API de Predicción de Precios del Sector Inmobiliario, basados en Modelos de Aprendizaje Supervisado

"Una imagen vale más que mil palabras y un número más que 10.000"

Índice

Introducción

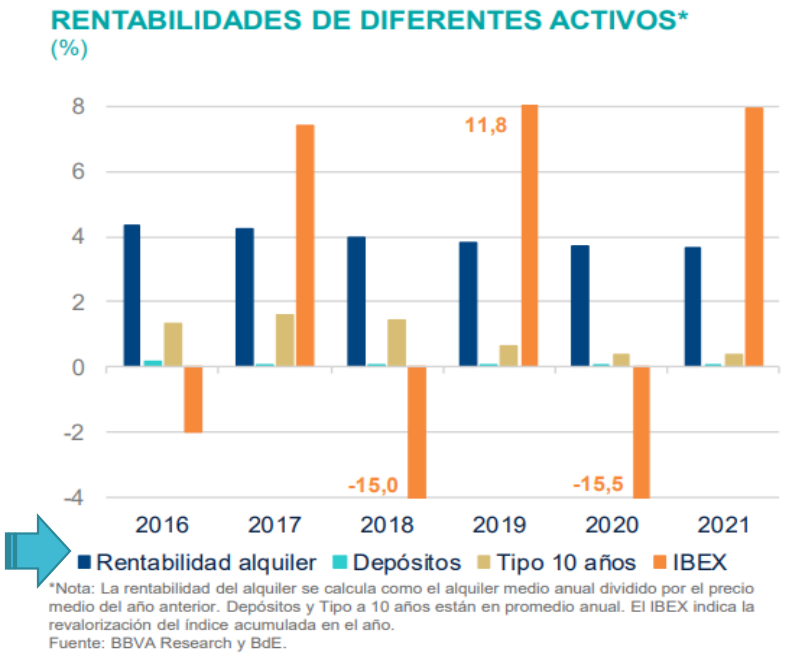
El sector de la construcción representa el 6.2% del PIB y emplea a más de 1,4 millones de personas

A 03/22 había 167.100 personas trabajando en el Sector Servicios de actividades inmobiliarias.

En 2021, se realizaron 678.000 operaciones de compra venta de inmuebles ante notario, según (BdE)

El esfuerzo que realizan las familias para la adquisición de una vivienda son 7 años en promedio.

Los activos inmobiliarios son los que producen un **retorno de la inversión más estable y consistente en el tiempo,** con una media del 4 por ciento anual en el periodo 2016-2021 y que bate la rentabilidad acumulada del resto de activos financieros.



Introducción

Motivación

- Evaluar la viabilidad de un modelo de negocio relacionado con el mercado inmobiliario y para definir el Mínimo Producto Viable (MPV) era necesario realizar un análisis previo de la oferta de inmuebles recopilados y mostrados por plataformas inmobiliarias diversas.

Objetivos

- Entrenar y optimizar dos modelos de aprendizaje supervisado, que sean capaces de predecir el precio medio de los inmuebles en función de las variables explicativas contenidas en el DataFrame.
- Desarrollar una API que permita explotar la información y las predicciones de nuestro modelo

State of Art

No se dispone de ninguna herramienta que permita el acceso en tiempo real con el ajuste de un modelo a nivel nacional .

Tampoco existe al nivel de detalle necesario que se persigue el presente estudio.

Se basan en una muestra obtenida mediante scraping con menos profundidad de la que hemos utilizado en nuestro estudio.

Se han identificado numerosos análisis de modelos, pero que se circunscriben al análisis de los inmuebles de una provincia en concreto.

