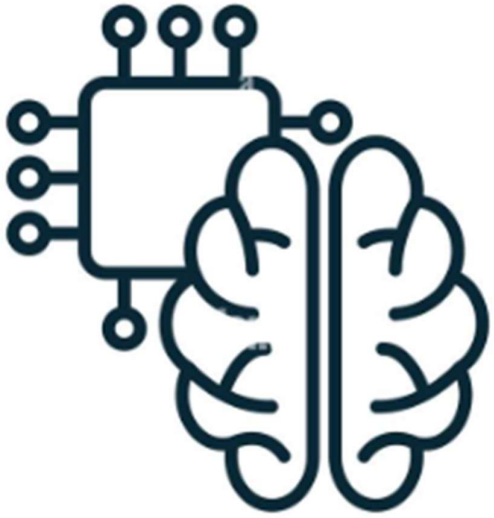
Informe Técnico:
“Machine Learning”

Aplicación de técnicas de Machine Learning para aprender los modelos que rigen el establecimiento del precio de las propiedades en Capital Federal, Argentina.



Autores:

- Tamagno, Giuliano
- Martino, Santiago

Fecha de publicación: 22/11/2022

Introducción y objetivos

En el presente informe se trabajará con un dataset con publicaciones de propiedades de Capital Federal.

El objetivo consiste en realizar un Modelo de Machine Learning para predecir el precio de las mismas de acuerdo a sus características.

Para lograrlo, se van a aplicar conceptos del ámbito de la Ciencia de Datos y Aprendizaje Estadístico, utilizando Python para procesar los algoritmos, a través de Jupyter Notebooks.

Descripción del dataset

Contamos con un dataset de 38.656 registros y 26 variables con distintas características de las propiedades publicadas. A continuación, se muestra el diccionario de ellas.

Diccionario dataset del Properati Regresion			
Nro	Variable	Significado	Dtype
1	Unnamed:	Indice del dataset	int64
2	id	número identificatorio de la publicación	object
3	ad_type	Tipo de publicacion	object
4	start_date	Fecha de inicio de la publicacion	object
5	end_date	Fecha de fin de la publicacion	object
6	created_on	Fecha creada de la publicacion	object
7	lat	Latitud de la propiedad	float64
8	lon	Longitud de la propiedad	float64
9	l1	Ubicación 1 de la propiedad	object
10	l2	Ubicación 2 de la propiedad	object
11	l3	Ubicación 3 de la propiedad	object
12	l4	Ubicación 4 de la propiedad	object
13	l5	Ubicación 5 de la propiedad	float64
14	l6	Ubicación 6 de la propiedad	float64
15	rooms	Cantidad de ambientes de la propiedad	float64
16	bedrooms	Cantidad de habitaciones de la propiedad	float64
17	bathrooms	Cantidad de banios de la propiedad	float64
18	surface_total	m2 total de la propiedad	float64
19	surface_covered	m2 cubiertos de la propiedad	float64
20	precio	Precio de la propiedad	float64
21	currency	Moneda del precio	object
22	price_period	periodo del precio	float64
23	title	Título de la publicación	object
24	description	Descripción de la publicación	object
25	property_type	Tipo de la propiedad	object
26	operation_type	Tipo de operación	object

Se puede observar que cuenta con 3 tipos distintos de datos int64, float64 y object.

Marcada en rojo se muestra la variable “precio” a predecir.

