DEPARTAMENTO
DE COMPUTACION

Facultad de Ciencias Exactas y Naturales - UBA

TP 3: Cuadrados Mínimos Lineales

7 de abril de 2022 — Métodos numéricos

Integrante	LU	Correo electrónico
Chami, Uriel Alberto	157/17	uriel.chami@gmail.com
Oliveira Gariboglio, Matias Nahuel	392/19	matiasnoliveira@gmail.com
Ciruzzi, Ramiro Augusto	228/17	ramiro.ciruzzi@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

1. Introducción

En el presente trabajo práctico estudiaremos el comportamiento de un dataset muy amplio de inmuebles en México. El mismo cuenta con múltiples propiedades como los metros cubiertos, el precio, si tiene piscina, la cantidad de baños, etc. Será de nuestro interés encontrar una función que estime el precio de un inmueble dados sus otros datos (y luego quizás estimar otras propiedades en función de las demás). Para lograr esto utilizaremos el método de cuadrados Mínimos Lineales (a partir de aquí CML).
El método de Cuadrados Mínimos Lineales es uno sumamente conocido y discutido en múltiples bibliografías. Se trata de una solución al sistema

$$Ax = b^1$$

con b^1 siendo la componente de $b \in Im(A)$. Utilizaremos para obtener resultados las famosas *ecuaciones normales* dado que se puede demostrar que x es solución de $Ax = b^1 \Leftrightarrow x$ es solución de $A^T Ax = A^T b$.

En nuestro caso A será la matriz de propiedades de los inmuebles con ciertas $\phi_i : \mathbf{R} \Rightarrow \mathbf{R}$ funciones continuas aplicadas en las columnas. Dichas ϕ_i serán gran parte de nuestro estudio.

La elección de estas funciones depende completamente del comportamiento de la variable dado que CML solo puede relacionar linealmente una variable respecto de la otra. Es por eso que a la hora de elegir la ϕ estamos interesados en el comportamiento de la variable y no en su comparativa respecto de las otras.

1.1. Data set y Segmentación

Como comentamos, en este trabajo estudiaremos un dataset de inmobiliarias mexicanas, el mismo posee 240 mil entradas. Las entradas en cuestión se refieren a muchas provincias y ciudades distintas, realidades completamente diferentes donde las variables pueden comportarse distinto. Por ejemplo: en una zona turística donde el ocio y lo recreativo importa mucho, la cantidad de escuelas cercanas puede ser un factor irrelevante a la hora de estimar el precio de la propiedad, dado que nadie vive en esa propiedad durante el año y su valor reside en la cantidad de piletas o si tiene mesa de ping pong o si está cerca del mar.

Habiendo entendido esto se devela por completo lo que haremos a lo largo de esta presentación. Buscamos ϕ's tales que estimen correctamente alguna variable mediante el estimador CML para algún segmento del dataset.
Eso requiere por un lado elegir inteligentemente los segmentos (teniendo en cuenta cuándo ciertas ciudades son comparables con otras), esto requerirá investigación para comprender las distintas realidades que se viven en México. Como también requiere de un estudio de los datos en sí para comprender distintos patrones y poder definir estas ϕ's.

1.2. Medidas de error

Para ver que tan eficientemente fueron nuestros métodos de aproximación, es decir, cuanto difieren nuestras estimaciones de los valores reales utilizaremos dos métricas: el *Root Mean Squared Error* (RMSE) y el *Root Mean Squared Log Error* (RMSLE).

Sea nuestro conjunto de observación de N muestras de la pinta (x_i, y_i), es decir, los y_i son los valores reales de la variable que queremos predecir; y sean los $\hat{y}_i$ nuestras predicciones, ambos errores se definen de la siguiente manera:

- RMSE

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

  Ventajas: Es una métrica bastante utilizada que penaliza exigentemente grandes errores.
  Desventajas: Los cambios en muestras altas tienen mucho más peso en el resultado que aquellos en las bajas, esto puede no ser lo ideal en algunos casos en los que se lograron buenos cambios pero en muestras poco significativas.

- RMSLE

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

Ventajas: La mejoras porcentuales sobre cualquiera de las muestras pesan lo mismo.
Desventajas: Exige que los y_i o los $\hat{y}_i$ sean mayores a -1 al estar los mismos dentro de un logaritmo.

En secciones venideras veremos más en profundidad que podemos inferir de la información que aportan estas métricas.

2. Experimentación - Primeros pasos

En primera instancia para empezar a conocer el dataset comenzamos a estudiar variable por variable. Es decir, a crear estimadores con CML de matrices A que en realidad son vectores columna con una sola propiedad. El objetivo de este estudio preliminar de los datos **no es obtener buenos estimadores**. El objetivo es en cambio **obtener buenas funciones** ϕ. Trataremos de minimizar (aunque sabemos que es muy dificil) el ruido que generan los segmentos. Nuestro objetivo para este experimento es hallar una idea (si es posible) de qué ϕ's son razonables para las variables más comunes y cuánto es 'poco' error.

2.1. Metros cubiertos

La primera variable que surge al estimar precio es indudablemente metros cubiertos. Es la variable que todos imaginamos que debe predecir el precio.

Hipótesis Los metros cubiertos están fuertemente correlacionados con el precio y permiten por si solos estimar el precio de una propiedad. La ϕ que hará que esto funcione es la identidad. Es decir $\phi(x) = x$.

