DEPARTAMENTO DE COMPUTACION

Facultad de Ciencias Exactas y Naturales - UBA

TP 3 - Regresión Lineal

Keywords: Linear Regression - RMSE - RMSLE

07/12/2020 Métodos Numéricos

Grupo TP: 4

Integrante	LU	Correo electrónico
Miodownik, Federico	726/18	fede@miodo.com.ar
Sujovolsky, Tomás	113/19	tsujovolsky@gmail.com

Facultad de Ciencias Exactas y Naturales
Universidad de Buenos Aires
Ciudad Universitaria - (Pabellón I/Planta Baja)
Intendente Güiraldes 2160 - C1428EGA
Ciudad Autónoma de Buenos Aires - Rep. Argentina
Tel/Fax: (54 11) 4576-3359
http://www.fcen.uba.ar

Índice

1. Introducción

1.1. Motivación

El problema de predicción se estudia en casi todas las áreas del conocimiento humano. Es una herramienta que tiene mucha utilidad práctica, aplicable a infinidad de problemas. Hoy en día, de la mano con los avances tecnológicos en capacidad de cómputo y manejo de grandes cantidades de datos, se siguen encontrando nuevas formas tanto elegantes como de fuerza bruta para predecir información de futuras muestras a partir de las ya conocidas, y para encontrar los patrones inherentes de cada conjunto.

1.2. Presentación

En este trabajo nos dedicaremos a crear un modelo de predicción de características de inmuebles, a partir de otras características de los mismos. Utilizando aproximaciones con el método de cuadrados mínimos lineales, el objetivo del mismo es encontrar el modelo que, entrenado en parte de la muestra, explique mejor las restantes. Para determinar el concepto de "mejor", tendremos en cuenta distintas métricas, que analizaremos en profundidad para decidir por un modelo sobre otro. Finalmente, utilizaremos otros métodos para crear modelos predictivos y compararemos su eficacia contra el original.

2. Desarrollo

2.1. Regresión Lineal

2.1.1. Algoritmo

Para realizar la regresión lineal utilizamos el modelo de clase presentado por la cátedra, LinearRegression. Esta clase tiene tres componentes o funciones principales: inicialización, fitteo y predicción.

La inicialización simplemente crea la clase pero no dispone de ningún valor para ninguna variable.

Para el fitteo de los datos, el algoritmo toma como parámetros una matriz A con las muestras y un vector b con sus resultados. Nosotros decidimos resolver el problema de cuadrados mínimos lineales por medio de ecuaciones normales. Utilizando la biblioteca *Eigen*, armamos el sistema $A^T Ax = Ab$ y utilizamos su función *solve* mediante el método de HouseHolder (QR) con pivoteo total. Guardamos la solución correspondiente como variable interna.

Para predecir, el algoritmo toma como parámetro una matriz y, fila por fila, realiza el producto interno con el vector solución que tenemos guardado y guardamos su resultado en un vector. El vector que devuelve la función es entonces aquel que en la i-ésima posición tiene la predicción de la i-ésima fila de la matriz pasada por parámetro.

2.2. Métricas

Las métricas que estaremos utilizando son:

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

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

$$R^2 = 1 - \left(\frac{\sum_{i=1}^{n}(\hat{y}_i - y_i)^2}{\sum_{i=1}^{n}(\overline{y} - y_i)^2}\right)$$

Usaremos RMSE como la métrica estándar de precisión por ser una medida absoluta, es decir, comparte las mismas unidades que el resultado. Su desventaja reside principalmente en ser sensible a los outliers, especialmente los de valores absolutos altos. Para minimizar este efecto, cuando sea posible usaremos RMSLE para acompañar el análisis[1], dado que es menos sensible a los errores de alta magnitud por el efecto del logaritmo. Estaremos usando R^2 principalmente para analizar si las características consideradas pueden explicar los resultados, entendiendo un valor de 1 como que correlacionan perfectamente, 0 como que no tiene correlación ni positiva ni negativa[2], y los valores negativos de la métrica como correlación negativa.

2.3. Análisis de los Datos

El database que vamos a usar en este TP contiene alrededor de 240,000 publicaciones de viviendas. Pero, como es frecuente en los problemas reales, tiene información faltante (algunas entradas contienen NaN) y no toda la información es cuantificable, ya que hay entradas categóricas y entradas descriptivas. Esto trae problemas a la hora de realizar un análisis numérico de los datos como lo es la regresión lineal. Es por esto que debemos tomar decisiones sobre cómo vamos a operar en estos casos.

Para resolver el problema de los NaN tenemos dos soluciones, o bien rellenamos los datos con valores siguiendo algún lineamiento, o bien no utilizamos la medición que lo contenga a la hora de la evaluación. Decidimos utilizar la segunda solución, que, si bien es la mas destructiva, nos asegura tener mediciones precisas. Incluso quitando todos los NaN del database, todavía contamos con alrededor de 50.000 mediciones, que confiamos serán suficientes. Igualmente, la cantidad de mediciones variará con el experimento a realizar y cuantos NaN contengan sus variables, así que con cada experimento detallaremos cuántas mediciones tendremos disponibles.

El problema de las entradas categóricas tiene varias soluciones, en particular nosotros evaluamos dos: utilizarlas para segmentar los datos o buscar valores con los cuales reemplazarlas, es decir, volverlas variables

[1] a veces no podremos porque el predictor dará valores negativos a algunas muestras, los cuales no se pueden procesar por esta métrica.

[2] se podría pensar alternativamente como que comportase como el valor promedio, más allá del input particular,

numéricas. Un ejemplo de segmentación es hacer un predictor por tipo de propiedad, y un ejemplo de reemplazar valores puede ser tomar sólo latitud y longitud en lugar de ciudad y provincia. En la siguiente sección (3.1) evaluaremos ambas y analizaremos sus ventajas y desventajas.

3. Experimentación

Para todas las siguientes experimentaciones utilizaremos Kfold Cross Validation con $K = 5$ como método para ver que nuestros resultados no sean un overfitting de los datos.

3.1. Preliminares

Considerando la amplia cantidad de variables presentes en cada inmueble, pensamos que una manera de reducir y simplificar el problema en cuestión podría ser intentar explicar la variable precio solamente con una de las restantes. En este espíritu de simplicidad, decidimos eliminar todas las entradas que contengan NaNs para este experimento preliminar, lo que nos dejó con una base de datos de aproximadamente cincuenta mil muestras. Ahora bien, varias características son no numéricas, lo que las hace inútiles para el método de cuadrados mínimos sin algún procesamiento (en este experimento), y otras, si bien son numéricas, carecen de sentido practico para predecir precios.

