FACULTAD DE CIENCIAS ECONÓMICAS

FAMAF
Facultad de Matemática, Astronomía y Física

UNC
Universidad Nacional de Córdoba

PRECIO DE VENTA DE LOCALES COMERCIALES EN CABA

Diplomatura en Ciencia de datos, inteligencia artificial y sus aplicaciones en Economía y negocios

TRABAJO FINAL: PRECIO DE VENTA DE LOCALES COMERCIALES

Diplomatura en Ciencia de datos, inteligencia artificial y sus aplicaciones en Economía y negocios.

Mariano Jesús Medaglia
Luis Wayar
Ignacio González Cabañas

INTRODUCCIÓN

El siguiente informe busca analizar las características de los inmuebles comerciales en la Ciudad Autónoma de Buenos Aires (CABA) y elaborar con ellos un modelo capaz de predecir el precio de venta de los locales.

DESCRIPCIÓN DEL PROBLEMA

Se presenta el problema de conocer cuales es y cuál va ser el precio de venta de los locales comerciales de la Ciudad Autónoma de Buenos Aires (CABA) eso mismo tomando a consideración las características distintivas que les son atribuidas propiamente e información georreferenciada las cuales son reconocidas por su afectación socioeconómica en la población objetivo bajo estudio (locales comerciales).

POSIBLE UTILIZACIÓN DE LAS PREDICCIONES

Buscamos dejar claramente denotado la importancia de conocer el precio de los locales comerciales para derivar en una íntegra comprensión de los posibles interrogantes que podemos abordar con el mismo, al predecir logramos hacer una **valoración relativa** del inmueble, también conseguimos hacer una aproximación del nivel económico de la región , lo mismo no es erróneo nombrar región antes que población, ya que no solo es un indicador de esta última, sino que a la vez nos orientara de sectores con diferencias en servicios públicos, lo cual sería información **crucial** tanto **para inversionistas o empresarios** al poder contrastar estos valores con la **proyección de flujos futuros** (Descuento de flujos de caja) como así **para políticas públicas** de carácter horizontal (asistencia general sin distinguir características propias del sector) y/o de carácter vertical (alícuotas diferenciales por región, valor, características del inmueble etc) .

Antes de mencionar otras de las posibles utilizaciones de este modelo, nos es de interés mencionar un hecho que es de conocimiento público, el cual es el alto nivel de volatilidad económica que tiene Argentina, lo que deriva entre otras cuestiones, en valuaciones

dolarizadas de inmueble, que esto no es más que considerar como factor externo no controlable y muy variable como es así el dólar, aunque no hablamos de valor como tal sino que de dolares fisicos, sabemos que no es el único factor en relación al bien considerado. Entonces podemos decir que este modelo provee una aproximación de los valores de los locales comerciales para un sector robusto y que es distintivo de la economía argentina, las PYMES.

El modelo nos daría el precio al que supuestamente el mercado valora el local, obteniendo una estimación de cuál es el efecto de las características de un inmueble sobre su precio, permitiendo las posibles comparaciones y proyecciones de precio que se pueden conseguir al aplicar diferentes combinaciones de variables sobre el modelo, logrando visualizar diferentes escenarios en evaluación.

Otra forma de verlo es pensar que el precio que nos devuelve el modelo es una guía para estimar si un inmueble está sobrevaluado o subvaluado, para identificar oportunidades de compra o de venta.

Si bien los datos originalmente debieron poseer una estructura temporal implícita, no se va a prestar atención a esa dimensión dado que no está incluida en el set de datos. Se les dará tratamiento de corte transversal.

ANÁLISIS EXPLORATORIO

En principio encontramos una base de datos bastante limpia y depurada, casi sin datos ausentes. Encontramos 3711 registros.

Nos interesa saber que nuestro *target,* que es el precio, posee una distribución asimétrica positiva con curtosis positiva otorgando la categorización de la misma como Leptocúrtica.

Realizamos una matriz de correlaciones para explorar la relación entre las distintas variables. Surge que las variables que están más correlacionadas con el precio son el total de metros cuadrados (0,72) y el total de metros cuadrados cubiertos (0,71). Esto es esperable dado que el tamaño del inmueble siempre tiene una enorme influencia en el precio, sin embargo lo que buscamos explorar es la relación del precio con otras variables más interesantes, por ende el tamaño será una variable de control importante.

