Valor de Terrenos
CABA

Facundo Iorfida, Tomás Revah y Agustín Varetti

Universidad Tecnológica Nacional

CABA, Buenos Aires, Argentina

***Abstract*—En este informe analizaremos datos de Terrenos de la Ciudad de Buenos Aires, que se encuentran a la venta. En base a sus características buscaremos predecir el valor a partir de ellas y veremos también de que manera impacta el entorno en él.** (*Abstract*)

***Keywords* —** *terrenos, precio, usd, m2, barrio, comuna*

I. Introducción y objetivo

El objetivo del siguiente paper es poder analizar distintos factores que afectan al valor de un terreno en CABA, y, determinar la posibilidad de predecir en un precio desconocido en base a dichas características.

II. Descripción del dataset

A. Dataset base[1]

El dataset principal se compone a su vez de cinco datasets. Cada uno de estos se corresponde con el valor de oferta que tuvieron diferentes terrenos a lo largo del año.

Con la intención de contar con una gran cantidad de muestras, se utilizaron los precios dados en el período 2014-2018. Cada muestra, además del precio en dólares del terreno, contaba con la siguiente información :

- Longitud
- Latitud
- Fecha
- Calle
- Número
- M^2
- Precio usd/m^2
- Precio usd
- Precio ars
- Barrio
- Comuna
- Fot
- Cambio
- Dirección
- Código Postal
- Código Postal argentino

Finalmente, se concatenaron dichos datasets para obtener el principal.

B. Datasets secundarios

Para complementar y enriquecer el análisis, se recolectó información de entorno referenciada a cada barrio de la Ciudad.

Para dicha tarea se utilizaron los siguientes dataset :

- Establecimientos educativos
- Espacios verdes públicos
- Oferta gastronómica
- Espacios culturales
- Delitos
- Estaciones de bicicletas
- Bocas de acceso al subterráneo

De esta manera, relacionamos cada muestra de valor de terreno con su entorno inmediato. Así analizaremos que tanto impacta cada uno.

C. Dataset principal

Una vez cargado todos los datasets, se procedió a realizar tareas de preprocesamiento.

En primer lugar, se verificó la correcta carga de cada uno. Se continuó con el tratamiento individual de los valores faltantes (o NaN). Finalmente, se normalizó el contenido de la información y corrigió errores de carga, concluyendo con la conversión de tipos de datos a numérica. Esta es necesaria para el posterior análisis.

Como se mencionó, los datasets base fueron concatenados para formar el principal. Por otro lado, para los datasets adicionales se generó una tabla pivot donde fue volcada la cantidad correspondiente a cada tipo de característica según el barrio en cuestión.

Para concluir, la tabla pivot se utilizó para anexar dicha información a cada muestra según el barrio correspondiente.

III. Análisis exploratorio de datos

Para llevar adelante el objetivo de este informe, se realizó un relevamiento sobre el dataset armado.

En esta revision, se buscó analizar:

A. Distribución de las muestras.

Para analizar la composición del dataset en cuanto a la distribución de las muestras se tuvieron en cuenta dos aspectos.

[1] Fuente https://data.buenosaires.gob.ar/dataset/terrenos-valor-oferta

El aspecto geográfico. Para dicha tarea se confecciona un scatterplot.

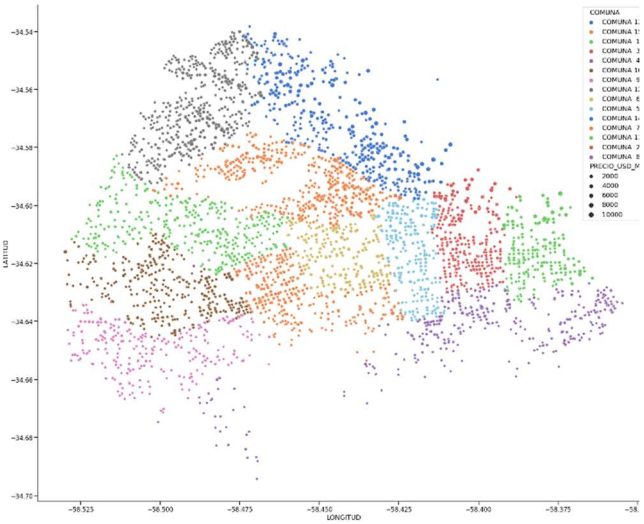

Fig. 1. Distribución geográfica de las muestras.

