DEPARTAMENTO DE COMPUTACION

Facultad de Ciencias Exactas y Naturales - UBA

TP3: Cuadrados Mínimos Lineales

Métodos Numéricos

Integrante	LU	Correo electrónico
Camjalli, Roni Ezequiel	12/18	rcamjalli@gmail.com
Rodriguez, Miguel	57/19	mmiguerodriguez@gmail.com
Itzcovitz, Ryan	169/19	ryanitzcovitz@gmail.com

Facultad de Ciencias Exactas y Naturales

Universidad de Buenos Aires

Ciudad Universitaria - (Pabellón I/Planta Baja)

Intendente Güiraldes 2610 - C1428EGA

Ciudad Autónoma de Buenos Aires - Rep. Argentina

Tel/Fax: (+54 11) 4576-3300

https://exactas.uba.ar

Índice

1. Introducción

1.1. Resumen

En este informe vamos a desarrollar una herramienta para para aproximar diferentes características de inmuebles. Se va a desarrollar un algoritmo de clasificación al cual entrenaremos con una base de 240.000 avisos de inmuebles con datos ya conocidos que luego nos permitirá predecir valores sobre inmuebles que no están presentes en nuestra base de datos. Luego del desarrollo realizaremos una serie de experimentos que nos permitirá ver diferentes formas de utilizarla y como optimizarla para obtener mayor precisión.

1.2. Introducción teórica

En la actualidad este tipo de herramientas se utiliza mucho en el mundo de los inmuebles ya que al tener una gran base de datos confiables se pueden clasificar de mejor forma los precios. Otro uso que se le puede dar con respecto a los inmuebles es poder segmentar las diferentes zonas con respecto a sus características, como por ejemplo metros totales, antigüedad, baños, entre otras. En otro rubro en el que es muy utilizado esta herramienta es en el automotriz, ya que cuenta con muchas características que hacen variar los precios del vehículo, y con esta herramienta desarrollada podremos adaptarla fácilmente a cualquier rubro que la necesite.

Contaremos con un conjunto de datos de avisos inmobiliarios de México. Este conjunto de datos tiene distintas características como las siguientes:

- id: Número identificador del aviso sin otra semántica particular
- titulo: Contiene el titulo del aviso inmobiliario
- descripción: Contiene una descripción en texto plano del aviso.
- tipodepropiedad: Variable categórica para definir el tipo (Casa, Apartamento, etc).
- Ubicación: varias características auto explicativas (ciudad, provincia, etc) sobre la ubicación del inmueble.
- Características numéricas: varias características cuantificables (baños, piscina, escuelas cercanas, etc).

Para llevar a cabo esta herramienta de clasificación vamos a utilizar la técnica de Cuadrados Mínimos Lineales y de esta forma aproximar las características de los inmuebles. Mediante esta técnica vamos a querer encontrar una función que "mejor aproxime" los datos. Dado un conjunto de funciones O_1,..., O_n linealmente independientes, definimos:

$$F = f(x) : \sum_{i=1}^{n} (C_i - Oi)^2$$

donde el problema de Cuadrados mínimos lineales (CML) es:

$$\min_{f \in F} \sum_{i=1}^{n} (f(x_i) - y_i)^2$$

El problema de los cuadrados mínimos siempre tiene solución ya que el resultado b existe en la imagen de la matriz de funciones. Sea M la matriz que contiene las funciones, queremos calcular $Mx = b$. Tomando $a = \text{mín}_x(||Mx - b||_2) \iff M^T Mx = M^T b$. Y que las columnas de M son linealmente independientes podemos calcular $(M^T M)^{-1}$ por ser cuadrada existe inversa y la solución de CML es $x = (M^T M)^{-1} * M^T b$. De esta forma podemos predecir de manera efectiva el problema de cuadrados mínimos.

Luego de realizar la aproximación correspondiente sobre una variable que queremos estimar, usamos un conjunto de datos que no hayan sido usadas en el entrenamiento para poder darle un uso a la herramienta desarrollada. Utilizaremos dos métricas: RMSE y RMSLE.

1.3. RMSE

La métrica Root Mean Squared Error (RMSE) es ampliamente utilizada, pero no suele ser ideal en ciertos casos, ya que pesan mucho mas las muestras con valores altos que las que tienen valores mas bajos. Mejorar la aproximación de una muestra alta, tiene mucha más influencia en la métrica RMSE que lograr el mismo objetivo en una muestra con valores menores.
Dado un modelo $\hat{f}$ y una observación $(x_{(i)}, y_{(i)})$, definimos $\hat{y}_{(i)} = \hat{f}(x_{(i)})$ y $e_{(i)} = y_{(i)} - \hat{y}_{(i)}$. Con esto, podemos calcular el RMSE del modelo $\hat{f}$ como:

$$RMSE(\hat{f}) = \sqrt{\frac{1}{N} \sum_{i=1}^{N} e_{(i)}^2}$$

1.4. RMSLE

Debido al problema de pesos en valores mas altos que bajos presentamos la métrica Root Mean Squared Log Error (RMSLE). Esta tiene como propiedad pesar de la misma forma la mejora porcentual sobre cualquiera de las muestras sin importar su valor absoluto.

RMSLE se calcula de la siguiente forma:

$$RMSLE(\hat{f}) = \sqrt{\frac{1}{N} \sum_{i=1}^{N} (\log(y_{(i)} + 1) - \log(\hat{y}_{(i)} + 1))^2}$$

1.5. Coeficiente de determinación

La métrica R^2 (o coeficiente de determinación) mide la proporción de la varianza en la variable dependiente que es predecible a partir de la variable o variables independientes. Sea $\bar{y}$ el promedio de los valores y_i y f_i el valor esperado para cierta predicción y_i

R^2 se calcula de la siguiente forma:

$$R^2(\hat{f}) = 1 - \frac{\sum_{i=1}^{N}(y_{(i)} - f_i)^2}{\sum_{i=1}^{N}(y_{(i)} - \bar{y})^2}$$

2. Desarrollo

2.1. K-Folds

A lo largo de todo el informe vamos a utilizar la técnica de K-Folds para realizar la experimentación, utilizando 80 % de los datos para entrenamiento y el 20 % restante para validación.

