INFORME ENCUESTA CONDICIONES DE VIDA 2016

Resumen ejecutivo

Este informe continua la investigación realizada en el anterior (ver aquí) sobre factores que inciden en el riesgo de pobreza y mecanismos para la predicción de dicho riesgo. En esta ocasión se emplean árboles de decisión con el objetivo de explicar de manera más nítida y transparente cómo se distribuye el riesgo de pobreza en España según la base de datos a la que se ha tenido acceso. Esta base de datos estudia la situación de 477 familia en las que se estudian dieciséis variables distintas a las que se añade la variable clasificadora "Riesgo de pobreza" que indica si la familia se encuentra o no en riesgo de pobreza. Finalmente se obtienen distintos árboles todos ellos con una tasa de acierto similar a la de la regresión logística (en torno al 70%) pero mucho más sencillos de interpretar.

El análisis previo inicial es análogo al del informe antes citado.

Se comienza este análisis explicando brevemente las distintas variables presentes en la base de datos así como los posibles valores que estas pueden tomar. La variable a predecir es el riesgo de pobreza y es dicotómico, al igual que la posibilidad de irse de vacaciones una semana al año o la capacidad para afrontar gastos imprevistos. Otras variables categóricas son el género de la persona de mayor edad en la unidad familiar, su ocupación, su régimen de tenencia de la vivienda (propia, hipotecada, en alquiler o cedida), la facilidad para llegar a fin de mes (medida en seis grados), la región donde habitan y la posesión de ordenador o TV a color (según se tenga, no se deseé o no sea posible adquirirla por el hogar). Como variables cuantitativas aparecen la renta de la familia en el año anterior, la cuantía de las ayudas recibidas, la renta neta percibida por los menores de 16 años, el número de miembros de la familia, el número de miembros adultos, la edad del mayor y el número de horas de trabajo semanal del conjunto de miembros del hogar.

Antes de comenzar a trabajar se eliminan dos variables; el código de identificación de hogar (no aporta ninguna información pues no se desean conocer resultados para individuos concretos) y la renta percibida durante el año anterior a la entrevista (el riesgo de pobreza se calcula como el 60% de la mediana de los ingresos anuales por unidad de consumo, así no tiene sentido emplear la renta como variable predictiva pues se incurriría en una definición circular).

Tras la supresión de estas dos variables se transforman en factores las variables categóricas para facilitar el trabajo con ellas en R. Se comprueba a continuación que todos los datos están correctamente cargados y que no existen huecos (no hay valores NA).

Tal y como se vio en el informe anterior las variables que presentan una mayor capacidad explicativa sobre el riesgo de incurrir en pobreza de la persona son: la ocupación, el régimen de tenencia de la vivienda y la capacidad para afrontar gastos imprevistos.

En este informe se plantean cinco árboles de clasificación diferentes: dos de ellos construidos sobre todas las variables siendo uno un árbol de decisión estándar y el otro un árbol de decisión basado en inferencia. Por otra parte se construyen de nuevo estos dos tipos de árboles para las variables seleccionadas mencionadas anteriormente. La idea del último árbol surge a partir de las conclusiones y se plantea como un árbol construido única y exclusivamente a partir de variables medibles de manera objetiva.

Con objeto de medir la bondad de los distintos árboles se emplea un conjunto de entrenamiento para ajustar el árbol y un conjunto (disjunto) de testeo para valorar la calidad de los distintos resultados. El método para comparar la bondad de estos árboles será la proporción de familias bien clasificadas respecto al total, es decir, se considerará como un error igual de grave un falso positivo (familia a la que se le asigna riesgo sin estarlo) como un falso negativo (familia considerada por el árbol fuera de riesgo cuando realmente sí lo está). El autor considera que un falso negativo es, en este caso, un error mucho más gravoso porque pone en riesgo la estabilidad económica de una familia, sin embargo esta idea se deja a la consideración del lector. Así se crea

una muestra con el 70% de las observaciones para entrenar el modelo y se deja aparte un 30% de familias empleadas para la validación.

Por último y previo a la construcción de los distintos árboles es esencial comprobar que las muestras (de entrenamiento y testeo) están bien balanceadas. Los resultados de esta comprobación devuelven en este experimento concreto (replicable pues se ha fijado una semilla aleatoria previa a su realización) una muestra de entrenamiento formada por 133 familias en riesgo de pobreza y 200 que no corren riesgo y una muestra de validación formada por 53 familias en riesgo y 91 sin él luego las condiciones de las muestras son oportunas para la realización del experimento.

