PROBLEM SET 3: MAKING MONEY WITH ML

Enlace GIT: https://github.com/dfosorio111/BD-ML---PS3.git

Por medio del presente documento, se hace descripción completa del modelo predictivo de *Machine Learning* desarrollado por Franco, Malkun y Osorio (2022) con el propósito de predecir de la manera más acertada posible el precio de las viviendas de la ciudad colombiana de Santiago de Cali, teniendo en cuenta que el objetivo final de los modelos es cerrar el mayor número de negocios posibles maximizando el porcentaje de inmuebles comprados en la ciudad de Cali con el gasto mínimo posible. Se hace uso de los datos de propiedad individual de las ciudades de Bogotá y Medellín provenientes de la plataforma de *Properati* para entrenar los modelos.

Además, se debe tener en cuenta el interés enfático de la compañía para la cual se desarrolla el presente informe en querer evitar una tragedia como la sucedida en el renombrado "fiasco de Zillow" donde se sobreestimó el valor de un gran número de propiedades y se reportaron pérdidas por más de $500 millones de dólares para la compañía.

Como resultado, se plantea una métrica de evaluación distinta a minimizar el error cuadrático medio (MSE) para entrenar el modelo. A partir del error, es decir la diferencia entre el precio real de la vivienda y el precio estimado, se define una función personalizada para esta aplicación propuesta por los autores, descrita en la sección de modelos y resultados. Como resultado, el modelo óptimo generado es *XGBoost* mediante el cual se logró concluir con éxito la compra del 66.1% de las viviendas ofertada, con un gasto efectivo de $765.30mil millones de pesos COP. El valor promedio de la métrica personalizada anteriormente descrita para el mejor modelo fue de 1.02 siendo un modelo que maximiza el número efectivo de negocios llevados a cabo en relación con el carrete de especificaciones diseñadas, minimiza el gasto efectivo en la compra de inmuebles mientras que genera un trade-off efectivo con el gasto. En contraste con modelos de regresión se tiene una ganancia de 5% en el número total de negocios llevados a cabo correspondiente a 250 casas de más que se están comprando, se genera un ahorro de $284.76mil millones de pesos COP por lo que es un modelo eficiente y personalizado que resulta efectivo para inversionistas y empresas del sector inmobiliario interesados en adquirir propiedades.

Con la información anterior, se entrenaron modelos de regresión con regularización, modelos de ensamble *Random Forest* con árboles de regresión y *XGBoost*. De los anteriores, el modelo *XGBoost* entregó los mejores resultados y el entrenamiento del modelo bajo la métrica antes sugerida presenta grandes ventajas frente a utilizar el error cuadrático medio.

La construcción de la base de datos utilizada para realizar el entrenamiento de los modelos se basó principalmente en 2 grupos de predictores: variables de las características de las viviendas, y datos externos derivados de variables disponibles en el censo, así como también predictores construidos a partir de los datos espaciales obtenidos de la geometría de las ciudades. En la siguiente Tabla se observan algunas estadísticas descriptivas de las características del hogar en Bogotá y Medellín.

Tabla 1. Estadísticas Descriptivas

	Promedio	Mediana	Desv.Est.	Mín.	Máx.
Metros	178.97	126	342.76	5	28,428
Cuartos	3.01	3	1.36	0	11
Baños	2.86	3	1.11	1	20

Como se muestra en la Tabla 1, en promedio un inmueble de la base de datos cuenta con 179 metros cuadrados, 3 cuartos y 3 baños. Sin embargo, hay alta variabilidad dentro de la muestra. Otra de las variables que será relevante para determinar el precio de la vivienda es el estrato. En la Figura 1 vemos la variación del logaritmo del precio del inmueble y el estrato. Entre los estratos 2 y 6 observamos que el precio de la vivienda incrementa a medida que lo hace el estrato, pero la relación se rompe en el estrato 1. Lo anterior se debe a la baja

representación del estrato en la muestra, pues tan solo 390 de los 51,437 inmuebles son de estrato 1. Además, probablemente hay cierto sesgo en la muestra, ya que se espera que los hogares de mayor ingreso, y por ende con inmuebles más caros, sean los que conozcan plataformas como *Properati* y suban sus propiedades.

