Solucion Datathon

Sahlre

1 de julio de 2017

Introducción

Lo que esta en negrilla es un link que llevará al respectivo documento, para que esto funcione debe descargar primero el pdf.

Los datos abiertos relacionados al sector educativo contienen diversas mediciones (variables o características) asociadas a los hogares que podrían dar solución al desafió de inversión educativa en Colombia. Una de ellas es el nivel educativo de las personas, variable que se convierte en la variable resultado o dependiente que se intentará predecir con las otras características asociadas a cada una de los individuos, como por ejemplo, sólo por mencionar algunas, el género, el material predominante de las paredes exteriores y estado laboral.

Con dichas características se construyeron diversos algoritmos supervisados para predecir el nivel educativo de las personas: *Uno vs todos Regresión Logistica Multivariada*, *Árbol de clasificación*, *Bagging con árboles de clasificación*, *Boosting con árboles de clasificación* y *Bosques Aleatorios de clasificación*. La decisión de incluir varios algoritmos es poder propener en esta solución ambos aspectos: predicción e interpretabilidad. Es sabido el *trade off* que existe entre esos aspectos, a mayor capacidad de predicción de un algoritmo menor será su interpretabilidad. Sin interpretabilidad es díficil propener solución alguna al desafio de inversión educativa.

Los principales resultados son: XXXXXXXXXXXXXX.

Este documento se organiza de la siguiente manera: XXXXXXXXXXXXXX.

Metodología

Se elegieron las bases de datos del mes de abril de la **Gran Encuesta Integrada de Hogares (GEIH) - 2017**. Luego con el objetivo de tener una base de datos a nivel nacional con la mayor cantidad de observaciones se seleccionaron las bases de datos: *Características Generales*, *Fuerza de trabajo*, *Otras actividades y ayudas en la semana*, *otros ingresos* y *Vivienda y hogares*, tanto para Areas metropolitanas como para Cabeceras y zonas Rurales. Con cada una de esas bases de datos se procedió a seleccionar aquellas variables con un porcentaje de valores pérdidos inferior al 10%. Tal decisión se toma con la intención de crear un modelo predictivo del nivel educativo que aproveche la mayor cantidad de observaciones de las bases de datos de la GEIH. De esta manera, la propuesta de este análisis de datos no se convierte en una respuesta única a la variación del nivel educativo de los colombianos, sino en una aproximación al mejor modelo predictivo del nivel educativo de los colombianos a partir de la GEIH. El código para cargar y unir las bases de datos está en **Carga de datos**, este produce una base de datos que se ha denominado *bd_Nac.csv*. El nombre de las variables, su significado y los valores asumidos, con su respectiva codificación se puede explorar en **Definicion de variables**

En la base de datos *bd_Nac.csv* aún hay presencia de valores pérdidos ya que estos no fueron representados a través de un valor vació sino como un valor númerico. Se procede a eliminarlos, además se crea la variables resultado o dependiente sobre la que se intentará realizar la predicción. Se decide transformar la variable P6210 (Nivel educativo mas alto alcanzado) en una variable con tres niveles: Bajo, Medio, Alto.

- El nivel bajo esta conformado por los valores: 1. Ninguno, 2. Preescolar, 3. B?sica primaria, 4. B?sica secundaria.
- El nivel medio por: 5. Media.
- El nivel alto por: 6. Superior o universitaria.

Este último procedimiento se encuentra soportado en el archivo **Preproceso de datos 1**. Allí el código produce la base de datos: *bd_Nac_final.csv*

Finalmente en el proceso de construcción de la base de datos sobre la que se realizará el entrenamiento de los algoritmos así como su evaluación de pronóstico, en el archivo **Datos de entrenamiento y de prueba** se encuentra el código que produce las bases de datos de entrenamiento, validación y prueba. La base de datos de entrenamiento *bd_Train.csv* tiene 54397 observaciones y 61 variables, y la base de datos de prueba *bd_Test.csv* tiene 13598 observaciones y 61 variables.

