**Modelo y resultados**

Las variables utilizadas para entrenar el modelo se encuentran en la siguiente tabla, con una breve descripción de cada una de ellas:

Tabla . Variables del modelo escogido

|  |  |  |
| --- | --- | --- |
| Tipo de atributo | Atributo | Explicación |
| Físico | Área construida de la vivienda | Variable continua que describe el tamaño en metros cuadrados de la vivienda |
| Físico | Tipo de propiedad | Variable categórica con dos clases: apartamento o casa |
| Físico | Cantidad de habitaciones | Variable discreta con la cantidad de habitaciones que tiene la vivienda |
| Físico | Cantidad de baños | Variable discreta con la cantidad de baños que tiene la vivienda |
| Físico | Existencia de terraza o balcón | Variable categórica con dos clases: tiene terraza o no tiene terraza |
| Físico | Existencia de parqueadero | Variable categórica con dos clases: tiene terraza o no tiene terraza |
| Físico | Existencia de depósito | Variable categórica con dos clases: tiene depósito o no tiene depósito |
| Ordenamiento | UPZ | Variable categórica con 112 clases, correspondientes a las UPZ existentes al 2019 |
| Servicios | Distancia la estación de TransMilenio | Variable continua que describe la distancia a la estación de TransMilenio más cercana. Polinomio orden 2 |
| Servicios | Distancia al parque | Variable continua que describe la distancia al parque más cercano. Polinomio orden 2 |
| Servicios | Distancia a la universidad | Variable continua que describe la distancia a la universidad más cercana. Polinomio orden 2 |
| Servicios | Distancia al hospital | Variable continua que describe la distancia al hospital más cercano. Polinomio orden 2 |
| Servicios | Distancia al colegio | Variable continua que describe la distancia al colegio más cercano. Polinomio orden 2 |
| Social/Económico | Estrato | Variable categórica con 6 clases, correspondientes a los estratos socioeconómicos en Colombia |
| Social/Económico | Cantidad de delitos en el 2019 en la UPZ | Variable continua que describe la cantidad de delitos cometidos en el 2019 por UPZ, descritos en la sección anterior |
| Social/Económico | Distancia al CAI | Variable continua que describe la distancia al Comando de Atención Inmediata (CAI). Polinomio orden 2 |

Nota: Las distancias estimadas están en metros y dentro del entrenamiento del modelo fueron incluidas como un polinomio de orden 2, luego de ver que, a través de varias modificaciones, el modelo tenía un mejor desempeño con esta especificación para todas las distancias.

Con estas variables el entrenamiento del modelo fue realizado a través de *Random Forest,* que es un método basado en árboles, en el que construyen una cantidad de árboles de decisión en muestras de entrenamiento que son generadas a través de un *Bootstrap*. Un parámetro crítico en este método es la cantidad de variables que son consideradas en cada división de los respectivos árboles, denominado como o en el código como .

Inicialmente, se definió el parámetro con base en la literatura [YYY1] teoría, en la que se indica que típicamente se define como la raíz cuadarada de la cantidad total de predictores. Para este caso, teniendo en cuenta la cantidad total de predictores (110), se definió inicialmente el como 10. Sin embargo, al realizar varias iteraciones con ajustes en el , se evidenció que la cantidad con el menor error absoluto (MAE, por sus siglas en inglés) era de 30.

Otro parámetro importante es la cantidad de dobleces o *folds* que se utilizan para la división de la base de datos de muestra para entrenar el modelo. A través de un proceso iterativo que comenzó con cinco *folds* y terminó con veinte *folds*, se identificó que cinco era la mejor cantidad en términos de desempeño y esfuerzo computacional (con diez y veinte *folds* las mejoras en MAE eran menos de 3 millones de pesos, lo que representaba una mejora de menos del 3% respecto al MAE del mejor modelo).

El último parámetro que fue tenido en cuenta para este proceso iterativo fue el tamaño del nodo o de la hoja de los árboles generados en el proceso. Para evitar el sobreajuste del modelo o un truncamiento forzado de la profundidad de los árboles, se evaluó el resultado con dos valores: 3 y 5.

La siguiente tabla muestra algunos de las varias alternativas estudiadas para escoger el mejor modelo dentro del método de *Random Forest* –que era el que tenía mejor desempeño con la base de datos de entrenamiento y con la validación en Kaggle–:

|  |  |  |  |
| --- | --- | --- | --- |
| mtry | Folds | Tamaño de nodos | MAE - Train |
| 30 | 10 | 5 | 96.776.686 |
| 20 | 10 | 5 | 99.325.911 |
| 10 | 20 | 5 | 118.459.734 |
| 10 | 10 | 5 | 119.370.083 |
| 10 | 5 | 5 | 121.089.040 |

Como se puede ver con esta muestra de opciones estudiadas, el parámetro más importante era *.* Adicionalmente, ya después de un valor de 30 el esfuerzo computacional era considerable y tomaba varias horas para que se identificara el mejor modelo[[1]](#footnote-1).

El mejor modelo fue escogido de 23 opciones que fueron sometidas a validación en Kaggle y varias más que no fueron subidas pero cuyo MAE en la base de datos entrenamiento fue el criterio para no ser subidos a la competencia. Dentro de estos intentos, se incluyeron los métodos vistos en clase, desde regresión lineal hasta *Boosting,* para ver cuál se comportaba mejor. La siguiente gráfica muestra el MAE para la base de datos de entrenamiento (*Train MAE*) y el resultado en Kaggle cuando se validaba:

Figura . Resultados de Error Promedio Absoluto (MAE) para los métodos utilizados. Para el modelo de Regresión Lineal el Test MAE fue de 4.133.457.140 (14 veces mayor al siguiente método); no se muestra en la gráfica para visualizar mejor la tendencia de los demás modelos.

Es importante notar que el resultado del mejor método también obedece a que en este fue el único que se incluyeron otras variables de servicios que se tenían alrededor de la vivienda (distancias a CAI, hospital, universidad, colegio, parque). Si se actualizarán los demás modelos con la especificación más completa se esperaría que los resultados mejoraran, pero no significativamente ya que antes de que se incluyeras estas variables, el método de *Random Forest* era el que tenía menor MAE[[2]](#footnote-2).

Bibliografía

[YYY1] James, G., Witten, D., Hastie T., Tibshirani, R. An Introduction to Statistical Learning with Application in R. Springer. Second Edition.

1. Se realizó un modelo con un parámetro de mtry de 50 y los cambios en el MAE fueron menos del XXX% [↑](#footnote-ref-1)
2. Con los métodos de regularización se hizo la inclusión de estas nuevas variables y sus MAE se redujeron en menos de 1 millón COP, es decir, menos del 1%. [↑](#footnote-ref-2)