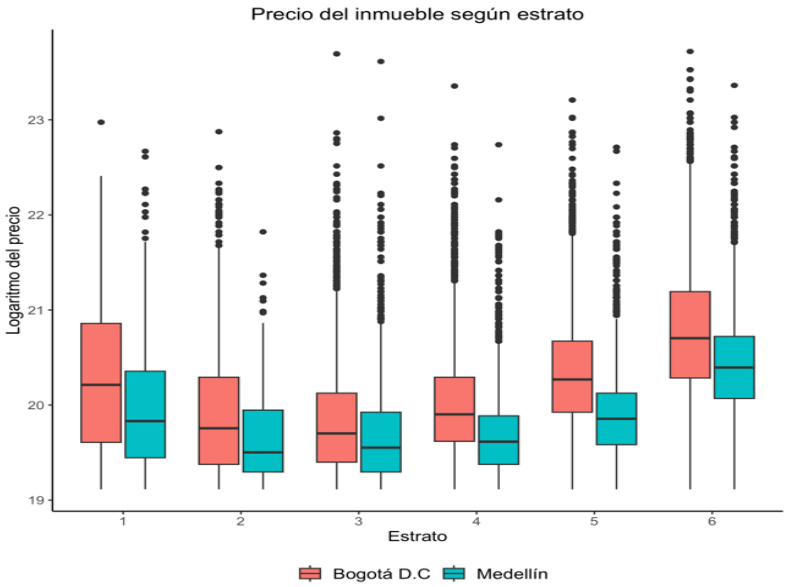
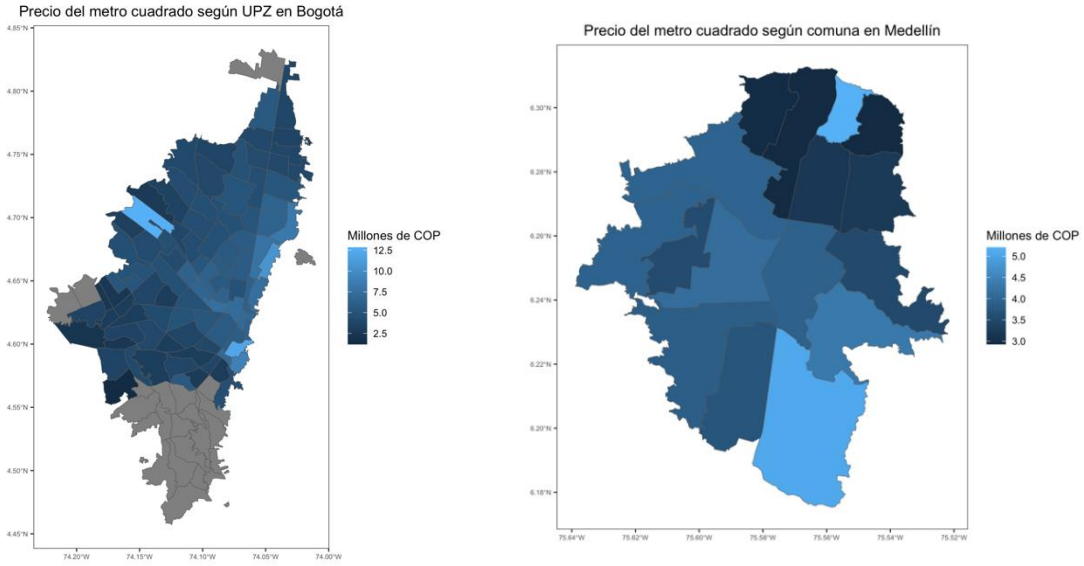


Figura 1: Precio del inmueble según el estrato en Bogotá y Medellín

Como ya mencionamos, los datos utilizados para el ejercicio se obtuvieron de la plataforma *Properati*. Tanto para Bogotá como para Medellín se cuenta con información de precios de los inmuebles. Como se observa en las Figuras 2 y 3, que detallan los precios por metro cuadrado, hay variabilidad relevante dentro de las ciudades del precio del metro cuadrado por ubicación. Además, el rango de precios en Bogotá es mucho más amplio que el de Medellín.



Figuras 2 y 3: Precio del metro cuadrado según UPZs de Bogotá y comunas de Medellín

Ahora bien, para los datos espaciales se tuvieron en cuenta los 'amenities' que corresponde a una categoría de características geográficas disponibles en la librería Open Street Maps en las que se puede encontrar información geográfica de diferentes rubros significativos tenido en cuenta al momento de estimar el precio de

una vivienda. Entre los más importantes se tuvieron en cuenta restaurantes, universidades, bancos, colegios, guarderías, hospitales, centros de policía, entre otros.

Para generar la base de datos espaciales inicialmente se construyó la caja de coordenadas para las ciudades. En seguida se extrajo para los rubros más significativos mencionado anteriormente las propiedades 'Simple Features' que describen una geometría y sus respectivos atributos. Luego se extrajeron los datos vectoriales de tipo polígono para las amenities consideradas, se filtraron los datos de las viviendas de los datos de entrenamiento dentro del polígono de las ciudades de Bogotá y Medellín. Finalmente se procedió a realizar el cálculo de la matriz de distancias entre los datos geográficos de los amenities y las viviendas en los datos de entrenamiento disponibles. Finalmente se extrajeron los predictores de las matrices de distancia mínima a cada uno de los amenities.

