IT ACADEMY

Barcelona Activa · Ajuntament de Barcelona

Data Science – Proyecto final

Scraping de un portal inmobiliario

Isabelle Pinot

2 de mayo de 2023

Abstract

Como en muchos otros sectores de la economía, el acceso al big data se ha convertido en una fuente muy valiosa de información para todos los agentes del sector inmobiliario . Sus aplicaciones son múltiples, para profesionales que quieren anticipar tendencias de mercado y detectar nuevas oportunidades de negocio, como usuarios finales que pretenden sacar el mayor partido a una vivienda en propiedad, o simplemente optimizar el proceso de búsqueda de una vivienda que cumpla con sus expectativas.

Acceder a esta información no es tarea fácil, y si bien abundan los informes y los análisis publicados diariamente por administraciones y medios de comunicación, suele ser una información basada en datos históricos que no cubre la demanda de datos en tiempo real característica de los usuarios de hoy en día.

En este contexto se están multiplicando las empresas de consultaría y servicios varios que pretenden hacer negocio con estos datos aprovechando la conyectura. En particular portales inmobiliarios que están creando nuevas divisiones orientadas a la creación de informes y tasación de propiedades, explotando los datos facilitados gratuitamente por sus anunciantes.

A través de este proyecto nos planteamos acceder a la web de uno de los principales portales inmobiliarios de España, descargar la máxima cantidad de información disponible respecto a las viviendas en venta y en alquiler en un momento dado, y generar unos datasets que se puedan explotar posteriormente con fines analíticos y/o predictivos.

1. Introducción

Las fases de nuestro proyecto se inscriben dentro del marco general de pautas a seguir para llevar a cabo un proyecto de Data Science, a saber :

- **Definición del problema**

Una vez enfocada la problemática general en el apartado anterior, nos planteamos acotar el proyecto a la ciudad de Barcelona, con tal de que la información se pueda procesar en un tiempo razonable.

- **Adquisición de datos**

Las posibles fuentes de datos son los portales inmobiliarios con más tráfico en España, por orden Idealista, Fotocasa y Habitaclia con una cantidad de visitas, respectivamente, de 62,1millones, 13,5millones y 9 millones en el mes de abril 2023. [1]

El acceso a la información de los portales se puede hacer a través de las APIs puestas a disposición de los usuarios, si las hay, o de software de web scraping siempre y cuando el portal nos deje acceder a su contenido con este tipo de herramientas.

- **Analísis de datos**

Nuestro objetivo es recuperar la información y construir los datasets con la ayuda de Pandas. Esta biblioteca de software de código abierto está diseñada específicamente para la manipulación y el análisis de datos en el lenguaje Python. Posteriormente, nos proponemos explorar y analizar los datasets, siempre con la ayuda de Pandas, para convertirlos en una fuente limpia y lista para ser explotada por algoritmos de machine learning.

- **Modelamiento de datos**

Con tal de comprobar su posible explotación, nos proponemos entrenar unos algoritmos de machine learning, en este caso de regresión, facilitados por la librería Scikit Learn, y valorar en que medida los resultados obtenidos son lo suficientemente precisos para ser usados como modelos de estimación de precio objetivo de venta y de alquiler de una vivienda en Barcelona.

2. Metodología

- **Web scraping con Selenium**

En un primer momento nos planteamos acceder al portal de Idealista al ser el más importante de lejos en número de visitas. Idealista propone al usuario usar su API, (Application Programming Interface), en la cual uno se puede registrar via solicitud directa a la empresa, aportando la explicación del proyecto para el cual se requiere este acceso. Mediante esta API el usuario puede descargar un máximo de 20 anuncios diarios…

Idealista dispone ahora de una división idealista/data cuyo objetivo es “poner al alcance del usuario información inmobiliaria estructurada, ordenada a nivel de un ámbito concreto homogéneo inmobiliariamente y en tiempo real’. A modo de ejemplo, el precio del informe de estimación de precio de un solo inmueble para particulares es 19,90€…

En consecuencia, nos planteamos usar un software de web scraping, en este caso Selenium , una herramienta que permite interactuar de forma elaborada con los websites, entre otras utilidades clicar links y botones, opciones imprescindibles para scrapear portales inmobiliaros.

Idealista rechaza el acceso a su web con Selenium, al detectar que su software está controlando el navegador.

El portal a scrapear será finalmente el de Habitaclia, www.habitaclia.com portal fundado en Mataró en 2005, con muy bien posicionamiento en la ciudad de Barcelona.

Creamos una función de referencia que nos permite acceder al url escogido, y desde allí, a los elementos del código html correspondiente a cada anuncio. Recogemos la información en una lista y desde allí convertimos la información a dataframe.

Dado el volumen de información, descartamos la opción de scrapear la información contenida en la página individual de cada anuncio, y nos limitamos a recoger la información contenida en el listado de los resultados de las búsquedas con la foto principal del inmueble y enlace a la página individual.

El código correspondiente esrá recogido en el notebook de Jupyter
web_scraping_selenium.ipynb

- **Creación, análisis exploratorio y impieza de los datasets**

El volumen de información no permite scrapear la totalidad de las viviendas anunciadas en una sola request, así que lanzamos una solicitud por distrito y por tipo de transacción (venta/alquiler). Fusionamos los 10 datasets obtenidos en uno para cada tipo de transacción.

Obtenemos dos datasets de 4486 registros para alquiler y 18753 registros para compra que se habrán reducido a 3725 y 12258 respectivamente después del proceso de limpieza de datos.

