DEPARTAMENTO
DE COMPUTACION
Facultad de Ciencias Exactas y Naturales - UBA

Metros Cuadrados Mínimos Lineales

Trabajo Práctico 3

19 de julio, 2020 — Métodos Numéricos

Grupo 4

Integrante	LU	Correo electrónico
López Menardi, Justo	374/17	juslopezm@gmail.com
Strobl, Matías	645/18	matias.strobl@gmail.com
Wehner, Tomás	67/17	tomi.wehner10@gmail.com
Yulita, Federico	351/17	fyulita@dc.uba.ar

Instancia	Docente	Nota
Primera entrega		
Segunda entrega		

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

Resumen

Se desarrolló un algoritmo de predicción para distintas caracteristicas de un inmueble utilizando una base de datos de 240000 muestras. Principalmente se utilizó el método de cuadrados mínimos para generar el estimador deseado. Se trabajó principalmente con la predicción de los precios de un inmueble en función de su ubicación, los metros cubiertos y el número de habitaciones. En segundo lugar también se experimentó con la predicción de metros cubiertos en función de los metros totales y la cantidad de ambientes de un inmueble. Se discutió cómo la correcta segmentación y tratamiento de las variables y los datos es necesaria para la correcta predicción del precio de los inmuebles y se presentaron ejemplos ilustrativos de estos métodos.

Palabras Clave: Regresión Lineal, Cuadrados Mínimos, Ecuaciones Normales

1. Introducción

En el presente trabajo se desarrollará y evaluará una herramienta de predicción de características de inmuebles. Para ello, se implementará un algoritmo de clasificación supervisado que será entrenado con una base de avisos de ventas de inmuebles con precios conocidos, que luego servirá para aproximar el precio de avisos de inmuebles no presentes en la base de datos de entrenamiento.

La motivación principal de este trabajo gira en torno al análisis y la experimentación de resolver problemas de regresión. Las regresiones pueden ser de suma importancia en casos donde se busque encontrar la manera de explicar datos que en principio se desconozca la familia de funciones que los responden.

Se cuenta con un set de datos de avisos inmobiliarios de México con las siguientes caracteristicas:

- Id: identificador del aviso
- Titulo del aviso inmobiliario
- Descripcion del aviso
- El tipo de propiedad (casa,apartamento,etc.)
- Ubicacion
- Caracteristicas cuantificables como cantidad de baños, piscina, escuelas cercanas, etc.

Se utilizará la técnica de Cuadrados Mínimos Lineales para aproximar una determinada característica en función de otras conocidas usando al precio como la variable que se quiere aproximar.[1] Dada una familida de funciones $\{\phi_1, \phi_2, \ldots, \phi_n\}$ y un conjunto de vectores $\{x_1, x_2, \ldots, x_m\} \subseteq \mathbb{R}^k$ y escalares $\{y_1, y_2, \ldots, y_m\} \in \mathbb{R}$ entonces tomamos $M \in \mathbb{R}^{m \times n}$ como la matriz con elementos $m_{ij} = \phi_j(x_i)$ e $y \in \mathbb{R}^m$ como el vector compuesto por los escalares y_i. A cada componente de un vector x lo llamamos *feature* y cada vector representa una medición de estos features. Es por eso que a cada medición x_i le corresponde un resultado y_i. Entonces, cuadrados mínimos es el problema que consiste en hallar el vector α que minimice $||M\alpha - y||_2$. Lo que esto logra es hallar la mejor combinación lineal de funciones ϕ_i que aproximan los valores de y con los vectores x asociados. En particular, para regresión lineal consideramos $\phi_1 = 1$ y $\phi_i = e_{i-1}^\dagger \ \forall i \leqslant k+1$, donde e_i es el i-ésimo vector canónico. En otras palabras, para regresión lineal consideramos una combinación lineal de los features de las mediciones más un término constante.

Una manera de resolver el problema de cuadrados mínimos es usando ecuaciones normales. Resulta que $\alpha = \text{argmin}_\beta \left(||M\beta - y||_2\right) \iff M^\dagger M\alpha = M^\dagger y$. Ya que en este caso la cantidad de mediciones es mucho mayor que la cantidad de features consideramos a $M^\dagger \cdot M$ una matriz inversible, ya que es linealmente independiente por columnas. Por lo tanto, vamos a usar que:

$$\alpha = \left(M^\dagger M\right)^{-1} My$$

Esta igualdad nos provee una manera fácil y relativamente efectiva de resolver el problema de cuadrados mínimos.

