia, si tiene franquicia y es menor a 300€ o si tiene franquicia y ésta supera los 300€.

En cuanto a la variable *Region* ésta recoge la distribución de las pólizas por comunidades autónomas. Se ha creado dentro de la variable, una categoría de referencia que incluye a Castilla y León, Cantabria, Galicia, Melilla, País Vasco, Navarra, Cataluña (categoría que tenía el mayor porcentaje de pólizas) junto con las pólizas de Andorra. Para la variable *Cobertura* se crearon tres categorías (*Básico*, *Terceros* y *Todoriesgo*) a partir de los principales paquetes de coberturas que comercializa la compañía aseguradora. Por último, las variables *varabsprima* e *IncrePor* recogen los incrementos de prima absolutos y porcentuales respectivamente. La variable *IncrePor* ha sido categorizada por tramos de incremento. Se resalta que las categorías creadas en

Valor y *Region* vienen determinadas porque los parámetros de las categorías previas resultaban ser no significativos respecto a la categoría de referencia.

3.2. Modelo de supervivencia

Para este modelo se dispone de una base de datos con pólizas del ramo Auto emitidas en España por la compañía aseguradora, que renovaron el contrato durante el año 20XX. Para calcular el número de días que está vigente la póliza, se tomará como fecha final el momento en que se emite el aviso de anulación o el 31 de diciembre de 20XX+3 que es la fecha final del estudio. De esta manera se obtiene información de un conjunto de pólizas durante 4 años aprox. Para este modelo se trabajará con las siguientes variables:

Nombre de la Variable	Descripción
Forma de pago	Fraccionamiento: A, B, C o D
Sexo	Género: Varón o Mujer
Age	Edad del conductor
Antig	Antigüedad de póliza
EdadVeh	Antigüedad del vehículo
Franq	Franquicia de la póliza
Valor	Valor del vehículo
Region	Comunidad Autónoma donde se suscribe la póliza
PRTOTAL	Prima pagada en 2010
Cobertura	Paquete de coberturas contratado
Canal	Canal de distribución
Siniestros	Número de siniestros ocurridos en 2010
Motivo	Motivo de anulación de la póliza
Lenfol	Número de días vigentes de la póliza
NORENOV	No renovación: 1 es no renovada y 0 es renovada

Tabla 2: Variables del modelo Proporcional Hazard. Fuente: Elaboración propia.

Siguiendo la misma metodología que en el modelo de retención, las variables *FPAGO*, *Valor*, *Franq*, *Region*, *Cobertura*, *Lenfol* y *NORENOV* son variables creadas o modificadas por el autor a partir de las variables originales de la base de datos. La categoría referente a la forma de pago D, ha sido agrupada con la forma de pago C por tener un parámetro no significativo y por recomendación del equipo actuarial de la compañía en base al pequeño número de registros que había. En la variable *Valor* se han creado cinco categorías en función del valor del vehículo (*0-10k*, *10k-20k*, *20k-50k*, *50k-100k*, *>100k*) a diferencia de las cuatro categorías que habían en el modelo anterior porque la categoría *>100k* resultó ser significativa. La variable *Franq* tiene las mismas tres categorías en función de si la póliza no tiene franquicia, o tiene franquicia y es menor a 300€ o tiene franquicia y ésta supera los 300€.

En cuanto a la variable *Region* que recoge la distribución de las pólizas por comunidades autónomas, se ha creado una nueva categoría de referencia que incluye a Asturias, Islas Canarias, Castilla-La Mancha, Madrid y Cataluña (categoría que tenía el mayor porcentaje de pólizas). Esta categoría de referencia fue creada a partir de la significación de los parámetros para cada comunidad autónoma. Aquellos que resultaban ser no significativos fueron agrupados con Cataluña. Para la variable *Cobertura* se crearon cuatro categorías (*Básico*, *Terceros*, *Todoriesgo* y *Otros*) a partir de los principales paquetes de coberturas que comercializa la compañía aseguradora. La categoría *Otros* recoge un número reducido de pólizas cuyo paquete de coberturas no es

estándar porque la póliza cubre diversas garantías de distintos paquetes como elaborar un “traje a medida”.

La variable *Lenfol* contabiliza el número de días que hay entre la fecha de renovación de la póliza en 20XX y el momento en que se conoce que la póliza será anulada o la fecha final del estudio. *NORENOV* es una variable dicotómica necesaria para el modelo que tomará valor 1 cuando no se renueva la póliza y toma valor cero si el cliente permanece en la empresa hasta el final del estudio.

4. Resultados

Para la aplicación de los modelos junto con las variables anteriormente descritas, se ha utilizado el software SAS 9.3 por su fácil aplicación e interpretación de los resultados.

4.1. Modelo de Retención

Del total de pólizas de la base de datos, aproximadamente el 80% de los contratos de la cartera permanecen en vigor al finalizar el año 20XX. En las siguientes tablas se muestran algunos descriptivos de algunas de las variables con las que se trabajará. En el **Anexo 3** se podrá ver la distribución de las pólizas en las categorías del resto de las variables.

Variable	Media	Desv Estándar	Mínimo	Máximo
VIGOR	0,8159764	0,3875038	0	1
Edad	51,020481	13,4913543	15	100
Antig. póliza	6,0323797	5,7921711	0	50
Franq	42,933997	135,0837331	0	6000
Antig. Veh.	8,0970841	5,5197925	0	85
Var. Prima	-9,211909	69,8687554	-3063,86	3074,86

Sexo	% Pólizas	Tipo de póliza	% Pólizas
Varón	71.43%	Cartera	82.79%
Mujer	28.57%	Nueva Producción	17.21%

Valor	% Pólizas	Siniestros	% Pólizas
0k-10k	13.62%	0	73.35%
10k-20k	46.41%	1	18.86%
20k-50k	37.28%	2	5.60%
>50k	2.69%	3 o más	2.19%

Tablas 3 a 7: Estadísticos descriptivos de algunas variables del modelo de Retención. Fuente: Elaboración propia.

Se observa que la cartera tiene un 81.59% de retención y la media de edad está en los 51 años. Esto indica que la mayoría de las pólizas son contratadas por personas mayores. Las pólizas tienen una media de 6 años de antigüedad en la compañía, mientras que la antigüedad de los vehículos tiene una media superior dos años. El valor de la mayoría de los vehículos asegurados ronda entre 10.000€ y 20.000€ con un tope máximo de 489.728€. La gran mayoría de las pólizas no tiene franquicia y en aquellas que si la tienen hay un gran número que tiene pactada una franquicia de 300€. La variación máxima de la prima supera ligeramente los 3000€ tanto en aumento como en disminución.