Por otro lado, se tuvieron en cuenta predictores del número de 'amenities' cercanas a la vivienda dentro de un radio de 300 metros, ya que se considera que tener un número significativo de rubros e indicadores económicos del hogar como restaurantes, bancos y centros comerciales cercanos a la vivienda valorizan el inmueble comercial. Para realizar lo anterior se realizó un buffer de la distancia anteriormente descrita utilizando la técnica de vecinos espaciales escogiendo el criterio de reina de adyacencia para obtener la matriz de pesos que genera la dependencia espacial.

Finalmente se anexaron a las bases de datos geométricos las matrices de distancia de las viviendas a los centros CBD de las ciudades extrayendo los datos geométricos. Estos predictores se escogieron porque diversos autores han reportado que en el modelo de ciudad monocéntrica para describir las dinámicas de los mercados urbanos el precio de la vivienda declina con la distancia al CBD. Del mismo modo está variable puede darnos una intuición sobre la densidad laboral de la ciudad en relación con el precio estimado de las casas.

Se escogieron los predictores de la matriz de distancia mínima a los diferentes amenities así como también el número de amenities en un radio determinado de 300 metros porque la dependencia de los datos espaciales a través de las matrices de pesos expresan las relaciones geográficas, socioeconómicas, geopolíticas entre otras de las viviendas y pueden ser indicadores significativos sobre el precio de las viviendas.

Adicional a las variables antes mencionadas sobre las características de la zona en que se ubican las viviendas, se construyeron predictores que representan particularidades propias del inmueble. Por un lado, la base de datos extraída de *Properati* aporta el número de habitaciones de forma completa, pero la superficie del inmueble, el número de baños y el total de cuartos cuentan con información parcial.

En consecuencia, para poder completar el número de baños y el tamaño de la superficie se inició haciendo uso de expresiones regulares que permitieran extraer las cantidades a partir del reconocimiento y extracción de diferentes patrones, junto con un procesamiento intermedio de la información extraída de la descripción otorgada por el usuario. Del mismo modo se desarrollaron funciones que permiten convertir la información extraída de la descripción de las viviendas e identificar patrones de números escritos en palabras, para dar un valor generado al proceso de pre- procesamiento y limpieza de las variables.

Aunque con lo anterior se logró llenar un gran número de datos que se encontraban faltantes, no fue suficiente. Por tanto, seguido al procedimiento de expresiones regulares, se utilizaron los vecinos espaciales utilizando el criterio de adyacencia de reina de forma que se asignaran los baños promedio de las propiedades ubicadas en 200 metros a la redonda, así como los metros cuadrados de aquellos inmuebles en el mismo radio, pero que compartieran además el mismo número de habitaciones.

Posteriormente, se utilizó el censo nacional del año 2018, realizado por el DANE para construir nuevas variables que dieran noción sobre las características propias de la vivienda y su entorno. Para el procedimiento, primero fue necesario asignar cada propiedad a las respectivas manzanas en que se encuentran, para posteriormente asignar el valor de alguna estadística de la manzana correspondiente. La estadística depende de la variable, si es categórica (como el estrato) se decidió utilizar la moda, si es numérica (como el número total de cuartos) se utilizó la mediana, para prevenir sobreestimaciones por algún dato atípico que pueda alterar el promedio. No

obstante, no todos los hogares lograron ser asignados a una manzana, por lo que nuevamente se requirió implementar la técnica de vecinos espaciales a partir de los pesos de la matriz de adyacencia.

En específico se obtuvo del censo el estrato de las viviendas, esta variable es de gran importancia debido a su fuerte relación con los precios, ya que usualmente en zonas de estratos más altos los precios de los inmuebles son más elevados. De igual forma se completa con el censo el número de cuartos totales de la vivienda, que junto con el metraje dan indicación de la amplitud de los espacios. También, se construye el número total de hogares en la vivienda, bajo la intuición de que en las viviendas más humildes es mayor la probabilidad de contar con más de un hogar y finalmente, se encuentra el número total de personas que habitan en la manzana, teniendo en cuenta que en las zonas donde habitan las personas con menores ingresos suele haber una mayor densidad poblacional.

Finalmente, algunas variables continuaron presentando valores faltantes, pero en ningún caso este número fue superior al 0.5% de los datos correspondientes. En consecuencia, dependiendo del tipo de variable y de su distribución, a criterio de los autores se completaron las imputaciones de la base haciendo uso de la media, la moda o la mediana de las demás observaciones de la variable correspondiente.

