Aplicación de un algoritmo de *machine learning* supervisado para la clasificación de conductores jóvenes. Identificación de factores de riesgo

JUAN MANUEL LÓPEZ ZAFRA

Actuario. Dr. en CCEE. Prof. Titular de estadística e Inv. Operativa
Co-Director del Master de Data Science para Finanzas. CUNEF. jmlopezzafra@cunef.edu

SONIA DE PAZ COBO

Actuario. Dra. En CCEE. Prof. Contratado Doctor en Economía Aplicada.
Fac. CC Jurídicas y Sociales. URJC. Sonia.depaz@urjc.es

RICARDO A. QUERALT

Dr. en CCEE. Máster en Hacienda Pública por IEF
Co-Director del Master de Data Science para Finanzas. CUNEF. ricardo.queralt@cunef.edu

Las técnicas de *machine learning* se emplean desde hace tiempo para aprovechar el conocimiento que se encuentra en grandes bases de datos. Se distinguen entre supervisadas y no supervisadas, entendiéndose por las primeras aquéllas que persiguen conocer las relaciones causa-efecto entre variables y las segundas aquellas otras que buscan establecer las relaciones subyacentes entre las variables y/o las observaciones. Nuestro objetivo es mostrar, de forma somera, cómo puede emplearse un algoritmo supervisado para clasificar conductores en virtud de sus características. En concreto, nos enfrentábamos al problema de establecer qué distingue un buen conductor, o conductor de mínimo riesgo, de un mal conductor, o uno de máximo riesgo. Para ello, empleamos datos reales de más de 150.000 conductores monitorizados mediante una baliza que recoge características, por cada desplazamiento, de velocidad (máxima y media), tiempo al volante, distancia recorrida, día de la semana y horario de conducción o tipo de vía. De esos conductores, anónimos, conocemos además su sexo, edad (todos menores de 30 años), la antigüedad de su permiso de circulación, la marca y modelo del vehículo asegurado, y su potencia, además del número de partes de siniestro entregados en el período objeto de estudio. La base de datos empleada contiene más de 42.000 millones de datos, correspondientes a más de 400 millones de desplazamientos con más de 2,8 billones de kilómetros estudiados.

Para entrenar al algoritmo se empleó alrededor de un 80% de la base de datos de conductores, que una vez depurada se estableció en alrededor de 90.000 de los 150.000 iniciales. El grupo de comprobación se fijó en unos 17.000 conductores. Desde el punto de vista de la accidentalidad, se establecieron cuatro grupos: los conductores que no habían presentado ningún parte de siniestro en el período de análisis, los que habían presentado sólo uno, los que sólo presentaron dos partes y todos aquellos que presentaron tres o más partes.

Es interesante observar que, siendo la edad un factor de discriminación tradicional en el seguro de automóvil, sólo se dieron diferencias significativas entre quienes presentaron tres o más partes respecto de quienes no presentaron ninguno, casi cuatro meses mayores (recordemos que se trataba de conductores menores de 30

Nuestro objetivo es mostrar, de forma somera, cómo puede emplearse un algoritmo supervisado para clasificar conductores en virtud de sus características

años). Resulta también interesante observar la distribución de frecuencias de la "edad" de los asegurados como conductores, comparada con la biológica: mientras que la antigüedad media de la licencia no llega a los 4 años, el tiempo medio vivido por encima de los 18 años supera los 6; y si el 75% de las licencias más recientes tiene 5 años, el de los conductores más jóvenes ha alcanzado ya los 8 años posteriores a la mayoría de edad. Esta situación es independiente del sexo del conductor, pues el reparto de los más inexpertos conductores es prácticamente idéntico en ambos sexos.

Entrando ya en la clasificación, optamos por implementar el CHAID. Este algoritmo de segmentación se basa en las técnicas de los árboles de decisión. Desarrollado por Gordon V. Kass, es el acrónimo de "CHi-squared Automatic Interaction Detector" o Detector Automático de Interacciones chi-cuadrado; se trata de una ampliación de los algoritmos AID (Automatic Interaction Detector) y THAID (THeta Automatic Interaction Detector).

Su ámbito principal de aplicación es el márketing, aunque su uso, como ponemos de manifiesto, es extensible a la explotación de cualquier tipo de base de datos. Es una técnica de segmentación poderosa que construye árboles de decisión no necesariamente binarios, siendo particularmente útil en todos aquellos problemas en que se quiera subdividir una población a partir de una variable dependiente y posibles variables predictoras que modifiquen esencialmente los valores de la variable dependiente en cada uno de los segmentos.

Además de para segmentar, esta técnica de *machine learning* se ha empleado para conocer cuáles, de entre el centenar de variables o factores potenciales de riesgo que hemos manejado, son las más significativas; es en este terreno un arma muy poderosa para poder separar *el polvo de la paja*. Asimismo, permite comprender el orden de importancia de los factores de riesgo en la caracterización de la siniestralidad. Permite igualmente entender cómo se modifican de forma recíproca ciertos factores de riesgo.

