VNIVERSITAT DE VALÈNCIA

unión alcoyana seguros

Detección del riesgo de fuga de clientes de una entidad aseguradora mediante algoritmos de machine learning

Septiembre 2018

José A. Álvarez-Jareño

José M. Pavía

Patricia García-Torres

Jorge Segura-Gisbert

Índice

Índice de tablas

Índice de gráficos

1. Introducción

Comparadores, agregadores, asistentes personales, sensores, etc. Los asegurados no habían tenido tanta información sobre sus pólizas, la compañía que los asegura y la competencia. Y todo esto gratis. La información que antes tenía un alto coste de oportunidad, se ha visto reducido a nada. El comportamiento del consumidor en general ha cambiado en los últimos años debido fundamentalmente a la tecnología, o como consecuencia de ella. Los clientes de cualquier empresa se informan sobre los productos que consumen a través de internet y en algunos casos puede llegar a saber más que los dependientes de las tiendas físicas donde los compran. Los clientes informados tienen un poder que nunca antes habían ostentado.

Fue en 2001 cuando George A. Akerlof, A Michael Spence y Joseph E. Stiglitz fueron galardonados con el Premio Nobel de Economía por sus análisis de los mercados con información asimétrica. Uno de esos mercados era y sigue siendo el mercado asegurador. Los asegurados conocen el riesgo que van a asegurar mucho mejor que la compañía aseguradora. Si se une que además tienen toda la información al alcance de su mano y gratis, su situación es privilegiada frente a las compañías aseguradoras.

Las empresas han tenido que efectuar un cambio de paradigma, pasando el foco del "producto" al "cliente". En el mercado asegurador, el centro deja de ser la póliza de seguro para ser el asegurado. Disponer de un buen producto (póliza de seguro) era sinónimo de éxito, sin embargo, hoy ya no es suficiente. Las compañías aseguradoras han empezado a aplicar sus conocimientos estadísticos y actuariales no sólo para determinar la prima o las reservas, sino para conocer a sus clientes y predecir su comportamiento. La aseguradora dispone de toda una serie de información sobre los asegurados y ha empezado a explotarla. El objetivo es cerrar la brecha de información entre asegurado y asegurador y conseguir una mejor gestión del negocio.

En este contexto, hay dos afirmaciones que deberían llamar la atención de los responsables de cualquier del negocio. La primera de Juran (2010) que expone la Ley de Pareto en el ámbito empresarial. El 20% de los clientes más rentables aportan el 80% del beneficio de la empresa; al mismo tiempo, el quintil menos rentable no solo no aporta ningún beneficio, sino que incluso destruye valor.

La segunda es de Torkzadeh et al. (2006) que afirman que el coste de conseguir un nuevo cliente puede ser hasta 12 veces el coste de retenerlo. Si se necesita hacer un esfuerzo tan grande para adquirir nuevos clientes, sería más oportuno retener a los clientes rentables que ya estén en la empresa.

Además, diferentes estudios muestran el valor económico de los clientes que permanecen en una compañía por un período largo de tiempo. Dawes y Swailes (1999) concluyen que el coste de obtener un nuevo cliente es muy superior al coste de retenerlo. Reichheld (1996) afirma que los clientes antiguos compran más que los clientes nuevos, y en la misma línea, Reichheld y Kenny (1990) llegan a la conclusión que para las entidades financieras, cuanto más duradera sea la relación con el cliente, mayor será su valor.

Kumar y Garg (2013) son de la opinión que la estrategia más importante de retención de clientes es identificar por adelantado aquellos que se quieran marchar, y una vez identificados, las

compañías deberían aplicar políticas de retención para disuadir de su marcha a estos clientes. Cabría añadir, más aún, si esos clientes corresponden al 20% que genera el 80% del beneficio.

El objetivo del presente trabajo será detectar y predecir el riesgo de fuga de la forma más precisa, y de esta forma centrar el esfuerzo en la labor de retención (Ali y Aritürk, 2014).

Berry y Linoff (2004) exponen que hay tres tipos de clientes que abandonan una empresa: voluntarios, involuntarios e inesperados. El abandono involuntario sería la anulación de la póliza por parte de la compañía por alguna causa (impago de la prima, excesiva siniestralidad, intento de fraude, etc.). El abandono inesperado sería cuando el cliente abandona la compañía sin intención, estos serían los casos de cierre del negocio, cambio de ciudad de trabajo, venta de la propiedad, etc. El cliente no tenía intención de abandonar la compañía pero se ha producido un cambio y se ve obligado a rescindir la póliza.

El abandono voluntario es el que realmente interesa a la compañía aseguradora, y es aquel que ocurre cuando un cliente deja una empresa deliberadamente para marcharse a otra por un producto o servicio mejor o más barato. Estos son los asegurados que cambian de compañía aseguradora por una póliza mejor o por una prima más reducida.

Se dispone de muchos ejemplos y en diferentes sectores para que una vez identificada la variable se utilice la minería de datos y el aprendizaje automático (machine learning) para la detección y predicción de los clientes en riesgo de fuga. Existe una amplia y extensa literatura científica que muestra cómo aplicar una gran variedad de algoritmos a este problema. En este sentido, el trabajo consistirá en identificar que algoritmo, o conjunto de algoritmos, permiten predecir con mayor exactitud el abandono de los clientes de la compañía aseguradora. Rodríguez (2018) nos trasmite que el aprendizaje automático y los algoritmos de la IA pueden hacer buenas predicciones, sí, pero igualmente importante es hacer las preguntas correctas.

En este caso, la pregunta correcta no sería quién se va a marchar de la compañía, sino por qué se va a marchar. Las metodologías de aprendizaje automático permiten identificar a las personas que tienen una alta probabilidad de abandonar la empresa, es decir, ¿quién? Sin embargo, si los algoritmos utilizados son complejos, los que normalmente obtienen mejores resultados, no existe la posibilidad de poder interpretar el modelo y explicar de forma lógica y razonada el comportamiento de los clientes, es decir, ¿por qué?

La finalidad de conocer el porqué del abandono es poder dotar al departamento comercial y/o de marketing de una nueva herramienta que permite realizar acciones individualizadas en función de las características de los diferentes asegurados de la cartera. El conocimiento que se puede extraer de los datos disponibles de los clientes de una cartera es muy amplio. La idea sería poder predecir el comportamiento del asegurado en base a todos esos datos.

Además hay un segundo motivo para hacer este análisis, ya que recientemente ha entrado en vigor el Reglamento Europeo de Protección de Datos (GDPR). Entre las principales novedades que ha introducido este reglamento es el "derecho de explicación" de los modelos a petición del asegurado. Es decir, sólo se podrán utilizar en producción modelos de aprendizaje automático que puedan ser explicados a los clientes. En este sentido, si el modelo utilizado es un árbol de decisión o una regresión lineal no habrá ningún problema porque son fáciles de

interpretar y de explicar. Sin embargo, si los modelos utilizados son aquellos conocidos como "caja negra", la interpretación no estadística del modelo es muy complicada.

Dado que la norma se conocía desde hace tiempo, y las necesidades del departamento de marketing no son nuevas, se están desarrollado una serie de herramientas estadísticas que permitan interpretar y explicar los modelos más complejos. Hacer uso de esas herramientas es el segundo objetivo del presente trabajo.

Resumiendo, el objetivo del presente trabajo es doble:

- Detectar y predecir que clientes están pensando en marcharse.
- Interpretar y explicar el modelo seleccionado para el riesgo de fuga.

Una vez fijados los objetivos, el trabajo se desarrollará según el esquema que se explica a continuación. Después de esta introducción donde se ha expuesto el problema a tratar, se desarrollará en el segundo apartado el análisis de las principales variables de la base datos y las metodologías para la selección de variables. El tercer apartado se dedicará a la metodología que se aplicará en la resolución del problema con especial énfasis en las posibles soluciones a los conjuntos de datos desequilibrados y los algoritmos de clasificación que se van a utilizar. La presentación de los resultados obtenidos se realizará en el apartado cuarto, realizando una valoración en base a las medidas de evaluación propuestas. En el quinto apartado se efectuará una interpretación de uno de los modelos considerados "caja negra". Poder interpretar y explicar los resultados de los modelos de ensamble será fundamentales para el departamento de marketing y para cumplir con el nuevo Reglamento de Protección de Datos Europeo. Se finalizará con las conclusiones alcanzadas en el apartado sexto.

2. Análisis y Selección de Variables

Tal como afirman Caballero y Martín (2015) "el fin último del Big Data no es acumular datos, sino extraer información útil a partir de los datos". Obtener información de un conjunto de datos precisa de una serie de etapas que se seguirán unas a otras como en un proceso de destilación. En este proceso, es de vital importancia la primera de las fases, que es la limpieza, acondicionamiento y análisis de la base de datos. Además, esta etapa es la que mayor tiempo consume, llegando a ser en algunos proyectos más del 80% del total (Baesens et al., 2015). Es imprescindible que se conozcan las variables y los problemas que presentan para que cuando se llegue a la fase de modelización se obtenga un resultado óptimo.

Silver (2014) afirma que "lo que distingue a la ciencia, y lo que hace que un pronóstico sea científico, es que tenga relación con el mundo objetivo. Cuando un pronóstico fracasa es porque sólo nos preocupamos por el método, la máxima o el modelo." La metodología por sí sola no obtendrá ningún resultado valioso, por lo que será necesario conocer los datos y el problema que se pretende resolver.

Después de un correcto análisis de las variables se procederá a seleccionar aquellas variables que permiten efectuar una modelización óptima. Aunque actualmente se dispone de una gran capacidad de cálculo, cuantas más variables se procesen mayor será el tiempo de computación. Si los modelos tienen que utilizarse con cierta asiduidad será mejor reducir el tiempo de computación para que sea más eficiente su utilización.

2.1 Los datos y la generación de nuevas variables

La Entidad Aseguradora ha facilitado para la realización del estudio una base de datos de las pólizas de automóviles correspondiente al período 2012-2016. La base de datos es un fichero de texto plano (.csv), compuesta por 173 variables y 473.356 instancias. Este formato permite trabajar con los datos en diferentes programas de análisis estadístico sin necesidad de realizar modificaciones. En particular, se persigue que los datos estén en los formatos apropiados y con las especificaciones correctas.

El período considerado es amplio y permite disponer de mucha información correspondiente a un determinado período consecutivo de tiempo. No obstante, es probable que se hayan producido cambios en los procesos de almacenamiento de la información y cambios en los hábitos de consumo de los asegurados. Por esta razón, se seleccionan los datos temporalmente para el entrenamiento de los modelos y posterior predicción. La base de datos confeccionada contiene información sociodemográfica, históricos de siniestralidad, características de los productos contratados (modalidad, forma de pago, sistemas Bonus-Malus aplicados, antigüedad del cliente, etc) así como una visión general de los datos: Frecuencia de cero en cada variable, los datos ausentes, el tipo de datos o el número de valores únicos para cada variable.

La primera labor que se efectuó fue la limpieza, acondicionamiento y análisis de la base de datos. Esta fase es de suma importancia para que los análisis posteriores sean lo más veraces posibles. Si los datos presentan errores o están mal tratados, se puede incurrir en el problema conocido como GIGO ("garbage in, garbage out") tal como expone Silver (2014), ya que no se puede esperar que salga un buen modelo si los datos de partida son malos.

A continuación, se procedió a extraer información disponible en los datos mediante transformación o generación de nuevas variables. Se realizaron transformaciones del formato de las fechas y se pasaron a factores las variables cualitativas.

La cuestión primordial del análisis que se va a realizar es definir que se considera "fuga" de un cliente. La base de datos inicialmente identifica todas las pólizas que se han anulado en el período de análisis, y el motivo de su anulación. Se procedió a clasificar los motivos en no renovación voluntaria ("fuga" del cliente o abandono) y no renovación involuntaria (cese de la actividad, alta siniestralidad, etc.).

Las variables más importantes que se generaron fueron las siguientes:

- Abandono: no renovación voluntaria. Esta será la variable objetivo en el problema de clasificación que se pretende resolver.
- Edad del asegurado: fecha de efecto de la póliza menos fecha de nacimiento. Se ha depurado la variable debido a que aparecen asegurados menores de 18 años, y se considera que la variable para esa instancia no está bien informada.
- Años carné de conducir: fecha de efecto de la póliza menos fecha de obtención del permiso de conducir. Se ha depurado la variable porque aparecen valores negativos de la variable, debido a problemas de fechas en la introducción de la fecha de obtención del permiso de conducción.
- Total siniestros de la anualidad de seguro: número de expedientes de siniestros aperturados en esa anualidad. Al considerarse una variable relevante, el cómputo de esta variable se ha efectuado por dos procedimientos diferentes.
- Total importe de la siniestralidad de la anualidad de seguro: suma de los importes de los expedientes aperturados. Si el siniestro estaba cerrado se tomaban los pagos, y si permanecía abierto, se tomaba la reserva.
- Día de la semana de ocurrencia del siniestro.
- Día de la semana de la apertura del siniestro.
- Día de la semana de cierre del siniestro.
- Diferencia en días desde la ocurrencia del siniestro hasta su comunicación a la compañía aseguradora.
- Diferencia en días desde la apertura del siniestro hasta su cierre.
- Diferencia en días desde la ocurrencia del siniestro hasta el cierre

2.2 Análisis de las principales variables

En este apartado se procede a exponer los resultados obtenidos del análisis de algunas de las principales variables de la base de datos. El estudio analiza todas las variables, sin embargo, sólo se expondrán aquellas que son más representativas para el trabajo que se está realizando. La principal variable objeto de estudio es el abandono voluntario de los clientes de la compañía.

Variable Abandono (variable objetivo)

Se inicia el análisis descriptivo por la variable objetivo. El porcentaje de abandono voluntario del total de pólizas analizadas es del 14,56%, frente al 4,44% de anulaciones por parte de la compañía aseguradora. Es decir, la variable dependiente está claramente desequilibrada.

La variable Abandono presenta un comportamiento claramente diferenciado por años, siendo los primeros años, los que muestran un mayor porcentaje de fuga de clientes. El período de tiempo analizado coincide con una crisis económica general, que fue más intensa en los años 2012 y 2013. Ante esta situación, los asegurados prestan más atención al precio y las compañías aseguradoras incrementaron de forma considerable su competencia.

Tabla 1. Distribución del Abandono Voluntario por Años Naturales

	Renovación	Abandono	Total	% Abandono
2012	76.292	17.391	93.683	18,56%
2013	73.358	15.694	89.052	17,62%
2014	75.885	13.289	89.174	14,90%
2015	85.768	10.173	95.941	10,60%
2016	93.135	12.371	105.506	11,73%

Fuente: Elaboración Propia. RStudio.

Una de las primeras cuestiones que se plantea es si existe relación entre las primas que se pagan y la fuga de clientes. Para determinar si existen evidencias de esta afirmación, se ha procedido a realizar un análisis de la varianza para ver si existe relación entre la prima del año en curso que fija la aseguradora y la decisión de abandonarla por parte del cliente. La hipótesis que se ha contrastado es la siguiente:

$$\begin{cases} H_0 : \mu_{Abandono} = \mu_{NoAbandono} \\ H_1 : \mu_{Abandono} \neq \mu_{NoAbandono} \end{cases}$$

Y los resultados se muestran en la siguiente Tabla.

