Análisis del mercado Inmobiliaria para la Ciudad de Buenos Aires

Integrantes: Abal Ignacio, Deyá Federico y Matias Tomesek

Abstract - Este proyecto consta de un análisis del mercado inmobiliario en CABA para identificar la correlación entre las diferentes variables disponibles en el dataset y el precio de las propiedades, con el objetivo de estudiar la viabilidad del desarrollo de un modelo que sirva como cotizador de departamentos.

INTRODUCTION

Para el estudio se utilizó un dataset que contiene información sobre la oferta de departamentos en CABA durante el año 2013 y se implementaron herramientas de Análisis Exploratorio de Datos (EDA) y Machine Learning (ML).

El dataset utilizado se encuentra disponible en https://data.buenosaires.gob.ar/dataset/departamentos-venta perteneciente al Ministerio de Desarrollo Urbano y Transporte, Subsecretaría de Planeamiento,Dirección General de Diagnóstico territorial y Proyección Urbana.

DATASET

El dataset utilizado se encuentra disponible en https://data.buenosaires.gob.ar/dataset/departamentos-venta perteneciente al Ministerio de Desarrollo Urbano y Transporte, Subsecretaría de Planeamiento, Dirección General de Diagnóstico territorial y Proyección Urbana.

La base de datos está compuesta por unas 17.000 muestras aproximadamente, cada una de ellas, contiene el detalle de 19 características, algunas de ellas numéricas y otras descriptivas. A continuación procedemos a describir brevemente a cada una de ellas:

'CALLE': Calle en la que se encuentra ubicado el departamento.

'NÚMERO': Altura de la calle en la que se encuentra.

'M2': Superficie en metros cuadrados.

'DOLARES': Cotización en dolares. Esta será nuestra variable a predecir durante el desarrollo del modelo.

'U_S_ M2': Precio por metro cuadrado en dólares.

'PESOS': Cotización del departamento en pesos.

'PESOS M2': Precio por metro cuadrado en pesos.

'AMBIENTES': Cantidad de ambientes.

'ANTIGÜEDAD': Cantidad de años desde que fue construido el edificio.

'ORIENTACIÓN': Describe la orientación de la vista del departamento. Los valores que toma son: “Frente”, “Contrafrente”, “Lateral”, “Frente y Contrafrente” e “Interno”.

'BAULERA': Explica si el departamento cuenta o no con una baulera. Es una característica del tipo binaria.

'COCHERA': Explica si el departamento cuenta o no con cochera.

'BAÑOS': Cantidad de baños.

'LAVADERO': Explica si el departamento cuenta o no con lavadero.

'TERRAZA': Describe el tipo de “terraza” del departamento. Los valores que toma son: “Balcon”, “Terraza”, “Balcón Terraza” y “Patio”, variando cada uno no solo en tamaño sino en disposiciones.

'BARRIO': Nombre del barrio en el que se encuentra.

'COMUNA': Número de comuna en la que se encuentra.

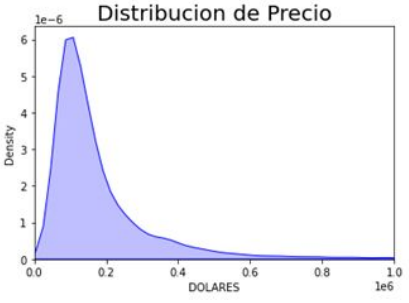
'LONGITUD': Longitud de la ubicación.

'LATITUD': Latitud de la ubicación.

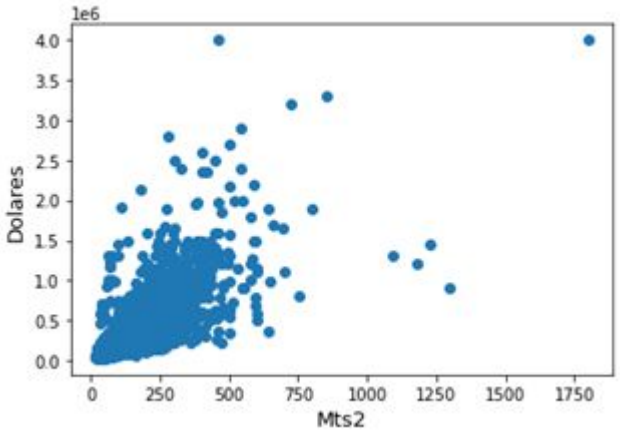
EDA

En esta etapa llevamos a cabo un análisis exploratorio de los datos para así poder comprenderlos con mayor profundidad y completar la limpieza efectuada.

