Daniel Casas Bautista

Código: **202120803**

Lucia Fillippo Aguillón

Código: **202213187**

Irina Andrea Vélez López

Código:

Miguel Ángel Victoria Simbaqueva

Código:

**Problem Set 3**

Big Data & Machine Learning

**Fecha de entrega: 30 de julio de 2023**

**Resumen:** El presente informe presenta la solución al Problem Set 3 de la clase Big Data & Machine Learning, en donde se aplicaron diversos conceptos y herramientas para la limpieza de bases de datos y el desarrollo de un modelo de predicción de la pobreza en territorio colombiano.

En el repositorio GitHub que abajo se referencia contiene el presente informe junto a los scripts de RStudio donde se generaron todos los resultados. Este repositorio se encuentra en: <https://github.com/irivelez/PS3_Predicting_Poverty.git>

1. **Introducción**

El desarrollo de políticas públicas orientado a mejorar el bienestar de la población pobre y vulnerable es fundamental para para fomentar una sociedad equitativa que brinde igualdad de oportunidades a sus habitantes. En este orden de ideas, el diseño de las políticas públicas será óptimo en la medida en que se focalice de mejor manera la población objetivo, para que la distribución de los recursos disponibles llegue a aquellas personas que más lo necesitan.

Teniendo presente lo anterior, el objetivo del presente documento es predecir la situación de pobreza de los hogares colombianos, con el fin de que las políticas públicas orientadas a esta población sean correctamente dirigidas, evitando errores de inclusión y de exclusión. Para lograr esto, se utilizará un modelo de predicción de la pobreza de los hogares, a partir de datos obtenidos de la Encuesta de Medición de Pobreza Monetaria y Desigualdad en el año 2018 por parte del Departamento Administrativo Nacional de Estadística – DANE. En el presente documento se utilizará un modelo predictivo que se basará en el siguiente modelo:

Donde la condición de pobreza existirá cuando el indicador *I* señale queel ingreso *Inc* sea menor a la línea de pobreza *Pl*. El modelo determinará la condición de pobreza por dos vías; en primer lugar, una estrategia de clasificación para predecir hogares pobres (1) y no pobres (0); en segundo lugar, se utilizarán regresiones para determinar el ingreso de los hogares y así determinar si se encuentra por debajo o por encima de la línea de pobreza. Dentro del ejercicio a realizar se tendrán que identificar ROC[[1]](#footnote-2), falsos positivos, falsos negativos y demás elementos para predecir de la manera más acertada los hogares que serían objeto de análisis de las políticas relacionadas con este problema.

1. **Datos**

Como se mencionó previamente, para el desarrollo del Problem Set se utilizarán datos obtenidos de la Encuesta de Medición de Pobreza Monetaria y Desigualdad en el año 2018 por parte del Departamento Administrativo Nacional de Estadística – DANE, la cual contiene información que permite realizar un análisis de la pobreza en Colombia, al contener datos de ingreso, de mercado laboral, sociodemográfica, entre otros. Por ejemplo, la variable de ingreso del hogar es una variable continua que cuenta con un total de 164.960 observaciones. El valor de ingreso en el primer cuartil es de $800.000 y el valor medio de ingreso es de $2.102.586, el cual es 1.6 veces el salario mínimo[[2]](#footnote-3) que aplicaba en dicho año. La línea de pobreza (Lp) refleja el límite de ingresos por debajo del cual un hogar es considerado pobre, señalando que el valor mínimo es de $167.222, el máximo de $303.8107 y la media de $271.605. El nivel de pobreza, teniendo en cuenta los anteriores datos de ingreso, reflejan un total de 33.024 personas en condición de pobreza monetaria en la muestra, mientras que hay 131.936 que no lo están. Para 2018, de acuerdo con el DANE, la línea nacional de pobreza monetaria fue de $257.433[[3]](#footnote-4).

1. **Modelo y resultados**

Para el modelo de regresión, se utilizarán las variables del modelo *test* que se encuentren simultáneamente en la de *train*. Se realizará el análisis de dichas variables.

El modelo es el siguiente:

Donde:

* : Es una variable dummy que toma el valor de 1 si la persona es pobre y 0 en caso contrario
* : Es una variable categórica que señala el número de personas por unidad de gasto
* : Es una variable continua que refleja la línea de pobreza que aplica al hogar, de acuerdo con su sitio de residencia
* : Es una variable categórica que refleja el tipo de vivienda y toma los siguientes valores. (a: Propia, totalmente pagada; b: Propia, la están pagando; c: En arriendo o subarriendo; d: En usufructo; e: Posesión sin título; f: Otra)
* : Es una variable que refleja la ciudad donde reside la persona
* : Es una variable categórica que refleja el número de cuartos que tiene el hogar

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| **Modelo** | **Alpha** | **Lambda** | **ROC** | **Sens** | **Spec** | **Accuracy** | **Kappa** |
| **Logit Lasso Upsample** |  | 0.0136766552 | 0.7735692 | 0.7016445 | 0.7067255 | 0.7041850 | 0.4083700 |
| **Logit Lasso Downsample** |  | 0.0130565713 | 0.7718717 | 0.6991289 | 0.7059481 | 0.7025385 | 0.4050769 |
|  |  |  |  |  |  |  |  |
|  |  |  |  |  |  |  |  |
|  |  |  |  |  |  |  |  |

1. **Conclusiones y recomendaciones**

* Haber utilizado datos geográficos como predictores ha sido útil para imputar datos a la base de datos que permite dar robustez al modelo y a los predictores elegidos;
* El modelo de Ranger ha sido el que menor MAE arrojó; no obstante, eso no se traduce en una mejor predicción, de hecho, ha sido el Boosting el modelo que, al compararlo con los datos reales en la plataforma Kaggle, tuvo un menor error de predicción;
* Para los modelos Boosting hay algunas ajustes que permiten mejorar la capacidad predictiva de un modelo, esto es, aumentar el número de árboles, permitir que el modelo tenga más capacidad para capturar patrones complejos en los datos reducir la tasa de aprendizaje, ajustar la profundidad máxima de los árboles;
* El resultado de las predicciones se pudo haber visto afectado por la heterogeneidad que existe en el sector de Chapinero en Bogotá; lo anterior, dado que se trata de una localidad que tiene muchos contrastes, al tener barrios de diferentes estratos socioeconómicos. Esto podría dificultar a los modelos identificar patrones correctamente en los datos.

Bla bla bla

1. La curva ROC (Receiver Operating Characteristic) y el área AUC (Area Under the Curve) son dos métricas comúnmente utilizadas para evaluar y comparar la calidad del rendimiento de un modelo de clasificación [↑](#footnote-ref-2)
2. De acuerdo con los valores de salario mínimo y subsidio de transporte aplicable a la fecha. Fuente: <https://www.portafolio.co/economia/empleo/salario-minimo-colombia-2017-109538> [↑](#footnote-ref-3)
3. Fuente: [Boletín técnico Pobreza Monetaria en Colombia 2018 (dane.gov.co)](https://www.dane.gov.co/files/investigaciones/condiciones_vida/pobreza/2018/bt_pobreza_monetaria_18.pdf) [↑](#footnote-ref-4)