Una vez realizada esta serie de comprobaciones previas se procede a la construcción de los árboles. El proceso para la construcción de los distintos árboles de decisión es el siguiente:

Se comienza construyendo el árbol de decisión para todas las variables disponibles. Para ello se genera un primer árbol con todas las variables. Dado que el árbol es demasiado grande e incluye variables muy poco significativas, se lleva a cabo una poda que pueda por una parte mejorar su poder predictivo y sobre todo simplificar la interpretación de dicho árbol (no debe olvidarse que una de las principales ventajas de los árboles de decisión frente a otros modelos es su facilidad para ser interpretados). A la hora de decidir cómo podar se lleva a cabo un proceso de validación cruzada estableciendo el punto de poda en el lugar donde se minimiza el error (esto es orientativo; en ocasiones se tolerará un mayor error en aras de obtener árboles más sencillos). Para calcular dicho punto se disponen de dos herramientas: por una parte el propio coeficiente de error proporcionado por R y por otra la construcción de un gráfico que asocie el CP con el error relativo. La **Figura 1** muestra un ejemplo de dicho gráfico asociado al árbol de decisión para las variables seleccionadas en el informe anterior.

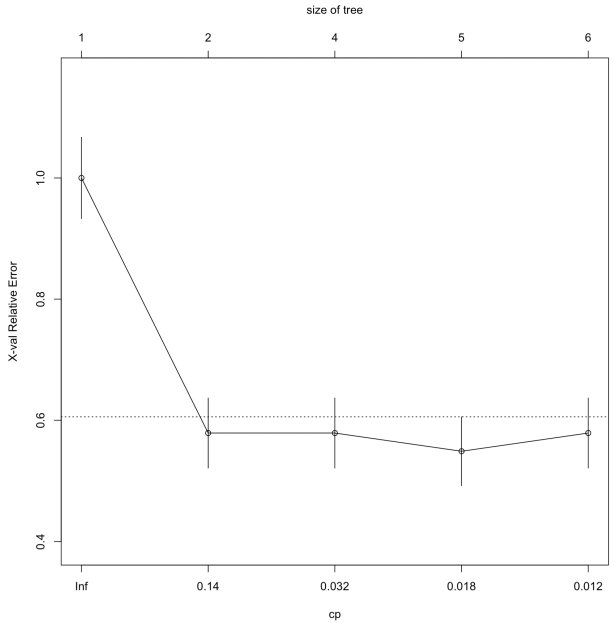

Figura 1. Gráfico para la elección de la altura de poda.

Para este caso se elige un CP con valor de 0.01503759 que minimiza el error de validación cruzada y proporciona unos resultados bastante satisfactorios como se observará más adelante.

Una vez elegido el punto de poda se procede a esta misma obteniendo un nuevo árbol de decisión mucho más sencillo que el anterior. Sobre este árbol podado se realizan las predicciones en el conjunto de validación para calcular la tasa de acierto asociadas a cada árbol.

Así a continuación se presenta una breve comparativa de los árboles obtenidos, su efectividad y un mayor desarrollo del mejor árbol:

En un primer lugar se construye un árbol a partir de todas las variables disponibles. Previo a la poda el árbol tiene una interpretación bastante compleja y remite acierto en 101 familias de un total de 144 (es decir, un 70,14%). Una vez podado éste, se obtiene un nuevo árbol que remite prácticamente la misma precisión (cien familia sobre el total, un 69,44%) y con una interpretación mucho más sencilla. (Este árbol permite clasificar a cada familia con un máximo de cuatro disyunciones). Este árbol se puede observar en el código adjunto y tiene una interpretación análoga a la siguiente.

Tras ello se construye un nuevo árbol de decisión empleando únicamente las variables que se concluyeron como relevantes en el informe anterior, es decir, sector de ocupación, número de miembros de la familia, dificultad para llegar a fin de mes, dificultad para afrontar gastos imprevistos y régimen de tenencia de la vivienda familiar. El árbol sin podar obtenido clasifica bien 105 familias (es decir, tiene una tasa de acierto del 72,92%). Tras la poda se obtiene un árbol que clasifica bien 101 familias, sin embargo al comparar ambos árboles se aprecia que la calidad de interpretación es la misma con lo cual se decide conservar el árbol sin podar que se muestra y explica en detalle a continuación (**Figura 2**) por ser el que una mayor tasa de acierto presenta:

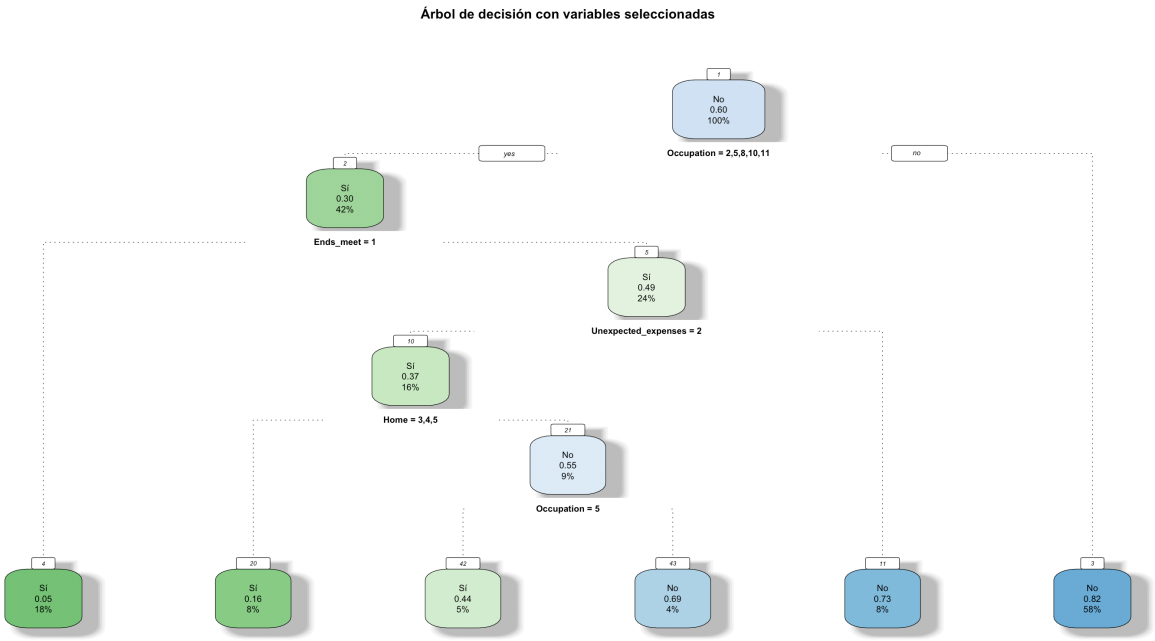

Figura 2. Árbol de decisión a partir de variables seleccionadas.

El árbol dibuja en azul los nodos en los que mayoritariamente no hay riesgo de pobreza y en verde los nodos en los que sí existe riesgo de pobreza. A mayor intensidad del color mayor es la pureza del nodo, es decir, en un nodo azul intenso como el nodo tres (esquina inferior derecha) los individuos se clasifican como fuera de riesgo (y el total de individuos de este nodo fuera de riesgo excede el 80%). En la esquina izquierda por el contrario los individuos (un 95% de los pertenecientes al nodo) se encuentran en riesgo de pobreza. Es interesante que precisamente estos dos nodos sean los más puros porque son los que concentran una mayor parte de población (un 58% y un 18% respectivamente).

La interpretación para este árbol se realizaría de la manera siguiente:

Si se elige un individuo al azar en un principio es más probable (por cómo está elegida la muestra) que se encuentre fuera del riesgo de pobreza. En primer lugar se señalan una serie de sectores que se podrían denominar de riesgo. Si la ocupación se encuadra en los sectores de silvicultura, extracción de antracita, hulla, lignito, piedra, arena o arcillas, industria de la alimentación o de la fabricación de bebidas, es necesaria una mayor inspección de los individuos, sin embargo, si trabajan en otros sectores (esto supone un 58% de la muestra) lo más probable es que se encuentren ajenos a este riesgo. Si además de trabajar en estos sectores sufren serias dificultades para llegar a fin de mes entonces se puede afirmar con una seguridad bastante alta que este hogar se encontrará en riesgo de pobreza. De no ser así, se considera igualmente que la familia se encuentra en riesgo de pobreza pero se comprueba las dificultades de la familia para afrontar gastos imprevistos. Si no presentan una gran dificultad se considera que el hogar está fuera de riesgo, si por el contrario presentan dificultades se continúa afirmando que está en riesgo de pobreza ahora con más firmeza. La siguiente variable a observar es el régimen de tenencia si la familia se encuentra en alquiler, hipoteca o cesión de vivienda se afirma con gran certeza que se encuentra en riesgo de pobreza en caso contrario se profundiza en su ocupación; si trabaja en extracción de carbón es más probable que esté en riesgo que en caso contrario.