En primer lugar, graficamos la distribución de los precios para obtener un primer pantallazo de la variable principal del estudio. Se puede apreciar que la media ronda los 120.000 USD y que posee una mayor dispersión hacia los precios mayores a la misma

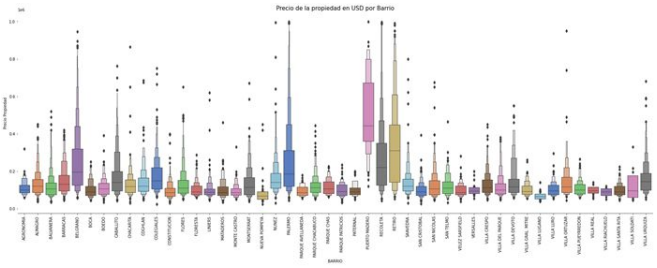


Continuando con el estudio de nuestra variable principal, graficamos la relación entre el valor en dólares y la superficie de las propiedades:



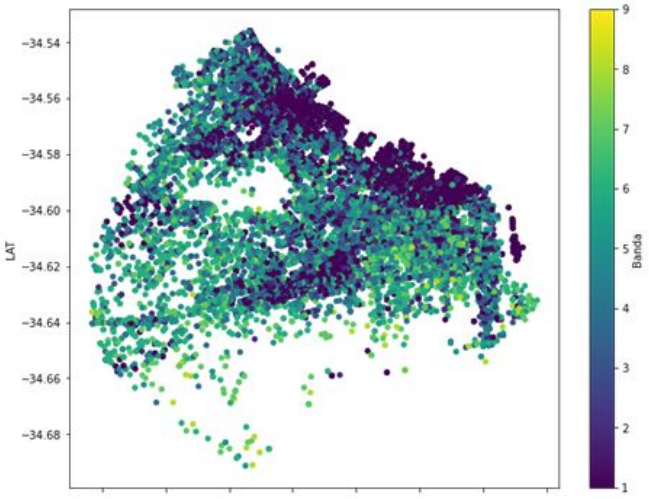
Si bien notamos una correlación entre ambas variables, como era previsible, los samples no se distribuyen cerca de alguna línea recta sino que hay una gran variabilidad en el precio que no parece estar explicada por la superficie y que abre la puerta al estudio de las demás características. También, observamos que había ciertos outliers que estaban ensuciando el análisis por lo que decidimos eliminarlos para poder tener mayor precisión. Se dejaron fuera del estudio las propiedades con precios superiores a 1 millón de dólares y superficies superiores a los 400 metros cuadrados, esto resultó en una reducción de 206 muestras (1% de la base de datos aproximadamente).

Por otra parte, nos pareció importante poder representar el valor de las propiedades según el barrio donde están ubicadas. Utilizamos un catplot, permitiéndonos ver cierta similitud entre algunas zonas, picos en la media, grandes variabilidades y otras muy reducidas.

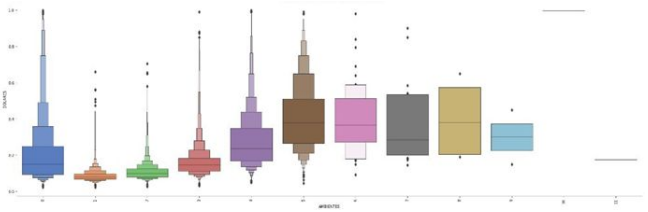


A simple vista, se puede notar que los barrios que cuentan con departamentos en venta de mayor valor son Puerto Madero, Retiro, Recoleta, Palermo y Belgrano, marcando una gran diferencia con respecto a los demás barrios. Para el caso particular de Puerto Madero, se puede ver que no cuenta, prácticamente, con propiedades que estén cerca de los valores de la media de los demás barrios y tiene la mediana más elevada. En cuanto a los barrios de precios más bajos, encontramos a Villa Lugano y Nueva Pompeya con las medianas más bajas. Se destaca también que coincide que estos barrios, además, son de los de menor variabilidad en sus precios.