A partir de la figura, podemos determinar que el dataset generado presenta una distribución geográfica aceptable.

La cantidad de muestras por barrio. Para dicha tarea se confeccionó un counplot.

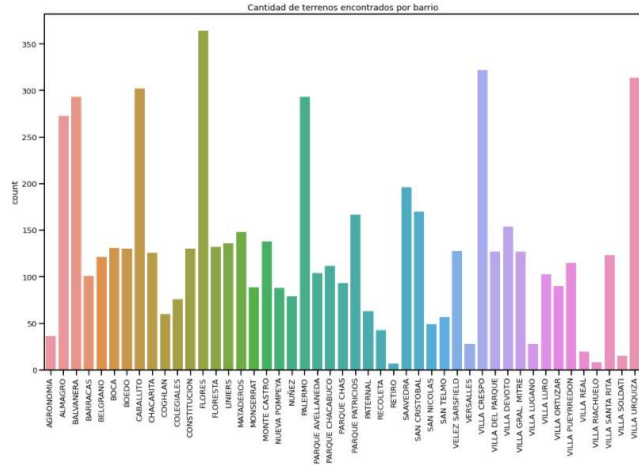

Fig. 2. Cantidad de muestras-terrenos por barrio.

A pesar de haber diferencias en las cantidades de muestras, el posible sesgo que puede tener lugar ante lo mencionado queda contrarrestado ante la generación de las features de entorno adicionales, dado que estas son independientes a la distribución en cantidad de muestras por barrio del dataset.

B. *Relación de las features con el precio del metro cuadrado en dólares*

Se estudió en detalle cada feature en relación con el precio del metro cuadrado en dólares.

1) En primer lugar, se confeccionó un gráfico de distribución del precio del metro cuadrado en dólares.

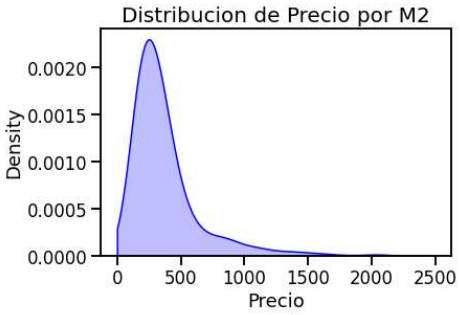

Fig. 3. Distribución del precio en dólares por metro cuadro

Se observa que la mayor cantidad de muestras-terrenos, poseen un valor de metro cuadro menor a 500 dólares.

2) En segundo lugar, generamos un catplot donde se relacione el precio del metro cuadrado en dólares con el barrio donde se ubica el terreno.

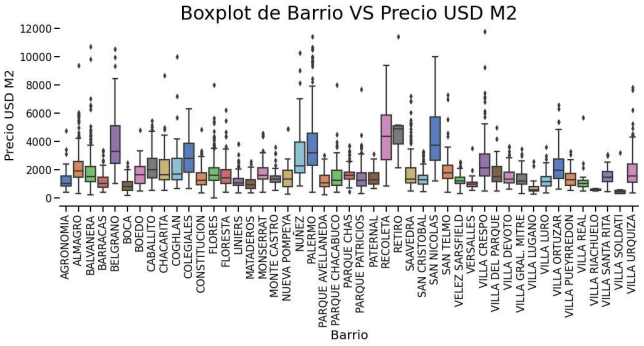

Fig. 4. Boxplot de Barrio vs Precio en dólares por m^2.

De este gráfico, podemos observar que a medida los valores de dólar por metro cuadrado son mayores, mayor es la variabilidad de los precios intra barrio. En lo contrario, para los barrios donde el valor del metro cuadrado es menor, su varianza es sustancialmente menor.

Finalmente, calculamos la correlación lineal entre cada una de las features. Para una mejor visualización, se gráfico con un heatmap donde se filtró por las diez features que más correlacionan.

