Executive Summary

Este informe tiene por objetivo la creación de un modelo de regresión logística basado en datos sobre condiciones de vida de 2016 extraídos del Instituto Nacional de Estadística.

Estos datos sobre las condiciones de vida permiten poner a disposición de la Comisión Europea un instrumento el cuál ayudará al estudio de la pobreza y desigualdad en general.

El objetivo por tanto será buscar un modelo de regresión logística usando como variable explicada 'Hogar en riesgo de pobreza', la cual tratamos como categórica.

A través del uso de variables tanto cualitativas como cuantitativas se construirá un modelo de entrenamiento con el 60% de los datos totales y otro de prueba con el resto (40%).

Tras esto realizaremos la prueba ANOVA para ver que variables son las que nos aportan más información para nuestro modelo. Además, se irán introduciendo y descartando variables de manera manual para afirmar o descartar que puedan introducir información extra en nuestro modelo.

Con los datos obtenidos en estas pruebas anteriores se procederá a realizar el Pseudo R2 de McFadden el cual busca una reducción proporcional de varianza del error.

Es necesario tener en cuenta que un modelo de clasificación debe ser capaz de predecir a qué clase va a pertenecer una supuesta nueva variable en base a nuestro modelo entrenado. Para ello se puede hacer una matriz de confusión para observar las variables que han sido bien o mal clasificadas en base a nuestros datos iniciales.

Finalmente, para evaluar el modelo en general podemos obtener su precisión como la proporción entre predicciones que han sido bien halladas en comparación con el total de variables de nuestro modelo.

Introducción

El modelo de regresión que vamos a construir permite relacionar una variable dependiente con otras independientes que pueden ser de cualquier tipo (como se ha especificado previamente). Para ello será necesario dividir el conjunto total de datos en una muestra de entrenamiento, la cuál llevará el 60% de los datos totales, y otra de prueba donde se contrastará la información de la primera y que contiene el 40% restante de los datos.

Como en todo análisis lo primero será un análisis exploratorio de variables: Nuestro conjunto inicial consta de 477 observaciones repartidas por 18 variables de las cuales vamos a descartar algunas ya que no aportan información al modelo.

Las variables eliminadas son:

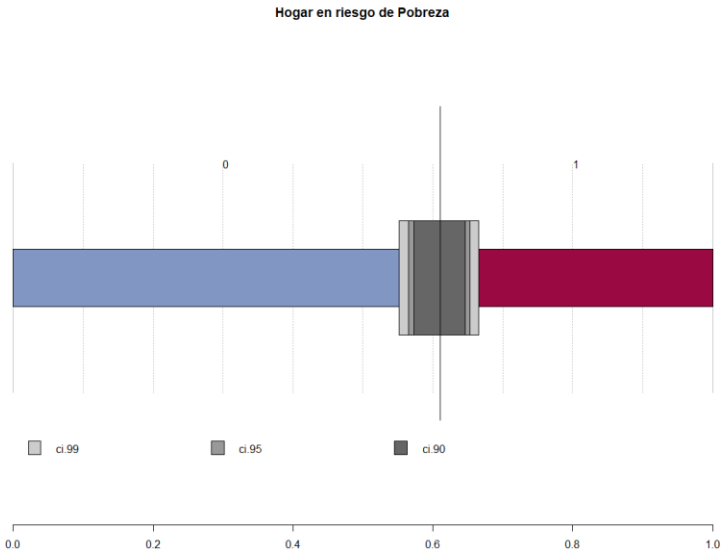
- Id del hogar
- Televisión de color
- Renta total del ejercicio anterior puesto ya está incluido en el modelo inicial
- Región
- Personas menores de 16 que reciben renta
- Sexo del mayor del hogar

El análisis exploratorio conlleva también las variables a su tipo correcto, esto implica cambiar de factor a numéricos en la mayoría de los casos.

Las categóricas se convierten a tipo factor como suele ser normal, entre ellas encontramos:

- Vacaciones fuera de casa
- Capacidad de afrontar pagos
- Ordenador
- Llegar a fin de mes
- Régimen de tenencia
- Actividad del mayor del hogar

A través del siguiente gráfico se puede apreciar el nivel de las variables en nuestro modelo.



Fuente: Elaboración propia

Si bien no está perfectamente nivelada no nos preocupa pues es un análisis exploratorio previo.

Modelo de regresión logística

El modelo se realizará sobre *Hogar en riesgo de pobreza* la cuál ayudará más tarde en la predicción de si hay riesgo de pobreza en ese hogar con la información que aportan el resto de las variables.

Como se ha explicado previamente, la muestra de entrenamiento contendrá 286 observaciones y la parte de prueba los 191 restantes.

Una vez obtenido el modelo linear con distribución binomial (recordemos que la variable es dicotómica), lo analizamos con *summary* y realizamos la prueba ANOVA que nos arroja que los valores que más influyen son los siguientes:

- Ayuda a familias
- Vacaciones fuera de casa
- Capacidad de afrontar los pagos
- Llegar a fin de mes
- Miembros
- Actividad del mayor

Seguidamente se analiza la bondad del ajuste a través del pseudo de McFadden el cual nos arroja un valor de 0.3789. En este estadístico un resultado entre 0.2 y 0.4 es un resultado más que bueno, lo cual muestra que nuestro modelo por ahora va por el camino adecuado.

Ahora es el momento de realizar predicciones sobre la muestra de *prueba,* estableciendo el corte en 0.68 ya que considero que es un valor más que bueno. El motivo de no dar un 0.5 es simplemente porque la discriminación de valores no sería correcta ya que este es la media (entre 0 y 1). La predicción se pone en forma de tabla y obtenemos los siguientes valores:

	Predicted	
Actual	0	1
0	102	7
1	42	40

Traducido a palabras esto nos indica que hemos realizado bien la predicción de 102 valores para el negativo y 40 para el positivo, teniendo solo un error en 49 variables.

Para comprobar que la precisión de nuestro modelo es adecuada lo que haremos será restar 1 menos el ratio de error del modelo general, obteniendo un 74.34%, lo cual es un valor más que aceptable para no tener que hacer más cambios en nuestro modelo.

Conclusiones

El objetivo general era realizar un modelo de regresión en base a un conjunto de datos, analizándolos y tratándolos previamente, además de hacer una separación en dos para la construcción de dicho modelo.

La variable objeto de estudio ha sido *Hogar en riesgo de pobreza*, explicada por doce variables independientes.

Tras realizar la prueba ANOVA y la prueba de McFadden hemos comprobado que obteníamos resultados apropiados y hemos procedido a la construcción del modelo. La matriz de confusión estimada es óptima y la precisión del modelo general está por encima del 70% concluyendo así nuestro modelo.