2.2. Segmentación

Uno de los problemas del método de Cuadrados Mínimos Lineales es que este algoritmo intenta explicar todos los datos con una única función y el conjunto de datos que se utiliza suele ser grande y muy variado, por lo que sería muy difícil conseguir una buena aproximación mediante este método para todos los datos. Para resolver este problema, recurrimos a la segmentación de datos. La idea de segmentación es enfocarse en conjuntos mas controlables. Por ende, logramos reducir la complejidad y al aplicar CML podremos predecir con mejor precisión. Para solucionar nuestro problema en específico una opción viable es segmentar el conjunto de datos, por ciudades, provincias, entre otras.

2.3. Feature Engineering

Al trabajar con este conjunto de datos, nos puede servir no solamente utilizar las características numéricas que tiene (como habitaciones, antigüedad, baños, metros totales, etc.) si no que también podemos extraer información a partir de el titulo y su descripción, y generar un nuevo conjunto de características a utilizar para poder predecir el precio de una vivienda.

3. Experimentación

3.1. Análisis de métricas RMSE y RMSLE

A continuación vamos a realizar el mismo experimento en dos ciudades diferentes para luego analizar sus métricas RMSE y RMSLE para ver por que se diferencian. Las dos ciudades que utilizaremos son Toluca y Cancún, seleccionamos estas ciudades ya que en una ejecución para predicción de precios únicamente con los metros totales segmentado por ciudades, notamos que el RMSLE entre estos dos es casi igual. A continuación se pueden ver los resultados de estas ejecuciones para cada ciudad:

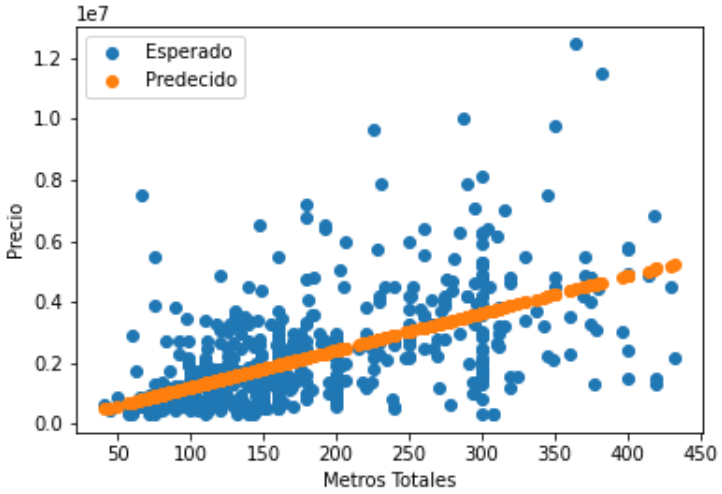


Figura 1: Metros Totales vs Precio en la ciudad de Cancún

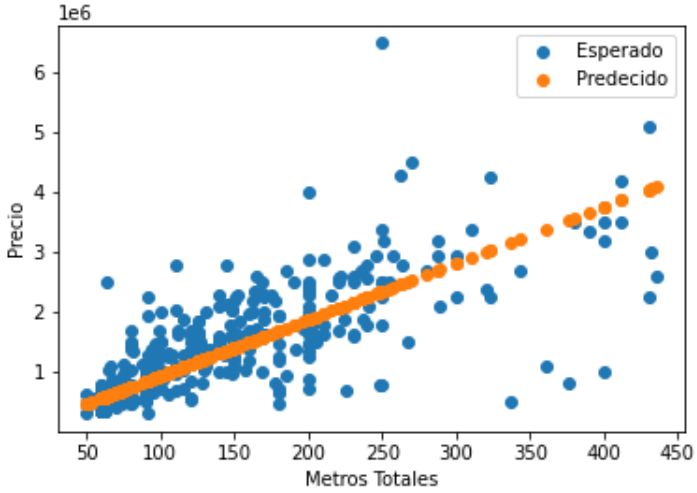


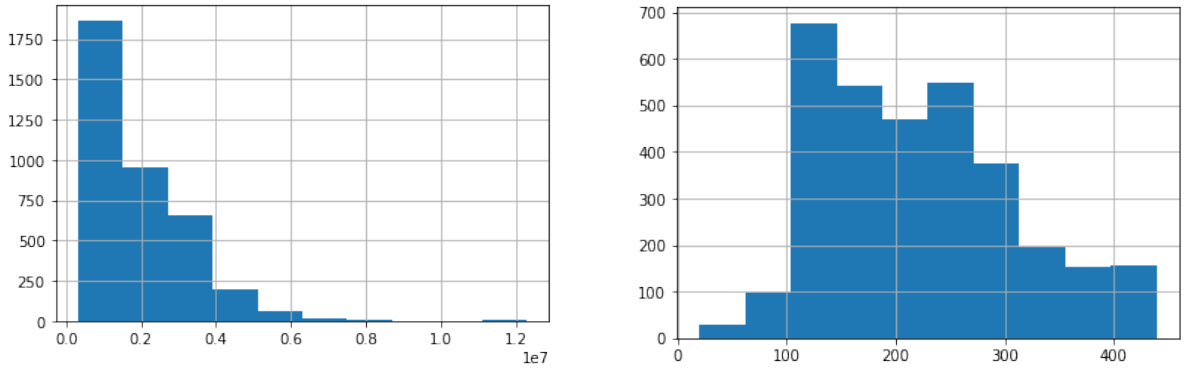
Figura 2: Metros Totales vs Precio en la ciudad de Toluca

En el gráfico de la ciudad de Cancún (1) podemos observar que hay muchos outliers en donde el inmueble es mucho mas caro en comparación con el resto, mientras que en el gráfico de la ciudad de Toluca (2) no hay tantos outliers. Llamamos outliers a aquellos inmuebles que luego de la segmentación y analizando en un categoría en especifico, por ejemplo el precio, son aquellos valores numéricos que están muy distantes del resto de los datos y podemos evaluarlos con el percentil, por ejemplo cortando el percentil 95, es por eso que en este experimento decidimos ver que ocurría cuando comparábamos dos ciudades en la cual una tenia mas outliers que el otro. Queremos observar como impactan estas diferencias en las métricas ya que con RMSLE ambas obtienen resultados relativamente similares, con 0.884 para Cancún y 0.833 para Toluca, pero cuando calculamos el RMSE notamos que en Cancún es de 2008586 y en Toluca 1269905, casi el doble. Esta diferencia en el RMSE se debe a que

tienen mas peso los precios mas altos en relación con los precios mas bajos, y como Cancún tiene mas outliers no es muy conveniente usar esta métrica para ver el error de la predicción.