Experimentación Primero graficamos m^2 $\times$ precio (Utilizamos un gráfico estilo heatmap mezclado con scatterplot para poder identificar los cúmulos de puntos y que la data sea más fácil de observar). En la figura 1 se puede observar la correlación. Las ciudades elegidas son las 3 con más datos: Querétaro, Benito Juárez y Mérida. Es bastante innegable y además nos da la pauta de que $\phi(x) = x$ será una elección razonable ya que se observa (más allá de la pendiente) un crecimiento lineal del precio respecto de los metros cubiertos.

Experimentación - CML Entonces intentemos estimar el precio de un inmueble en estas tres ciudades únicamente utilizando los metros cubiertos. Utilizamos k-fold cross validation para asegurarnos que los resultados no sean engañosos y RMSLE como medida de error. Los resultados obtenidos son

- Queretaro : 0.818
- Benito juarez : 0.773
- Merida : 0.755

Por ahora estos errores no nos dicen nada. A medida que avancemos entenderemos si es o no es un error *bueno*.

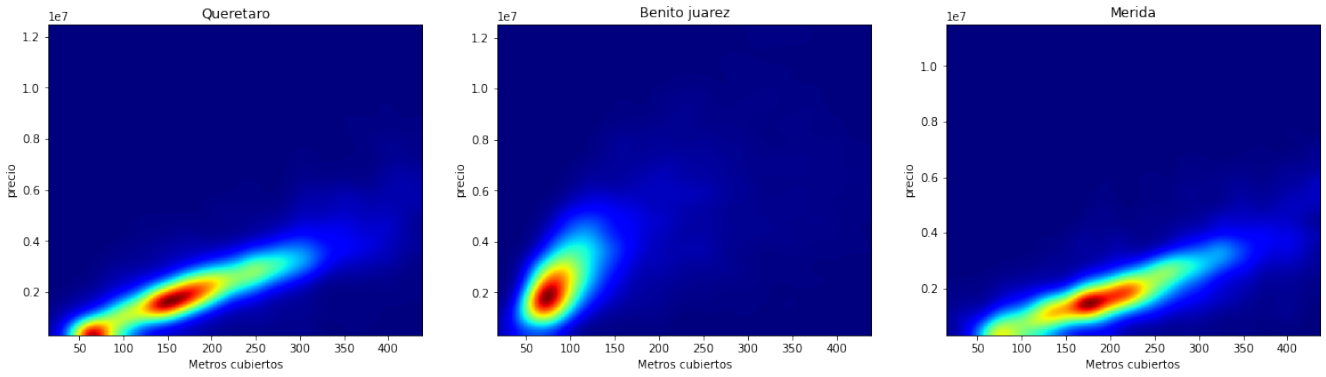


Figura 1: m^2 x precio

Para confirmar que efectivamente $\phi(x) = x$ es razonable, veamos como se comportaría si estimáramos $\phi(x) = x^2$.

- Queretaro : 1.295
- Benito juarez : 1.559
- Merida : 1.160

No sabemos aún si 0,8 es un buen error, pero sí sabemos que la identidad es la función apropiada para metros cubiertos. Los gráficos nos lo dicen, las comparaciones también.

2.2. Antigüedad

La antigüedad es una variable un poco más compleja. Sabemos que no toda ciudad es igual, y sabemos que esta no será la ϕ final o la única. Quizás hasta descubramos que no existe correlación y la variable no sirve para predecir precio. En la figura 2 se puede observar el gráfico de antigüedad por precio. Primero vale aclarar que al ser una variable discretizada (es decir, expresada en años siempre enteros) los valores se agrupan pues no pueden valer otra cosa. Sin emabrgo podemos observar cierta relación que combinada con nuestra intuición del significado de la variable nos indica que cuanto más antiguo sea un inmueble, menos debe costar. Es por esto que utilizaremos algo del orden de $\phi(x) = \frac{1}{x}$. Sin embargo hay muchas antigüedades de valor 0 y queremos que ϕ esté definida, entonces agregamos un parche para permitir esto quedando así $\phi_1(x) = \frac{1}{x+1}$. Definida la idea aún nos queda decidir por qué esa ϕ y no otra similar. Probamos con $\phi_2(x) = \frac{1}{\sqrt{x+1}}$, $\phi_3(x) = \frac{1}{(x+1)^2}$ y $\phi_4(x) = (x+1)$.

Hipótesis Viendo la figura 2 ya sabemos que no será un caso de éxito rotundo. Y es razonable que no lo sea ya que no mirar otras variables significa que **solo** sabiendo hace cuánto fue construído un inmueble podemos decir cuánto vale y eso es una locura, una casa de 15 habitaciones por más que tenga 20 años desde su construcción no vale menos que un departamento de un ambiente con 0 años de antiguedad. Será de interés encontrar cuál de las ϕ's se comporta mejor para proceder. En conclusión, creemos que obtendremos resultados mediocres que sin embargo pueden llegar a ser útiles cuando combinemos con más variables.