Una vez hecha la aproximación de la variable a estimar, se pondrá a prueba el algoritmo con un conjunto de validación con muestras que no hayan sido usadas durante el entrenamiento. Se trabajará de esta forma con cuatro distintas métricas: RMSE, RMSLE, R^2 y MAE. [2]

1.1. RMSE

La métrica RMSE (*Root Mean Squared Error*) es la raíz cuadrada del promedio de errores cuadrados. Dado un modelo $\hat{f}$ y una observación (x_i, y_i), se define $\hat{y}_i = \hat{f}(x_i)$. Entonces,

$$\text{RMSE} = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(y_i - \hat{y}_i)^2} \tag{1}$$

El problema que esta métrica tiene es que las muestras con valores altos pesarán más que aquellas con valores bajos. Por ello, en algunas ocasiones tiene más sentido considerar a la métrica RMSLE.

1.2. RMSLE

La métrica RMSLE (*Root Mean Squared Log Error*) es definida como

$$\text{RMSLE} = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(\ln(y_i + 1) - \ln(\hat{y}_i + 1))^2}. \tag{2}$$

Esta última métrica tiene la propiedad de pesar de la misma manera la mejora porcentual sobre cualquiera de las muestras sin importar su valor absoluto.

1.3. R^2

El metodo R^2 (coeficiente de determinación) mide la proporción de la varianza total de la variable explicada por la regresion. Sea $\bar{y}$ el promedio de los valores de y_i entonces definimos el coeficiente de determinación como

$$R^2 = 1 - \frac{\sum_{i=1}^{N}(y_i - \hat{y}_i)^2}{\sum_{i=1}^{N}(y_i - \bar{y})^2} \tag{3}$$

1.4. MAE

La métrica MAE (*Median Absolute Error*) es definida como

$$\text{MAE} = \text{MED}(|y_i - \hat{y}_i|), \tag{4}$$

donde MED es la mediana de un conjunto de datos. Esta métrica tiene la propiedad de no ser muy afectada por outliers en el conjunto de datos, ya que encuentra el valor que es "más frecuente".

2. Desarrollo

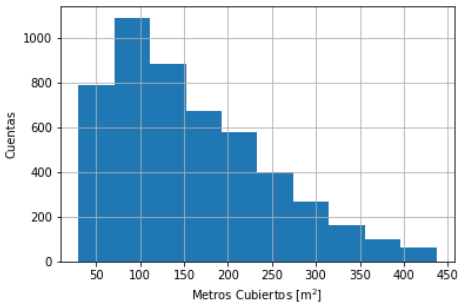
Los algoritmos utilizados se implementaron en C++ usando la biblioteca Eigen para facilitar las operaciones con matrices y vectores. Para estudiar el conjunto de datos se utilizaron *notebooks* de Jupyter Python en el que se corría el código de C++ y se analizaban los datos con las bibliotecas NumPy, Matplotlib, Pandas y SciKit Learn. Para resolver el problema de regresión lineal se implimentó el resultado hallado con ecuaciones normales ya descrito en la introducción.

Uno de los problemas presentes es que el método de Cuadrados Mínimos Lineales es un algoritmo que intenta explicar todos los datos con una sola función y el conjunto de datos a utilizar es bastante extenso y heterogéneo, por lo que será muy difícil conseguir una buena aproximación mediante cuadrados mínimos de todos los datos.

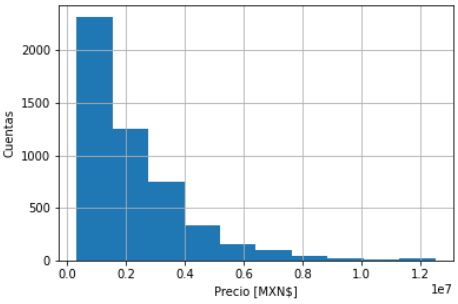
Para resolver este problema, se recurre a la segmentacion de datos. Con esto se logra enfocar el problema en un espacio más controlable, es decir, en el que menos variables estén en juego. Por ende, redujimos la complejidad, y al aplicar CML podemos predecir con mejor precisión.

Otro de los procesos con los que se experimentará es el de *feature engineering*, que consiste en producir características nuevas a raíz de los datos existentes. Esto se aplica para poder profundizar el estudio y a su vez intentar encontrar patrones o correlaciones que con los meros datos crudos uno no descubriria.

Para los ajustes se usaron dos conjuntos de datos: uno de entrenamiento y otro de test. Estos conjuntos se tomaron como subconjuntos del conjunto de entrenamiento provisto con 240000 muestras. No se usó el conjunto de test ya que no tiene los precios de las viviendas. Para el ajuste se usó el conjunto de



(a) Metros cubiertos de la muestra de viviendas tomada.



(b) Precios de la muestra de viviendas tomada.