Tabla 2. Análisis de la Varianza Prima/Abandono

```
                Df     Sum Sq  Mean Sq  F value    Pr(>F)
Abandono         1  7.660e+05   765977    27.06  1.97e-07 ***
Residuals   222860  6.308e+09    28306
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

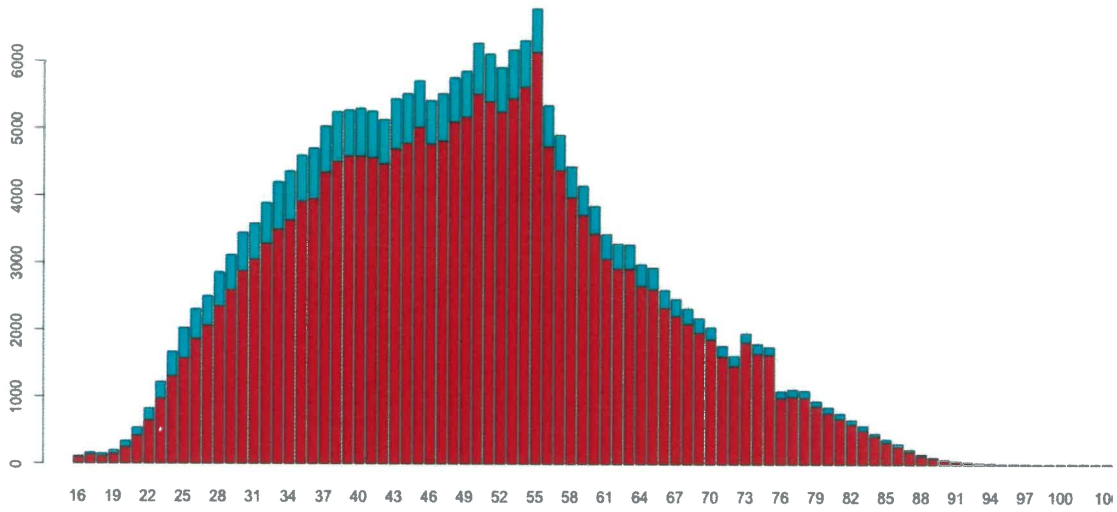
Fuente: Elaboración Propia. RStudio.

Si la probabilidad del contraste F es inferior al nivel de significación (normalmente 5%) se rechaza la hipótesis nula que, en este caso, es que las primas medias de los asegurados que abandona la compañía son diferentes de los asegurados que permanecen. La causa de la fuga de clientes podría deberse, si no todo en parte, a la fijación de unas primas más elevadas.

Variable edad del conductor

En el seguro de automóviles una de las principales variables es la edad del conductor. Al efectuar una representación gráfica, inmediatamente se observa que el número de asegurados se incrementa paulatinamente hasta los 55 años, y luego decrece.

Gráfico 1. Distribución del Abandono por Edades



Fuente: Elaboración Propia. RStudio.

Lo mismo le ocurre a la tasa de abandono. A la edad de 18 años la tasa de abandono es superior al 25%, e irá disminuyendo hasta quedar por debajo del 10% para los mayores de 67 años.

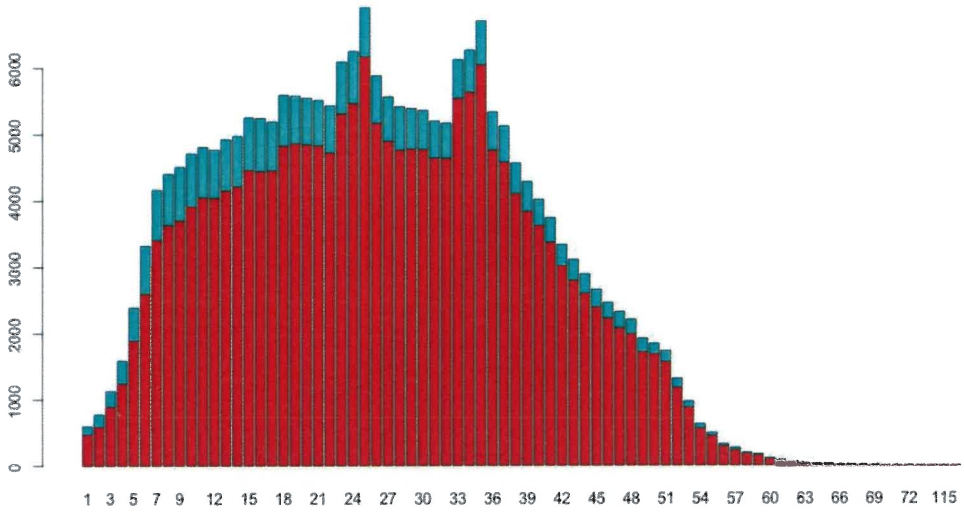
Gráfico 2. Tasas de Abandono por Edades

Fuente: Elaboración Propia. Microsoft Excel.

Variable antigüedad del carné de conducir

La distribución en función de la antigüedad en el carné de conducir sigue una distribución muy parecida a la de la edad del asegurado.

Gráfico 3. Distribución del Abandono por Antigüedad del Carné



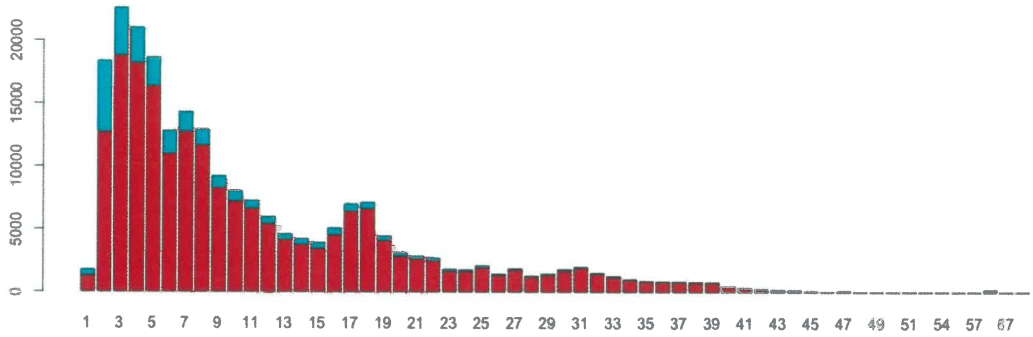
Fuente: Elaboración Propia. RStudio.

Variable antigüedad del cliente

La variable antigüedad del cliente mide el grado de fidelidad de los asegurados a la compañía aseguradora. La vida media de una póliza en la Entidad Aseguradora es de 10,87 años, cifra muy elevada, si se compara con la media del sector asegurador.

La moda de la variable es 2 años, valor en el que coincidirá la mayor tasa de abandono (25,85%). A partir del segundo año, el número de asegurados descenderá paulatinamente. De igual forma la tasa de abandono, pero con una curiosidad, ya que se produce un repunte en la tasa entre los 13 y los 16 años de antigüedad. Esta variable debe estar relacionada con la vida útil de los vehículos asegurados, motivo por el cual se produzca este máximo local.

Gráfico 4. Distribución del Abandono por Antigüedad del Cliente



Fuente: Elaboración Propia. RStudio.

Gráfico 5. Tasas de Abandono por Antigüedad del Cliente

Fuente: Elaboración Propia. Microsoft Excel.

Variable siniestralidad

Las variables fundamentales en el análisis de datos del sector asegurador son la frecuencia y la intensidad de los siniestros acaecidos. En cuanto a la frecuencia, la distribución de los siniestros es la siguiente:

Tabla 3. Distribución del Número de Siniestros por año

Núm. de Siniestros	2013	2014	2015	Totales
0	50.636	51.798	56.660	159.094
1	12.956	14.243	16.197	43.396
2	4.052	4.465	5.251	13.768
3	1.284	1.436	1.602	4.322
4	463	415	529	1.407
5	174	148	204	526
6	79	61	78	218
7	28	23	22	73
8	13	10	14	37
9	5	7	2	14
10	2	3	0	5
11	0	2	0	2
Total	**69.692**	**72.611**	**80.559**	**222.862**

Fuente: Elaboración Propia. RStudio.

El coste medio de los siniestros de los asegurados que abandonan la compañía es superior al coste medio del resto. Para el período medio de los tres años se obtiene 608,16 € por siniestro para los asegurados que se marchan y 465,95 € para los que permanecen, siendo la diferencia estadísticamente significativa. No obstante, si realizamos el mismo análisis sobre las tres anualidades, no existen diferencias significativas.

Tabla 4. ANOVA coste medio de los siniestros por anualidad

	Df	Sum Sq	Mean Sq	F value	Pr(>F)
Abandono	1	1.196e+08	119628673	86.015	<2e-16 ***
Año	1	1.125e+06	1124993	0.809	0.368
Residuals	222859	3.100e+11	1390792		

Fuente: Elaboración Propia. RStudio.

Las medias de siniestralidad estimada para las tres anualidades son 509,90 € para el año 2013, 481,75 € para el 2014 y 472,62 € para el último año. El coste medio de los siniestros se fue reduciendo en estos años.

Variable Valor Cliente

La aseguradora en base al conocimiento y la experiencia que tiene de sus asegurados ha efectuado una clasificación de los mismos. Esta clasificación debería ser un buen predictor para otras variables como prima de la anualidad en curso, o incluso abandono.

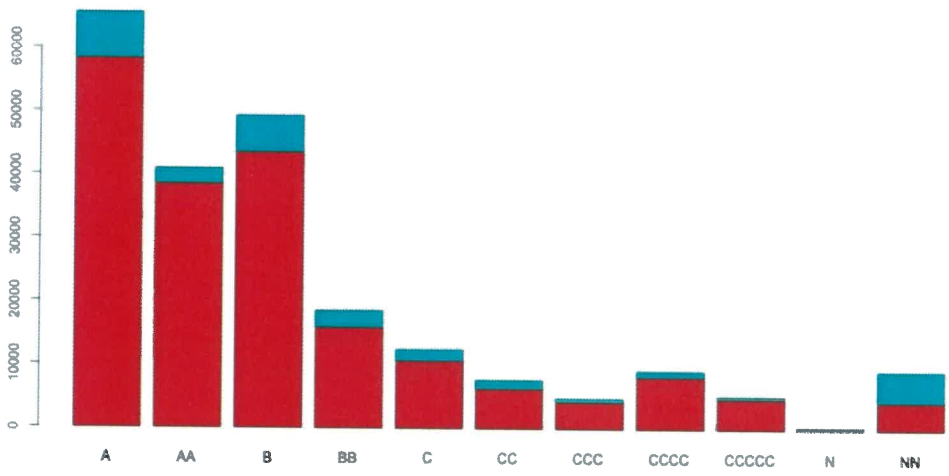
Tabla 5. Distribución Abandono / Score Valor Cliente

	No Abandono	Abandono	Total	% Abandono
A	58.511	7.252	65.763	11,03%
AA	38.645	2.410	41.055	5,87%
B	43.658	5.763	49.421	11,66%
BB	15.906	2.656	18.562	14,31%
C	10.678	1.760	12.438	14,15%
CC	6.260	1.363	7.623	17,88%
CCC	4.229	545	4.774	11,42%
CCCC	8.214	931	9.145	10,18%
CCCCC	4.780	411	5.191	7,92%
N	159	219	378	57,94%
NN	4.386	4.874	9.260	52,63%
	195.426	**28.184**	**223.610**	

Fuente: Elaboración Propia. RStudio.

Destaca la clase CC que tiene una tasa de abandono superior al 17%. También las clases N y NN tienen una tasa muy elevada, pero el motivo es que se asignan estas clases cuando no había valoración para los clientes en los años 2012 y 2013.

Gráfico 6. Distribución del abandono por clases de Valor Cliente



Fuente: Elaboración Propia. Gráfico generado con RStudio.

El porcentaje de abandono no es el mismo en las diferentes clases analizadas. Efectuado el contraste de asociación, se rechaza la hipótesis nula, por lo que no existe relación entre ambas variables.

Se ha procedido a realizar un análisis de la varianza tomando como variable dependiente la prima de la anualidad en curso, y como variables independientes Abandono y la clasificación

de Valor Cliente. Los resultados que se muestran en la siguiente Tabla indican que existen evidencias para rechazar la hipótesis nula.

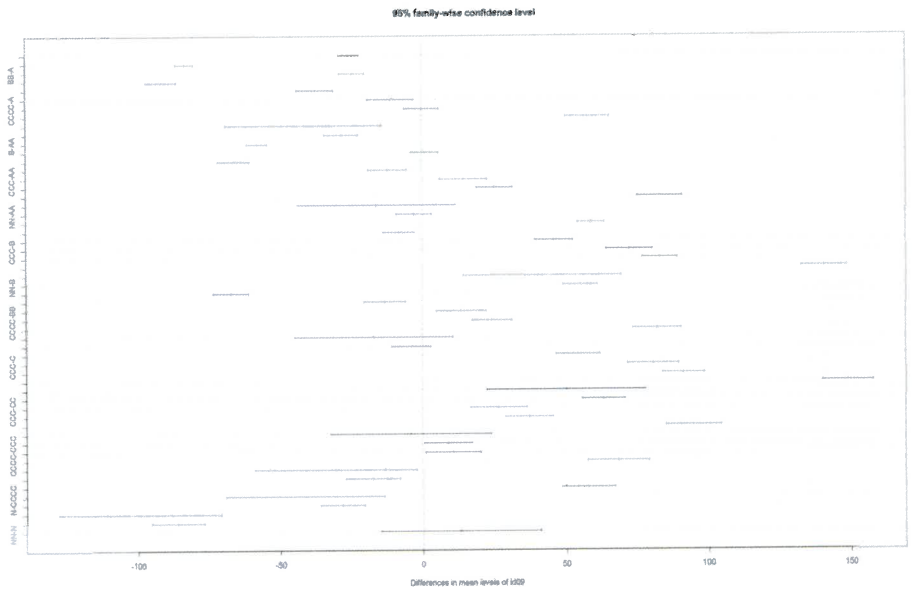
Tabla 6. ANOVA coste medio de los siniestros por anualidad

	Df	Sum Sq	Mean Sq	F value	Pr(>F)
Abandono	1	7.660e+05	765977	28.39	9.92e-08 ***
Valor Cliente	10	2.960e+08	29603292	1097.30	< 2e-16 ***
Residuals	222850	6.012e+09	26978		

Fuente: Elaboración Propia. RStudio.

En este caso al analizar dos factores se han calculado los contrastes por pares mediante el ajuste de los p-valores por el método de Tukey, rechazándose la hipótesis nula para la mayoría de los pares. Los resultados se pueden observar a través del siguiente gráfico.

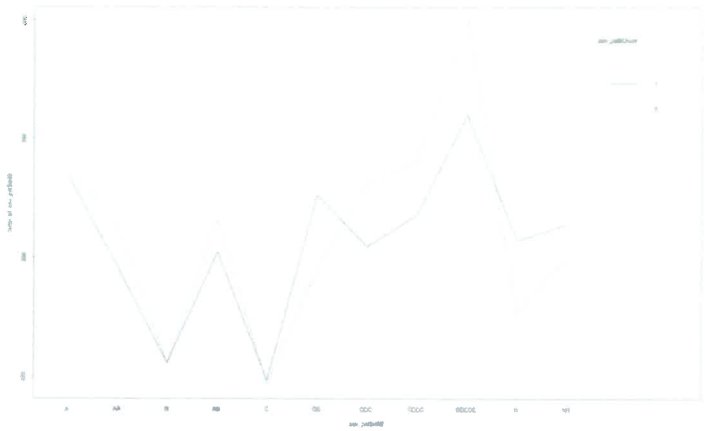
Gráfico 7. Intervalos de confianza para la diferencia de medias (Tukey)



Fuente: Elaboración Propia. Gráfico generado con RStudio.

Si en el análisis de la varianza se incluye la interacción entre ambos factores, éste sale significativo, es decir, existen evidencias de interacción. El gráfico de interacción de ambos factores con la prima se muestra a continuación.

Gráfico 8. Interacción entre los factores Abandono y Valor Cliente



Fuente: Elaboración Propia. Gráfico generado con RStudio.

Cuando las líneas se cruzan en este tipo de gráficos indicaría que hay interacción entre los factores. Se observa que las líneas se cruzan en varias ocasiones. La línea discontinua son los asegurados que no abandonan, y la línea continua los que se marchan. En algunos casos las medias son iguales, pero en otros la diferencia entre ellas es muy elevada. Por ejemplo, para el grupo CCCCC la prima media para los que se quedan es mucho más elevada que para los que se marchan. Y a la inversa, ocurre en el grupo N.

Variable Score Global

La compañía utiliza una segunda valoración del cliente a través de un proveedor externo.

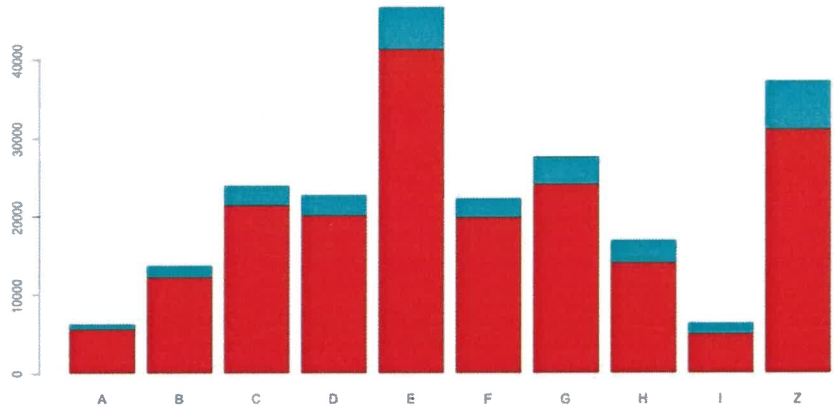
Tabla 7. Distribución Abandono / Score Global

	No Abandono	Abandono	Total	% Abandono
A	5.577	581	6.158	9,43%
B	12.290	1.390	13.680	10,16%
C	21.523	2.428	23.951	10,14%
D	20.209	2.526	22.735	11,11%
E	41.483	5.317	46.800	11,36%
F	19.969	2.334	22.303	10,46%
G	24.194	3.417	27.611	12,38%
H	14.093	2.803	16.896	16,59%
I	4.965	1.307	6.272	20,84%
Z	31.120	6.081	37.201	16,35%
	195.423	**28.184**	**223.607**	

Fuente: Elaboración Propia. RStudio.

La clase I es la que muestra un mayor porcentaje de abandonos, aunque es una de clases más minoritarias de todas. Solo la clase A tiene menos instancias, aunque la diferencia es mínima. La clase Z se utiliza para los asegurados que no han podido ser clasificados según esta valoración.

Gráfico 9. Distribución del abandono por clases Score Global



Fuente: Elaboración Propia. Gráfico generado con RStudio.

Los resultados obtenidos en el contraste de asociación son muy similares a la variable valor cliente de la compañía aseguradora.

Variable Forma de Pago

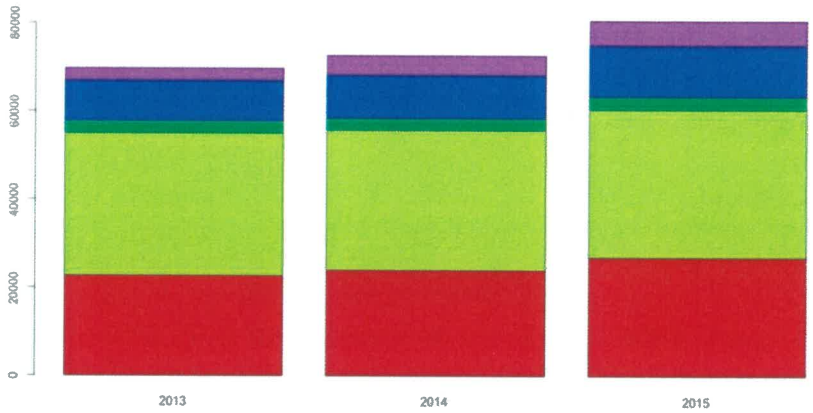
Tabla 8. . Distribución Abandono / Forma de Pago

	No Abandono	Abandono	Total	% Abandono
Anual Domiciliado	66.223	7.653	73.876	10,36%
Anual Efectivo	83.503	14.313	97.816	14,63%
Semestral 1 Pago Efect.	6.965	1.333	8.298	16,06%
Semestral Domiciliado	27.311	3.339	30.650	10,89%
Semestral Efectivo	11.424	1.546	12.970	11,92%
	195.426	**28.184**	**223.610**	

Fuente: Elaboración Propia. RStudio.

El pago en efectivo presenta una mayor tasa de abandono que el pago domiciliado. El comportamiento es de fácil explicación, ya que si no se paga la póliza queda anulada de acuerdo con la Ley de Contrato de Seguro.

Gráfico 10. Estructura de la forma de pago por años



Fuente: Elaboración Propia. Gráfico generado con RStudio.

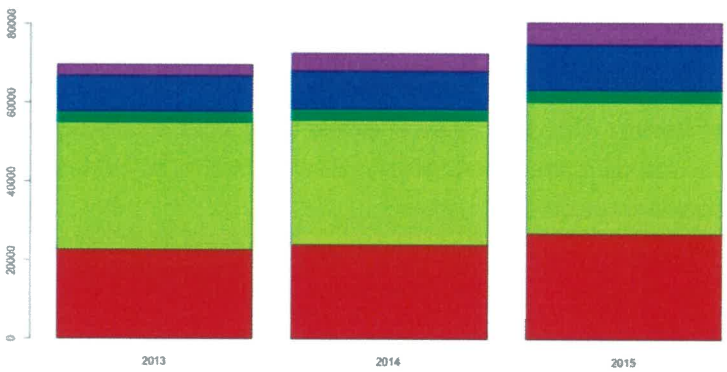
Actualmente la domiciliación bancaria es superior al inicio del período de análisis. En 2013 la domiciliación solo representaba el 41,64%, mientras que en el año 2016 ascendió hasta el 50,4%. El pago anual es el más utilizado, frente al fraccionamiento de las primas en semestrales.

También en este caso se rechaza la hipótesis de igual de abandono entre las diferentes formas de pago, siendo el pago semestral en efectivo en el primer recibo el que tiene una mayor tasa de abandono.

Variable Canal de Distribución

La distribución del seguro de automóviles se centra en los Agentes exclusivos (49%) y los Corredores (40%), sin embargo, el abandono será mayor en los corredores ya que tienen la opción de ofertar el mismo seguro con otras compañías aseguradoras. La estructura de la distribución no ha cambiado en el período de análisis (2012-2016).

Gráfico 11. Estructura de los Canales de Distribución por años



Fuente: Elaboración Propia. Gráfico generado con RStudio.

Tabla 9. Distribución Abandono / Canal de Distribución

	No Abandono	Abandono	Total	% Abandono
Agentes Exclusivos	103.378	11.400	114.778	9,93%
Agentes Vinculados	7.819	1.157	8.976	12,89%
Corredores	67.669	13.987	81.656	17,13%
Empleados	5.687	503	6.190	8,13%
Oficinas Entidad	10.870	1.137	12.007	9,47%
	195.423	**28.184**	**223.607**	

Fuente: Elaboración Propia. RStudio.

También en este caso se rechaza que la tasa de abandonos sea igual en los diferentes canales de distribución, siendo el canal corredores el que muestra una tasa más elevada (17,13%) casi el doble que los Agentes Exclusivos (9,93%).

Interacción Forma de Pago y Canal de Distribución

La domiciliación es mayor entre los corredores, sin embargo, su tasa de abandono es mayor en corredores con pago en efectivo. El porcentaje de pago en efectivo entre agentes exclusivos es muy alto, pero la tasa de abandonos es muy baja en este canal.

Los Agentes Vinculados y los empleados tienen un porcentaje de domiciliación muy alto, frente a las Oficinas de Entidad donde el pago es principalmente en efectivo.

Tabla 10. Distribución Forma de Pago / Canal de Distribución

	A. Exclusivos	Corredores	A. Vinculados	Empleados	O. Entidad	Totales
Anual Domiciliado	37.542	25.903	3.058	2.696	4.677	**73.876**
Semestral Domiciliado	12.977	13.905	1.545	808	1.415	**30.650**
Anual Efectivo	55.426	32.202	2.947	2.290	4.948	**97.813**
Semestral 1er Pago E.	2.821	4.675	550	62	190	**8.298**
Semestral Efectivo	6.012	4.971	876	334	777	**12.970**
Totales	**114.778**	**81.656**	**8.976**	**6.190**	**12.007**	**223.607**

Fuente: Elaboración Propia. RStudio.

En el análisis de la varianza efectuado para la prima y los factores Abandono, Forma de Pago y Canal de Distribución, se alcanza los mismos resultados, se rechaza la hipótesis nula para ambos factores y para la interacción.

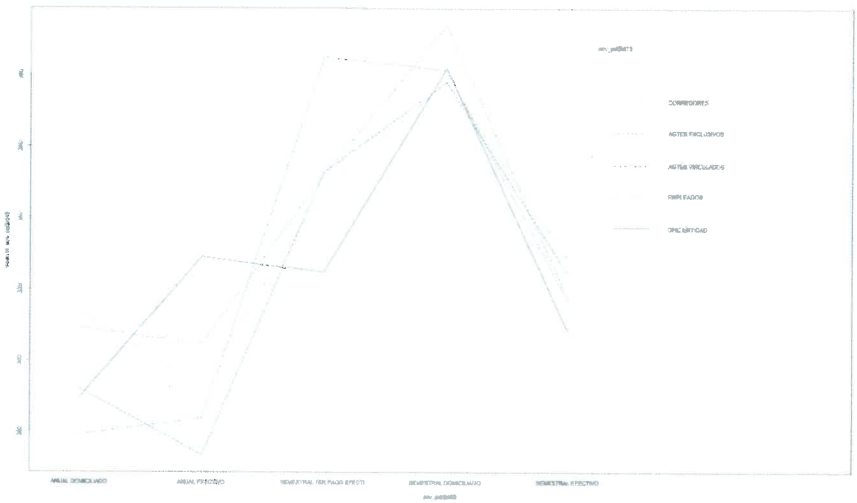
Tabla 11. ANOVA

	Df	Sum Sq	Mean Sq	F value	Pr(>F)	
Abandono	1	7.663e+05	766255	28.25	1.07e-07	***
Canal D.	4	3.245e+07	8113612	299.09	< 2e-16	***
Forma P.	4	2.167e+08	54186571	1997.44	< 2e-16	***
CanalD:FormaP	16	1.393e+07	870686	32.09	< 2e-16	***
Residuals	222833	6.045e+09	27128			

Fuente: Elaboración Propia. RStudio.

Finalmente se ha procedido a representar gráficamente la interacción entre ambos factores (Canal de distribución y Forma de pago), observándose una clara interacción entre ambos.

Gráfico 12. Interacción entre los factores Canal de distribución y Forma de pago



Fuente: Elaboración Propia. Gráfico generado con RStudio.

Variables ocurrencia del evento por día de la semana

Al disponer de diferentes fechas para ocurrencia del siniestro, declaración a la compañía y cierre del mismo se ha procedido a realizar un análisis de los días de la semana en los que ocurren cada uno de ellos por si fuera significativo para el análisis que se está realizando.

Los resultados del análisis son los siguientes:

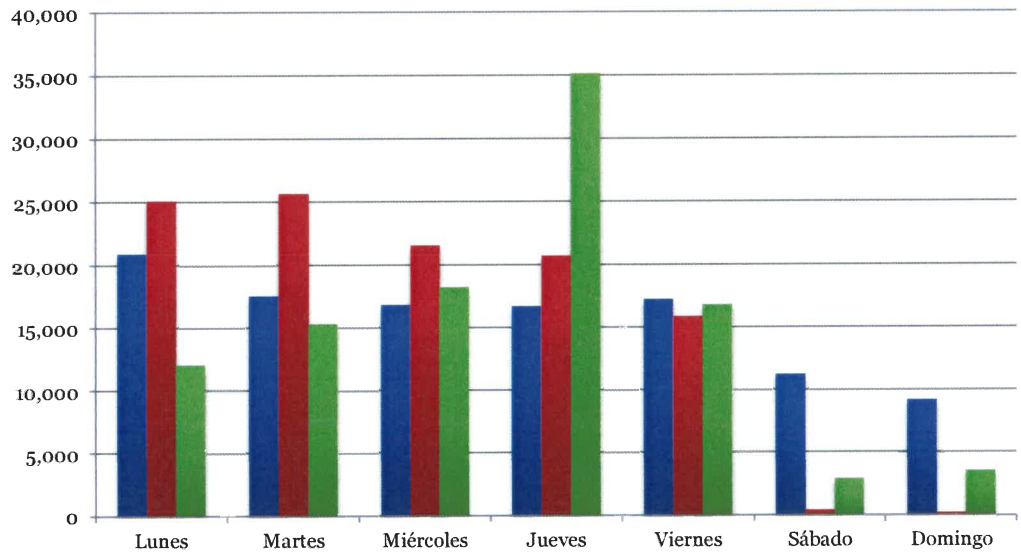
Tabla 12. Distribución de ocurrencia del evento por día de la semana

	Siniestro	Comunicación	Cierre
Lunes	20.862	25.092	12.049
Martes	17.522	25.663	15.302
Miércoles	16.809	21.528	18.198
Jueves	16.671	20.703	35.134
Viernes	17.227	15.872	16.778
Sábado	11.216	410	2.912
Domingo	9.153	192	3.525
	109.460	109.460	103.898

Fuente: Elaboración Propia. RStudio.

Se supone que existe un patrón de comportamiento en la ocurrencia y comunicación de los siniestros que puede condicionar el comportamiento a la hora de tomar la decisión de abandonar voluntariamente la compañía.

Gráfico 13. Distribución de ocurrencia del evento por día de la semana



Fuente: Elaboración Propia. Microsoft Excel.

En el gráfico se observa que la distribución más uniforme sería la de ocurrencia de siniestros (barras azules). Los lunes es cuando más siniestros ocurren, y conforme avanza la semana se va reduciendo la ocurrencia hasta finalizar el domingo con algo menos de la mitad de siniestros que el lunes.

La mayor parte de los siniestros se comunican a la compañía entre semana (barras rojas), siendo muy escasos los que se comunican en fin de semana. Finalmente, las barras verdes representan la fecha de cierre del siniestro, que muestra una peculiaridad, ya que el jueves es el día de la semana que más siniestros se cierran. Un tercio (33,82%) de todos los siniestros se cierran los jueves.

2.3 Selección de Variables

El aprendizaje automático funciona con una regla muy sencilla: GIGO (garbage in, garbage out). La materia prima son los datos por lo que de la calidad de los mismos dependerá el resultado final del proceso de aprendizaje. La mayoría de la metodología estadística utilizada en estos procesos tiene más de 60 o 70 años de antigüedad. Son técnicas contrastadas y que se sabe de su correcto funcionamiento, pero si no se les proporcionan los datos adecuados, el resultado no será bueno. Si los datos tienen demasiado ruido, o están demasiado sucios, podrían entorpecer el proceso de aprendizaje, o llegar a conclusiones erróneas.

Cuando se dispone de un gran número de variables, este principio se vuelve todavía más importante. Aunque la potencia de cálculo de los modernos equipos informáticos, no es necesario utilizar todas las variables para la modelización de un hecho. Los modelos sencillos y fáciles de explicar se impondrán a los modelos complejos y denominados "caja negra". Se impone el principio de parsimonia o la navaja de Ockham. Algunas variables serán importantes, otras aportarán algo de conocimiento, y otras muchas serán irrelevantes, incluso algunas pueden aportar únicamente ruido, lo que sería contraproducente. La práctica nos ha enseñado, aunque no esté de acuerdo con algunos planteamientos teóricos, que algunas veces un subconjunto de variables obtiene mejores resultados que todas las variables disponibles. Se ha introducido mucho ruido y el algoritmo no es capaz de detectar la señal.