Para este modelo se estableció una categoría de referencia en función de aquellas categorías que tuviesen un mayor número de pólizas dentro de su respectiva variable. Así pues, la categoría de referencia es una póliza con forma de pago A, cuyo tomador es un Varón, el valor del vehículo asegurado oscila entre los 10.000€ y 20.000€, sin franquicia, la residencia del tomador está en la zona de referencia descrita en el punto

3.1., suscrita en el mes 7 a través del Canal 1, su coberturas están dentro del grupo Terceros, que no ha tenido siniestros y cuya renovación implicó un aumento de prima entre el 0% y el 2.5%. En la **Tabla 8** se pueden observar los parámetros estimados a través del modelo LOGIT³:

Parámetro	Estimación	St. Error	Pr > ChiSq	Odds Ratio
Intercept	0.9355	0.0251	<0.0001	-
fpago, <i>B</i>	-0.2921	0.00988	<0.0001	0.747
fpago, <i>C</i>	-0.3778	0.0179	<0.0001	0.685
Sexo, <i>M</i>	0.1736	0.00902	<0.0001	1.190
Age	0.0176	0.000333	<0.0001	1.018
Antig	0.0207	0.00105	<0.0001	1.021
Valor, <i>0-10k</i>	-0.0633	0.0148	<0.0001	0.939
Valor, <i>20k-50k</i>	0.1046	0.00905	<0.0001	1.110
Valor, <i>>50k</i>	0.1502	0.0266	<0.0001	1.162
Franq, Menor a 300	-0.2360	0.0218	<0.0001	0.790
Franq, Mayor a 300	-0.4813	0.0250	<0.0001	0.618
Region, <i>Andalucía</i>	-0.2385	0.0119	<0.0001	0.788
Region, <i>Aragón</i>	0.0448	0.0200	0.0252	1.046
Region, <i>Asturias</i>	-0.1747	0.0451	0.0001	0.840
Region, <i>Castilla La Mancha</i>	-0.1610	0.0233	<0.0001	0.851
Region, <i>Ceuta</i>	-0.7822	0.0865	<0.0001	0.457
Region, <i>Comun. Valenciana</i>	-0.0475	0.0127	0.0002	0.954
Region, <i>Extremadura</i>	-0.0709	0.0275	0.0099	0.932
Region, <i>Islas Baleares</i>	0.1825	0.0161	<0.0001	1.200
Region, <i>Islas Canarias</i>	0.1750	0.0155	<0.0001	1.191
Region, <i>La Rioja</i>	0.3194	0.0451	<0.0001	1.376
Region, <i>Madrid</i>	-0.1617	0.0235	<0.0001	0.851
Region, <i>Murcia</i>	-0.3202	0.0364	<0.0001	0.726
EdadVeh	-0.0315	0.000852	<0.0001	0.969
Mes, <i>1</i>	-0.2660	0.0156	<0.0001	0.766
Mes, <i>2</i>	-0.2831	0.0156	<0.0001	0.753
Mes, <i>3</i>	-0.2389	0.0153	<0.0001	0.787
Mes, <i>4</i>	-0.2071	0.0156	<0.0001	0.813
Mes, <i>5</i>	-0.1372	0.0155	<0.0001	0.872
Mes, <i>6</i>	-0.0677	0.0155	<0.0001	0.935
Mes, <i>8</i>	0.0104	0.0167	0.5338	1.010
Canal, <i>Canal 2</i>	-0.2241	0.0100	<0.0001	0.799
Canal, <i>Canal 3</i>	-0.1321	0.0103	<0.0001	0.876
New_Ren, <i>New</i>	0.0146	0.0109	0.1817	1.015
Cobertura, <i>Básico</i>	0.0671	0.0199	0.0007	1.069
Cobertura, <i>TodoRiesgo</i>	0.4386	0.0182	<0.0001	1.551
Siniestros, <i>1</i>	-0.1690	0.00987	<0.0001	0.844
Siniestros, <i>2</i>	-0.3645	0.0155	<0.0001	0.695
Siniestros, <i>3</i>	-0.4822	0.0230	<0.0001	0.617
varabsprima	0.00103	0.000117	<0.0001	1.001
IncrePor, <i>>-10%</i>	2.3725	0.0335	<0.0001	10.724
IncrePor, <i>-10% to -5%</i>	1.5482	0.0106	<0.0001	4.703

³ Se ha escogido el modelo LOGIT debido a que los criterios de información AIC y SC eran menores a los del modelo PROBIT.

IncrePor, -5% to -2.5%	0.1350	0.0147	<0.0001	1.145
IncrePor, -2.5% to 0%	-0.1410	0.0106	<0.0001	0.868
IncrePor, 0%	-0.2666	0.0196	<0.0001	0.766
IncrePor, 2.5% to 5%	-0.2015	0.0131	<0.0001	0.818
IncrePor, 5% to 10%	-0.0998	-0.0998	<0.0001	0.905
IncrePor, >10%	0.1710	0.0299	<0.0001	1.187

Tabla 8: Parámetros estimados utilizando el modelo LOGIT. Fuente: Elaboración propia.

Se observó que todos los parámetros resultan ser significativos (a un nivel de 5%) para explicar la retención a excepción de la categoría de las pólizas que son suscritas en el mes 8 y la variable que distingue entre Nueva Producción y Cartera. Aquellos parámetros con valor positivo indican un aumento de la probabilidad de renovar mientras que los negativos indican un descenso.

Así pues, si se analizan los parámetros y la estimación de los Odds-ratios se observa que hacer el pago de una forma fraccionada B o C disminuye la probabilidad de renovar hasta casi un 25%. Mientras que si el tomador es una mujer la probabilidad incrementa un 19% respecto a la de los hombres. Esto puede explicarse con las diferencias que existen entre hombres y mujeres a la hora de contratar una póliza de seguros o productos financieros, tal como Larivière y Van den Poel (2004) demostraron que los hombres tienen una mayor probabilidad de cambiar de compañía. La edad del tomador o la antigüedad en la compañía incrementan la probabilidad un 2% aproximadamente.

En cuanto al valor del vehículo, la estimación indica que a medida que va aumentando el valor del vehículo, es más probable que renueve. La diferencia entre asegurar un coche cuyo valor está entre 10.000€ y 20.000€ y uno que vale más de 50.000€ es de un 16% posiblemente debido a que son coches más caros y a que los tomadores buscan compañías que ofrezcan productos más específicos para ellos. El hecho de tener franquicia también disminuye la probabilidad de renovar causado quizás por el riesgo que está asumiendo el tomador o la influencia de una compañía externa que le haga una oferta sin franquicia.

