Universidad Nacional de Colombia
Facultad de Ciencias
Maestría en Actuaría y Finanzas
Proyecto final Introducción a la ciencia de datos
Bogotá, Septiembre de 2021

Modelo de propensión de fuga para seguros de automóvil voluntarios

Evelyn Johanna Romero [1]

Introducción

La retención de clientes es trascendental para mantener e incluso optimizar la posición de una compañía en el mercado. Si bien para el crecimiento de un negocio es importante la captación de nuevos usuarios, lo es también el conservar los existentes, teniendo en cuenta que en términos de costos generalmente es más complicado adquirir nuevos consumidores que retenerlos. En diversas industrias el riesgo de retiro de clientes se presenta de manera constante, por lo que identificar los posibles factores que conllevan a dicha pérdida es un criterio de bastante relevancia para las decisiones que se deban llevar a cabo en cuanto a la estabilidad y posicionamiento del negocio a corto y largo plazo. Con la presencia de contingencias como lo es la actual pandemia del Covid-19 que genera un cambio en los patrones de consumo y la priorización de gastos, dicha identificación adquiere un peso mucho mayor para la ejecución de medidas preventivas efectivas en la conservación de clientes.

A pesar de que en la literatura se han llevado a cabo diversas investigaciones para tratar con esta situación, es de interés llevarlo a cabo en el contexto de America Latina, más específicamente en Colombia, ya que se ha encontrado de acuerdo a la Federación de Aseguradores Colombianos *Fasecolda* [1] que para finales de Agosto de 2019 el 90.3 % de los hogares colombianos ha tenido algún tipo de esquema de protección o aseguramiento, sin embargo solo el 30.3 % adquieren un seguro voluntario. La misma entidad afirma que dichos seguros voluntarios se concentran en hogares de nivel socio-económico alto donde el seguro de vida y el seguro de vehículo son los que más se adquieren, estableciendo así que la industria aseguradora tiene una gran oportunidad de expansión en el país, sobretodo en el ramo de automóviles, teniendo en cuenta que el 70 % de los vehículos que circula a nivel nacional no está asegurado frente a eventualidades como el robo, como lo afirma el presidente de la Asociación del Sector Automotriz y sus Partes *Asopartes* [7]. En base a dicho comportamiento en seguros voluntarios es vital no solo incrementar la cultura del seguro en el país para aumentar la adquisición de este tipo de productos, sino también evitar que se presenten abandonos que generen aún más la falta de aseguramiento.

Teniendo en cuenta el comportamiento de los seguros voluntarios en el país, así como la trascendencia que tiene el determinar los factores de riesgo de un posible abandono en los usuarios, se optará por la construcción de un modelo de propensión de fuga, más concretamente un modelo de sobrevivencia o riesgos proporcionales de Cox para la detección de deserción de clientes en la compañía AON Risk Services, especialmente para el área Affinity que tiene a su cargo la comercialización de seguros voluntarios masivos, y en el ramo de automóviles que es uno de los que tienen mayor movimiento dentro de dicha área. Hasta la fecha no se cuenta con una herramienta para medir e identificar los segmentos o perfiles de las personas más propensas a retirarse de una póliza de seguro, por lo que la única opción es recurrir a la experiencia que se tiene por parte del equipo comercial y a partir de ella diseñar estrategias de retención, sin embargo el problema es que se implementan a toda la población, requiriendo mayores gastos en ciertos recursos que si se implementaran a un grupo o segmento específico, por esta razón a pesar de ser la única solución llevada a cabo en la compañía no es la más acertada o viable en términos económicos, prácticos y comerciales.

Metodología y datos

La construcción de un modelo de churn permitirá la adecuada segmentación de los consumidores con un potencial retiro para posteriormente establecer estrategias y acciones anticipadas que permitan reducir esta tasa de retiros. Además de esto la industria en general se ha visto enfrentada a la ausencia de la cultura del seguro a pesar de las diferentes estrategias e

[1] E-mail: ejromeroo@unal.edu.co

incentivos empleados, por lo que es primordial seguir dando solución a las situaciones que hacen que este problema continúe. Para realizar todo el desarrollo e implementación requerido se definirán algunos conceptos para comprender el transfondo de la técnica de machine learning a emplear, esta información se encuentra en el anexo de este documento. Para el contexto del presente estudio se utilizará la metodología de análisis de supervivencia con el objetivo de analizar y predecir el tiempo que los clientes pueden tardar en cancelar su póliza de automóvil. Dicho proceso se llevará a cabo en el software `R`

Datos

Para iniciar con el procesamiento y limpieza de la data se tiene por un lado la información sociodemográfica de afiliados y retirados de toda la compañía, es decir todas las personas afiliadas a alguna de las pólizas de seguro disponibles. Como primer paso se seleccionaron solo los individuos que están o estuvieron afiliadas a un seguro de automóvil. Al analizar el contexto del problema se decidió eliminar la variable ciudad ya que había una variable mucho más general llamada región que contenía menos categorías y podía ser más manejable. Adicionalmente se identificaron los registros con inconsistencias en variables como la edad, y se procederá a eliminar el número de identificación de cada persona para anonimizar la información una vez hecho el cruce y consolidación de la información, así como sus nombres y apellidos. Finalmente para esta base permanecieron las variables Edad, Estado civil, Género, Fecha de afiliación, Región (Localización del usuario), Beneficiarios, Número de convenios y Póliza de vehículos (1 si seguía activo o 0 si se había presentado retiro)

Para la data de retirados no se presentaron inconsistencias significativas, sin embargo es necesario hacer una unificación de las categorías en variables como la marca del vehículo ya que hay que categorías que por errores de digitación no tienen el mismo nombre a pesar de referirse al mismo tipo. Una vez realizado el cruce entre ambas bases se eliminará la identificación y datos de contacto como la placa y número de póliza, también nombres y apellidos e incluso variables que pueden tener menor relevancia como lo son la referencia del vehículo (con la marca es suficiente al ser más general) y su color. Por tanto las variables finales a tratar en esta base son Marca, Modelo, Valor asegurado total, Prima media mensual y Fecha de retiro. Se utilizarán las variables Fecha de afiliación y Fecha de retiro para formar una única columna que se llame **Antigüedad** e indique en años el tiempo en que la persona ha permanecido o permaneció en la compañía.

Adicionalmente, dentro de las variables originales no se encontraba alguna que manifestara como es el salario o ingreso económico que tienen los usuarios a pesar de que podría ser una característica bastante relevante, sin embargo poco tiempo después de empezar con este estudio se hizo entrega por parte de la compañía de un dato adicional que es la ocupación del afiliado, que variaba entre estudiante, persona con posgrado, profesor universitario entre otros, por lo que se decidió inferir a partir de esta información un posible rango salarial, que se clasificó en tres categorías; salario medio, medio alto y alto, o como se denotó **MEDIUM**, **MEDIUM HIGH** y **HIGH**.

Teniendo en cuenta las inconsistencias presentadas se decidió eliminar los datos atípicos de la variable edad, es decir valores inferiores a 16 años y superiores a 100, así mismo se eliminaron registros faltantes en el valor de la prima y el valor asegurado dejando un total de 4461 individuos. A partir del análisis exploratorio que se hizo también fue notorio el poco impacto que tenía la categoría de **NO REGISTRA** en la variable género, más que todo porque para estos clientes tampoco se contaba con información de estado civil, por eso mismo también se decidió eliminar 59 registros con este inconveniente dejando un total finalmente de 4462 usuarios para el correspondiente estudio.