3.2. Metros Totales vs Precio

En el siguiente experimento analizaremos la relación entre los metros totales de un inmueble con su precio. La motivación de este experimento es poder ver que tan buena es la predicción del precio cuando usamos los metros totales de todos los inmuebles de la base de entrenamiento luego de la segmentación por ciudad. Para este experimento vamos a segmentar por todos los inmuebles del dataset a aquellos que se encuentren en la ciudad de Chihuahua para poder concentrar todo en una mas pequeña y mejorar la precisión. Las distribuciones de los datos con los que vamos a trabajar son las siguientes:



(a) Precios (b) Metros totales

Figura 3: Histogramas de los datos usados en el experimento

Luego de llevar a cabo el experimento obtuvimos los siguientes resultados:

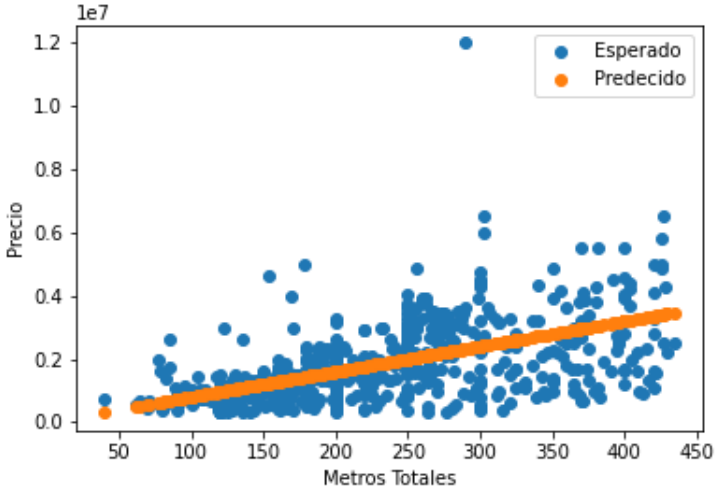


Figura 4: Metros Totales vs Precio en la ciudad de Chihuahua

En la figura 4 podemos ver la linea naranja que es la función que mejor predice el precio de un inmueble luego del entrenamiento, ya que es la que se calcula haciendo CML siendo la recta que minimiza la distancia con todos los puntos del gráfico. Siempre van a haber una minoría de puntos que no serán representados por la recta, esto se debe ya que a pesar de hacer segmentación por ciudades podemos seguir teniendo un conjunto de datos heterogéneo, por ejemplo inmuebles que sean mucho mas caros que el resto, por lo que la función no la va a alcanzar ya que no es muy probable que sucedan esos casos. A su vez podemos observar como esta linea pasa cerca de la mayoría de los inmuebles de validación, por ende los datos de entrenamiento fueron de gran ayuda, mas allá de que sigan habiendo outliers al predecir los precios de otros inmuebles desconocidos.

Las diferentes métricas observadas fueron:

- RMSE: 1413402
- RMSLE: 0.839
- R^2: 0.302

El siguiente paso en el experimento va a ser eliminar los precios mas caros que suponemos que nos desbalancean la predicción, y de esta forma ver si logramos disminuir el error. De la misma base de datos utilizada anteriormente segmentamos por los inmuebles de la ciudad de Chihuahua quedándonos con el percentil 93 de los datos, los cuales se corresponden con aquellos inmuebles cuyo precio es menor a 4.000.000 MXN. Tomamos esta decisión ya que luego de observar el primer gráfico notamos que había una minoría de datos que sobrepasaban ese precio lo que podía causar una desvío en el cálculo de los precios. Luego de ejecutar el experimento modificado obtuvimos el siguiente gráfico:

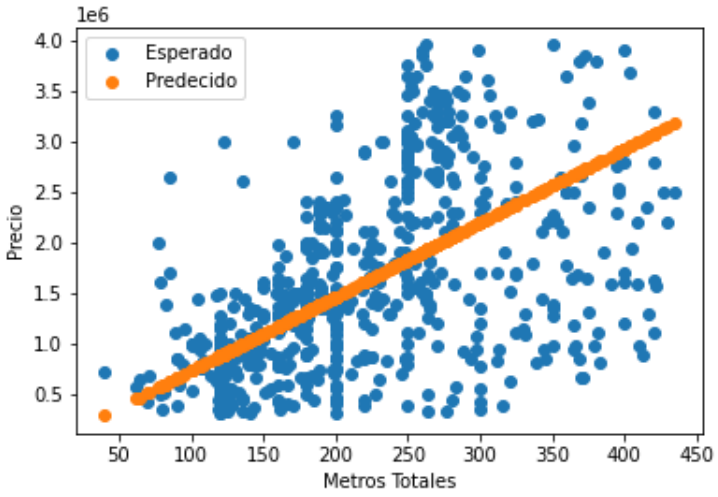


Figura 5: Metros Totales vs Precio en la ciudad de Chihuahua sin outliers

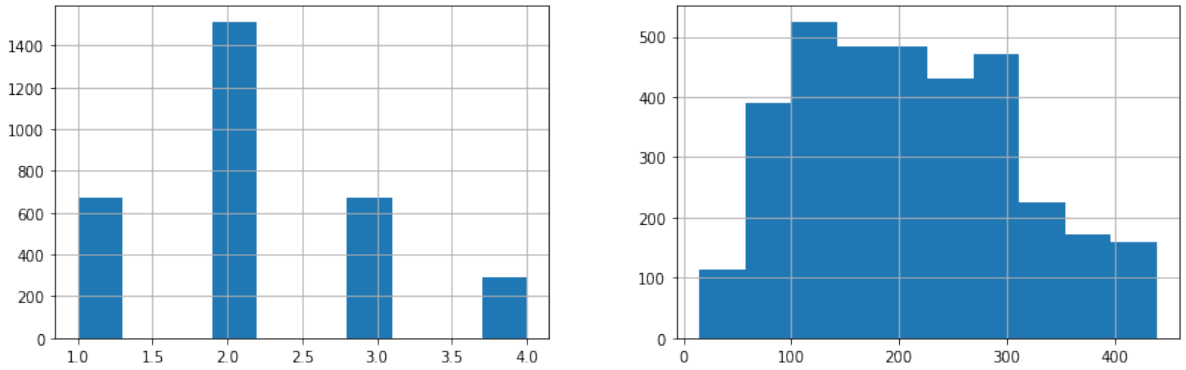
En la figura (5) podemos notar que ya no tenemos outliers significantes, que la función de predicción sigue siendo muy parecida a la de (4), por lo que ahora analizaremos si hubo un cambio con respecto a las métricas.