Las principales razones para realizar una selección de variables son:

- El algoritmo de aprendizaje automático se entrenará más rápidamente. Con la llegada de los ordenadores cuánticos esto dejaría de ser un problema, ya que la capacidad de cálculo se multiplica exponencialmente.
- Reduce la complejidad del modelo y facilita su interpretación. Empiezan a desarrollarse herramientas que permitan comprender o al menos interpretar correctamente algoritmos o modelos complejos.
- Mejora la precisión de un modelo si se selecciona el subconjunto adecuado. Es evidente que alguna información disponible no tiene relación con el problema que se pretende solucionar, e incluirla generaría más ruido en los datos, y consecuentemente, se reduciría su precisión.
- Reduce la sobre-estimación. El objetivo del aprendizaje automático es realizar predicciones o previsiones, y para ello, los modelos deben generalizar sin ser demasiado simples. La sobre-estimación explica muy bien los datos del conjunto de entrenamiento, pero tiende a facilitar malas predicciones.

A continuación se exponen varias metodologías y técnicas que permitirán seleccionar las variables con la finalidad de conseguir modelos más eficientes y que funcionen mejor.

2.3.1 Métodos Indirectos (filter)

Los métodos indirectos se utilizan generalmente como un paso del pre-procesamiento de los datos. La selección de variables es independiente de cualquier metodología de aprendizaje automático que se aplique a un conjunto de datos. Las variables se seleccionan sobre la base de las valoraciones obtenidas en varias pruebas de correlación estadística con la variable dependiente. La correlación dependerá del tipo de variables que se esté utilizando.

- Coeficiente de Correlación de Pearson: se utiliza para cuantificar la dependencia lineal entre dos variables cuantitativas. El valor oscilará entre -1 y 1, siendo el valor 0 el que representa la incorrelación entre las variables.

$$\rho_{x,y} = \frac{\sigma_{x,y}}{\sigma_x \sigma_y}$$

- ADL: el Análisis Discriminante Lineal se utiliza para determinar una combinación lineal de variables que caracteriza o separa a dos o más clases de una variable cualitativa.
- ANOVA: el Análisis de la Varianza es similar al ADL con la diferencia en que en este caso la variable dependiente es cuantitativa, y las variables independientes son factores. Proporciona un contraste estadístico para determinar si existen diferencias en la media entre los diferentes factores.
- El contraste Chi-cuadrado se utiliza para determinar el grado de asociación entre dos variables cualitativas o atributos basándose en la distribución de frecuencias.

Una cuestión importante en este apartado es recordar que los métodos de filtrado no eliminan la multicolinealidad, por lo que será necesario tratar los problemas de colinealidad de las variables antes de la modelización.

2.3.2 Métodos Directos (wrapper)

Esta metodología consiste en seleccionar un subconjunto de variables para entrenar un modelo, y en función de los resultados se decide añadir o eliminar variables del mismo. El problema es buscar el mejor subconjunto de variables. Estos métodos suelen ser intensivos en cálculo.

A continuación se exponen algunos de los principales métodos directos, también denominados wrapper en la literatura anglosajona:

- Selección hacia adelante es un método iterativo en el que se empieza sin ninguna variable y en cada iteración se va añadiendo la variable que más mejora el modelo. Se dejará de añadir variables cuando la adición de una nueva variable no mejore el rendimiento del modelo.
- Eliminación hacia atrás será la metodología inversa. Se inicia el proceso con todas las variables y se van eliminado iterativamente la variable menos significativa, y consecuentemente, se mejora el rendimiento del modelo. Se repetirá este procedimiento hasta que no se observe ninguna mejoría al efectuar la eliminación de una variable.
- Eliminación de variables recursivas se efectúa con un algoritmo que busca encontrar el subconjunto de variables con mejor rendimiento. De forma repetida se van creando modelos y se mantiene a un lado la mejor y la peor variable en cada iteración. El siguiente modelo se construye con las variables de la izquierda hasta que se agotan todas. A continuación se clasifican las variables según su orden de eliminación.

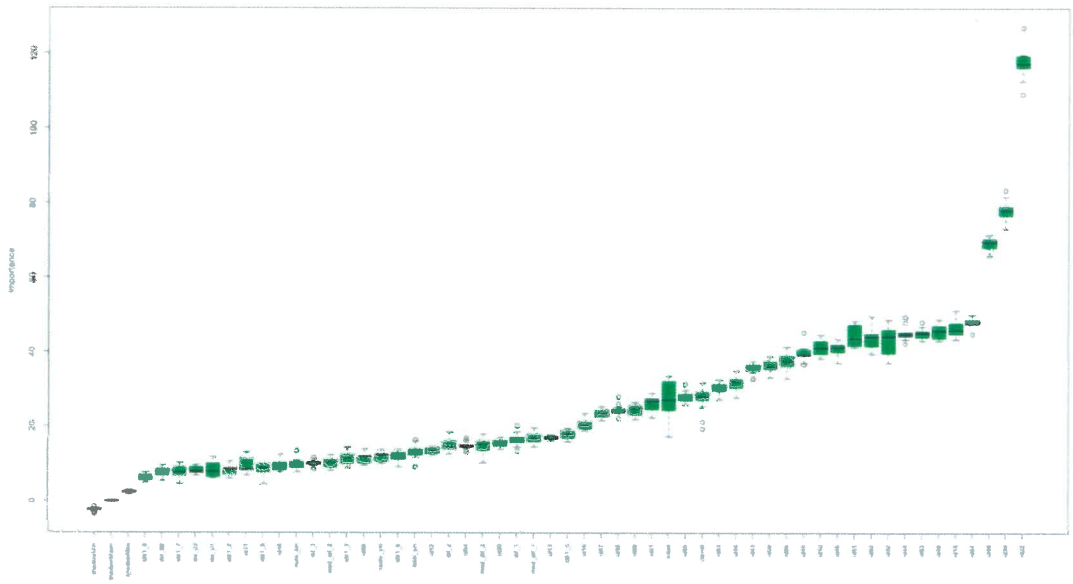
Una de las mejores formas de implementar la selección de variables con métodos directos es utilizar el algoritmo Boruta, que se aplicará con la librería que lleva el mismo nombre y fue dsarrollada por Kursa y Rudnicki (2010). El algoritmo funciona de la siguiente manera:

1. En primer lugar se agrega aleatoriedad al conjunto de datos mediante la creación de copias barajadas de las variables, las cuales se denominan variables sombra.
2. Luego se entrena un bosque aleatorio (random forest) con el conjunto de datos ampliado y se obtiene una media de la importancia de las variables (la medida por defecto es la precisión de disminución de la media, Mean Decrease Accuracy).

3. En cada iteración se comprueba si la variable real tiene mayor importancia que la variable sombra, y se eliminan las variables que se consideran poco importantes.
4. El algoritmo se detendrá cuando todas las variables se hayan aceptado o rechazado.

Se aplica el algoritmo Boruta al conjunto de entrenamiento para determinar que variables entrarán a formar parte del conjunto final para el entrenamiento de los modelos.

Gráfico 14. Importancia de las variables (algoritmo Boruta de R)



Fuente: Elaboración Propia. Gráfico generado con RStudio.

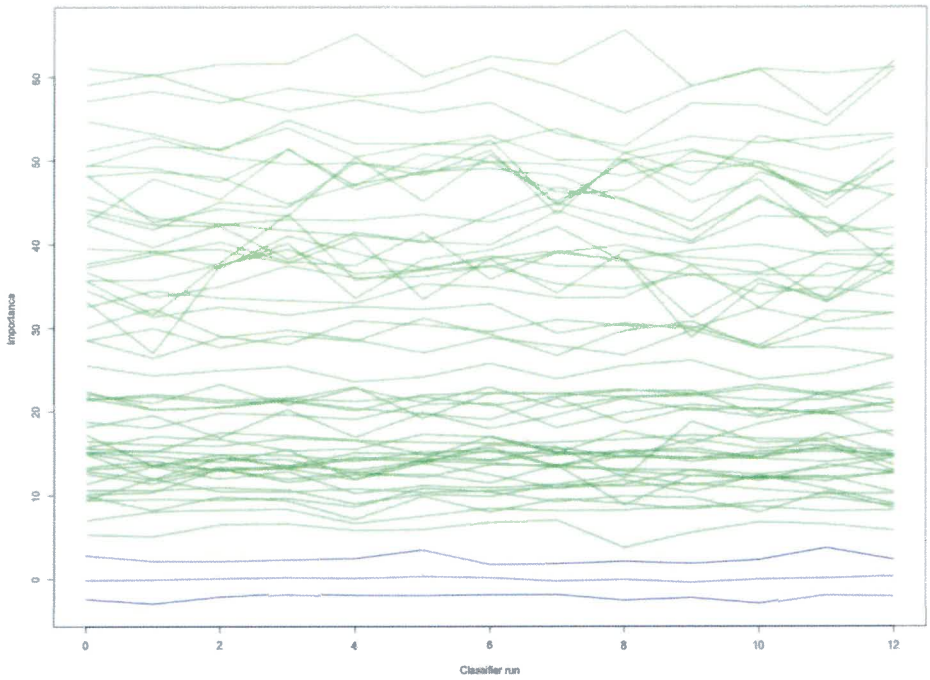
La representación gráfica muestra las variables que deben permanecer en el modelo en color verde, en rojo las que se deben eliminar, en amarillo las que quedan a decisión del investigador, y finalmente, las variables "sombra" son representadas en azul.

Todas las variables se muestran en verde, excepto las variables "sombra". No se eliminará ninguna de las variables para la modelización del "riesgo de fuga".

Las variables que muestran una mayor importancia según esta metodología son:

- Número de pólizas en vigor de todos los ramos.
- Número de ramos contratados.
- Número de pólizas anuladas del cliente.
- Año de matriculación del vehículo.
- Antigüedad del cliente en años.
- Prima del año en curso.

Gráfico 15. Evolución de la Importancia de las variables (algoritmo Boruta de R)



Fuente: Elaboración Propia. Gráfico generado con RStudio.

También se puede conocer la evolución de la importancia de las variables en función del número de ejecuciones del algoritmo. Se puede comprobar que en este caso particular se ha comportado con bastante estabilidad.

2.3.3 Métodos Integrados (Ridge y LASSO)

Los métodos integrados combinan los métodos directos e indirectos. Se implementan a través de algoritmos que tienen sus propios métodos de selección de variables incorporados. Algunos de los más populares son la regresión LASSO y RIDGE que tienen funciones de penalización intrínsecas para reducir el sobre ajuste.

- La regresión LASSO realiza la regularización L1 que agrega una penalización equivalente al valor absoluto de la magnitud de los coeficientes.
- La regresión RIDGE la regularización L2 que agrega una penalización equivalente al cuadrado de la magnitud de los coeficientes.

3. Metodología

Nisbett (2016) afirma "*El mejor predictor del comportamiento futuro es el comportamiento pasado*". Y sobre esta premisa se pretende fundamentar la metodología para ofrecer una solución al problema objeto de estudio.

Una vez que se ha seleccionado el conjunto de variables que formarán parte del análisis, se continuará con el proceso de extracción de la información de los datos. La siguiente etapa a cubrir será la división del conjunto de datos en dos subconjuntos, uno de entrenamiento y otro de comprobación. No se debe perder de vista que el objetivo es la predicción, y no la simple descripción. A continuación se seleccionarán los algoritmos con los que primero se entrenarán los modelos y posteriormente clasificarán las instancias del conjunto de entrenamiento. Como ya se ha indicado, el conjunto de datos está desequilibrado en su variable dependiente, por lo que se revisarán las metrologías que permitirán solucionar, sino totalmente, al menos parcialmente el problema. Se finalizará este apartado estableciendo cuáles serán las medidas de evaluación de los diferentes modelos propuestos. Poder comparar los resultados permitirá seleccionar el mejor de ellos, y dadas las espaciales características de la muestra no será suficiente con una sola medida de precisión.

3.1 División del conjunto de datos

El procedimiento habitual en el aprendizaje automático o machine learning es la división de los datos en dos subconjuntos: uno de entrenamiento y otro de comprobación o test. Al tratarse de datos que tienen una temporalidad la división se efectuará en función de los períodos que se pretendan predecir.

En función de los algoritmos utilizados nos podemos encontrar modelos que explican muy bien los datos de entrenamiento, tienen una escasa capacidad predictiva. Nuestro principal objetivo es poder predecir con antelación a su marcha que clientes tienen un mayor riesgo de abandono para poder llevar a cabo acciones comerciales para retener el cliente.

Una vez determinados los períodos de entrenamiento y predicción, se generarán los dos subconjuntos para construir los modelos que explicarían el abandono de los clientes. Al disponer de una base de datos que abarca 3 años, se va a proceder a realizar un análisis en cascada. Se utilizará como datos de entrenamiento un período inicial de dos años, y se realizarán las predicciones sobre el trimestre inmediatamente posterior.

Este procedimiento se repetirá cuatro veces, con diferentes conjuntos de entrenamiento y de comprobación, como si se estuviera avanzando en el tiempo. Este procedimiento permitirá analizar si el modelo varía con el tiempo y su comportamiento temporal. La capacidad predictiva del modelo se obtendrá al comparar los resultados reales de la muestra con las predicciones del modelo.

De acuerdo con los datos disponibles los conjuntos de entrenamiento y comprobación para cada una de las cuatro repeticiones se muestran en el siguiente gráfico.

Gráfico 16. División de los Datos para Entrenamiento y Comprobación

Fuente: Elaboración Propia. Gráfico generado con Microsoft Excel.

3.2 Datos desequilibrados

En castellano tenemos una expresión para indicar que un suceso es muy raro, y se dice que "es más difícil de encontrar que una aguja en un pajar". Este problema es bastante habitual en el aprendizaje automático, porque el grupo que se pretende identificar es muy pequeño dentro de un colectivo enorme. Los arcos de seguridad en los aeropuertos es el ejemplo perfecto para estos casos. De los millones de pasajeros que vuelan todos los años, sólo un número muy pequeño son terroristas o delincuentes, sin embargo, todo ellos tienen que pasar por los controles de seguridad. Cuando se tiene una variable con dos categorías donde una es mucho más grande que la otra, se dice que los datos están desequilibrados.

Los delitos o los fraudes son un conjunto pequeño del total de la actividad humana, al menos en los países desarrollados. Si las acciones fraudulentas que se van a examinar son únicamente el 1% del total de la actividad generada, un modelo naif que no tuviera en cuenta la actividad delictiva tendría un porcentaje de acierto del 99%. El problema es que todos los delincuentes seguirían en la calle, y nadie podría asegurar que mañana hubiera un 2% de delitos. Es decir, la clase minoritaria es la clase más importante tal como exponen Yen y Lee (2009).

Weiss (2004) y López et al. (2013) exponen algunas razones por las que los datos desequilibrados reducen la precisión de los algoritmos de aprendizaje automático, y son las siguientes:

- La distribución desigual de la variable dependiente. La clase no abandono cuenta con un gran número de instancias por lo que formará un conjunto más homogéneo y con menor dispersión que la clase abandono. Esta diferencia no será únicamente por el número de instancias en cada clase, sino también, por las características de cada una de ellas.
- El rendimiento de los clasificadores se distorsionan hacia la clase mayoritaria. Generalmente, los algoritmos tienden a reducir el error general que se produce en la

modelización, y consecuentemente, los errores de la clase minoritaria contribuyen muy poco al error general.

- Los algoritmos tienen como objetivo minimizar el error general al que la clase minoritaria contribuye muy poco.
- Los algoritmos parten del supuesto de clases equilibradas.
- Los errores obtenidos por diferentes clases tienen la misma ponderación. No todos los errores tienen porque tener la misma ponderación, y deberán ser valorados en función de su importancia.
- El ruido tiene un mayor impacto en los casos raros que en los casos comunes.

Para evitar este problema se disponen de diferentes métodos para balancear los datos. Se producirá una modificación del tamaño de la muestra original para generar un nuevo conjunto de datos balanceado. Estos métodos han conseguido gran importancia después de que muchos investigadores hayan probado que los resultados de los datos balanceados en un proceso de clasificación mejoran sustancialmente en comparación con los datos desequilibrados.

De acuerdo con López, et al. (2013) hay tres grandes categorías para solucionar el problema de datos desequilibrados:

1. Muestreo de los datos: Esta solución pretende balancear las clases de una variable mediante la adición o la eliminación de instancias a través de un procedimiento predeterminado. A su vez, se pueden diferenciar dos subcategorías:
    a. Submuestreo y Sobremuestreo: eliminar o añadir instancias para equilibrar las clases.
    b. Generación de datos sintéticos: crear en base a los datos disponibles instancias de la clase minoritaria que se parezcan mucho a los datos, pero que sean diferentes.
2. Aprendizaje sensible a costes: Inicialmente, todos los errores tienen la misma ponderación, pero con este enfoque, se asignan diferentes costes a los diferentes errores, modificándose el objetivo de la modelización. El error deja de ser la variable a minimizar, pasando a ser el coste asociado a la clasificación.
3. Modificación del algoritmo: Este procedimiento estaría orientado a adaptar el método de aprendizaje para que este mejor sintonizado con la clase desequilibrada. Esta solución implicaría la modificación interna de los algoritmos para fijar un objetivo que no sea la precisión general del modelo. En el presente trabajo no se aplica esta metodología.

Chawla et al. (2002), Kotsiantis et al. (2006) y Yen y Lee (2009) concluyen que la generación de datos sintéticos trabaja mucho mejor que el método de sobremuestreo aleatorio y previene el sobre-ajuste. Para la generación de datos sintéticos se dispone de dos algoritmos: SMOTE (Synthetic Minority Oversampling TEchnique) y ROSE (Random Over-Sampling Examples) que es una técnica proyectada por Lunardon, Menardi, Torelli (2014). El algoritmo SMOTE ha sido propuesto por Chawla et al. (2002) y desarrollado para R en la liberia DMwR (Torgo, 2010).

Existen otros métodos avanzados para balancear conjuntos de datos desequilibrados, como por ejemplo, el basado en clústers, el muestreo sintético adaptativo, la línea de borde SMOTE, SMOTEboost, DataBoost-IM, métodos basados en kernel y muchos más. El trabajo básico en estos algoritmos es similar a los anteriores, con algunas pequeñas diferencias.