No alcanzan el nivel de puesta en producción del modelo

Metodología

Base de datos

- Obtenida por Web Scraping :2019
- Inicial: 100.000 – 39 variables explicativas (características del inmueble, geográficas, demográficas y económicas)

Análisis Exploratorio

- Correlaciones: No hay multicolinealidad. Y las >0.4 coinciden con los predictores más importantes

Procesamiento de la información

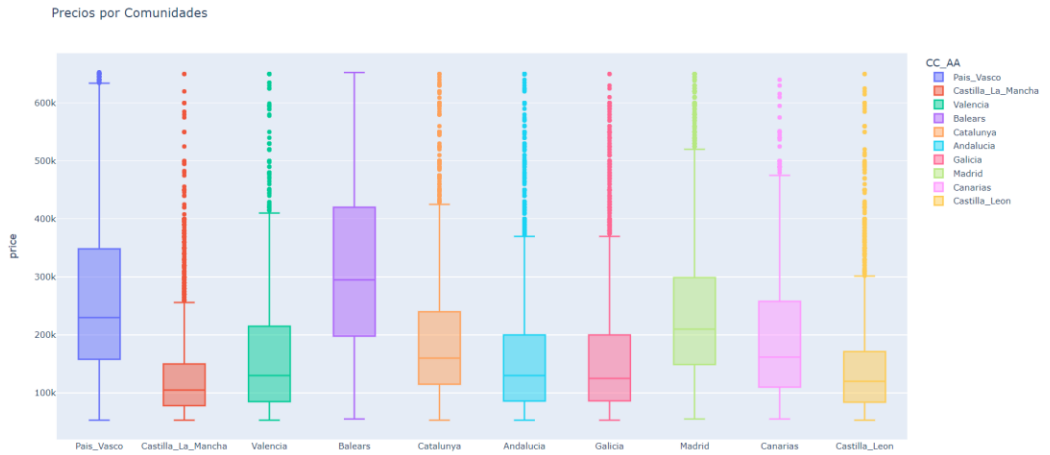
- No hay distribución normal de las variables
- 7 variables categóricas convertidas en [0,1]
- NaN: Criterio de la moda, K-Nearest Neighbor- KNN (float, m2_útiles)
- Estandarización 6 variables y Robust Scaler a m2- reales y útiles
- No se han modificado las variables a nivel provincial: económicas, demográficas(8 en total)
- BBDD Modelo: +75.000 - 39 variables - (alquiler, NaN no procesados, geográficas categóricas) = 22 var.

Metodología

Comunidad Autónoma	Número
Pais_Vasco	28.342
Balears	15.199
Galicia	6.539
Andalucia	5.386
Castilla_León	4.732
Castilla_La_Mancha	4.719
Madrid	4.272
Catalunya	2.619
Valencia	2.489
Canarias	971
Total	**75.268**

Tipo de Inmueble	Número
Piso	48.561
Casa o chalet	10.522
Chalet adosado	7.590
Ático	2.737
Casa rural	2.631
Dúplex	2.262
Estudio	631
Finca rústica	291
Otros	27
Masía	16
Total	**75.268**

99,5%



Elección de las métricas

- R2:explica cuán cerca están los datos de la línea de regresión ajustada
- MSE: Es una medida ponderada de la diferencia al cuadrado entre la predicción de nuestro modelo y el valor real.
- RMSE: se interpreta como la desviación estándar de la varianza inexplicada, y tiene la propiedad de estar en las mismas unidades que la variable target.

Metodología - Selección de Modelos

- El árbol de decisión > regresión lineal, cuando la relación de las variables explicativas en relación al target es compleja.
- Son más fáciles de explicar, y genera los predictores de forma automática (te indica las tendencias).
- Decision Tree Regressor: En el entrenamiento de un árbol de regresión, las observaciones se van distribuyendo por bifurcaciones (nodos) generando la estructura del árbol hasta alcanzar un nodo terminal.
- Random Forest: está formado por un conjunto de árboles de decisión individuales, cada uno entrenado con una muestra ligeramente distinta de los datos de entrenamiento generada mediante bootstrapping. Las ventajas este modelo son:
    a) No es necesario que se cumpla ningún tipo de distribución específica de las variables explicativas.
    b) Permiten identificar de forma rápida y eficiente las variables (predictores) más importantes de forma automática
    c) No se ven muy influenciados por outliers.
    d) Gracias al Out-of-Bag Error, puede estimarse su error de validación sin necesidad de recurrir a la validación cruzada.