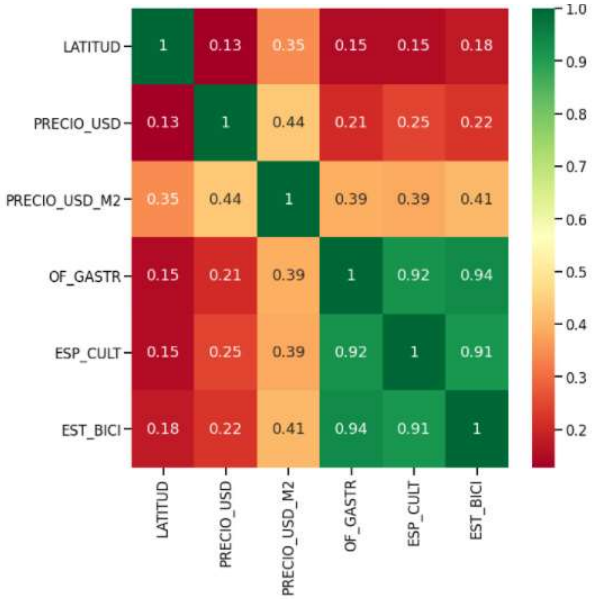

Fig. 5. Heatmap-Matriz de correlación.

Para destacar de esta matriz, es que observamos que el precio del m^2 correlaciona positivamente en 0.35 con la latitud. Esto establece que al subir la latitud (más al norte), sube el valor.

IV. MATERIALES Y MÉTODOS

A. Materiales

El análisis y procesamiento de la información se realizó a través del uso de la herramienta "Google Colab". Dentro de ella, se recurrió a diferentes librerías como numpy, matplotlib, y seaborn para el eda, mientras que para machine learning se utilizó scikit-learn (preprocessing, train_test_split, LinearRegression, Ridge, SVR, KNeighborsRegressor, r2_score, mean_squared_error, mean_absolute_error, GridSearchCV).

B. Métodos

Para completar un análisis integral, se utilizaron distintos modelos de regresión. Estos son :

- Linear Regression [lr]
- Ridge Regression [rr]
- Support Vector Regression [svr]
- K-Nearest Neighbor Regression [knn]

1) KNN Regression

Es algoritmo de regresión donde durante el entrenamiento se determinan los K vecinos más cercanos por distancia euclídea (distancia par a par).

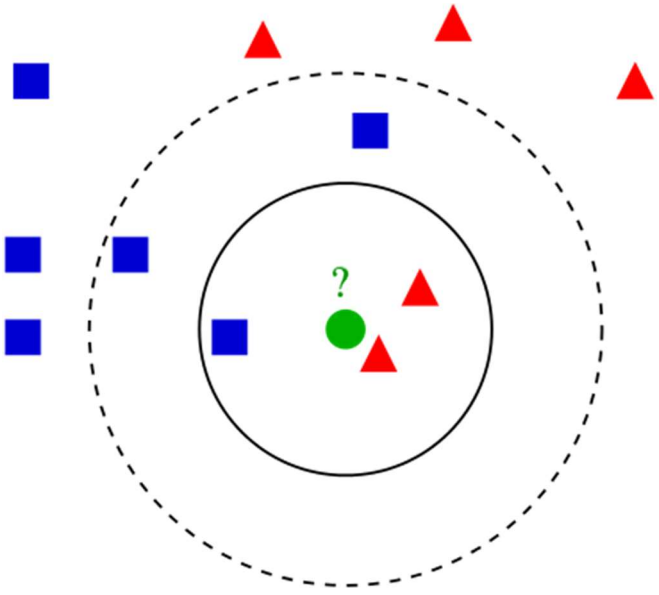

Fig. 6. K-Nearest Neighbors[2]

El y_i a predecir se determina por la interpolación de los y en los vecinos.

Los pesos w indican cómo se interpolara cada K vecino : uniforme o por distancia.

$$d(x_a, x_b) = \sqrt{(x_{a1} - x_{b1})^2 + (x_{a2} - x_{b2})^2 + ... + (x_{ap} - x_{bp})^2}$$

3) Métricas

Cada uno de estos fue entrenado y, comparando predicciones con las muestras de test, se les calculó tres tipos de error :

- MAE : mean absolute error.
- MSE : mean square error.
- RMSE : root mean square error.

4) Cross Validation

Para la elección de los hiper parámetros de cada modelo, se utilizó la técnica de cross validation.