Por último, se realizó una matriz de correlación (Figura 1) con las variables númericas para conocer si se presenta multicolinealidad en lo que más adelante se emplearan como covariables. Se encontró que dos de de estas variables tienen una correlación significativa que podría afectar el análisis, estas variables son **Beneficiarios** y **Número de convenios**, lo cual tiene mucho sentido ya que a mayor número de productos que el cliente tenga con la compañía mayor será la cantidad de beneficiarios que tendrá a su cargo, es por ello que al identificar que ambas variables representan algo similar se procederá a eliminar una de ellas, la cual será la variable **Beneficiarios** ya que la otra variable puede ser un poco más específica y así mismo brindar mejores resultados. Con el restante de variables se identificó que no presentan una correlación fuerte por lo que todas irán al análisis

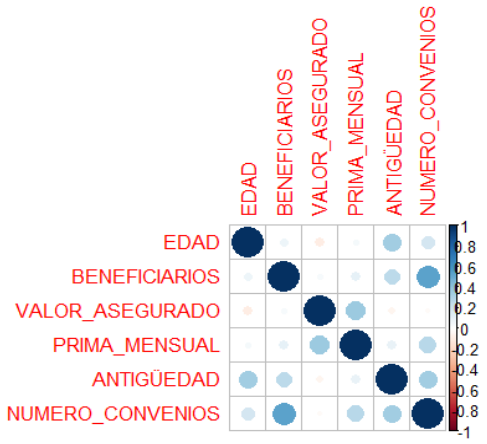

Figura 1: Matriz de correlación

Una vez realizado el análisis descriptivo y exploratorio de la información consolidada se decidió para efectos prácticos categorizar variables como edad y modelo, además de inferir qué gama de vehículo del afiliado podría ser de acuerdo a su valor asegurado, prima y marca para así crear una sola categoría que pueda aportar más al modelo. A continuación se presenta los grupos creados

Edad	Inferior a los 40 años: Crecimiento De 40 a 60 años: Apasionados Superior a 60 años: Expertos
Modelo	Inferior a 2005: Antiguos De 2005 a 2015: Intermedios Superior a 2015: Modernos
Gamas	Inferior a una valor asegurado de 50'000.000: Baja Entre 50'000.000 y 80'000.000: Media Superior a 80'000.000: Alta

Cabe resaltar que la variable creada **Gamas** también tuvo en cuenta las marcas que en el mercado son más costosas y la prima media cancelada hasta el momento por los usuarios, en donde predomina en el grupo de gama alta marcas como Audi, BMW y Mercedes Benz y de gama baja autos como algunos Renault o Chevrolet, no obstante se clasificó de acuerdo al valor asegurado porque esta característica reflejaba el tipo de vehículo del afiliado. Así mismo, la categorización de la variable edad está influenciada tanto en unos conceptos comerciales que maneja la compañía que proporcionó los datos para el estudio, así como en la distribución de dicha variable una vez removidos los outliers, por tanto en base al análisis las variables finales para la implementación del modelo son:

A. Edad

B. Género (F, M)

C. Estado civil (6 categorías incluyendo NO REGISTRA)

D. Valor asegurado, marca y prima mensual $\rightarrow$ Gama de vehículo (Alta, media y baja)

E. Número de convenios

F. Modelo (Antiguos, Intermedio, Moderno)

G. Antiguedad (Años de permanencia en la compañía) y Salida (Variable con valor de 1 si la persona ya se retiró, 0 en caso contrario)

H. Salario (Medio, Medio alto y Alto)

Resultados

Una vez categorizadas las variables que se decidieron transformar, se procede a realizar la partición de los datos entre conjuntos de entrenamiento y de prueba para la adecuada creación y replicación del modelo, así como la obtención del mejor desempeño posible. Inicialmente la distribución para las personas retiradas y activas, identificadas con 0 y 1 respectivamente en la variable **SALIDA**, se tiene que el 61 % del total siguen con su póliza mientras que el 39 % restante ya se retiraron. Teniendo en cuenta este comportamiento se decide emplear el 70 % del total de registros para el conjunto de entrenamiento y el 30 % para los datos que se emplearán como test.

	SALIDA	n	percent
1	0	2852	0.61
2	1	1789	0.39

Cuadro 1: Distribución de variable SALIDA para el total de la población

Estimación por el método de Kaplan - Meier

Para empezar el análisis de supervivencia se crea una variable de tipo *survival* que requiere tanto la variable tiempo, en este caso ANTIGUEDAD, como la variable evento, SALIDA para este estudio, tal como lo muestra la función en `R`, `Surv(time, event)`. Este objeto creado tiene como propósito emplear la censura que se pueda presentar, como se arroja en esta salida

ANTIGÜEDAD	SALIDA	surv
6	0	6+
13	1	13
6	0	6+
4	1	4
6	0	6+

Figura 2: Salida variable de tipo *survival*

Por ejemplo para el primer registro en donde la variable salida es 0, es decir aún la persona se encuentra activa, y donde el valor de antigüedad es 6, se obtiene un valor para la variable creada de **6+** lo que significa que como el usuario aún se encuentra activo hasta el momento presenta 6 años de permanencia pero podrían ser más hasta que se presente el retiro correspondiente, caso contrario a las filas 2 y 4, en donde como el cliente ya no se encuentra con su póliza entonces su tiempo de permanencia definitivo fueron de 13 y 4 años respectivamente. Con la creación de una variable de este tipo se procede a estimar como primera medida la función de sobrevivencia por medio del método de Kaplan - Meier, donde aún sin tener en cuenta ninguna variable en específico se obtiene lo presentado en 3. A partir de esta estimación se tiene que hay una probabilidad del 50 % de que a los 25 años se presente un retiro.

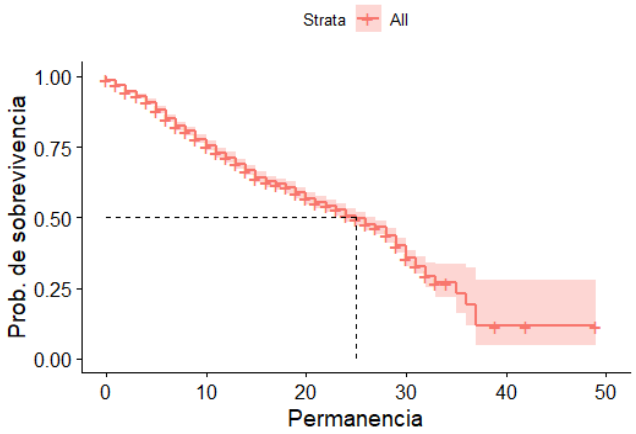

Figura 3: Curva de supervivencia general

Así mismo se puede apreciar que un poco antes de los 40 años de permanencia la probabilidad de retiro se estabiliza, lo que indica que a partir de que una persona, independientemente de sus características, llegue mínimo a ese tiempo con la compañía la probabilidad de que abandone su póliza no incrementará más. Adicionalmente, a partir de este método de estimación se puede obtener para cada año el número de personas en riesgo hasta el tiempo t, número de eventos ocurridos hasta esa fecha y la probabilidad de permanencia para ese tiempo correspondiente, como lo indica la columna *surv* de la tabla 2. Por ejemplo a los 10 años de tener en funcionamiento la póliza la probabilidad de seguir con la compañía es de 75 %

Si bien esta estimación y curva general da un panorama amplio de como se puede comportar la población, es de interés conocer dichas probabilidades a partir de ciertas características específicas. Empezando con la segmentación por género se obtiene lo siguiente