Resultados

Importancia de los predictores

	Predictor Decision Tree Regressor	Importancia
17	Valor_catastral_medio	39,89%
2	bath_num	36,86%
8	m2_real	21,51%
7	lift	1,02%
15	number_of_companies_prov	0,69%
5	floor	0,04%

	Predictor Random Forest Regressor	Importancia
17	Valor_catastral_medio	22,18%
8	m2_real	22,05%
2	bath_num	19,40%
9	m2_useful	7,69%
5	floor	4,25%
15	number_of_companies_prov	3,19%
11	room_num	3,09%
18	CC_AA_no	2,65%
7	lift	1,83%
16	population_prov	1,81%
13	swimming_pool	1,59%
14	terrace	1,51%
3	built_in_wardrobe	1,47%
12	storage_room	1,46%
4	condition	1,33%
0	air_conditioner	1,15%
6	garden	1,14%
1	balcony	1,10%
10	reduced_mobility	0,64%
19	indoor	0,50%

Resultados y optimización

Métricas de los Modelos Iniciales

	Model	R2	MSE	RMSE	R2 - Cross Val.
0	Decision Tree Regressor	0.498637	9,77E+09	98.836,54	0.364789
1	Radom Forest	0.689052	6,06E+09	77.836,82	0.489874

Random Forest

Optimización de parámetros Out-off-bag

	oob_r2	max_depth	max_features	n_estimators
0	0.692404	NaN	5.0	150.0
3	0.692404	100.0	5.0	150.0
1	0.690757	NaN	7.0	150.0
4	0.690757	100.0	7.0	150.0

Métricas del Modelo Optimizado

	Model	R2	MSE	RMSE
0	Random Forest Optimizado	0.703150	5,78E+09	76.051,88
1	Radom Forest	0.689052	6,06E+09	77.836,82

Discusión de resultados

Mejoras sobre la Base de datos:

- Web Scraping: es posible obtener la información de forma más completa y de forma más directa a través del código postal.
- Ampliar la información de las provincias de las que no están bien representadas.
- Depurar la información relativa a los m2, reales y útiles, ya que son el segundo predictor más importante en ambos modelos y, que tienen un gran número de inconsistencias.

Mejoras de las Métricas del Modelo Final:

- Valor catastral medio si se incluye información de esta variable a niveles inferiores, ya sea por localidades, distritos o barrios, es muy probable que el modelo óptimo mejore significativamente sus métricas.
- Añadir la variable IBI (impuesto sobre Bienes Inmuebles) aportaría mayor calidad al ajuste del modelo.

Mejoras de las predicciones de la API:

- Completar las informaciones con datos del Registro de la Propiedad y del Registro de Transacciones Inmobiliarias, del Registro de notarios sobre precios reales de las compraventas, también mejorará la estimación de precios final del API.

Discusión de Resultados: modelo ajustado a Provincia y Budget

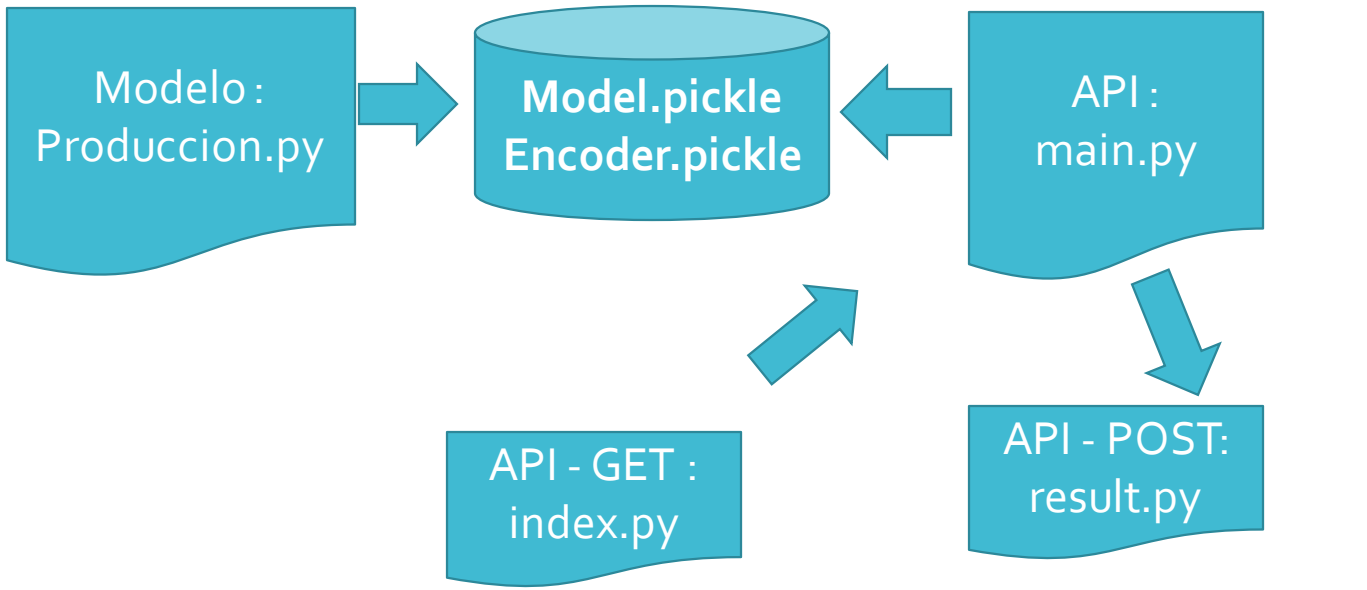
¿Sirve nuestro modelo cuando descendemos en el detalle geográfico?

