Problem Set N°2 - Machine Learning
UNLP

Cantero, Lara Sofía
Temossi, Francisco
Vollmer, Candela

Repositorio de GitHub

1)

CONTEXTO

Este problema se inspiró en una competencia reciente organizada por el Banco Mundial: Pruebas Pover-T: Predicción de la pobreza. La idea es predecir la pobreza en Colombia. Como se afirma en la competencia, "medir la pobreza es difícil, lleva tiempo y es costoso. Al construir mejores modelos, podemos realizar encuestas con menos preguntas y más específicas que midan de manera rápida y económica la efectividad de nuevas políticas e intervenciones. Cuanto más precisos sean nuestros modelos, con mayor precisión podremos orientar las intervenciones e iterar sobre las políticas, maximizando el impacto y la relación costo-efectividad de estas estrategias".

El objetivo principal es construir un modelo predictivo de la pobreza de los hogares. Nótese que un hogar se clasifica como:

$$Pobre = I(Inc < PL)$$

donde I es una función indicadora que toma uno si el ingreso familiar está por debajo de una cierta línea de pobreza.

Esto sugiere dos maneras de predecir la pobreza. Primero, abordarla como un problema de clasificación: predecir ceros (no pobre) y unos (pobre). Segundo, como un problema de predicción de ingresos. Con los ingresos previstos, se puede utilizar la línea de pobreza y obtener la clasificación. Se explorarán ambas rutas en este trabajo.

INTRODUCCIÓN

VER ANTECEDENTES La identificación de hogares pobres representa uno de los grandes desafíos de las economías latinoamericanas, en las cuales aún persisten altas tasas de pobreza. Por un lado, la pobreza es un fenómeno que puede ser abordado desde un enfoque monetario, así como también desde una perspectiva multidimensional, contemplando cuestiones relativas a la calidad de vida de los hogares. Por el otro, estos países cuentan con escasos recursos fiscales disponibles para la recolección de datos. En este contexto, la identificación de las variables y el diseño de un modelo que permita predecir la condición de un hogar con el menor número de variables posible, representa una oportunidad para la

focalización y evaluación de impacto de políticas destinadas a reducir la pobreza. En este sentido, la literatura ... Los datos utilizados se obtuvieron del reporte "Medición de Pobreza Monetaria y Desigualdad" que extrae información de las bases de microdatos de la Gran Encuesta Integrada de Hogares (GEIH) de 2018 que elabora el Departamento Administrativo Nacional de Estadística (DANE) y de la misión Empalme de las Series de Empleo, Pobreza y Desigualdad - MESEP. Los datos contienen cuatro conjuntos divididos en datos de entrenamiento y de prueba a nivel de hogar e individuo.

La Gran Encuesta Integrada de Hogares es una encuesta que recolecta información sobre las condiciones de vida de los habitantes de Colombia (condiciones de la vivienda, educación, nutrición, composición demográfica del hogar, tenencia de activos, etc.), además de las características generales de la población como sexo, edad y estado civil. La GEIH proporciona información a nivel nacional, cabecera - resto, regional, departamental, y para cada una de las capitales de los departamentos.

Las características más importantes de esta encuesta son:

- Unidad de observación: Hogares.
- Muestra (total anual): 240.000 hogares, aproximadamente.
- Periodicidad: trimestral y anual.
- Temas: algunos son ingresos, condiciones sociales, nutrición, propiedad, ahorro/inversión, empleo y desempleo, situación económica, economía rural, vivienda, fertilidad y uso de servicios públicos.
- Cobertura: cobertura nacional que permite obtener resultados por zona urbana y rural, cinco grandes regiones (Región Atlántica, Región Oriental, Región Central, Región Pacífica y Región Bogotá) y total por departamentos: 24 departamentos, las 13 grandes ciudades con sus áreas metropolitanas y 11 ciudades intermedias.
- Universo: población civil no institucional, residente en todo el territorio nacional.

Una de las principales ventajas de esta fuente de datos es la amplitud de los aspectos relevados, que abarcan una variedad de temáticas demográficas y socioeconómicas. Esto la convierte en una herramienta particularmente valiosa para analizar cuestiones relacionadas con la pobreza e indicadores sociales y distribución del ingreso, lo que la hace especialmente adecuada para la construcción de un modelo de predicción de pobreza.

Los resultados obtenidos al usar una muestra de entrenamiento del 30 % de la muestra total indican que el mejor modelo para predecir pobreza .....