En términos geográficos, un hallazgo importante es la existencia de una fuerte correlación negativa entre la distancia al obelisco y la cantidad de robos (-0,72), de locales gastronómicos (-0,64) y de centros culturales (-0,54).

Examinando la distribución de cada variable numérica, encontramos que muchas (precio del metro cuadrado, distancia al obelisco, robos, espacios verdes, gastronomía y centros culturales) presentan una distribución del tipo X^2 (chi-cuadrado).

Para solucionar el problema de los datos extremos que no permite una clara identificación del comportamiento de cada variable o de nuestro target resultante, generamos una nueva base que es igual a la original pero sin outliers sobre este último.

DECISIONES TOMADAS AL LIMPIAR LOS DATOS

Observamos que al extraer los datos outliers de la muestra nos da que el intervalo de confianza de la media en precios (USD) se encuentra en:

```
(294035.97779275454, 309165.8113743975)
```

Cuando en la base pura se encontraría en:

```
(399991.45293881965, 440004.86772946385)
```

Siendo la misma comparativa para m2(USD):

```
(2847.8776784495835, 2964.950628494505)

(2976.226629270523, 3106.327014766123)
```

Es claro que los datos extremos ampliaban el intervalo en que se encontraba nuestra media objetivo de precio USD.

En el caso de la variable *cantidad de ambientes,* encontramos que toma valor 0 en la mayoría de los casos, esto parece ser un defecto de medición que puede justificar la exclusión de la variable.

Del conjunto de datos, separamos nuestro target (preciosUSD) de nuestras características o inputs del modelo, se asigna el 80% al conjunto de entrenamiento.

MODELO APLICADO

Análisis de regresión lineal

Se realiza una regresión múltiple y a continuación se tiene el reporte de la regresión

OLS Regression Results

Dep. Variable:	precioUSD	R-squared:	0.657
Model:	OLS	Adj. R-squared:	0.656
Method:	Least Squares	F-statistic:	591.3
Date:	Tue, 14 Dec 2021	Prob (F-statistic):	0.00
Time:	09:52:13	Log-Likelihood:	-52783.
No. Observations:	3711	AIC:	1.056e+05
Df Residuals:	3698	BIC:	1.057e+05
Df Model:	12		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
const	-2.548e+05	3.68e+04	-6.926	0.000	-3.27e+05	-1.83e+05
antig	-479.4636	279.879	-1.713	0.087	-1028.197	69.269
m2total	1072.3881	62.837	17.066	0.000	949.189	1195.587
m2cub	487.5780	62.841	7.759	0.000	364.371	610.785
banios	2.831e+04	4745.458	5.966	0.000	1.9e+04	3.76e+04
comuna	631.6966	1541.324	0.410	0.682	-2390.233	3653.626
m2precioUSD	106.4397	3.102	34.316	0.000	100.358	112.521
comisaria_dista	12.5710	18.393	0.683	0.494	-23.491	48.633
obelisco_dista	0.1482	3.099	0.048	0.962	-5.928	6.224
nrobos	58.5201	136.811	0.428	0.669	-209.713	326.753
sup_espacio_verde	0.1349	0.080	1.683	0.092	-0.022	0.292
count_gastronomia	1040.7442	188.176	5.531	0.000	671.805	1409.684
count_culturales	-866.3106	593.505	-1.460	0.144	-2029.939	297.318

Omnibus:	5185.932	Durbin-Watson:	1.649
Prob(Omnibus):	0.000	Jarque-Bera (JB):	7430230.269
Skew:	7.401	Prob(JB):	0.00
Kurtosis:	221.710	Cond. No.	7.05e+05

Siguiendo lo indicado por los valores p (grado de significación), la variable comuna no es estadísticamente significativa (lo que es correcto dado que solamente es categórica).

Por otro lado, la distancia a la comisaría, la distancia al obelisco, la cantidad de robos y la cantidad de centros culturales tampoco muestran significación estadística, siendo esta última variable la que tiene mejor grado de significación.

