Proyecto 1

Análisis del precio de las viviendas de Barcelona

Juan Pablo Delzo, Carlos Arregui, Alan Rocamora

Descripción breve

Desarrollo de un análisis completo y un modelo predictivo para los precios de viviendas en Barcelona, utilizando datos extraídos del portal Fotocasa y aplicando técnicas de extracción, manipulación y análisis de datos, así como algoritmos de Machine Learning, para predecir los precios de las viviendas en función de diversas características  
  
El código de Python utilizado para realizar este trabajo se encuentra en:  
https://github.com/carlosarreguib/PROYECTO\_1  
  
Archivos relevantes:  
- Barcelona\_Fotocasa\_HousingPrices.csv  
- Pre-processing.ipynb  
- MachineLearning.ipynb

# Análisis exploratorio de los datos

El análisis exploratorio realizado aporta una visión global de la situación del mercado inmobiliario de Barcelona. Los datos proporcionados contienen la siguiente información de un total de 8188 viviendas:

* Precio
* Número de habitaciones
* Número de baños
* Ascensor en finca
* Tipo de vivienda
* Barrio
* Precio por metro cuadrado

Para empezar, se grafica cuantos tipos de viviendas hay por cada barrio. Queda evidenciado que la vivienda modal en Barcelona es la del tipo “flat” en todos los barrios. Como el conjunto de datos no es demasiado extenso y el resto de los valores de la fila son de utilidad, se decide que los valores nulos de la columna se rellenan con este valor categórico.

Gráfico

Descripción generada automáticamente

Para medir y visualizar la relación entre varias variables del conjunto de datos, se grafica la matriz de correlación eliminando el precio de la vivienda.

Gráfico, Gráfico de barras

Descripción generada automáticamente

Se observa que el coeficiente de correlación entre square\_meters y bathroom es alto. Por tanto, se puede excluir bathroom en el diseño de modelos de predicción. Se debe de omitir el análisis a los atributos que provienen de boleanas.

En el análisis exploratorio se ha realizado un análisis por atributo. El objetivo es graficar a cada atributo respecto al precio.

## Análisis por atributo: neighborhood

El coeficiente de correlación entre barrio y precio es 0.09. De acuerdo con la correlación, se puede afirmar que no existe diferencia del precio de la vivienda por distrito.

Gráfico, Gráfico de barras

Descripción generada automáticamente

Gráfico, Gráfico de cajas y bigotes

Descripción generada automáticamente

## Análisis por atributo: real estate

El coeficiente de correlación entre el tipo de vivienda y el precio es -0.25. Se puede descartar que no existe una relación entre el precio y tipo de vivienda.

Gráfico, Gráfico de barras, Gráfico en cascada

Descripción generada automáticamente

Gráfico, Gráfico de cajas y bigotes

Descripción generada automáticamente

## Análisis por atributo: rooms

El coeficiente de correlación entre el número de habitaciones y el precio es de 0.35. No se puede descartar que no haya una relación entre el número de habitaciones y el precio, y los valores que prevalecen son 2 y 3 habitaciones.

Gráfico, Gráfico de barras

Descripción generada automáticamente

Gráfico, Gráfico de cajas y bigotes

Descripción generada automáticamente

## Análisis por atributo: bathroom

El coeficiente de correlación entre el número de baños y el precio de la vivienda es de 0.58. Es un coeficiente mayor que el de las habitaciones, con la moda del atributo en 1 baño

Gráfico

Descripción generada automáticamente con confianza media

Gráfico, Gráfico de cajas y bigotes

Descripción generada automáticamente

## Análisis por atributo: lift

El coeficiente de correlación entre el ascensor y el precio de la vivienda es 0.06. Se comprueba que la mayoría de las fincas en Barcelona disponen de ascensor, aunque no afecte en casi ninguna medida al precio.

Gráfico

Descripción generada automáticamente

Interfaz de usuario gráfica, Aplicación, Teams

Descripción generada automáticamente

## Análisis por atributo: terrace

El coeficiente de correlación entre la terraza y el precio de la vivienda es 0.17. La mayoría de pisos no disponen de terraza.

