ÁRBOLES DE CLASIFICACIÓN

Máster en Data Science

Jesús Cerro de los Santos

Marta Divassón Carribero

Fecha: 16 diciembre 2018

Introducción y Resumen

El objetivo de este informe es plantear un modelo de predicción a partir de un árbol de decisión que permita visualizar el modelo desde una perspectiva diferente a la tradicional Esto es, de manera supervisada, se busca la clasificación de una variable dependiente a partir de el conjunto de variables explicativas de la base de datos. La respuesta del modelo es continua, por lo que se califican como árboles de regresión. Para ello utilizamos dos conjuntos de datos: uno de entrenamiento y otro de validación y así poder predecir sobre ellos para calcular la efectividad del modelo.

El objetivo de este análisis es identificar las combinaciones de variables (las ayudas a la familia, la renta percibida por los menores, la capacidad de enfrentarse a gastos imprevistos, la capacidad de llegar a fin de mes, los miembros que constituyen el hogar, la edad del mayor de la casa, las horas de trabajo semanales y aquellos que son mayores de 16 años) que predicen mejor la asignación de cada individuo a la categoría de tener el riesgo de caer o no en pobreza.

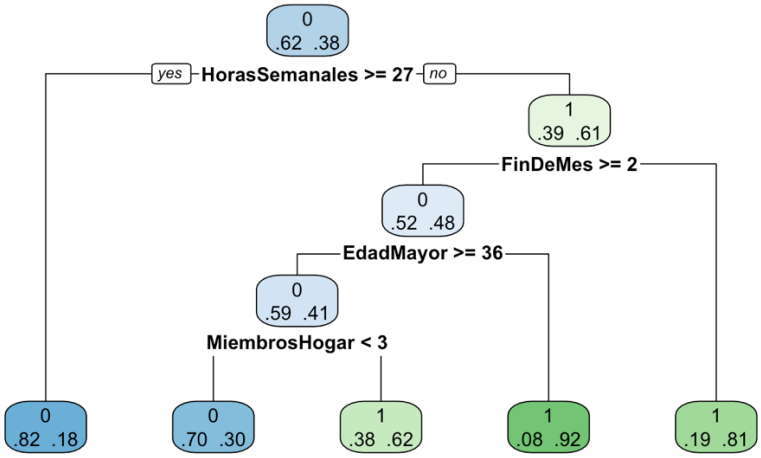
Análisis

A partir de la constitución de un conjunto de entrenamiento, formado por el 60% de los datos y un conjunto de validación (que constituye el 40% restante) se crea un árbol de decisión con la variable "*vhPobreza*" que indica aquellos hogares en riesgo de pobreza.

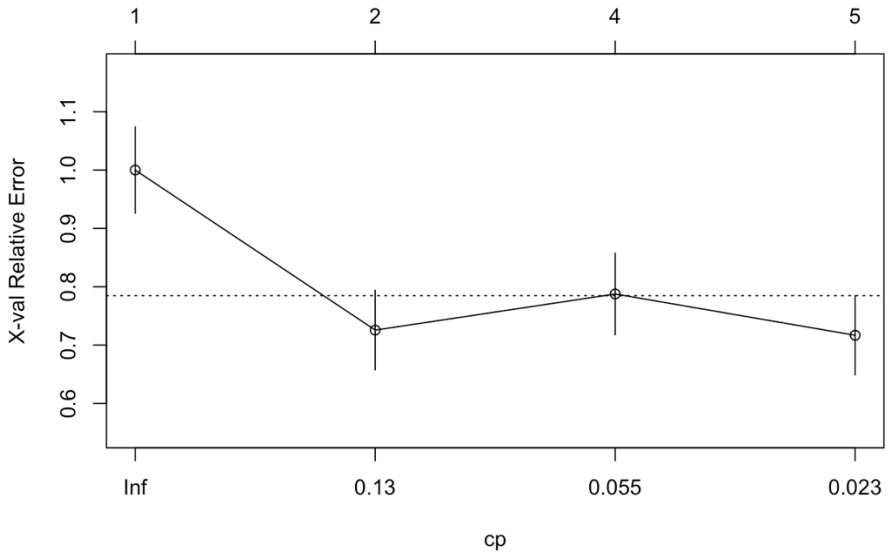
Al graficar el árbol, obtenemos que del conjunto de 298 observaciones (del conjunto de entrenamiento), obtenemos un corte (Split) de 113 observaciones que tienen riesgo de incurrir en pobreza. A estas les corresponde 38% de incurrir en pobreza, mientras que un 0.62080537 de ellos no incurren en pobreza según los datos analizados. No obstante, el mejor corte del árbol se encuentra en aquellas personas que trabajan mas de 27 horas semanales, a la que le corresponde la mayoría de los datos (0.81987578 u 81.98%) como aquellos que no tienen riesgo de incurrir en pobreza.