Para expresar de una manera más visual esta información, ya que contamos con los valores de latitud y longitud de cada departamento, generamos un mapa de la ciudad en base a ellos. Creamos nueve bandas de precios con el objetivo de configurar la escala de colores del gráfico. Siendo la Banda 1, la de mayor precio (superior a 200.000 USD) y la Banda 9 la de menor precio (inferior a 25.000 USD). El resto de las bandas están divididas en intervalos iguales de 25.000 USD cada una.



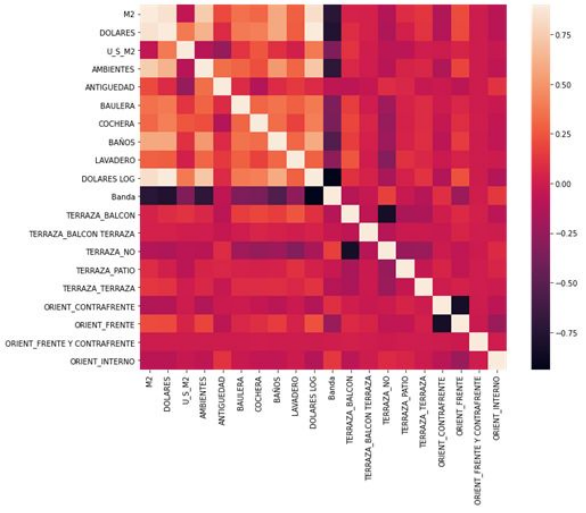
Otro enfoque que se le dio a este estudio fue el de realizar un catplot para ver la distribución de los precios discriminado por la cantidad de ambientes.



Vemos que hay departamentos de 10 o más ambientes, a los cuales los consideramos como outliers y otros de 0 ambientes, con una gran dispersión en sus precios. En las propiedades de 1, 2 y 3 ambientes vemos que las muestras están considerablemente concentradas cerca de su precio promedio, lo cual indica, por un lado, un alto impacto de la cantidad de ambientes en el precio y, por otro lado, que asignar una cantidad de ambientes para los que cuentan con 0 ambientes generaría mucho ruido en el dataset y empeoraría el desempeño del modelo de predicción. Es por eso que, en esta etapa, se dejaron afuera más de 1000 samples de nuestra base de datos, dejándola con un número apenas superior a las 15000 muestras.

Para finalizar el EDA, buscamos entender la correlación que hay entre todas las variables del dataset mediante un heatmap. Para esto tuvimos que convertir en numéricas las variables categóricas ('ORIENTACIÓN', 'TERRAZA', 'BARRIO') mediante la herramienta

get_dummies. De esta manera, se agregó una feature al dataset por cada uno de los valores que contenían estas variables y se asignó valor 1 cuando la muestra contase con ese valor en la feature original y 0 cuando no. Posteriormente, se eliminaron las variables categóricas mencionadas y se ejecutó el heatmap. Dejamos afuera de esta representación a las variables correspondientes a los barrios ya que complicaría la visualización de la información y ya fueron estudiadas en detalle anteriormente.



Lo que se ve en la imagen es una representación gráfica de la matriz de correlación de Pearson, la cual asigna un valor entre -1 y 1 para expresar el nivel de relación que mantienen 2 variables entre sí. Siendo que -1 indica una proporcionalidad inversa y 1 una proporcionalidad directa. Entonces, constatamos la alta correlación que tienen la cantidad de ambientes y metros cuadrados con el precio de la propiedad, pero además podemos observar que existe también una correlación moderada con la cantidad de baños y la disponibilidad de lavadero, cochera y baulera. Mientras tanto, observamos que la antigüedad no guarda correlación alguna con el precio, por lo que procedimos a descartarla para el desarrollo del modelo. También se puede observar que la alternativa más preciada para la orientación es hacia el frente y para la terraza es una terraza convencional, levemente por encima de un balcón terraza y con mayor diferencia por sobre un simple balcón o un patio.

