Proyecto Data Science

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

Antonia María Sánchez Moreno
amariasm@telefonica.net

IT Academy
Barcelona Activa

4 de octubre de 2022

Resumen

La motivación de este proyecto surge cuando nos planteamos evaluar la viabilidad de un modelo de negocio, relacionado con el mercado inmobiliario y para definir el Mínimo Producto Viable (MPV).

El objetivo es obtener una Application Programming Interface (API) que proporcione la simulación de estimaciones de precios de los inmuebles, en función de las variables explicativas determinadas y el modelo de aprendizaje supervisado óptimo.

Definir un modelo de predicción de precios de la vivienda es la parte central del API y permitirá a los vendedores determinar el precio promedio al que deben poner al inmueble en venta, a las inmobiliarias asesorar a sus clientes en el proceso de negociación del precio entre el comprador y el vendedor, y a los compradores, puede ayudarles a encontrar el precio promedio correcto para comprar la casa, evaluando un mayor número de parámetros, de forma más objetiva.

Palabras clave: Aprendizaje Supervisado, Decision Tree Regressor, Random Forest Regressor, Mercado Inmobiliario, API, Python, Flask.

Abstract

The motivation of this project arises when we consider evaluating the viability of a business model, related to the real estate market and to define the Minimum Viable Product (MPV).

The objective is to obtain an application programming interface (API) that provides the simulation of real estate price estimates, based on the explanatory variables determined and the optimal supervised learning model.

Defining a housing price prediction model is the central part of the API and will allow sellers to determine the average price at which they should put the property up for sale, real estate agents to advise their clients in the price negotiation process between the buyer and seller, and buyers, can help them find the correct average price to buy the house, evaluating a greater number of parameters, in a more objective way.

Keywords: Supervised Learning, Decision Tree Regressor, Random Forest Regressor, Real Estate Market, API, Python, Flask.

---

1. Introducción

El sector de la construcción representa en España el 6.2 por ciento del PIB, emplea a más de 1,4 millones de personas y según la última Encuesta de Población Activa (EPA), a cierre de marzo había 167.100 personas trabajando en el Sector Servicios de actividades inmobiliarias.

En el año 2021, se realizaron un total de 678.000 operaciones de compra venta de inmuebles ante notario, según los datos del Banco de España (BdE) y el esfuerzo que realizan las familias para la adquisición de los mismas son 7 años en promedio.

Pero la clave del interés en analizar la evolución del sector inmobiliario, la oferta disponible y los precios actualizados en tiempo real, se explica porque 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: Depósitos, Deuda pública a 10 años, y Bolsa Española-Ibex, tal y como se refleja en la Figura 1 - [5] Bbva Research

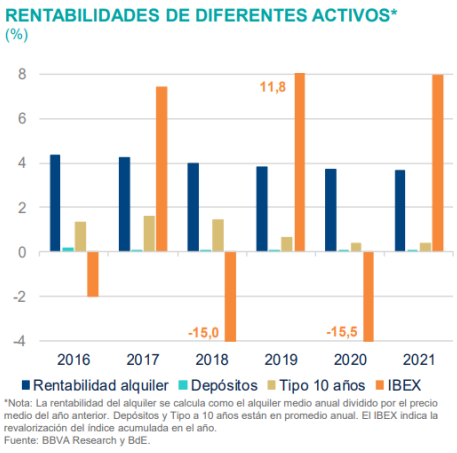

Figura 1. Rentabilidad de los activos. BBVA Reserch

1.1. Motivación

La motivación de este trabajo surgió cuando nos planteamos 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.

1.2. Objetivos

Los objetivos de este proyecto se pueden dividir en dos líneas de trabajo. La primera será 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. Para ello, realizaremos un análisis exploratorio de datos, ingeniería de características principales de las variables y ajuste de hiperparámetros para mejorar el rendimiento del modelo final. El segundo objetivo será desarrollar una API que permita explotar la información y las predicciones de nuestro modelo, siguiendo el esquema de la Figura 2.

