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.

1. **Linear model (LM)**

Este modelo contiene como variable Y el Logaritmo natural del Precio (Ln\_Price) y como variable predictora el tamaño de la vivienda en metros cuadrados. Este primer resultado implementó un modelo de regresión lineal simple (LM) como estimación preliminar para predecir el precio de las viviendas de Chapinero. La elección del logaritmo natural del precio como variable dependiente se debe a la necesidad de reducir la asimetría en la distribución de los precios o la estabilización de la varianza en los residuos del modelo. En ese sentido, la idea es que al conocer el valor del área de una vivienda, el modelo pueda proporcionar una estimación del precio considerando una normalización por el logaritmo natural, que luego podría ser transformado nuevamente para obtener una estimación del precio real de la vivienda. En este ejercicio inicial, el Mean Abolute Error – MAE fue de: 0.1336067 (87521698 COP)

1. **Leave One Out Cross Validation (LOOCV)**

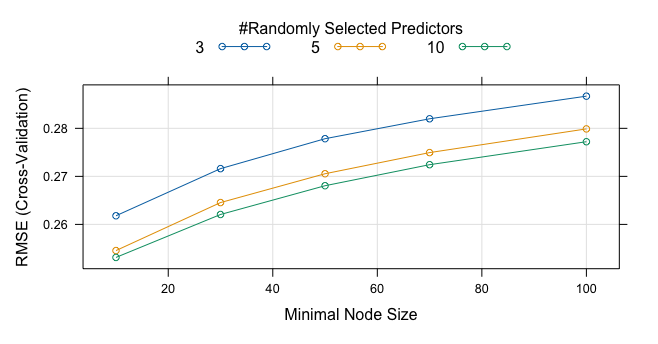
En esta ocasión se tomó el mismo modelo al anterior; no obstante, por la naturaleza del modelo se implementaron varias iteraciones para la información de muestra y de entrenamiento. En este caso el MAE fue de: 0.1336067 (87521698 COP)

1. **Leave One Out Cross Validation (LOOCV) – Modelo II**

Este modelo contiene como variable Y el Logaritmo natural del Precio (Ln\_Price) y como variables predictoras tiene: área en metros cuadrados, Número cuartos, Número de baños, Parqueadero, Ascensor, Baño privado; Balcón; Vista; Remodelado; Tipo de propiedad; Cercanía a parques; Área del parque más cercano; Cercanía a centros deportivos y Cercanía a piscinas. En este caso el MAE fue de: 0.1609716 (88643254 COP)

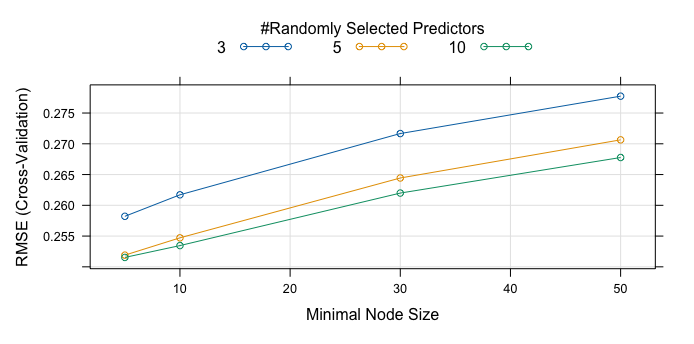
1. **Primer modelo de bosques son hiper parámetros (Ranger)**

Este modelo contiene como variable Y Ln\_Price y como predictoras las mismas variables del anterior modelo; no obstante, se adicionaron como hiper parámetros un MIN.NODE.SIZE con un rango desde (10) a (100), y un MTRY de (3), (5) y (10). Los resultados del RMSE, a la luz de la inclusión de los hiper parámetros, se observa en la siguiente gráfica, donde se muestra que el nodo en 10 es el de menor RMSE, mientras que el de 3 tiene mayor RMSE. Producto de lo anterior, en este caso el MAE fue de: 0.08934838 (59397009 COP)

****

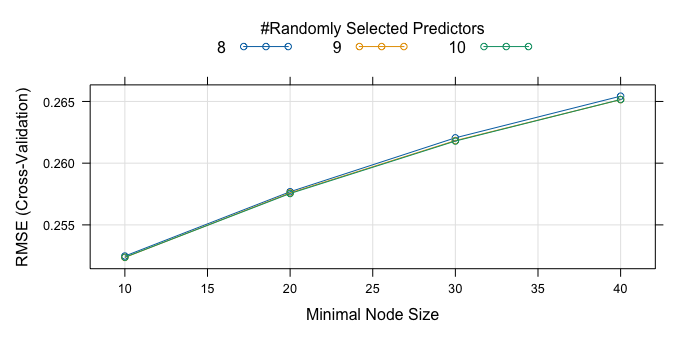
1. **Segundo modelo de bosques son hiperparámetros (Ranger)**