Las variables utilizadas en el entramiento/ajuste de los modelos de aprendizaje supervisado para los modelos de regresión y regularización consistió en el consolidado de las variables de las bases extraídas de los datos extraída de *Properati* después de realizado el procesamiento intrínsecas a la vivienda incluyendo los metros cuadrados, el número de baño, el tipo de vivienda, los predictores derivados de las matrices de la distancia mínima de los inmuebles a los amenities principales disponibles en OSM, las variables relativas a los pesos de la dependencia en los datos especiales entre el número de amenities en un radio de 300 metros predeterminados, y las variables del censo anteriormente descritas.

Para realizar el entrenamiento de los modelos de regresión y regularización se generó a partir de los datos de entrenamiento provenientes de la ciudad de Bogotá y Medellín se dividió en 3 grupos: datos de entrenamiento con los que se realizó el ajuste de los modelos, datos de validación mediante los cuales se realizó una evaluación preliminar de repertorio de modelos disponibles para contrastar las métricas de evaluación propuestas así como el promedio de la función del error para realizar la escogencia del mejor modelo, y los datos de prueba sobre los cuales se presentan las predicciones, resultados y análisis de resultados presentados en el documento.

Para los modelos de regresión y regularizados sobre la base con las variables mencionadas anteriormente se realizó la obtención de la ‘sparse matrix’ con las variables categóricas correspondientes a los predictores de estrato (1 -6) y tipo de vivienda (casa o apartamento). De manera análoga se estandarizaron las variables continuas para obtener una distribución de la variable con media 0 y desviación estándar 1, para controlar la magnitud de los estimadores generados por el modelo en relación con el orden de magnitud de los predictores. Los modelos de ensamble como *Random Forest* y *XGBoost* no fueron estandarizados, teniendo en cuenta la independencia del modelo siguiendo la intuición que el mismo modelo se encarga de escoger la importancia de los predictores al momento de ajustarlos.

En el presente documento se reporta un total de 9 especificaciones, de los cuales 3 son *XGBoost,* 2 *Random Forest,* 3 regresiones lineales regularizadas y una regresión lineal sin regularizar. Como se reporta en la Tabla 1, el mejor modelo fue el *XGBoost* denominado como modelo 1.

La selección del mejor modelo se basó en la observación de 3 métricas principales, a partir de la motivación descrita en el inicio de este documento. En primer lugar, se quería un modelo que permitiera cerrar la mayor cantidad de negocios posibles. Como se puede notar, los modelos de regresión permitían concretar la compra de poco más del 61% de las viviendas en venta, los modelos de *Random Forrest* completaban la compra del 63% de los inmuebles, los modelos de *XGBoost* entrenados para minimizar la raíz del error cuadrático medio (RMSE) permitían la compra de más del 64% de las viviendas, pero el modelo *XGBoost* entrenado bajo la fórmula personalizada descrita más adelante conduce a la compra de más del 66% de las propiedades de la plataforma, generando un valor agregado de aproximadamente 5 puntos porcentuales correspondientes a la compra de 250 inmuebles más que los resultados obtenidos mediante el modelo de regresión lineal.

En segundo lugar, se buscaba un modelo que en el agregado representara las menores pérdidas posibles de dinero. Como se puede notar en la Tabla 1 los modelos de regresión lineal, aun cuando efectúan la menor cantidad de compras de inmuebles, son los que mayores pérdidas presentan (reportadas en miles de millones de pesos) siendo los más ineficientes, esto debido a la naturaleza y simplicidad al ajustar los modelos.

Es relevante notar que los modelos de *Random Forrest* cuentan con menores pérdidas que los modelos *XGBoost.* No obstante, la diferencia es cercana a tan solo el 9%, por lo que se decidió dar prioridad a la compra de propiedades, dada la cercanía en el monto de las pérdidas.

Para realizar el entrenamiento del modelo 1 *XGBoost* seleccionado se basó en el algoritmo de ajuste de árboles por el método de *Boosting* en el cual cada arbol entrenado depende de la información del arbol anterior mientras que el arbol siguiente se ajusta con una versión modificada de los datos de entrenamiento originales. En pocas palabras el algoritmo sigue la intuición de que ajusta un árbol de decisión con el vector de residuales actual $r_i = (y_i - y_{pred})$, luego se suma el nuevo árbol de decisión a una función ajustada para actualizar los residuos. Por ende, se va ajustando arboles reiterativamente al vector de residuales mientras que se va mejorando el arbol base en las áreas en que tiene errores de predicción, generando un modelo exhaustivo con mayor costo computacional.