Primero dejamos a un lado las variables no numéricas, si bien no descartamos procesar alguna en un paso posterior porque podrían contener información valiosa. Éstas son *titulo, descripción, tipodepropiedad, dirección, ciudad, provincia, fecha*. Si bien podríamos interpretar esta última como una variable numérica, decidimos desestimar su poco probable pero no imposible poder predictivo[3].

Y entre las numéricas que decidimos desestimar para esta primer experimentación se encuentran: *id, idzona*. Las mismas identifican con un número único a cada inmueble pero, hasta lo que entendemos, no contienen ninguna esencia del inmueble en si ni tampoco poseen un sentido intrínseco de distancia entre las muestras.

Decidimos considerar las siguientes características booleanas como numéricas, presumiendo que, por sí solas, no tendrán suficiente fuerza predictiva, pero confiando en que en un futuro, se podrán combinar para segmentar la base de datos o encontrar *features* interesantes: *gimnasio, usosmultiples, piscina, escuelascercanas, centroscomercialescercanos*

Por ultimo, listamos las restantes. Son las variables numéricas que creemos que darán los mejores resultados predictivos para la variable precio: *habitaciones, banos, metroscubiertos, metrostotales, antiguedad, garages, lat, lng*. Éstas últimas las incluiremos pero no confiamos en que, por sí mismas, resulten informativas.

	Caracteristica	RMSLE	RMSE (x 1000)	R^2
0	antiguedad	6.67	2356	-0.68
1	**habitaciones**	**0.79**	**1795**	**0.03**
2	**garages**	**3.49**	**1669**	**0.16**
3	**banos**	**0.65**	**1550**	**0.27**
4	**metroscubiertos**	**0.58**	**1433**	**0.38**
5	**metrostotales**	**0.65**	**1549**	**0.28**
6	lat	-1.00	1903	-0.09
7	lng	-0.22	1836	-0.02
8	gimnasio	13.77	2682	-1.17
9	usosmultiples	13.69	2697	-1.20
10	piscina	13.58	2691	-1.19
11	escuelascercanas	8.23	2206	-0.47
12	centroscomercialescercanos	8.58	2205	-0.47
13	precio	0.00	0	1.00

Cuadro 1: Resultados preliminares de prediccion de precio

[3]Un ejemplo donde se vuelve relevante podría ser si hubiera considerable inflación anual.

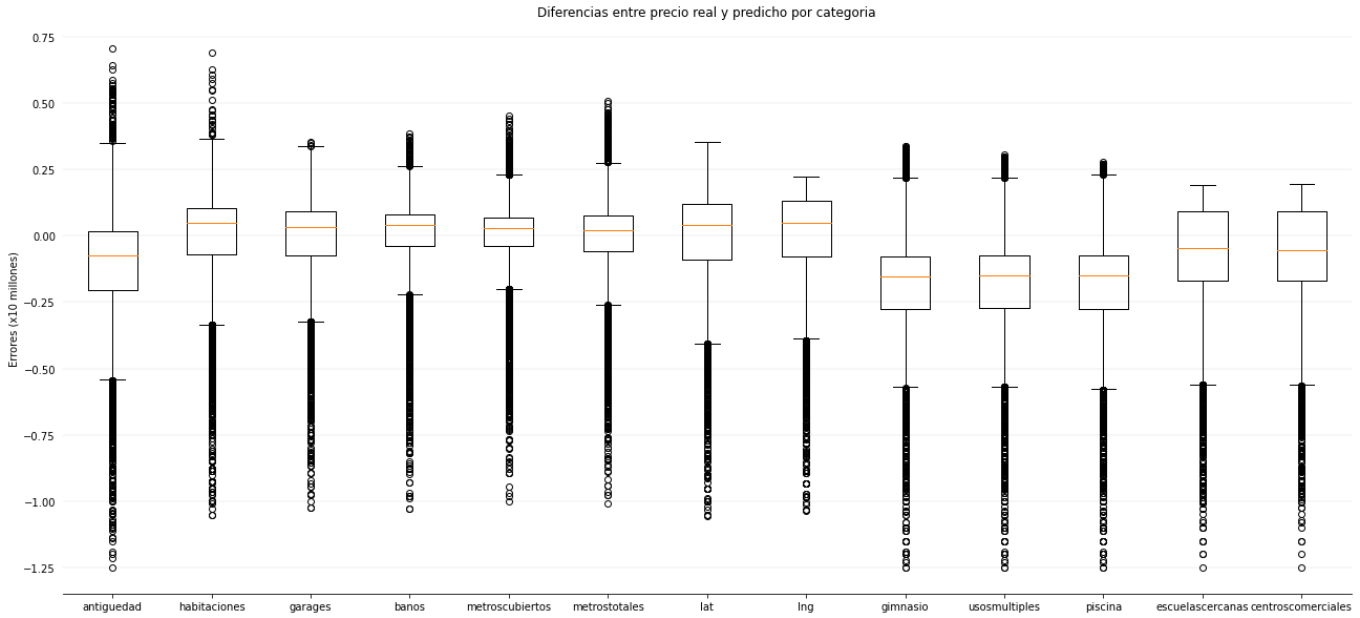
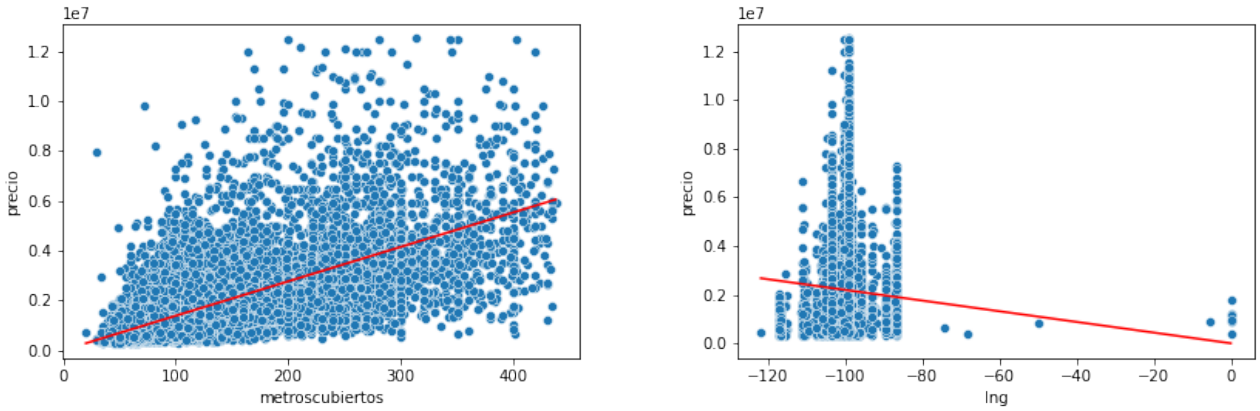