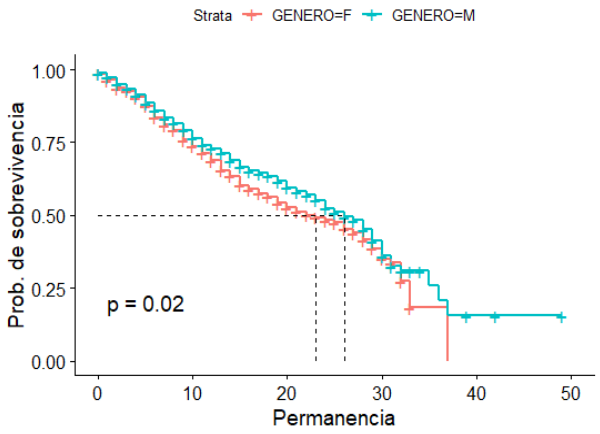

Figura 4: Curva de supervivencia de acuerdo al género

time	n.risk	n.event	n.censor	surv	std.err	upper	lower
0.00	3207.00	43.00	37.00	0.99	0.00	0.99	0.98
1.00	3127.00	56.00	75.00	0.97	0.00	0.97	0.96
2.00	2996.00	74.00	175.00	0.94	0.00	0.95	0.94
3.00	2747.00	42.00	65.00	0.93	0.00	0.94	0.92
4.00	2640.00	61.00	69.00	0.91	0.01	0.92	0.90
5.00	2510.00	78.00	71.00	0.88	0.01	0.89	0.87
10.00	1762.00	56.00	69.00	0.75	0.01	0.77	0.74
15.00	1215.00	47.00	66.00	0.64	0.02	0.66	0.62
20.00	854.00	28.00	52.00	0.57	0.02	0.59	0.55
25.00	556.00	13.00	40.00	0.50	0.02	0.52	0.47
30.00	249.00	28.00	78.00	0.36	0.04	0.39	0.33
35.00	7.00	1.00	0.00	0.23	0.19	0.34	0.16
36.00	6.00	1.00	0.00	0.19	0.26	0.32	0.12
37.00	5.00	2.00	0.00	0.12	0.45	0.28	0.05
39.00	3.00	0.00	1.00	0.12	0.45	0.28	0.05
42.00	2.00	0.00	1.00	0.12	0.45	0.28	0.05
49.00	1.00	0.00	1.00	0.12	0.45	0.28	0.05

Cuadro 2: Estimación de probabilidades de supervivencia a nivel general

A partir de la Figura 4 se puede inferir que las mujeres tienen mayor probabilidad de retiro a comparación de los hombres, y que relacionado con la gráfica anterior (3), las personas de género masculino son las que tienen una probabilidad de retiro constante a partir de aproximadamente los 37 años de permanencia. En esta figura se presenta adicionalmente un valor denotado con la letra p, que hace referencia al p-valor obtenido en una prueba conocida como *Log-Rank test* o de rango logarítmico,

la cual permite comparar las curvas de supervivencia de dos grupos, en otras palabras es un test de hipótesis estadístico que prueba la hipótesis nula de que las curvas de supervivencia de dos poblaciones no difieren. En este caso, $p < 0.05$ indicaría que los dos grupos de género son significativamente diferentes en términos de supervivencia, es decir esta variable posiblemente sí influye en el retiro de un cliente.

Debido a que se categorizó la edad dicha estimación y curvas de supervivencia también se pueden emplear con esta variable. Como se presenta en la Figura 5 para los primeros años de permanencia las categorías de expertos y apasionados tienen probabilidades bastante similares y por ello el entrelazamiento que se logra identificar, no obstante para estos dos grupos a partir de los 20 años ya se presenta una diferencia significativa en sus probabilidades. Por otro lado en la categoría apasionados hay una probabilidad de riesgo un poco mayor a comparación del grupo de expertos, y además este último grupo se estabiliza a partir de los 37 años aproximadamente, como en curvas pasadas. Sin dejar de lado los clientes en crecimiento estos abandonarán su póliza antes de los 40 años de permanencia, así mismo a los 11 años ya tendrán un 50 % de posibilidad de retiro. Finalmente para esta variable, la H_0 del test de rango logarítmico se rechaza indicando así que estos grupos de edad son significativamente diferentes.

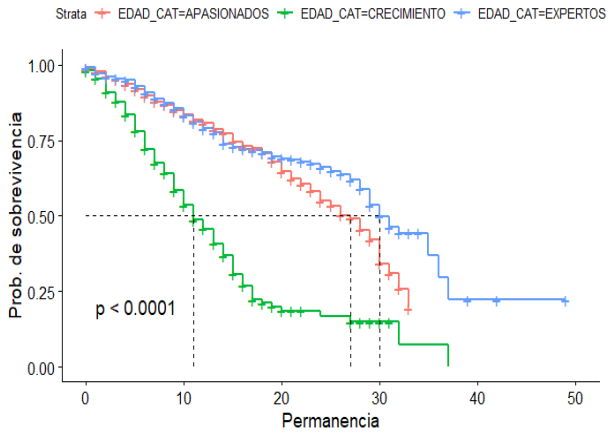

Figura 5: Curva de supervivencia de acuerdo a los grupos de edad

Otra característica de interés a nivel univariado es el salario, cuya estimación de la curva de probabilidades se expone a continuación

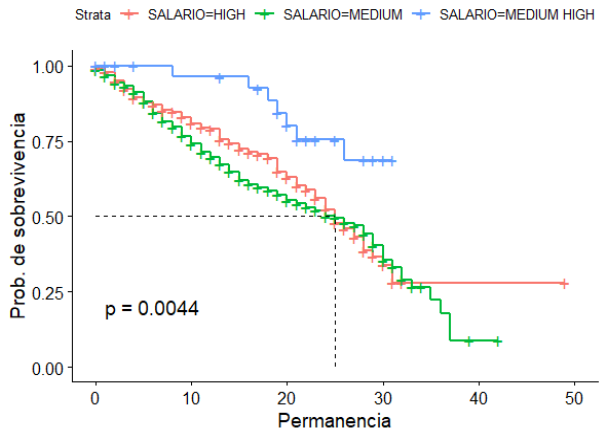

Figura 6: Curva de supervivencia de acuerdo al rango salarial

De esta figura se puede identificar que entre los tres grupos las personas de salario medio tienen mayor probabilidad de

cancelar su póliza y que a los 25 años de continuidad en la compañía existe un 50 % de probabilidad de finalmente abandonar, tal como sucede también con los usuarios de salario alto. Para este último grupo la probabilidad de deserción se estabiliza a partir de los 31 años aproximadamente y para la categoría de usuarios con salarios medios altos la probabilidad de retiro es sorprendentemente baja, es decir, aún con 30 años en la compañía la probabilidad de aún conservar su póliza es de casi el 75 %, lo que indicaría que son usuarios que pueden llegar a ser bastante fieles con su póliza de seguro.

Para finalizar este tipo de análisis no solo con características sociodemográficas del usuario, se realizará el mismo procedimiento con las categorías de gamas presentadas, obteniendo lo presentado a continuación

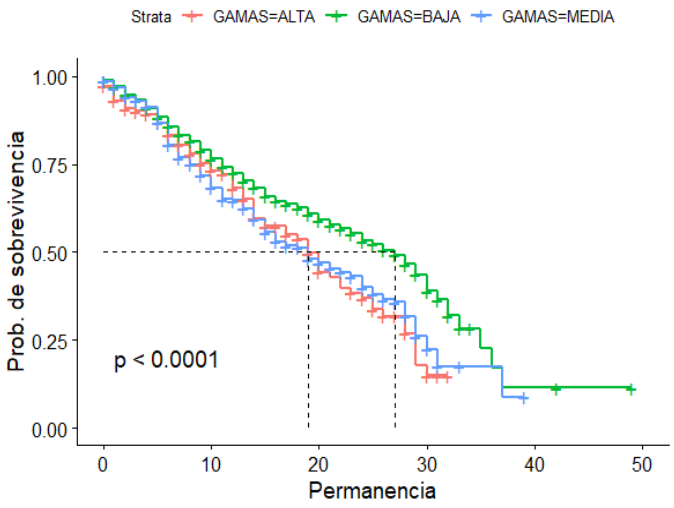

Figura 7: Curva de supervivencia de acuerdo a la gama del vehículo

De acuerdo con la Figura 7 se tiene que la condición de abandono para los vehículos de gama baja es constante a partir de los 37 años a comparación de los otros grupos, así mismo para esta categoría existe un 50 % de probabilidad de deserción hasta casi los 30 años de estancia, mientras que para las gamas altas y medias esta probabilidad de deserción se presenta antes de los 20 años, permitiendo identificar que las personas con vehículo de gama baja pueden ser más fieles debido a que las primas a cancelar suelen ser más bajas y esto puede implicar menos complicación en el pago de este monto, contrario a vehículos de gama alta o media que por las características de su vehículo tienen que cancelar una prima mayor y por ende su estancia en la compañía corre más peligro.