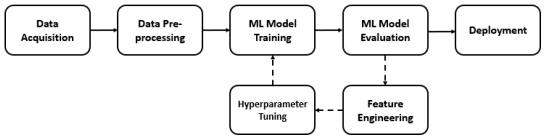

Figura 2. Machine Learning Pipeline

2. State of Art

En relación a análisis del Sector inmobiliario existen numerosas instituciones que realizan informes detallados y numerosas plataformas que recogen bases de datos con características diversas, pero no se dispone de ninguna herramienta que permita el acceso en tiempo real con el ajuste de un modelo a nivel nacional y el nivel de detalle necesario que se persigue el presente estudio.

Después de una prospección de mercado se han identificado numerosos análisis que se basan en el mismo esquema que se ha seguido en este estudio, pero que se circunscriben al análisis de los inmuebles de una provincia en concreto, y no alzanzan el nivel de puesta en producción del modelo, o que se basan en una muestra obtenida mediante scraping con menos profundidad de la que hemos utilizado en nuestro estudio.[3] (Bizkaia y Guipuzcua) y [7](Valencia)

El estudio que más se aproxima a nuestro enfoque es [1] Rentabilidad de activos inmobiliarios, pero no han desarrollado una API propia y se basa en la información de acceso libre de la API de Idealista.com, que se limita a facilitar una base de 100 inmuebles que no cubre las necesidades de nuestro análisis.

3. Metodología

3.1. Base de Datos

El Dataframe empleado en este análisis se ha obtenido de un repositorio github extraído de la plataforma idealista.com, mediante técnicas de Web Scraping. Las fechas de extracción de la información relativa a veinte provincias españolas se realizaron entre marzo y abril de 2019 con un total de 100.00 registros, 39 variables explicativas y price como variable objetivo o target.[8]Rullán

Hemos unido las bases provinciales y completado con datos económicos y demográficos adicionales, obtenidos también mediante Web Scraping del portal del Ministerio de Hacienda y de la Hacienda Foral Vasca.[9] Sánchez

3.2. Análisis Exploratorio

3.2.1. Variables Explicativas

Las 40 variables originales, se han agrupado en: Variables relativas a las características del inmueble, variables geográficas y variables demográficas y económicas.

Variables por Categorías		
Inmueble	Geográficas	Económicas
m2 reales/útiles	CC.AA	Val.Catastral
rooms/bath/storage	Provincia	CC.AA.no
floor/lift	Zona	Población
air conditioner	Ciudad	Empresas
energetic certif		
balcony		
condition/indoor		
reduced mobility		
terrace/pool/garden		
etc..		

Tabla 1. Grupos de Variables por categorías.

Se ha depurado la información para obtener el mayor número de datos completos, analizando la tipología de cada cada variable y eliminado los registros de datos erróneos y no consistentes, y con el objetivo de reducir el número de variables categóricas se han reemplazado los los valores no numéricos por numéricos, y se han agrupado las categorías homogéneas.

En relación a las variables geograficas se ha optado por agrupar a nivel de Comunidad Autónoma y Provincia, ya que los datos de localización a nivel de zonas geográficas implicaba generar 94 dummies y 1.138 a nivel de Ciudades y Distritos.

La concreción del análisis exploratorio nos ha permitido reducir el número de variables de 39 a 22, y la matriz de correlaciones nos permite cubrir tres objetivos de los siguientes pasos del análisis exploratorio: analizar la existencia o no de multicolinealidad, evaluar la consistencia de las correlaciones entre las variables explicativas y en relación al target y obtener una primera aproximación a los predictores del modelo de regresión.