Figura 1: Histogramas de los metros cubiertos y del precio de las viviendas de la muestra tomada.

entrenamiento y para evaluar el ajuste se usó el conjunto de test. Para lidiar con outliers en el conjunto de datos se hizo lo siguiente. Primero, se hizo un ajuste en el conjunto de entrenamiento. Luego, usando ese ajuste se hallaron las distancias de cada punto al ajuste. Luego, se descartaron aquellos datos cuya distancia al ajuste era dos desviaciones estándar mayor que la distancia promedio. Esto se hizo en el conjunto de datos de entrenamiento y de test. Finalmente, se hizo un nuevo ajuste en el conjunto de entrenamiento y se lo evaluó con el conjunto de test. Para medir la calidad de los ajustes en el conjunto de entrenamiento de datos usamos K-Fold Cross Validation con las métricas descritas, ya que nos permite obtener un resultado más robusto y menos sesgado. [3] [4]

3. Resultados y Discusión

3.1. Caso 1 - Metros Cubiertos vs. Precio

En este caso de experimentación se obtuvo una relación lineal entre los metros cubiertos y el precio de las viviendas. Debido a que las viviendas más grandes suelen ser más caras se esperaba hallar una relación positiva entre ambas variables.

Se tomó una muestra al azar de 100 viviendas del conjunto de entrenamiento de datos para el conjunto de test y 5000 para el conjunto de entrenamiento. En la **Figura 1** se encuentran histogramas de los metros cubiertos y los precios de las viviendas del conjunto de entrenamiento. Notemos que la gran mayoría de las viviendas son de hasta 250 m^2 y salen menos de 3 millones de pesos. Esperamos entonces que la densidad de puntos para bajos precios y para viviendas chicas sea mucho mayor que para altos precios o viviendas grandes. Luego, con el algoritmo implementado se obtuvo un ajuste lineal de los datos y se lo usó para formar el gráfico de la **Figura 2**. En esta figura se muestra el ajuste del conjunto de entrenamiento con los datos del conjunto de test. Notemos que el ajuste es adecuado para la muestra y se obtuvo una relación positiva entre los datos como se esperaba. Además, la densidad de puntos para los distintos rangos de valores corresponde con lo que vimos en los histogramas. Sin embargo, debido a que los datos no están determinados por este único factor la cantidad de outliers es significante, especialmente para viviendas más grandes.

Se usó K-Fold Cross Validation en el conjunto de datos de entrenamiento para verificar la precisión del ajuste obtenido con un método robusto. Se usaron las fórmulas (1) para calcular el RMSE, (2) para calcular el RMSLE, (3) para calcular el R^2 y (4) para calcular el MAE. Se obtuvieron los valores:

- RMSE $= (1{,}45 \pm 0{,}05) \times 10^6$ MXN$
- RMSLE $= (0{,}594 \pm 0{,}009)$
- $R^2 = (0{,}37 \pm 0{,}03)$
- MAE $= (6{,}4 \pm 0{,}3) \times 10^5$ MXN$

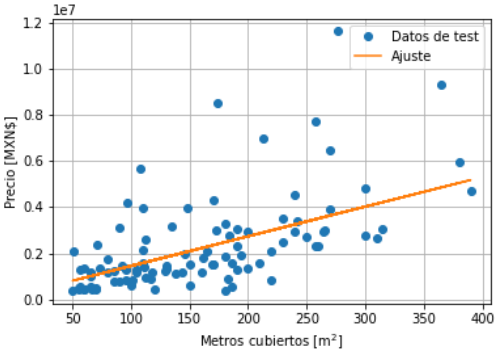


Figura 2: Precio de las viviendas en función de su tamaño en metros cubiertos.

con $K = 10$. Notemos que el RMSE es mucho mayor que el RMSLE. Esto es razonable, ya que el RMSE es del orden del precio de las viviendas (del millón de pesos) debido a los outliers. Sin embargo, el RMSLE no se ve tan afectado por outliers ya que es esencialmente el logaritmo del error relativo del ajuste; no del error absoluto. Además, notemos que el R^2 indica una relación con los datos bastante baja, lo cual indica que la relación no es estrictamente lineal. Esto puede deberse a que hay muchas otras variables que afectan al precio.

Veamos ahora los errores del ajuste en el conjunto de test:

- RMSE $= 1{,}62 \times 10^6$ MXN$
- RMSLE $= 0{,}615$
- $R^2 = 0{,}35$
- MAE $= 6{,}3 \times 10^5$ MXN$

Notemos que el RMSE y RMSLE no están contenidos en los errores provenientes de K-Fold pero el R^2 y MAE si. Esto nos demuestra que las primeras dos métricas son las que más se ven afectadas por los outliers y que estas últimas dos son las más estables ante este cambio del conjunto de datos.