Esta, se realiza con las muestras de entrenamiento y consiste en dividir los datos de forma aleatoria en k grupos de aproximadamente el mismo tamaño.

k-1 grupos se emplean para entrenar el modelo y uno de los grupos se emplea como validación. Este proceso se repite k veces utilizando un grupo distinto como validación en cada iteración. El proceso genera k estimaciones del error cuyo promedio se emplea como estimación final.

De esta manera, logramos preservar para cada modelo el hiper parámetro que menor error promedio de cross validation genere.

C. Grid search

En esta estrategia el proceso se automatiza realizando una búsqueda, no aleatoria en todo el espacio formado por las posibles combinaciones de los parámetros, sino en puntos de dicho espacio repartidos de forma espaciada en una rejilla.

Esta herramienta en combinación con la de cross validation nos brindará entre todas las combinaciones de hiper parámetros, la de mayor train accuracy resulte.

V. RESULTADOS

Una vez calculado los tres tipos de error para cada método de regresión, se confecciona la siguiente tabla :

TABLE I. COMPARACIÓN DE MODELOS DE REGRESIÓN

	Modelo	R2	MSE	RMSE	MAE
0	LR	0.349623	1.172242e+06	1082.701094	737.261813
1	RR	0.342877	1.184401e+06	1088.302043	737.025688
2	SVR	0.435441	1.017563e+06	1008.743219	623.955106
3	KNNR	0.522206	8.611771e+05	927.996285	547.958489

[2] https://www.unite.ai/what-is-k-nearest-neighbors/

Se le agregó a la tabla el valor de r2, coeficiente de determinación, el cual muestra la proporción de la varianza total de la variable explicada por la regresión y refleja la bondad del ajuste de un modelo a la variable que pretender explicar.

Finalmente, una vez comparados los resultados de entrenamiento se optó por el que presentaba mejores métricas con sus predicciones : **KNN Regression**.

VI. Conclusiones

Finalizado el análisis, procedemos a realizar unas conclusiones generales :

- En primer lugar, consideramos valioso que el gobierno, en este caso el de la Ciudad de Buenos Aires, ponga a disposición la numerosa cantidad de datasets que hoy en día ofrece en la página web : https://data.buenosaires.gob.ar/

- Luego, a pesar de lo mencionado anteriormente, observamos que muchos de los datasets ofrecidos no se encuentran preparados para la utilización de métodos de aprendizaje estadístico. En ocasiones, se detecta la falta de datos esenciales. También se observa cambios en los campos entre las mismas bases de datos, pero de diferentes años. Finalmente, algo que es usual en toda base de datos, es la carga errónea de valores. Todas las cuestiones nombradas dificultan el trabajo y requieren de un exhaustivo pre procesamiento de los datos previo al modelado.

- Es destacable la correlación lineal que presenta el valor del terreno (usd/m^2) con :
    - Estaciones de bicicletas
    - Oferta gastronómica
    - Espacios culturales

  Esto demuestra la incidencia positiva sobre el valor del terreno, cuando en su entorno geográfico existen espacios sociales y recreativos a donde la gente puede concurrir, en conjunto con la accesibilidad y movilidad brindada por las estaciones de bicicletas públicas.

  También, como ya se mencionó en el cuerpo del informe previamente, vale la pena volver a remarcar la diferencia económica, social y de infraestructura que se da entre el norte y el sur de esta.

- A pesar de haber seleccionado al modelo KNN Regression como el de mejor rendimiento para lograr la predicción de valor, consideramos que la precisión aún se puede mejorar, y el error por consiguiente reducirse. Para lograr este cometido, consideramos que sería oportuno el agregado de nuevas features que aporten información, como por el ejemplo el valor de FOT. Otra posibilidad sería la suma de nuevas muestras al entrenamiento y al testeo.

Reconocimiento

Este informe fue realizado en el marco de la materia Ciencia de Datos, en la facultad UTN Regional Buenos Aires, a través de la cátedra de Cluster AI.

Referencias

[1] "An Introduction to Statistical Learning" Gareth James, Daniela Witten, Trevor Hastie, Robert Tibshirani.

[2] "Python Data Science Handbook" Jake VanderPlas.

[3] "The Elements of Statistical Learning Data Mining, Inference, and Prediction" Trevor Hastie Robert Tibshirani, Jerome Friedman

[4] "Hands-On-Machine Learning with Sckikit-learn and TensorFlow" Aurelien Geron