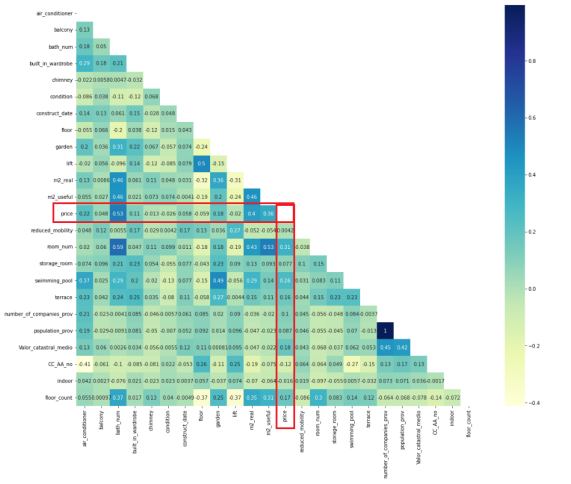

Figura 3. Matriz de Correlaciones

Como se observa en la Figura 3, no existe multicolinealidad significativa entre las variables explicativas, salvo en el caso de Población y Nímero de Empresas por provincias, pero hemos mantenido las dos variables ya que son interesantes para obtener un modelo ampliado en base a la localización aunque el presente artículo no lo abarca.

Esperábamos una mayor correlación entre los m2 reales y útiles, pero el resutado se ajusta al comportamiento de ambas variables, ya que los m2 reales presentan outliers muy significativos y en ocasiones no justificables, y la variable m2 útiles tenía un gran número de datos no informados.

En general, las correlaciones superiores al 0.4 entre las variables explicativas y el targert (price) se aproximarán a los Predictores de los modelos, salvo en el caso de la variable Valor-Catastral-Medio que tiene una correlación inferior de 0.18, pero que resultará determinante como predictor de los modelos.

3.2.2. Target - Price

La variable objetivo price, también presenta numerosos outliers que hemos analizado, y en todos los casos que hemos encontrado que se trataba de errores los hemos subsanado, pero en general los precios son coherentes con la descripción de los anuncios, dada la diversidad de tipologías de inmuebles que se encuentran en la base de datos.

También existe una gran diferencia entre las Comunidades, donde se observa que aquellas en las que se ha realizado un scraping mas profundo reflejan datos con menos outliers que las que se han obtenido menos registros. Figura 4

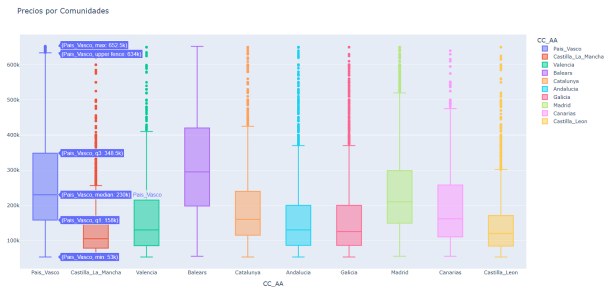

Figura 4. Precios Comunidades-20 provincias

3.3. Preprocesamiento de la información

3.3.1. Tratamiento de los valores nulos

En el tratamiento de los valores nulos se han seguido tres estrategias diferentes: a)Se han eliminado las variables kitchen y unfurnished con porcentaje de NaN de 100 por cien, y porque están relacionadas con las viviendas de alquiler.

b) Dos variables que consideramos importantes, m2 usefull y floor, las hemos completado mediante K-Nearest Neighbor. KNN es un algoritmo que para predecir una observación, identifica las K observaciones del conjunto de entrenamiento que más se asemejan a ella en base a sus predictores, y se emplea como valor predicho el promedio de la variable que deseamos completar.

c) Completados con la moda: indoor (interior o exterior), sistema de calefacción(heating) y certificado enérgético.

3.3.2. Normalización, Estandarización y Robust Scaler

No existen variables que tengan una distribución normal en el Data set, y se ha estandarizado 6 variables: bath-num, condition,construct-date, floor, room-num y floor-count. Robust Scaler se ha aplicado a m2-real y m2-usefull ya que presentaban un gran número de outliers.