A partir de todo lo analizado con algunas de las variables de interés que por medio del test de rango logarítmico han demostrado que pueden ser características que influyan en el retiro de un usuario, surge también interés por conocer como son dichas curvas si se emplearan al menos dos factores. Para realizar esto se realizará la estimación con la combinación de algunas variables para conocer como es el comportamiento. El primero de ellos será la curva resultante entre las diferentes categorías de género y salario, arrojando lo siguiente

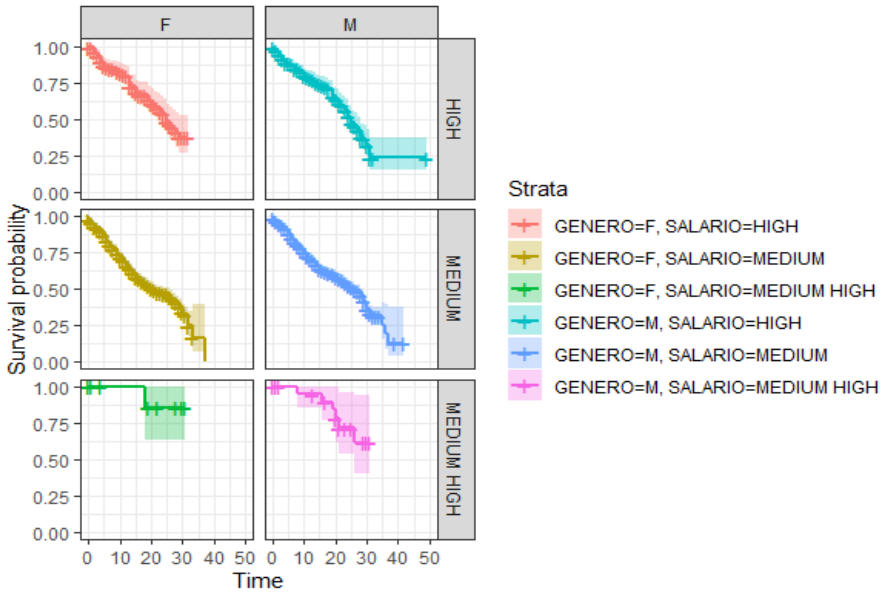

Figura 8: Curva de supervivencia de acuerdo al género y rango salarial

Con base en esta gráfica se puede identificar que las personas de género femenino y con salario medio tienen mayor probabilidad de retiro, así mismo para los clientes de este mismo género pero con salario medio alto la probabilidad de continuar con la póliza es de más del 80 %, pero más allá de indicar son un grupo de usuarios más fieles también nos hace ver que posiblemente no hay suficientes datos para llegar a una estimación confiable, por lo que se sugeriría aumentar el tamaño de muestra para observar un comportamiento más razonable, algo similar ocurre con el mismo rango salarial pero para los hombres.

Con la variable género también podemos hacer el análisis atado al estado civil, ya que esta última en los análisis univariados no se tuvo en cuenta al ser una variable con más de tres categorías. Los resultados arrojan que en casos como las combinaciones masculino-no registra, femenino-viudo y masculino-viudo se obtienen cambios abruptos de un tiempo a otro en la probabilidad o el retiro tiene una posibilidad de ocurrencia baja, esto probablemente al mismo problema mencionado anteriormente y es la falta de registros en estas categorías específicas que impiden una estimación más completa. En general para las mujeres solteras el tiempo más tardío de retiro es a los 36 años de estancia en la compañía, mientras que para los hombres existe un mayor tiempo de permanencia en el estado civil casado. Todo esto se puede identificar en la Figura 9

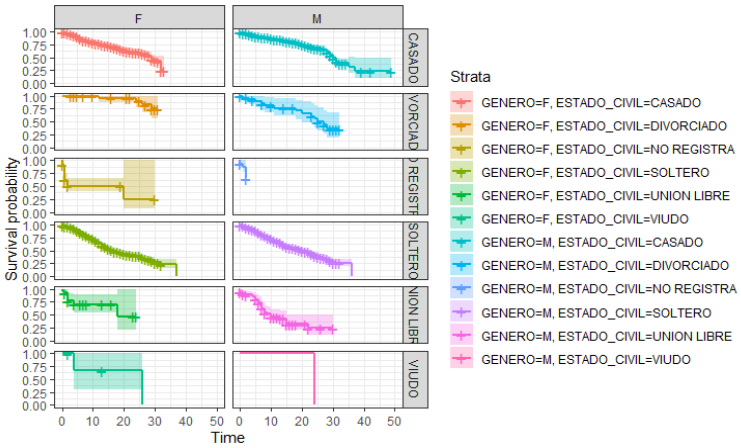

Figura 9: Curva de supervivencia de acuerdo al género y estado civil

Para el caso de estado civil - salario también se encuentra ausencia de datos en algunas categorías impidiendo una mejor estimación de la curva, pero con la data obtenida se tiene que el retiro más temprano puede suceder en las personas de unión libre con salarios altos

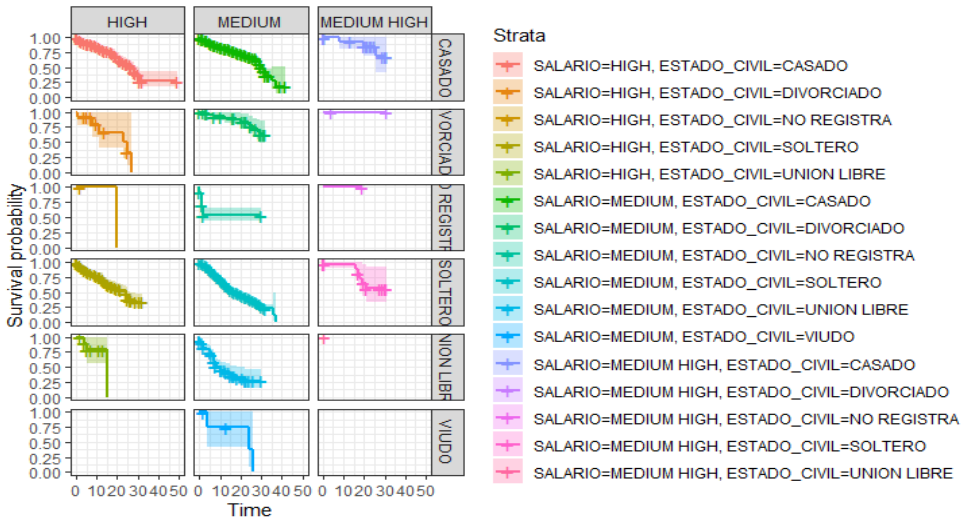

Figura 10: Curva de supervivencia de acuerdo al salario y estado civil

En el análisis para las variables salario y edad la mayor posibilidad de abandono ocurre en las personas de edad con categoría crecimiento, específicamente para salarios altos y medios, aunque la probabilidad de retiro más temprano ocurre en la primer categoría salarial mencionada.