Figura 1: Boxplot comparativo de errores por categoría



(a) Precio en decenas de millones vs. Metros cubiertos

(b) Precio en decenas de millones vs. Longitud

Figura 2: Scatters y regresiones lineal de metroscubiertos y longitud

En una mirada general, notamos que los errores obtenidos varían de manera considerable de una variable a otra, lo cual se refleja en los coeficientes de correlación; algunas variables correlacionan positivamente, otras negativamente, y algunas casi no correlacionan (*antiguedad*, *lat* y *lng*) con el precio. Ahora bien, los errores absolutos, determinados por el RMSE, no necesariamente son proporcionales al índice de correlación. Observando la figura 1, se puede apreciar que algunas variables contienen outliers que devuelven errores de predicción de alrededor de 12 millones de pesos, los cuales pesan fuertemente en la métrica RMSE, incrementando el error total, mientras que otras, si bien también tienen varios outliers, devuelven diferencias de menor magnitud, resultando en un mejor RMSE.

En lo particular, también podemos observar que:

- las variables booleanas por sí solas no pueden explicar los precios, tal como intuíamos, ya que sus RMSLE y RMSE son muy altos y su R^2 es muy negativo. Que tenga sólo dos opciones (tener o no gimnasio) podría ser un indicador de si una propiedad es cara o no, pero no es muy útil para predecir un precio por sí solas, pues todas las muestras, tanto las de precio alto como bajo, se aglomeran en solamente los dos valores de la característica.
- las variables numéricas (en negrita) correlacionan positivamente con el precio, aunque con un error RMSE alto, de alrededor de 1,400,000 pesos. Se evidencia fácilmente en la figura 2a, pues si bien existe una tendencia proporcional entre los *metroscubiertos* y *precio* también se aprecia la amplia

variabilidad de los datos para con esta característica. Sabiendo que esta imagen presenta a la variable con mejores métricas entre las realizadas, queda claro que habrá que proseguir con segmentación o inteligentes combinaciones de variables, para poder mejorar el predictor.

- *lat y lng* son características que casi no correlacionan con los precios (R^2 cercano a 0), pero que curiosamente resultan en errores de RMSE más pequeños que otras variables, como las booleanas, las cuales correlacionan negativamente. Creemos que esto se evidencia al observar el scatter de los datos de Precio vs. Longitud (figura 2b). Se puede apreciar que la pendiente del regresor es poco pronunciada, de manera que, casi desestimando el valor de longitud de una entrada, el predictor devolverá un precio cercano al promedio general. Sin emargo, los outliers sí son afectados, logrando que devuelva precios negativos cuando su longitud ronda el valor 0.

A partir de los resultados obtenidos, decidimos realizar otra experimentación donde combinemos las mejores variables para explicar nuevamente el precio de los inmuebles. Esperamos poder reducir significativamente los errores encontrados por las métricas en los casos univariables. Continuaremos con el mismo manejo de NaNs que utilizamos hasta ahora, observando atentamente la cantidad de muestras con las que contamos al empezar a combinar las variables.

Luego de experimentar con varias combinaciones, la que demostró tener los errores mínimos fue *antiguedad, habitaciones, garages, banos, metroscubiertos y metrostotales*[4]. Llamaremos a este conjunto *combinación 1.*

También queremos mencionar la experimentación de *combinación 1 + lat + lng*, ya que parecía que incluirlas podría reducir los errores, pero también sucede que introduce nuevos NaNs, que por nuestra política de manejo de NaNs nos reduce la cantidad de muestras cerca de un 50 %, lo cual puede generar una pérdida de precisión para el método. Sin embargo, este efecto no se aprecia, en ambos casos el error absoluto se logró reducir frente al mejor de los singulares, cerca del 10 %, que si bien es mejor a cualquier resultado individual, no indica tampoco una mejora sustancial. Además, se introducen algunas predicciones negativas, generadas al aplicar los coeficientes con outliers, que imposibilitan calcular el RMSLE.[5]

Creemos que la combinación de variables por sí sola no resulta suficiente para explicar los precios de los inmuebles. Esto se debe a que estamos intentando agrupar datos muy variados en un mismo modelo, y por más que incluyamos más variables, no encontraremos un regresor lineal que explique bien todas.

Experimento	Cantidad de muestras	RMSE x1000	RMSLE	Rˆ2
metroscubiertos	49881	1433	0.58	0.38
combinación 1	120510	1274	-1.0	0.47
combinación 1 + lat + lng	58776	1285	-1.0	0.48

Cuadro 2: Resultados combinando variables

3.1.1. Segmentación

Hasta lo que hemos visto, incluso combinando las variables mejor correlacionadas con el precio, seguimos reportando errores altos. Para continuar mejorando, debemos encontrar buenas maneras de segmentar la base de datos, es decir, partir las muestras por categorías relevantes que permitan realizar un fitteo entre muestras con más variables en común.

La segmentación como herramienta tiene el potencial de ayudar significativamente separando muestras diametralmente distintas, las cuales, hasta ahora, han causado que la solución de cuadrados mínimos venga de la mano con errores cuadráticos gigantes. Sin embargo, encontrar las mejores segmentaciones no tiene por qué ser simple, por lo que realizaremos varios experimentos para encontrar algunas que nos parezcan adecuadas con la base de datos actual, y que creamos que sean suficientemente generales como para predecir también muestras futuras.

Comenzaremos realizando dos segmentaciones distintas y luego veremos cómo se comportan si se analizan juntas. La primera será segmentación por ubicación y la segunda será por tipo de propiedad. Aclaramos

[4]Estos resultados así como otras combinaciones se pueden reproducir en la notebook *Preliminares_vol2.*

[5]el algoritmo de RMSLE retorna -1 cuando no se puede calcular, un valor inválido.

que para estas segmentaciones notamos efectos positivos en la precisión agregando las variables booleanas, efectos que sin segmentar la base de datos no pudimos observar.

Segmentación por Ciudad

Antes de segmentar por ciudades, analizamos un poco la estructura de las mismas. El dataset tiene publicaciones de 875 ciudades distintas, pero muchas contienen pocos o incluso un sólo dato y además introducen nuevos NaNs. Por eso, decidimos analizar de manera individual a las ciudades que superen una cierta cantidad de datos sin NaN, y analizar en conjunto al resto. Esta cantidad de datos de corte la dispusimos en 1000 para este experimento. Al localizar las predicciones en cada suburbio, esperamos ver una mejora en las métricas con respecto al análisis del cuadro 2.