3.3.3. Data Frame Procesado

Finalmente dado que no hemos modificado ni tratado los outliers de la variable target-price, porque después de analizarlos eran consistentes con las descripciones de los anuncios, lo que hemos hecho es reducir parte del numero de categorías de Otros, porque incluía un gran número de inmuebles del tipo hotel, fincas, castillos, masías y casas rurales,etc.. pero se ha aplicado un criterio sobre el precio, que ha sido excluir aquellos inmuebles que se situaban por encima del percentil 0.9 y por debajo del 0.05, rango que coincide con el boxplot de la Comunidad de Baleares, que es la que más edificios, en particular Hoteles, tenía. Figura 4.

También hemos excluido del anális los inmuebles en alquiler, ya que requerirían un análisis previo para asignarles un valor de mercado, en función de la renta mensual informada en los anuncios. El Data frame final contiene 39 variables explicativas, el target-price y 75.268 registros que se distribuyen por Comunidades y Tipos de inmueble en la Figura 5.

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**

Figura 5. Data Frame depuradol

3.4. Elección de las métricas

Elegir la métrica de evaluación correcta nos ayuda a evaluar el rendimiento de nuestro modelo. Para los problemas de regresión, las métricas de evaluación de referencia son: Coefficient of determination (R2),

Mean squared error (MSE) y, Root mean square deviation (RMSE).

a) R2:Es una medida estadística que explica cuán cerca están los datos de la línea de regresión ajustada, y explica cuánta variabilidad del target puede ser causada por su relación con las variables explicativas.

b) MSE: Es una medida del promedio de la diferencia al cuadrado entre la predicción de nuestro modelo y el valor real.

c) RMSE: Es una raíz cuadrada del MSE, se puede interpretar como la desviación estándar de la varianza inexplicada, y tiene la propiedad de estar en las mismas unidades que la variable target.

3.5. Selección de los Modelos

Para determinar los modelos que vamos a aplicar a nuestro análisis hemos tenido en cuenta las caraterísticas de los modelos que mejor se ajustan a nuestros datos y objetivos.

Cuando la relación de las variables explicativas en relación al target es compleja y no lineal, el árbol de decisión facilita mejores resultados que un método clásico de regresión. Además, si se quiere construir un modelo que sea fácil de explicar, y que genere los predictores de forma automática, entonces un modelo de árbol de decisión será mejor que un modelo lineal.

En base a estos parámetros, hemos planteado la evaluación de dos modelos basados en árboles de decisión: Decision Tree Regressor y Random Forest Regressor.

3.5.1. Decision Tree Regressor Model

Decision Tree Regressor [6] es un subtipo de árboles de predicción que se aplica cuando la variable respuesta es continua. 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. Cuando se quiere predecir una nueva observación, se recorre el árbol acorde al valor de sus predictores, hasta alcanzar uno de los nodos terminales. La predicción del árbol es la media de la variable respuesta de las observaciones de entrenamiento que están en ese mismo nodo terminal.

3.5.2. Random Forest Regressor Model

El modelo Random Forest [6] 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. La predicción de una nueva observación se obtiene agregando las predicciones de todos los árboles individuales que forman el modelo. Las ventajas este modelo son:

a) No es necesario que se cumpla ningún tipo de distribución específica de las variables explicativas, y requieren mucha menos limpieza y pre-procesado de los datos.

b) Son muy útiles en la exploración de datos ya que permiten identificar de forma rápida y eficiente las variables (predictores) más importantes de forma automática y no se ven muy influenciados por outliers.

c) Gracias al Out-of-Bag Error, puede estimarse su error de validación sin necesidad de recurrir a la validación cruzada.

3.6. Desarrollo del API

Para desarrollar un MPV basado en el aprendizaje automático, es necesario crear servicios que otros equipos puedan usar o un producto donde los usuarios puedan interactuar.