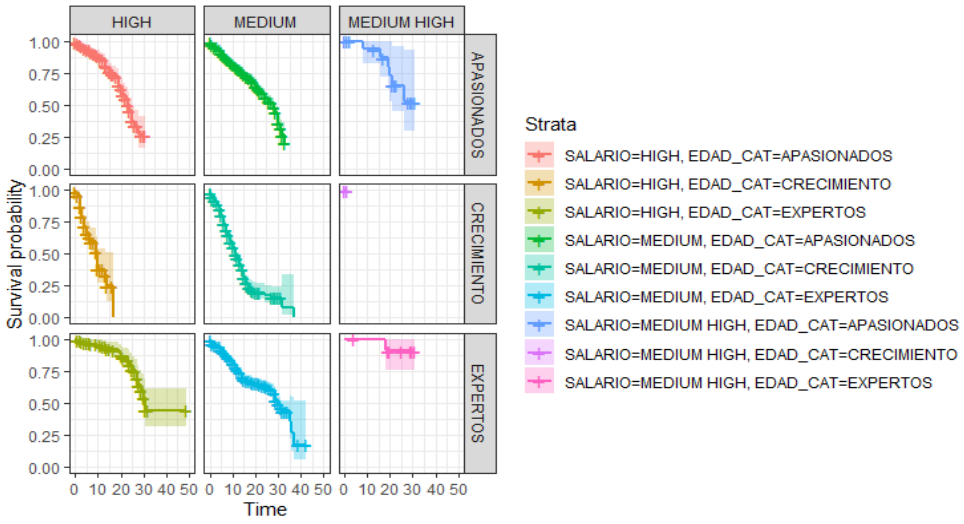

Figura 11: Curva de supervivencia de acuerdo al salario y edad

Replicando el ejercicio para genero y edad se puede identificar que la mayor probabilidad de deserción está en las personas de género masculino y en crecimiento, presentándose a los 30 años de permanencia, así mismo para los expertos antes de los 40 años de estancia la probabilidad permanece constante, posiblemente porque esta categoría de edad tiene personas que ya tienen estabilidad significativa con su póliza

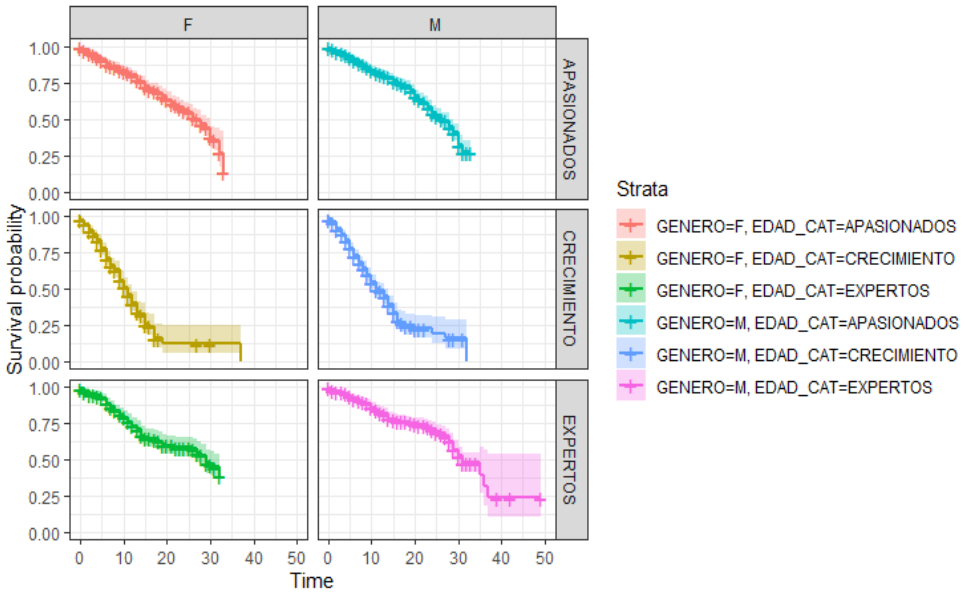

Figura 12: Curva de supervivencia de acuerdo al género y edad

Este análisis de combinación de factores puede realizarse con más de las categorías y variables presentes, pero para efectos del presente estudio se exponen únicamente estos. A pesar de que la estimación por el método de Kaplan-Meier arroja resultados interesantes y amplía el panorama del comportamiento que tiene la población respecto a la problemática a resolver, no tiene en cuenta el efecto que puede tener más de una variable, es por ello que se procede a realizar el análisis empleando el modelo de Cox de riesgos proporcionales.

Modelo de riesgos proporcionales de Cox

Para este método se empezará con un modelo que tenga en cuenta todas las variables que se consideran de interés, como lo son estado civil, edad categorizada, género, valor asegurado, prima mensual, salario y número de convenios, además se realizará con los datos de entreno para posteriormente obtener las probabilidades con los datos de test. Características de la personas como la región en donde habita, que era una de las variables a emplear, no se incluirá en el modelo ya que tiene una frecuencia alta en una sola categoría por lo que puede no tener un buen efecto en el desempeño de la regresión. La estimación de parámetros obtenida para este modelo general, teniendo en cuenta que la columna *estimate* corresponde al ratio de hazard, es decir $exp(b_i)$, es la presentada en la tabla 3

term	estimate	std.error	statistic	p.value
ESTADO_CIVILDIVORCIADO	0.94	0.24	-0.26	0.80
ESTADO_CIVILNO REGISTRA	14.72	0.18	15.19	0.00
ESTADO_CIVILSOLTERO	1.49	0.07	5.57	0.00
ESTADO_CIVILUNION LIBRE	1.98	0.15	4.42	0.00
ESTADO_CIVILVIUDO	7.27	0.71	2.78	0.01
EDAD_CATCRECIMIENTO	2.09	0.07	9.88	0.00
EDAD_CATEXPERTOS	0.95	0.08	-0.63	0.53
GENEROM	0.91	0.06	-1.49	0.14
VALOR_ASEGURADO	1.00	0.00	7.89	0.00
PRIMA_MENSUAL	1.00	0.00	-24.84	0.00
SALARIOMEDIUM	0.87	0.08	-1.79	0.07
SALARIOMEDIUM HIGH	0.49	0.39	-1.83	0.07
NUMERO_CONVENIOS	0.65	0.03	-13.25	0.00

Cuadro 3: Estimación de parámetros modelo 1

Se puede identificar a partir de esta salida que variables como el valor asegurado y la prima mensual no tienen efecto alguno en el retiro de una persona ya que su hazard ratio es de 1, por lo que en lugar de estas variables numéricas se empleará la categoría creada anteriormente para la clasificación de gamas del automóvil. También se puede observar que la variable género no es significativa, no obstante se seguirá teniendo en cuenta para el siguiente modelo que se realice. El segundo modelo tendrá la variable de gama de automóvil así como la de modelo categorizada como se describió anteriormente, la estimación de parámetros es la siguiente

term	estimate	std.error	statistic	p.value
ESTADO_CIVILDIVORCIADO	0.89	0.24	-0.48	0.63
ESTADO_CIVILNO REGISTRA	13.01	0.18	14.24	0.00
ESTADO_CIVILSOLTERO	1.54	0.07	6.00	0.00
ESTADO_CIVILUNION LIBRE	2.09	0.15	4.78	0.00
ESTADO_CIVILVIUDO	3.82	0.71	1.88	0.06
EDAD_CATCRECIMIENTO	1.94	0.07	8.93	0.00
EDAD_CATEXPERTOS	0.85	0.08	-2.10	0.04
GENEROM	0.90	0.06	-1.74	0.08
SALARIOMEDIUM	0.78	0.08	-3.11	0.00
SALARIOMEDIUM HIGH	0.36	0.39	-2.65	0.01
NUMERO_CONVENIOS	0.43	0.03	-24.50	0.00
GAMASBAJA	0.87	0.13	-1.08	0.28
GAMASMEDIA	0.89	0.14	-0.83	0.41
MODELO_CATINTERMEDIO	0.73	0.11	-2.93	0.00
MODELO_CATMODERNOS	1.53	0.06	6.60	0.00

Cuadro 4: Estimación de parámetros modelo 2

Para esta segunda salida se observa que la variable de gamas del automóvil asegurado no es siginificativa por lo que se removerá, contrario con la variable género que en esta ocasión teniendo en cuenta un nivel de significancia del 10 % es significativa por lo que se dejará por ahora. El tercer modelo teniendo en cuenta que se eliminará la variable **Gamas** arroja las siguientes estimaciones del hazard ratio

term	estimate	std.error	statistic	p.value
ESTADO_CIVILDIVORCIADO	0.89	0.24	-0.48	0.63
ESTADO_CIVILNO REGISTRA	13.14	0.18	14.34	0.00
ESTADO_CIVILSOLTERO	1.54	0.07	5.99	0.00
ESTADO_CIVILUNION LIBRE	2.09	0.15	4.77	0.00
ESTADO_CIVILVIUDO	3.79	0.71	1.87	0.06
GENEROM	0.90	0.06	-1.68	0.09
EDAD_CATCRECIMIENTO	1.94	0.07	8.92	0.00
EDAD_CATEXPERTOS	0.85	0.08	-2.12	0.03
SALARIOMEDIUM	0.78	0.08	-3.10	0.00
SALARIOMEDIUM HIGH	0.36	0.39	-2.67	0.01
NUMERO_CONVENIOS	0.42	0.03	-24.76	0.00
MODELO_CATINTERMEDIO	0.73	0.11	-2.99	0.00
MODELO_CATMODERNOS	1.56	0.06	7.13	0.00