Se usó el ajuste obtenido para eliminar outliers de ambos conjuntos de datos y se ajustó de vuelta. En la **Figura 3** puede verse un gráfico del conjunto de test sin outliers con el nuevo ajuste hecho en el conjunto de entrenamiento sin outliers. Si lo comparamos con el gráfico anterior es evidente que hay menos puntos aislados de alto precio. Esperamos entonces que los errores de este ajuste sean menores que en el anterior.

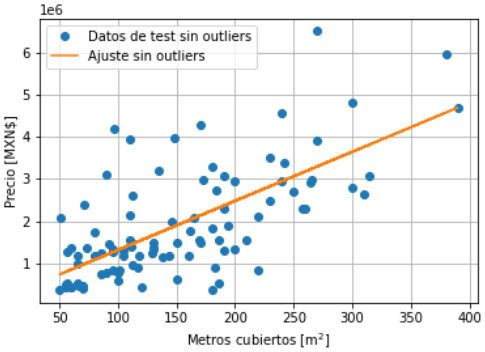


Figura 3: Precio de las viviendas en función de su tamaño en metros cubiertos (sin outliers).

Se usó K-Fold Cross Validation en el nuevo conjunto de entrenamiento y se obtuvo lo siguiente:

- RMSE $= (9{,}9 \pm 0{,}3) \times 10^5$ MXN$
- RMSLE $= (0{,}524 \pm 0{,}008)$
- $R^2 = (0{,}50 \pm 0{,}04)$
- MAE $= (5{,}3 \pm 0{,}3) \times 10^5$ MXN$

Notemos que estos errores son mucho menores que los anteriores. Los errores en el conjunto de test son:

- RMSE $= 9{,}6 \times 10^5$ MXN$
- RMSLE $= 0{,}548$
- $R^2 = 0{,}44$
- MAE $= 5{,}1 \times 10^5$ MXN$

Esta vez los errores del conjunto de test se acercan más que a los del conjunto de entrenamiento, por más de que algunos sigan sin estar contenidos entre el intervalo dado por los errores. Es por eso que usamos esta técnica en el resto de los casos para eliminar outliers.

3.2. Caso 2 - Latitud y Longitud vs. Precio

En este caso de experimentación se obtuvo una relación lineal entre la longitud y latitud y el precio de las viviendas. Esto nos permitiría hallar una dirección cardinal en la que el precio de las viviendas de México incrementa.

Se tomó una muestra al azar de 500 viviendas del conjunto de entrenamiento de datos. Luego, con el algoritmo implementado se ajustaron los datos de la muestra y se obtuvo una dirección en la que los precios incrementan. Esta información se refleja en la **Figura 4**. Notemos que los puntos de la figura superpuestos en la imagen del mapa de México reflejan las viviendas de la muestra siendo el color el precio de cada vivienda, mientras que el gradiente de colores refleja el ajuste obtenido. Notemos que el ajuste

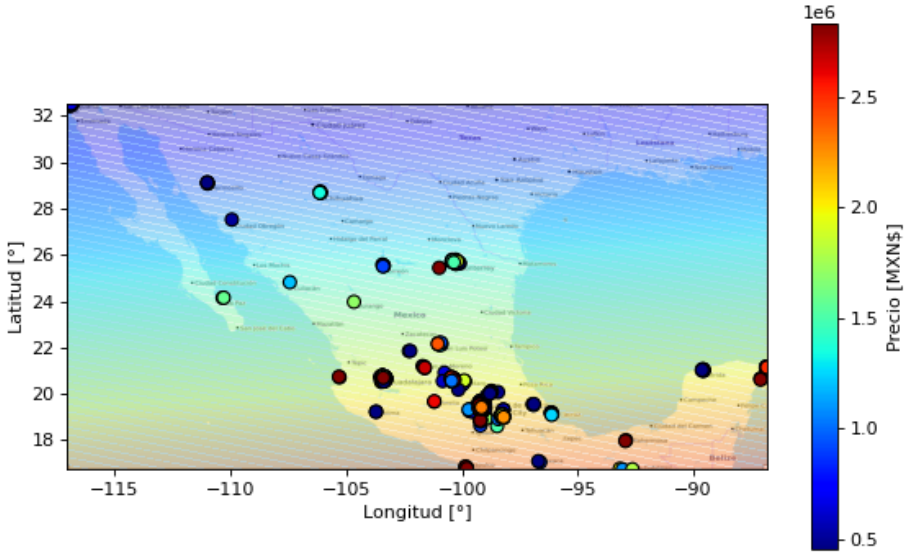


Figura 4: Precio de viviendas en México por ubicación.

muestra que las viviendas más caras están en la costa sur del atlántico. Esto es razonable, ya que esta zona es la del caribe de México con gran atractivo turista.