Para complementar nuestro proyecto, se ha desarrollado un API, que es la forma en que los sistemas informáticos se comunican entre sí, actuando como un agente que lleva la información del usuario al servidor y luego nuevamente del servidor al usuario devolviendo la respuesta.[4]

Flask es una librería de Python que proporciona esa capacidad, actuando a su vez como una API entre nuestro modelo predictivo, el archivo HTML de GET POST y RESPONSE.[2]

En nuestro caso la API abarca:

a) Protocolo de transferencia HTTP: es la forma principal de comunicar información en la web.

b) Métodos: GET: este método permite obtener información de la base de datos o de un proceso, POST: permite mandar información, ya sea para añadir información a una base de datos o para pasar el input de un modelo de machine learning, como es nuestro caso.

4. Resultados

Se ha evaluado la precisión de las predicciones mediante train-test y Cross validation y se han optimizado los parámetros para seleccionar el modelo final óptimo.

4.1. Importancia de los predictores

4.1.1. Decision Tree Regressor

El árbol se divide en función de las variables que definen los nodos y la importancia de los predictores se calcula en función del numero de veces que aparecen en cada nodo. Figura 6

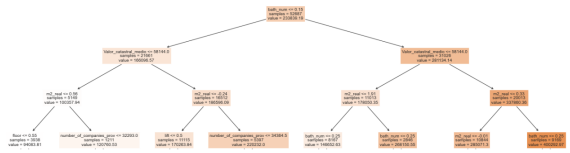

Figura 6. Esquema del Árbol

El punto más significativo de estos resultados es que el predictor más importante para este modelo es el Valor-catastral-medio, que es una variable que hemos incorporado al data set inicial y que tiene 20 valores únicós por cada provincia. La importancia de este predictor es del 39.89 por ciento y que junto a bath-num y m2-reales explican más del 90 por cien de las prediciones del modelo. Figura 7

	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%

Figura 7. Predictores Decision Tree Model

El peso de la Variable Valor catastral medio se explica porque es un valor sintético ajustado del precio de los inmuebles, ya que en su definición se evalúa un gran numero de parámetros de los mismos, de forma objetiva y pormenorizada.

De la misma forma que el número de baños sintetiza las variables de m2 reales y útiles.

4.1.2. Random Forest Regressor

El modelo de Random Forest distribuye el peso de los predictores entre las 20 variables explicativas, y otorga un peso del 22.18 por ciento a la variable Valor catastral medio. Figura 8

	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%

Figura 8. Predictores-Random Forest

4.2. Métricas de los Modelos y Validación Cruzada

La precisión de las predicciones iniciales se han obenido mediante train-test y se han contrastado con Cross validation. El R2 de los modelos iniciales es de 0.4986 para Decision Tree y 0,6890 en Random Forest. Las cifras de Cros Validation han resultado significativamente inferiores, con (R2) de 0.3658 y 0.4899 respectivamente. Figura 9

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

Figura 9. Métricas de los modelos

4.3. Optimización de parámetros

La optimización de parámetros la hemos realizado sobre Random Forest Regressor, ya que es el modelo que ofrece mejores métricas en train-test y Cross-validation, y se han optimizado mediante Out-of-Bag Error los siguientes parámetros para seleccionar el modelo final:

a) max-deph: define la profundidad máxima del árbol, es decir, el número de divisiones de la rama más larga, en sentido descendente, del árbol.

b) max-features : Indica cuántas de las características o predictores, se deben considerar en cada división.

c) n-estimators: Número de árboles a crear y generalizar sobre Random Forest. Figura 10

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

Figura 10. Métricas optimización-parámetros

4.4. Métricas Modelo Final

Las métricas de nuestro modelo final optimizado presentan un R2 del 0.7012, con un RMSE de 76.05 mil €, lo que supone una desviación respecto al precio medio de 195.00 mil €, de 0.38, que es coherente con el coeficiente de determinación obtenido. Figura 11

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

Figura 11. Métricas modelo optimizado

4.5. Análisis de los residuos

La idea detrás de una gráfica QQ-plot es simple. Si los residuos se ajustan a lo largo de una línea recta aproximadamente en un ángulo de 45 grados, éstos se distribuyen de manera aproximadamente normal. En nuestro gráfico QQ-plot los residuos tienden a desviarse, especialmente en los extremos de la cola. Aunque un gráfico QQ no es una prueba estadística formal, es una forma visual de verificar visualmente si los residuos están distribuidos normalmente o no.