Chen et al. (2004) propone dos maneras de tratar el problema de los datos desequilibrados: el aprendizaje sensible a costes y el balanceado mediante la generación sintética de datos. En esta ocasión, se ha optado por el segundo tratamiento por considerarlo más robusto y el análisis se efectuará en dos etapas:

- Los datos tal cual han sido facilitados sin realizar ningún tipo de transformación o aplicación de costes (Datos Desequilibrados).
- Generación de datos sintéticos en la clase minoritaria con el algoritmo SMOTE.

3.3 Técnicas de Machine Learning

Los métodos de aprendizaje automático se vienen empleando en la resolución de problemas de gestión de clientes desde distintas vertientes. El problema de este proyecto de investigación es el de clasificación de los distintos clientes, en particular, clientes con probabilidad de abandono y clientes que permanecerán en la compañía.

Para llevar a cabo esta clasificación se analizan distintos algoritmos. En primer lugar, se ha procedido a efectuar una revisión bibliográfica de los métodos que han obtenido mejores resultados en problemas análogos. Se ha partido de las revisiones efectuadas por Vafeaiadis et al. (2015) y Kumar y Garg (2013), para posteriormente revisar otra literatura más específica (Burez y Van den Poel, 2009; Günther et al., 2014; Huigevoort, 2015; Sundarkumar y Ravi, 2015; Gordini y Veglio, 2017).

En base a esta revisión se ha procedido a aplicar diferentes algoritmos que habían obtenido buenos resultados en problemas similares. No obstante, en este trabajo sólo se mostrarán los resultados de tres algoritmos, que son:

- Árbol de decisión con poda.
- Bosque Aleatorio.
- Extreme Gradient Boosting.

El análisis se realizará sobre la base de datos con un total de 53 variables (52 variables independientes y una variable endógena).

Tradicionalmente, los algoritmos de bosques aleatorios han destacado en las competiciones de ciencia de datos pero recientemente ha sido postergado por el algoritmo Extreme Gradient Boosting. Su éxito se debe no sólo a su precisión en la clasificación de instancias sino a la velocidad de procesamiento. Aunque no existe literatura que apoye la aplicación de esta metodología para la resolución de problemas de riesgo de fuga, se ha decidido incluirla por los buenos resultados obtenidos.

3.3.1 Árboles de Decisión

Santamaría (2006) define un árbol de decisión (DT: Decision Tree) como "*un conjunto de condiciones organizadas en una estructura jerárquica, de tal manera que la decisión final a tomar, se puede determinar siguiendo las condiciones que se cumplen desde la raíz del árbol hasta sus hojas*". El árbol de decisión se puede fácilmente traducir a reglas, y por lo tanto, fácil de comprender (Huang y Hsueh, 2010 y Shmueli et al., 2011).

Además se puede hacer una representación del mismo, permite entender el modelo, y posteriormente, transformar en reglas de decisión las ramas y las hojas del árbol. El árbol de decisión clasifica las instancias de acuerdo con un objetivo.

De acuerdo con Shmueli et al. (2011), la finalidad de la técnica de árboles de decisión es clasificar instancias y reducir la dimensión de los datos. Los árboles de decisión se pueden utilizar tanto para problemas de clasificación, indicando una probabilidad de pertenecer a una clase, como para problemas de regresión, estimando el valor de la variable objetivo.

Los árboles de decisión parten del nodo raíz con todas las instancias, y desarrollan dando lugar a distintas ramas, una por cada valor disponible del atributo. En primer lugar, el algoritmo seleccionará un atributo para confeccionar una división del total de instancias, generando dos o más subconjuntos con unos valores particulares del atributo. Si el valor del atributo en un subconjunto tiene la misma clase se convierte en hoja, en caso contrario será un nuevo nodo en el que se tendrá que continuar discriminando.

Los algoritmos de árboles de decisión particionan los datos recursivamente hasta que se cumple alguna condición, como la minimización de la entropía o la clasificación de todas las instancias. Debido a este procedimiento, la tendencia es a generar árboles con muchos nodos y nodos con muchas hojas, lo que supone un sobreajuste o sobreentrenamiento.

Se deberá solucionar este problema con un procedimiento de poda a posteriori de la construcción del árbol. La idea para un procedimiento de postpoda es medir el error estimado de cada nodo, de modo que si el error estimado para un nodo es menor que el error estimado para sus subnodos, entonces los subnodos se eliminan.

Se dispone de diferentes algoritmos para la generación de los árboles de decisión. Los más utilizados son:

- CART (Classification and Regression Trees) desarrollado por Breiman en 1984. Se basa en el lema "divide y vencerás", y está basado en el criterio de Gini. El procedimiento de poda se realizará a través de una estimación de la complejidad del error.
- ID3 fue propuesto por Quinlan en 1986 y se puede considerar el árbol de decisión más simple, usa la ganancia de información como criterio de división. El árbol crece hasta encontrar un nodo final, sin procedimientos de poda.
- El algoritmo C4.5. es la evolución del ID3, y fue presentado por Quinlan en 1993. Utiliza como criterio de separación el ratio de ganancia.

Hay disponibles diferentes algoritmos en R para generar árboles de decisión, y para el presente trabajo se ha optado por la librería rpart desarrollada por Therneau y Atkinson (2018), y para su representación gráfica se ha utilizado la librería rpart.plot de Milborrow (2018).

3.3.2 Bosques Aleatorios

La evolución natural de los árboles de decisión individuales son los bosques aleatorios o Random Forest (Breiman, 2001). La estrategia de los bosques aleatorios es seleccionar de forma aleatoria un subconjunto de predictores para construir árboles de decisión, cada árbol se desarrolla sobre una muestra bootstrap del conjunto de entrenamiento. El número de variables para construir el árbol es inferior al total de variables disponibles, por lo que la división que se haga de los nodos será diferente y los árboles menos profundos (Larivière y Van den Poel, 2004). Como cada árbol depende de los valores de una muestra aleatoria de variables e instancias y el bosque aleatorio estándar es una combinación de las predicciones de los árboles con la misma distribución de todos los árboles en el bosque (Breiman, 2001), el bosque aleatorio no trabaja correctamente con conjuntos de datos que están muy desequilibrados, como los problemas de riesgo de fuga.

Los bosques aleatorios han sido utilizados para modelizar el riesgo de fuga por Larivière y Van den Poel (2004), Burez y Van den Poel (2009) y Xie et al. (2009) entre otros. La librería que se ha utilizado es randomForest que se ha llevado a cabo por Liaw y Wiener (2002).

3.3.3 Extreme Gradient Boosting

El Extreme Gradient Boosting (Friedman, 2001) se basa en la idea de entrenar el algoritmo mediante la actualización de los pesos de las observaciones pertenecientes a las clases del suceso de interés a través de la optimización en dirección descendente de una función de pérdida o error determinada, consiguiendo dar mayor relevancia en cada iteración a las observaciones mal clasificadas en pasos anteriores. Es por tanto, un ensamble de métodos de aprendizaje que combina el poder de predicción de diferentes métodos en un único modelo agregado.

Es, por tanto, un algoritmo adaptativo en el que cada clasificador se construye en base a los resultados obtenidos en los clasificadores previos mediante la asignación de pesos a cada uno de los casos de entrenamiento. Aquellas instancias que no han sido correctamente clasificadas por los clasificadores anteriores o han tenido peores resultados, tendrán más importancia a la hora de construir el nuevo clasificador. De esta manera, los clasificadores se centran en aquellos casos que son más difíciles de etiquetar correctamente por el conjunto de entrenamiento.

De acuerdo con Hidalgo (2014), el extreme gradient boosting se definió en su origen como un método enfocado a reducir significativamente el error de cualquier algoritmo débil (se entiende como algoritmo débil todo aquel cuyo error es un poco menor que el clasificador aleatorio) pero, en la práctica, es una técnica que se combina con árboles de decisión, los cuales no se consideran débiles.

XGBoost (Chen et al., 2018) es una implementación muy popular del gradient boosting ya que tiene algunas características que lo hacen muy interesante:

- Estructura de bloques para el aprendizaje en paralelo: XGBoost puede hacer uso de los múltiples núcleos de la CPU.
- XGBoost ha sido diseñado para hacer un uso óptimo del hardware.

- Está obteniendo los mejores resultados en las competiciones de ciencia de datos que se llevan a cabo en Kaggle y otras plataformas.

Aunque no se dispone de literatura científica que sustente la utilización de esta metodología, se ha optado por incluirla en base a los buenos resultados obtenidos.

3.4 Medidas para la evaluación de los modelos

Lo más normal es no poder construir modelos perfectos que permitan clasificar correctamente todas las instancias del conjunto de comprobación. Entonces, se deberá elegir el modelo de clasificación que mejor se adapte a las necesidades y mejor funcione en el dominio del problema (Burez y Van den Poel, 2009).

La matriz de confusión o matriz de clasificación, es útil en el aprendizaje automático para comparar los valores reales de una variable con los valores estimados por el modelo construido con los datos de entrenamiento. Al construirse como una matriz las filas representan los valores de predicción del modelo y las columnas los valores reales de la variable. De esta forma se pueden identificar cuatro categorías: falso positivo, verdadero positivo, falso negativo y verdadero negativo.

La importancia de la matriz estriba en que no solo nos indica que instancias están correctamente clasificadas, sino donde están clasificadas aquellas que lo han sido incorrectamente.

Seleccionar la métrica adecuada es un aspecto crítico del trabajo con datos desequilibrados. La mayoría de los algoritmos calculan la exactitud basada en el porcentaje de observaciones correctamente clasificadas. Con datos desequilibrados, los resultados son engañosos porque la clase minoritaria tiene muy poco efecto sobre la exactitud global.

		Previsión	
		Positivo	**Negativo**
Actual	**Positivo**	Verdadero Positivo (TP)	Falso Negativo (FN)
	Negativo	Falso Positivo (FP)	Verdadero Negativo (TN)

Partiendo de la matriz de confusión, las métricas comúnmente más utilizadas son la Exactitud y la Tasa de Error:

Exactitud (Accuracy) = (TP + TN) / (TP + TN + FP + FN)

Tasa de Error = 1- Exactitud = (FP + FN) / (TP + TN + FP + FN)

Sin embargo, como indican Provost y Fawcatt (2013) la calidad de este indicador es muy simple cuando se trabaja con datos desequilibrados, ya que se precisa de indicadores que no sean tan sensibles a cambios en los datos. De la matriz de confusión se podrán obtener otra serie de indicadores:

- La sensibilidad, que medirá cuantas instancias de la categoría positiva se etiquetan correctamente.

  Sensibilidad (Sensitivity) = TP / (TP + FN)

- Especificidad (Specificity) = TN / (TN + FP)
- La Precisión es una medida de exactitud que indica como de buena es la predicción del modelo para las instancias etiquetadas como positivas (Keramati et al., 2014).
  Precisión (Precision) = TP / (TP + FP)

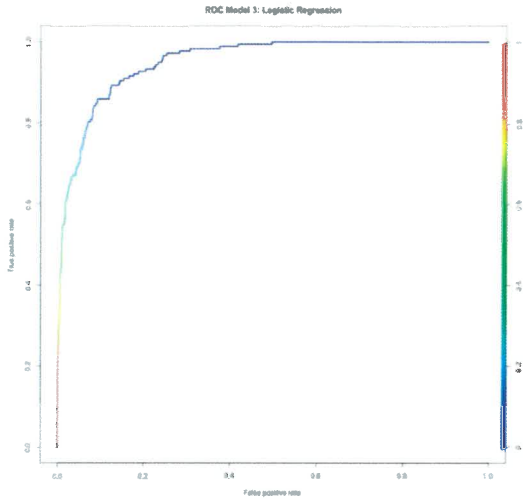
El estadístico Kappa de Cohen es un índice que compara la precisión global del modelo con la precisión que se obtendría si el modelo clasificase de forma aleatoria las instancias (Kaymak et al., 2012). Kappa se define como la diferencia entre la precisión general y la precisión esperada dividida por 1 menos la precisión esperada. Si se obtiene un valor de +1 indicaría total coincidencia (el valor ideal), valores de 0 indican que la coincidencia es la misma que se puede esperar por casualidad y valores de −1 expresarían total desacuerdo.

El estadístico Kappa de Cohen es un índice que compara el nivel de coincidencia entre varios expertos con el nivel de coincidencia que se podría dar por casualidad. Si se obtiene un valor de +1 indicaría total coincidencia (el valor ideal), valores de 0 indican que la coincidencia es la misma que se puede esperar por casualidad y valores de −1 expresarían total desacuerdo.
La prueba de McNemar se aplica en matrices de confusión de dos por dos para determinar si las frecuencias marginales de las filas y las columnas son iguales. Lo introdujo Quinn McNemar en 1947 para determinar que el número de falsos positivos no es igual al número de falsos negativos, es decir, no hay evidencia de efecto tratamiento.

La curva ROC (Receiver Operating Characteristics) y el área bajo la curva ROC (AUC) se utilizan para medir la precisión de una clasificación tal como lo expone Bradley (1997). Esta curva ROC se obtienen representado la tasa de TP (sensibilidad) y la tasa de FP (especificidad). Cualquier punto en el gráfico ROC, corresponde al rendimiento de un solo clasificador en una distribución dada. Es útil porque proporciona una representación visual de los beneficios (TP) y los costes (FP) de una clasificación (Kim et al., 2014). Cuanto mayor sea el área bajo la curva, mayor será la precisión.

En el siguiente gráfico se muestra un ejemplo de la Curva ROC para un modelo de clasificación que tiene un valor de 0.9524 para su AUC. Los valores del AUC oscilarán entre 0 y 1; el valor 0,5 vendría representado por la línea diagonal y los valores próximos a 1 indicarían un buen ajuste del modelo respecto a los datos.

Gráfico 17. Curva ROC de un modelo de clasificación



Fuente: Elaboración Propia. Gráfico generado con RStudio.

Huigevoort (2015) afirma que con los cuatros parámetros de AUC, AUK, la precisión y la sensibilidad, se podrá seleccionar el mejor modelo para cada una de las diferentes técnicas y luego compararlos con otra metodologías.

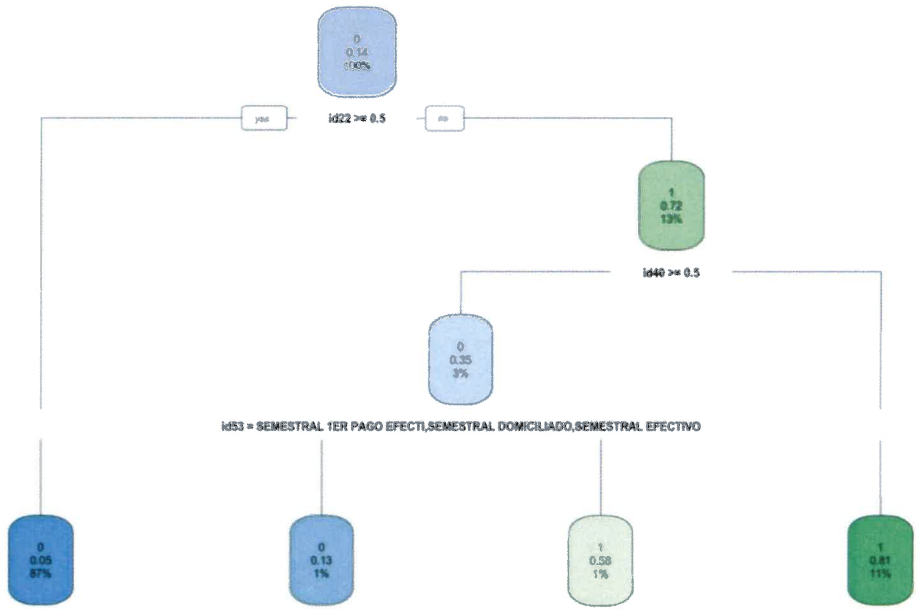
4. Resultados de las técnicas de predicción

Los modelos se han entrenado con los datos correspondientes a un período de dos años y posteriormente se han comprobado las predicciones sobre el trimestre inmediatamente posterior al período de entrenamiento. Además, se han generado los modelos con datos desequilibrados y posteriormente con la inclusión de datos sintéticos mediante el algoritmo SMOTE. En el primer subapartado se presentan los resultados de los datos desequilibrados, que son los datos originales, de los tres modelos y en el segundo subapartado únicamente se muestran los resultados para el modelo realizado con el algoritmo XGBoost con datos SMOTE.

4.1. Resultados de los modelos con datos desequilibrados

Se ha realizado el primer análisis del árbol de decisión mediante el algoritmo rpart (Therneau y Atkinson, 2018) y rpart.plot (Milborrow, 2018) de R, estos algoritmos sirven para clasificar las instancias de entrenamiento mediante árboles de decisión, y se ha obtenido el árbol de decisión que se muestra a continuación, el cual se podría transformar con cierta facilidad en reglas de decisión.

Gráfico 18. Árbol de Decisión con datos desequilibrados



Fuente: Elaboración Propia. Gráfico generado con RStudio.

El algoritmo permite representar gráficamente el árbol de decisión, y de esta forma, tener una visión, fácilmente comprensible, del modelo que se ha ajustado. El gráfico incluye las variables que se utilizan en cada nodo para realizar la división, así como, el porcentaje de instancias y la probabilidad de abandono (azul no hay riesgo de fuga y verde si que lo hay). Como se puede comprobar la rama de la derecha del árbol es la que identifica el riesgo de fuga, mientras que la rama de la izquierda caracteriza los clientes que permanecerán en la compañía.

La salida de RStudio de los resultados del árbol de decisión en el conjunto de datos de entrenamiento sin ninguna transformación, es la siguiente:

```
n= 138666
node), split, n, loss, yval, (yprob)
      * denotes terminal node

 1) root 138666 19585 0 (0.85876134 0.14123866)
   2) id22>=0.5 120185  6365 0 (0.94703998 0.05296002) *
   3) id22< 0.5 18481  5261 1 (0.28467074 0.71532926)
     6) id40>=0.5 3797  1315 0 (0.65367395 0.34632605)
      12) id53=SEMESTRAL 1ER PAGO EFECTI,SEMESTRAL DOMICILIADO,SEMESTRAL EFECTIVO 2002   265
0 (0.86763237 0.13236763) *
      13) id53=ANUAL DOMICILIADO,ANUAL EFECTIVO 1795   745 1 (0.41504178 0.58495822) *
     7) id40< 0.5 14684  2779 1 (0.18925361 0.81074639) *
```

El árbol de decisión muestra con muy pocas ramas y hojas, una alta capacidad explicativa de los datos analizados. Sin embargo, lo que se busca es la capacidad predictiva de los modelos.

```
       Cross-Validated (10 fold) Confusion Matrix
       (entries are percentual average cell counts across resamples)
                 Reference
       Prediction    0    1
                0 83.5  5.0
                1  2.4  9.2
       Accuracy (average) : 0.9268
```

Con esta modelización se obtiene en promedio una precisión del 92,68%. Si de los datos porcentuales pasamos a los datos del número de instancias, se podría conocer que porcentaje de verdaderos negativos se está clasificando correctamente. El porcentaje de verdaderos positivos es del 94,35%, mientras que el de verdaderos negativos no alcanza el 80% (79,3%). Estos resultados se reducirán cuando se aplique el modelo al conjunto de comprobación para evaluar su capacidad predictiva.

Los resultados obtenidos con este modelo al realizar la predicción sobre el conjunto de comprobación son los siguientes:

```
Confusion Matrix and Statistics
   tree.preds.2
        0     1
  0 16051   687
  1   709  1180

               Accuracy : 0.9251
                 95% CI : (0.9212, 0.9288)
    No Information Rate : 0.8998
    P-Value [Acc > NIR] : <2e-16

                  Kappa : 0.5867
 Mcnemar's Test P-Value : 0.5741

            Sensitivity : 0.9577
            Specificity : 0.6320
         Pos Pred Value : 0.9590
         Neg Pred Value : 0.6247
             Prevalence : 0.8998
         Detection Rate : 0.8617
   Detection Prevalence : 0.8986
      Balanced Accuracy : 0.7949

       'Positive' Class : 0
```

Aunque la precisión es aproximadamente la misma, los verdaderos positivos han ascendido hasta el 95,9% y los verdaderos negativos se han reducido hasta el 62,47%. El modelo predice con bastante precisión los asegurados que no se marcharán pero no hace lo mismo con los que abandonarán la compañía. Esto se produce, en gran medida, por el desequilibrio de las dos clases que se están analizando.

La segunda metodología aplicada es Random Forest o bosques aleatorios que es un ensamble de árboles de decisión y que se ha aplicado con la librería randomForest (Liaw y Wiener, 2002) de R. En lugar de tomar las 52 variables independientes para construir un árbol de decisión, se seleccionan aleatoriamente un número de variables (normalmente raíz cuadrada del número de variables independientes) y se construye un árbol. Este proceso se repite un número grande de veces (100 o más veces) y se establece el resultado del modelo ensamblado mediante votación. Una instancia será clasificada como abandono, cuando la mayoría de los árboles la hayan clasificado de esta forma.

El método es muy robusto porque además se ha utilizado validación cruzada con 10 carpetas (fold), que se repetirá tres veces (repeated), y se modificará el tamaño de la muestra.

El modelo con Random Forest es el siguiente:

```
Random Forest
138666 samples
    52 predictor
     2 classes: 'NO', 'SI'

No pre-processing
Resampling: Cross-Validated (10 fold, repeated 3 times)
Summary of sample sizes: 124799, 124800, 124800, 124800, 124799, 124798, ...
Resampling results:
Accuracy  Kappa
  0.92993   0.684117
Tuning parameter 'mtry' was held constant at a value of 7.745967
```

El número de instancias procesadas es de 138.666, con 52 variables independientes. El número de variables en los diferentes árboles del ensamble indica que es de 7,75. La precisión se ha fijado en 92,99% y el estadístico Kappa en 0,6841. Los resultados son muy parecidos al árbol de decisión del apartado anterior, sin embargo, el tiempo de computación se ha disparado. Si los resultados que se obtengan en la predicción no mejoran sustancialmente el modelo más sencillo, se descartará esta metodología.

Al igual que en el caso anterior la capacidad explicativa del modelo es del 92,99%, y la matriz de confusión sobre los datos de entrenamiento es la siguiente:

```
Cross-Validated (10 fold, repeated 3 times) Confusion Matrix
 (entries are percentual average cell counts across resamples)
Reference
Prediction   NO   SI
        NO 83.8  4.9
        SI  2.1  9.2
Accuracy (average) : 0.9299
```

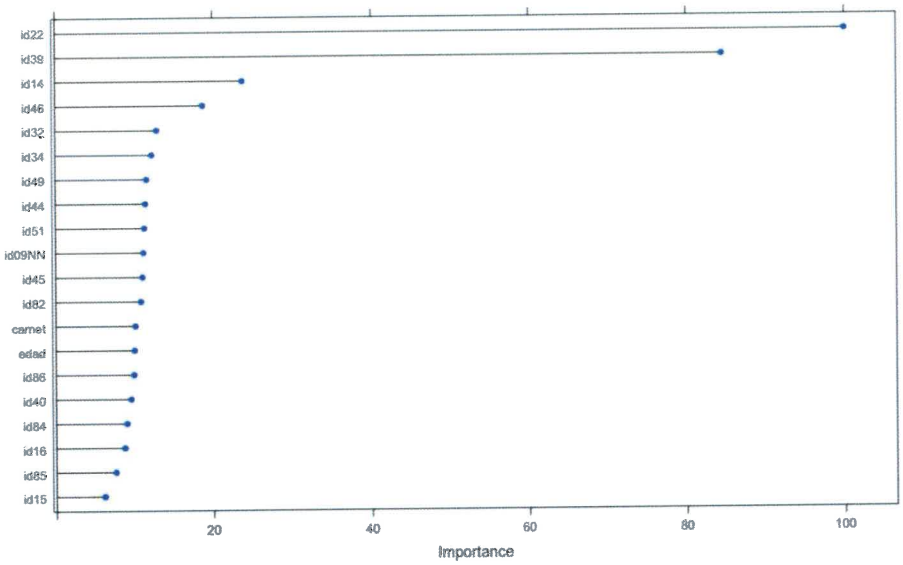
Los resultados obtenidos con este modelo al realizar la predicción sobre el conjunto de comprobación son los siguientes:

```
Confusion Matrix and Statistics
       0     1
  0 16165   573
  1   734  1155

               Accuracy : 0.9298
                 95% CI : (0.9261, 0.9335)
    No Information Rate : 0.9072
    P-Value [Acc > NIR] : < 2.2e-16

                  Kappa : 0.5999
 Mcnemar's Test P-Value : 9.613e-06

            Sensitivity : 0.9566
            Specificity : 0.6684
         Pos Pred Value : 0.9658
         Neg Pred Value : 0.6114
             Prevalence : 0.9072
         Detection Rate : 0.8678
   Detection Prevalence : 0.8986
      Balanced Accuracy : 0.8125

       'Positive' Class : 0
```

La predicción sobre el conjunto de comprobación es ligeramente inferior al árbol de decisión. El porcentaje de verdaderos positivos es casi igual, 95,66%, pero el porcentaje de verdaderos negativos se ha reducido hasta el 61,14%. No parece razonable un incremento en tiempo de computación enorme y una reducción de la capacidad predictiva de la clase de interés.

Como el bosque aleatorio es un ensamble de modelos será deseable conocer la importancia de las variables, que se puede observar en el siguiente gráfico:

Gráfico 19. Importancia de las Variables en el Bosque Aleatorio



Fuente: Elaboración Propia. Gráfico generado con RStudio.

En el gráfico de importancia de las variables, únicamente se han representado las 20 más importantes.

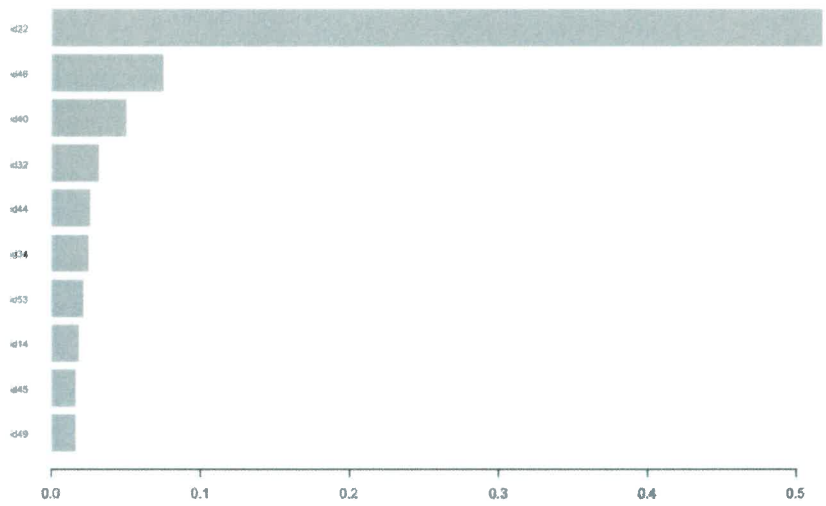
El algoritmo Extreme Gradient Boosting también es un ensamble de árboles de decisión, pero aplicando una técnica diferente a la del random forest. En este caso el boosting es interpretado como un algoritmo de optimización, tal como propuso Breiman (1997). Se optimiza una función de coste sobre el espacio de la función mediante la elección iterativa de una función que apunta en la dirección del gradiente negativo. En el algoritmo xgboost (Chen et al., 2018) se utilizarán árboles de decisión y la función a optimizar será el área bajo la curva ROC, también conocida como AUC.

Se empezará por entrenar el modelo con los datos correspondientes a dos años, y se selecciona el modelo que obtiene mejor resultado para el área bajo la curva ROC. Del modelo se puede obtener la relación de variables más importantes, bien analíticamente, bien gráficamente.