Con respecto a los resultados globales, tenemos que el grado de significación del estadístico F es 0.00 lo que indica que la regresión es significativa en conjunto. Además tenemos un R cuadrado de 0.657 (65,7%) y un R cuadrado ajustado de 0.656 (65.6), lo que indica un muy buen nivel de ajuste y que el aumento de variables no baja el nivel de precisión del modelo. La superficie y la superficie cubierta total, si bien presentaban una alta correlación que podía inducir a imprecisión, se muestran ambas estadísticamente significativas, esto es seguramente a raíz de que el gran tamaño de la muestra consigue compensar la correlación entre ambas. Finalmente, la variable precio en dólares del metro cuadrado es la que tiene mayor significancia, esto es esperable dado que posee alta correlación lineal con el precio del local.

Repetimos, ahora con una regresión robusta a la heterocedasticidad y quitando algunas de las variables:

OLS Regression Results

Dep. Variable:	precioUSD	R-squared:	0.657
Model:	OLS	Adj. R-squared:	0.656
Method:	Least Squares	F-statistic:	43.53
Date:	Wed, 15 Dec 2021	Prob (F-statistic):	8.87e-75
Time:	20:31:08	Log-Likelihood:	-52783.
No. Observations:	3711	AIC:	1.056e+05
Df Residuals:	3701	BIC:	1.056e+05
Df Model:	9		
Covariance Type:	HC1		

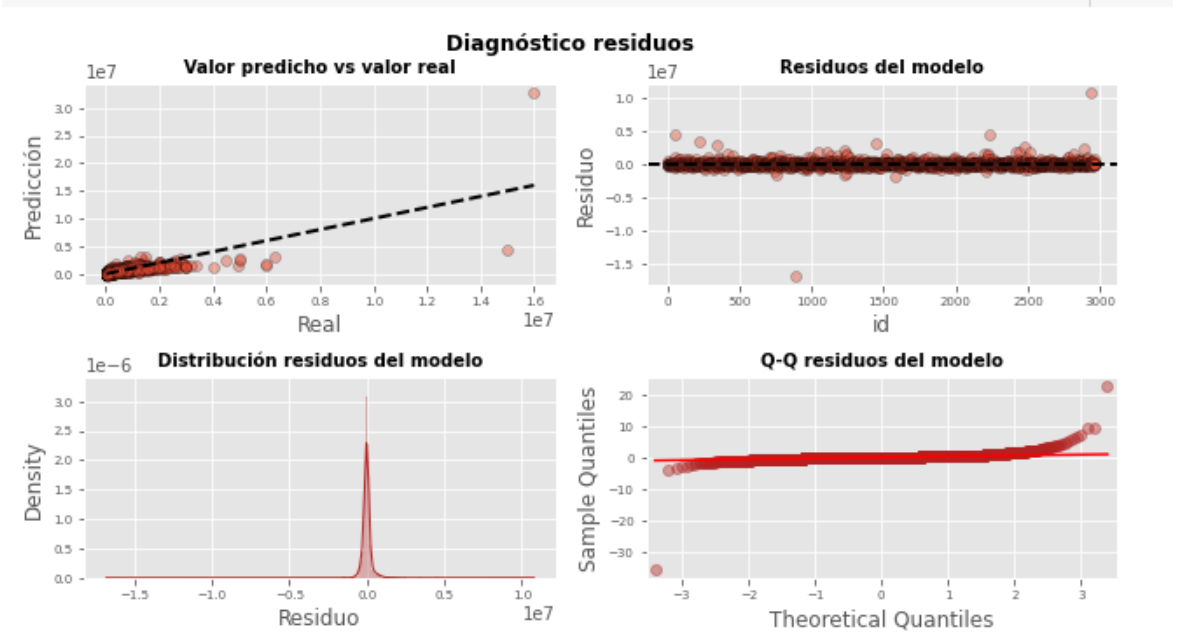
	coef	std err	t	P>|t|	[0.025	0.975]
const	-2.326e+05	5.61e+04	-4.150	0.000	-3.43e+05	-1.23e+05
antig	-485.8555	320.019	-1.518	0.129	-1113.287	141.576
m2total	1070.2973	331.686	3.227	0.001	419.992	1720.603
m2cub	489.5749	349.950	1.399	0.162	-196.539	1175.689
banios	2.855e+04	1.32e+04	2.168	0.030	2733.043	5.44e+04
m2precioUSD	106.5123	9.398	11.334	0.000	88.087	124.937
nrobos	10.3649	99.562	0.104	0.917	-184.836	205.566
sup_espacio_verde	0.1246	0.063	1.988	0.047	0.002	0.247
count_gastronomia	1021.0838	336.721	3.032	0.002	360.906	1681.261
count_culturales	-906.1963	662.490	-1.368	0.171	-2205.077	392.685

