SEGUROS DE VIDA

TÉCNICAS DE CLASIFICACIÓN

Beltrán Aller López

24/11/2019

CONCLUSIONES

A raíz de los resultados obtenidos, se comprueba la existencia de ciertas variables de gran relevancia a la hora de prever si un hogar es poseedor de un seguro de salud privado. El número de hijos menores de 14 años, así como los ingresos del sustentador principal y, los ingresos en general del hogar; resultan variables de gran importancia. Existen otras variables que también explican en buena medida la posibilidad de poseer un seguro de vida privado, como por ejemplo la edad o el nivel de estudios.

La tenencia o no de un seguro de vida privada se relaciona directamente con variables de carácter económico (nivel de ingresos), carácter social (nivel educativo) y demográfico (lugar de residencia y edad). Estas contribuyen positivamente a la probabilidad de poseer un seguro de vida privado frente a no poseerlo.

El resultado del modelo presenta una precisión del 91% aproximadamente, sin embargo, la tasa de verdaderos positivos predicha resulta baja, por lo que convendría una reformulación del modelo.

RESUMEN

El dataset está compuesto por un total de 16 variables y 21.395 observaciones. Los datos resultan obtenidos de la Encuesta de Presupuestos Familiares correspondiente al año 2018. El objetivo es predecir la tenencia o no de un seguro de salud privado.

DESARROLLO

El primer paso, previo a realizar o estimar cualquier tipo de modelo es la limpieza del dataset. Se comprueba la existencia de valores omitidos o cuya codificación resulte ilógica respecto a lo esperado. Realizado esto, obtenemos un dataset de 20.020 observaciones.

En principio, incluiré todas las variables presentes, no se desea a priori perder porcentaje alguno de varianza explicada. Una vez llevado a cabo el cálculo del modelo, se observará que variables son más importantes en la predicción y cuales menos, pudiendo realizarse una reestimación del modelo con un número distinto de variables.

Algunas de las variables son susceptibles de ser recodificadas, por ejemplo, el tipo del hogar o el régimen de tenencia de la vivienda.

En el primer caso, se procede a una recodificación de la variable de tal forma que, de 12 posibles valores, pasamos a únicamente 5. Se agrupa en hogares unipersonales, hogares en pareja sin hijos, en pareja con hijos y otros hogares.

Para el régimen de tenencia de la vivienda se opta por una recodificación en 3 grupos, si se posee la vivienda en propiedad, ya sea hipotecada o no; si esta se encuentra en alquiler o si se trata de una cesión.

En cuanto al encoding del resto de variables de dataset, observamos un ordinal encoding, es decir, algunas de las variables si bien pueden ser susceptibles de transformación a variables dicotómicas, se aprecia una estructura lógica subyacente en estas. Ejemplifico con el nivel de estudios.

En la variable nivel de estudios se aprecian 4 posibles valores, inferior a educación secundaria, primera y segunda etapa de educación secundaria y estudios superiores, calificadas de 1 a 4 en ese orden respectivamente. Podrían crearse 4 variables dicotómicas que adoptasen los valores 1 y 0 en función de si se encuentran ubicado en tal nivel o no, pero en este caso, la estructura inherente de la variable ya nos aporta esa información, además en un orden lógico, el menor valor, es decir, 1; se corresponde con el menor nivel de estudios y, el mayor, 4; con el mayor nivel de estudios.

Al igual que ocurre con el nivel de estudios, pasa lo mismo con los ingresos, por ejemplo.

Una vez recodificado el dataset, se procede a la estimación de una regresión logística, se aprecia que variables como el sexo o, el tipo de hogar carecen de relevancia en el modelo. Otro tipo de variables como el nivel educativo o los ingresos, resultan claves.