- RMSE: 1129817
- RMSLE: 0.784
- R^2: 0.257

Podemos ver que la métrica RMSE bajo notablemente luego de quitar los outliers, mientras que el RMSLE no fue tan notorio el cambio ya que se trata sobre un error mas porcentual. Podemos concluir que para la ciudad de Chihuahua el análisis de la relación metros totales con el precio nos da la misma función de predicción con y sin outliers y las métricas se corresponden con el hecho de haber o no.

3.3. Metros Cubiertos vs Baños

A continuación llevaremos a cabo un experimento en el que intentaremos predecir la cantidad de baños de un inmueble en base a la cantidad de metros cubiertos. La motivación de este experimento es poder analizar que tan bueno es el sistema para predecir en base a datos con un rango de valores variado, otros que tienen un rango muy pequeño. Al igual que en el experimento anterior, segmentamos de todo el dataset por la ciudad de Chihuahua para obtener datos mas precisos. Las distribuciones de los datos con los que vamos a trabajar son las siguientes:



(a) Baños

(b) Metros cubiertos

Figura 6: Histogramas de los datos usados en el experimento

Luego de llevar a cabo el experimento obtuvimos los siguientes resultados:

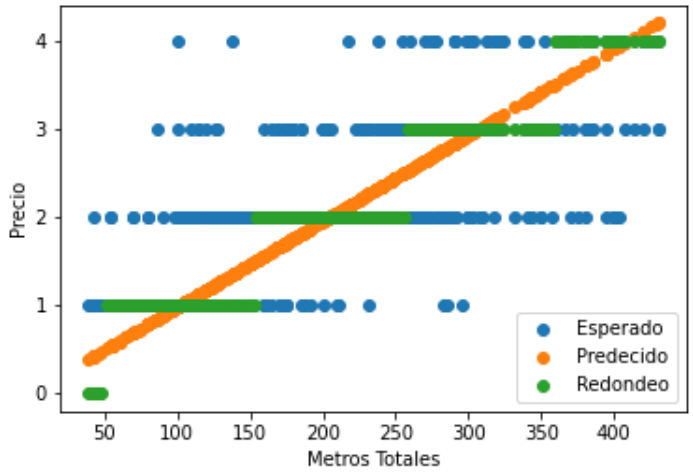


Figura 7: Metros Cubiertos vs Baños en la ciudad de Chihuahua

El la figura 7 podemos ver la linea naranja que es la función que mejor predice el precio de un inmueble luego del entrenamiento. Al obtener una función lineal tal vez no sea la mejor forma de observar si es un buen modelo para predecir los baños por lo que le agregamos un sistema de redondeo de la predicción para poder llevarlo a un mundo real en el que tenga sentido la predicción. Por ejemplo si el valor es 1.5 se redondeara a 1, pero si la predicción es 2.6 se redondea a 3. Las lineas verdes corresponden a ese sistema recién explicado, y podemos observar que hace un buen trabajo ya que diría lo que se puede esperar, que a mayor cantidad de metros cubiertos hay mas baños. Entonces con ese modelo desarrollado obtendrá el mejor resultado posible sobre inmuebles desconocidos.

Luego de la experimentación se calcularon las métricas obtenidas fueron:

- RMSE: 1.276
- RMSLE: 0.442
- R^2: 0.300

En conclusión, podemos decir que una función lineal no siempre es la mejor forma de predecir cualquier característica, ya que puede pasar que los valores a predecir sean enteros y una predicción con "coma" puede sacarla del mundo en el que se la estudia. Para hacer mas real el sistema podemos integrar un redondeo de datos y así poder obtener datos acordes los valores que se esperan.

3.4. Predicción Cantidad de Habitaciones

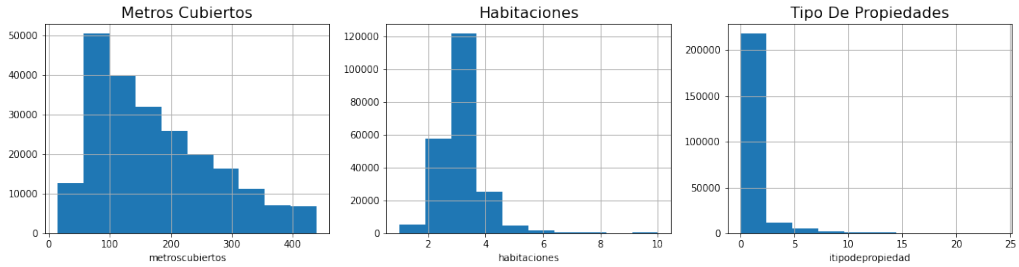
En este experimento nos va a interesar ver si es posible predecir la cantidad de habitaciones y, de ser posible, analizar cual es la mejor manera. Como eje principal nos va a interesar hacerlo a partir de los metros cubiertos y de los tipos de propiedades. Vamos a realizar una segmentación por tipo de propiedad para poder generar resultados que se ajusten mejor al mundo real.

En base a esto sin mucho análisis probamos utilizando *metroscubiertos* para las predicciones y segmentar todo el dataset por tipo de propiedad. En la tabla 1 vemos los resultados de las métricas; Podemos observar que los resultados arrojados no son muy buenos, por ejemplo el RMSE da mas de 2 que eso nos habla que el error medio es de 2 Habitaciones. Si hacemos un análisis de porque sucede esto, una buena hipótesis es que el conjunto de datos a predecir no tiene un comportamiento lineal, y pensamos que segmentando todavía mas, seguro podemos obtener una mejor aproximación.