En cuanto al análisis por comunidades autónomas se observó que la diferencia no es significativa entre las comunidades de la zona de referencia que son Castilla y León, Cantabria, Galicia, Melilla, País Vasco, Navarra, Cataluña (categoría de referencia inicial) y las pólizas de Andorra. Mientras que en Aragón, Islas Baleares, Islas Canarias y La Rioja es más probable que se renueve una póliza en comparación con las de la zona de referencia. Lo contrario pasa en las comunidades de Andalucía, Asturias, Castilla La Mancha, Ceuta, la Comunidad Valenciana, Extremadura, Madrid y Murcia. El Odd-ratio del parámetro de la antigüedad del vehículo indica que por cada año de más que tenga el coche, la probabilidad de renovar la póliza disminuye un 3.1%. Es un claro resultado que está vinculado a la vida útil de los coches porque llegará un momento en el que el vehículo sea sacado de circulación o abandonado. El modelo indica que todas las pólizas suscritas en los meses anteriores al mes 7 tienen una disminución del 6.5% o más en la probabilidad de renovar mientras que el mes 8 tiene un parámetro no significativo respecto al mes 7.⁴

⁴ Dado que las pólizas que son suscritas en los inicios del año tienen un mayor tiempo para decidir si renuevan o no, se ha estimado un modelo que recogiera este efecto. Los resultados indicaban una estimación de los parámetros similar a la que ya se tenía aunque los criterios de información (AIC y SC) eran mayores y se descartó.

Cuando las pólizas son suscritas por medio del canal 1 tienen una mayor probabilidad de renovar en comparación a los otros canales donde la probabilidad disminuye casi un 20% en Canal 2 mientras que en Canal 3 la disminución no supera el 13%. Hay un incremento significativo de la probabilidad de casi un 56% cuando se contrata alguno de los productos del grupo Todo Riesgo y cuando es alguno de los del grupo Básico el incremento es de sólo un 7%. En cuanto al número de siniestros es evidente la disminución progresiva de la probabilidad de renovar a medida que va aumentando la cantidad de siniestros. Su disminución es un claro reflejo de que a cuantos más siniestros se declaren, es más previsible un aumento de la prima y un más que posible cambio de compañía.

Por último, la variable referida al incremento porcentual de prima que tiene nueve posibles categorías muestra resultados interesantes. Teniendo en cuenta que la categoría de referencia son las pólizas que han tenido un incremento entre el 0% y el 2.5%, el modelo indica que las pólizas que no han tenido incremento o que incluso han tenido un decremento de hasta un 2.5% es menos probable que renueven. Esto puede explicarse por una posible influencia de la competencia que les ofrece precios más bajos dado que son buenos riesgos. Otro resultado interesante tiene que ver con las pólizas que tienen un incremento mayor al 10% para las cuales se esperaría que la probabilidad de renovar fuera menor pero el modelo indica que hay un aumento del 18%. Este resultado puede deberse a un incremento en las garantías ofrecidas o un cambio en el mismo riesgo asegurado.

De esta manera entonces se puede observar que las variables que mayor influencia tienen en este modelo son: la forma de pago, el sexo, tener una franquicia o no, la comunidad autónoma donde se suscribe la póliza, el mes de suscripción, la cantidad de siniestros, el incremento porcentual de prima y el tipo de producto contratado. Dentro de estas variables, el parámetro que indica que hay un mayor decremento de la probabilidad de renovar es el que corresponde a suscribir una póliza en Ceuta (54.3%). Cuando la póliza ha tenido tres o más siniestros, o tienen una franquicia superior a 300€ la probabilidad de renovar cae aproximadamente un 38%, mientras que tener dos siniestros o hacer el pago de forma C reducen ésta probabilidad un 31%. Tener un incremento del 0% en la prima, suscribir la póliza en el mes 1 o el mes 2, o haberla suscrito en Murcia o haber hecho el pago de la forma B disminuirá en un 25% aproximadamente la probabilidad de renovar.

Por el lado contrario, cuando la póliza ha tenido una disminución en la prima superior al 5% la probabilidad de renovar aumenta más del 400% e incluso más del 1000% cuando la disminución supera el 10% por razones evidentes. Suscribir la póliza en La Rioja o tener un producto de los que están en el grupo de Todo Riesgo hacen incrementar la probabilidad de renovar un 37.6% y un 55.1% respectivamente respecto a la categoría de referencia. Finalmente, cuando el tomador es una mujer, o las pólizas suscritas tanto en las Islas Canarias como en las Islas Baleares o un incremento en prima superior al 10% es un 20% más probable que renueven que la categoría de referencia.

4.2. Modelo de Supervivencia

En esta base de datos, del total de pólizas que estaban en vigor en 20XX, queda en vigor aproximadamente un 44.11% del total de pólizas iniciales al finalizar el año 20XX+3. A continuación se podrán observar estadísticos descriptivos y tablas que recogen la

distribución de pólizas de algunas variables. En el **Anexo 4** se verá esta información más ampliada incluyendo el resto de variables.

Variable	Media	Desv Estándar	Mínimo	Máximo
Edad	49,911573	13,601955	14	97
Antig. Póliza	5,5984933	5,5147193	1	49
Valor	17873,92	13398,15	0	3.373.000
Prima	426,47	263,2063	0	10861,79
Siniestros	0,1936505	0,5355858	0	12
Nº Días	915,8951	401,3331	1	1461

FPAGO	% Pólizas	Canal	% Pólizas
A	72.94%	Canal 1	38.14%
B	20.95%	Canal 2	30.04%
C*	6.11%	Canal 3	31.82%

*C y D

Tabla 9, 10 y 11: Estadísticos descriptivos y distribución de pólizas para las variables FPAGO y Canal.
Fuente: Elaboración propia.

A partir de estos estadísticos y en comparación a las variables del modelo anterior, se observa que la media de edad baja en dos años, la media de antigüedad de póliza está un poco por debajo de los 6 años y la media de antigüedad del vehículo también baja un año. La prima media de la cartera se sitúa por encima de los 400€, mientras que el valor del vehículo se sitúa en los 17.000€ de media y cuyo valor está en la categoría de referencia de esta variable. En cuanto al tiempo que permanece una póliza, se obtuvo que el tiempo promedio en la compañía es de dos años y medio. Teniendo en cuenta que el tiempo de estudio son cuatro años, es un buen resultado porque representa un 62% del tiempo total.

En la **Tabla 10** y **Tabla 11** hay evidencia de que las categorías que tenían el mayor número de pólizas en el modelo anterior, aún lo siguen teniendo. Por lo tanto, la única diferencia en cuanto a las categorías de referencia que se han establecido en los modelos, tiene que ver con la comunidad autónoma donde es suscrita la póliza. En el punto **3.1** y el punto **3.2** puede verse cuáles son las comunidades que están dentro de la zona de referencia y cuáles no, según el modelo que se esté aplicando.