En nuestro caso la varianza se explica por la existencia de outliers que aunque hemos preprocesado, no hemos podido ajustar para alcanzar un nivel de R2 superior al 0.7 y tiene su origen en la calidad de la información adicional del data set. Figura 12

Para reducir las desviaciones de los extremos se puede considerar realizar una transformación de las variables, usando por ejemplo la raíz cuadrada o el logaritmo de las mismas, se podría utilizar pruebas de igualdad de varianza (complementarias a los análisis gráficos) o modelar la heterogeneidad encontrada con modelos generalizados (GLM) o modelos mixtos (MM).

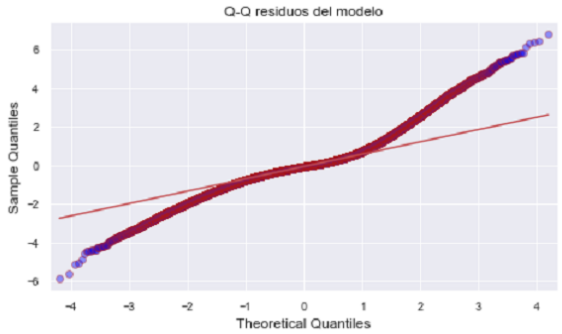

Figura 12. Análisis de los residuos del Modelo General

4.6. Resultados del API

A partir de los predictores obtenidos, se ha realizado un modelo abreviado con cinco variables explicativas: m2-reales, número de baños, número de habitaciones, si tiene o no ascensor y tipo de inmueble residencial.

Con esta información que el usuario completa en un formulario html, el API obtiene mediante un GET las variables, procesa la mismas aplicando el modelo Random Forest y la transformación de variables definidas, y obtiene a partir de la base de datos reducida de los inmuebles de la provincia de Madrid, el precio estimado de tipo de inmueble en función de los parámetros definidos. Tal y como se recoge en las capturas de pantalla de la Figura 13.

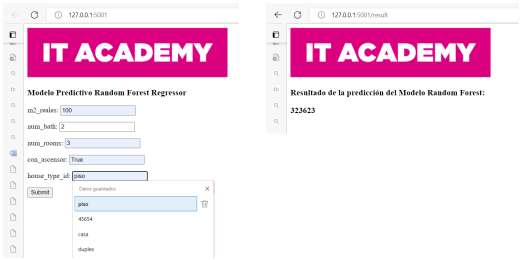

Figura 13. Api Modelo predictivo de precios

5. Discusión de resultados

5.1. Mejoras sobre la Base de datos y las Métricas del Modelo Final

a) En relación al Web Scraping, es posible obtener la información de forma más completa aplicando un bucle con el código postal y posteriormente accediendo al (id) de cada anuncio, de esta forma se reduce una capa del Scraping en relación a la metodología usada en las extracción de la base de datos de este estudio y sería más rápido extraer toda la información.

b) Ampliando la información de las provincias y completando el Scraping de las que no están bien representadas, obtendríamos mejores resultados para poder generalizar el análisis a todo el territorio nacional.

c) Dada la importancia que los dos modelos otorgan a la variable Valor catastral medio, que es un dato a nivel provincial, 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.

En su defecto, se intuye que añadir la variable IBI (que es el impuesto sobre Bienes Inmuebles que se paga a los ayuntamientos y tiene como base el valor catastral del inmueble) en la descripción de los anuncios a completar por los vendedores aportaría mayor calidad al ajuste del modelo.

d) 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. Los datos originales presentan distorsiones en la correlación, que en estas dos variables debería ser evidente y en cuanto a los outliers no son justificados, e intuimos que se debe a la mala calidad de la información que competan los vendedores en el portal de idealista.

e) Completar las informaciones con datos del Registro de la Propiedad en relación a los valores catastrales y del Registro de Transacciones Inmobiliarias, del Registro de notarios sobre precios reales de las compraventas, también mejorará significativamente nuestras predicciones.

