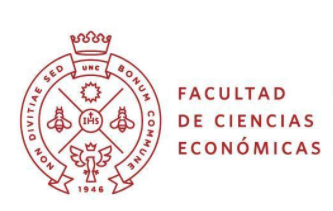
****

**UNIVERSIDAD NACIONAL DE CÓRDOBA**

**FACULTAD DE CIENCIAS ECONÓMICAS**

**Diplomatura en Ciencia de Datos, Inteligencia Artificial y sus aplicaciones en Economía y Negocios**

*“Predicción de precios de venta de locales comerciales de la ciudad de Buenos Aires”*

**COORDINADORA:** Bergamasco, Julieta

**AUTORES:** Bocco, Alessio

Bonvin, Mauricio Gabriel

Buffa, Agustin

Chehade, Santiago Gabriel

Diciembre 2021, Córdoba

INTRODUCCIÓN

En el siguiente informe se pretende predecir el precio de venta de locales comerciales en la Ciudad Autónoma de Buenos Aires (CABA) a partir de características propias de estos inmuebles (metros cuadrados, cantidad de ambientes, antiguedad, etc) como así también una serie de variables contextuales (relacionadas con su ubicación) construidas a partir de información georreferenciada (parques, atracciones turísticas, sitios gastronómicos, seguridad, etc.).

Se parte de un análisis exploratorio de los datos y variables para preparar la base para la aplicación de modelos predictivos utilizando los métodos de regresión lineal simple por mínimos cuadrados y XGBoost.

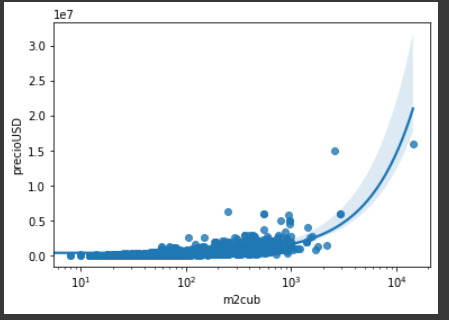
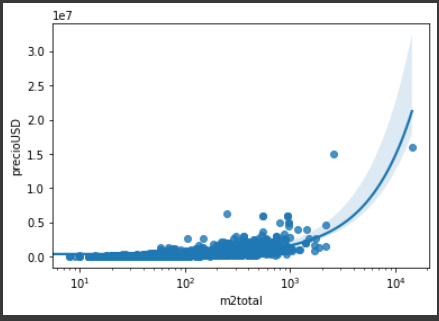
ANÁLISIS EXPLORATORIO Y LIMPIEZA DE DATOS

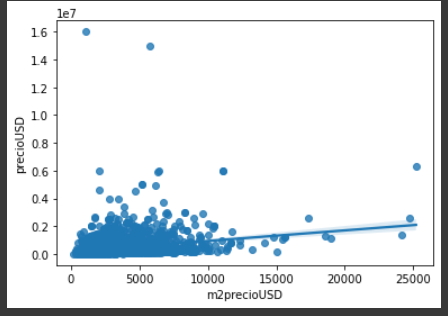
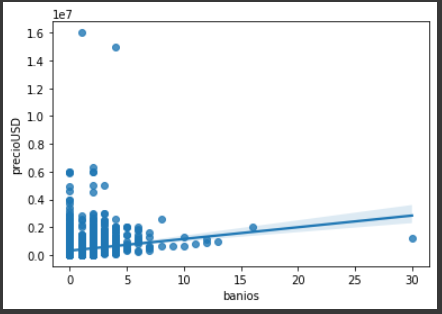
Se trabajó con un DataSet de 19 variables, las mismas no poseen datos faltantes excepto en la variable dirección la cual es corregida con los datos espaciales proporcionados. Para ello, se realizó una conversión a datos espaciales con las coordenadas planares obteniendo un GeoDataFrame. A través de la función get\_adress y utilizando las coordenadas, se corrigieron las direcciones faltantes. La corrección de comunas y barrios se realizó a través de la base de datos del gobierno de CABA.

