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

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 conjuntos de datos de capacitación 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