Cuadro 5: Estimación de parámetros modelo 3

Para este modelo vemos que en general todas las variables son significativas, si bien algunas categorías en específico como la categoría de divorciado no lo es, en general con un α del 10 % todas las estimaciones lo son. Para identificar qué modelo tiene el mejor desempeño se emplearán algunos criterios de selección, el primero de ellos será el valor del criterio AIC que es una medida de la calidad relativa del modelo, por lo que dado un conjunto de modelos candidatos el favorito a seleccionarse es el que tiene el valor mínimo en el AIC. Como se presenta en el siguiente cuadro si bien el modelo 1 arrojó el menor AIC de los tres, el modelo 3 obtuvo un valor no tan lejano e incluso un menor valor respecto al modelo 2, además de tener todas las variables significativas, por lo que puede ser un candidato sólido a ser el modelo definitivo.

AIC Modelo 1	AIC Modelo 2	AIC Modelo 3
15774.77	16640.64	16637.77

Cuadro 6: Valor criterio AIC

Si nos remitimos a otras medidas de desempeño como lo son la concordancia o el test de rango logarítmico se obtiene lo siguiente:

```
### MODELO 1 ###
Concordance= 0.844  (se = 0.006 )
Likelihood ratio test= 2070  on 13 df,   p=<2e-16
Wald test            = 659.4  on 13 df,   p=<2e-16
Score (logrank) test = 2308  on 13 df,   p=<2e-16

### MODELO 2 ###
Concordance= 0.783  (se = 0.007 )
Likelihood ratio test= 1208  on 15 df,   p=<2e-16
Wald test            = 1138  on 15 df,   p=<2e-16
Score (logrank) test = 1273  on 15 df,   p=<2e-16

### MODELO 3 ###
Concordance= 0.783  (se = 0.007 )
Likelihood ratio test= 1207  on 13 df,   p=<2e-16
Wald test            = 1134  on 13 df,   p=<2e-16
Score (logrank) test = 1269  on 13 df,   p=<2e-16
```

La significancia global del modelo se tiene en base a la prueba de razón de verosimilitud, la prueba de Wald y el test de rango logarítmico. Estos tres métodos son asintóticamente equivalentes, es decir, para un tamaño de muestra N suficientemente grande darán resultados similares. El p valor para las tres pruebas generales es significativo para los tres modelos, indicando así que son regresiones en donde al menos uno de los parámetros es significativo. Estas pruebas evalúan la hipótesis nula de que todas las β son 0, en este caso para todos los modelos se rechaza y en general no sería un criterio de desempate.

Por otro lado la concordancia es un criterio que funciona también como medida de desempeño, su valor resultante muestra la capacidad para predecir de un par de observaciones cuál tendrá la ocurrencia del evento antes, en este caso el abandono, sin embargo no indica necesariamente cuánto antes o qué proporción de la varianza de los tiempos de los eventos se explica por el modelo. Para la comparación de los tres modelos se tiene que a pesar de que el valor del primero es el más alto, lo que sería ideal ya que es un criterio que mide una capacidad de predicción (entre más grande mejor), los otros dos no disminuyen mucho y adicionalmente siguen siendo valores significativos. Por ende, teniendo en cuenta estas medidas de desempeño y la significancia de la estimación de los parámetros el modelo escogido es el 3.

Con esta elección de modelo se procede a la interpretación de parámetros, que se puede obtener a partir del siguiente gráfico

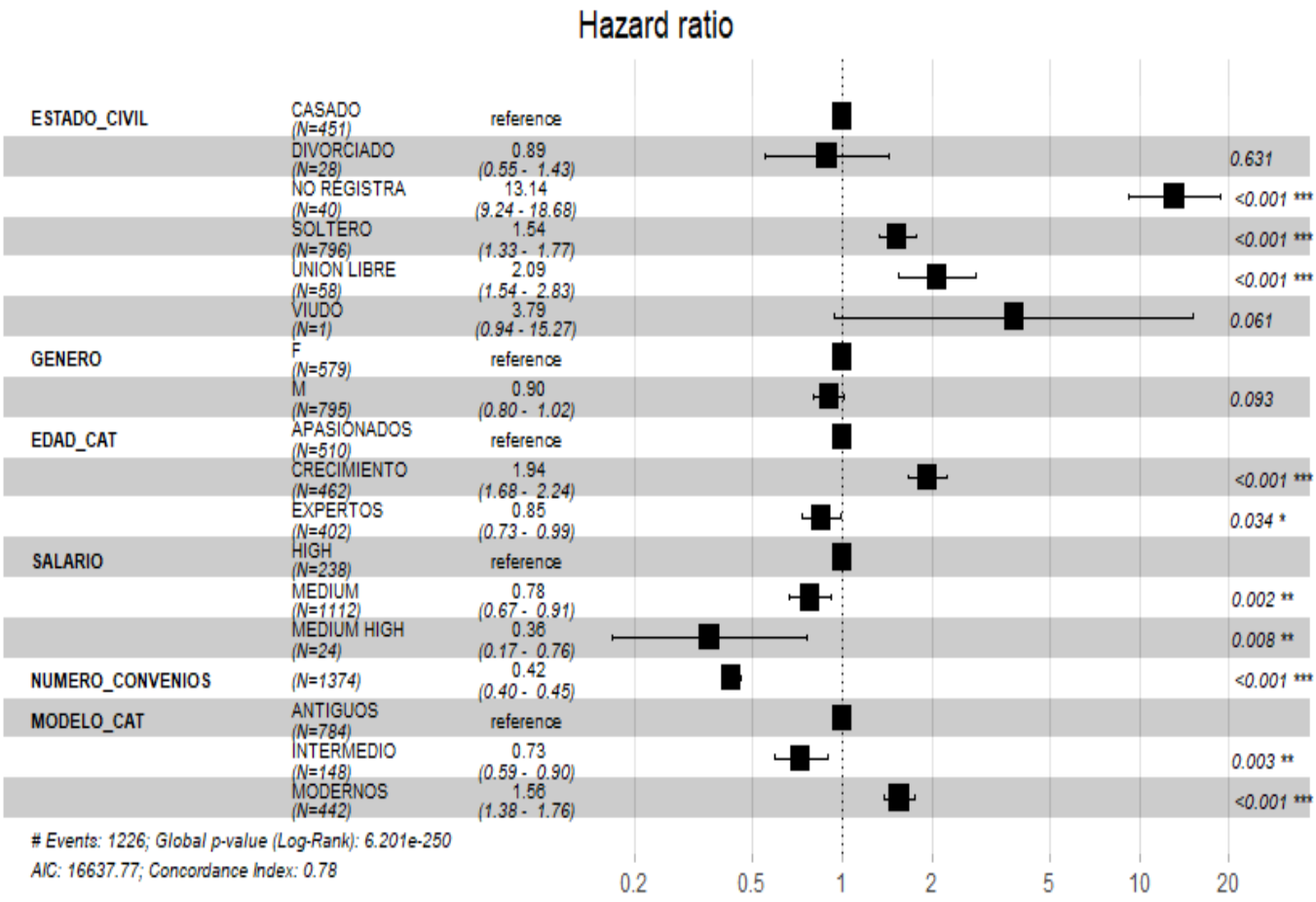

Figura 13: Hazard ratio modelo de Cox

Para la variable género se puede identificar que las categorías No registra, soltero y unión libre son factores que incrementan el riesgo de retiro, por ejemplo, para el caso de las personas solteras hay un 54 % más de chance de abandono respecto al caso base o referencia, es decir respecto a los casados. Por otro lado para el género masculino se reduce el riesgo un 10 % con respecto al género femenino. Así mismo las personas en categoría de edad de crecimiento incrementan el riesgo de abandono en un 94 %, contrario a los expertos que lo reducen en un 15 %, todo esto respecto a los del grupo apasionados. Para el salario tanto las categorías de medio y medio alto reducen el riesgo de abandono con respecto a las personas de salarios altos, en un 22 y un 64 % respectivamente. Para el caso de número de convenios, si se dejan el resto de covariables constantes, como se ha asumido en las interpretaciones anteriores, se tiene un hazard ratio de 0.42 indicando una relación débil entre esta característica y el abandono de la póliza, es decir, mayor número de convenios implican un menor chance de retiro, lo cual tiene sentido ya que si la persona tiene más de un producto con la compañía garantiza fidelización. Finalmente modelos modernos del vehículo implican un mayor riesgo de deserción respecto a los antiguos, probablemente por el alto costo de la prima a cancelar, caso contrario de los vehículos con modelo intermedio que reducen el riesgo.