Este modelo contiene como variable Y Ln\_Price y como predictoras las mismas variables del anterior modelo. En cuanto a los hiper parámetros se tiene un MIN.NODE.SIZE con un rango de (5), (10), (30) y (50) y un MTRY de (3), (5) y (10). Los resultados del RMSE, como se observa en la gráfica, muestra que el nodo en 10 sigue siendo el de menor RMSE y varía en la medida en que se incrementa el tamaño del nodo. En este caso el MAE fue de: 0.0826849 (54991319 COP)

****

1. **Tercer modelo de bosques son hiperparámetros (Ranger)**

Este modelo contiene como variable Y Ln\_Price y como predictoras las mismas variables del anterior modelo. En cuanto a los hiper parámetros se tiene un MIN.NODE.SIZE con un rango de (10), (20), (30) y (40) y un MTRY de (8), (9) y (10). Los resultados del RMSE, como se observa en la gráfica, muestra que el nodo en 10 sigue siendo el de menor RMSE, incluso, si se sube a más de 10, el RMSE se incrementa. En este caso el MAE fue de: 0.09132121 (60677661 COP)

****

1. **Primer modelo con árboles (RPart)**

Este modelo contiene como variable Y Ln\_Price y como predictoras las mismas variables del anterior modelo. En este caso el MAE fue de: 0.1944964 (125098502 COP)

1. **Boosting Model I**

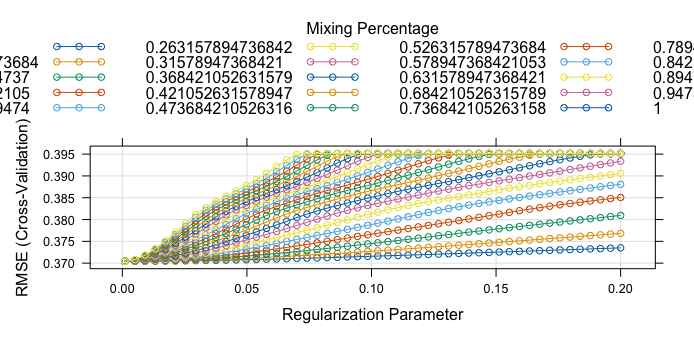
Este modelo contiene como variable Y Ln\_Price y como predictoras las mismas variables del anterior modelo. En cuanto a hiper parámetros se tiene una learn\_rate de (0.001); un ntrees de (100); un max\_depth de (6); un min\_rows de (70) y un col\_sample\_rate de (0.2). En este caso el MAE fue de: ­­0.1324681 (88168556 COP)

1. **Boosting Model II**

Este modelo contiene como variable Y Ln\_Price y como predictoras las mismas variables del anterior modelo. En cuanto a hiper parámetros se tiene una learn\_rate de (0.1), (0.01) y (0.001); un ntrees de (50), (100) y (500); un max\_depth de (10); un min\_rows de (70) y un col\_sample\_rate de (0.2). En este caso el MAE fue de: 0.1315953 (84721894 COP)

1. **Elastic net**

Este modelo contiene como variable Y Ln\_Price y como predictoras las mismas variables del anterior modelo. El model es LMNET y el MAE fue de: ­­0.1324681 (88168556 COP)



A continuación se muestra el resumen de los datos obtenidos para cada modelo:

|  |  |  |
| --- | --- | --- |
| Modelo | MAE (logaritmo) | MAE (COP) |
| Linear model (LM) | 0.1336067 | 87521698 |
| Leave One Out Cross Validation (LOOCV) | 0.1336067 | 87521698 |
| Leave One Out Cross Validation (LOOCV) – II | 0.1609716 | 88643254 |
| Ranger I | 0.08934838 | 59397009 |
| Ranger II | 0.0826849 | 54991319 |
| Ranger III | 0.09132121 | 60677661 |
| RPart | 0.1944964 | 125098502 |
| Boosting – I | 0.1337077 | 88910529 |
| Boosting – II | 0.1315953 | 84721894 |
| Elastic net | 0.1324006 | 84836876 |

Si bien el menor MAE en la presente tabla aparece en el modelo Ranger II, al comparar los datos en la lista en Kaggle, el modelo que mejor se aproximó a los datos reales es el Boosting I.

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)