Análisis exploratorio de datos

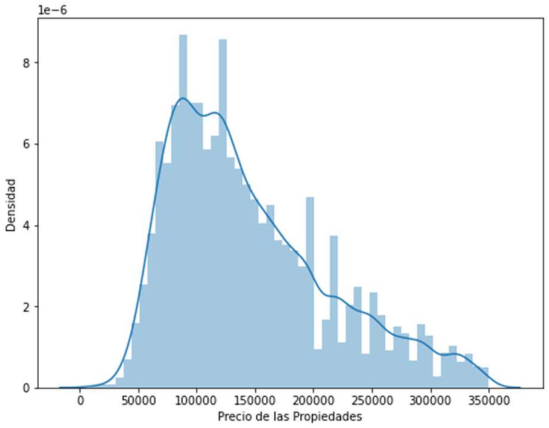
Resumen preprocesamiento

- Se eliminan las columnas **"l5", "l6" y "price_period"** ya que se encuentran vacías.
- Se eliminan las columnas **"id", "title", "description", "adtype", "l1", "l2", "currency", "operation_type"** debido a que se trata de texto de las publicaciones que no aporta información útil al modelo o todos sus registros se encuentran repetidos.
- Se rellenan los registros de las variables **"surface_total"** y **"surface_covered"** con su media y **"rooms"** con su mediana.

- Se eliminan 8248 registros que cuentan con el id de la publicacion **“Unnamed 0”** duplicado.
- Se eliminan outliers en función de la variable **“precio”**.

Resultados del Análisis

De acuerdo al objetivo final de este trabajo, vamos a iniciar el análisis entendiendo la distribución del conjunto de datos obtenido.



La mayor proporción de los datos se encuentran a partir de, aproximadamente, 30.000 USD y luego hay una gran concentración en los valores cercanos a 100.000 USD.

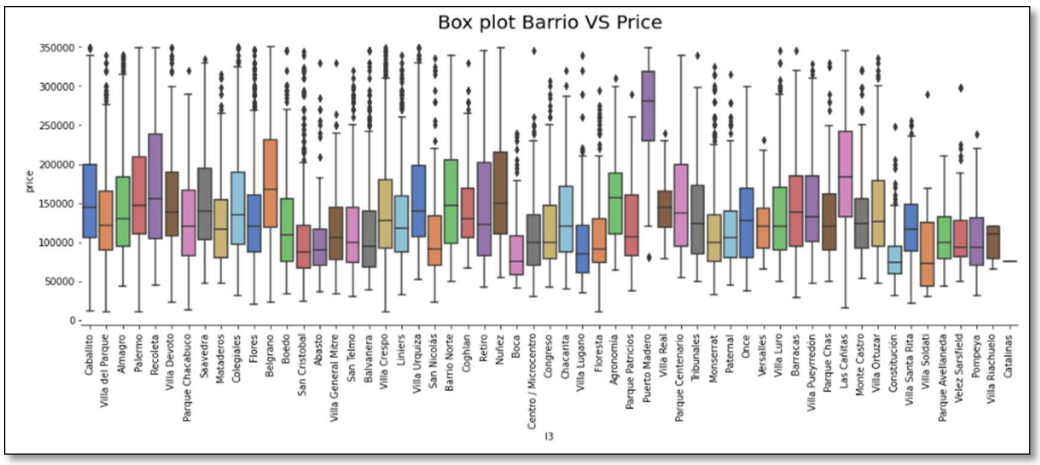
Ahora bien, es importante conocer qué variables de los datos tienen cierta correlación entre sí, ya que son las que más inciden a la hora de determinar el precio de las propiedades. Para esto, realizamos un mapa de correlación entre cada par de variables y descubrimos que solo la cantidad de *habitaciones* tiene relación lineal con el *precio*, con un valor de 0,6.

Otra información importante que tenemos que entender de los datos es la que refiere a lo estadístico.

Las variables realmente relevantes para este análisis son *rooms, surface y price*.

- Variable *rooms*: Del total de propiedades publicadas, la mediana de habitaciones con las que cuentan es de 2 (es lógico si estamos analizando departamentos de Capital Federal).
- Variable *surface*: En promedio, la superficie total es de 114 m^2 y la cubierta desciende a 88 m^2.
- Variable *price*: El precio medio de las propiedades es de 145 mil dólares, con un desvío estándar de 69 mil. No obstante, la mitad de las propiedades publicadas se encuentra por debajo de los 129 mil dólares, lo que evidencia que hay relativamente pocas propiedades con precio muy alto que eleva el valor de la media.