Al igual que en el caso anterior, se usó K-Fold Cross Validation para calcular las métricas. Se obtuvo:

- RMSE $= (1{,}30 \pm 0{,}04) \times 10^6$ MXN$
- RMSLE $= (0{,}75 \pm 0{,}01)$
- $R^2 = (0{,}057 \pm 0{,}009)$
- MAE $= (9{,}2 \pm 0{,}2) \times 10^5$ MXN$

con $K = 10$. Notemos que para este caso los errores son muy grandes y que R^2 es muy bajo. Esto nos indica que el modelo de ajuste utilizado es muy malo para el conjunto de datos que tenemos. Esto puede deberse al hecho de que solo tomamos una dirección cardinal como variable de ajuste y que utilizamos datos de viviendas de todo el país, sin importar región, tamaño ni ninguna otra variable relevante para el precio de casas. Sin embargo, notemos que el ajuste logra apuntarnos en la dirección donde las viviendas suelen ser más caras en México, que es la costa sur del atlántico.

Para el conjunto de test obtuvimos:

- RMSE $= 1{,}39 \times 10^6$ MXN$
- RMSLE $= 0{,}76$
- $R^2 = 0{,}053$
- MAE $= 9{,}1 \times 10^5$ MXN$

Estos resultados se corresponden a los del conjunto de entrenamiento de la misma forma que en el caso anterior.

3.3. Caso 3 - Precios de Viviendas en Cancún

En este caso de experimentación se obtuvo una relación cuadrática entre la longitud y latitud y el precio de la viviendas pero solo para la zona céntrica y costera de la ciudad de Cancún. Es decir, ajustamos el precio de esta zona a las variables $long$, lat, $long^2$, lat^2 y $long \cdot lat$; donde $long$ es la longitud y lat la latitud. Esto nos permitió no solo encontrar una dirección cardinal en donde los precios son más altos sino que también nos perimitió dividir esta zona con mayor precisión ya que al tener estas variables adicionales

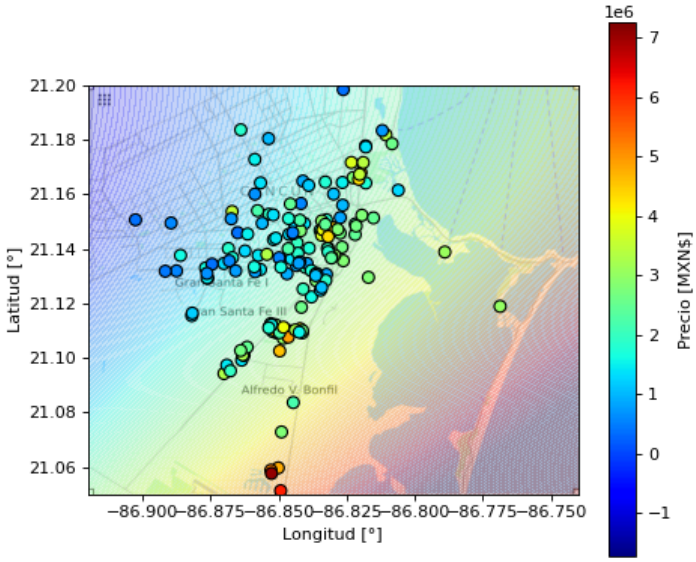


Figura 5: Precio de viviendas en Cancún por ubicación.

podemos dividir zonas con curvas cuadráticas. Elegimos esta ciudad ya que nos pareció interesante ver cómo cambian los precios a medida que las viviendas se alejan de la costa. Esperamos hallar que las viviendas más cercanas a la costa sean las más costosas de la ciudad.

Se tomó una muestra al azar de 178 viviendas para el conjunto de test y 1800 para el conjunto de train y restringimos las viviendas a los rangos $long \in [-86{,}92, -86{,}74]$ y $lat \in [21{,}05, 21{,}20]$. Esto lo hicimos para tener viviendas de la zona central de Cancún. Luego, se eliminaron outliers y se ajustaron los datos de la muestra y se obtuvo el ajuste mostrado en la **Figura 5**. Notemos que el ajuste es adecuado a los datos de la muestra y que además la curvatura hallada corresponde con la curvatura natural de la costa de Cancún que separa adecuadamente por zonas los precios de las viviendas. Es decir, las viviendas azules se encuentran en el relleno azul del ajuste, la banda de viviendas verdes y amarillas se encuentran en la banda verde y amarilla de relleno y las viviendas más costosas se encuentran en la costa donde la banda de relleno es roja. Esto se corresponde con lo que esperábamos hallar. Esta división precisa entre las viviendas es algo que no se obtuvo en el caso anterior ya que solo habíamos obtenido una dirección cardinal donde los precios tienden a ser mayores.