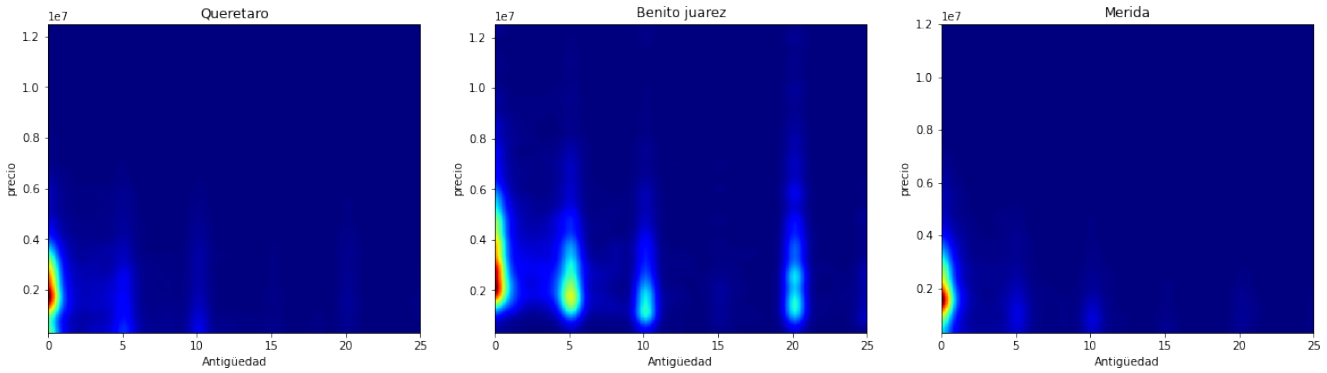


Figura 2: Antigüedad x precio

Resultados:

Error RMSLE según $\phi(x)$				
Ciudad	$\phi_1(x) = \frac{1}{x+1}$	$\phi_2(x) = \frac{1}{\sqrt{x+1}}$	$\phi_3(x) = \frac{1}{(x+1)^2}$	$\phi_4(x) = (x+1)$
Querétaro	1.385	0.856	2.769	1.766
Benito Juárez	1.853	0.899	4.116	1.772
Mérida	1.255	0.802	2.439	2.168

Nuestro análisis no es concluyente y entonces probamos con $\frac{1}{\sqrt[3]{x+1}}$:

Ciudad	$\phi(x) = \frac{1}{\sqrt[3]{x+1}}$
Querétaro	0.761
Benito Juárez	0.725
Mérida	0.714

Al ver una mejora empezamos a dudar de qué pasaría si seguimos subiendo el k de $\frac{1}{\sqrt[k]{x+1}}$

Error RMSLE misma ϕ distintos k			
Ciudad	$k = 10$	$k = 100$	$k = 100000$
Querétaro	0.711	0.708	0.708
Benito Juárez	0.637	0.631	0.630
Mérida	0.658	0.652	0.652

Es claro que $\frac{1}{\sqrt[n]{(x+1)}} \to 1$ ya que, como ningun inmueble pasa de los 99 años de antigüedad, x+1 no pasa nunca de 100. Recordemos que todos los experimentos fueron ejecutados sobre las mismas 3 ciudades, con 3-fold cross validation. Es entonces que este experimento nos da una idea no de que tan buena es la antiguedad estimando, si no de cuál sería nuestro *benchmark* para llamar a algo 'buena' estimación. Básicamente estamos diciendo que si el error es mayor a 0.7 (para RMSLE) entonces el estimador es malo ya que haciendo algo tan simple como estimar con una constante ($\phi(x) = 1$) se obtiene un error de esa magnitud.

También podemos concluir que la antiguedad pareciera comportarse como $\phi(x) = \frac{1}{x}$ pero comete errores tan groseros que no podemos concluir para cual k el estimador $\frac{1}{\sqrt[k]{x+1}}$ es mejor. Más adelante intentaremos encontrar dicho k utilizando un CML de más de una variable.

¿Pero qué significa estimar con una constante? Resulta que estimar con $\phi(x) = 1$ es de hecho estimar con el promedio. Verifiquemos esto, A será un vector columna con n filas (donde n es la cantidad de inmuebles de train) y recordemos que b tiene los precios de cada inmueble

$$A^t Ax = A^T b$$

$$\begin{pmatrix} 1 & \cdots & 1 \end{pmatrix} \begin{pmatrix} 1 \\ \vdots \\ 1 \end{pmatrix} x = \begin{pmatrix} 1 & \cdots & 1 \end{pmatrix} b$$

$$(\sum_{i=1}^{n} 1)x = (\sum_{i=1}^{n} b_i)$$

$$nx = (\sum precios)$$

$$x = \frac{(\sum precios)}{n}$$

Con esta demostración reforzamos la idea de que estos valores de error relativo son un *benchmark* de lo que definimos como 'bueno'. El precio promedio es conceptualmente el valor fijo que mejor aproxima los precios, nuestro objetivo es encontrar algo más dinámico que mejore dicho benchmark. De lo contrario, podríamos directamente estimar utilizando el promedio ya que es mucho más fácil de calcular.

2.3. Piscina

Otra variable que nos parece que puede llegar a generar efecto en la estimación es si la casa posee piscina. Creemos que este tipo de lujo debe venir de la mano con un precio más alto.

Hipótesis Tener piscina está asociado con una diferencia significativa de precio.

Resultados Para este caso lo que analizamos fue el precio promedio de los inmuebles con piscina y sin piscina. Como tener una piscina es un rasgo raro utilizamos provincias completas como segmentos.