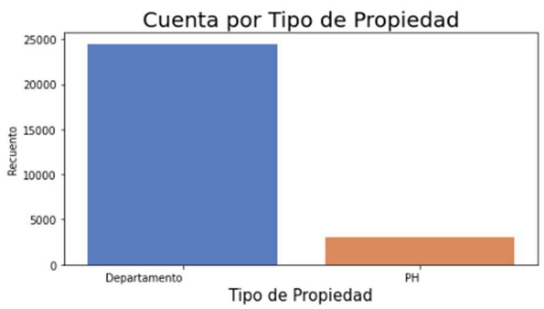
Realizamos un “BoxPlot” de la variable *precio* distinguiendo el barrio donde reside la propiedad publicada. El resultado fue el siguiente:

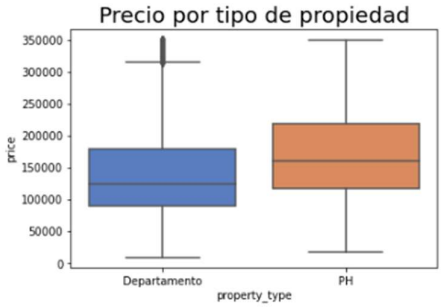


Del precedente gráfico podemos ver algunos datos relevantes.

- Precios más altos*:* El barrio con mayor media de precio es *Puerto Madero* seguido por *Las Cañitas*. También tiene altos precios *Recoleta*, pero para este último podemos ver que la variedad de precios es muy alta, teniendo una media relativamente normal.
- Precios más bajos*:* Como los barrios con precios más bajos, encontramos a *Constitución*, *Villa Soldati* y *Boca*

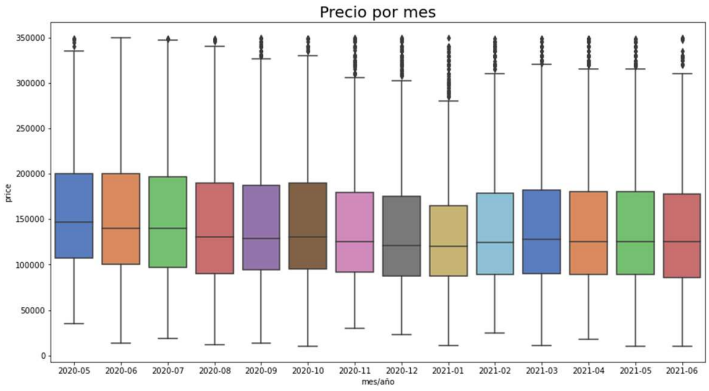
Pasamos ahora a realizar el mismo análisis, pero distinguiendo según el tipo de propiedad.





Podemos observar que la media del precio de los PentHouse es mayor a la de los departamentos, si bien hay departamentos que cuentan con valores de precio mayores.

Finalmente, utilizamos la fecha de publicación de la propiedad para construir dos nuevas columnas, referentes a la fecha de publicación. Gracias a su implementación, pudimos realizar un nuevo BoxPlot que nos diga cómo fue variando la media de los precios a lo largo del tiempo.



Resulta llamativo en el contexto argentino ver que la media de los precios vaya bajando, de acuerdo a la inflación que tenemos todos los años. Sin embargo, a causa de la pandemia, se vio una gran reducción en el precio de los alojamientos.

Materiales y métodos

Para la elaboración de las predicciones utilizaremos los siguientes algoritmos:

1. *Ridge Regression*: Modelo matemático que vamos a utilizar para realizar el aprendizaje de los datos de entrenamiento y la subsecuente evaluación.
2. *Support Vector Regression*: Modelo matemático para realizar el aprendizaje de los datos de entrenamiento y la subsecuente evaluación.
3. *Principal Components Analysis*: Algoritmo de reducción de la dimensionalidad para facilitar el análisis y aprendizaje de los modelos.
4. *GridSearch y CrossValidation*: Para la selección de los mejores hiper parámetros para los modelos y el cálculo de los pesos de la función en base a ellos.

La razón de usar dos modelos matemáticos distintos es contrastar la performance de cada uno de ellos para poder identificar el que tenga menor error de predicción para nuestro conjunto de datos.