Al igual que en los casos anteriores, se usó K-Fold Cross Validation para calcular las métricas. Se obtuvo:

- RMSE $= (9{,}7 \pm 0{,}4) \times 10^5$ MXN$
- RMSLE $= (0{,}55 \pm 0{,}03)$
- $R^2 = (0{,}41 \pm 0{,}04)$
- MAE $= (6{,}48 \pm 0{,}05) \times 10^5$ MXN$

Notemos que este ajuste es significativamente mejor que el anterior. Esto se debe a que le permitimos mayor libertad con las variables extras de orden cuadrático y a la segmentación de viviendas de todo México a viviendas de la zona central de Cancún. Además, al igual que en el ajuste anterior se obtuvo que la zona más cara es la más cercana a la costa, lo cual es razonable. Para el conjunto de test obtuvimos:

- RMSE $= 1{,}01 \times 10^6$ MXN$
- RMSLE $= 0{,}59$
- $R^2 = 0{,}36$
- MAE $= 6{,}47 \times 10^5$ MXN$

Estos errores son similares a los que obtuvimos para el conjunto de entrenamiento.

3.4. Caso 4 - Metros cubiertos

Se obtuvo estimador para los metros cubiertos de un inmueble. Para ello en primer lugar se realizó una regresión lineal en función de la cantidad de habitaciones. Se utlizó K-fold Croos validation con K=10 sobre el conjunto de entrenamiento para aproximar el error del ajuste de forma más robusta.Y utilizando las ecuaciones explicadas en la introducción se obtuvieron las siguientes métricas:

- RMSE = $(73{,}274 \pm 0{,}497)$
- RMLSE = $(0{,}460 \pm 0{,}003)$
- R2 = $(0{,}274 \pm 0{,}013)$
- MAE = $(47{,}107 \pm 0{,}952)$

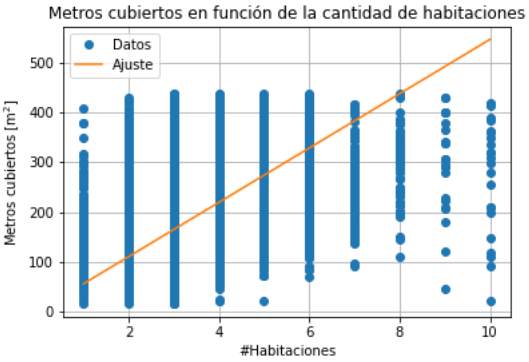


Figura 6: Metros Cubiertos en func de # habitaciones

Observando la Figura 6 vemos que si bien existe una varianza entre inmuebes con la misma cantidad de ambientes se puede observar una proporcionalidad directa entre ambas variables, a más ambientes más metros cubiertos. Esto se puede formalizar al observar la covarianza entre ambas variables.En particular se obtuvo una covarianza entre ambas variabes de 37.07 indicando una correlación directmente positiva. Es decir a más metros cubiertos, mayor cantidad de habitaciones.

En segundo lugar se realizó feature engineering creando una nueva caracteristica llamada #ambientes la cual es el resultado de la suma de la cantidad de habitaciones y baños.Se decidieron utilizar estos features ya que ambas se relacionan de manera directa con los metros cubieros de un ambiente ya que espacio que ocupe cada uno de estos ambientes será considerado dentro de metros cubiertos. Si observamos la covarianza el número de ambientes y el numero de ambientes se obtiene un valor de 86,67 lo cual indica una relación positiva entre ambas variables.A más cantidad de ambientes más metros cubiertos.

Se utilizo K-Fold cross validation con K=10 sobre el conjunto de entrenamiento para aproximar el error generado por este ajuste de forma más robusta el cual se puede observar en la figura 7. Se utliizó las ecuaciones (1) , (2) ,(3) y (4) para obtener asi los valores de RMSE y RMSLE ,RMLSE R2 y MAE respectivamente. Obteniendo como resutados:

- RMSE = $(61{,}771 \pm 0{,}668)$
- RMLSE = $(0{,}37564 \pm 0{,}00358)$
- R2 = $(0{,}48410 \pm 0{,}01276)$
- MAE = $(34{,}756 \pm 0{,}425)$

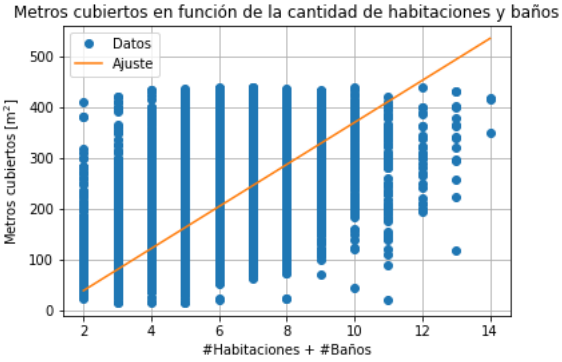


Figura 7: Metros Cubiertos con feature engineering

Continuando con la experimentación, se decidió observar otra carecterista extra. Se observó la cantidad de metros cubiertos en función de los Metros totales y ambientes.Se realizo un regresión lineal en función a estas variables. Si bien se utilizó a todo el conjunto de train para la experimentación se utilizó un sample 100 inmuebles para obtener la **Figura** 8 para poder observar con una mayor faciidad como se realacionan las variables.