Posteriormente, se realizó un análisis univariado de cada variable de la base, graficando histogramas y boxplot para observar el comportamiento de cada variable en forma independiente. Cabe destacar los siguientes hallazgos y correcciones realizadas a la base:

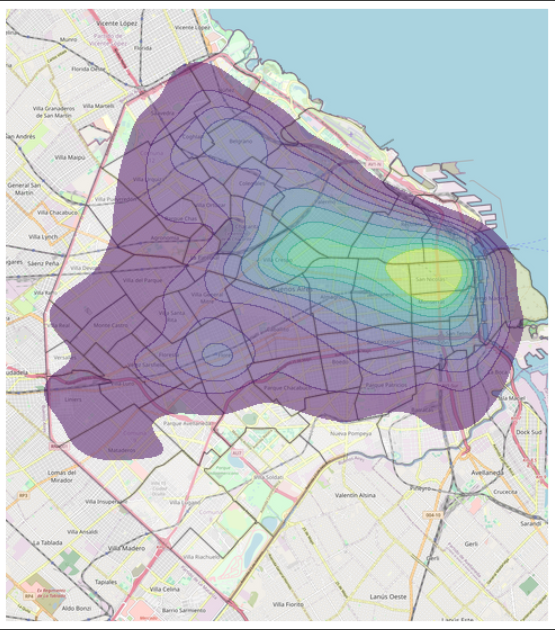
* La variable superficie total (m2total) presenta 41 outlier, considerando el percentil 0,99 del total de locales. Hemos ubicado y graficado los mismos en un mapa. Hemos decidido dejarlos fuera del análisis eliminando las 41 observaciones.
* Además, observamos valores de la variable “m2total” con 0m2 de superficie, valor que es ilógico. Los valores iguales a 0 de superficie total se corrigieron con los valores de superficie cubierta de la variable “m2cub”.
* En la variable cantidad de baños (banios) se identificó que la mayoría de las observaciones se encuentran entre 0 y 4 baños siendo bajo el porcentaje de locales con más de 4 baños por lo que decidimos agrupar la variable en las categorías 0, 1, 2, 3,y 4 o + baños.

Luego, se realizó un análisis sobre las variables numéricas. Primero se transformó el GeoDataFrame en un PandasDataFrame. Se realizó una correlación de las variables del DataFrame con respecto a la variable objetivo “precioUSD” identificando aquellas que tengan una correlación superior a 0,20 como las variables con mayor correlación en relación al precio en USD a decir, ”m2total”, “m2cub”, “m2precioUSD”, “banios”.





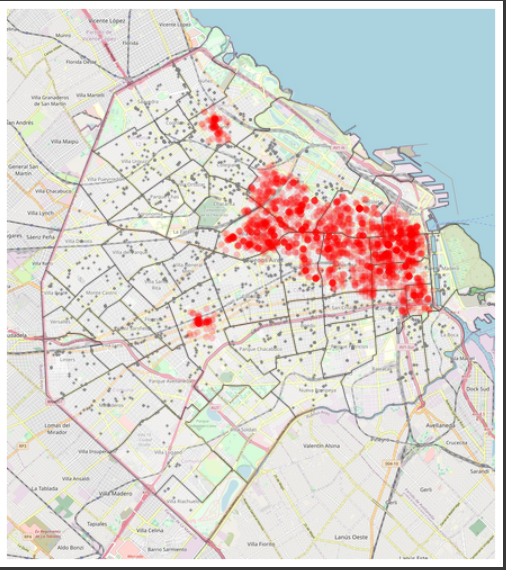
ANÁLISIS EXPLORATORIO ESPACIAL



Como parte del análisis exploratorio se graficó la distribución de los locales en el mapa mediante hexabin. Además, se utiliza la centrografía que consiste en el análisis de la centralidad de un patrón de puntos. Por centralidad se refiere a la ubicación general de la distribución de puntos y al patrón de dispersión. Si el hexabin podría considerarse como un histograma espacial, la centrografía sería el equivalente espacial a las medidas de tendencia central como la media o mediana y a las medidas de dispersión. Estas medidas son útiles ya que permiten resumir la distribución espacial de un único punto.