Para los B arboles de *Boosting* se tiene la siguiente función que aproxima los árboles, donde λ es la tasa de aprendizaje:

$$f^*(x) = f^*(x) + \lambda f^{*m}(x)$$

Finalmente se actualiza el vector de residuales con base a la siguiente ecuación:

$$r_i = r_i - \lambda f^{*m}(x)$$

El modelo final de *Boosting* que es la base de *XGBoost* es el promedio de los B modelos generados:

$$f_{boost}(x) = \frac{1}{B}\sum_{1}^{B} \lambda f^{*m}(x)$$

El ajuste mediante $XGBoost$ que es el caso de Boosting extremo penaliza de manera activa la optimización del modelo en cada paso de la función objetivo generando la siguiente función de costo, donde $\Omega(f_k)$ penaliza la complejidad del modelo, T es el número de hojas en el árbol y w es la predicción/el score en la hoja:

$$L = \sum_{1}^{N} L(y_i, \hat{y_i}) + \sum_{1}^{m} \Omega(f_k)$$

$$\Omega(f_k) = \gamma T + \frac{1}{2} n \, |w|_2$$

Para realizar la sintonización de hiper parámetros se realizó un grid search exhaustivo contemplando las posibles combinaciones de parámetros que construyen el modelo. Los resultados obtenidos en computación para los hiper parámetros se detallan en seguida. Para los parámetros de *Boosting* de número de árboles y tasa de aprendizaje se obtuvo 2000 y 0.01 respectivamente. Estos parámetros se escogieron buscando obtener un trade-off entre complejidad y exhaustividad del modelo, así como también generando un control para evitar el sobreajuste de este, debido a que se busca generalizar las predicciones de los precios de los inmuebles para la ciudad de Cali. En la grilla de parámetros se iteró entre $0.01, 0.05$ para los valores de tasa de aprendizaje teniendo en cuanta que entre menor sea este hiper parámetro más efectivo el aprendizaje en el ajuste para el arbol siguiente, sin embargo, deben tener suficiente número de arboles para ajustar los mismo.

Para los parámetros de arbol se tuvo en cuenta la profundidad máxima de los árboles para valores entre $5, 7, 10$ debido a que diferentes autores y artículos basados en aplicación de Computer Science reportan que para estos niveles se tienen resultados robustos, a su vez se controla la complejidad del modelo.

La profundidad se determinó en 10. Finalmente, los parámetros de $gamma$ y peso mínimo de nodo hijo se establecieron en $1\ y\ 100$ respectivamente, teniendo en cuenta que diversas aplicaciones reportan que el mínimo número de simples en nodo debe ser entre $0.5 - 1\%$ del total de observaciones.

Para definir la función del error se conserva la forma funcional cuadrática cuando se sobreestima el precio es decir cuando el error de la predicción es positivo, pero cuando el precio predicho es menor al real, la penalización se hace de forma lineal (por lo que toma un valor negativo). Con lo anterior, el modelo va a tender a subestimar un poco más sus predicciones o en caso de sobreestimar, hacerlo en menor proporción. No obstante, no se pretende abstenerse de comprar propiedades con el pretexto de minimizar gastos, pues de ser así, sencillamente no habría negocio. Para lo anterior, se establece una penalidad más alta una vez se subestiman los precios por más de 40 millones de pesos, ya que en ese caso no sería posible realizar la compra del inmueble, generando así un truncamiento agresivo sobre la función del error definida a continuación:

$$error = precio_{predicho} - precio_{real}$$

$$métrica = \begin{cases} error^2 & si\ error > 0 \\ error^6 & si\ error < -40\ millones \\ error & de\ lo\ contrario \end{cases}$$

Como se evidencia de la forma funcional de la métrica es deseable su minimización. La Tabla 1 reporta nuevamente que el modelo de *XGBoost "modelo 1"* presenta el menor valor para este criterio, aunque es claro que, al ser un modelo entrenado justamente con esta función, en contraste con los otros modelos *XGBoost* que se entrenaron para minimizar el RMSE, su desempeño debe ser mejor, obteniendo los resultados esperados. Cabe resaltar que los *Random Forest* también se entrenaron para minimizar la métrica personalizada.

Lo anterior puede indicar que en la predicción de precios las no linealidades son de gran importancia, así como también predictores no lineal de índole social y socio económico contemplado en las bases definidas. Por ejemplo, dado que la proporción de personas que utiliza el transporte público suele aumentar a medida que el estrato disminuye, para las personas de menores estratos puede ser más valioso estar más próximos al transporte público, mientras que para algunas personas de estratos más altos esto puede resultar indiferente. Estas interacciones y demás no linealidades no son contempladas en los modelos de regresión lineal.

A partir de la Tabla 2 se observa que el modelo 1 de *XGBoost* óptimo se obtuvo las siguientes métricas: raíz del error cuadrático medio de 481.9, error absoluto medio de 2002.32 mientras que el promedio de la función del error propuesta para el modelo 1 fue de 1.02. En contraste el modelo 2 presenta un RMSE de 481.9, MAE de 202.32 y valor promedio de la función del error personalizada de 2.42 es decir más de 2.4 veces del error promedio generado, sin embargo, las pérdidas totales en miles de millones de pesos con el modelo 2 es de $\$757.60\ COP$ que es menor a $\$765.30\ COP$, sin embargo, se compran 1.4% menos casas.