Del mismo modo que se realizó para la experimentación anterior se utilizo K-fold cross con K= 10 validation para obtener los valores de RMSE y RMSLE ,RMLSE R2 y MAE. Se utliizaron las ecuaciones (1) , (2) ,(3) y (4). Obteniendo como resutados:

- RMSE = $(48{,}834 \pm 0{,}446)$
- RMLSE = $(0{,}30324 \pm 0{,}00415)$

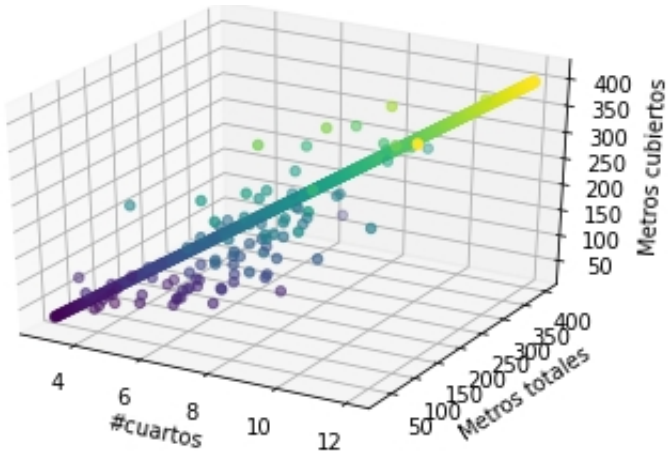


Figura 8: Metros Cubiertos en función de la cantidad de ambientes y metros totales

- R2 = $(0{,}67762 \pm 0{,}0043)$
- MAE = $(25{,}3082 \pm 0{,}3466)$

Finalmente comparando los errorres entre los diferentes estimadores observados para los metros cubiertos de un inmueble se pudo observar una mejora sinificativa al hacer feature engineering con la cantiad de baños y cantidad de habitaciones.

En particular si observamos el RMSE hubo un reducción de un 16 % del valor inicial medido para el primer estimador. Y luego al agregar como variable extra la cantidad de metros totales se observó una mejora significativa en comparación con el model anterior. Si se observa el RMSE hubo una reducción del 21 % en comparació del segundo estimador utilizado(numero de ambientes). Y en comparación con el primer estimador se observó una reducción del 43,5 % del valor inicial de RMSE.

3.5. Caso 5: Ajuste con Muchas Variables

Finalmente, veamos el ajuste de todas estas variables que estuvimos estudiando en la zona céntrica de Cancún que ya vimos. Vamos a considerar la latitud y longitud hasta orden cuadrático, los metros cubiertos y la cantidad de habitaciones. Esperamos que este ajuste al considerar tantas variables sea mucho mejor que el resto, y al estar bien segmentado en cuanto a región y outliers los errores deberían ser relativamente bajos.

Se tomaron 1300 datos para el conjunto de entrenamiento y 116 para el conjunto de test, segmentadas de la misma forma que en el caso 3. En la **Figura 9** puede verse un gráfico de las viviendas del conjunto de test en el mapa de Cancún.

En la **Figura 10a** puede verse un gráfico del ajuste obtenido y de los datos del conjunto de test para el precio en función de la cantidad de habitaciones de las viviendas y en la **Figura 10a** puede verse un gráfico del ajuste obtenido y de los datos del conjunto de test para el precio en función de la cantidad de metros cubiertos de las viviendas. Notemos que el ajuste es adecuado para ambos gráficos, solo fallando en predecir los precios de las casas más caras, aun habiéndose filtrado los outliers del conjunto de datos.