En el Distrito Federal el precio promedio de los inmuebles sin piscina es de 3.38 M, mientras que para los inmuebles con piscina es de 5.33 M, una diferencia del 57 %.
Además este promedio es de verdad significativo ya que proviene de 2627 inmuebles con piscina en esta provincia.
En Jalisco sin embargo la diferencia no es tan significativa: 2.25 M sin piscina vs 3.28 M con piscina, un 45,8 %.
Esto puede explicarse entendiendo estas provincias, el Distrito Federal es la capital y de las provincias más prosperas de todo México, allí las piscinas probablemente serán de lujo y quizás hasta climatizadas. Además la temperatura más alta alcanzada ronda los 30ºC en los meses de marzo a mayo. Por otro lado Jalisco es una provincia más fría con temperaturas máximas de 23ºC y tampoco goza del mismo nivel de prosperidad que el DF.
Habiendo visto la significancia, veamos si podemos observar aún más diferencia en una ciudad playera y turística como lo es Cancún. Allí se observa un promedio de 3.2 M con piscina y 2.05 M sin piscina, una diferencia del 55 %, similar a la del DF.
Habiendo formado tal seguridad sobre la significancia de esta variable nos faltaría decidir cuán importante es. Es decir, entendemos que esta correlacionada de manera directa con el precio, pero en qué escala es esta correlación, ¿Es acaso $\phi(x) = x$?

Uno de los desafíos que surge es que si utilizáramos x como ϕ aquellos inmuebles sin piscina tendrían un precio estimado de 0, ya que no existe un coeficiente que pueda aportar CML que agrande 0. Por esto el parche que surge es utilizar $x + 1$, pero luego tener piscina vale 2 y no tenerla vale 1, esta diferencia nos suena poco singificativa por eso creemos que el mejor ϕ será $(x + 1)^2$. Ya que estamos diciendo que tener piscina vale 4 mientras que no tenerla vale 1.

Comparemos entonces la calidad de los resultados de aplicar CML (con 3-fold y RMSLE como error) para estas dos $\phi(x)$

Provincia	$\phi(x) = x + 1$	$\phi(x) = (x + 1)^2$
Distrito Federal	0.835	1.026
Jalisco	0.871	0.932
Quintana Roo - Cancún	0.835	1.026

Como podemos ver, x+1 parece hacer un mejor trabajo en todos los casos, logrando un error logarítmico promedio por debajo de aceptable utilizándolo por si mismo.

3. Feature engineering - Análisis de descripciones y lectura de 'humor'

Una de las propuestas ni bien fue planteado este trabajo era, además de intentar aproximar con las características ya presentes en el set de datos, producir nuevas características para también utilizarlas en los métodos de aproximación. Este método es lo que se conoce como *Feature engineering*.

Uno de los temas que más nos llamaba la atención era el análisis de las variables con campos de texto, ya que creíamos que extrayendo la información presente en los mismos se podían crear nuevas características útiles a la hora de estimar otras variables.

El conjunto de datos esta compuesto de únicamente dos características que consisten en su totalidad (excepto campos vacíos) de párrafos en texto plano: descripción y título. Si bien los títulos parecen englobar características de forma más concisa lo que queríamos era contar con una cantidad de palabras más robusta y como a simple vista vimos que la descripciones eran en general más largas que los títulos decidimos extraer la información para nuestro nuevo $feature$ de las mismas.

Como en un principio nuestro foco estaba en estimar el precio nos propusimos encontrar características que estuvieran relacionadas con un aumento o decremento del mismo. En afán de lograr esto nos planteamos crear una lista de palabras que para nosotros estaban estrictamente ligadas a un cambio en el precio del inmueble, lista que en un futuro usaremos para crear la nueva característica.Pero no queríamos que estas palabras fueran producto únicamente de nuestras imaginaciones o solo viendo algunas descripciones, a fin de evitar esto se nos ocurrió crear un ranking de las palabras más usadas en el total de las descripciones de los inmuebles.

Visto y considerando que el tamaño tanto del conjunto de datos del que disponemos como el de las descripciones de cada inmueble era bastante extenso nos vimos obligados a segmentar los datos para poder hacer dicho ranking. Es por esto que las palabras que usamos para llenar la lista fueron sacadas solamente del ranking de la ciudades Cancún y Puebla. Lo que hicimos entonces fue revisar aquellas palabras que se encontraban en el top 200 del ranking obtenido en ambas ciudades y meter a la lista aquellas que según nuestro criterio tenían una fuerte incidencia en el aumento del precio de los inmuebles, como por ejemplo 'lujo', 'piscina' o 'alberca'. A esta lista la llamamos **palabrasCaras**.

Con esta lista en nuestro haber creamos un nuevo *feature* que llamamos *valor* cuya idea consiste en asignar un puntaje a cada inmueble, siendo puntaje el resultado de sumar la cantidad de apariciones de cada palabra de **palabrasCaras** en las descripciones de cada inmueble.

Una vez obtenida esta nueva característica nos dispusimos a intentar estimar el precio de los inmuebles por su *valor*.

Hipótesis El precio de los inmuebles va a estar medianamente bien correlacionado con el *valor* de cada inmueble pues confiamos en que que una descripción tenga una gran cantidad de **palabrasCaras** esta ligado a un aumento de precio. Sin embargo, sabemos que el estimador puede fallar ya que la lista de palabras que creamos fue producto de una segmentación de los datos y los vocablos relacionados con mayor precio en otras ciudades o provincias pueden ser distintos. Además de que podría ser que varias de las **palabrasCaras** que seleccionamos no se condigan con un aumento de precio ya que el método de selección de las mismas fue bastante subjetivo.

Resultados Al requerir, el método para obtener el *valor* de cada inmueble, una cantidad de operaciones bastante grande [1] tuvimos que segmentar el set de datos por ciudad. Es por esto que solo trabajamos con las ciudades Zapopan, San Luis Potosí y Monterrey; que a pesar de ser solo un fragmento de nuestros datos las elegimos de forma tal que representen una buena parte de los mismos.