Parámetro	Estimación	St. Error	Pr > ChiSq	Hazard Ratio
FPAGO, <i>B</i>	0.31672	0.00347	<0.0001	1.373
FPAGO, <i>C</i>	0.30178	0.00583	<0.0001	1.352
SEXO, <i>M</i>	-0.07426	0.00337	<0.0001	0.928
Age	-0.00881	0.0001222	<0.0001	0.991
Antig	-0.02214	0.0003733	<0.0001	0.978
EdadVeh	0.0008725	0.0000987	<0.0001	1.001
Franq, <i>Menor a 300</i>	0.16896	0.00720	<0.0001	1.184
Franq, <i>Mayor a 300</i>	0.34830	0.00793	<0.0001	1.417
Valor, <i>0-10k</i>	0.08999	0.00542	<0.0001	1.094
Valor, <i>20k-50k</i>	-0.10821	0.00336	<0.0001	0.897
Valor, <i>50k-100k</i>	-0.37126	0.01090	<0.0001	0.690

Valor, >100k	-1.19625	0.03720	<0.0001	0.302
Region, <i>Andalucía</i>	0.06705	0.00438	<0.0001	1.069
Region, <i>Andorra</i>	0.20257	0.04959	<0.0001	1.225
Region, <i>Aragón</i>	-0.09539	0.00769	<0.0001	0.909
Region, <i>Baleares</i>	-0.03576	0.00619	<0.0001	0.965
Region, <i>Cantabria</i>	-0.03852	0.01894	0.0420	0.962
Region, <i>Castilla y León</i>	-0.05096	0.00799	<0.0001	0.950
Region, <i>Ceuta</i>	0.12619	0.03775	0.0008	1.134
Region, <i>Comun. Valenciana</i>	0.00898	0.00455	0.0487	1.009
Region, <i>Extremadura</i>	0.03666	0.01123	0.0011	1.037
Region, <i>Galicia</i>	-0.08290	0.00720	<0.0001	0.920
Region, <i>La Rioja</i>	-0.26476	0.01705	<0.0001	0.767
Region, <i>Melilla</i>	-0.34717	0.03721	<0.0001	0.707
Region, <i>Murcia</i>	0.05020	0.01292	0.0001	1.051
Region, <i>Navarra</i>	-0.51016	0.01070	<0.0001	0.600
Region, <i>País Vasco</i>	-0.04464	0.00806	<0.0001	0.956
PRTOTAL	0.0007247	5.09E-6	<0.0001	1.001
Cobertura, <i>Básico</i>	-0.03487	0.00736	<0.0001	0.966
Cobertura, <i>Otros</i>	-0.09236	0.03380	0.0063	0.912
Cobertura, <i>TodoRiesgo</i>	-0.50762	0.00603	<0.0001	0.602
Canal, <i>Canal 2</i>	0.15078	0.00368	<0.0001	1.163
Canal, <i>Canal 3</i>	-0.03119	0.00369	<0.0001	0.969
Siniestros	0.22242	0.00232	<0.0001	1.249

Tabla 12: Parámetros estimados a partir del Proportional Hazard model. Fuente: Elaboración propia.

En la **Tabla 12** está la estimación del Proportional Hazard model y cuyos resultados se comentarán a continuación. Todos los parámetros resultan ser significativos a un nivel del 5% pero en estimaciones previas los parámetros asociados a categorías como forma de pago D o algunas comunidades resultaban ser no significativos. Por esta razón y como ya se ha mencionado previamente, la forma de pago D se agrupó con la forma de pago C y las comunidades “no significativas” con la categoría de referencia (véase punto 3.2). En primer lugar, se observa que las formas de pago B y C tienen un hazard ratio similar y superior a 1, lo que indica que estas categorías reducen el tiempo esperado de permanencia de una póliza en la compañía. Esto se debe principalmente a que el asegurado puede tomar la decisión de no pagar los fraccionamientos de prima que les faltan por pagar. El resultado por género indica que las mujeres tienden a permanecer más tiempo en la compañía y que confirma los resultados obtenidos en el modelo anterior y los obtenidos por Brockett et al. (2008). En el **Gráfico 1** se observa la pequeña pero significativa diferencia entre la supervivencia de varones y mujeres.

En cuanto a la edad del conductor y a la antigüedad de la póliza se observa que los clientes con mayor edad o aquellos que tienen más años en la compañía tienen una mayor probabilidad de permanecer en la compañía que los clientes más jóvenes o con pólizas contratadas recientemente. Lo contrario pasa con la antigüedad del vehículo ya que el modelo indica que cuanto más viejo sea el automóvil, se espera que la póliza durará menos tiempo en la compañía.

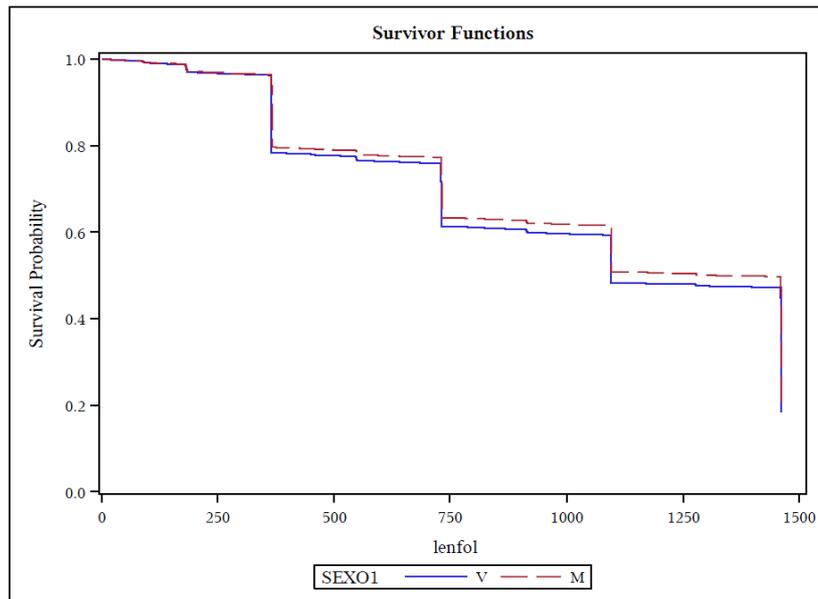


Gráfico 1: Curvas de Supervivencia entre varones y mujeres. Fuente: Elaboración propia.

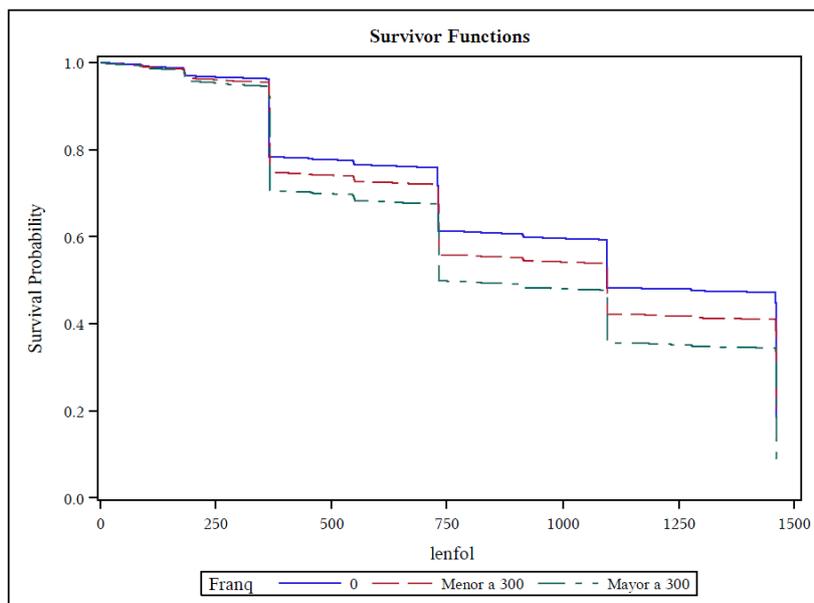


Gráfico 2: Curvas de supervivencia en función de la Franquicia contratada. Fuente: Elaboración propia.