Al igual que en los casos anteriores, se usó K-Fold Cross Validation para calcular las métricas. Se obtuvo:

- RMSE = $(5{,}5 \pm 0{,}3) \times 10^5$ MXN$
- RMSLE = $(0{,}41 \pm 0{,}06)$
- $R^2 = (0{,}79 \pm 0{,}03)$
- MAE = $(3{,}5 \pm 0{,}3) \times 10^5$ MXN$

Notemos que estos errores son mucho más bajos que los que vimos para el resto de los casos. Esto se corresponde con nuestra hipótesis, al tener más variables en consideración y al haber segmentado correctamente los resultados son mucho mejores. Para el conjunto de test se obtuvo:

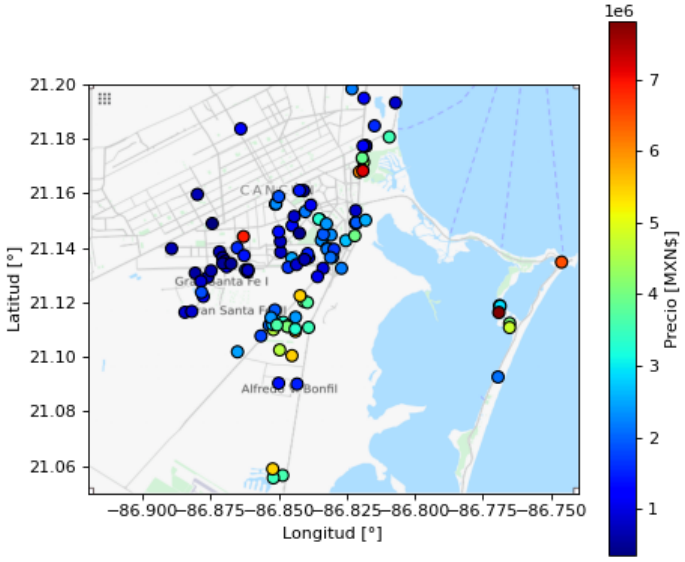
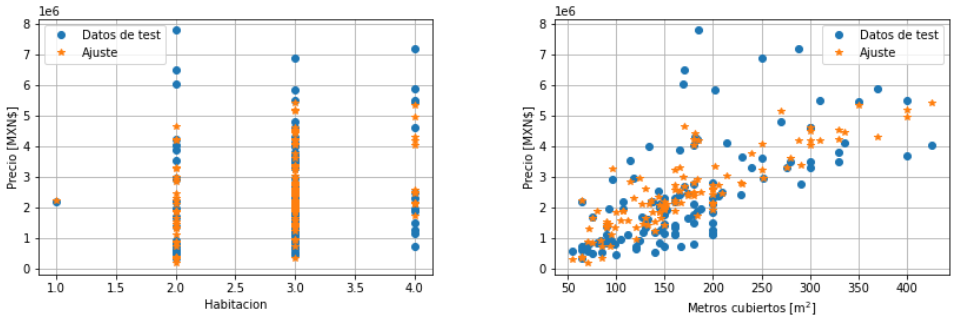


Figura 9: Precio de viviendas en Cancún por ubicación.



(a) Precio de viviendas en Cancún en función del número de habitaciones.

(b) Precio de viviendas en Cancún en función de la cantidad de metros cubiertos.

Figura 10: Precio de viviendas en Cancún en función del número de habitaciones y de la cantidad de metros cubiertos

- RMSE $= 6{,}0 \times 10^5$ MXN$
- RMSLE $= 0{,}40$
- $R^2 = 0{,}78$
- MAE $= 4{,}0 \times 10^5$ MXN$

Estos errores se corresponden a los obtenidos con el conjunto de entrenamiento.

4. Conclusiones

Se estudiaron cinco distintos casos en los que las variables independientes fueron distintas y se hicieron regresiones lineales con el objetivo de ajustar una variable dependiente, en general el precio de las viviendas. En el primer caso se estudiaron los metros cubiertos de las viviendas, en el segundo la ubicación, en el tercero se segmentó a la ciudad de Cancún, en el cuarto se estudiaron los metros cubiertos en función de la cantidad de habitaciones y en el quinto se usaron los metros cubiertos, ubicación y cantidad de habitaciones para predecir el precio.

En todos los casos se observó que mientras mejor sea la segmentación de los datos y la cantidad y significancia de las variables independientes mejor era el ajuste, presentando errores menores. Estos errores se midieron usando el RMSE, RMSLE, R^2 y MAE. Además, se observó una mejora significativa en la predicción del precio del inmueble implementando la segmentación utilizada y outliers para el modelo observado de Cancún, y utilizar de forma conjunta en la predicción varias variables diferentes observadas anteriormente (habitaciones, metros cubiertos y ubicación).

Referencias

[1] A. S. Goldberger *et al.*, *Econometric theory*. New York: John Wiley & Sons., 1964.

[2] D. Ruppert, *Statistics and data analysis for financial engineering*, vol. 13. Springer, 2011.

[3] D. M. Allen, "The relationship between variable selection and data agumentation and a method for prediction," *Technometrics*, vol. 16, pp. 125–127, 1974.

[4] T. Hastie, R. Tibshirani, and J. Friedman, *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*. Springer, 2009.