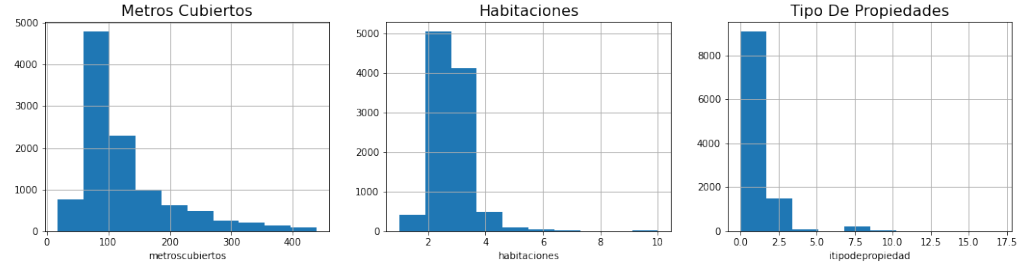
Ciudad	**RMSLE**	**RMSE**
Todas	0.519909	2.045370

Cuadro 1: Comparación de métricas obtenidas para la predicción de habitaciones segmentando por tipo de propiedad y los metros cubiertos, para todo el dataset.

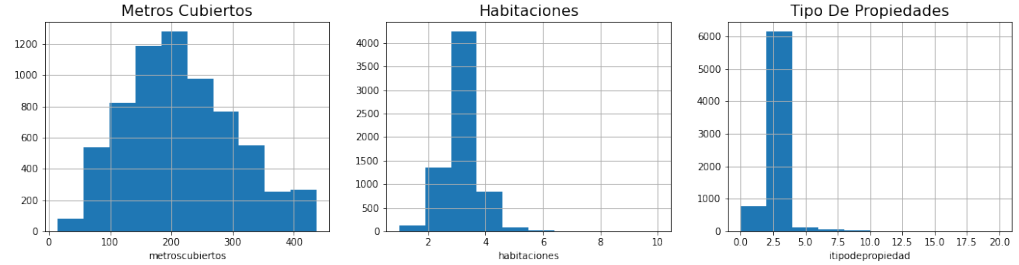
Entonces vamos a realizar un análisis de la distribución de nuestras variables realizando cortes por segmentación. En la figura 8 podemos ver la distribución de las variables Metros Cubiertos, Habitaciones y Tipo de propiedad tomando distintos cortes sobre el dataset. En la imagen 8a tenemos las distribuciones para todo el dataset, nos va a servir para comparar con los otros cortes. Luego vemos en la imagen 8b donde podemos observar la distribución analizando solo la ciudad de Benito Juárez, y podemos ver como tanto los metros cubiertos como las habitaciones cambian, los metros cubiertos tienen mayor proporción de propiedades chicas y para las habitaciones, a diferencia del dataset completo hay tantas propiedades con 2 habitaciones como de 3. Siguiendo con la imagen 8c vemos la distribución solo para la ciudad de Mérida, donde los metros cubiertos tienen una clara distribución normal que es muy diferente a la de Benito Juárez y el dataset, y además el tipo de propiedad tiene también una diferencia de distribución. Por lo tanto solamente observando estas dos ciudades podemos ver que cada ciudad tiene comportamientos distintos, entonces a lo mejor realizando una segmentación por ciudad obtendremos mejores resultados.



(a) Usando el dataset completo.



(b) Solo para la ciudad Benito Juárez.



(c) Solo para la ciudad Merida.

Figura 8: Histogramas de distribución variables para distintos cortes del dataset.

Luego de esto realizamos unas primeras pruebas con estas dos ciudades para ver si efectivamente logramos una mejora. Obtenemos métricas sobre estas pruebas y comparamos en el cuadro 2, donde observamos que realizando corte en las ciudades obtenemos métricas mas bajas para dichas pruebas.

Ciudad	**RMSLE**	**RMSE**
Todas	0.519909	2.045370
Benito Juárez	0.406600	1.678228
Mérida	0.340390	1.097243

Cuadro 2: Comparación de métricas obtenidas para la predicción de habitaciones a partir de los metros cubiertos, con segmentación por tipo de propiedad, para las ciudades de Mérida y Benito Juárez.

Luego de esto realizamos pruebas con todas las ciudades, calculamos métricas para cada una y tomamos la media entre todas; En la tabla 3 donde vemos como la tendencia mejora

al segmentar en comparación con no hacerlo, por ende podemos afirmar que segmentar por ciudades es un mejor modelo.

Segmentación	**RMSLE**	**RMSE**
Ninguna	0.519909	2.045370
Ciudad	0.370575	1.335785

Cuadro 3: Comparación de métricas obtenidas para la predicción de habitaciones a partir del tipo de propiedad y los metros cubiertos, con segmentación.

En conclusión creemos que es posible realizar una estimación de la cantidad de habitaciones, segmentando por tipo de propiedad y tomando los metros cubiertos. Y además si realizamos una segmentación extra por ciudad podemos obtener mejores resultados.

3.5. Predicción del precio por metro

En el ámbito inmobiliario los precios de las propiedades se fijan en base a muchas características, pero principalmente si tenemos que hablar sobre el valor de una propiedad hablamos del precio por metro. Esta característica nos permite comparar el valor de la propiedad independientemente de su tamaño, si uno tuviese el precio por metro de una propiedad luego simplemente sabiendo la cantidad de metros y multiplicando obtiene su precio. Esta característica suele estar muy relacionada a la ubicación de la propiedad, por ejemplo en las grandes ciudades donde la demanda es muy grande el valor suele ser mas alto que en las afueras de la ciudades. Por esto nos pareció interesante ver como podemos predecir este valor en base a la ubicación en el mapa de la propiedad. Definimos la característica $precioxmetro = \frac{precio}{metrostotales}$ y realizamos una primera visualización en el Distrito Federal sobre la distribución del precio por metro en la figura 9. Efectivamente nos muestra nuestra suposición sobre que el precio por metro esta relacionado con la ubicación de la propiedad, ya que vemos que en las zonas mas céntricas los valores aumentan.