En el **Gráfico 2** se refleja la diferencia entre las categorías referentes a la franquicia. Se observa como la diferencia entre las curvas se va acentuando a medida que van pasando los años. Por lo tanto, el tener pactada una franquicia con la compañía aseguradora reduce la duración de la póliza en la compañía y esta reducción vendrá relacionada directamente a la cuantía de la franquicia. En el **Gráfico 3** se indica que a medida que va aumentando el valor del vehículo, va aumentando la probabilidad de supervivencia, lo que resulta ser positivo para la compañía porque son clientes que pagarán primas más altas durante más tiempo y que tendrán una baja siniestralidad porque serán clientes más cuidadosos porque sus coches son caros.

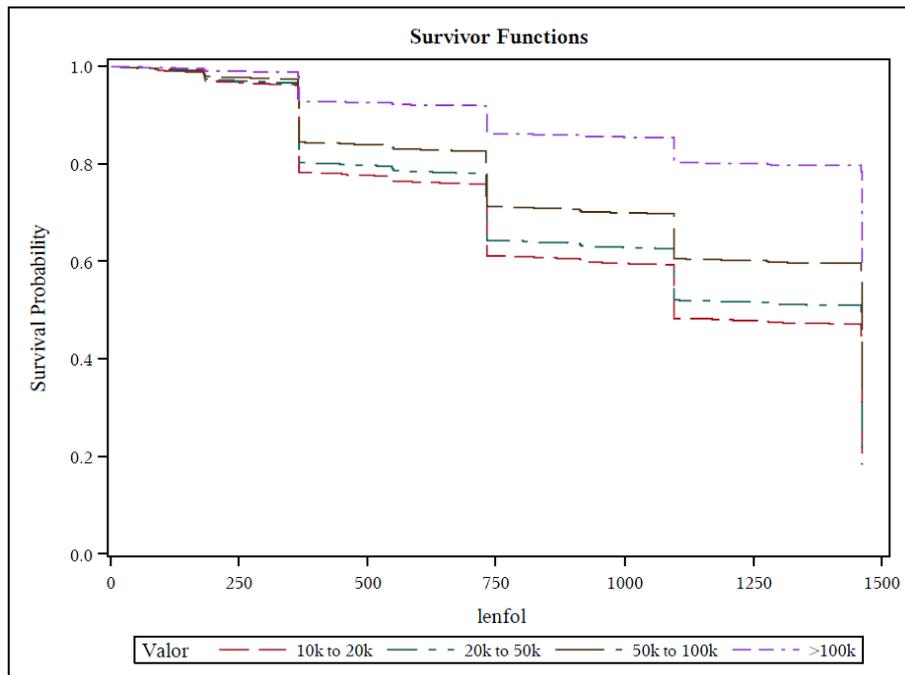


Gráfico 3: Curvas de supervivencia en función del valor del vehículo. Fuente: Elaboración propia.

En el modelo se ven las diferencias derivadas de la comunidad autónoma donde se suscribe la póliza. Así pues, las pólizas contratadas en Aragón, Islas Baleares, Cantabria, Castilla y León, Galicia, La Rioja, Melilla, Navarra o País Vasco han reflejado tener una duración mayor que las comunidades agrupadas en la zona de referencia, mientras que las que están contratadas en Andalucía, Andorra, Ceuta, Comunidad Valenciana, Extremadura o Murcia tienen una duración menor.

El parámetro de la variable Prima muestra que a medida que va aumentando la prima irá aumentando el tiempo esperado de la póliza, resultado que se puede contrastar con los parámetros del paquete de coberturas contratado que indican una mayor probabilidad de supervivencia de las pólizas con Todo Riesgo en comparación a las otras categorías. Además, aquellas pólizas agrupadas en el grupo de Terceros son las que menos tiempo esperado tienen.

Una póliza contratada a través del Canal 2 tiene una probabilidad menor de durar en la compañía en relación a los otros dos canales que son ligeramente diferentes (véase **Gráfico 4**). En cuanto a los siniestros, el parámetro resulta ser significativo y positivo, lo cual es un indicador de que el tiempo esperado de permanencia de la póliza se reducirá en función del número de siniestros que haya tenido la póliza. Así entonces, en el **Gráfico 5** pueden observarse cuatro curvas de supervivencia que se diferencian en el número de siniestros. A medida que va aumentando la cantidad de siniestros, la curva se va desplazando paralelamente hacia abajo.

En todos los gráficos se observa que hay grandes disminuciones de la probabilidad en tres puntos antes de llegar a la fecha final del estudio. Estas disminuciones se encuentran en los días que completan un año (365, 730 y 1095 días) y son causadas porque la mayoría de las pólizas son anuladas en el vencimiento y para la otra minoría se emiten notificaciones de anulación antes de cumplir el año.

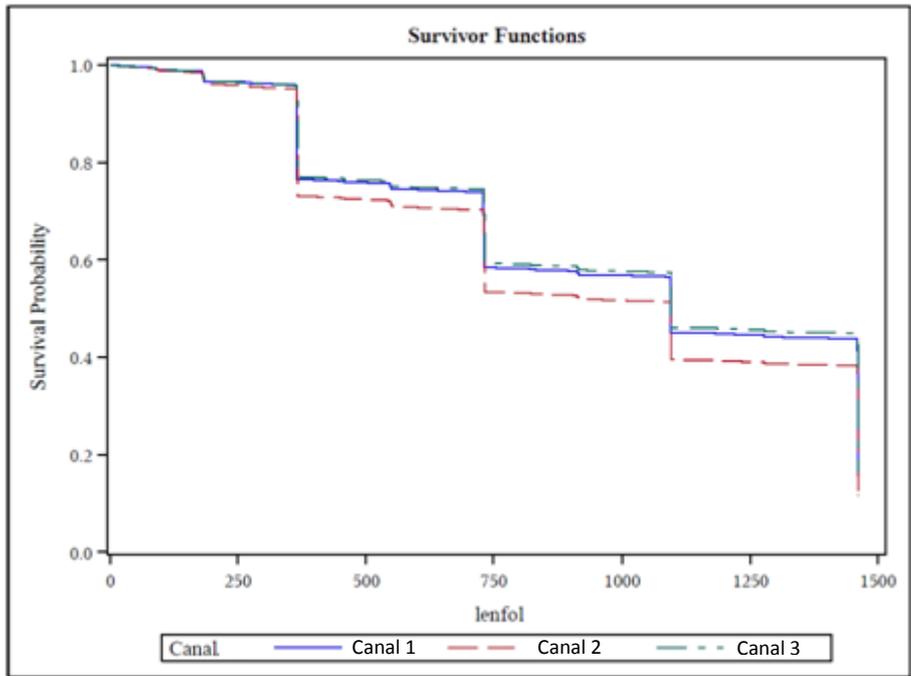


Gráfico 4: Curvas de supervivencia según el canal de distribución. Fuente: Elaboración propia.