Lo primero que hicimos fue comparar el precio promedio de aquellos inmuebles con un *valor* menor a 4 contra su opuesto, elegimos el 4 como separador pues es el promedio redondeado de las medianas de la variable *valor* para las 3 ciudades. Obtuvimos estos resultados:

Ciudad	$valor < 4$	$valor \geq 4$
Zapopan	2.11 M	3.11 M
San Luis Potosí	1.77 M	2.29 M
Monterrey	2.37 M	3.44 M

Con una diferencia del 47 % en Zapopan y una del 45 % en Monterrey podríamos decir que se ve una correlación entre *valor* y *precio*. Por otro lado en San Luis Potosí al ser la diferencia de precios promedios de 29 % no parecería haber una correlación muy buena entre dichas variables. Para entender mejor a que escala correlaciona la variable *valor* con la variable *precio*, decidimos mirar los gráficos para cada ciudad de *valor* por *precioPromedio*. En la figura 3 se pueden ver los mismos.

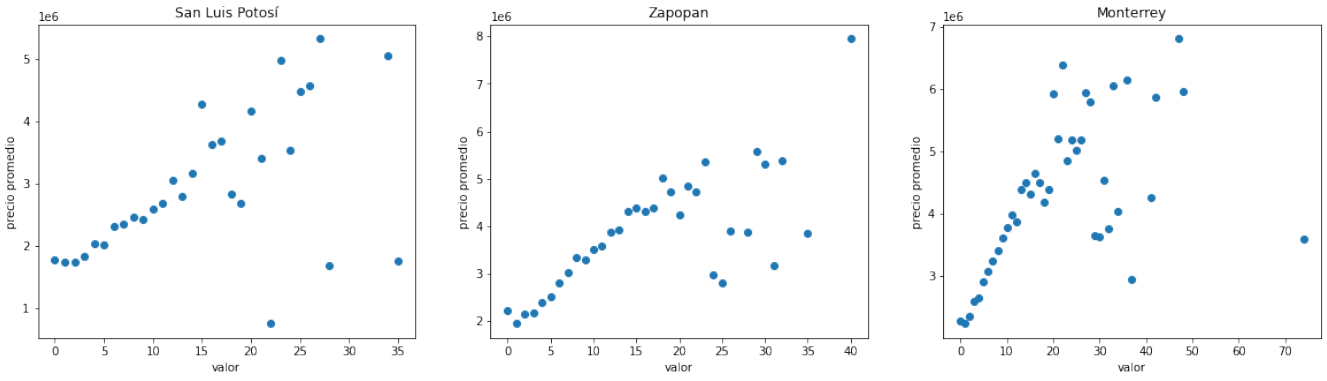


Figura 3: valor x precio promedio

[1] A modo de ejemplo, uno de los pasos en la obtención de la variable *valor* es 'traducir' las descripciones de lenguaje html a texto.

Teniendo presentes los gráficos de la imagen 3 podemos decir que *valor* y *precio* se correlacionan de una forma aproximadamente lineal. Entonces la primera $\phi(x)$ posible que se nos ocurre es $\phi(x) = x$. Pero, así como en el caso de la variable *piscina* antes vista, aquellos inmuebles con $valor = 0$ estimarían un $precio = 0$ pues CML no puede aportar un coeficiente que agrande 0. Aún sabiendo esto probamos CML con $\phi(x) = x$ con el afán de, al comparar la misma con la que nosotros creemos que funcionaría mejor $\phi(x) = x + 1$, ver que estabamos en lo cierto. El error RSMLE resultado de aplicar CML con 3-fold para ambas ϕ nos dio:

Ciudad	$\phi(x) = x$	$\phi(x) = x + 1$
Zapopan	3.724	1.069
San Luis Potosí	5.713	1.062
Monterrey	3.816	1.055

Como esperabamos usar $x + 1$ nos dio resultados más prometedores. Aún así un error logarítmico promedio de 1 no nos deja del todo conformes y nos lleva a preguntarnos cómo perfeccionar nuestro método de aproximación.

Discutiremos ideas de mejora en la sección 5.1.

4. Feature engineering - Percepción de inseguridad

Por lo mencionado anteriormente, se planteo generar un nuevo *Feature Engineering* utilizando información obtenida de fuentes externas, en este caso puntual se utilizo un *índice de percepción de inseguridad por ciudad*, el cual fue obtenido de la pagina del INEGI (Instituto Nacional de Estadística y Geografía) de México.
Como el índice nos ofrece un valor por ciudad, se opto por segmentar el conjunto de información por tipo de propiedad, tomando los tipos de propiedades con mayor cantidad de datos una vez agregado el índice al dataset, ya que así tendríamos distintos porcentajes dentro de cada segmento.
Para la estimación del precio se utilizaron las variables de antigüedad, metros totales y el índice propiamente dicho, donde para las primeras dos se mantuvieron los ϕ ya mencionados anteriormente, mientras que para el índice se definió un $\phi(x) = \frac{1}{x^2}$ suponiendo que mientras mas alto sea el valor del mismo, mas barato debería ser el inmueble.

Resultados:

Tipo de propiedad	RMSLE
Casa	0.481
Apartamento	0.547
Casa en condominio	0.360
Local Comercial	0.676

Viendo los resultados, podríamos decir que los resultados no son tan "buenos", pero creemos que esto se debe al hecho de que el índice se encuentra dividido por ciudad, lo que pensamos es muy vago a la hora de predecir una propiedad de una zona especifica. Pero creemos que de poseer un índice mas especifico, por provincia o barrio de ser posible, esto daría mucho mejores resultados.