Al analizar la distribución de locales en el mapa se determinó mediante la función 𝐹, que los datos se encuentran agrupados.

Al confirmar la presencia de grupos de datos, éstos pueden identificarse mediante técnicas de clustering similares a las utilizadas durante la Diplomatura. Sin embargo, debe quedar claro que estos agrupamientos surgen de la posición de los locales en el espacio y nada tiene que ver con las variables discutidas en secciones anteriores.

Desde el punto de vista de la técnica usada, DBSCAN, un cluster es una concentración de al menos “n” puntos, cada uno dentro de una distancia “r” de al menos otro punto dentro del clúster. Bajo esta premisa, el algoritmo clasifica tres categorías:

* Ruido: puntos fuera del cluster.
* Núcleos: puntos dentro del cluster de al menos “m” puntos dentro de la distancia “r”.
* Bordes: puntos dentro del cluster con menos “m” puntos dentro de la distancia “r”.

Con esta configuración se identificaron 1656 núcleos comerciales (centroides) alrededor de los cuales se agrupan los demás locales. La configuración del algoritmo es bastante arbitraria y se necesitaría de un experto en el domino para definir mejor las distancias y cantidad de observaciones por grupo.

Los puntos en rojos pertenecen a un cluster mientras que los grises son observaciones aisladas que no pertenecen a ninguno. Se observa como las áreas comerciales más importantes se ubican en el noreste de la ciudad y en proximidades a avenidas importantes.

VARIABLES ADICIONALES

Se procedió a agregar variables adicionales al dataset de la siguiente manera:

* Variables no espaciales:
  + SobreAvenida: aquellas observaciones con valores 0 se encuentran sobre una avenida, con valores 1 no se encuentran en una avenida.
  + Aestrenar: adquieren valor 1 aquellos locales a estrenar y valor 0 aquellos con mayor antigüedad.
  + monoambiente: adquieren valor 1 aquellos locales con 1 solo ambiente y valor 0 aquellos con más ambientes.
  + clas\_ambientes: Reduce la categoría de ambientes a 0, 1, 2, 3, 4 y 5 o + ambientes.
  + clas\_banios: Reduce la categoría de banios a 0, 1, 2, 3 y 4 o + banios.
* variables espaciales: Para determinar las variables espaciales se utiliza la librería OSMnx que permite descargar información geoespacial desde OpenStreetMap (OSM).
  + count\_gastronomía: se crea una variable categórica que indica la cantidad de puntos de interés gastronómico (POI) dentro de un área de 500m (buffer) relacionado al ID de cada local. Los puntos de interés se toman de las categorías de OSM (Restaurantes, bares, fast foods, cafes, etc).

Lo mismo aplica para las variables count\_educación, count\_roads, count\_salud, count\_transporte,

SonCosteros: Para generar esta variable se obtuvieron datos del IGN y del gobierno de CABA. Se asigna valores 1 a los locales en barrios cercanos a la costa y valores 0 a los locales que no están en barrios cercanos de la costa.

zonas\_EAH: se crea la variable categórica ordenando los barrios alfabéticamente y se clasifica manualmente en zona norte, centro y sur.

* Mediante el uso de medidas resumen espaciales se puede determinar las interacciones entre los distintos locales comerciales. Para ello se obtienen las distancias entre locales bajo la premisa de que locales cercanos con características similares compiten entre sí por el mismo comprador potencial.
* Cardinality: Mediante el uso de matriz de pesos espaciales se determina la cantidad de vecinos. A mayor cantidad de observaciones cercanas se infiere que hay mayor cantidad de vecinos.
* Mediante el uso de la función lag\_categorical de la librería”libpysal”, que contempla cada valor dentro de la categoría y selecciona el valor más común entre vecinos y se determina una especie de moda espacial. Si por ejemplo, aplicamos esta función a la variable ambientes, se observa que el caso más común, un monoambiente está rodeado principalmente por otros locales que poseen solo un ambiente. En el otro extremo del espectro, los locales con más ambientes están más aislados y sólo rodeados por grandes superficies. Esto indicaría que los locales de las zonas más densamente "pobladas" tienden a ser de un sólo ambiente mientras que las grandes superficies están más aisladas y seguramente en la periferia de la ciudad. Se genera la variable “lagged\_ambientes”.