RMSE x1000	RMSLE	R^2
862	-1.0	0.62

Cuadro 3: Promedio ponderado de los resultados de todas las ciudades

Si se compara el resultado preliminar con el promedio de todas la ciudades (cuadro 3) las mejoras del método de segmentación se hacen notables. Con alrededor de las mismas cantidades de datos y usando las mismas variables se obtuvo una mejora del RMSE del 32 %. Yendo a mayor detalle, si se observa la tabla 17, ubicada en el apéndice 5.1 debido a su gran tamaño, se puede apreciar que varias ciudades obtuvieron una mejora sustancial reflejada en las métricas con respecto al análisis general descripto en el cuadro 2. Sin embargo, otras todavía conservan errores sustanciales (mayores a 1.500.000 millones de pesos, según su RMSE), y en general, la variabilidad entre las ciudades es alta.

Intuimos que la mejora de los resultados se debe al aumento de significancia que se obtiene de los datos al compararlos con sus pares. Es decir, al crear segmentaciones que dividen a las muestras en ciudades, se hace mas fácil distinguir las cualidades y variables que hacen que una propiedad tenga un valor especifico en esa ciudad, que podrían ser cualidades que tienen una lectura distinta en otra ciudad. Un ejemplo de esto puede ser la lectura de la variable metros cubiertos. En general, el valor que se le asigna al metro cuadrado disminuye a medida que uno se aleja de las ciudades ya que las propiedades están menos compactadas y el espacio deja de ser un problema. Es por esto que dos casas con exactamente las mismas propiedades (de las que analizamos), ubicadas en dos lugares diferentes (como por ejemplo, CABA y la costa argentina) tienen precio diferente. Por ende, vemos que la ubicación otorga valor al inmueble. Esta segmentación por ciudad podría, implícitamente, estar teniendo en cuenta este valor en el análisis.

Para cerrar el análisis, el último problema potencial de segmentación por ciudades es la cantidad de muestras por ciudad. Teniendo tan pocos datos, creemos que nuevas predicciones pueden resultar erróneas por casos severos de overfitting. Y asimismo se hace difícil de evitar subiendo el mínimo de muestras por ciudad. Si se quiere tener al menos 5000 datos por cada una, sólo podríamos distinguir Zapopan, Querétaro y **otros**. Es para saltear este factor de error que probaremos segmentar por provincia.

Segmentación por Provincia

Si bien creemos que obtendremos peor precisión en las métricas, al segmentar por provincia, tendremos más confianza en ellas ya que se realizan con más datos.

El dataset contiene información de 32 provincias: 31 estados y una capital federal. Copiando la metodología de las ciudades, definimos un punto de corte en 5000 datos. Los resultados completos, incluyendo el RMSLE, se encuentran en el apéndice 5.2.

	provincia	datos totales	datos sin NaN	RMSE promedio (x1000)	$R^2 promedio$
0	Distrito Federal	58790	26382	1580	0.49
1	Jalisco	21238	13251	827	0.69
2	Edo. de México	41607	23855	986	0.71
3	Nuevo León	15324	6593	1051	0.66
4	Querétaro	16988	8877	510	0.76
5	Puebla	10421	5883	672	0.75
6	Otros	75632	35669	686	0.63

Cuadro 4: Resultados segmentación por provincias

RMSE x1000	RMSLE	R^2
962	-1.0	0.63

Cuadro 5: Promedio ponderado de los resultados de todas las provincias

Encontramos, como previmos, que la segmentación por provincias no supera los resultados más altos de la segmentación por ciudad (como por ejemplo Tlajomulco de Zúñiga, con un RMSE de 422,000 pesos y un R^2 de 0,84), pero sí es más consistente, con el agregado de contar con una sustancial cantidad de datos por provincia. La baja variabilidad creemos que se debe a que la provincia agrega suficiente información geográfica para las propiedades, y no llegar a los resultados máximos se puede deber a la diferencia de información que otorga la provincia con respecto a la ciudad. Pero, como los resultados (sin contar el D.F.) demuestran una precisión aceptable, podemos asumir que para estas provincias y sus respectivas ciudades, la diferencia de información no es tan grande. La gran excepción es México D.F., donde la segmentación recae en las mismas imprecisiones que detallamos en la segmentación por ciudad. Es natural que una ciudad con tales características e historia (como lo es CABA en Argentina), muestre una diferencia de precios muy marcada entre los mismos barrios y calles de la ciudad. Para predecir bien precios en el D.F. sería necesaria una segmentación extra con respecto al barrio e, incluso, probablemente también con respecto a las direcciones de las propiedades.

Segmentación por Tipo de Propiedad

Otra segmentación posible es por tipo de propiedad. Con esta entrada categórica, ocurre algo interesante. Hay tres entradas que dominan el dataset: Casas, Casas en Condominios y Apartamentos. El resto de las entradas o tienen pocos datos cargados o no utilizan los datos que pedimos nosotros (vienen con NaNs). Un gran ejemplo de esto último son los Terrenos, que sólo utilizan metros totales como métrica principal. El modelo que nosotros queremos aplicar sólo describe cierto tipo de propiedades. Volviendo al ejemplo de Terrenos, hay alrededor de 9000 entradas, pero de ellas solo 21 contienen información de las características que nosotros pedimos.

Con esto en mente, y con la metodología vista en la segmentación de ciudades y provincias, realizamos la experimentación. Esta vez, el punto de corte fue 300, que si bien es muy bajo, solo los tres tipos principales superan esa cantidad. Esperamos resultados mejores que los preliminares y peores que con segmentación por ciudad o provincia, ya que pensamos que la información de ubicación es más valiosa que la de tipo de propiedad. Fuera de los tres tipos de propiedades principales, esperamos que los otros tipos (conglomerados) obtenga resultados poco precisos, en el orden de los preliminares, ya que no se agrupan por tener cualidades en común si no por tener pocas muestras de cada tipo de propiedad.

	tipo de propiedad	datos sin NaN	RMSE x 1000	R^2
0	Apartamento	22972	1328	0.65
1	Casa en condominio	11711	1112	0.64
2	Casa	84633	979	0.59
3	Otros	1194	1513	0.39

Cuadro 6: Resultados de segmentación por tipo de propiedad

RMSE x1000	RMSLE	R^2
1063	-1.0	0.60

Cuadro 7: Promedio ponderado de los resultados de todos los tipos de propiedades

Estos resultados son acordes a nuestras hipótesis y nos empujan a realizar un experimento más con respecto a la segmentación. ¿Que pasaría si mezclásemos la segmentación por provincia con la segmentación por tipo de propiedad?