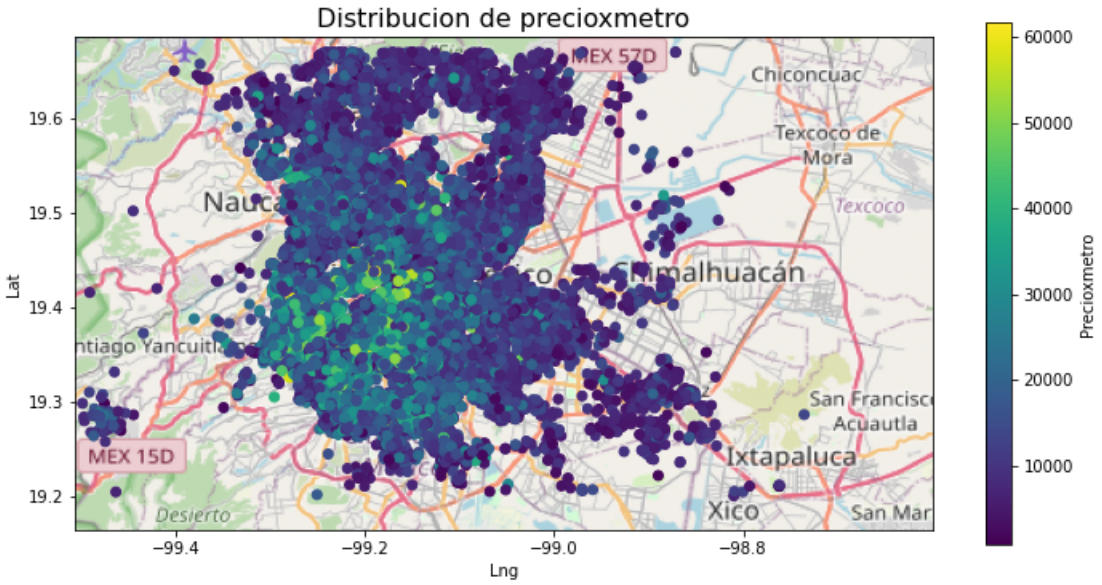


Figura 9: Distribución del precio por metro en el distrito federal.

Luego de ver esto realizamos una primera predicción en la zona del distrito federal, tomando la latitud y longitud como valores de entrada para entender como se comportaría el modelo. En la figura 10 vemos la comparación dibujado en el mapa entre los datos esperados y las predicciones. Podemos observar que la predicción no es tan buena a simple vista. Por eso mismo nos preguntamos si capaz realizando una segmentación logramos mejorar esto, entonces entendimos que seguramente una segmentación por el tipo de propiedad nos puede ayudar ya que los precios por metro varían mucho dependiendo si hablamos de casas, departamentos u oficinas por ejemplo. Así que realizamos las pruebas nuevamente segmentando y vemos en la figura 11 los resultados. A priori vemos una predicción mas natural, sin embargo sigue la misma tendencia anterior a darle precios mas altos a las propiedades que se encuentran mas por debajo, pero esta vez al separar el tipo de propiedad evitamos esa masa homogénea de precios y empezamos a ver un comportamiento mas parecido al esperado donde esta todo mezclado.

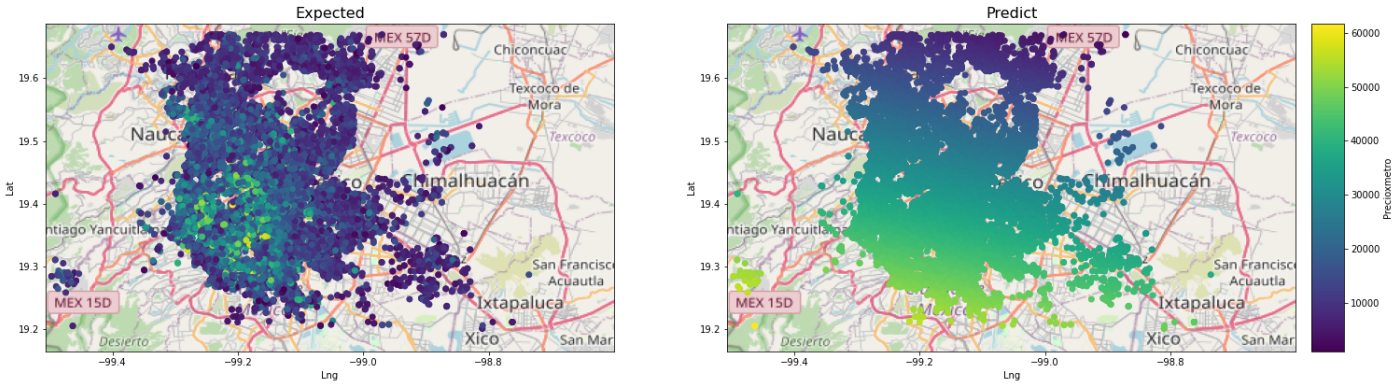


Figura 10: Distribución del precio por metro vs. Predicción del modelo en el distrito federal.

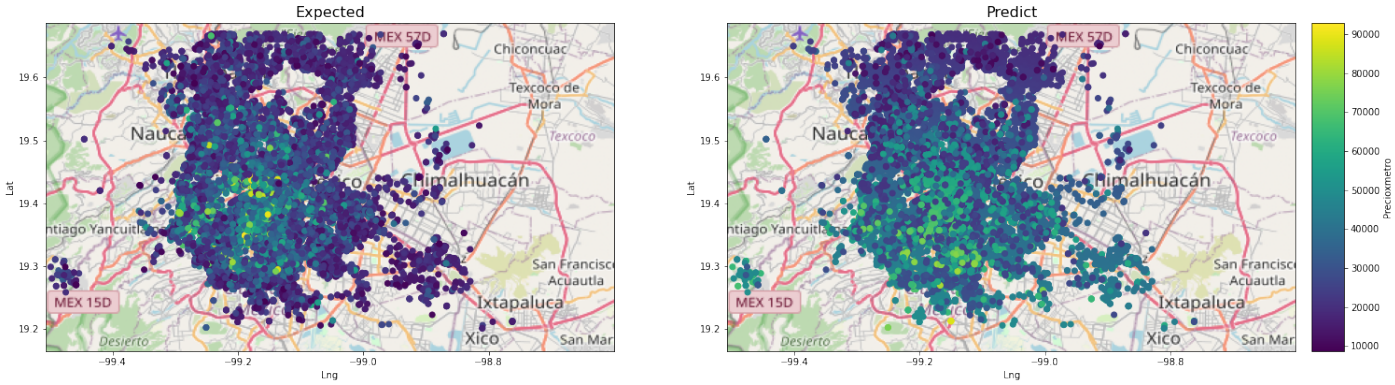


Figura 11: Distribución del precio por metro vs. Predicción del modelo en el distrito federal, segmentado por tipo de propiedad.