5. Experimentación - Combinación de variables

Estudios similares a los presentados en la sección 2 nos ayudaron a encontrar algunas opciones razonables de $\phi(x)$ para otras variables. Variables que viendo su comportamiento y correlación con el precio, concluimos relevantes.

Estas variables son:

- Metros cubiertos: $\phi(x) = x$
- Antigüedad: $\phi(x) = \frac{1}{x+1}$
- Piscina: $\phi(x) = (x + 1)^2$
- Centros comerciales cercanos: $\phi(x) = (x + 1)$

- Usos múltiples: $\phi(x) = (x+1)^2$
- Gimnasio: $\phi(x) = (x+1)^2$
- Escuelas cercana: $\phi(x) = (x+1)$

Probamos 5-fold para las siguientes ciudades: Querétaro (RMSLE: 0.321), Benito Juárez (RMSLE: 0.411), Zapopan (RMSLE: 0.426), San Luis Potosí (RMSLE: 0.359). Como podemos ver los resultados son muy alentadores, logramos mejorar a nuestro temido *benchmark* que había mostrado ser mejor que cualquiera de estas variables por si solas.

¿Cuán buen error es RMSLE 0.3? La medida logarítimica nos ayuda mucho a comparar resultados entre distintas ciudades o provincias donde los precios son muy diferentes. Sin ir más lejos, en promedio el costo de un inmueble en San Luis de Potosí es de 1.9 M de pesos mexicanos mientras que en Benito Juárez es de 3.3 M. Si mirarámos el error absoluto, veríamos números más elevados casi con seguridad para Benito Juárez.

Sin embargo si observamos únicamente el RMSLE, obtendremos aquel que es el mejor, pero no podremos descifrar qué significa ser así de bueno. Únicamente tendremos información comparativa pero no cualitativa.

Habiendo obtenido el resultado que comparativamente es mejor, observemos su RMSE, para ver por cuántos pesos mexicanos estamos errando al estimar.
En Querétaro, donde el promedio es de 2.27 M erramos por 781 K (Un 34 %).
En Benito Juárez, donde el promedio es de 3.3 M erramos por 1.3 M (Un 39 %).
En Zapopan, donde el promedio es de 2.7 M erramos por 1.1 M. (Un 42 %).
En San Luis Potosí, donde el promedio es de 1.9 M erramos por 657 K (Un 33 %).

Estos resultados nos dan otra visión más cualitativa, donde nos preguntamos intuitivamente qué significan los errores que obtuvimos. En nuestra opinión estos resultados son aún un poco pobres, un 40 % (como el de Benito Juárez) de error promedio implica que dado un inmueble, habiendo estimado su precio en 4 M de pesos mexicanos, este podría haber costado en realidad 2.6 M - 5.3 M, pero además esta fluctuación es optimista, podría ser un outlier en el que falle por más aún.
El tasado de inmuebles no es un trabajo sencillo, y es muy dependiente de las condiciones de cada lugar. Dependiendo de la zona, ciertas características son más o menos valiosas. Es por esto que aunque logremos mejorar estos valores, nunca podremos hacer una diferencia significativa si probamos en lugares tan dispares con el mismo estimador. Puede que logremos error bajo en alguna ciudad, pero probablemente sea el caso contrario en otra.

5.1. Incorporación de *palabrasCaras* en el estimador integral

Nuestro objetivo es analizar si el feature *palabrasCaras* puede ayudar a mejorar el estimador que ya encontramos en la sección previa.
Utilizaremos la lista de **palabrasCaras** encontrada previamente. La misma posee 40 palabras y no diferencia a ninguna de ellas.

Nuestros primeros resultados con esta lista, no fueron muy alentadores, obteniendo la misma precisión que antes. Es decir, CML consideró a nuestro feature poco significativo y le asignó un coeficiente muy bajo. En consecuencia casi no tiene efecto en la estimación.

Pudimos verificar que esto era asi cambiando la $\phi(x)$ que inicialmente fue $\phi(x) = x + 1$, probamos con $\phi(x) = (x+1)^2$ y $\phi(x) = 1/(x+1)$. Ninguno de estos cambios generó diferencias significativas en la precisión de los estimados.
Nuestra única conclusión posible es creer que este feature no es suficientemente bueno como para competir con metros cubiertos y metros totales. Para poder observar si este feature genera impacto se nos ocurren 2 posibilidades. Como primera opción surge quitar las propiedades más correlacionadas y luego ver si el feature genera impacto en ese contexto, esta opción no nos parece muy buena, ya que aunque pudieramos observar impacto, este sería irrelevante ya que si se trata de estimar el precio, esta variable no ayuda más que otras. Otra posiblidad es hacerle mejoras al feature, de forma tal que logre un impacto real en la estimación, este será el objetivo de la parte final de esta sección.

En la figura 4 podemos ver que existe una correlación aparentemente fuerte entre el indice de palabrasCaras y el precio.