Esperamos ver una visible mejora en los datos de los tres principales tipos de propiedades de cada provincia, ya que estamos agregando una información geográfica, mientras que esperamos ver que los de categoría *otros* se comporten de manera similar y sean poco precisos. Los puntos de corte serán 5000 datos válidos para las provincias y 2000 para las propiedades.

Presentamos, para conveniencia del lector, los resultados unificando las provincias, realizando un promedio ponderado con la cantidad de muestras y métricas de cada una:

	tipo de propiedad	datos sin NaN	RMSE x 1000	R^2
0	Apartamento	21318	1189	0.71
1	Casa en condominio	10412	876	0.69
2	Casa	84633	828	0.66
3	Otros	4147	1106	0.62

Cuadro 8: Resumen de segmentación por provincias + tipo de propiedad (Cuadro 21)

RMSE x1000	RMSLE	R^2
906	-1.0	0.67

Cuadro 9: Promedio ponderado de segmentación por provincias + tipo de propiedad

Comparando el cuadro 7 y 8 encontramos que realizar primero una segmentación de los datos por provincias para luego realizar la subsegmentación por tipo de propiedad resulta en predicciones más precisas. El RMSE por tipo de propiedad se reduce un 10 % para apartamentos, 21 % para casas en condominio, 15 % para casas y un 27 % para la categoría *otros*. Este resultado evidencia que encontrar segmentaciones efectivas tiene mayor efecto en la precisión del modelo que combinar distintas variables. Hilando más fino, es decir, utilizando los resultados no unificados en el apéndice 5.3, también se aprecia una mejora considerable en la predicción en un segmento hasta ahora difícil como lo es México D.F., ya que poder segmentar los apartamentos, el tipo de propiedad más presente en la capital, atrapa cualidades de los mismos que antes se nos escapaban.

Entre las desventajas que podemos encontrar con la doble segmentación, reaparece la falta de datos en las provincias donde originalmente ya habían pocos sin la segunda segmentación; éstos fueron agrupados en la categoría *otros*.

Si se compara los resultados de la doble segmentación con los de segmentación por provincia, se aprecia sólo una mejora del RMSE del 6 %. Esto significaría que la segmentación por tipo de propiedad no agregaría mucha información nueva, mientras que provincia sí lo hace. Creemos que esto se debe a su correlación con los datos numéricos que estamos analizando. Un tipo de propiedad es fácilmente descriptible por las características del inmueble, pero sus datos geográficos no tanto. Por ejemplo, un apartamento suele tener menos metros cuadrados que una casa, por lo que podemos asumir que las propiedades con menor cantidad de metros cuadrados son apartamentos. Por lo contrario es muy difícil determinar con los metros cuadrados donde está ubicada una propiedad.

3.1.2. Reemplazo de Datos

Otra propuesta para mejorar la precisión de la predicción, es reemplazar las variables categóricas, por variables numéricas. La ubicación (todas las variables geográficas, como ciudad o provincia), por ejemplo, puede ser reemplazada por latitud y longitud. De esta manera, podría ser posible generalizar más la predicción sin tener que segmentar por alguna variable geográfica.

Repetimos el mismo experimento de segmentación por tipo de propiedad, pero agregamos como característica a analizar, la latitud y longitud.

	tipo de propiedad	datos totales	datos sin NaN	RMSE x 1000	R^2
0	Apartamento	57341	12217	1313	0.65
1	Casa en condominio	19297	5949	1168	0.63
2	Casa	141717	39965	977	0.60
3	Otros	21645	645	1614	0.33

Cuadro 10: Segmentación por tipo de propiedad incluyendo lat y lng

	tipo de propiedad	datos totales	datos sin NaN	RMSE x 1000	R^2
0	Apartamento	57341	22972	1328	0.65
1	Casa en condominio	19297	11711	1112	0.64
2	Casa	141717	84633	979	0.59
3	Otros	21645	1194	1513	0.39

Cuadro 11: Segmentación por tipo de propiedad sin lat y lng (copia del cuadro 8)

Si se compara ambas tablas, no se puede apreciar una gran diferencia en el RMSE y el R^2. Lo que sí se puede apreciar es que agregando latitud y longitud, hay alrededor de un 50 % menos de datos para analizar. Que las métricas se mantengan con la mitad de los datos, nosotros lo tomamos como un indicio de que el reemplazo de variables podría llegar a ser factible, pero no nos alcanza para afirmarlo. Creemos que para mejorar este experimento, estaría bueno agregarle su latitud y longitud correspondiente a los datos que no las tienen y analizar si el resultado continúa similar, o si realmente es mejor.

3.2. feature engineering

Ante las dificultades de reducir de manera consistente los errores en la predicción de precios fruto de segmentaciones básicas, decidimos crear nuevas combinaciones de características buscando encontrar mayores similitudes entre los datos.

Una propuesta que se nos ocurrió, fue poder aprovechar las diferentes provincias como segmentación, pero agrupándolas en función de su PBI anual. Referenciando los datos oficiales del PBI del año 2019 en Informacion Estadistica y Geografica de Jalisco s.f. modificamos cada entrada de provincia por su respectivo PBI, y luego segmentamos la base de datos en:

- Provincias de PBI bajo (menor a 300,000 millones de pesos anuales)
- Provincias de PBI medio (entre 300,000 millones y un billon[6] de pesos anuales)
- Provincias de PBI alto (entre uno y tres billones de pesos anuales)
- y Mexico D.F., la unica entidad federativa con más de tres billones de pesos anuales.

	RMSE x1000	RMSLE	R^2
bajo	852	-1	0.51
medio	618	-1	0.70
alto	1007	-1	0.66
D.F.	1617	-1	0.46

Cuadro 12: segmentacion por PBI de provincias

Encontramos que esta nueva segmentación trae aparentes beneficios para todas las provincias menos Mexico D.F. (la cual contiene 26,382 datos por sí sola) reduciendo los errores absolutos por debajo del

[6]billon = un millon de millones

millón para las primeras. Además, podríamos implementar segmentaciones subsiguientes dentro de ésta, ya que la cantidad de datos bajo cada segmento permite todavía continuar hilando más fino.