Ahora veamos un poco de métricas de estas pruebas en la tabla 4 donde vemos que obtuvimos mejores métricas con la segmentación donde baja el RMSLE y el RMSE, pero en R2 Score no mejora ya que se acerca al 0 y debemos estar atentos a esto.

Segmentacion	RMSLE	RMSE	R2 Score
Ninguna	0.626523	11547.674194	0.381613
Tipo de propiedad	0.523907	9980.993296	0.182326

Cuadro 4: Comparación de métricas obtenidas para la predicción de habitaciones a partir del tipo de propiedad y los metros cubiertos en el Distrito Federal, con segmentación y k-folds cross validation.

Ahora que vimos que el modelo podría funcionar, vamos a escalarlo a todo el territorio de México y vamos a comparar las métricas con y sin segmentación a ver si esta vez podemos obtener un modelo ganador.

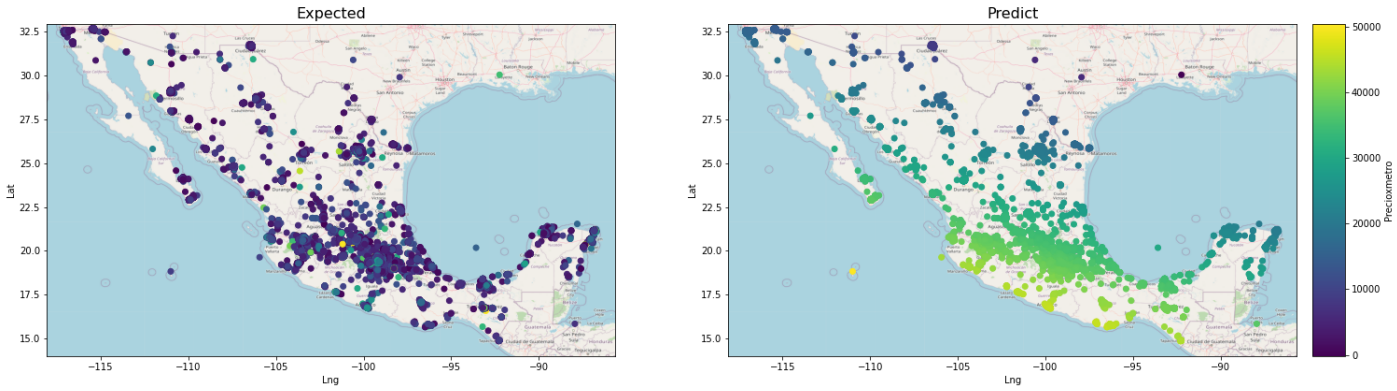


Figura 12: Distribución del precio por metro vs. Predicción del modelo en el todo el territorio.

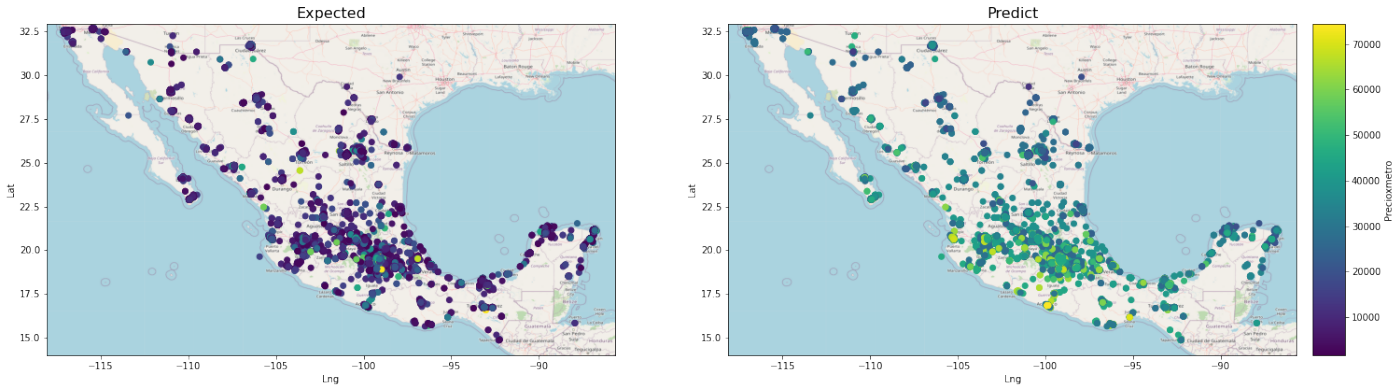


Figura 13: Distribución del precio por metro vs. Predicción del modelo en todo el territorio, segmentado por tipo de propiedad.

Luego de realizar las pruebas para todo el territorio de México vemos en la figura 12 el resultado sin segmentación y en la figura 13 los resultados con la segmentación. Logramos ver en ambos resultados que el modelo no se ajusta correctamente a todo el mapa sin segmentación alguna, nos genera la misma estimación que los precios del norte son mas caros que los del sur, en cambio cuando segmentamos por tipo de propiedad logramos una mejora. En la tabla 5 podemos ver las métricas obtenidas para estas pruebas. Vemos que la tendencia que vimos en Distrito Federal sigue marcada, pero vemos que esta vez la diferencia entre R2 Score no es tan marcada como antes, con lo cual nos hace pensar que a pesar que el modelo no le gana, a mayor territorio logra una mejor performance.