Un dato de gran importancia será la interpretación de los odd ratio, los cuales se muestran a continuación:

Tabla 1: Odd ratio

TAMAÑO DEL MUNICIPIO	0.88
N.º MIEMBROS DEL HOGAR	0.96
N.º MIEMBROS <14 AÑOS	1.27
N.º MIEMBROS OCUPADOS	1.07
TIPO DE HOGAR	0.91
EDAD	1.02
SEXO	1.04
NIVEL EDUCATIVO	1.34
OCUPACION PRE-ENCUESTA	1.06
INGRESOS SUSTENTADOR PRINCIPAL	1.25
TENENCIA VIVIENDA	0.92
SUPERFICIE VIVIENDA	1.75
INGRESOS FAMILIA	1.27

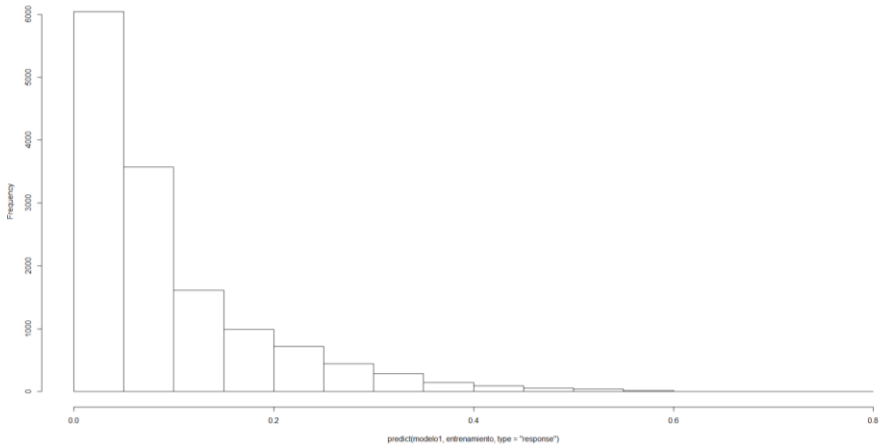
Aquellos odd ratio superiores a uno, indican que un incremento de una unidad en esa variable aumenta la probabilidad de poseer un seguro de vida privado frente a no poseerlo. Mientras que aquellos odd ratio inferiores a 1, indican lo contrario.

Por ejemplo, de un municipio de 100.000 habitantes o más, a un municipio de entre 50.000 y 100.000 habitantes, el efecto sería 0.88^2, lo que da como resultado 0.77, es decir, la probabilidad de tener seguro privado frente a no tenerlo se reduciría un 23%. Si atendemos a otra variable, por ejemplo, el nivel educativo; pasar de un nivel educativo inferior a la secundaria, al siguiente, primera etapa de la secundaria, el efecto sería 1.34^2, lo que da como resultado 1.79. Esto quiere decir que la probabilidad de tener un seguro privado frente a no tenerlo se incrementa un 79%.

En verdad lo que interpretamos es e^beta, de aquí obtenemos las variaciones en la ventaja relativa.

Represento un histograma de la predicción del modelo, con ello observo donde marcar el cut off y asignar a la probabilidad de cada individuo un resultado final, 1 o 0; poseer seguro privado o no poseerlo.

Ilustración 1: Histograma frecuencia de probabilidades



Una vez visualizado el histograma establezco un cut off en 0.4, es decir, cualquier individuo con una probabilidad mayor de 0.4 doy por hecho que tiene seguro privado, en caso contrario no.

Una vez establecido el cut off, se procede a la predicción del modelo sobre la muestra de validación, el resultado es la siguiente matriz de confusión.

Tabla 2: Matriz de confusión

	Predicho	
Actual	Sin seguro	Con seguro
Sin seguro	5431	59
Con seguro	477	39

El modelo calculado presenta una precisión del 91% aproximadamente en las predicciones sobre la muestra de entrenamiento. Sin embargo, el resultado es engañoso; puesto que, aunque la tasa de verdaderos negativos es alta, la de verdaderos positivos resulta baja.

Bibliografía

James, G., Witten, D., Hastie, T., & Tibshirani, R. (2013). *An Introduction to Statistical Learning.* New York: Springer.