El modelo 1 es genera el mayor número de negocios cerrados con el 66.1% de inmuebles adquiridos, mientras que el modelo 4 de Random Forest ajustado con 500 árboles, con una profundidad máxima de 10 capas genera las mínimas pérdidas de $\$700{,}0$ mil millones, sin embargo, se dejan de adquirir el 3% de las viviendas correspondientes a aproximadamente 150 generando una penalización severa sobre los objetivos y alcance del proyecto.

En la Figura 4, se observa la distribución del error de predicción del precio de los inmuebles para Bogotá y Medellín (se muestra el 97% de los datos, omitiendo valores atípicos muy distantes). Como se puede observar, si bien la mayoría de los datos se encuentran cercanos al 0 (línea negra punteada), aún existe una gran cantidad de precios subestimados y sobre estimados. Las líneas punteadas de color rojo representan errores en valor absoluto de 40 millones de pesos (subestimación o sobreestimación según sea el caso). En primer lugar, a la izquierda de la línea que demarca una subestimación de 40 millones de pesos se encuentra poco menos del 34% de los inmuebles, que representan negocios que no se cierran por ofrecer precios muy bajos. En segundo lugar, entre las 2 líneas rojas se agrupa el 29% de los inmuebles, por lo que aproximadamente un 37% de los precios sobrepasan la barrera de los 40 millones de pesos.

Si todos los negocios se cerraran, la función objetivo planteada lleva a un ahorro neto y una buena predicción de los precios, en promedio. No obstante, la sobreestimación es particularmente problemática debido a que la distribución del error realmente se ve truncada al no poderse ejecutar la compra de inmuebles que se subestimen en más de 40 millones, pero sí se lleva a cabo la adquisición de propiedades sin importar el valor de la sobreestimación. En consecuencia, la verdadera distribución del error se trunca en la línea roja del lado izquierdo y conserva todos los valores que se encuentran a su derecha. Si bien podría penalizarse con mayor fuerza el sobrepasar un umbral de sobreestimación, en la práctica (aunque no se reporta en este informe) se culmina en un menor número de inmuebles adquiridos debido a que disminuye el costo de oportunidad de subestimar los precios.

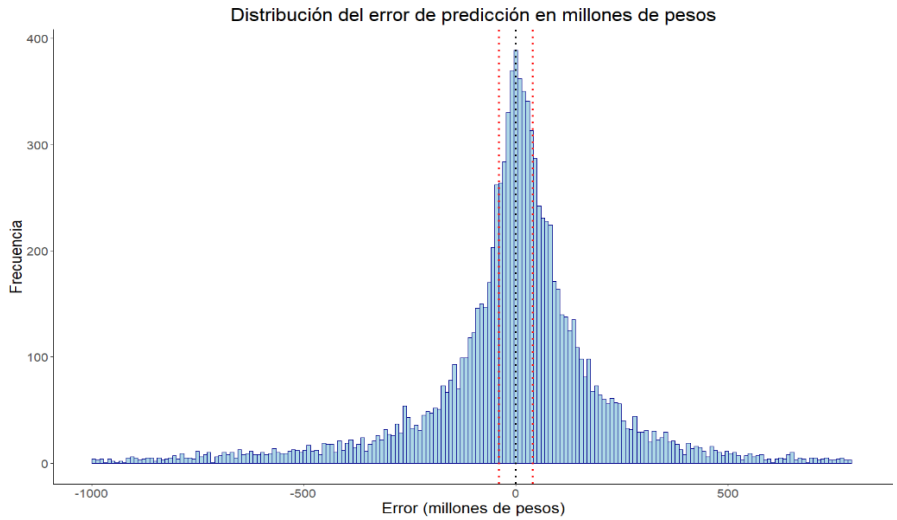


Figura 4: Distribución del error de predicción.
Frecuencia del error en millones de pesos COP

Por su parte, se destaca la superficie total de los inmuebles en metros cuadrados, el total de baños, habitaciones, cuartos y el estrato socio económico como los principales predictores que representan características propias de la vivienda. Igualmente, la distancia a las instalaciones religiosas, los cines, los moteles, los parqueaderos, los bancos y los lugares de cuidado para niños también resultan determinantes para predecir el precio.

Finalmente, teniendo en cuenta todo lo anterior, se hizo la predicción correspondiente para la ciudad de Santiago de Cali. En la Figura 5 se observa un histograma de color azul y un histograma de color rojo. El histograma de color azul representa las predicciones originales y la línea punteada del mismo color el promedio de los precios predichos. Empero, se hace reconocimiento de que el precio promedio de la vivienda en Cali es menor al precio promedio de las viviendas en Medellín o en Bogotá. Como consecuencia, teniendo en cuenta los valores de la Tabla 3 se ajustó la predicción a los valores presentados en el histograma de color rojo, con un valor promedio predicho de $555.3 millones representado por la línea punteada del mismo color.

