Conjunto de problemas 2: Predicción de la pobreza “Las guerras de las naciones se libran para cambiar los mapas. Pero las guerras de la pobreza se libran para mapear el cambio” M. Ali

1. Introducción

Este conjunto de problemas se inspiró en una competencia reciente organizada por el banco mundial: Pover-T Tests: Predicting Poverty. La idea es predecir la pobreza en Colombia. Como afirma la competencia, “medir la pobreza es difícil, requiere mucho tiempo y es costoso. Al construir mejores modelos, podemos realizar encuestas con menos preguntas y más específicas que miden de manera rápida y económica la efectividad de las nuevas políticas e intervenciones. Cuanto más precisos sean nuestros modelos, con mayor precisión podremos orientar las intervenciones e iterar las políticas, maximizando el impacto y la rentabilidad de estas estrategias”.

El objetivo es predecir la pobreza a nivel de los hogares. Los datos, sin embargo, se proporcionan a nivel de hogar e individual. Puede usar información a nivel individual para crear variables adicionales para mejorar su predicción.

Los datos provienen del DANE y la misión para el “Empalme de las Series de Empleo, Pobreza y Desigualdad - MESE”. Los datos contienen cuatro conjuntos divididos en capacitación y pruebas a nivel de hogar e individual. Puede usar la variable id para fusionar hogares con individuos. Notará que faltan algunas variables en los conjuntos de datos de prueba; esto está diseñado para hacer las cosas un poco más desafiantes. Más información sobre los datos está disponible en el sitio web del concurso.

Una dimensión esencial para los formuladores de políticas es que pueden medir la pobreza de manera rápida y económica. Al construir su modelo, apunte a un modelo que use la cantidad mínima de variables.

Hay dos resultados esperados:

1. Un documento .pdf.

2. Envíos con los pronósticos de tu equipo en Kaggle en el siguiente enlace.

1.1 Instrucciones generales

El objetivo principal es construir un modelo predictivo de la pobreza de los hogares. Tenga en cuenta que un hogar se clasifica como

$$\text{Pobre} = I(\text{Inc} < \text{Pl}) \tag{1}$$

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

Esto sugiere dos formas de predecir la pobreza. En primer lugar, plantéelo como un problema de clasificación: prediga ceros (no pobres) y unos (pobres). En segundo lugar, como un problema de predicción de ingresos. Con los ingresos previstos, puede utilizar la línea de pobreza y obtener la clasificación. Explorarás ambas rutas en este conjunto de problemas.

El documento debe contener las siguientes secciones:

- Introducción. En la introducción se expone brevemente el problema y si existen antecedentes. Describe brevemente los datos y su idoneidad para abordar la pregunta del conjunto de problemas. Contiene una vista previa de los resultados y las conclusiones principales.

- Datos.[1] . Al redactar esta sección, debe:

    1. Describa la idoneidad de los datos para resolver la pregunta predictiva, el proceso de construcción de la muestra, incluido cómo se limpiaron y combinaron los datos y cómo se crearon nuevas variables.

    2. Incluir un análisis descriptivo de los datos. Como mínimo, debe incluir una tabla de estadísticas descriptivas con su interpretación. Sin embargo, espero un análisis profundo que ayude al lector a comprender los datos, su variación y la justificación de sus elecciones de datos. Utilice su conocimiento profesional para agregar valor a esta sección. No la presente como una lista "seca" de ingredientes.

- Modelos y Resultados. En esta sección se presentan las especificaciones y modelos utilizados para el tareas predictivas. Aquí debes incluir tres subapartados:

    1. Modelos de clasificación. En esta subsección se describe el enfoque de clasificación, es decir, su intento de predecir directamente ceros (no pobres) y unos (pobres).

    2. Modelos de regresión de ingresos. Esta subsección describe el enfoque de predicción de ingresos, es decir, su intento de predecir los ingresos primero y luego, indirectamente, predecir ceros (ningún pobre) y unos (pobre).

[1] Esta sección se encuentra aquí para que el lector pueda comprender su trabajo, pero probablemente debería ser la última sección que escriba. ¿Por qué? Porque vas a hacer elecciones de datos en los modelos estimados. Y todas las variables incluidas en estos modelos deben describirse aquí.

3. Modelo definitivo. Aquí describe los modelos que seleccionó como su presentación final en la competencia. Hasta 2 presentaciones contarán para el puntaje de la tabla de clasificación. Si se seleccionan menos de 2, Kaggle seleccionará automáticamente entre sus envíos con mejor puntuación. Esta subsección debe incluir:

   – Una explicación detallada de los modelos finales elegidos para la evaluación en Kaggle. La explicación debe incluir cómo se entrenó el modelo, la selección de hiperparámetros y cualquier otra información relevante.

   – Una comparación con al menos otras 2 especificaciones, para cada enfoque.

   – Una descripción de las variables utilizadas en el modelo y discutir su relación importancia en la predicción.

   – Una descripción de cualquier estrategia de submuestreo utilizada para abordar los desequilibrios de clase.

• Conclusiones y Recomendaciones. En esta sección, establece las principales conclusiones de tu trabajo.

2 Directrices adicionales

• Las predicciones deben enviarse en Kaggle. Consulte el sitio web de la competencia para obtener más información. información.

• Convierte un documento .pdf en Bloque Ne´on. El documento no debe tener más de 10 (diez) páginas e incluir, como máximo, 8 (ocho) anexos (tablas y/o figuras). La bibliografía y las exhibiciones no cuentan para el límite de páginas. Puede agregar un apéndice, pero el documento principal debe ser independiente. Específicamente, un lector debe poder seguir el análisis en el documento y estar convencido de que es correcto y coherente solo con el texto principal, sin consultar el apéndice.

• El documento debe incluir un enlace a su repositorio de GitHub.

   – El repositorio debe seguir la plantilla.

   – El LÉAME debería ayudar al lector a navegar por su repositorio. Un buen README ayuda a que su proyecto se destaque de otros proyectos y es el primer archivo que una persona ve cuando se encuentra con su repositorio. Por lo tanto, este archivo debe ser lo suficientemente detallado para enfocarse en su proyecto y cómo lo hace, pero no tanto como para que pierda la atención del lector. Por ejemplo, Proyecto Impresionante tiene una lista seleccionada de archivos README interesantes.

   – Incluya instrucciones breves para replicar completamente el trabajo.

   – La rama del repositorio principal debe mostrar al menos cinco (5) contribuciones sustanciales ciones de cada miembro del equipo.

   – El código tiene que ser:

      Totalmente reproducible.

Legible e incluir comentarios. En la codificación, como en la escritura, un buen estilo de codificación es fundamental. Te animo a que sigas la guía de estilo de tidyverse.

- Las tablas, figuras y escritos deben ser lo más prolijos posible. Etiquete todas las variables incluidas. Si tiene algo en sus figuras o tablas, espero que se aborden en el texto. Las tablas deben seguir el formato AER.