En la siguiente representación grafica, observamos que el 38% de la muestra corresponde a que son hogares con riesgo de caer en el umbral de pobreza mientras que el 62% no corre ese peligro. Sobre esos datos, a partir de responder la pregunta de "¿Trabajan 27 o más horas semanales?". En el caso de que se cumpla esa condición entonces obtenemos que la mayoría (82%) no caerá en pobreza, mientras que el 18% lo hará. En la rama derecha del árbol, es decir, si no trabaja 27 o más horas semanales entonces la probabilidad de que no tengan riesgo de pobreza es del 39%, mientras que la de tener riesgo de pobreza es del 61%. El resto de los datos se ve de igual forma, es decir, a medida que

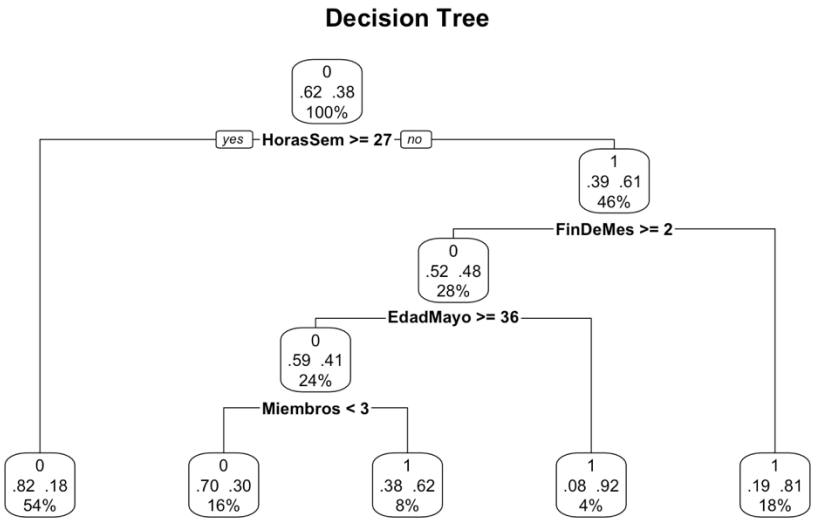
haya menos dificultad para llegar a fin de mes, entonces el riesgo de pobreza es del 52%, si hay cierta dificultad entonces el riesgo de pobreza es del 81%. Dentro de aquellos con mas facilidad para afrontar los gastos a fin de mes, aquellos de 36 a más mayores no caerán en pobreza con una probabilidad del 59%, mientras que si son menores si lo harán con una probabilidad del 92%. Finalmente, si los miembros del hogar son menos de 3 en total, no correrán ese riesgo con un 70% mientras que aquellos que sean más si la tendrán con un 62%.



Cuando la información es de profundo análisis, se procede a la poda del árbol en el punto en el que el modelo empieza a incrementar el error. Es decir, el punto en el que el error deja de disminuir es el punto ideal para la poda del árbol. Se lleva a cabo, observando los cortes del árbol y sobretodo los "cp". El error aumenta entre el 2 y el 3, como se observa en el grafico siguiente y el cp que le corresponde a esto es de un 0.053097.



Tras realizar la poda se hace un grafico en el que se reflejan las estadísticas anteriores y se asigna una probabilidad a cada rama del árbol. Las probabilidades de cada una corresponden al porcentaje de la probabilidad anterior, es decir del 100% una parte de la muestra, el 46% se asigna a no y un 54% se asigna a si (sobre las horas semanales).



Posteriormente, se evalúa la capacidad de predicción del árbol, importante para conocer el ajuste del modelo. Al predecir sobre los datos iniciales y sobre los datos de validación sobre la variable a predecir, se obtiene que, sobre los datos verdaderos, 41 se han clasificado bien como sin riesgo a caer en pobreza mientras que 15 que si tienen riesgo se han clasificado como que no tienen riesgo. Por otra parte, de los que no tienen riesgo, 32 han sido clasificados como que si lo tienen y 91 de ellos han sido bien clasificados.

Matriz de Confusión	**0**	**1**
0	**91**	**32**
1	**15**	**41**

Con estos datos se lleva a cabo la estadística del modelo, que implica la evaluación de la efectividad del modelo a partir de la división entre el total del numero de aciertos entre el numero de predicciones. Al calcular el porcentaje de como acierta el modelo se obtiene que ha clasificado de manera efectiva con un 73.74302 %.

Conclusión

Si comparamos los datos obtenidos con el modelo lineal de regresión logística frente a los obtenidos con la elaboración del árbol de decisión, se observa que en el primero clasifica con una exactitud del 71.73% frente a un 73.74% de exactitud en el segundo caso. Se ha demostrado en este ejemplo que el modelo más exacto es el segundo ya que clasifica de forma más acertada y es el mismo el que se encarga de asignar probabilidades por los criterios que crea convenientes que suelen ser más acertados. No obstante, hay que revisar los criterios de clasificación en el segundo modelo ya que puede que el análisis exploratorio de datos no se haya realizado correctamente y algunos datos no se asignen correctamente. Asimismo, una de las ventajas de este método es la anticipación a la hora de tomar decisiones, es decir, si se pretende predecir sobre unos datos para obtener una solución concreta entonces se puede ver de antemano cual es el coste de esa decisión siempre sujeto a una serie de suposiciones, así como, nos permiten observar cómo de balanceados están los datos.