Los conductores que presentan un número de siniestros no superior a dos tienen una antigüedad de carné estadísticamente igual, que podemos fijar en el entorno de los 3 años y 10 meses. Sin embargo, en cuanto presentan tres o más accidentes, la antigüedad del carné cae hasta los 3 años y 5 meses; esos 5 meses pueden parecer muy pocos; sin embargo, es necesario tener en cuenta el contexto en el que nos movemos: se trata de conductores con una antigüedad de 41 meses (3 años y 5 meses), en los que 5 meses adicionales supone incrementar su experiencia en más de un 12%. Algo que creemos no es en absoluto despreciable, dadas las circunstancias.

Uno de los principales hitos de nuestro trabajo fue el desmitificar la relación peso-potencia como factor incremental de la siniestralidad. El 25% de los coches con menor relación peso-potencia (los más potentes y ligeros) presentó una siniestralidad media que no era significativamente superior a la del resto de tramos; o, lo que es lo mismo, que no puede afirmarse que tengan más accidentes quienes conducen los coches con me-

nor cociente peso-potencia. Esta situación se hace aún más patente cuando analizamos el valor extremo de esta variable, y nos centramos en el uno por ciento de los coches más potentes. Este grupo de conductores presenta una siniestralidad media idéntica a la del resto de conductores. En cambio, en cuanto a la antigüedad del carné se refiere, los primeros conductores son quienes mayor valor presentan, más de cuatro años de experiencia frente a 7 meses menos de antigüedad media entre quienes usan los coches más pesados y menos potentes. Como hemos comprobado a lo largo de la investigación, son los más jóvenes (y con menor antigüedad del carné y menor experiencia) quienes emplean los coches más pesados y menos potentes. En cuanto a quienes conducen el 1% de los coches más potentes, la antigüedad media de su licencia es 9 meses mayor que la del resto (4 años y medio por 3 años y 9 meses).

Una de las principales conclusiones de nuestra investigación es la gran relación entre siniestralidad declarada y uso del vehículo. Efectivamente, un incremento en el uso diario del coche provoca inmediatamente un incremento en la siniestralidad (siempre hablando en términos medios). Así, el 25% de quienes menos usan el coche lo emplea menos de la mitad de veces que el 25% que más los usa (2,45 desplazamientos diarios frente a 5,25); y declara asimismo la mitad de accidentes (0,64 frente a 1,27). Son, quienes menos usan el coche, claramente mayores que quienes más lo usan (más de 24 años y medio por menos de 23 años y nueve meses). Y, además, la antigüedad de la licencia de quienes más usan el coche, en consonancia con su edad, es claramente inferior que el resto: menos de tres años y cinco meses por casi cuatro años los demás. Así pues, no sólo son más jóvenes, también más inexpertos quienes más usan el coche.

El efecto se amplifica enormemente cuando analizamos el 1% de los conductores que más usa el coche. El número medio de usos diarios es aún mayor (8,7 frente a 3,6 del 99% restante), su siniestralidad declarada casi dos veces y media mayor (2 siniestros frente a 0,89), y la edad casi año y medio más baja (22 años y 9 meses frente 24 años y casi cuatro meses). Así pues, no es de extrañar que la antigüedad media de su licencia sea también muy inferior, dos años y siete meses por tres años y casi diez meses los demás; más de un año de diferencia.

Como elemento claramente distintivo entre quienes tienen mayor o menor riesgo de accidente aparece siempre el número medio de trayectos diarios, esto es, la intensidad de uso del vehículo privado. A mayor cantidad de recorridos diarios, permaneciendo constantes el resto de factores, mayor incremento de la posibilidad de accidentes. Junto con esa característica se observa que el incremento en el número de desplazamientos va acompañada, habitualmente, y como es normal esperar, de una disminución en la distancia media recorrida en cada trayecto; una reducción muy significativa en la edad media de los conductores involucrados (especialmente intensa en el caso de los hombres), con el consiguiente incremento de la participación relativa de este grupo de conductores en el total del grupo, y una modificación reseñable en los patrones de comportamiento vial respecto del conductor medio: el conductor de riesgo tienda a circular mucho más por carreteras locales y ciudades y menos por la consideradas de mayor seguridad (vías con límites de 100 ó 120 kph). El incremento de la siniestralidad declarada, además de con los factores anteriores, está claramente ligado a un uso constante del vehículo (independientemente del sexo, el grupo de conductores con tal calificación siempre ve incrementada su presencia relativa en los grupos de riesgo). Por último, es destacable asimismo el incremento de la presencia masculina según se incrementa la posibilidad de accidente.