Predicción del nivel educativo

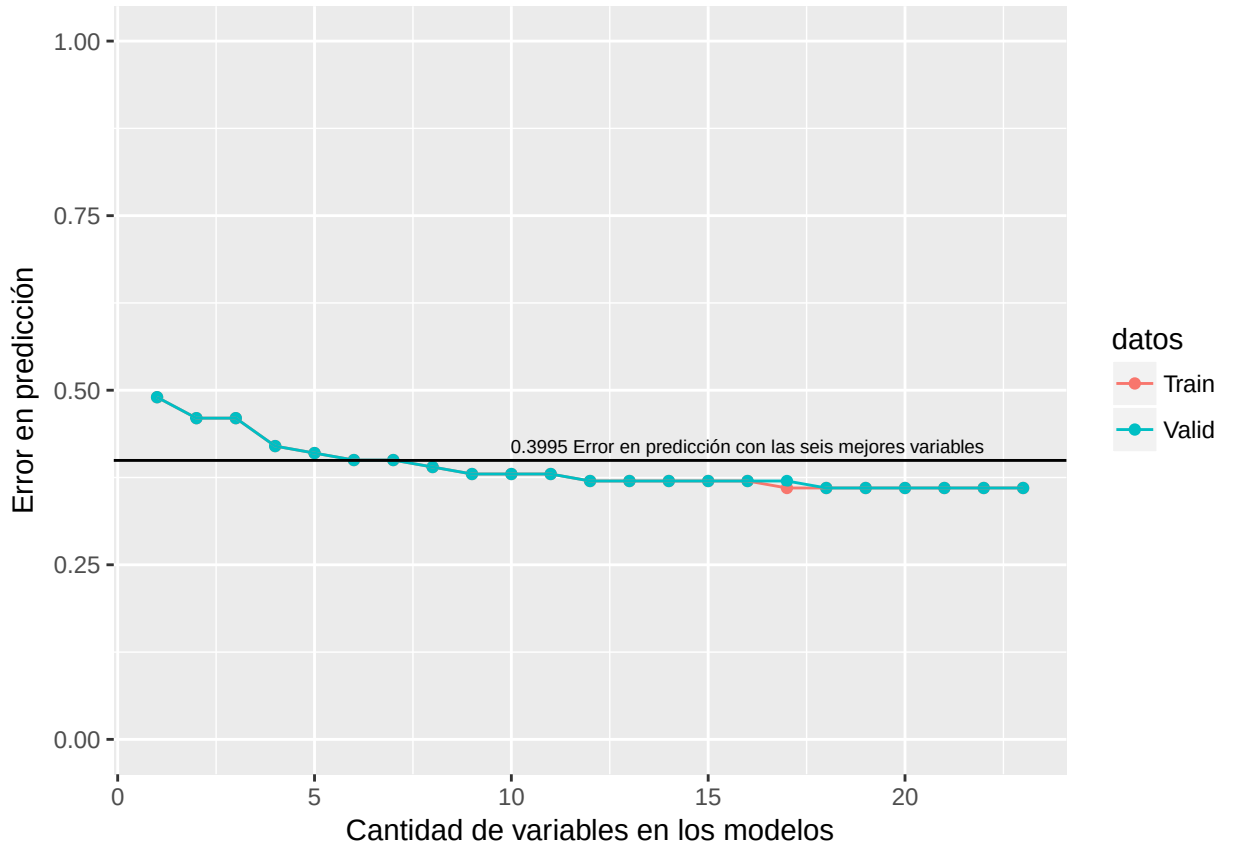
Uno vs todos Regresión Logistica Multivariada

El código para estimar el algorítmo se encuentra en el archivo FSS Uno vs Todos

Cómo el nivel educativo, variable dependiente, asume tres niveles: Bajo-Medio-Alto, la predicción de esta variable no se puede ejecutar a través de la Regresión Logistica Multivariada adecuada para una variable resultado binaria. Este algoritmo consiste en aplicar una regresión logística multivariada por cada nivel asumido en la variable dependiente. De esta manera se estiman tres modelos: el primero con la variable dependiente asumiendo sólo dos valores: "Bajo" y "El resto", el segundo con la variable dependiente: "Medio" y "El resto", y el tercero con: "Alto" y "El resto".

Con el propósito de crear un modelo con mayor nivel de interpretación se utilizó la técnica de selección de variables: Forward Stepwise Selection (FSS).

Los resultados de las primeros 23 modelos se presentan en la siguiente gráfica.



Se observa que a partir de las seis mejores variables el error de predicción tanto para datos de entrenamiento (Train) como para datos de validación (Valid) disminuye lentamente. De esta manera se eligen las primeras seis variables seleccionadas a partir del método FSS que son: P5210s16, P6100, P6040, P6240, P6170, P4020.

Al estimar el modelo con las seis variables se producen los siguientes indicadores:

- Error de clasificación de entrenamiento = 0.3995
- Error de clasificación de prueba = 0.4002

Este modelo es aceptable si se tiene en cuenta que al estimar el modelo con todas (60) las variables los indicadores son:

- Error de clasificación de entrenamiento = 0.3590
- Error de clasificación de prueba = 0.3602

También es importante mencionar que en la gráfica anterior se observa que la línea de error de predicción para datos de validación y datos de entrenamiento se solapan. Esto sugiere que hay alto sesgo. Es decir la hipótesis del algoritmo planteado es muy simple para modelar el nivel educativo. Otra señal de alto sesgo es cuando el error de clasificación es alto y el error de clasificación en datos de entrenamiento y en datos de validación son muy similares.

Resultados que indican varios caminos a seguir si se desea mejorar la predicción:

- Permitir al modelo capturar relaciones no lineales. Esto se lograría a través de interacción entre variables y aumentar el grado de polinomio de aquellas variables que sean continuas.
- Usar algorítmos como las Redes Neuronales, las Máquinas de Soporte Vectorial o Bosques Aleatorios, reconocidos por la capacidad de predicción en escenarios complejos.
- Adicionar nuevas variables.

Árbol de Clasificación.

El código para estimar el algorítmo se encuentra en el archivo Tree

Los métodos basados en árboles consisten en segmentar el espacio de variables predictoras (independientes) dentro de un número más simple de regiones. Luego de estimar el árbol de clasificación y realizar la respectiva podación se obtiene:

- Error de clasificación de entrenamiento = 0.3945
- Error de clasificación de prueba = 0.4029

Con el algortimo Árbol de clasifiacación sucede algo similar al resultado del anterior algoritmo. Los errores de clasificación en datos de entrenamiento y de prueba son altos y similares, lo que indica alto sesgo. De esta manera, para aumentar el nivel de predicción sería necesario utilizar algorítmos que produzcan hipótesis no lineales complejas como lo son las Redes Neuronales, las Maquinas de Soporte Vectorial o los Bosques Aleatorios.

Bagging de Árboles de clasificación

Uno de los problemas más recurrentes de los árboles de decisión es la débil precisión en el pronóstico. El método Bagging o Bootstrap aggregation es un procedimiento para reducir la varianza de los algorítmos supervisados.