Una vez calibrado el modelo se procederá a hacer las estimaciones correspondientes para los datos de test, obteniendo la salida de 7 y curva de supervivencia a nivel general

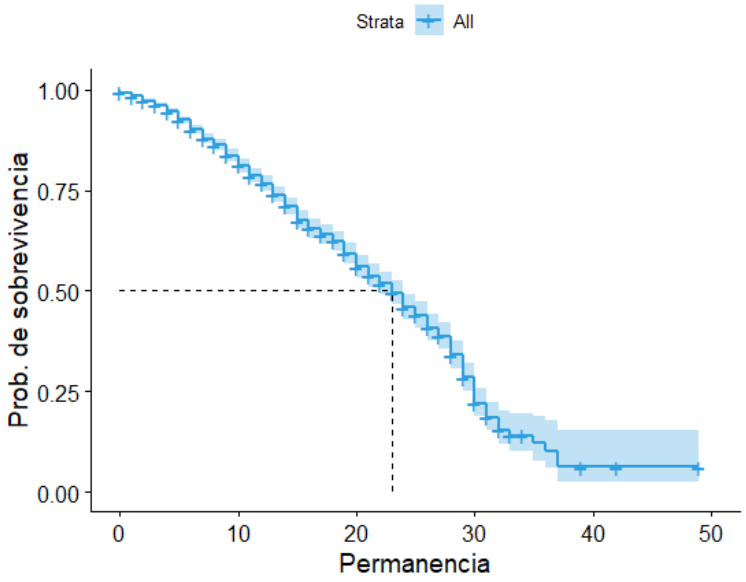

Figura 14: Curva de supervivencia a nivel general para datos de test

Como se puede observar en la curva 14 y la salida de la tabla 7 para este caso, a comparación de la curva inicial estimada por Kaplan - Meier que se encuentra en la figura 3, habrá un 50 % de probabilidad de retiro para el año número 23 de aseguramiento con la póliza, tal como lo muestra también la salida de abajo aplicando la función `survfit`. Igualmente, la curva tiene un comportamiento un poco diferente a 3 pero rasgos característicos como la constancia en la probabilidad de salida a partir de aproximadamente los 37 años.

```
Call: survfit(formula = fitcoxph3)

    n  events  median 0.95LCL 0.95UCL
 3207    1226      23      22      24
```

Cabe resaltar también que si se desea obtener una probabilidad de abandono en un tiempo específico así como el número de eventos que podría llegar a presentarse se puede recurrir a la tabla 7, la cual tiene los resultados de que sucedería en un rango de tiempo de hasta 49 años de permanencia con el seguro. También es importante recalcar que este último análisis o estimación se hizo a nivel general, pero con los mismos datos de test se pueden obtener probabilidades para grupos más específicos empleando exclusivamente el género, estado civil u otras variables como se hizo en el análisis por Kaplan - Meier.

time	n.risk	n.event	n.censor	surv	std.err	upper	lower
0.00	3207.00	43.00	37.00	0.99	0.00	1.00	0.99
1.00	3127.00	56.00	75.00	0.99	0.00	0.99	0.98
2.00	2996.00	74.00	175.00	0.97	0.00	0.98	0.97
3.00	2747.00	42.00	65.00	0.96	0.00	0.97	0.96
4.00	2640.00	61.00	69.00	0.95	0.00	0.95	0.94
5.00	2510.00	78.00	71.00	0.93	0.00	0.93	0.92
6.00	2361.00	85.00	101.00	0.90	0.01	0.91	0.89
7.00	2175.00	66.00	90.00	0.88	0.01	0.89	0.87
8.00	2019.00	43.00	68.00	0.86	0.01	0.88	0.85
9.00	1908.00	64.00	82.00	0.84	0.01	0.85	0.82
10.00	1762.00	56.00	69.00	0.81	0.01	0.83	0.80
11.00	1637.00	53.00	66.00	0.79	0.01	0.80	0.77
12.00	1518.00	34.00	58.00	0.77	0.01	0.79	0.75
13.00	1426.00	47.00	69.00	0.74	0.01	0.76	0.72
14.00	1310.00	43.00	52.00	0.71	0.02	0.73	0.69
15.00	1215.00	47.00	66.00	0.68	0.02	0.70	0.65
16.00	1102.00	24.00	51.00	0.65	0.02	0.68	0.63
17.00	1027.00	16.00	37.00	0.64	0.02	0.66	0.62
18.00	974.00	17.00	38.00	0.62	0.02	0.65	0.60
19.00	919.00	30.00	35.00	0.59	0.02	0.62	0.57
20.00	854.00	28.00	52.00	0.56	0.03	0.59	0.53
21.00	774.00	19.00	40.00	0.54	0.03	0.57	0.51
22.00	715.00	15.00	34.00	0.52	0.03	0.55	0.49
23.00	666.00	17.00	27.00	0.49	0.03	0.53	0.47
24.00	622.00	25.00	41.00	0.46	0.03	0.49	0.43
25.00	556.00	13.00	40.00	0.44	0.04	0.47	0.41
26.00	503.00	18.00	36.00	0.41	0.04	0.44	0.38
27.00	449.00	12.00	40.00	0.39	0.04	0.42	0.36
28.00	397.00	24.00	33.00	0.34	0.05	0.38	0.31
29.00	340.00	28.00	63.00	0.28	0.06	0.32	0.25
30.00	249.00	28.00	78.00	0.22	0.08	0.26	0.19
31.00	143.00	11.00	95.00	0.19	0.10	0.22	0.15
32.00	37.00	4.00	20.00	0.15	0.13	0.20	0.12
33.00	13.00	1.00	4.00	0.14	0.17	0.19	0.10
34.00	8.00	0.00	1.00	0.14	0.17	0.19	0.10
35.00	7.00	1.00	0.00	0.12	0.22	0.19	0.08
36.00	6.00	1.00	0.00	0.10	0.29	0.18	0.06
37.00	5.00	2.00	0.00	0.06	0.47	0.15	0.02
39.00	3.00	0.00	1.00	0.06	0.47	0.15	0.02
42.00	2.00	0.00	1.00	0.06	0.47	0.15	0.02
49.00	1.00	0.00	1.00	0.06	0.47	0.15	0.02

Cuadro 7: Estimación de probabilidades a nivel general para los datos de test

Conclusiones

Se encontró que para este caso en particular variables de la población activa y retirada como la edad, género, salario, estado civil, el modelo del vehículo asegurado e incluso el número de productos de seguro que tienen los usuarios impactan en el retiro que se pueda presentar en una póliza de automóvil voluntaria, independientemente si son factores de riesgo o de reducción contra la deserción. Otras características del auto como valor asegurado y prima mensual no tuvieron efecto en la estimación de la deserción por lo que se decidió crear una nueva variable categórica que tuviera en cuenta este tipo de atributos para realizar una clasificación por gamas y así tener una sola variable que podría resumir la posible influencia que tenía el tipo de vehículo, no obstante esta nueva variable creada a pesar de representar una estimación más coherente no fue significativa en el modelo final por lo que se decidió no tenerla en cuenta. Lo que sí se tuvo en cuenta fue el modelo del vehículo que de

igual forma se categorizó. En general se logró identificar que las personas solteras o en unión libre pueden incrementar el riesgo de retiro, también los usuarios con edad en el grupo crecimiento (inferior a los 40 años) y autos modernos, caso contrario de clientes de género masculino, con más de 60 años, afiliados a otros productos o pólizas y con salarios medios y medio altos.