Omnibus:	5182.153	Durbin-Watson:	1.651
Prob(Omnibus):	0.000	Jarque-Bera (JB):	7417479.509
Skew:	7.390	Prob(JB):	0.00
Kurtosis:	221.523	Cond. No.	3.67e+05

Entonces, podemos afirmar que el precio depende positivamente del tamaño del inmueble, del precio del metro cuadrado (estos dos hallazgos son más bien triviales), de la cantidad de superficie cubierta, de la cantidad de baños, del espacio verde, de la cantidad de locales gastronómicos cercanos y depende negativamente de la antigüedad y de la cantidad de locales culturales, todos estos casos en la cuantía que se muestra en la tabla.

Métrica seleccionada

Realizamos análisis de validación cruzada

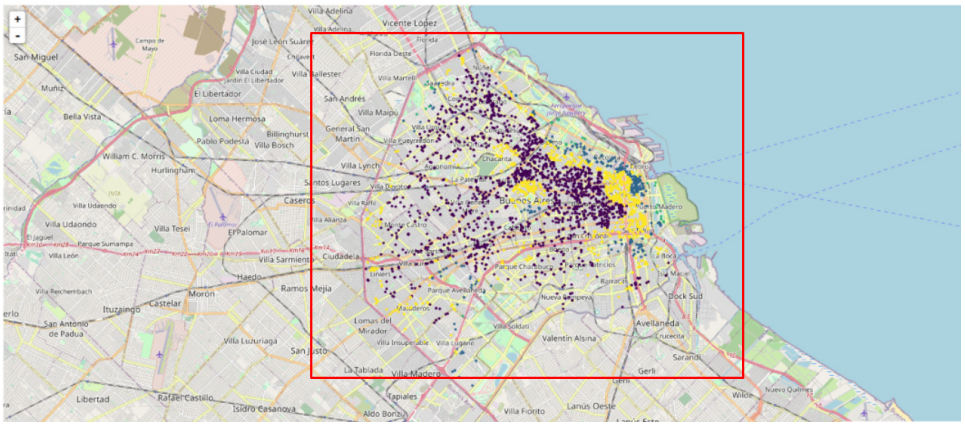


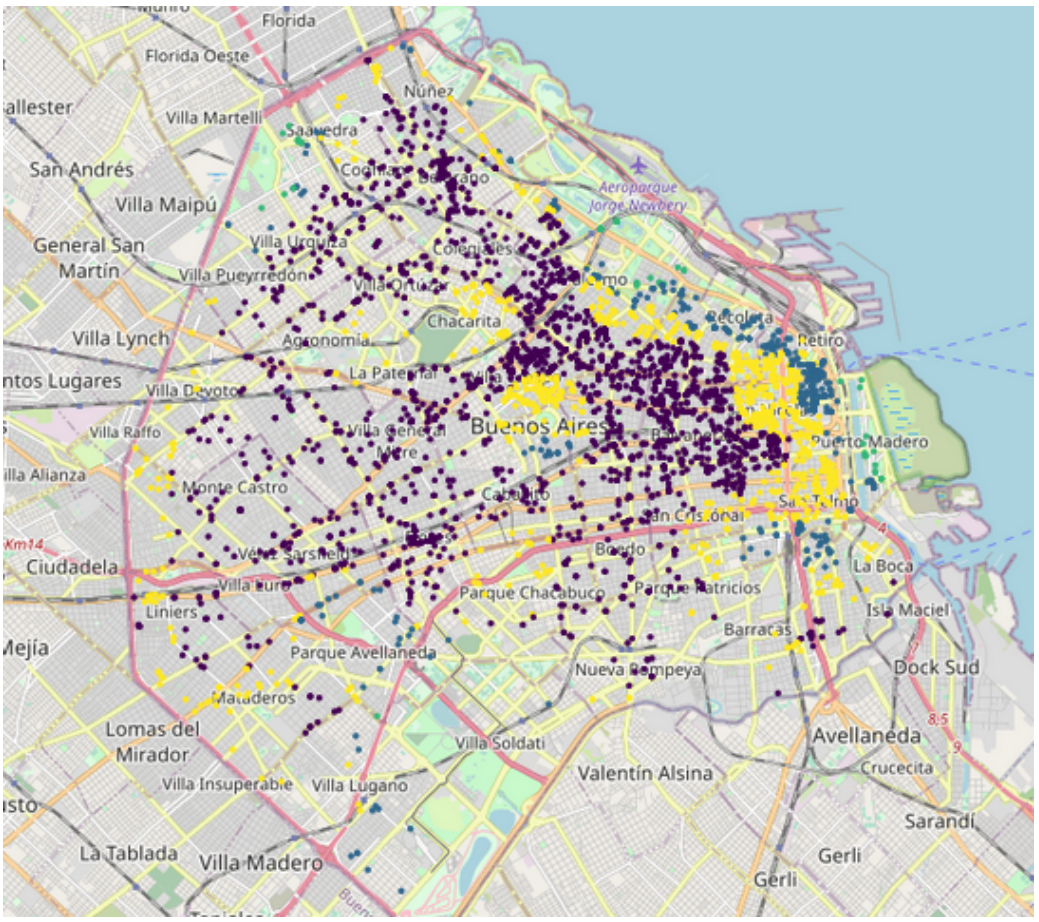
Obtuvimos un diagnóstico de los residuos para evaluar la normalidad, como se ve se distribuyen de una manera compatible con una distribución normal.

ANÁLISIS DE CLUSTERS

Se decidió abordar un análisis por cluster para identificar factores correctivos o no del modelo de regresión así como proporcionar una visión macro de la distribución de nuestra variable objetivo de análisis.

Generamos un modelo de cluster con k-medias que nos devuelva cuatros tipos de aglomerados, basándose en las características de los locales.





Se observan distintas aglomerados bastante diferenciados, uno (amarillo) que podría considerarse mas céntrico, otro (azul) que recorre desde la *city* hasta retiro y recoleta. En color verde se notan locales que están en zonas donde predomina el espacio verde (bosques de Palermo, reserva ecológica) y de color púrpura locales mas cercanos a zonas residenciales.