PBI	tipo de propiedad	datos totales	datos sin NaN	RMSE x 1000	R^2
bajo	Apartamento	4820	1410	1316	0.65
bajo	Casa en condominio	3250	1759	531	0.72
bajo	Casa	24944	13394	611	0.64
bajo	Otros	3064	156	907	0.49

Cuadro 13: Segmentacion por PBI bajo + tipo de propiedad

PBI	tipo de propiedad	datos totales	datos sin NaN	RMSE x 1000	R^2
medio	Apartamento	5859	1803	753	0.62
medio	Casa en condominio	4751	2760	527	0.80
medio	Casa	49253	28917	594	0.71
medio	Otros	6945	222	1121	0.44

Cuadro 14: Segmentacion por PBI medio + tipo de propiedad

PBI	tipo de propiedad	datos totales	datos sin NaN	RMSE x 1000	R^2
alto	Apartamento	12782	5641	1048	0.80
alto	Casa en condominio	6696	4409	756	0.77
alto	Casa	51611	33146	890	0.69
alto	Otros	7080	503	1016	0.48

Cuadro 15: Segmentacion por PBI alto + tipo de propiedad

PBI	tipo de propiedad	datos totales	datos sin NaN	RMSE x 1000	R^2
D.F.	Apartamento	33839	14118	1281	0.67
D.F.	Casa en condominio	4596	2783	1548	0.47
D.F.	Casa	15812	9168	1619	0.36
D.F.	Otros	4543	313	1985	0.28

Cuadro 16: Segmentacion por PBI maximo + tipo de propiedad

Observando los resultados, podemos apreciar mejoras generales en las predicciones, para todas las segmentaciones de PBI. En particular, tanto *Casas* como *Casas en condominio* redujeron el RMSE entre un 0 y un 40 %, con las mejoras mas apreciables en las tres primeras segmentaciones, donde la cantidad de datos en estas categorías son las mayores. En lineas similares en Mexico D.F., la única provincia de PBI máximo, en la categoría de *Apartamentos*, el predictor reduce el RMSE un 21 %. Si bien esta mejora no se aprecia en los demás tipos de propiedad de la provincia, es en *Apartamentos* donde se encuentra la mayor densidad de datos (un poco mas de la mitad). Igualmente, el D.F. acarrea los errores que le atribuimos anteriormente, y es que es una ciudad con gran variabilidad de precios entre sus propios barrios y requiere una segmentación extra para ser predicho de manera satisfactoria.

En conclusión, el feature de PBI con segmentación por tipo de propiedad resulta en mejores predicciones para una gran parte de los datos, a costa de la pérdida en la precisión para las minorías. Sin embargo, desde una perspectiva utilitaria, las mejoras son apreciables, porque las muestras que se escapan al predictor son pocas en comparación a las que está acertando mejor.

3.3. Comparación con otros métodos

3.3.1. Redes Neuronales

Como método de comparación para poder analizar la performance de cuadrados mínimos, nos propusimos sumarnos a la moda de las redes de aprendizaje profundo. Utilizando la guía de https://github.com/Harshita9511 s.f. construimos tres modelos de red neuronales[7]:

- model1: cuatro capas con sesgos aleatorios.
- model2: cuatro capas sin sesgos.
- model3: tres capas sin sesgos.

El modelo que mejor performance obtuvo fue el tercero, y fue el que continuamos optimizando hasta llegar a un mínimo local del cual ya no pudo continuar mejorando en un tiempo de cómputo razonable.

La predicción a partir de redes neuronales funciona de manera muy distinta a la regresión lineal. La primera capa tiene una neurona por cada input (característica del dataset) y presentará una activación entre 0 y 1 (tuvimos que normalizar todas las variables entre cero y el máximo encontrado). La segunda capa de tres neuronas se conecta con cada una de la capa anterior, y dependiendo de su activación y el peso de cada neurona (en un principio aleatorio) obtendrá su propio valor. Finalmente estas tres neuronas se combinan de la misma forma para obtener una última neurona que será la salida, el precio predicho.

Frente al error obtenido en cada paso del entrenamiento, la red realiza ajustes de los pesos de cada conexión, comenzando del final hacia el comienzo, de manera de obtener una combinación que, con el mismo input, genere menor error en el futuro. Esto se repite una cantidad fija de veces, y al terminar prueba su efectividad en el conjunto de validación. Una vez hecho esto, termina lo que llamaremos una **generación**.

Cabe mencionar, que debido al proceso auto correctivo que realiza la red, al completar el entrenamiento se termina fijando alguna conexión implícita entre las variables iniciales y el precio final del inmueble, pero ésta no se puede extrapolar como información valiosa sobre los datos, ya que debido a la inicialización aleatoria, al repetir el proceso desde cero se llegarían a resultados similares, pero con conexiones completamente diferentes.

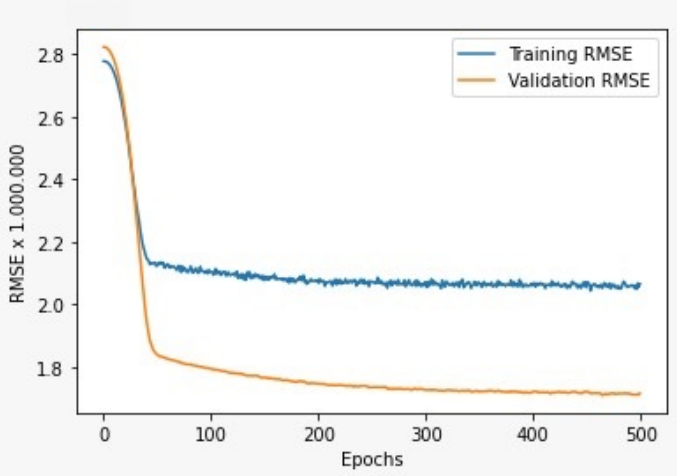


Figura 3: Primeras 500 generaciones

[7] se pueden encontrar en la notebook *Neural Network*

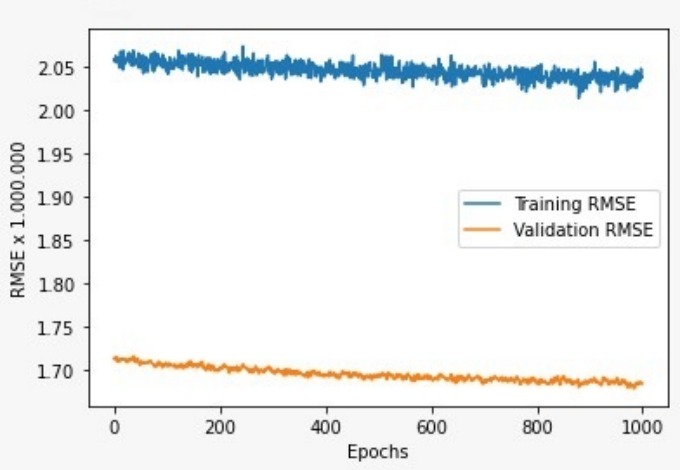


Figura 4: Últimas 1000 generaciones

En las dos imágenes podemos observar una característica muy común en el aprendizaje de redes neuronales, que es su rápida optimización en las primeras generaciones, y la creciente dificultad de continuar mejorando cercano al mínimo local. Para ilustrarlo en números, entre la generación 1 y 100 se reduce el RMSE un 150 % mientras que entre la 100 y 1500 solamente mejora un subsiguiente 5 %.