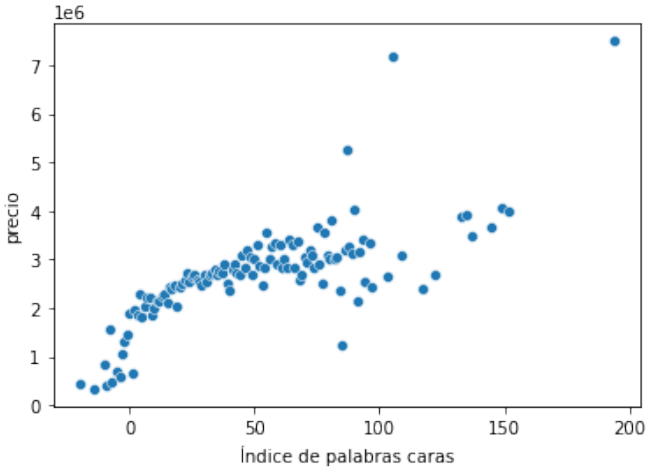


Figura 4: Índice de palabrasCaras x precio

Algunas ideas que teniamos para mejorar la calidad del estimador eran:

- Idea 1: Ejecutar este mismo experimento (con la misma lista) para los titulos.
- Idea 2: Crear una lista especial mirando el ranking de aparación de palabras en los títulos
- Idea 3: Cambiar la lista para diferenciar el valor de ciertas palabras respecto de otras, para poder expresar la idea de que 'excelente' es más significativo que 'bueno' o que 'amplio' no es tan lujoso como 'jacuzzi'.
- Idea 4: Notamos que cuanto más grande nuestra lista de palabras, mejor fue el estimativo, así que una idea es recorrer aún más los mencionados rankings y extraer más palabras.

Estas 4 ideas fueron probadas paralelamente, ninguna de ellas fue exitosa obteniendo en todos los casos resultados muy similares a la prueba original, aproximadamente un RMSLE de 0.36-0.42.
Creemos que el feature tiene mucho potencial pero que debido al poco tiempo que pudimos dedicarle a sus mejoras el mismo no mostró un aporte para la estimación de precio. Algunas otras ideas que no pudimos probar es quizás hacer un análisis de frases o de interconexión de palabras para mejorar la precisión, se podría hacer un CML interno para la predicción de humor tomando como features a cada palabra o quizás con PCA llevar las descripciones al espacio latente de palabras.

6. Experimentación - Estimar otras propiedades

Uno de los objetivos de este trabajo era, además de ser capaces de estimar el precio de un inmueble, que seamos capaces de estimar alguna otra propiedad. Ahora que ya poseemos un mayor conocimiento del comportamiento de algunas variables y los ϕ razonables de las mismas, podemos tratar de realizar la tarea de estimar algunas propiedades distintas a la de precio.
En este caso vamos a intentar estimar la cantidad de habitaciones de un inmueble y si tiene piscina. Para ambos experimentos se utilizó 5-fold pero como cada uno tiene sus consideraciones particulares, explicaremos cada uno por separado.

6.1. Estimación de habitaciones

Para este experimento se empezó por analizar el rango posible de valores, de lo cual se obtuvo que todos los inmuebles poseen entre 1 y 10 habitaciones. Se decidió segmentar la información por ciudades, en especial se tomaron

las cinco con mas datos y la estimación se realizo utilizando las variables de antigüedad, baños, metros cubiertos y precio.
Donde para el caso de antigüedad y metros cubiertos se mantuvieron los ϕ anteriormente definidos y para las demás se definieron los siguientes:

- Baños: $\phi(x) = \frac{x}{2} + 1$
- Precio: $\phi(x) = \log(x)$

Resultados:

Ciudad	RMSLE	RMSE
Querétaro	0.134	0.617
Benito Juárez	0.159	0.603
Zapopan	0.143	0.656
San Luis Potosí	0.15	0.721
Mérida	0.141	0.592

Como podemos ver el error obtenido con RMSLE es muy bajo, lo cual es muy alentador, pero lo mas apreciable es el error de RMSE donde vemos que el promedio de error es menor a 1, siendo este un gran estimador.

6.2. Estimación de piscina

En este experimento los únicos valores posibles de piscina son 0 y 1 (tiene o no tiene).
La idea es poder estimar si una propiedad posee piscina basado en su longitud y latitud, a tales fines vemos que es poco practico segmentar por ciudad, ya que dentro de cada una la variación de estos es muy chica o nula, por esa razón se decidió segmentar por tipo de propiedad, en particular tomando 6 tipos específicos. Basándonos en la idea de como cambia la temperatura media en base a la latitud y longitud, y asumiendo que la tendencia seria que las propiedades con piscina estuvieran mas cerca del ecuador (latitud 0), planteamos los siguientes ϕ:

- Latitud: $\phi(x) = \frac{x}{20}$
- Longitud: $\phi(x) = x$

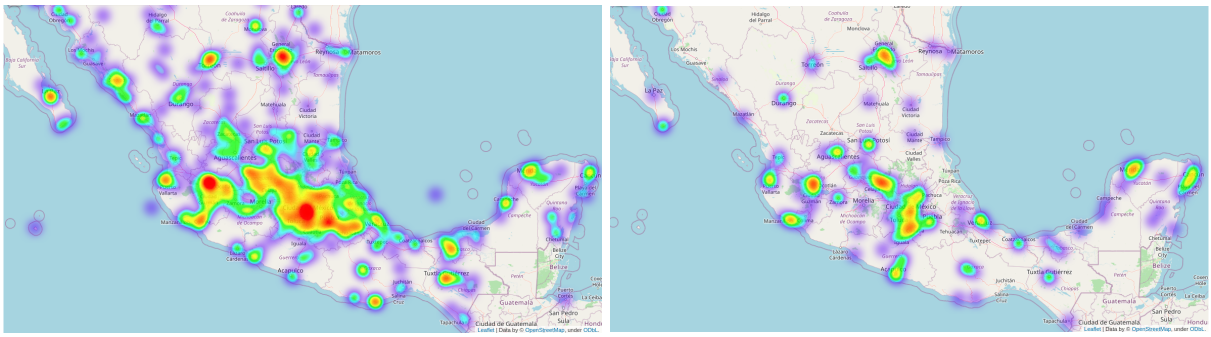
Tipo de propiedad	RMSLE
Casa	0.170
Apartamento	0.245
Casa en condominio	0.255
Terreno	0.155
Villa	0.372
Quinta Vacacional	0.328