CONCLUSIONES FINALES

La distribución de locales, como por su cuantía (mayor o menor) era esperada que se encuentre coincidente responder a patrones geográficos y socioeconómicos dando de forma resultante cuatros grupos claramente diferenciados que corresponden a barrios como Barracas, Flores y Palermo entre otros, notablemente los parques y espacios verdes de Barracas tienen un aspecto histórico, y es claro que sus habitantes tienen interés en mantener espacios de estas calificaciones. Esto último a modo ejemplificativo.

No es de desconocer tampoco en esta conclusión que nuestro precio medio en CABA rondaría (usd 2976.22- 3106.32), pero quedarnos con este análisis tampoco es del todo acertado por las segmentaciones identificadas, lo cual recomendamos realizar un análisis por grupos en caso de encontrarnos con la necesidad de esta información más detallada, no se llevó a cabo el mismo al no responder al interrogante de estudio.

En esta etapa del informe hemos de concluir el mismo, destacando que el modelo ajusta en un 65,7% de los casos, y podemos decir que la existencia de observaciones que no aportan un grado de información coherente se puede deber a valores alzados/aminorados por factores de valuación según flujo de fondos futuros, dado que hablamos de locales comerciales, no obstante el mismo busca proyectar los valores de los mencionados en CABA y a pesar de las limitaciones de los factores nombrados provee de una aproximación con un buen nivel fidedigno.