A esta altura del estudio, ya se tiene una idea un poco más clara de las características generales del mercado inmobiliario y las interrelaciones que lo estructuran y el dataset se encuentra listo para ser preprocesado y entrar en el desarrollo de un modelo de predicción.

El objetivo era desarrollar un modelo de aprendizaje que sea capaz de cotizar un departamento en función de sus características (las variables analizadas anteriormente). Esto no es más que una predicción de la feature 'DÓLARES' a partir de las demás mediante una regresión. Procedimos, entonces, a probar distintos modelos de aprendizaje hasta encontrar el que mejor se adapte a nuestro caso de estudio.

Antes de ingresar el dataset a un modelo de aprendizaje, fue necesario llevar a cabo un preprocesamiento. De esta manera se equilibran las variaciones de las distintas features, de modo que el sistema no interprete, por ejemplo, que vale lo mismo 1 unidad de la variable ‘M2’ que de la variable ‘AMBIENTES’. La distribución de nuestras variables no tendía hacia una distribución gaussiana, por lo tanto, elegimos normalizar nuestras features mediante la herramienta MinMaxScaler. Esta herramienta transforma al dataset haciendo que todos sus valores queden entre 0 y 1, el cálculo que realiza para cada feature en cada sample es el siguiente:

$$X_{new} = \frac{X_i - \min(X)}{\max(x) - \min(X)}$$

A continuación probamos los siguientes modelos:

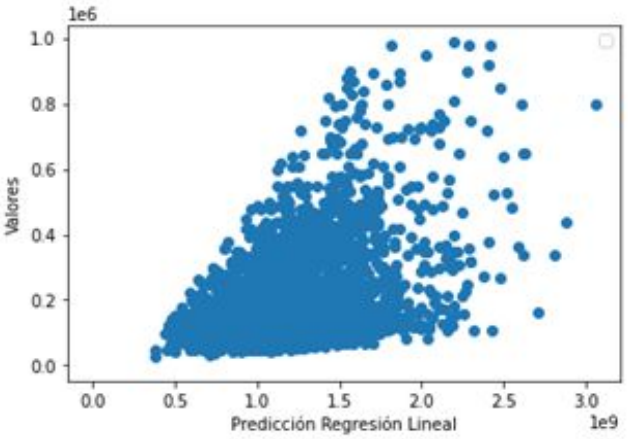
- Lineal Regressor: Modelo lineal de regresión.
- SVR (Support Vectors Machine): Estimador lineal basado en una serie de puntos centrales del dataset.
- KNN Regressor: Estimador no lineal que interpola entre los K samples más cercanos (neighbours).
- Ridge Regressor: Modelo lineal de regresión regularizada (disminuye sobreajuste al set de entrenamiento)

Los mismos fueron definidos mediante un Grid Search con Cross Validation. Evaluamos, luego, el rendimiento de cada uno en función de las siguientes métricas:

- R2: porción de la variabilidad que se encuentra explicada por el modelo. Valor menor a 1, siendo 1 su valor ideal.
- MSE: Error cuadrático medio. Se busca que sea lo menor posible.
- MAE: Error absoluto medio. Se busca que sea lo menor posible

	Model	R2	MSE	MAE
0	Regresión lineal	-8.979961e+07	1.411551e+18	1.146365e+09
1	SVR	-4.035431e-01	2.206215e+10	1.290024e+05
2	KNN	6.746505e-01	5.114136e+09	3.974467e+04
3	Ridge	-9.599846e+00	1.666179e+11	3.955118e+05

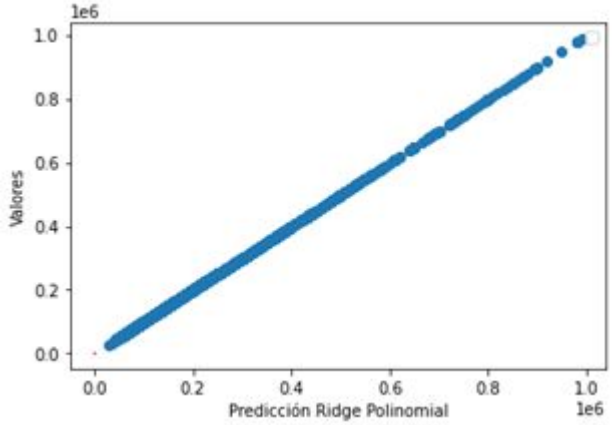
Como se puede observar, los resultados no fueron para nada favorables. A modo de ejemplo, en el siguiente gráfico, se puede visualizar la diferencia entre los valores de los precios a predecir con los devueltos por el Linear Regressor.