```
    Feature       Gain      Cover  Frequency
 1:    id22 0.51870458 0.10027796 0.04297448
 2:    id46 0.07563338 0.12336406 0.06233500
 3:    id40 0.05060930 0.03953042 0.01408038
 4:    id32 0.03205677 0.08802678 0.07993546
 5:    id44 0.02614136 0.03566822 0.03505427
 6:    id34 0.02505504 0.05205216 0.06614843
 7:    id53 0.02174950 0.02616083 0.01657378
 8:    id14 0.01866177 0.05021967 0.03974773
 9:    id45 0.01664631 0.02307191 0.02845409
10:    id49 0.01621714 0.03525867 0.04121443
```

La importancia de las variables es diferente a la obtenida con el bosque aleatorio, sin embargo, de las 10 primeras variables coinciden 7, aunque con distinta importancia. La estructura de variables para determinar el modelo predictivo es muy similar en ambos casos, aunque la metodología es diferente. Todos los algoritmos utilizados tienen como base los árboles de decisión con diferentes enfoques metodológicos.

Gráfico 20. Importancia de las variables en el modelo XGBoost con datos desequilibrados



Fuente: Elaboración Propia. Gráfico generado con RStudio.

Al efectuar la predicción sobre el conjunto de comprobación, y mostrar los resultados a través de la matriz de confusión y sus estadísticos, se comprueba que aunque la precisión del modelo no se ha incrementado en comparación con el árbol de decisión con poda y el bosque aleatorio, sí que hay una mejora en los verdaderos negativos. Por lo tanto, los verdaderos positivos se habrán visto reducidos. La mejora con respecto al bosque aleatorio es de 4,61% y con respecto al árbol de decisión del 3,28%.

Los resultados de la predicción obtenida con este modelo son los siguientes:

```
Confusion Matrix and Statistics
          0     1
  0 16061   677
  1   647  1242

                Accuracy : 0.9289
                  95% CI : (0.9251, 0.9326)
    No Information Rate : 0.897
    P-Value [Acc > NIR] : <2e-16

                   Kappa : 0.6127
 Mcnemar's Test P-Value : 0.4255

            Sensitivity : 0.9613
            Specificity : 0.6472
         Pos Pred Value : 0.9596
         Neg Pred Value : 0.6575
             Prevalence : 0.8970
         Detection Rate : 0.8622
   Detection Prevalence : 0.8986
      Balanced Accuracy : 0.8042

       'Positive' Class : 0
```

La capacidad predictiva del modelo en los verdaderos positivos es de 65,75%. Es decir, se identifican adecuadamente 2 de cada 3 asegurados que se marchan de la compañía. De acuerdo con estos resultados, trimestralmente se deberían realizar 1.919 acciones individuales encaminadas a retener a otros tantos asegurados en "riesgo de fuga". Los asegurados se identifican por el número de póliza y con el análisis que efectúa xgboostExplainer se pueden realizar acciones individuales y personalizadas para cada uno de ellos.

4.2. Resultados de los modelos con datos sintéticos

Con la técnica SMOTE (Synthetic Minority Oversampling TEchnique) se consigue balancear el conjunto de datos generando datos artificiales, así que sería un forma de sobremuestreo pero con mejores condiciones. Esta técnica genera un conjunto aleatorio de observaciones de la clase minoritaria para cambiar el sesgo de aprendizaje del clasificador hacia la clase minoritaria.

Se utilizará la función SMOTE de la librería DMwR (Torgo, 2010) para balancear los datos de la muestra. Los resultados de su aplicación es un nuevo conjunto de datos con parte de los datos originales y parte de datos generados sintéticamente. Los datos sobre los que se volverán a entrenar los modelos tienen 78.340 instancias (57,14%) de la clase No abandono, y 58.755 instancias (42,86%) de la clase Si abandono.

Cuando se utilizan datos balanceados con SMOTE el número de iteraciones que se precisan para alcanzar un equilibrio es mayor, y por tanto, mayor tiempo de computación. En el caso actual, se han fijado en 1.500 el límite máximo de iteraciones, y se han consumido todas. El tiempo de cálculo es superior a los árboles de decisión, pero muy inferior al bosque aleatorio, ya que como se ha indicado aprovecha todos los núcleos de la CPU.

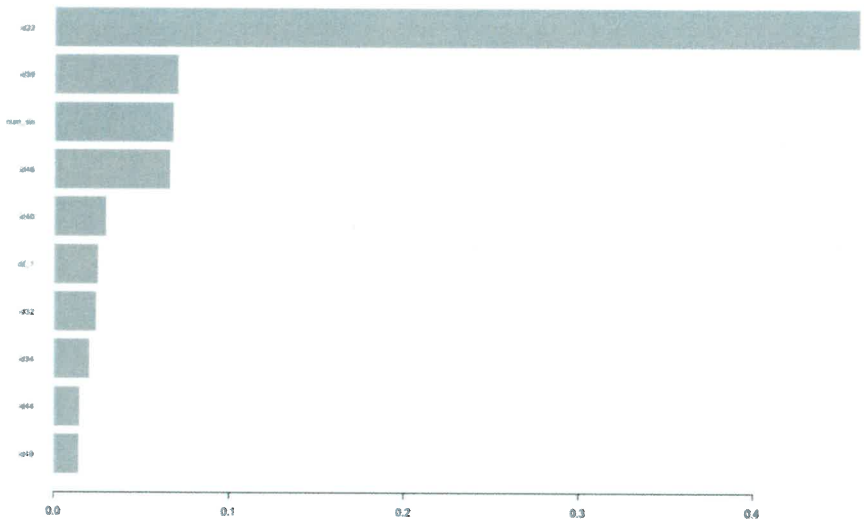
La importancia de las variables en este nuevo modelo es la siguiente:

	Feature	Gain	Cover	Frequency	Importance
1:	id22	0.46077780	0.08223236	0.026066811	0.46077780
2:	id39	0.07031216	0.01705662	0.009687976	0.07031216
3:	num_sin	0.06755848	0.01066015	0.003499336	0.06755848
4:	id46	0.06605265	0.06146208	0.041311603	0.06605265
5:	id40	0.02958264	0.02415433	0.009428766	0.02958264
6:	dif_1	0.02522691	0.01233999	0.006399248	0.02522691
7:	id32	0.02385514	0.09902378	0.080695331	0.02385514
8:	id34	0.02017567	0.07486265	0.077406603	0.02017567
9:	id44	0.01473672	0.03208957	0.032061044	0.01473672
10:	id49	0.01449346	0.05110152	0.058581473	0.01449346

Igual que en el caso anterior, y tomando las 10 primeras variables por importancia, se tiene que 7 variables se repiten si lo comparamos con el xbgoost sin SMOTE, y 6 coinciden con el bosque aleatorio.

Las variables que se incluyen ahora entre las 10 primeras y que no estaban antes son el número de siniestros de la póliza en la anualidad en curso (num_sin) y la diferencia entre la fecha de ocurrencia del siniestro y la fecha de cierre del mismo (dif_1). Estas variables tendrían una sencilla explicación en los casos de fuga. Un asegurado con siniestros que se alargan en el tiempo su resolución sería un cliente poco receptivo a la renovación de la póliza sobre todo si la prima de renovación espera que sea más elevada.

Gráfico 21. Importancia de las variables en el modelo XGBoost con SMOTE



Fuente: Elaboración Propia. Gráfico generado con RStudio.

Con el nuevo modelo se procede a realizar las predicciones sobre el conjunto de comprobación para el trimestre inmediatamente inferior al conjunto de dos años que se ha utilizado para entrenar el modelo. Los resultados de la matriz de confusión y sus estadísticos se muestran a continuación:

```
Confusion Matrix and Statistics
    xgb.preds.smote2
        NO    SI
  NO 15700  1038
  SI   506  1383

               Accuracy : 0.9171
                 95% CI : (0.9131, 0.921)
    No Information Rate : 0.87
    P-Value [Acc > NIR] : < 2.2e-16

                  Kappa : 0.5957
 Mcnemar's Test P-Value : < 2.2e-16

            Sensitivity : 0.9688
            Specificity : 0.5713
         Pos Pred Value : 0.9380
         Neg Pred Value : 0.7321
             Prevalence : 0.8700
         Detection Rate : 0.8429
   Detection Prevalence : 0.8986
      Balanced Accuracy : 0.7700

       'Positive' Class : NO
```

La precisión del modelo se ha reducido en comparación con su homólogo sin datos balanceados, sin embargo, el resultado de verdaderos negativos se ha incrementado hasta un 73,21%. La mejora es de un 7,46%, lo que significa que ahora el modelo predice correctamente 3 de cada 4 asegurados en "riesgo de fuga".

El problema es que ha empeorado el ratio de verdaderos positivos, como era de esperar, y por tanto habrá que ver cuál es el coste que se deberá pagar para obtener mejores predicciones de los verdaderos negativos. El número de acciones que se deberían llevar ahora sería de 2.421 en un trimestre, lo que supone 502 acciones más que en el modelo xgboost sin datos balanceados.

Sería necesario un análisis de coste/beneficio para determinar cuál de las dos modelizaciones optimiza el ingreso de la compañía aseguradora. Para ello sería necesario conocer las acciones dirigidas a retener a los clientes y su coste, así como el coste de captar a un cliente de otra compañía.

4.3. Evaluación de los resultados

La comparación se deberá realizar a dos niveles. El primero, en cuando a la capacidad descriptiva de los modelos, no presentan diferencias sustanciales, y todos tienen una precisión global del modelo similar.

El segundo, en relación con la capacidad predictiva, la metodología SMOTE con el algoritmo xgboost presenta una mejora importante respecto del resto de modelos. Siendo el objetivo clasificar correctamente el máximo de asegurados en "riesgo de fuga".

Tabla 13. Resúmenes Matrices de Confusión de los principales modelos

	DT (UD)		RF (UD)		XGB (UD)		XGB (SMOTE)	
	NO	YES	NO	YES	NO	YES	NO	YES
NO	16.051	687	16.165	573	16.061	677	15.700	1.038
YES	709	1.180	734	1.155	647	1.242	506	1.383
Correctly Classified		92,51%		92,98%		92,89%		91,71%
Incorrectly Classified		7,49%		7,02%		7,11%		8,29%
TP Rate (No)		95,90%		96,58%		95,96%		93,80%
TN Rate (Yes)		62,47%		61,14%		65,75%		**73,21%**
FP Rate (No)		37,53%		38,86%		34,25%		26,79%
FN Rate (Yes)		4,10%		3,42%		4,04%		6,20%

Fuente: Elaboración Propia. Resultados obtenidos con RStudio.

La tasa de verdaderos negativos que se obtiene con el modelo xgboost con SMOTE es la más elevada (73,21%), lo que representa que se identifica a 3 de cada 4 clientes que abandonarán la aseguradora. La precisión general del modelo se ha visto reducida ya que el bosque aleatorio es el que ha obtenido los mejores resultados con un 92,98% de las instancias correctamente clasificadas.

Tabla 14. Medidas de Precisión de los Modelos

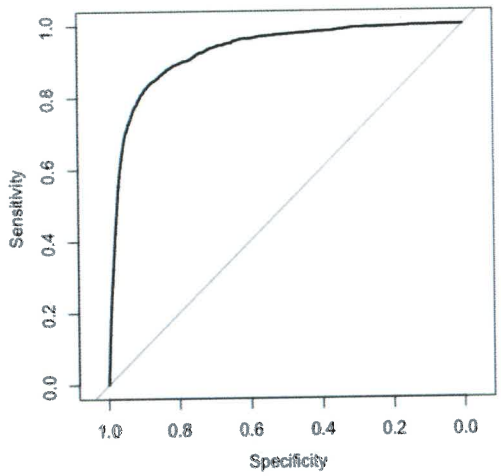
	AUC	Kappa	Balanced Accuracy
DT (UD)	0,7918	0,5867	0,7949
RF (UD)	0,7886	0,5999	0,8125
XGB (UD)	0,9296	0,6127	0,8042
XGB (SMOTE)	0,955	0,5967	0,77

Fuente: Elaboración Propia. Resultados obtenidos con RStudio.

El área bajo la curva ROC (AUC) con el modelo xgboost con SMOTE es mejor que las obtenidas con los datos originales, sin embargo, el estadístico Kappa se ha visto reducido, al igual que la precisión balanceada. El hecho de que la precisión balanceada se reduzca era de esperar ya que lo que se pretende es incrementar la tasa de verdaderos negativos a cambio de empeorar la tasa de verdaderos positivos.

Finalmente se muestra gráficamente la Curva ROC del modelo xgboost con datos generados sintéticamente mediante el algoritmo SMOTE.

Gráfico 22. Curva ROC del algoritmo xgboost con SMOTE



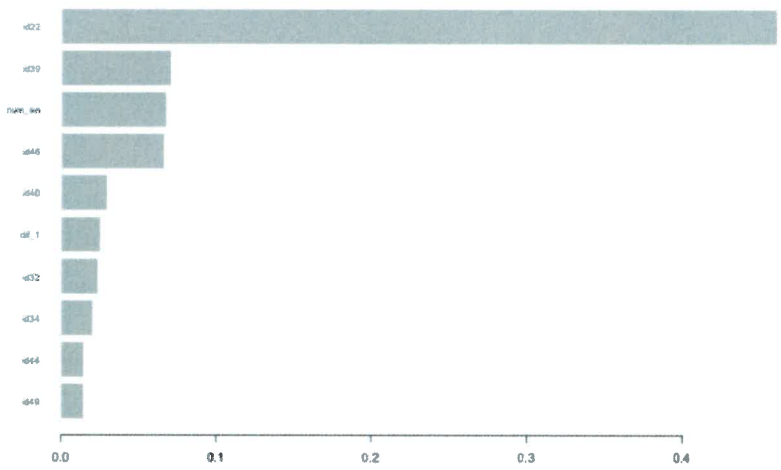
Fuente: Elaboración Propia. Gráfico generado con RStudio.

5. Interpretación de los modelos de riesgo de fuga

El algoritmo xgboost permite la obtención de una probabilidad asociada a la variable objeto de análisis. Es decir, que se calcula la probabilidad que cada cliente se vaya a marchar a su próximo vencimiento, aunque finalmente se clasificará a cada cliente como abandono (probabilidad superior al 50%[1]) o permanencia.

Este algoritmo es considerado como "caja negra", ya que es difícil explicar cómo se obtiene la probabilidad final. Mientras que un único árbol de decisión permite una explicación clara, a través de reglas, un ensamble de árboles, inicialmente, es inescrutable. Una vez efectuadas las predicciones sobre el conjunto de entrenamiento se puede obtener el gráfico de importancia de cada variable. Si se toman las predicciones del modelo xgboost para datos balanceados con SMOTE se obtiene el siguiente gráfico:

Gráfico 23. Importancia de las variables en el algoritmo XGBoost



Fuente: Elaboración Propia. Gráfico generado con RStudio.

El gráfico se ha reducido a las 10 variables ordenadas con mayor importancia en la predicción. De igual forma, se pueden obtener los valores numéricos, que muestran la variable más importante para todas las predicciones, pero no se puede garantizar que sea la más importante para un cliente en particular. La importancia de las variables para un asegurado en concreto, dependerá de los valores que tomen cada una de las variables que se incluyen en el modelo. Los valores que se representan son la ganancia de información (Gain) que se obtienen al dividir cada hoja en otras dos de acuerdo con la siguiente formula:

[1] El porcentaje de probabilidad se podría modificar al alza para intentar reducir los falsos negativos.

$$Gain = \frac{1}{2}\left[\frac{G_L^2}{H_L+\lambda}+\frac{G_R^2}{H_R+\lambda}-\frac{\left(G_L+G_R\right)^2}{H_L+H_R+\lambda}\right]-\gamma$$

Lo que se tiene es la puntuación en la hoja de la izquierda, la puntuación en la hoja de la derecha, la puntuación en la hoja original y la regularización en la hoja adicional. Si la ganancia es menor sería mejor no agregar esa rama.

Los 10 primeros valores obtenidos de la función son estos:

	Feature	Gain	Cover	Frequency	Importance
1:	id22	0.46077780	0.08223236	0.026066811	0.46077780
2:	id39	0.07031216	0.01705662	0.009687976	0.07031216
3:	num_sin	0.06755848	0.01066015	0.003499336	0.06755848
4:	id46	0.06605265	0.06146208	0.041311603	0.06605265
5:	id40	0.02958264	0.02415433	0.009428766	0.02958264
6:	dif_1	0.02522691	0.01233999	0.006399248	0.02522691
7:	id32	0.02385514	0.09902378	0.080695331	0.02385514
8:	id34	0.02017567	0.07486265	0.077406603	0.02017567
9:	id44	0.01473672	0.03208957	0.032061044	0.01473672
10:	id49	0.01449346	0.05110152	0.058581473	0.01449346

Las variables que muestran una mayor importancia según esta metodología son:

- Id22: Número de pólizas en vigor de todos los ramos.
- Id39: Número de ramos contratados.
- num_sin: Número de siniestros en la póliza del año en curso.
- Id46: Frecuencia histórica de siniestros.
- Id40: Número de pólizas anuladas del cliente.
- dif_1: Diferencia en días desde la ocurrencia del siniestro al cierre.

El hecho de que los modelos no sean interpretables genera dos problemas importantes. El primero sería legal, ya que con la entrada en vigor del Reglamento General de Protección de Datos el pasado 25 de mayo de 2018, sustituyendo a la anterior Directiva de la UE de 1995 de Protección de Datos (DPD), es necesario que los modelos se puedan explicar a petición de los clientes interesados. Es decir, este "derecho a la explicación" obligaría a que sólo se puedan utilizar modelos que se puedan interpretar, y consecuentemente explicar al cliente. Esta limitación es muy importante para poder utilizar determinados algoritmos que son conocidos como "cajas negras".

El segundo problema sería de conocimiento interno. Aunque es un gran paso poder identificar a los clientes con la mayor probabilidad de abandono, esta información no permite que el departamento de marketing pueda diseñar medidas individualizadas para cada cliente. En definitiva, que se conoce que clientes se van a marchar, pero no se sabe cuál o cuáles son los motivos de su abandono. Obviamente, cada asegurado se marchará de la compañía por una razón diferente, ya que las características de los mismos son diferentes. Unos clientes se

marcharán porque han sido mal atendidos en la tramitación de un siniestro, y otros se marcharán por el incremento de las primas de la siguiente anualidad de seguro.

Para solucionar estos dos problemas se están desarrollando una serie de herramientas que permitan la comprensión e interpretación de los modelos más complejos. Actualmente son muchos los algoritmos que se están desarrollando para extraer conocimiento de los modelos, entre las que cabría destacar pdp (Greenwell, 2017), ALEPlot (Apley, 2018), randomForestExplainer (Paluszynska y Biecek, 2017), xgboostExplainer (Foster, 2017) o, live (Staniak y Biecek, 2018).

Como el modelo que ha obtenido el mejor ratio de verdaderos negativos ha sido el xgboost, se utilizará el algoritmo xgboostExplainer desarrollado por David Foster (2017) para determinar cómo se ha obtenido la probabilidad para cada uno de los clientes de la Entidad Aseguradora.

La librería xgboost de R tiene una función (xgb.model.dt.tree) que recoge los cálculos que el algoritmo está utilizando para generar las predicciones. La librería xgboostExplainer extiende estos cálculos para expresar cada predicción individual como la suma de la importancia de cada variable. En este caso, la importancia no es un coeficiente estático como en la regresión logística o la regresión lineal, sino que la importancia de una variable dependerá de la ruta específica que cada asegurado a través del conjunto de árboles irá tomando en función de los valores que toma cada variable para este individuo.

A continuación se muestran dos ejemplos que se han obtenido de las predicciones con los datos balanceados con SMOTE, el primero con una probabilidad inferior al 50% (permanece en la compañía), y el segundo con una probabilidad superior (abandona la compañía).

5.1 Interpretación de los modelos denominados "caja negra"

Si se selecciona un asegurado al azar, se pueden comprobar los valores de las variables del mismo, y posteriormente realizar la predicción con el modelo seleccionado. Los valores que toman las diferentes variables del modelo de clasificación para un individuo concreto se muestran en la siguiente tabla:

Tabla 15. Valores que toman las diferentes variables del modelo de clasificación para un individuo concreto

	id09	id10	id11	id12	id13	id14	id15	id16	id20	id21	id22
15530	CCCC	E	Z	PROYECTA CORREDORES	19	125	57	CASTELLON	LA PLANA BAIXA	1	
	id32	id34	id39	id40	id43	id44	id45	id46	id48	id49	id51	id53	id54
15530	6670.63	3382.58	1	0	1	0.26	0.39	0.5	0	274.21	-194.62	ANUAL	EFECTIVO	1
	id82	id83	id84	id85	id86	id87	id88	id91_2	id91_3	id91_4	id91_5	id91_6	id91_7	id91_8
15530	34.94	0	2002	4184	105	5	G	NO	NO	SI	NO	SI	NO	NO
	edad	carnet	num_sin	total_sin	coste_sin	dia_01	dia_02	dia_03	dif_1	dif_2	dif_3		
15530	70	52	2	325.02	162.51	5	3	7	161	153	10		
	med_dif_1	med_dif_2	med_dif_3										
15530	80.5	76.5	5										

Con estos datos se introducen en la función de la librería xgboostExplainer para extraer la ruta especifica que seguirá a través del conjunto de árboles. A continuación, se muestra el resultado para este asegurado:

