Universidad de los Andes

Maestría Economía Aplicada: Big Data and Machine Learning for Applied Economics

Grupo: Laura Natalia Capacho (202121025), Sebastián David Beltrán (202121021) y

Yurani Gonzalez (201212100)

***Problem Set 1: Predicting Income***

El objetivo es construir un modelo predictivo para el ingreso individual, a partir de la Gran Encuesta Integrada de Hogares del DANE para 2018 en la ciudad de Bogotá.

**1. Adquisición de datos**

(a) El proceso de adquisición de datos está descrito en el *script* de R adjunto al documento.

(b) No hay restricciones para acceder a la información puesto que no existe ningún archivo (robots.txt) asociado a la raíz del sitio web que indique restricciones para rastrear la información de la página. Sin embargo, al momento de realizar el *web scraping* para obtener las bases de datos, nos enfrentamos a un problema debido a que la página web es dinámica y no estática. Por lo anterior, no es posible extraer las tablas de forma directa, en la medida que la información no carga inmediatamente. Las tablas realmente provienen de una promesa *json*, asociada a un link externo que es de donde pudimos extraer los datos.

(c) Para adquirir los datos seguimos el siguiente proceso:

1. Crear una lista vacía “Links” en la que se van a guardar los links de cada una de las páginas que contienen las tablas con la información
2. Crear una lista vacía “base” en la que se van a almacenar las tablas que se extraen de cada página web
3. Crear un *loop* que para i = 1, 2, …, 10, ejecuta los siguientes pasos:
   1. Guardar en la lista “Links” cada uno de los links de las 10 páginas web, definiéndolo como un documento de .*html*
   2. Descargar y preprocesar las páginas web descargadas de cada link (*read\_html)*
   3. Convertir los nodos de tipo tabla de la página web en tablas de R (*html\_table)*
   4. Almacenar cada una de las tablas en la lista “base”
4. Renombrar la primera columna de todas las bases, que está vacía, para evitar errores
5. Convertir cada una de las tablas en la lista “base” a un formato *tibble*
6. Hacer un *append* de las 10 bases almacenadas en la lista “base”, para obtener una base de datos completa con toda la información

**2. Limpieza de datos (N)**

Luego de hacer una exploración de las variables disponibles en la base de datos se hizo una selección de aquellas que consideramos más relevantes para explicar el ingreso individual de una persona. Destacamos que en este trabajo se hace referencia al ingreso total, constituido tanto por ingresos laborales provenientes de salarios o independientes, como por ingresos de ayudas, subsidios, bonificaciones y diversas fuentes que serán detalladas más adelante.

En este sentido, los determinantes del ingreso son tanto características de la persona o su hogar, como del tipo de trabajo que desempeñe (rama de actividad económica, condiciones de formalidad, tipo de empresa, entre otros). Así como otros factores que pueden aumentar la probabilidad de recibir otros ingresos (subsidios o auxilios, pensiones, entre otros).

Variables de interés

En primer lugar, se tomó un conjunto de características sociodemográficas de los individuos: **edad, edad al cuadrado** (para capturar el efecto decreciente de la edad sobre el ingreso), **género, máximo nivel educativo, relación con el jefe del hogar** y **estrato**. Dentro de estas se hicieron algunas modificaciones y se crearon nuevas variables:

1. Creamos una variable de **educación** medida en años que construimos a partir del máximo nivel educativo alcanzado que se reportó en la encuesta. En efecto, si este nivel es ninguno se asignaron 0 años, si es preescolar 3 años, si es primaria incompleta 7 años, si es primaria completa 8 años, si es secundaria incompleta 13 años, si es secundaria completa 14 años y si es terciaria 19 años.
2. Construimos un *proxy* de **experiencia laboral**[[1]](#footnote-1). Debido a que no se tiene información reportada de los años de experiencia laboral, utilizamos el concepto de “experiencia potencial”, que se crea a partir de la edad, los años de educación y los años de iniciación en el mercado laboral (Aristizábal & Ángel, 2017).
3. Generamos una variable *dummy* que caracteriza si la persona es **jefa del hogar** o no.

En segundo lugar, se consideraron un conjunto de variables correspondientes al trabajo del individuo: trabajador **asalariado o independiente, formalidad, oficio** y **tamaño de la empresa** en que labora. En este caso también se realizaron modificaciones y se crearon variables:

1. Se genera una variable *dummy* que indica si la persona trabaja en una **microempresa** o no, a partir del tamaño de la empresa, medido por el número de trabajadores. En concreto, se denotan como microempresas aquellas compañías que tengan personal no superior a 10 trabajadores (Ministerio de Comercio, Industria y Turismo [MINCIT], 2007).