Segmentacion	**RMSLE**	**RMSE**	**R2 Score**
Ninguna	0.734876	9361.971349	0.315564
Tipo de propiedad	0.579330	7796.620572	0.248053

Cuadro 5: Comparación de métricas obtenidas para la predicción de habitaciones a partir del tipo de propiedad y los metros cubiertos en todo el territorio, con segmentación.

En conclusión pudimos ver que es posible predecir el precio por metro basado en la ubicación de la propiedad, y si utilizamos segmentación logramos mejorar las predicciones. En base a esto se podría utilizar esta información para poder definir el precio de las propiedades sin tener que utilizar el tamaño de la propiedad como input al modelo.

3.6. Utilización de Feature Engineering

Para este experimento, lo que hicimos fue agregar 2 características mas al conjunto de datos a partir de información que ya teníamos, pero que estaba dispersada o no eran valores numéricos.

Lo primero que hicimos fue segmentar por cada ciudad del conjunto de datos y ejecutar en cada una, dos predicciones, una con y una sin feature engineering. Una vez observados los resultados, calculamos la media de la mejora porcentual obtenida, la cual fue del 11 %. En el mejor caso, obtuvimos una mejora del 33 % para la ciudad de Tultitlán, el gráfico unificado obtenido para las predicciones de esta ciudad, utilizando K-fold cross validation es el siguiente:

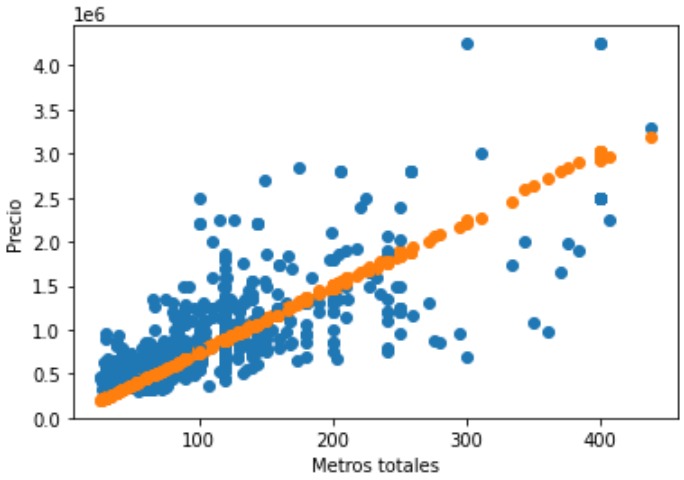


Figura 14: Metros Totales vs Precio sin Feature Engineering

Al igual que en los experimentos anteriores, obtuvimos predicciones que representan una función lineal, y los errores obtenidos fueron:

- RMSE: 789964
- RMSLE: 0.778
- R^2: 0.541

Luego, para intentar obtener predicciones mas acertadas, utilizamos la técnica explicada anteriormente de feature engineering, para esto agregamos 2 columnas al conjunto de datos. Una de ellas fue *buena_desc* que representa si en el titulo o en la descripción de la publicación se encuentran palabras claves que hablen bien de la casa, como *preciosa*, *cómoda*, *moderna*, *nueva*, etc. Las palabras que utilizamos para decidir si el inmueble tiene una buena descripción provino de observar que estas palabras se encontraban reiteradas veces en las descripciones de los inmuebles además de ser características que las distinguen sobre otros inmuebles. La otra columna agregada representaba si una vivienda estaba ubicada en una zona céntrica, esto era generado a partir de si habían escuelas y centros comerciales cercanos, ya que normalmente estos establecimientos suelen encontrarse en el centro de la ciudad, por ende, es un buen filtro para la decisión.

Los resultados obtenidos al aplicar estas dos columnas fueron los siguientes:

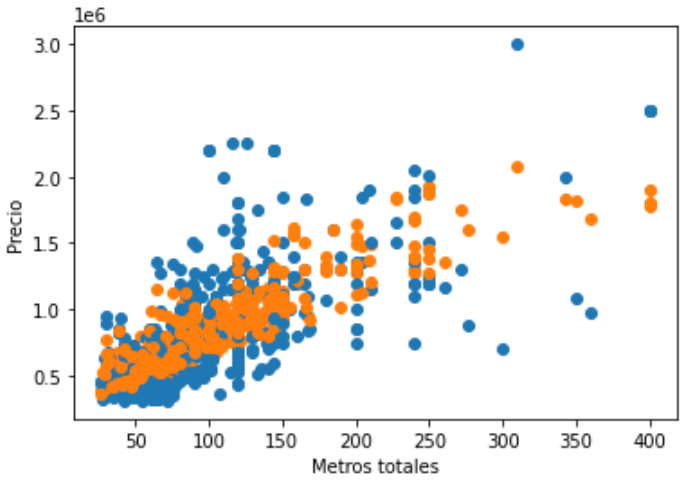


Figura 15: Metros Totales vs Precio con Feature Engineering

Podemos observar que ahora, nuestras predicciones ocupan un espacio mas amplio que la anterior que se correspondía con una linea prácticamente recta. Los errores obtenidos fueron los siguientes:

- RMSE: 528741
- RMSLE: 0.586
- R^2: 0.632

Como conclusión para este experimento, podemos decir que agregar nuevos datos que antes no eran tenidos en cuenta, o generar columnas nuevas a partir de datos que ya tenemos nos sirve, y es un gran método para poder predecir de forma mas exacta nuestro conjunto de datos. En nuestro caso, obtuvimos mejoras de hasta mas del 30 % en algunas ciudades y en promedio una mejora del 11 %. Segmentando por provincia, esta mejora fue del 8 %.

4. Conclusiones

A lo largo de los experimentos realizados pudimos analizar diferentes características (metros totales, baños, antigüedad, etc) entrenando modelos que se adecuen lo mejor posible. En general observamos que con la segmentación correcta obteníamos mejores resultados ya que segmentar por provincia tenia peores resultados que segmentando por tipo de propiedad. Estos comportamientos también ocurrieron a la hora de hacer feature engineering con los datos y expandir nuestro conjunto de datos para predecir los precios. Para determinar como mejoraban las experimentos utilizamos diferentes métricas de cálculo de errores como: RMSE, RMSLE y R^2.

Referencias

[1] `What's the Difference Between RMSE and RMSLE? - Analytics Vidhya`