```
Extracting the breakdown of each prediction...
|================================================================================| 100%
DONE!
Prediction:  0.2036737
Weight:  -1.36349
Breakdown
   intercept         id14      num_sin        dif_1         id22         id43         id39
-0.546457432 -1.377476216 -1.075111479  1.016117414 -0.901037527  0.886464288  0.763691161
        id34         id44         id85    coste_sin         id53         id16         id88
-0.748940141  0.573658563 -0.564217950  0.552677058  0.541535632 -0.508742425  0.488516868
       dif_3         id46       carnet         id82         id83         id49         id32
-0.479454384 -0.472818878  0.455348182 -0.413119494 -0.370939760  0.364338785  0.324406200
        id10         id15    med_dif_1         id84         id13         edad         id87
-0.286105941  0.273587205  0.266013565  0.231135716  0.213178925 -0.209122364 -0.183538369
        id40         id09         id12         id21         id51       dia_02         id45
 0.181777297  0.168158769 -0.147637316 -0.145566535 -0.127171147 -0.114017661  0.083883221
      dia_01       id91_2       id91_3         id54         id20         id86       id91_6
 0.083118341 -0.062166622  0.055249914 -0.054580586 -0.050240101  0.038824047 -0.033835880
   med_dif_3       dia_03         id11    total_sin         id48       id91_8       id91_5
-0.031189413 -0.026392257  0.025427834  0.021633758 -0.018355230 -0.014300218 -0.013247035
      id91_4    med_dif_2        dif_2       id91_7
-0.012359432  0.007472367  0.006749301  0.001687509
```

La predicción que facilita el modelo xgboost con SMOTE es del 20,36% de abandonar la compañía aseguradora. Conociendo la probabilidad que le asigna el modelo se puede obtener el peso (weight) a través de la función logística:

$$Peso = -ln\left(\frac{1 - Prob}{Prob}\right) = -ln\left(\frac{1 - 0{,}2036}{0{,}2036}\right) = -1{,}3635$$

Este será el peso de la predicción total que se podrá descomponer en función de las variables y los valores que toman. Aparecerá en primer lugar el término contante (intercept) que será igual para todas las predicciones, y que viene a representar el porcentaje de asegurados que abandonaron la compañía aseguradora. Se debe tener en cuenta que estas predicciones se han realizado entrenando un conjunto de datos que ha sido balanceado con SMOTE, por lo que este valor no representa el porcentaje de abandono en los datos desequilibrados. A continuación, irán representándose en un gráfico de cascada las diferentes variables, ordenadas por su importancia y con su peso específico en función del valor de la variable.

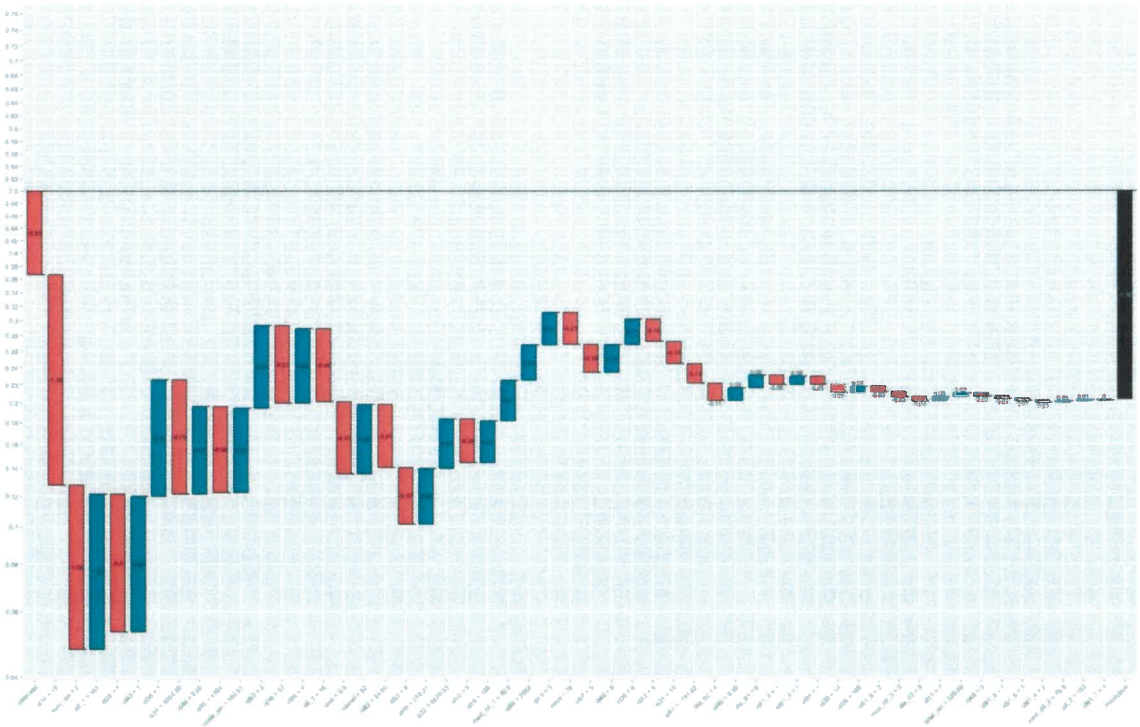
La ponderación para cada una de las variables de este individuo está en función del modelo y de los valores que toma en concreto en este caso. La variable id14 (antigüedad del cliente en años) toma un valor de 19 años para este asegurado y es la variable que tiene para él la mayor ponderación -1,3775. En primer lugar, indicar que el peso de esta variable es negativo, es decir, favorece que el asegurado se quede en la compañía. En el gráfico se representará con una barra roja en dirección descendente.

Si se conoce el peso, se podrá conocer cuál es la probabilidad que aporta esta variable al resultado final. No obstante, se debe tener en cuenta que los pesos se van acumulando, y que la probabilidad de la variable se obtendrá de la ponderación acumulada, tal como se muestra en la Tabla 16.

Como se puede observar en el gráfico, el asegurado tiene una probabilidad inferior al 50% que viene marcada por la recta negra horizontal. El término constante reduce esta probabilidad,

porque en el conjunto de datos SMOTE presenta una mayor proporción de asegurados que no se marchan, y se representa con una barra roja en sentido descendente. Las variables id14 y num_sin (número de siniestros 2) son las primeras en importancia y colaboran negativamente a la marcha del asegurado, también están en color rojo. La primera variable que genera un impacto positivo a la marcha de este asegurado es dif_1 (la compañía tardo 161 días en cerrar el siniestro, es decir, casi medio año) que está representada en verde y en dirección ascendente. Las variables irán perdiendo importancia, ya que están ordenadas todas menos el intercept, y su impacto en el resultado final será menor.

Gráfico 24. Descomposición de la probabilidad del "riesgo de fuga" en las variables del modelo



Fuente: Elaboración Propia. Gráfico generado con RStudio.

La última barra negra es el impacto agregado de todas las variables más el término constante. El peso del conjunto de variables es negativo porque es inferior al 50%, y queda por debajo de la línea negra horizontal.

Finalmente en la siguiente Tabla se muestran los pesos que se han obtenido con esta nueva librería, los pesos acumulados y la probabilidad acumulada para el peso acumulado. La última probabilidad acumulada deberá coincidir con la predicción y el peso acumulado con el peso (weight) que le corresponde.

Tabla 16. Descomposición del peso y la probabilidad

Prediction:	0,2037		
Weight:	-1,3635		
Breakdown			

	Peso	Peso A.	Prob. A.
intercept	-0,5465		36,67%
id14	-1,3775	-1,9239	12,74%
num_sin	-1,0751	-2,999	4,75%
dif_1	1,0161	-1,9829	12,10%
id22	-0,901	-2,884	5,30%
id43	0,8865	-1,9975	11,95%
id39	0,7637	-1,2338	22,55%
id34	-0,7489	-1,9827	12,10%
id44	0,5737	-1,4091	19,64%
id85	-0,5642	-1,9733	12,20%
coste_sin	0,5527	-1,4206	19,46%
id53	0,5415	-0,8791	29,34%
id16	-0,5087	-1,3878	19,98%
id88	0,4885	-0,8993	28,92%
dif_3	-0,4795	-1,3788	20,12%
id46	-0,4728	-1,8516	13,57%
carnet	0,4553	-1,3962	19,84%
id82	-0,4131	-1,8094	14,07%
id83	-0,3709	-2,1803	10,15%
id49	0,3643	-1,816	13,99%
id32	0,3244	-1,4916	18,37%
id10	-0,2861	-1,7777	14,46%
id15	0,2736	-1,5041	18,18%
med_dif_1	0,266	-1,2381	22,48%
id84	0,2311	-1,0069	26,76%
id13	0,2132	-0,7938	31,14%
edad	-0,2091	-1,0029	26,84%

	Peso	Peso A.	Prob. A.
id87	-0,1835	-1,1864	23,39%
id40	0,1818	-1,0046	26,80%
id09	0,1682	-0,8365	30,23%
id12	-0,1476	-0,9841	27,21%
id21	-0,1456	-1,1297	24,42%
id51	-0,1272	-1,2569	22,15%
dia_02	-0,114	-1,3709	20,25%
id45	0,0839	-1,287	21,64%
dia_01	0,0831	-1,2039	23,08%
id91_2	-0,0622	-1,266	21,99%
id91_3	0,0552	-1,2108	22,96%
id54	-0,0546	-1,2654	22,01%
id20	-0,0502	-1,3156	21,16%
id86	0,0388	-1,2768	21,81%
id91_6	-0,0338	-1,3106	21,24%
med_dif_3	-0,0312	-1,3418	20,72%
dia_03	-0,0264	-1,3682	20,29%
id11	0,0254	-1,3428	20,71%
total_sin	0,0216	-1,3211	21,06%
id48	-0,0184	-1,3395	20,76%
id91_8	-0,0143	-1,3538	20,53%
id91_5	-0,0132	-1,367	20,31%
id91_4	-0,0124	-1,3794	20,11%
med_dif_2	0,0075	-1,3719	20,23%
dif_2	0,0067	-1,3652	20,34%
id91_7	0,0017	-1,3635	20,37%

Fuente: Elaboración Propia. Resultados obtenidos RStudio y Microsoft Excel.

Si se realiza un segundo análisis con un asegurado cuya predicción es una probabilidad superior al 50%, los resultados cambiarán completamente.

Si se comparan estos resultados con los obtenidos en el anterior apartado, se verán claramente las diferencias entre ambas predicciones.

Tabla 17. Valores que toman las diferentes variables del modelo de clasificación para un individuo con riesgo de fuga

	id09	id10	id11	id12	id13	id14	id15	id16	id20	id21	id22	id32
17979	BB	F	Z	PROYECTA CORREDORES	3	125	71	CASTELLON	LA PLANA ALTA	0	609.72	

	id34	id39	id40	id43	id44	id45	id46	id48	id49	id51	id53	id54	id82
17979	483.13	0	0	0	0.33	0.33	0.5	0	243.85	-203.68	ANUAL	EFECTIVO	1 32.7

	id83	id84	id85	id86	id87	id88	id91_2	id91_3	id91_4	id91_5	id91_6	id91_7	id91_8	edad
17979	0	2002	4108	71	5	D	NO	NO	SI	NO	SI	NO	NO	57

	carnet	num_sin	total_sin	coste_sin	dia_01	dia_02	dia_03	dif_1	dif_2	dif_3	med_dif_1
17979	29	1	52.09	52.09	5	7	2	19	17	3	19

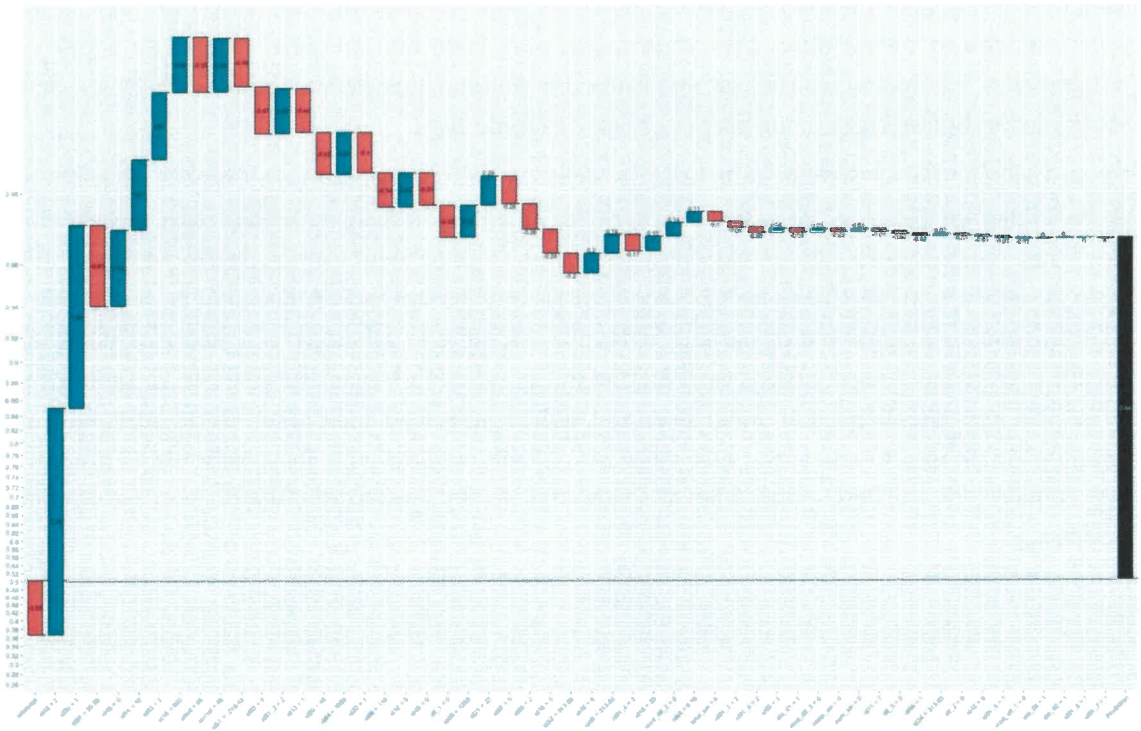
	med_dif_2	med_dif_3
17979	17	3

La importancia de las variables ha cambiado, y ahora las barras verdes ascendentes predominan sobre las barras rojas descendentes. La variable más importante para este asegurado es identificada como la id43 (histórico de siniestros en ramo de diversos que es 0), a continuación id39 (número de ramos contratados que también es 0), id46 (frecuencia siniestral histórica en

automóvil 0,5), id14 (antigüedad del cliente, 3 años), id 53 (forma de pago, anual efectivo), etc. La probabilidad que le ha asignado el modelo xgboost con SMOTE a este asegurado es del 96,88%, por lo tanto, se marchará casi con total seguridad.

Con estos datos se genera el gráfico de cascada que se muestra a continuación:

Gráfico 25. Descomposición de la probabilidad del "riesgo de fuga" en las variables del modelo



Fuente: Elaboración Propia. Gráfico generado con RStudio.

En esta ocasión la barra negra final queda por encima de la línea negra horizontal, mostrando que la probabilidad del riesgo de fuga de este asegurado es muy alta (96,98%). La mayoría de las variables, por los valores que toman, generan un impacto negativo con unos pesos muy elevados. En el gráfico predomina el color verde, y cuando aparece el color rojo es de escasa intensidad, lo que hace que al final los impactos negativos sean muy superiores a los impactos positivos.

5.2 Análisis de resultados para el departamento de marketing

El hecho que se pueda explicar a cada asegurado, si él lo solicitase, como le afecta el modelo en función de los valores de las variables de cada individuo, permite solventar el problema que plantea el nuevo Reglamento General de Protección de Datos. Sin embargo, serán muy pocos los asegurados que se dirijan a la compañía para que les expliquen cómo les afecta a ellos el modelo que se está aplicando.

El departamento de marketing de una aseguradora que esté interesado en retener a aquellos asegurados que son rentables y que obtienen una alta probabilidad de riesgo de fuga, querrá saber cuáles son las variables más importantes para cada uno de ellos, cuál es su signo y su ponderación. El objetivo es realizar una estrategia de retención personalizada en función de los resultados individuales de cada asegurado en riesgo de fuga.

La idea es sencilla, se extraen de la base de datos todos los asegurados que hayan obtenido una probabilidad de riesgo de fuga superior a 50%. Se calculan los pesos para cada uno de ellos. Con estas ponderaciones de las variables y sus signos se generan grupos de asegurados con las mismas características. Como las variables por las que han tomado la decisión de marcharse en un grupo son las mismas o muy parecidas, las acciones que se deberían realizar serían parecidas. Es decir, el grupo de asegurados que tiene como ponderación más alta la de la variable prima de la próxima anualidad con signo positivo, indicaría que se marcharán de la compañía por cuestión de precio. Conociendo que su principal motivación para marcharse es el precio, el departamento de marketing podrá llevar a cabo una acción comercial concreta con ese conjunto de clientes.

Si se toma como referencia no sólo la variable con mayor ponderación, sino 3 ó más variables se podrían establecer perfiles de abandono a los que se podrían asignar acciones comerciales concretas.

No todos los perfiles tienen porque tener acciones con un elevado coste económico (reducción de primas, regalo de algún objeto si renueva la póliza, etc.). El seguro es un servicio basado en la confianza de ambas partes, por lo que el asegurador deberá ganarse esa confianza en todas las ocasiones que entre en contacto con el cliente. Crear un vínculo con el asegurado en base a demostrar un interés verdadero por él, incrementará la fidelidad del cliente y reducirá el riesgo de fuga. No sólo se dispone de incentivos económicos, también hay incentivos sociales e incentivos morales (Levitt y Dubner, 2009) que pueden funcionar igual o mejor que los económicos.

Obviamente, algunos asegurados tendrán una alta probabilidad de abandono, y la compañía aseguradora no hará nada para retenerlos. Es más sabiendo que algunos malos clientes tienen intención de marcharse, se les puede dar un "nudge" final para que se marchen definitivamente (Thaler y Sunstein, 2017). Si el cliente es sensible a la prima y no es rentable para la compañía, se le podrá incrementar la prima para que no le quede ninguna duda a la hora de abandonar la compañía.

Para terminar este apartado, es importante recordar, que una vez el departamento de marketing empiece a realizar incentivos para retener a sus clientes, el modelo que se ha entrenado dejará de ser válido. Como la duración de las pólizas de automóvil son renovables anualmente, se

dispondrá de un año para recopilar la información generada con las acciones comerciales desarrolladas. Durante este año el modelo de riesgo de fuga ajustado seguirá siendo válido, pero una vez, se alcance la siguiente renovación de los primeros asegurados a los que se les incentivo para permanecer o marcharse de la compañía, el modelo estará sesgado y no servirá para clasificar la realidad de ese momento.

Es imprescindible que el departamento de marketing incluya las acciones realizadas con cada asegurado y el resultado de la misma. Este feedback es el que permitirá modificar el modelo y poder continuar realizando una política de marketing adaptada a la realidad de cada momento.

Para poder medir adecuadamente la efectividad de las políticas de marketing acometidas con los individuos en riesgo de fuga, deberá crearse aleatoriamente un grupo de control. El grupo de control no tiene por qué ser muy grande y sobre él no se llevará a cabo ninguna acción comercial. El objetivo es saber cómo se comportan los clientes cuando no tienen ningún tipo de incentivo y de esta forma conocer cuál es la efectividad de las acciones de marketing realizadas.