De esta manera y teniendo en cuenta los hazard ratios, tener una franquicia superior a 300€, hacer el pago de la forma B o C, tener uno o más siniestros o suscribir la póliza en Andorra son los factores que más influencia tienen a la hora de reducir la probabilidad de supervivencia en comparación a la categoría de referencia.

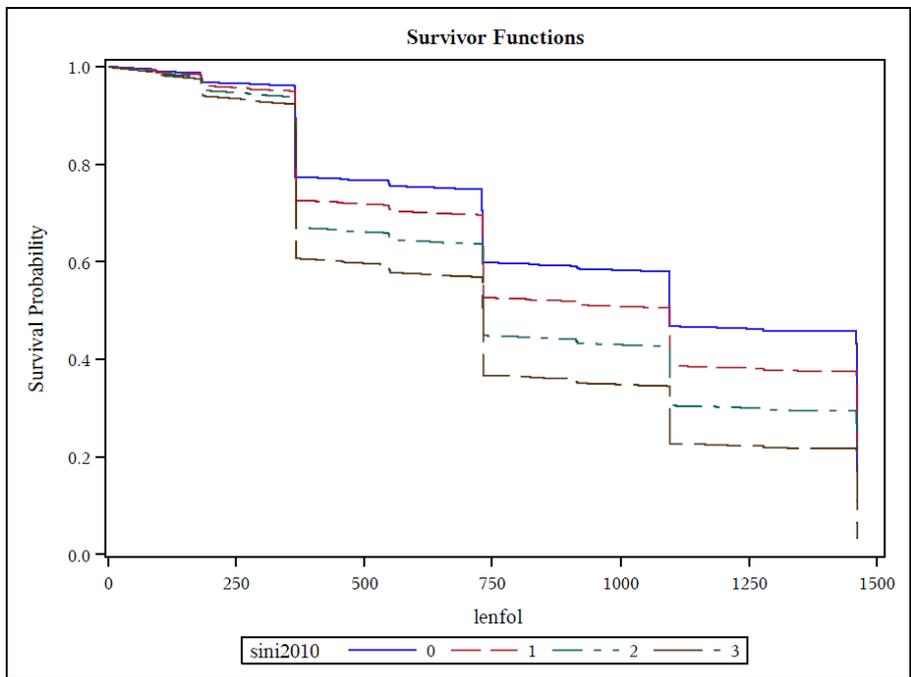


Gráfico 5: Curvas de supervivencia en función de la cantidad de siniestros. Fuente: Elaboración propia.

Mientras que tener asegurado un vehículo con un valor superior a los 50.000€, tener contratado un paquete de coberturas del grupo Todo Riesgo, suscribir la póliza en La Rioja, Melilla o Navarra son los factores que más influyen para aumentar la probabilidad de supervivencia del cliente dentro de la compañía.

5. Conclusiones

El sector asegurador a nivel europeo es un mercado muy competitivo y ésta característica se ve claramente en Autos que es uno de los ramos donde hay un mayor número de pólizas, una mayor cantidad de primas y una continua lucha por ganar más terreno dentro del sector atrayendo clientes de otras compañías por medio de fuertes campañas de marketing. De ahí viene la importancia de estudiar la retención porque es primordial para las empresas aseguradoras poder identificar tanto a sus mejores clientes como a los peores porque sería muy perjudicial para una compañía tener el mayor porcentaje del sector si en su cartera sólo tiene a los clientes con peor rendimiento ya sea por una alta frecuencia o una siniestralidad elevada. Así pues, el estudiar la retención es un primer paso en una estrategia de una compañía para identificar bien a su cartera de clientes y cómo poder influir sobre ella para retener a las pólizas con buen rendimiento y sanearla cartera de las pólizas que tengan mal rendimiento. Lo que se pretendía con este trabajo era: en una primera parte detectar las variables que podrían explicar la retención y su influencia sobre la probabilidad de renovar en un momento determinado del tiempo para luego comparar los resultados con los obtenidos en la segunda parte del trabajo que consistió en hacer un análisis de la supervivencia mediante el modelo del riesgo proporcional o Proportional Hazard.

En el primer modelo del trabajo se realizó una categorización de las variables con el fin de encontrar agrupaciones que tuvieran parámetros significativos para explicar la retención. De esta manera, categorías que resultaban ser no significativas fueron agrupadas con otras que eran significativas. En la estimación final, la variable que distinguía si la póliza pertenecía a la cartera o si era nueva producción resultó ser no significativa junto con la categoría de las pólizas suscritas en el mes 8.

El modelo estimaba que factores como suscribir la póliza en Ceuta o Murcia, tener 2 o más siniestros, una franquicia superior a los 300€, forma de pago B o C, un incremento de la prima casi nulo (0%) o suscribir la póliza en los dos primeros meses del año provocaban una disminución superior al 25% en la probabilidad de renovar respecto a la categoría que se había establecido como referencia en función del número de pólizas. Mientras que factores como un decremento de la prima superior al 5% o un incremento superior al 10%, suscribir la póliza en La Rioja o en alguna de las Islas (Baleares o Canarias), tener un paquete de coberturas incluidas en el grupo de Todo Riesgo o que el tomador sea mujer hacen aumentar la probabilidad de renovar en más de un 20% respecto a la misma categoría de referencia.⁵

Respecto a los dos últimos factores, se puede contrastar lo que algunos autores habían encontrado en estudios previos. Por ejemplo Gandolfi y Miners (1996), Mittal y Kamakura (2001) o Larivière y Van den Poel (2004) ya habían demostrado que las mujeres tenían un comportamiento diferente al de los hombres a la hora de decidir sobre productos financieros o de seguros ya que las mujeres tienen una probabilidad mayor a permanecer más tiempo en la compañía mientras que los hombres buscan alternativas diferentes o las compañías normalmente buscan al hombre como cabeza de familia para ofrecer sus productos. Respecto a las coberturas, Larivière y Van den Poel también mostraron que en el ámbito financiero, los productos más básicos tenían una tasa de retención menor a la que tienen productos más sofisticados. En este trabajo se ha

⁵ Categoría de referencia descrita en el punto 4.1.

llegado a un resultado similar en cuanto a que los productos con sólo coberturas frente a Terceros tienen una probabilidad menor a los productos con coberturas de Todo Riesgo.