El proceso de limpieza consiste principalmente en eliminar registros duplicados (anuncios repetidos, viviendas anunciadas por varias inmobiliarias,etc…), eliminar datos nulos o erróneos (anuncios con datos erróneos de superficie, dormitorios, baños, etc…) e incorporar nuevas variables en base a la información recogida en columnas de texto como el título del anuncio o la descripción de la vivienda, para construir variables que no se han podido scrapear como tal (número de baño, temporalidad del alquiler, cambio de precio, etc…)

El código correspondiente está recogido en los notebooks de Jupyter
exploración_datasets_alquiler.ipynb y exploración_datasets_comprar.ipynb

- **Modelos de machine learning con Scikit Learn**

Construimos cuatro modelos de regresión con los algoritmos KNN, regresión lineal Ridge y Lasso, Random Forest y Gradient Boosting Tree.

Previamente al entrenamiento realizamos un preprocesado de las variables. Empleamos RobustScaler en el caso de las variables numéricas ya que presentan todas una cantidad importante de outliers, y OneHotEncoder para las variables categóricas que no son de tipo binarias.

Guardamos el preprocesamiento y el modelo en un pipeline para su posterior explotación.

Después de buscar los mejores parámetros con el método gridsearch CV pasamos a determinar el grado de error de cada modelo.Las métricas empleadas son r2, RMSE y MAPE.

El código correspondiente está recogido en los notebooks de Jupyter modelos_machine_learning_comprar.ipynb y modelos_machine_learning_alquiler.ipynb

3. Resultados

Resumen de un dataset con todas las variables obtenidas con el web scraping :

Dataset final

```
df_Alquiler.info()
```

```
<class 'pandas.core.frame.DataFrame'>
Int64Index: 3814 entries, 0 to 3813
Data columns (total 21 columns):
 #   Column          Non-Null Count  Dtype 
---  ------          --------------  ----- 
 0   id              3814 non-null   object
 1   url             3814 non-null   object
 2   subtype         3814 non-null   object
 3   transaction     3814 non-null   object
 4   owner           3814 non-null   object
 5   sell-type       3814 non-null   object
 6   provider        3814 non-null   object
 7   title           3814 non-null   object
 8   location        3814 non-null   object
 9   features        3814 non-null   object
 10  description     3814 non-null   object
 11  price           3814 non-null   object
 12  price2          3814 non-null   int64 
 13  m2              3814 non-null   int64 
 14  rooms           3814 non-null   int64 
 15  amueblado       3814 non-null   object
 16  parking         3814 non-null   object
 17  distrito        3814 non-null   object
 18  temporalidad    3814 non-null   object
 19  cambio_precio   3814 non-null   object
 20  baños           3814 non-null   int64 
dtypes: int64(4), object(17)
memory usage: 655.5+ KB
```

Resultados comparativos de los errores de modelos de regresión :

Viviendas a la venta :

	KNN	Linear Model Ridge	Random Forest	XGBoost
r2	0.71	0.70	0.73	0.75
rmse	372038.05	378894.26	356920.79	347896.98
mape	0.31	0.37	0.29	0.29

Viviendas en alquiler :

	KNN	Linear Model Ridge	Random Forest	XGBoost
r2	0.70	0.62	0.70	0.68
rmse	1072.56	1202.49	1065.53	1094.23
mape	0.26	0.29	0.26	0.26

R2, el coeficiente de determinación, es una medida de la precisión general de un modelo de regresión, evalua lo bien que el modelo se ajusta a los datos reales. Es la proporción de la varianza total de la variable explicada por la regresión.

RMSE, la raíz cuadrada del error cuadrático medio, básicamente mide el error promedio de nuestras predicciones. Para cada punto, calcula la diferencia entre las predicciones y el objetivo y luego promedia esos valores. Cuanto mayor sea este valor, peor es el modelo.

MAPE es el error de porcentaje medio absoluto, que es una medida relativa del error de previsión media del modelo en comparación con los datos reales.

Los resultados obtenidos no varían mucho de un modelo al otro, si nos quedamos con la métrica MAPE, podemos concluir que en el caso de las viviendas a la venta, el error medio de predicción del precio es del 29% en el mejor de los casos, y en el caso de las viviendas en alquiler del 26%.

4. Conclusión

Hemos conseguido acceder con éxito al portal y descargar la información de nuestro interés. Sin lugar a duda, la recogida de estos datos y su almacenamiento de forma estructurada nos ofrece la posibilidad de analizar datos del mercado en profundidad y optimizar la búsqueda de viviendas de forma ordenada, sea cual sea el objetivo del usuario final.

Las métricas de evaluación de los resultados facilitados por los modelos de machine learning, que son muy similares de un modelo a otro, ponen de relieve que es necesario profundizar en este aspecto y hacernos las siguientes preguntas :

Es el sector inmobiliario un mercado guiado por unos criterios cartesianos , que nos permite afirmar que es posible prever el precio al que un anuanciante pondrá una propiedad en alquiler o a la venta? Es posible calcular el precio objetivo de una vivienda en base a los datos del mercado? Son las variables escogidas suficientes como para construir un modelo preciso?

5. Referencias:

[1] datos obtenidos de Similarweb https://www.similarweb.com/

Selenium Software https://www.selenium.dev/

Idealista https://www.idealista.com/

Idealista/data https://www.idealista.com/data/

Habitaclia https://www.habitaclia.com/