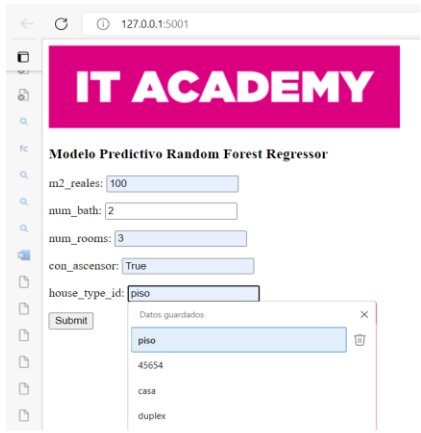
	Predictor Random Forest Regressor	Importancia
8	m2_real	23,75%
2	bath_num	18,71%
9	m2_useful	13,18%
5	floor	11,19
7	lift	9,42%
11	room_num	4,62%
4	condition	3,42%
3	built_in_wardrobe	3,16%
12	storage_room	3,02%
14	terrace	2,83%
1	balcony	2,34%
6	garden	1,96%
10	reduced_mobility	1,05%
19	indoor	0,95%
0	air_conditioner	0,21%
13	swimming_pool	0,19%
15	number_of_companies_prov	0,19%
16	population_prov	0,19%
17	Valor_catastral_medio	0,19%
18	CC_AA_no	0,19%

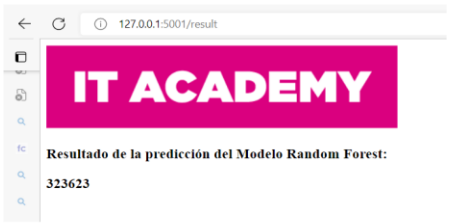
Métricas del Modelo Optimizado (Bizkaia, Budget= 250 mil €)

	Model	R2	MSE	RMSE
0	Random Forest Optimizado	0.524342	1,10E+09	33.190,27
1	Radom Forest	0.508411	1,14E+09	33.741,51

Esquema y Resultados del API







Conclusiones

Alineación de resultados y objetivos

- Desarrollo de una API basado en un modelo de aprendizaje automático, Random Forest con un coeficiente de determinación (R2) del 0.7%
- Se han indicado opciones de mejora : en la extracción de la información y en su tratamiento, que permitirán mejorar las métricas del modelo.
- Se ha demostrado que el modelo puede ajustarse para el análisis de subconjuntos del subset del Data Frame , que era una de las premisas del MPV que dio origen a este estudio.

Funcionalidad de la API

- Aunque se encuentra en una fase muy previa, cubre el objetivo de facilitar una predicción de precio medio estimado en función de un número reducido de variables explicativas.
- Es fácil ampliar el número de predictores al total de las variables del Data Frame.,
- También es escalable, para extraer información particular de los links de los anuncios de los inmuebles, que es otra de las premisas importantes previstas en el MPV.