Con este método logramos obtener, sin ninguna segmentación, un RMSE de 1,683,000 pesos, contra un RMSE de cuadrados mínimos de 1,274,000 pesos bajo las mismas condiciones.[8] En comparatoria, no resultó superior a la regresión lineal, aunque queda abierta la pregunta de si, con mayor tiempo de cómputo, se podría mantener la tendencia de mejora suave hasta superar cuadrados mínimos, o incluso, si la implementación particular fue una buena elección entre las posibles.

[8]Este valor se extrae del cuadro **??**

4. Conclusiones

Queremos comenzar remarcando el enorme efecto que hace a la precisión de un análisis predictivo, la calidad y cantidad de los datos disponibles. Trabajar con un dataset real, con muchos agujeros de información y datos faltantes, supuso un desafío enorme para sortear dificultades que hasta ahora no nos habiamos enfrentado en otros trabajos de prediccion.

Si bien propusimos a lo largo de la experimentación diferentes soluciones y enfoques para mejorar el predictor de precios, creemos que no pudimos encontrar una propuesta que nos deje satisfechos, confiable y que correlacione adecuadamente para cada segmento. Debemos en parte esto a una dificultad propia del dataset, donde costó encontrar los patrones donde se podían agrupar mejor los datos, y también a la fuerza de los outliers fruto de esta primer dificultad, los cuales aumentaron el error global. Al decidir no utilizar los NaNs, creímos, como explicamos en 2.3, que contaríamos con suficientes muestras. Sin embargo, los problemas que conllevaron a la necesidad de segmentar se incrementaron debido a esto.

Para cerrar, escribiremos nuestra propuesta de modelo para realizar predicciones sobre precios de inmuebles a partir de todo lo aprendido. Creemos que lo más importante es tener un dataset de entrenamiento sólido, que tenga bien definidas las variables (la menor cantidad de NaNs posibles) y una buena cantidad de datos por cada separación categórica. Sabemos que en la realidad esto es difícil. Pero, como la base del predictor es el dataset de entrenamiento, hay que prestarle la mayor atención posible.

El segundo paso sería explorar el dataset. Identificar a fondo qué describe, para segmentar de la mejor manera posible. Un ejemplo seria el tipo de propiedad *Terrenos*, que creemos debería utilizar un predictor distinto al de los tipos residenciales, ya que se diferencian en demasiadas variables. Otro ejemplo: México D.F., la capital federal, vista en la tabla de segmentación por provincias, requiere de una segmentación extra en barrios o en tipos de propiedad (esta última mejoró las métricas, pero siguieron siendo bajas). Entender qué requiere el dataset y cómo dividirlo para realizar las mejores predicciones por categoría, es nuestra segunda gran recomendación, ya que fue lo que mejores resultados nos devolvió.

El tercer paso es sacar a la vista información oculta o agregar información faltante a la muestra. Una vez identificado cómo se puede aprovechar el dataset, hacer feature engineering puede mejorar un poco las métricas, o por lo menos ayudar a segmentar en bloques más manejables, como nuestra segmentación por PBI.

Con nuestro dataset y problemática específica, creemos que la mejor manera de encarar el problema es separando terrenos y sectores industriales de los residenciales, luego dividir por provincias y además subdividir al D.F. por tipos de propiedad. Creemos que con estos lineamientos se podrían conseguir mejores datos, minimizando el riesgo de overfitear. Termina siendo un compromiso entre resultados y confianza, los mejores resultados posibles manteniendo muestras razonablemente grandes.

5. Apéndice

5.1. Tabla de Segmentación, Ciudades

	ciudad	datos totales	datos sin NaN	RMSE promedio (x1000)	$R^2 promedio$
0	Benito Juárez	11014	4831	1259	0.54
1	Zapopan	10360	6421	781	0.75
2	Coyoacán	5293	2837	1339	0.55
3	San Luis Potosí	7925	2546	470	0.83
4	Querétaro	12646	6405	516	0.78
5	Naucalpan de Juárez	6554	3653	1083	0.53
6	Monterrey	6946	2623	783	0.72
7	Cancún	3779	1563	917	0.65
8	Puebla	4636	2851	561	0.75
9	Miguel Hidalgo	5795	2238	2045	0.55
10	Mérida	7162	2397	560	0.64
11	Huixquilucan	5718	2073	1512	0.54
12	Atizapán de Zaragoza	5783	2824	830	0.69
13	Tlalpan	5721	2685	1212	0.53
14	Cuautitlán Izcalli	3408	2405	391	0.71
15	Metepec	1996	1114	636	0.71
16	Cuauhtémoc	6614	2636	1724	0.50
17	Alvaro Obregón	6633	2841	1649	0.55
18	Tlajomulco de Zúñiga	3254	2078	422	0.84
19	San Andrés Cholula	3805	1825	756	0.70
20	Chihuahua	3757	1523	486	0.77
21	Cuernavaca	3775	1391	651	0.66
22	Tlalnepantla de Baz	3588	2142	681	0.56
23	Hermosillo	2590	1216	531	0.77
24	Gustavo A. Madero	3141	1600	983	0.58
25	Tijuana	2863	2155	280	0.50
26	Guadalajara	4006	2372	1025	0.63
27	Cuajimalpa de Morelos	3020	1161	1732	0.53
28	Iztapalapa	3093	1580	643	0.62
29	Durango	2048	1074	325	0.73
30	Toluca	2026	1329	398	0.72
31	Ecatepec de Morelos	2159	1488	387	0.57
32	Pachuca	1386	1020	421	0.76
33	Corregidora	2018	1255	369	0.78
34	Saltillo	1845	1232	423	0.77
35	Otros	73643	39126	856	0.56