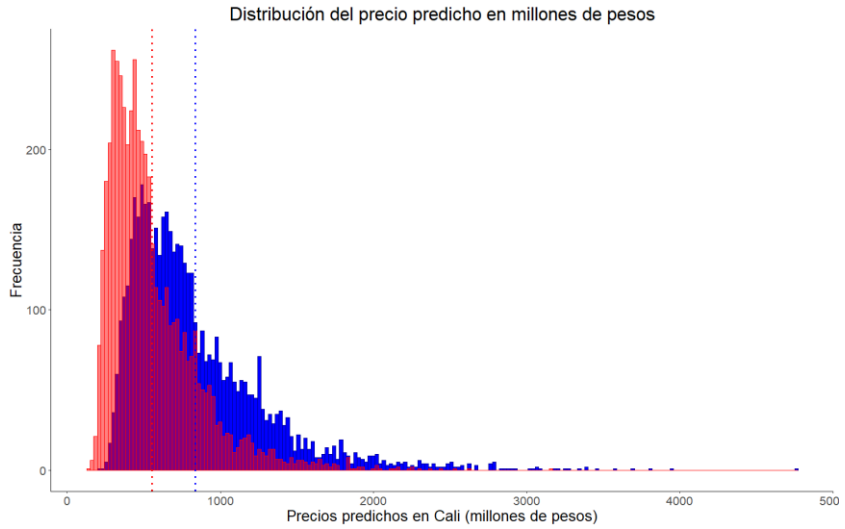


Figura 4: Distribución del precio estimado en millones de pesos de la Ciudad de Cali. Frecuencia del error en millones de pesos COP

CONLCUSIONES

Los modelos de predicción muestran que las variables más significativas para predecir el precio de los inmuebles son los metros cuadrados, el número de baños y el estrato con base en el proceso de importancia de variable derivados por modelos basados en árboles. A partir de los distintos modelos desarrollados encontramos que los árboles de regresión hacen predicciones más acertadas que los de regresión lineal, lo que sugiere que las interacciones entre las variables y no linealidades son relevantes, del mismo modo se observa que contrario a lo esperado que para aplicaciones en las que se construyen modelos de regresión para predecir precios donde normalmente los modelos de regresión lineal se ajustan de manera pertinente, los modelos de ensamble *Random Forest* y *XGBoost* están capturando las relaciones no linealidades intrínsecas a las variables socioeconómicas así como también las covariables derivadas de los predictores generados con datos espaciales, que se evidencia en un aumento de 5 puntos porcentuales en la compra de inmuebles y $284.76mil millones de pesos que se ahorran en gasto en relación con modelos de regresión lineal. A pesar de que los modelos de ensamble con *Random Forest* generan menores pérdidas significativamente en un ahorro de $65.30mil millones de pesos en relación con el modelo 4, se están dejando de comprar 150 inmuebles por lo que los autores concluyen en que es mejor tener mayor gasto, pero lograr los objetivos de comprar la mayor cantidad de inmuebles

En cuanto al modelo con *XGBoost*, vemos que los resultados obtenidos mejoran considerablemente al entrenar con la métrica personalizada de la función definida a trozos. De esta manera, la métrica es nuestra principal contribución frente al fiasco de Zillow. A pesar de que los resultados obtenidos son satisfactorios aún hay varios puntos de mejora en futuros ejercicios. Aún hay varios negocios que no se cierran y dado el truncamiento de la distribución del error, la sobreestimación resulta costosa, por lo que se puede buscar una forma de mejorar las predicciones reduciendo la sobreestimación sin sacrificar las compras realizadas. Por otro lado, se propone realizar una relajación en el momento de sobre ajustar entre $60 - 70M\ COP$ de modo que se maximice el número de viviendas adquiridas, o del otro lado en la región negativa relajando el truncamiento por un margen de $10 - 15M\ COP$ a la izquierda.

Por otro lado, para futuros proyectos que involucren datos espaciales se propone realizar el entrenamiento de los modelos mediante validación cruzadas mediante la construcción de $K - Fold$ bloques o clusters que agrupen los datos espaciales, así como también generan un *buffer* determinado durante el entrenamiento para evitar introducir sesgo o ruido en las predicciones. Por otro lado, para futuros proyectos se pudiese haber ajustado los modelos con los datos de entrenamiento sobre la ciudad de Bogotá, realizar el ajuste sobre las predicciones utilizando la información de Medellín como datos de validación para finalmente realizar las predicciones sobre la ciudad de Cali, generando modelos más robustos con menor dispersión.