En consecuencia, se transformó la base de datos mediante la herramienta Polynomial Features. Como resultado, se obtuvieron nuevas features que son combinaciones lineales entre las originales o variables elevadas al cuadrado y se pudo buscar relaciones no lineales entre las distintas variables. El grado polinomial utilizado fue 2. Los resultados obtenidos para los nuevos estimadores fueron ampliamente superiores:

	Model	R2	MSE	MAE
0	Regresión lineal	-8.979961e+07	1.411551e+18	1.146365e+09
1	SVR	-4.035431e-01	2.206215e+10	1.290024e+05
2	KNN	6.746505e-01	5.114136e+09	3.974467e+04
3	Ridge	-9.599846e+00	1.666179e+11	3.955118e+05
4	RIDGE Poly	9.999996e-01	5.850210e+03	2.440957e+01
5	KNN Poly	6.746505e-01	5.114136e+09	3.974467e+04
6	RIDGE Poly	9.999997e-01	5.225982e+03	1.676600e+01

Los resultados mejoraron considerablemente. Encontramos que el Ridge Regression polinomial explica la variabilidad de nuestro dataset a un nivel muy alto, ya que arrojó un resultado muy próximo a 1 (0,9999997). Por lo tanto se le realizó un leve perfeccionamiento, buscando un valor aún más óptimo para su hiper parámetro 'alpha' (relacionado con el nivel de regularización del modelo), incrementando las alternativas en el grid search. Esta gráfica muestra cómo se apegan las predicciones del modelo a los valores del set de testeo, formando, prácticamente, una línea recta de pendiente igual a 1.



CONCLUSIONES

Pudimos comprender, a lo largo de este informe, las distintas variables que condicionan el precio final de una propiedad en la ciudad de Buenos Aires. Vimos como el precio se ve principalmente afectado por la superficie, la cantidad de ambientes y el barrio en el que se encuentra, pero constatamos, también, que hay otras variables que tienen un peso moderado en el mismo. De esta información se obtuvo un modelo que logra explicar la variabilidad de los datos con alta precisión y realiza predicciones acertadas. Esto significa que puede llegar a ser utilizado como cotizador de referencia, pero, de cualquier manera, cuenta con sus limitaciones dadas por la antigüedad de la información manipulada (recordamos que el dataset era de 2013) y por la dificultad de seguir el rastro la dinámica del mercado y de tener en consideración variables estéticas no cuantificables.

REFERENCIAS

- EDA: Tukey, J. W. (1977). Exploratory data analysis (Vol. 2, pp. 131-160).

http://theta.edu.pl/wp-content/uploads/2012/10/exploratorydataanalysis_tukey.pdf

- Supervised Learning Regression: Criminisi, A., Shotton, J., & Konukoglu, E. (2012). Decision forests: A unified framework for classification, regression, density estimation, manifold learning and semi-supervised learning. Foundations and Trends® in Computer Graphics and Vision, 7(2–3), 81-227.

https://dl.acm.org/doi/abs/10.1561/0600000035

- Polynomial Features: Florescu, D., & England, M. (2019). Algorithmically generating new algebraic features of polynomial systems for machine learning. arXiv preprint arXiv:1906.01455.

https://arxiv.org/abs/1906.01455

- Minmax scaler (normalization): Shaheen, H., Agarwal, S., & Ranjan, P. (2020). MinMaxScaler Binary PSO for Feature Selection. In First International Conference on Sustainable Technologies for Computational Intelligence (pp. 705-716). Springer, Singapore.

https://link.springer.com/chapter/10.1007/978-981-15-0029-9_55