En cuanto al modelo de supervivencia, se llegaron a resultados muy similares a los que se han obtenido en el modelo de retención. Teniendo en cuenta que la categoría de referencia general es casi idéntica en ambos modelos se puede hacer una comparación de los resultados. La principal diferencia radica en las comunidades que componen la zona de referencia en los modelos. Mientras que en el modelo de retención la zona de referencia comprendía a las comunidades de Castilla y León, Cantabria, Galicia, Melilla, País Vasco, Navarra, Cataluña y Andorra, en el modelo de supervivencia la zona de referencia estaba formada por Asturias, Islas Canarias, Castilla-La Mancha, Madrid y Cataluña. Estas agrupaciones fueron realizadas en función de la significación de los parámetros estimados.

En base al modelo de riesgo proporcional estimado y los ratios de riesgo obtenidos (hazard ratios) se evidenció que factores como tener una franquicia pactada, hacer el pago de la prima de una forma diferente a la A, el hecho de haber tenido siniestros o tener suscrita la póliza en Andorra, eran los que más hacían disminuir la probabilidad de supervivencia. Son resultados similares a los obtenidos en el modelo de retención. Esta disminución por franquicia puede ser causada por una influencia de compañías externas que ofrezcan las mismas coberturas a un precio más bajo con el objetivo de ganar cuota de mercado. En cuanto a la forma de pago, es posiblemente causada por el no pago de las fracciones de prima que han quedado pendientes motivo por el cual se anula la póliza. Finalmente, se sabe que ante la ocurrencia de siniestros es previsible que la compañía aumente la prima y el tomador haga una búsqueda de mejores precios en el mercado.

En sentido contrario, los factores que más pueden influenciar el aumento en el tiempo esperado de un cliente en la compañía aseguradora son tener asegurado un vehículo cuyo valor económico supere los 50.000€, o tener una póliza con un paquete de coberturas incluido en el grupo Todo Riesgo o tener la póliza suscrita en La Rioja, Navarra o Melilla. Incluyendo el parámetro de la variable Sexo, se observan las similitudes nuevamente con el modelo de retención. Los dos primeros factores tienen relación con la prima pagada por parte de estos clientes. Es posible que la compañía lleve a cabo acciones con el fin de retener estos clientes siempre que estos tengan un buen rendimiento dentro de ella.

A partir de aquí y de cara a futuras líneas de investigación, sería interesante poder disponer de variables que recojan información más personal sobre los clientes de la compañía como por ejemplo el nivel de ingresos del tomador, su nivel de educación, o su estado civil porque de esta manera se pueden obtener resultados en base a factores que pueden influir más en el momento de tomar la decisión de renovar la póliza o no. Con los resultados obtenidos en estos modelos y con otros que vengan derivados de un estudio conjunto entre modelos de recuento (para estimar un número esperado de siniestros) y modelos probabilísticos para estimar el coste de los siniestros se podría obtener una información bastante completa de la cartera de pólizas de una compañía. Esta información sería muy útil a la hora de planificar una estrategia con el objetivo de generar más ingresos reteniendo a los clientes que interesan por su buen rendimiento y atrayendo a otros “clientes objetivo” a través de una política de tarificación que permita ofrecer unas primas que sean atractivas en el mercado y que generen beneficios. La

información también serviría para ir deshaciéndose de aquellos “clientes no deseados” con aumentos de prima que reportarán más ingresos hasta cuando el cliente definitivamente se vaya y una disminución en los costes gracias a ese saneamiento.

En el modelo de retención se tuvo la limitación de sólo poder disponer de datos de ocho meses del año 20XX con lo cual puede haber cierta distorsión en los resultados aunque al estimar el modelo de supervivencia se mostró que las variables tenían el mismo efecto positivo o negativo sobre la variable dependiente. Es probable que al poder disponer de una base de datos que abarque el año entero, el porcentaje de aumento o disminución en la probabilidad aumente o disminuya pero no cambie de signo.

En la base de datos utilizada en el modelo de supervivencia se disponía de la variable *Motivo* que recoge los diferentes motivos por los que una póliza ha sido anulada. Pero debido a que desde la misma compañía advirtieron de la poca fiabilidad de la información recogida en esta variable, no se realizaron estimaciones sobre ésta y se espera poder hacer un estudio posterior para analizar las posibles diferencias en cuanto a la influencia y la significación de las variables según el motivo de anulación. Puede darse que una variable sea significativa en la estimación del modelo para un determinado motivo y que no lo sea en la estimación de otro. Para ello, es necesario dividir la base de datos en muestras según el motivo y hacer estimaciones individuales para cada muestra.

Para terminar, se quiere dar un agradecimiento a la compañía aseguradora y en especial al equipo actuarial por facilitar las bases de datos que han hecho posible la realización de este trabajo.

6. Bibliografía

- Au, T., Li, S. And Ma G. (2003). Applying and Evaluating Models to Predict Customer Attrition Using Data Mining Techniques, *Journal of Comparative International Management*, Vol. 6, No. 1, pp. 10-22.
- Brockett, P. L., Golden, L. L., Guillen, M., Nielsen, J. P., Parner, J. and Perez-Marin, A. M. (2008). Survival Analysis of a Household Portfolio of Insurance Policies: How Much Time Do You Have to Stop Total Customer Defection?, *The Journal of Risk and Insurance*, Vol. 75, No. 3, pp. 713-737.
- Frees, E.W. (2010). Generalized Linear Models. En *Regression Modeling with Actuarial and Financial Applications*, Cambridge University Press. 362-377. Cambridge.
- Frees, E.W. (2010). Proportional Hazard model. En *Regression Modeling with Actuarial and Financial Applications*, Cambridge University Press. 392-395. Cambridge.
- Gandolfi, A.S., and Miners, L. (1996). Gender-based Differences in Life Insurance Ownership, *The Journal of Risk and Insurance*, Vol. 63, No. 4, pp. 683-693.
- Greene, W.H. (2002). Models for Discrete Choice. En *Econometric Analysis*. Ed, Prentice Hall. 663-688. Upper Saddle River, New Jersey (USA).
- Guillen, M., Parner, J., Densgsoe, C. and Perez-Marin, A. M. (2003). Using Logistic Regression Models to Predict and Understand Why Customer Leave an Insurance Company. In *Intelligent and Other Computational Techniques in Insurance. Theory and Applications*, Jain, L. and Shapiro, A., (eds.). 465-490. Hackensack, New Jersey (USA).
- Guillen, M., Nielsen, J.P., Scheike, T. and Perez-Marin, A.M. (2011). Time-varying effects in the analysis of customer loyalty: a case study in insurance, *Expert Systems with Applications*, 39, pp. 3551-3558.
- Larivière, B. and Van den Poel, D. (2004). Investigating the role of product features in preventing customer churn, by using survival analysis and choice modelling: The case of financial services, *Expert Systems with Applications* 27, pp. 277-285.
- Larivière, B. and Van den Poel, D. (2005). Investigating the post-complaint period by means of survival analysis, *Expert Systems with Applications* 29, pp. 667-677.
- Liao, T. F. (1994) Interpreting Probability Models: Logit, Probit, and other Generalized Linear Models. En *Quantitative applications in the Social Sciences*. 101. A Sage university papers series. Thousand Oaks, California (USA).
- Mittal, V. and Kamakura, W. A. (2001). Satisfaction, repurchase intent, and repurchase behaviour: Investigating the moderating effect of customer characteristics, *Journal of Marketing Research*, Vol. 38, No. 1, pp. 131-142.
- Reichheld, F. F. and Sasser, W. E. (1990). Zero defections: Quality Comes to Services, *Harvard Business Review*, Vol. 68, No. 5, September/October, pp. 105-111.
- Smith, K. A., Willis, R. J. and Brooks, M. (2000). An analysis of customer retention and insurance claim patterns using data mining: a case study, *Journal of the Operational Research Society* 51, pp. 532-541.