Siegel (2014) lo expone de la siguiente manera: "Como pasa con la medicina, el éxito en el marketing (o la ausencia de éste) se pone de manifiesto al compararlo con un grupo de control, que es un conjunto de individuos a los que se ha suprimido el tratamiento (o se ha administrado un placebo, en el caso de la medicina). Por consiguiente, necesitaremos recopilar dos conjuntos de datos para poder efectuar el análisis correcto."

Los modelos de respuesta incremental (uplift modelling) tienen en cuenta que un tratamiento o una campaña sólo es eficaz si es mejor que no hacer nada. Es tan importante aprender del grupo sobre el que se realiza la campaña de marketing como del grupo de control. Esto supone un cambio de paradigma, ya que es necesario trabajar con dos conjuntos de datos para poder extraer el máximo de conocimiento.

Al poder interpretar los ensamble de modelos, permitirá que el departamento de marketing pueda realizar acciones individualizadas, y a partir de este momento, se deberá cambiar la forma en la que se miden la eficiencia de estas acciones comerciales. El departamento de marketing tendrá un mayor conocimiento de sus clientes, por lo que su estrategia de actuación ya no podrá ser la de "café para todos". Los avances técnicos y metodológicos obligan a cambiar el paradigma de las políticas de marketing.

6. Conclusiones

Después de efectuados los análisis de los resultados se muestran las conclusiones que se han obtenido:

- Dado el gran número de datos facilitados, la precisión global de los modelos es muy similar, no habiendo diferencias significativas entre ellos. En líneas generales, es más importante la información que el algoritmo utilizado, es decir, es preferible incrementar la muestra analizada que seleccionar un algoritmo más complejo.
- El objetivo fundamental es obtener la mejor predicción posible de los verdaderos negativos, que será lo que ayude al departamento de marketing para realizar un acción comercial más eficiente.
- Los ratios de verdaderos negativos con datos desequilibrados permanecen por debajo del 65%, y únicamente la metodología xgboost es capaz de alcanzarlo.
- La generación de datos sintéticos mediante el algoritmo SMOTE permite balancear las dos clases y evitar que los algoritmos de clasificación primen a la clase mayoritaria.
- El mejor resultado en la predicción de verdaderos negativos se obtiene con algoritmo xgboost y datos balanceados mediante SMOTE.
- El algoritmo identifica correctamente a uno de cada cuatro asegurados en "riesgo de fuga", siendo el porcentaje de 73,21%.
- Los falsos negativos también se incrementan cuando se balancean los datos, por este motivo sería necesario un análisis coste/beneficio de cada una de las soluciones obtenidas.

La decisión final, como casi siempre, deberá ser analizada por la empresa en base a las acciones que se llevarán a cabo para la retención de clientes. La solución tecnológicamente mejor no tiene por qué ser la mejor solución para la empresa, y dependerá de la política general de la misma. Evidentemente, desde el momento en el que se conocen las probabilidades de abandono de cada uno de los asegurados y la importancia relativa de las diferentes variables, es necesario fijar una política de retención para determinar qué asegurados vale la pena retener y cuáles serán las acciones a realizar en base a los resultados obtenidos de cada uno de ellos.

Para este último punto, será de gran ayuda los nuevos algoritmos para interpretar los *ensamble* de modelos. Los resultados obtenidos con el mismo, permiten interpretar la importancia relativa de cada una de las variables para cada asegurado en función de los valores de las mismas. Agregando las importancias relativas se obtiene mediante la función logística la probabilidad del "riesgo de fuga".

Esta personalización para cada asegurado de la predicción de abandono, permitirá conocer cuáles son las variables que tienen una mayor ponderación para cada asegurado, y consecuentemente el departamento de marketing podrá llevar a cabo unas acciones comerciales individualizadas. Esta personalización de las estrategias comerciales incrementará la eficiencia y la eficacia del departamento de marketing.

La introducción de incentivos para la retención de asegurados cambiará el paradigma en la forma de medir la eficacia de las acciones de marketing. Se deberán introducir los modelos de respuesta incremental para dar tratar las nuevas situaciones, siendo necesario incluir un grupo de control para su correcto análisis.

Por último indicar que se podría mejorar la predicción incluyendo fuentes de bases externas que aportasen más información socio-económica de los asegurados. En función del perfil de los asegurados, se puede intentar extraer información de redes sociales para mejorar la información disponible del individuo. Evidentemente sería necesario un perfil de asegurados jóvenes, ya que las personas mayores utilizan las redes sociales de una forma muy limitada. No obstante, la aseguradora debería recopilar información de sus asegurados en diferentes ámbitos para mejorar el conocimiento del colectivo. Aunque si bien es cierto que determinadas variables no se pueden utilizar para la tarificación de los productos, serían de gran ayuda para el análisis en otros campos de interés para la Entidad Aseguradora.

7. Bibliografía

- Ali, Ö.G. y Aritürk, U. (2014). Dynamic churn prediction framework with more effective use of rare event data: The case of private banking. Experts Systems with Applications, 41, pp. 7889-7903.
- Apley, D. (2018). ALEPlot: Accumulated Local Effects (ALE) Plots and Partial Dependence (PD) Plots. R package version 1.1. URL: https://CRAN.R-project.org/package=ALEPlot
- Baesens, B.; Van Vlasselaer, V. y Verbeke, W. (2015). Fraud Analytics. Using Descriptive, Predictive and Social Network Techniques. A guide to data science for fraud detection. New Jeresey, John Willey & Sons.
- Bradley, A.P. (1997). The use of the area under the ROC curve in the evaluation of machine learning algorithms. Pattern Recognition, 30 (7), pp. 1145-1159.
- Breiman, L. (2001). Random Forests. Machine Learning, 45 (1), pp. 5-32.
- Burez, J. Y Van den Poel, D. (2009). Handling class imbalance in customer churn prediction. Experts Systems with Applications, 36, pp. 4626-4636.
- Caballero, R. y Martín, E. (2015). Las Bases del Big Data. Catarata, Madrid.
- Chawla, N. V., Bowyer, K. W., Hall, L. O., Kegelmeyer, W. P. (2002). Smote: Synthetic minority over-sampling technique. Journal of Artificial Intelligence Research, 16, pp. 321-357.
- Chawla, N.V. (2010). Data mining for imbalanced datasets: An overview. In Oded Maimon and Lior Rokach, editors, Data Mining and Knowledge Discovery Handbook, pp. 875-886. Springer US, Boston.
- Chen, C.; Liaw, A. y Breiman, L. (2004). Using random forests to learn imbalanced data. Technical Report 666. Statistics Deparment of University of California at Berkley.
- Chen, T.; He, T.; Benesty, M.; Khotilovich, V.; Tang, Y.; Cho, H.; Chen, K.; Mitchell, R.; Cano, I.; Zhou, T.; Li, M.; Xie, J.; Lin, M.; Geng, Y. y Li, Y. (2018). xgboost: Extreme Gradient Boosting. R package version 0.71.2. https://CRAN.R-project.org/package=xgboost
- Dawes, J y Swailes, S. (1999). Retention sans frontiers: Issues for financial service retailers. International Journal of Bank Marketing, 17 (1), pp. 36-43.
- Foster, D. (2017). xgboostExplainer: XGBoost Model Explainer. R package version 0.1.
- Friedman, Jerome H. (2001). Greedy function approximation: A gradient boosting machine. The Annals of Statistics, 29 (5), pp. 1189–1232.
- Greenwell, B.M. (2017). pdp: An R Package for Constructing Partial Dependence Plots. The R Journal, 9 (1), 421-436. URL: https://journal.r-project.org/archive/2017/RJ-2017-016/index.html.
- Günther, C.C.; Tvete, I.F.; Aaas, K.; Sandnes, G.I. y Borgan, O. (2014). Modelling and predicting customer churn from an insurance company. Scandinavian Actuarial Journal, 2014 (1), pp. 58-71.
- Hidalgo Ruiz-Capillas, S (2014). Random Forests para detección de fraude en medios de pago. Trabajo Final de Máster. Universidad Autónoma de Madrid.

- Huang, Ch-F. y Hsueh, S-L. (2010). Customer behavior and decision making in the refurbishment industry: a data mining approach. Journal of Civil Engineering and Management, 16 (1), pp. 75-84.
- Huigevoort, Chantine (2015). Customer Churn prediction for an insurance company. Eindhoven University of Technology. Master Thesis. https://pure.tue.nl/ws/portalfiles/portal/47019808 [Último acceso: 23 de septiembre de 2018]
- Juran, J.M. (2011). Juran's Quality Control Handbook. 6th Edition. McGraw-Hill, New York.
- Kaymak, U.; Ben-David, A. y Potharst, R. (2012). The AUK: A simple alternative to the AUC. Engineering Applications of Artificial Intelligence, 25 (5), pp. 1082-1089.
- Keramati, A.; Jafari-Marandi, R.; Aliannejadi, M.; Ahmadian, I.; Mozaffari, M. y Abbasi, U. (2014) Improved churn prediction in telecommunication industry using data mining techniques. Applied Soft Computing, 24, pp. 994-1012.
- Kim, K.; Jun, Ch-H. y Lee, J. (2014). Improved churn prediction in telecommunication industry by analyzing a large network. Expert Systems with Applications, 41 (15), pp. 6575-6584.
- Kotsiantis, S.; Kanellopoulos, D. y Pintelas, P. (2006). Handling imbalanced datasets: A review. GESTS International Transactions on Computer Science and Engineering, 30 (1), pp. 25-36.
- Kumar, D. y Garg, A. (2013). A study of Data Mining Techniques for churn prediction. International Journal of Science, Engineering and Computer Technology, 3 (1), pp. 1-4.
- Kursa, M.B. y Rudnicki, W.R. (2010). Feature Selection with the Boruta Package. Journal of Statistical Software, 36(11), 1-13. URL: http://www.jstatsoft.org/v36/i11/
- Larivière, B. Y Van den Poel, D. (2004). Investigating the role of products features in preventing customer churn, by using survival analysis and choice modelling: The case of financial services. Experts Systems with Applications, 27 (2), pp. 277-285.
- Levitt, S.D. y Dubner, S.J. (2009). Freakonomics. Ediciones B, Barcelona.
- Liaw, A. y Wiener M. (2002). Classification and Regression by randomForest. R News 2(3), pp 18-22.
- López, V.; Fernández, A.; García, S. Palade, V. y Herrera, F. (2013). An insight into classification with imbalanced data: Empirical results and current trends on using data intrinsic characteristics. Information Sciences. Vol. 250, pp. 113-141.
- Lunardon, N.; Menardi, G. y Torelli, N. (2014). ROSE: a Package for Binary Imbalanced Learning. The R Journal, 6(1), pp. 82-92.
- Milborrow, S. (2018). rpart.plot: Plot 'rpart' Models: An Enhanced Version of 'plot.rpart'. R package version 3.0.4. URL: https://CRAN.R-project.org/package=rpart.plot
- Nisbett, R.E. (2016). Mindware. Herramientas para pensar mejor. Editorial Debate, Barcelona.
- Paluszynska, A. y Biecek, P. (2017). randomForestExplainer: Explaining and Visualizing Random Forests in Terms of Variable Importance. R package version 0.9. URL: https://CRAN.R-project.org/package=randomForestExplainer

- Provost, F. y Fawcatt, T. (2013). Data Science for Business: What you need to know about data mining and data-analytic thinking. O'Reilly Media, Inc.
- Reichheld, F.F. (1996). Learning form customer defections. Harvard Business Review, 74, pp. 56-69.
- Reichheld, F.F. y Kenny, D.W. (1990). The hidden advantages of customer retention. Journal of Retail Banking, 12 (4), pp. 19-23.
- Rodríguez, P. (2018). Inteligencia Artificial. Cómo cambiará el Mundo (y tu vida). Ediciones Deusto, Barcelona.
- Shmueli, G.; Patel, N.R. y Bruce, P.C. (2011). Data mining for business intelligence: concepts, techniques, and applications in microsoft office excel with xlminer. John Wiley and Sons, second edition.
- Siegel, E. (2014). Analítica Predictiva. Predecir el futuro utilizando Big Data. Ediciones Anaya Multimedia, Madrid.
- Silver, N. (2014). La Señal y el Ruido. Ediciones Península, Barcelona.
- Staniak, M. y Biecek, P. (2018). "Explanations of Model Predictions with live and breakdown Packages." _ArXiv e-prints_. 1804.01955, URL: https://arxiv.org/abs/1804.01955 [Último acceso: 23 de septiembre de 2018]
- Thaler, R.H. y Sunstein, C.R. (2017). Un Pequeño Empujón (Nudge): El impulso que necesitas para tomar las mejores decisiones en salud, dinero y felicidad. Editorial Taurus, Barcelona.
- Therneau, T. y Atkinson, B. (2018). rpart: Recursive Partitioning and Regression Trees. R package version 4.1-13. URL: https://CRAN.R-project.org/package=rpart
- Torgo, L. (2010). Data Mining using R: learning with case studies, CRC Press (ISBN: 9781439810187). URL: http://www.dcc.fc.up.pt/~ltorgo/DataMiningWithR [Último acceso: 23 de septiembre de 2018]
- Torkzadeh, G; Chang, J.C. y Hansen, G.W. (2006). Identifying issues in customer relationship management at Merck-Medco. Decision Support Systems, Vol. 42 (2), pp. 1116-1130.
- Vafeiadis, T.; Diamantaras, K.I.; Sarigiannidis, G. y Chatzisavvas, K.Ch. (2015). A comparison of machine learning techniques for customer churn prediction. Simulation Modelling Practice and Theory, 55, pp. 1-9.
- Xie, Y.; Li, X; Ngai, E.W.T. y Ying, W. (2009) Customer churn prediction using improved balanced random forest. Experts Systems with Applications, 36, pp. 5445-5449.
- Yen, S.J y Lee, Y.S (2009). Cluster-based under-sampling approaches for imbalanced data distributions. Experts Systems with Applications, 36 (3), pp. 5718-5727.