Gráfico, Gráfico de embudo

Descripción generada automáticamente

Interfaz de usuario gráfica, Aplicación

Descripción generada automáticamente

## Análisis por atributo: square meters

El coeficiente de correlación entre los metros cuadrados del piso y el precio es 0.69. Hay una alta correlación y la distribución se ajusta a una campana de Gauss.

Interfaz de usuario gráfica, Aplicación

Descripción generada automáticamente con confianza media

Gráfico, Gráfico de dispersión

Descripción generada automáticamente

## Análisis por atributo: square meters price

El coeficiente de correlación entre el precio por metro cuadrado y el precio de la vivienda es 0.58, menor que respecto a los metros cuadrados. El histograma es claramente asimétrico desplazado hacia los precios bajos.

Imagen que contiene Aplicación

Descripción generada automáticamente

Gráfico, Gráfico de dispersión

Descripción generada automáticamente

## Eliminación de atípicos

Se detectan varios atípicos, que se eliminan del dataframe original. Las filas seleccionadas son: 6951, 1772, 2427, 2754, 4220, 4750, 7646, 7928.

# Modelos de predicción

La distribución de los precios de la vivienda tiene un sesgo positivo muy evidente. Se aplica una transformación logarítmica para eliminar el sesgo y aproximar más la distribución a una distribución normal.

Gráfico, Histograma

Descripción generada automáticamente

Se aplican los modelos de predicción a ambas distribuciones para comparar si algún modelo se ve mejorado. Los modelos aplicados para comparar los errores y ver cuál es el que predice mejores resultados han sido los siguientes:

* Regresión lineal simple
* Lasso
* Ridge
* ElasticNet
* K-nearest neighbor
* Random Forest
* Gradient Boosting

En todos los ajustes se ha utilizado una fracción del 80% de los datos para entrenar el modelo y un 20% para testear.

A continuación, se presenta una tabla resumen de los resultados de los ajustes de todos los modelos para las dos distribuciones mencionadas más arriba.

|  |  |  |
| --- | --- | --- |
| **Modelo** | **Error distribución original** | **Error distribución con transformación log.** |
| *Regresión lineal simple* | 165.17 € | 256.03 € |
| *Lasso* | 164.90 € | 612.21 € |
| *Ridge* | 223.51 € | 256.00 € |
| *ElasticNet* | 165.15 € | 612.21 € |
| *K-nearest neighbor* | 68.63 € | 70.52 € |
| *Random Forest* | 22.27 € | 24.11 € |
| *Gradient Boosting* | 22.62 € | 20.53 € |

Queda evidente que la transformación logarítmica únicamente ayuda en el modelo Gradient Boosting, donde el error se disminuye en 2.09 € respecto a la distribución original.

Para enseñar el ajuste de todos los modelos, se presentan todos los gráficos de todos los modelos para la distribución original y para la distribución con transformación logarítmica.

## Regresión lineal simple

Gráfico, Gráfico de dispersión

Descripción generada automáticamente Gráfico, Gráfico de dispersión

Descripción generada automáticamente

## Lasso

Gráfico, Gráfico de dispersión

Descripción generada automáticamente Gráfico, Gráfico de líneas

Descripción generada automáticamente

## Ridge

Gráfico, Gráfico de dispersión

Descripción generada automáticamente Gráfico, Gráfico de dispersión

Descripción generada automáticamente

## ElasticNet

Gráfico, Gráfico de dispersión

Descripción generada automáticamente Gráfico, Gráfico de líneas

Descripción generada automáticamente

## K-nearest neighbor

Gráfico, Gráfico de dispersión

Descripción generada automáticamente Gráfico, Gráfico de dispersión

Descripción generada automáticamente

## Random Forest

Gráfico, Gráfico de dispersión

Descripción generada automáticamente Gráfico, Gráfico de dispersión

Descripción generada automáticamente

## Gradient Boosting

Gráfico, Gráfico de dispersión

Descripción generada automáticamente Gráfico, Gráfico de dispersión

Descripción generada automáticamente