Se replica el método para crear las variables 'lagged\_banios', 'lagged\_antig', 'lagged\_m2total', 'lagged\_m2cub', 'lagged\_comisaria\_dista', 'lagged\_obelisco\_dista', 'lagged\_nrobos', ‘lagged\_sup\_espacio\_verde', 'lagged\_count\_gastronomia', 'lagged\_count\_educacion', 'lagged\_count\_roads', 'lagged\_count\_salud', 'lagged\_count\_transporte'.

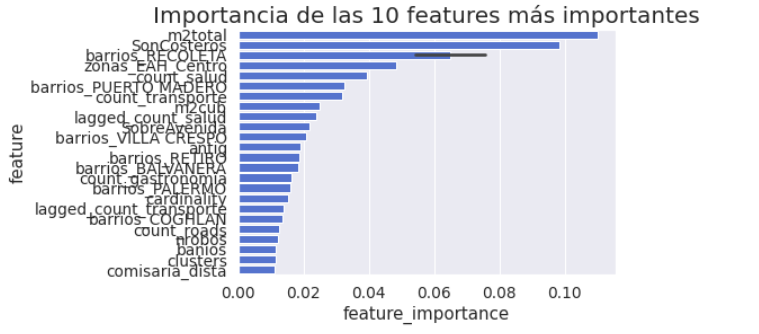
MODELOS PREDICTIVOS UTILIZADOS

* Modelo de regresión lineal:

La primera prueba corresponde a una regresión lineal simple por mínimos cuadrados sin considerar la relación espacial entre las observaciones. El modelo lineal explica poco más del 55% de la variabilidad presente en los datos. Con respecto a la significancia de las variables, no todas son significativas como la distancia a las comisarías, la superficie de espacios verdes, el acceso a avenidas, entre otros. Claramente la antigüedad y la superficie están muy ligados al precio de venta del local independientemente de la ubicación del mismo. Al trabajar con una relación log - lin, las anteriores variables se interpretan como: un aumento de la antigüedad del inmueble de un año disminuye su valor un 0.6% (100 \* beta\_antig), mientras que para la superficie, cada metro cuadrado extra aumenta el valor en un 0.3%.

Los residuos por barrio muestran que hay algunos, como Puerto Madero,cuyo precio es fuertemente subestimado por el modelo lineal.

* XGBoost:



Al igual que con el modelo lineal, la variable más importante fue la superficie ocupada por el local. También se suman otras, como si el barrio se encuentra sobre el río de la Plata o no y si el local se encuentra en el barrio de Recoleta. Por último, las comunas que el GCBA categoriza como zona centro en función de sus características socioeconómicas también tiene una gran influencia.

CONCLUSIÓN

Se probaron principalmente dos modelos: el modelo de regresión lineal y el modelo XGBoost teniendo en cuenta el error cuadrático medio (aproximadamente U$D 300.000). Ambos modelos arrojan un resultado similar. Dicha variabilidad sugiere que los modelos no son capaces de capturar la variabilidad por lo que se concluye que falta información para predecir los precios de los locales comerciales de la ciudad de Buenos Aires.

Al arrojar ambos modelos resultados similares se puede determinar que el uso del modelo XGBoost no se justifica en este caso por el tiempo de cómputo que conlleva.

Cabe destacar que la variable que mayor peso tiene para la predicción de los precios de los locales comerciales es para ambos modelos: la superficie del local. A mayor superficie mayor precio. Hay una dependencia espacial de los datos y es evidente que los barrios con mayor poder adquisitivo (Palermo, Puerto Maderos, Belgrano, etc.) cuentan con locales comerciales más caros.