Finalmente, se toman todas las variables referentes a **ingresos** **totales** de la persona: **ingreso laboral** (salario o independiente) y **otros ingresos**, conformados por: ingreso por intereses o dividendos, por jubilaciones o pensiones, de ayudas de hogares e instituciones, por arriendos, por horas extra, por bonificaciones, por auxilios alimentarios o de transporte, por subsidios familiares o educativos, por primas de servicios, navidad o de vacaciones, por pensiones alimentarias, viáticos, accidentes o de cualquier otra fuente. En este conjunto de variables se hicieron algunas agrupaciones:

1. Se crea una variable de **ayuda de hogares** a partir de la suma de variables que denotan dinero recibido de otros hogares o personas residentes en el país y fuera del país.
2. Se crea la variable de **primas** que agrupa los ingresos por primas de servicios, navidad y de vacaciones.

Tratamiento de valores faltantes

En el procesamiento de los datos encontramos que hay muchas observaciones con datos faltantes (*missings)*, en este sentido, dimos un tratamiento especial a las variables correspondientes de acuerdo con ciertos criterios. En principio se dio un tratamiento a las variables existentes en la base de datos, a continuación, se detalla el procesamiento:

* La variable **oficio** es una variable categórica que denota la ocupación de la persona y toma valores de 1 a 99. En este caso los *missings* se reemplazaron por 0, que corresponde a una nueva categoría de ocupación que agrupa “otras ocupaciones”.
* La variable **formal** es una variable categórica que toma el valor de 1 si se trata de un trabajador formal y 0 en otro caso. Para los *missings* en este caso se asignó el valor de la variable p6090 (¿Está afiliado, es cotizante o es beneficiario de alguna entidad de seguridad social en salud?). Lo anterior, puesto que el empleo informal se refiere a los trabajadores que, entre muchos factores, pertenecen a una empresa o desempeñan un trabajo sin contrato laboral y sin aportes a seguridad social (Departamento Administrativo Nacional de Estadística [DANE], 2009).
* La variable **iof1es**corresponde al ingreso por intereses y dividendos imputado, y para sus valores faltantes se asignó el valor de la variable iof1 que es la misma variable, pero antes de imputación.
* La variable **iof2es**corresponde al ingreso por jubilaciones y pensiones imputado, y para sus valores faltantes se asignó el valor de la variable iof2 que es la misma variable, pero antes de imputación.
* La variable **y\_vivienda\_m** representa la renta por vivienda, en este caso a los *missings* se les asignó el valor reportado en la variable p7500s1 (Valor asociado a ¿El mes pasado, recibió pagos por: a. arriendos de casas, apartamentos, fincas, lotes, vehículos, equipos etc.?)
* Las variables **y\_horasExtras\_m**, **y\_bonificaciones\_m**, **y\_auxilioAliment\_m**, **y\_auxilioTransp\_m**, **y\_subFamiliar\_m**, **y\_subEducativo\_m**, **y\_primaServicios\_m**, **y\_primaNavidad\_m**, **y\_primaVacaciones\_m**, **y\_primas\_m**, **y\_viaticos\_m**, **y\_accidentes\_m**, **y\_total\_m** son ingresos por conceptos de horas extra, bonificaciones, auxilios de alimentación y transporte, subsidios familiar y educativo, primas de servicios, navidad y vacaciones, primas totales, viáticos, accidentes e ingreso salarial y de independientes, respectivamente. Para los valores faltantes en estas variables se asignó el valor cero (0). Lo anterior, partiendo del supuesto de que, si no hay valor en estos ingresos, es porque no existió o fue igual a cero.

Posteriormente, procedimos a crear algunas variables de interés, y así mismo se hizo un tratamiento sobre sus valores faltantes.

* La variable **Micro\_empresa** que creamos para denotar si la persona trabaja en una microempresa o no. En este caso, se decidió eliminar las observaciones para las cuales esta variable tenía *missings,* esto puesto que no tenemos información suficiente para asignar a microempresa o no arbitrariamente.
* En la variable **educ** que creamos y mide el nivel de educación en años, utilizamos el máximo nivel de educación, y en ese caso solamente se encontró un valor faltante, al cual se le asignó el valor de cero (0).

1. Para las personas con educación se utiliza la siguiente aproximación de experiencia (X): Si 18 < edad < 22, X = edad – 18; si edad > 22, X = edad – educación – 6. Y para personas sin educación terciaria se aproxima como sigue: si edad > 18, X = edad – 18. [↑](#footnote-ref-1)