Experimentos y resultados

Realizamos un *Pipeline* de aprendizaje supervisado sobre el conjunto de datos, iniciando por la preparación de los mismos para el correcto uso de los modelos, y terminando con la medición de la performance de cada uno de ellos.

Tomando el dataframe resultante del preprocesamiento, procedimos a generar las variables "dummies" de modo que las variables categóricas puedan ser utilizadas por los modelos matemáticos para la predicción.

Hecho esto, realizamos la separación del dataframe en conjuntos de entrenamiento y evaluación, en una proporción de 70/30, respectivamente, y la estandarización de los datos para que todos tengan la misma magnitud.

Support Vector Regression

Definimos los siguientes hiper parámetros posibles: *Kernel* (Lineal y sigmoide), *C* (100 y 1000) y *gamma* (0,01 y 0,1). Mediante Cross Validation se realizaron los aprendizajes y el que mejor performance tuvo fue el conjunto {kernel: lineal, C: 1000, gamma: 0,01} con una media de *score=0,4294*.

Luego se realizaron las predicciones con el modelo entrenado y se obtuvo la siguiente performance: *RMSE = 50.346,44.*

Ridge Regression

Primero se definió la función del modelo. Luego, establecimos el hiper parámetro *lambda = 30* y procedimos a calcular los pesos de la función, utilizados para las predicciones del modelo con el conjunto de evaluación.

Estas últimas dieron como resultado el siguiente indicador de performance: *RMSE = 49.277,70.*

Como se puede observar a simple vista, el modelo de regresión "Ridge" tuvo un resultado mejor.

PCA – Reducción de la dimensionalidad

Finalmente, realizamos el procedimiento para aplicar un algoritmo de reducción de la dimensionalidad, conocido como "Análisis de Componentes Principales", que consiste en obtener un espacio de menor dimensión que el original, manteniendo una variabilidad alta que permite un mejor aprendizaje de la información.

La reducción que propusimos para empezar con este análisis fue de pocas dimensiones, pasamos de un conjunto de 62 dimensiones a uno menor de 55 dimensiones, para ver si se lograba mejorar la performance de los modelos.

Una vez realizada la reducción, pasamos a repetir el aprendizaje de los modelos con el nuevo conjunto de variables. El resultado del modelo *Support Vector* fue similar, pero peor, obteniendo un *RMSE=50.361,206.* Sin embargo, el resultado del modelo *Ridge Regression* cambió mucho, aunque también para peor, obteniendo una nueva medida de error, *RMSE=153.654,21*.

Al ver que los modelos se desmejoraron luego de la reducción, decidimos finalizar este análisis.

Discusión y conclusiones

A continuación, vamos a realizar un pequeño análisis de los resultados obtenidos a lo largo de todo lo experimentado.

Al momento del Análisis Exploratorio, pudimos entender la variación de los precios de las propiedades de acuerdo a su ubicación geográfica y la fecha de publicación, lo cual se corresponde perfectamente con el contexto real que se vive en Capital Federal.

Finalmente, relacionado al aprendizaje supervisado, trabajamos con pocas variables numéricas y muchas columnas *dummies* que representaban a dos variables categóricas.

Los algoritmos fueron aplicados sin problemas, evidenciando la superioridad del modelo matemático *Ridge Regression*, que tuvo una performance superior al modelo *SVR* en un 2%.

A la hora de reducir la dimensionalidad del conjunto de datos, el problema fue que, al intentar reducir algunas dimensiones del conjunto, la performance de los modelos se redujo, lo cual hace que no sea útil, ya que estoy perdiendo calidad en la predicción de los mismos.

Referencias

Para el desarrollo del informe se consultó informacion de las siguientes fuentes:

- *Gareth James, Daniela Witten,Trevor Hastie ,Robert Tibshirani. "An Introduction to Statistical Learning with Applications in R"*
- *Christopher M. Bishop, "Pattern Recognition and Machine Learning"*
- *Wold, S., Esbensen K., & Geladi, P (1987). "Principal component analysis. Chemometrics and laboratory systems"*