Anexos

Modelo	Descripción	AVG	RMSE	MAE	% Casas compradas	$ perdido
Modelo 1	XGBoost con 2000 árboles; máxima profundiadad de 10; tasa de aprendizaje de 0.01; gamma = 1; min_child_weight = 100; proporción de columnas por árbol de 0.7; remuestreo con el 0.6. Utilizando la métrica personalizada.	1.02	481.9	202.32	66.1%	$ 765.30
Modelo 2	XGBoost con 2000 árboles; máxima profundiadad de 10; tasa de aprendizaje de 0.01; gamma = 1; min_child_weight = 100; proporción de columnas por árbol de 0.7; remuestreo con el 0.6. Utilizando el RMSE como métrica.	2.45	524.4	215.33	64.7%	$ 757.60
Modelo 3	XGBoost con 1500 árboles; máxima profundiadad de 10; tasa de aprendizaje de 0.005; gamma = 0; min_child_weight = 50; proporción de columnas por árbol de 0.7; remuestreo con el 0.6. Utilizando el RMSE como métrica.	2.82	539.4	223.4	64.1%	$ 785.30
Modelo 4	Random Forrest con 500 árboles, profundidad máxima igual a 10, 9 variables a probar en cada nodo (mtry), mínimo tamaño de nodo = 50.	3.91	606.4	243.4	63.0%	$ 700.00
Modelo 5	Random Forrest con 200 árboles, profundidad máxima igual a 10, 9 variables a probar en cada nodo (mtry), mínimo tamaño de nodo = 50.	3.96	607.7	243.8	62.9%	$ 700.07
Modelo 6	Modelo de regresión lineal con regularización Lasso. Lambda = 0.01	2.92	708.0	326.7	61.5%	$ 1,025.26
Modelo 7	Modelo de regresión lineal con Elastic Net. Alpha = 0.11; lambda = 0.02	3.00	706.2	326.1	61.3%	$ 1,028.10
Modelo 8	Modelo de regresión lineal con regularización de ridge. Lambda = 0.03	3.11	708.7	326.4	61.3%	$ 1,011.35
Modelo 9	Modelo de regresión lineal	2.89	703.2	325.8	61.2%	$ 1,050.06

Tabla 1: Especificaciones de modelos predictivos para estimar el precio de las viviendas. Se presenta la identificación del modelo, descripción con hiper parámetros y características, principales métricas de evaluación RMSE, MAE y función personalizada promedio AVG, porcentaje de casas compradas y gasto total del modelo.

Tabla 1. Amenities por ciudad

	Bogotá	Medellín	Cali
Bancos	215	32	44
Estaciones de bus	380	35	52
Clinicas & hospitales	252	143	62
Restaurantes & comidas rápidas	917	95	59
Universidades	79	59	9
Supermercados	41	13	7

Tabla2: Número de amenities para diferentes rubros, por ciudad

Ciudad	Promedio	Desviación
Bogotá	$ 869,755,897.00	$ 899,818,886.00
Medellín	$ 639,246,711.00	$ 623,222,205.00
Cali	$ 555,314,430.00	$ 601,842,533.00

Tabla3: Valor promedio de las casas por ciudad en millones de pesos COP.
Valor promedio y desviación estándar

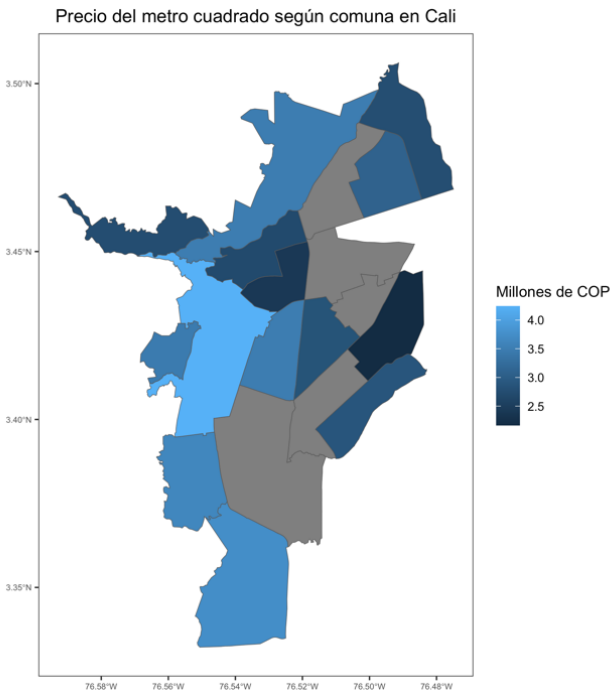


Figura 5: Precio del metro en la ciudad de Cali
Frecuencia del error en millones de pesos COP