En un futuro a corto plazo quedan algunos objetivos a cumplir como lo es adaptar estos resultados a un entorno más comercial y en un dashboard dinámico para su fácil lectura, de esta manera habrá una adecuada interpretación para todo el equipo de trabajo involucrado y se establecerá de mejor forma el camino para el diseño e implementación de estrategias preventivas. De igual manera, se espera a partir de este primera versión del modelo recibir feedback del efecto que produzca para darle mejora, por ejemplo se puede establecer una categoría más exactas de la variable gama del auto o conseguir más data para obtener resultados más concluyentes y completos. Finalmente si toda el desarrollo realizado brinda solución a la tasa de retiros para el ramo de automóviles se replicará la metodología para otros ramos como vida u hogar, que si bien tienen otros factores a considerar como la cobertura del producto, también tienen retiros considerables que se deben reducir.

Referencias

[1] Óscar Augusto Vargas Acosta. "Seguros voluntarios: la gran oportunidad para aumentar la protección de las familias". En: *Revista Fasecolda* (2018), págs. 80-85.

[2] Taane G Clark y col. "Survival Analysis Part I: Basic concepts and first analyses". En: *British Journal of Cancer* 89.3 (2003), 232–238.

[3] Taane G Clark y col. "Survival Analysis Part II: Multivariate data analysis – an introduction to concepts and methods". En: *British Journal of Cancer* 89.3 (2003), 431–436.

[4] D. R. Cox. "Regression Models and Life-Tables". En: *Journal of the Royal Statistical Society. Series B (Methodological)* 34.2 (1972), 187–220.

[5] E.L Kaplan y Paul Meier. "Nonparametric estimation from incomplete observations". En: *Journal of the American Statistical Association* 53.282 (1958), 457–481.

[6] Mauricio Henao Madrigal, Diego Restrepo Tobón y Henry Laniado. *Customer churn prediction in insurance industries : a multiproduct aproach*. 2020.

[7] Portafolio. *El 70 % de los autos en Colombia no está asegurado*. URL: https://www.portafolio.co/innovacion/el-70-de-los-autos-en-colombia-no-esta-asegurado-518996.

[8] Emily Zabor. *Survival Analysis in R*. URL: https://www.emilyzabor.com/tutorials/survival_analysis_in_r_tutorial.html.

Anexo

Conceptos básicos

De acuerdo a [2], definiendo la variable aleatoria T como la duración del cliente en la empresa; la función de supervivencia $S(t)$ y la función de riesgo $h(t)$ se definen de la siguiente manera:

- **Función de supervivencia:** Mide la probabilidad de retiro más allá de un tiempo t

$$S(t) = 1 - F(t) = P(T > t) \tag{1}$$

- **Función de riesgos (Hazard function):** Se define como la tasa de eventos en el momento t condicionada a la supervivencia hasta el momento t o posterior

$$h(t) = \frac{f(t)}{S(t)} \tag{2}$$

Para el análisis univariado se puede recurrir al estiamdor de Kaplan-Meier descrito por Edward Kaplan y Paul Meier y publicado conjuntamente en 1958 en el Journal of the American Statistical Association [5], el cuál es un método no paramétrico que nos permite estimar la función de supervivencia. Se define como

$$\hat{S(t)} = \prod_{j|t_j \leq t} \frac{n_j - d_j}{n_j} \tag{3}$$

Donde n_j representan los clientes en riesgo para el tiempo t_j y d_j es el número de clientes retirados al tiempo t_j. La curva de supervivencia de Kaplan-Meier que representa la probabilidad de supervivencia frente al tiempo, proporciona un resumen útil de los datos que se pueden utilizar para estimar medidas como la mediana del tiempo de supervivencia.

Así mismo, también se debe tener en cuenta el concepto de censura, como se mencionó anteriormente, el análisis de supervivencia se enfoca en la duración esperada de tiempo hasta que ocurra un evento de interés (por ej. muerte). Sin embargo de acuerdo a lo plasmado en [2] y [6], es posible que el evento no se observe para algunas personas dentro del período de tiempo del estudio, lo que produce observaciones censuradas que pueden derivarse de tres casos. Uno de ellos es si una persona no ha experimentado (todavía) el evento de interés dentro del período de tiempo del estudio, también si un individuo se pierde durante el seguimiento durante el período de estudio o si el individuo experimenta un evento diferente que hace imposible un seguimiento posterior. Para este caso se empleará la censura hacia la derecha que describe la situación en que una persona aún siga activa con su póliza en la compañía, es decir que no haya experimentado el evento de interés.

Modelo de riesgos proporcionales de Cox

Métodos como el de Kaplan-Meier describen generalmente la supervivencia de acuerdo a un solo factor o variable como se mencionó anteriormente, es decir es univariado, además funciona de manera adecuada únicamente cuando la variable es categórica, por lo que se debe recurrir a una metodología que tenga en cuenta también variables cuantitativas y que permita evaluar simultáneamente el efecto de varios factores de riesgo en el tiempo de supervivencia. Un método alternativo que cumpla dichos requisitos es el Modelo de riesgos proporcionales de Cox propuesto en [4] y ejemplificado en [3].

El modelo de Cox se expresa mediante la función de riesgo denotada por $h(t)$ y se puede estimar de la siguiente manera

$$h(t) = h_0(t) \times exp(b_1 x_1 + b_2 x_2 + ... + b_p x_p) \tag{4}$$

Donde

- t representa el tiempo de sobrevivencia
- $h(t)$ es la función de riesgo determinada por un conjunto de p covariables $(x_1, x_2, ..., x_p)$
- El término h_0 se denomina *baseline hazard*. Corresponde al valor de riesgo o hazard si todos los x_i son iguales a cero (la cantidad exp(0) es igual a 1).

Las cantidades $\exp(b_i)$ se denominan razones de riesgo o *Hazard Ratio* (HR). Un valor de b_i mayor que cero, o equivalentemente una razón de riesgo mayor que uno, indica que a medida que aumenta el valor de la covariable i-ésima, aumenta el riesgo del evento. Es decir, una razón de riesgo por encima de 1 indica una covariable que está asociada positivamente con la probabilidad del evento y, por lo tanto, asociada negativamente con la duración de la supervivencia.

Para resumir se tiene

- HR = 1: Sin efecto
- HR $<$ 1: Reducción de riesgo de retiro
- HR $>$ 1: Aumento de riesgo de retiro

Ahora, como se muestra en [8] si se consideran dos usuarios k y k^* que difieren en sus valores de x, la función de hazard correspondiente se puede escribir simplemente de la siguiente manera para ambos tipos de clientes

$$h_k(t) = h_0(t) \times exp(\sum_{i=1}^{n} \beta_i) \tag{5}$$

$$h_k^*(t) = h_0(t) \times exp(\sum_{i=1}^{n} \beta_i^*) \tag{6}$$

Por lo tanto el hazard ratio para dichos clientes se denotaría como

$$\frac{h_k(t)}{h_k^*(t)} = \frac{h_0(t) \times exp(\sum_{i=1}^{n} \beta_i)}{h_0(t) \times exp(\sum_{i=1}^{n} \beta_i^*)} = \frac{exp(\sum_{i=1}^{n} \beta_i)}{exp(\sum_{i=1}^{n} \beta_i^*)} \tag{7}$$

Como se puede identificar dicha razón o ratio no depende del tiempo por tanto el modelo de Cox es un modelo de riesgos proporcionales, es decir, el riesgo del evento en cualquier grupo es un múltiplo constante del riesgo en cualquier otro. Esta suposición implica que las curvas de riesgo para los grupos deben ser proporcionales y no pueden cruzarse.