5.2. Resultados del modelo ajustado a Provincia y Budget

Hemos evaluado el comportamiento de nuestro modelo a nivel de subset de Provincia y con limitación de Budget de compra, hemos seleccionado los datos de la provincia de Bizkaia y un presupuesto de 250 mil € sobre un total de 10.101 inmuebles.

El modelo reduce sus métricas óptimas del 0.70 al 0.50 y eso se debe a que la variable que antes explicaba el 20 por ciento del ajuste, Valor catastral medio, en este caso al ser un dato similar para toda la provincia, solo explica el 0.19 por ciento del precio estimado. Figuras 14 y 15

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

Figura 14. Métricas (Bizkaia, Budget<250m)

En relación a la importancia de los predictores, en este caso en particular, prevalecen las variables asociadas a las características de los inmuebles y el predictor más significativo es m2 reales con un 23.75 por ciento de peso en la predicción. Figura 15

6. Conclusiones

6.1. Alineación de resultados y objetivos

El presente proyecto ha cubierto todos los ámbitos previstos en los objetivos iniciales, con el desarrollo de una API basado en un modelo de aprendizaje automático de regresión múltiple, Random Forest con un coeficiente de determinación (R2) sobre la base global del 0.7.

Adicionalmente se han encontrado opciones de mejora tanto sobre la extracción de la información como

	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%

Figura 15. Predictores (Bizkaia y Budget<250m)

de su tratamiento, que permitirán mejorar las métricas del modelo para el análisis de subconjuntos del subset del Data Frame que era una de las premisas del MPV que dió origen a este estudio.

6.2. Funcionalidad de la API

En cuanto a la API aunque se encuentra en una fase muy previa, sí que cubre el objetivo inicial planteado de facilitar una predicción de precio medio estimado en función de un número reducido de variables explicativas, pero es fácil ampliar el número de predictores al total de las variables del Data Frame, así como extraer información particular de los links de los anuncios de los inmuebles, que es otra de las premisas importantes previstas en el MPV.

Referencias

[1] Javier Holguera Crespo. *Sistema de análisis de rentabilidad de activos inmobiliarios*. UAM. Departamento de Ingeniería Informática, 2019. URL: https://repositorio.uam.es/handle/10486/692893.

[2] Ander Fernández Jauregui. *Cómo crear una API en Python*. 2022. URL: https://anderfernandez.com/blog/como-crear-api-en-python/.

[3] Haritz Laboa. *Compra tu casa de forma inteligente*. 2019. URL: https://www.hlaboa.com/post/Compra_tu_casa_de_forma_inteligente_1_web_scraping/.

[4] Gael Varoquaux y otros. *API design for machine learning software: experiences from the scikit-learn project*. 2013. DOI: https://doi.org/10.48550/arXiv.1309.0238.

[5] Bbva Reserch. *Observatorio Inmobiliario, Ier. Trimestre*. 2022. URL: https://www.bbvaresearch.com/publicaciones/espana-analisis-de-la-demanda-de-vivienda-tras-la-irrupcion-de-la-covid-19/.

[6] Joaquín Amat Rodrigo. *Random Forest con Python*. 2020. URL: https://www.cienciadedatos.net/documentos/py08_random_forest_python.html.

[7] Alejandro Antón Ruíz. *Predicción del precio en el mercado de la vivienda en la ciudad de Valencia mediante redes neuronales en el años 2020*. UPV Facultad de Administración de Empresas, 2020. URL: https://riunet.upv.es/handle/10251/152158.

[8] Marcos Rullán. *Spain Housing Crawler*. 2019. URL: https://github.com/trueuoc/spa_housing_crawler.git.

[9] Antonia M Sánchez. *Sprint 10: Web Scraping*. 2022. URL: https://github.com/amariasm/web_scrapping.git.