Cuadro 17: Resultados segmentación por ciudades

	ciudad	RMSLE 1	RMSLE 2	RMSLE 3	RMSLE 4	RMSLE 5
0	Benito Juárez	0.385	0.392	0.390	0.401	0.382
1	Zapopan	-1.000	0.406	-1.000	-1.000	0.377
2	Coyoacán	0.417	0.371	0.374	0.387	0.393
3	San Luis Potosí	-1.000	0.299	0.278	0.314	0.318
4	Querétaro	0.260	0.236	0.243	0.242	0.237
5	Naucalpan de Juárez	0.336	0.314	0.312	0.316	0.318
6	Monterrey	-1.000	-1.000	0.361	0.395	0.345
7	Cancún	-1.000	0.431	-1.000	-1.000	-1.000
8	Puebla	0.399	-1.000	0.476	-1.000	-1.000
9	Miguel Hidalgo	-1.000	0.504	0.542	-1.000	0.508
10	Mérida	-1.000	0.280	-1.000	0.312	-1.000
11	Huixquilucan	0.343	0.313	0.313	0.320	0.295
12	Atizapán de Zaragoza	0.355	0.374	0.345	0.336	0.328
13	Tlalpan	0.384	0.351	0.397	0.360	-1.000
14	Cuautitlán Izcalli	0.308	0.326	0.303	0.337	0.303
15	Metepec	0.425	0.287	0.259	0.280	0.265
16	Cuauhtémoc	0.549	0.595	0.530	0.527	0.550
17	Alvaro Obregón	-1.000	-1.000	-1.000	-1.000	0.456
18	Tlajomulco de Zúñiga	0.310	-1.000	-1.000	0.300	-1.000
19	San Andrés Cholula	0.267	-1.000	0.265	0.265	0.261
20	Chihuahua	-1.000	-1.000	0.438	-1.000	0.364
21	Cuernavaca	0.273	0.284	0.284	0.293	0.262
22	Tlalnepantla de Baz	0.303	0.312	0.313	0.310	0.318
23	Hermosillo	-1.000	0.476	-1.000	-1.000	0.439
24	Gustavo A. Madero	0.412	0.438	0.397	0.420	0.404
25	Tijuana	0.303	-1.000	0.311	0.338	-1.000
26	Guadalajara	-1.000	-1.000	-1.000	-1.000	-1.000
27	Cuajimalpa de Morelos	-1.000	0.402	0.424	-1.000	0.507
28	Iztapalapa	0.402	0.376	0.362	0.377	0.374
29	Durango	0.264	0.279	0.277	0.287	0.297
30	Toluca	0.263	0.299	0.288	0.284	0.268
31	Ecatepec de Morelos	0.307	0.280	0.284	0.282	0.303
32	Pachuca	0.302	0.337	0.306	-1.000	0.333
33	Corregidora	0.185	0.209	0.211	0.253	0.210
34	Saltillo	-1.000	-1.000	0.390	0.330	-1.000
35	Otros	-1.000	-1.000	-1.000	-1.000	-1.000

Cuadro 18: Resultados segmentación por ciudades (RMSLE)

5.2. Tabla de Segmentacion, Provincias

	provincia	datos totales	datos sin NaN	RMSE promedio (x1000)	$R^2 promedio$
0	Distrito Federal	58790	26382	1580	0.49
1	Jalisco	21238	13251	827	0.69
2	Edo. de México	41607	23855	986	0.71
3	Nuevo León	15324	6593	1051	0.66
4	Querétaro	16988	8877	510	0.76
5	Puebla	10421	5883	672	0.75
6	Otros	75632	35669	686	0.63

Cuadro 19: Resultados segmentación por provincias

	provincia	RMSLE 1	RMSLE 2	RMSLE 3	RMSLE 4	RMSLE 5
0	Distrito Federal	0.477	0.479	-1.000	0.478	-1.000
1	Jalisco	-1.000	-1.000	-1.000	-1.000	-1.000
2	Edo. de México	-1.000	-1.000	-1.000	-1.000	-1.000
3	Nuevo León	-1.000	-1.000	-1.000	-1.000	-1.000
4	Querétaro	0.256	0.240	0.253	0.254	0.244
5	Puebla	-1.000	-1.000	-1.000	-1.000	-1.000
6	Otros	-1.000	-1.000	-1.000	-1.000	-1.000

Cuadro 20: Resultados segmentación por provincias (RMSLE)

5.3. Tabla de Segmentacion, Tipos De Propiedad y Provincia

	Provincia	propiedad	datos totales	datos sin NaN	RMSE promedio (x1000)	R^2
0	Distrito Federal	Apartamento	33839	14118	1252	0.69
1	Distrito Federal	Casa en condominio	4596	2783	1539	0.47
2	Distrito Federal	Casa	15812	9168	1610	0.37
3	Distrito Federal	Otros	4543	313	1980	0.29
4	Jalisco	Casa	14196	10108	697	0.73
5	Jalisco	Otros	7042	3143	1040	0.67
6	Edo. de México	Apartamento	8297	3516	992	0.83
7	Edo. de México	Casa en condominio	4717	3082	780	0.78
8	Edo. de México	Casa	25938	16994	932	0.70
9	Edo. de México	Otros	2655	263	912	0.60
10	Otros	Apartamento	12567	3684	1137	0.66
11	Otros	Casa en condominio	8079	4547	535	0.77
12	Otros	Casa	85771	48363	671	0.68
13	Otros	Otros	11948	428	1075	0.47

Cuadro 21: Resultados segmentación por provincias y tipo de propiedad

	Provincia	propiedad	RMSLE 1	RMSLE 2	RMSLE 3	RMSLE 4	RMSLE 5
0	Distrito Federal	Apartamento	-1.000	-1.000	-1.000	-1.000	-1.000
1	Distrito Federal	Casa en condominio	-1.000	0.419	0.382	0.376	0.385
2	Distrito Federal	Casa	0.473	-1.000	-1.000	0.460	0.474
3	Distrito Federal	Otros	0.612	0.475	0.493	0.533	0.475
4	Jalisco	Casa	-1.000	-1.000	-1.000	-1.000	-1.000
5	Jalisco	Otros	0.459	-1.000	0.471	-1.000	0.443
6	Edo. de México	Apartamento	-1.000	-1.000	-1.000	-1.000	-1.000
7	Edo. de México	Casa en condominio	0.357	-1.000	0.383	0.366	0.384
8	Edo. de México	Casa	-1.000	-1.000	-1.000	-1.000	-1.000
9	Edo. de México	Otros	0.477	0.461	0.528	0.520	0.586
10	Otros	Apartamento	-1.000	-1.000	-1.000	-1.000	-1.000
11	Otros	Casa en condominio	0.284	0.289	-1.000	-1.000	0.290
12	Otros	Casa	-1.000	-1.000	-1.000	-1.000	-1.000
13	Otros	Otros	0.600	0.506	0.575	0.550	0.511

Cuadro 22: Resultados segmentación por provincias y tipo de propiedad (RMSLE)

Referencias

[1] https://github.com/Harshita9511. Predicting household prices using Keras Tensorflow. URL: https://medium.c (accessed: 07.12.2020).

[2] Instituo de Informacion Estadistica y Geografica de Jalisco. PIB anual 2018 por entidad federativa - Mexico. URL: https://iieg.gob.mx/ns/wp-content/uploads/2020/02/Boletin_economico_anual_2019.pdf. (accessed: 03.12.2020).