DATOS

En el presente trabajo la predicción de la pobreza se llevará a cabo de dos maneras distintas. Primero, se abordará como un problema de clasificación, es decir, la predicción de la pobreza como variable binaria; y segundo, como un problema de predicción de ingresos, donde con los ingresos previstos y la línea de pobreza se obtendrá la clasificación.

Para la primera parte se seleccionaron 7 variables. Es importante aclarar que se va a estar midiendo pobreza multidimensional, por ello se agregaron variables proxys de posibles dimensiones que afecten la determinación de si una persona es pobre o no, como por ejemplo educación, salud, condiciones de vida, hacinamiento, etc. La variable numérica incluída es la edad; mientras que las variables categóricas son la educación máxima alcanzada por

los miembros del hogar, el tipo de ocupación de los individuos, los regímenes de seguridad social en salud al que están afiliados los miembros del hogar (subsidiado u otros), si reciben subsidio familiar y el número de personas que duermen en una misma habitación.

Como proxy de educación, se utilizó el máximo nivel educativo alcanzado por los miembros del hogar. Esta variable asigna a todos los miembros del hogar el mismo nivel educativo del individuo más educado. Niveles bajos de educación pueden señalar hogares que se encuentren en una situación de pobreza. Como proxy de acceso a la salud, se contempló a cuáles regímenes de seguridad social en salud estaban afiliados los miembros del hogar. Esta variable asigna un valor 1 a aquellos individuos que estén bajo un régimen subsidiado y un valor 0 a aquellos que estén bajo cualquier otro régimen (contributivo o especial). Dadas las características del sistema de seguridad social en salud colombiano, en el régimen subsidiado se encuentran todas las personas más pobres y vulnerables **(cita a la documentación de la base)**. Como proxy de condición de vida se utilizaron distintas variables categóricas: la tenencia o no de algún subsidio familiar, donde un valor de 1 implica que el individuo lo tiene y 0 en caso contrario. Si la distribución del beneficio es pro-pobre, el acceso a estos programas de subsidios puede indicar que el hogar se encuentra en una situación de pobreza. Además, se incorporó la edad de los individuos y su tipo de ocupación. La primera, dado que influye en la etapa del ciclo de vida en que se encuentra una persona, lo que termina afectando su situación económica, pero principalmente porque es una proxy de la experiencia laboral; y la segunda, debido a que algunas ocupaciones ofrecen más estabilidad laboral y beneficios que otras, entonces ocupaciones precarias o informales están más asociadas con la pobreza.. Por último, como proxy de hacinamiento crítico se utilizó el número de personas por cuarto. Ésta se creó en función de otras dos variables referentes a la cantidad de cuartos en el hogar (incluyendo sala-comedor) y el número de habitantes de ese hogar. De acuerdo con la DANE (2012), son considerados hogares con hacinamiento crítico aquellos en donde el número de personas por cuarto es superior a tres, contando la sala y el comedor pero excluyendo los baños, garajes y los cuartos utilizados para negocio.

Para la segunda parte se seleccionaron 4 variables. Las variables numéricas consideradas son la edad y la cantidad de horas trabajadas[1]. También se han incluido variables categóricas como el sexo y el máximo nivel de educación.

Estas variables se consideran pertinentes por varias razones:

- Edad: influye en la experiencia laboral y el nivel salarial.
- Horas trabajadas: afecta directamente los ingresos.
- Sexo: es relevante porque existen diferencias salariales significativas entre hombres y mujeres en muchos sectores laborales, por lo que incluirla ayuda a capturar estas disparidades en el modelo.
- Máximo nivel de educación: es uno de los factores más importantes en la predicción de ingresos, ya que generalmente, a mayor nivel educativo, mayores son las oportunidades de empleo y salarios.

La base de datos final se obtuvo mediante cinco simples pasos. Primero, se crearon las variables de máximo nivel educativo alcanzado por los miembros del hogar, de régimenes de seguridad social en salud y de hacinamiento. Segundo, se transformaron los datos a formato numérico entero. Tercero, se renombraron las columnas a nombres más indicativos

[1]Se considera el total de las horas trabajadas la semana previa a la encuesta.

de cada variable. Cuarto, se filtraron los valores NA. Y por último, se agruparon los datos por id de hogar, para obtener una base a un nivel más agregado.

Tabla 1: Estadísticas descriptivas de la submuestra seleccionada

Variable	n	Media	S.E.	Mínimo	Máximo	25 %	50 %	75 %	Máximo

Elaboración propia en base GEIH 2018 (DANE)

Referencias

Departamento Administrativo Nacional de Estadística (DANE) (2012). Atlas Estadístico 2012.