Anexo 1

Si tenemos un conjunto de n observaciones individuales independientes, Y sigue una distribución Binomial de parámetros $B(n, \pi)$ que es de la familia exponencial con las siguientes características:

$$\theta = \ln \frac{\pi}{1 - \pi}$$

$$b(\theta) = n * \ln(1 + e^\theta)$$

$$E(Y) = n\pi$$

$$Var(Y) = n\pi(1 - \pi)$$

Donde θ es el parámetro canónico. La función de enlace canónica es el logit, de forma que:

$$\theta = \ln \frac{\pi}{1 - \pi} = x_i' \beta$$

Odds ratios asociados a cada variable: es el cociente de probabilidades relativas de una opción respecto a la otra cuando se incrementa en una unidad la variable explicativa. Se pueden calcular a partir de los coeficientes asociados a cada una de las variables:

$$P_i = P(Y_i = 1) = \frac{\exp(x_i' \beta)}{1 + \exp(x_i' \beta)}$$

$$odds_i = \frac{P_i}{1 - P_i} = \frac{\frac{\exp(x_i' \beta)}{1 + \exp(x_i' \beta)}}{\frac{1}{1 + \exp(x_i' \beta)}} = \exp(x_i' \beta)$$

$$\frac{odds_i(x_j + 1)}{odds_i(x_j)} = \exp(\beta_j)$$

Anexo 2

De forma similar se puede calcular la probabilidad conjunta de los tiempos de fallecimiento (no renovar) de cada uno de los individuos, como una función del vector de parámetros β dados los valores de las variables X_i .

$$L(\beta) = \prod_{i=1}^n \frac{\exp(x_i' \beta)}{\sum_{i=1}^n \exp(x_i' \beta)}$$

Los métodos de máxima verosimilitud intentan hallar los valores del vector de parámetros β que maximizan ésta probabilidad, es decir, los parámetros que ofrecen la mayor probabilidad conjunta de los tiempos de fallecimiento con los valores de las variables asociadas.

Anexo 3

Descriptivos de las variables utilizadas en el modelo de retención y tablas de distribución de pólizas.

Variable	Media	Desv Estándar	Mínimo	Máximo
VIGOR	0,8159764	0,3875038	0	1
Age	51,020481	13,4913543	15	100
Antig	6,0323797	5,7921711	0	50
Franq	42,933997	135,0837331	0	6000
Valor	20016,59	12982,35	1	489728
EdadVeh	8,0970841	5,5197925	0	85
Siniestros	0,366237	0,6894902	0	3
Var. Prima	-9,211909	69,8687554	-3063,86	3074,86

Sexo	% Pólizas
Varón	71.43%
Mujer	28.57%

Tipo de póliza	% Pólizas
Cartera	82.79%
Nueva Producción	17.21%

Franquicia	% Pólizas
Sin franquicia	87.91%
Menor a 300€	8.22%
Mayor a 300€	3.86%

Cobertura	% Pólizas
TodoRiesgo	18.95%
Terceros	66.40%
Básico	14.66%

FPAGO	% Pólizas
A	78.44%
S	17.08%
T	4.48%

Canal	% Pólizas
Canal 1	41.68%
Canal 2	30.00%
Canal 3	28.32%

Valor	% Pólizas
0k-10k	13.62%
10k-20k	46.41%
20k-50k	37.28%
>50k	2.69%

Siniestros	% Pólizas
0	73.35%
1	18.86%
2	5.60%
3 o más	2.19%

CCAA	% Pólizas	CCAA	% Pólizas	CCAA	% Pólizas
Andalucía	15.59%	Com. Valenciana	13.78%	Madrid	2.98%
Aragón	5.38%	Extremadura	2.26%	Murcia	1.23%
Asturias	0.81%	I. Baleares	7.64%	Zona REF	36.84%
Castilla-La Mancha	3.60%	I. Canarias	8.56%		
Ceuta	0.13%	La Rioja	1.20%		

Anexo 4

Estadísticos descriptivos de las variables utilizadas en el modelo de supervivencia.

Variable	Media	Desv Estándar	Mínimo	Máximo
Edad	49,911573	13,601955	14	97
Antig	5,5984933	5,5147193	1	49
EdadVeh	7,1127511	41,893764	-7989	2006
Valor	17873,92	13398,15	0	3.373.000
Prima	426,47	263,2063	0	10861,79
Siniestros	0,1936505	0,5355858	0	12
Nº Días	915,8951	401,3331	1	1461

Sexo	% Pólizas
Varón	73.22%
Mujer	26.78%

Franquicia	% Pólizas
Sin franquicia	479.841
Menor a 300€	44.888
Mayor a 300€	21.076

FPAGO	% Pólizas
A	72.94%
S	20.95%
T	6.11%

Valor	% Pólizas
0k-10k	20.57%
10k-20k	44.99%
20k-50k	32.35%
50k-100k	1.92%
>100k	0.17%

Cobertura	% Pólizas
TodoRiesgo	20.49%
Terceros	65.69%
Básico	13.53%
Otros	0.29%

Canal	% Pólizas
Canal 1	38.14%
Canal 2	30.04%
Canal 3	31.82%

Siniestros	% Pólizas
0	85.39%
1	11.09%
2	2.66%
3 o más	0.87%

CCAA	% Pólizas	CCAA	% Pólizas	CCAA	% Pólizas
Andalucía	15.81%	Ceuta	0.12%	Melilla	0.17%
Andorra	0.14%	Com. Valenciana	14.04%	Murcia	1.38%
Aragón	5.06%	Extremadura	1.97%	Navarra	3.03%
Baleares	6.60%	Galicia	4.97%	P. Vasco	3.68%
Cantabria	0.66%	La Rioja	1.13%	Zona REF	36.69%
Castilla y León	4.55%				