En este caso creemos que al solo tener dos valores posibles, usar RMSE no nos dice mucho del error, por lo que optamos por solo comparar usando RMSLE.
Podemos observar que los resultados son alentadores, pero nos gustaría profundizar un poco mas en la idea planteada y ver que tanta relación existe entre estos datos y si una propiedad posee piscina.

Tomando esto en cuenta, si bien la idea original era estimar solo usando esas dos variables, luego de analizarlas pudimos ver en la Figura 5 como nuestra idea no parece cumplirse, y tanto propiedades con o sin piscina se encuentran igualmente distribuidas.

Por esta razón se opto por agregar dos variables mas para la estimación, siendo estas metros totales y metros cubiertos, el valor en usar estas variables esta en conocer los metros descubiertos que posee una propiedad (dejamos que CML se encargue de realizar dicha relación), ya que a mayor espacio descubierto, es mas probable poseer piscina. En base a esta relación que se plantea se definieron los siguientes ϕ:

- Metros totales: $\phi(x) = x^2$
- Metros cubiertos: $\phi(x) = \sqrt{x}$



(a) Sin piscina (b) Con piscina

Figura 5: Mapa de las propiedades

Resultados:

Tipo de propiedad	RMSLE
Casa	0.156
Apartamento	0.231
Casa en condominio	0.249
Terreno	0.158
Villa	0.274
Quinta Vacacional	0.371

Finalmente podemos observar como el utilizar solo dos variables mas, las cuales podríamos agrupar como metros descubiertos, vemos que obtenemos resultados mejores, siendo en algunos casos poco perceptibles pero en otros generando una gran diferencia.

7. Segmentación - Pueblos mágicos

México es un país de mucha tradición y de muchísima historia y cultura. El gobierno allí proclama "pueblo mágico."[a] toda ciudad que rememore y mantenga intacta su historia. Sitios que son museos vivientes. Lugares donde la gente vive como hace varios siglos, dignos de ver por los curiosos turistas (en general yankees). Estos lugares serán el foco de nuestra investigación, pero dada su naturaleza, debemos aunarlos para estudiarlos como un gran conjunto ya que cada pueblo por si solo no tiene suficientes datos como para poder realizar un estudio en uno solo de ellos.

Obtuvimos y emparejamos 71 pueblos mágicos formando así nuestro grupo de estudio, en él intentaremos estimar el precio en función de las otras variables.

Hipótesis Al ser pueblo de interés histórico, es posible que la antigüedad esté correlacionada positivamente con el precio. Es decir, que cuanto más antigüo, el inmueble sea más caro ya que es de más interés.

Resultados Estimamos la antigüedad con $\phi(x) = \sqrt{x+1}$ y descubrimos que la variable garage se comportaba como $\phi(x) = (x+1)^5$. También incluimos metros cubiertos, metros totales, centros comerciales cercanos y habitaciones en el estimador, utilizando ϕ's que ya probamos que son efectivas para estas variables.

El resultado fue un RMSLE de 0.388 y RMSE de 747 K (un 32 % del precio promedio). Pudimos notar una leve mejora al estimar con la antiguedad negativamente relacionada, pero no es muy significativa, estimando con $\phi(x) = 1/(x+1)$ obtuvimos RMSLE de 0.378 y RMSE de 756 K.

Como conclusión, ni siquiera en donde se supone que la antigüedad sería más importante pudimos notar una correlación significativa entre precio y antigüedad. Esta variable pareciera ser irrelevante para el estudio y estimación de precios. En cada estudio, con cualquier ϕ que usemos CML pareciera siempre asignarle un coeficiente muy bajo.

8. Conclusiones

En el presente trabajo presentamos nuestro estudio de un dataset muy muy amplio. El mismo nos enseñó cómo los datos siempre se resisten a ser homogeneos y homogeneizarlos es una tarea ardua. Primero analizamos ciertas ϕ's que luego utilizamos en el estimador a gran escala. También descubrimos que cuando una variable estima muy mal, estimar con esa variable únicamente y utilizar una ϕ que le quite importancia a la variable achica el error engañosamente. Para estas variables que podrían ayudar, pero no correlacionan tanto, la única forma de probar es en conjunto con otras variables.
En el proceso encontramos un *benchmark* de error que nos ayudaría más adelante a poder saber cuándo el error cometido comenzaba a ser aceptable. Analizamos con distintas métricas de error (RMSLE y RMSE), encontramos un estimador integral y creamos múltiples features extras. También estimamos variables distintas al precio, en particular las piscinas y las habitaciones, logrando resultados muy alentadores.

Como conclusión, la técnica de Cuadrados Mínimos Lineales puede ser útil para la estimación de problemas no tan obvios como inicialmente sugiere. La capacidad de minimizar el error es muy poderosa, sin embargo la herramienta es limitada y depende completamente de la elección de ϕ's.