Es curioso que si se procede a la poda del árbol aunque se mantiene el número de preguntas necesarias para clasificar a los individuos las variables a tener en cuenta varían (gráfico en el código adjunto) siendo relevantes por el siguiente orden: la ocupación, la dificultad para llegar a fin de mes, la región donde habita la familia y la edad del miembro más mayor.

Los árboles basados en inferencia presentan en ambos casos (con todas las variables y solo con variables seleccionadas) el mismo porcentaje o peor porcentaje de acierto que los árboles estándar y no aportan una mayor interpretabilidad por lo que aunque se encuentran presentes en el código adjunto no serán estudiados en este informe.

Las diferentes variables destacadas según el árbol elegido hacen reflexionar sobre hasta qué punto son fiables variables cómo dificultad para llegar a fin de mes por no ser esta una variable medible en sí, si uno una percepción por parte del encuestado. Por ello y como mera curiosidad se construyó un árbol eliminando tanto esta variable como capacidad para afrontar gastos imprevistos y manteniendo solo las variables objetivamente ciertas. Sorprendentemente tras la poda esta técnica devuelve un árbol extremadamente sencillo (**Figura 3**) que tiene una tasa de acierto del 70,14% (solo dos puntos por debajo de la mejor tasa obtenida).

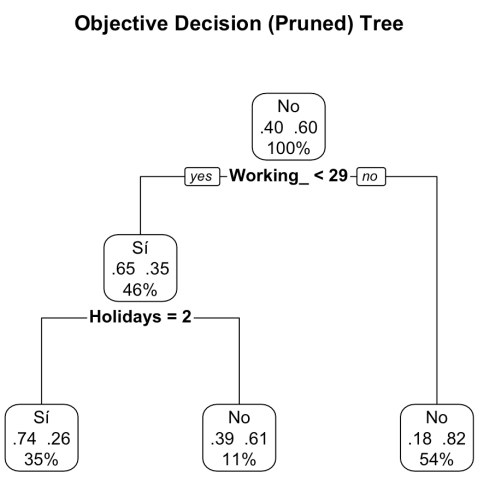

Figura 3. Árbol de decisión podado construido a partir de variables objetivas.

Este árbol mide tan solo dos factores: si el número de horas trabajadas es inferior o superior a 29 (descartándose a los hogares donde es mayor) y si el hogar ese puede permitir o no unas vacaciones (descartando de nuevo a aquellas familias que se las pueden permitir). Si bien es cierto que los nodos terminales obtenidos mediante este árbol no presentan ni por asomo la pureza que presentaban los del resultado principal, lo cual reduce su fiabilidad en comparación con los demás, se considera interesante por su simplicidad y por estar construido únicamente a partir de variables objetivas.

Conclusiones

En este caso concreto los árboles de decisión y la regresión logística parecen devolver resultados similares en cuanto a la calidad de la clasificación. En el informe anterior, el modelo presentado, construido a partir de una regresión logística, clasificaba correctamente un 72.92% de las familias, precisión idéntica a la que devuelve el mejor árbol de decisión construido en este experimento. Dado que las medidas de calidad son idénticas, desde un punto de vista didáctico o pensando en la comunicación de resultados, parece que los árboles de decisión suponen una mejor opción porque devuelven un resultado parecido pero mucho más fácil de interpretar y comunicar como se ha visto anteriormente.

En cuanto a las variables esenciales cabe destacar que varios de los árboles de decisión introducen una variable que en su momento la regresión logística había descartado como es la región donde se ubica el hogar. En mi opinión, esta idea sería bastante interesante pues dichos árboles plantean una precisión ínfimamente menor que el presentado (en algunos de ellos solo cambia la clasificación de una familia) y sin embargo introducen una clasificación que da bastante idea de la desigualdad entre las regiones. Además, desde un punto de vista científico parece más fiable confiar la clasificación a variables como sector ocupador o región que son puramente objetivas, que a variables como dificultad para llegar a fin de mes o capacidad para afrontar gastos imprevistos que están notablemente sesgadas por el encuestado al no ser variables puramente métricas.

Si se construye un árbol a partir de las variables métricas se perdería parte de la calidad clasificativa pero a cambio se obtendría un árbol quizá demasiado sencillo que permitiría la clasificación de las familias con solo dos preguntas y un tasa de acierto de 70,14%.

Código

En este enlace se puede consultar el código sobre el que se sostiene este informe. Tras pasarlo a PDF ocupaba 31 páginas y presentaba algunas dificultades para su lectura por lo que esta se consideró una mejor opción.