

1,772 x 1

Aplicación de Data Science en la Valoración de Viviendas en Madrid España



Desafío Idealista - MDS

Proyecto final de Máster

Autores

Alejandro González López
Arq. Leonardo Velásquez Estupiñan

Director: David Rey, Chief Data Officer de Idealista

Resumen

Este documento presenta el proyecto "Aplicación de Data Science en la Valoración de Viviendas en Madrid España", un modelo de valoración de viviendas desarrollado como parte del Master de Data Science y Big Data Analytics de The Valley Business School en colaboración con Idealista como solución al desafío lanzado por Idealista de nombre "¿Cuánto vale mi casa?" como proyecto final de Máster. Utilizando una amplia base de datos de anuncios de viviendas en Madrid en 2018, el proyecto ofrece un enfoque detallado, resultados obtenidos y detalles sobre la implementación del modelo en un entorno en la nube.

Capítulo 1: Introducción al Desafío Idealista y Enfoque Propuesto.....	1
Abstract.....	1
Recursos de datos disponibles.....	1
Enfoque.....	2
Caso de uso propuesto.....	3
Marco Teórico.....	4
Estado del Arte y referentes.....	5
Capítulo 2: Análisis Exploratorio e Ingeniería de Variables.....	6
Estadísticas de Resumen.....	6
Outliers y corrección de datos.....	9
Target Encoding – Locationname (MEAN).....	10
Próximos pasos en Ingeniería de variables.....	11
Matriz de Correlación.....	11
Escala de las variables.....	12
Variables definitivas.....	13
Capítulo 3: Modelado.....	14
1. Random Forest Regression.....	14
2. Regresión Lineal.....	17
3. XGBoost.....	17
Selección final de modelo.....	18
Discusión.....	18
Modelado 2.0.....	19
Capítulo 4: Productivización.....	22
AI Idealista – App.....	22
Inferencia automática de variables de distancia a puntos urbanos.....	23
Streamlit.....	24
Transparencia.....	24
Productivización 2.0.....	24

Capítulo 1: Introducción al Desafío Idealista y Enfoque Propuesto

Abstract

Este documento presenta el producto de datos desarrollado como parte del desafío "¿Cuánto vale mi casa?" lanzado por la empresa Idealista como proyecto final del Master de Data Science y Big Data Analytics de The Valley Business School. El objetivo principal de este proyecto es crear un modelo de valoración de viviendas en el mercado de compraventa que permita estimar el precio al que se encontraría un anuncio de una vivienda publicado en el portal de Idealista.

El desafío se basa en una amplia base de datos que incluye 94,815 anuncios de inmuebles en venta en la ciudad de Madrid, España, durante el período de 2018. Esta base de datos proporciona información detallada sobre las características de las propiedades, como el número de metros cuadrados construidos, su distribución y más. Además, se han puesto a disposición conjuntos de datos complementarios que enriquecen las variables utilizadas para entrenar el modelo, incluyendo datos geográficos, distancias a puntos de interés e información catastral.

En este documento, se detalla el enfoque y los resultados alcanzados por los autores del proyecto final, Alejandro López y Leonardo Velásquez bajo la dirección de David Rey, CDO de Idealista. Se explican las fases de desarrollo del modelo, desde su concepción hasta su implementación, y se presentan los resultados obtenidos. Además, se describe cómo se ha desplegado el modelo en un entorno cloud para su uso y evaluación.

Recursos de datos disponibles

Como recursos de datos, el desafío Idealista puso a disposición una variedad de fuentes valiosas que se utilizaron en este proyecto:

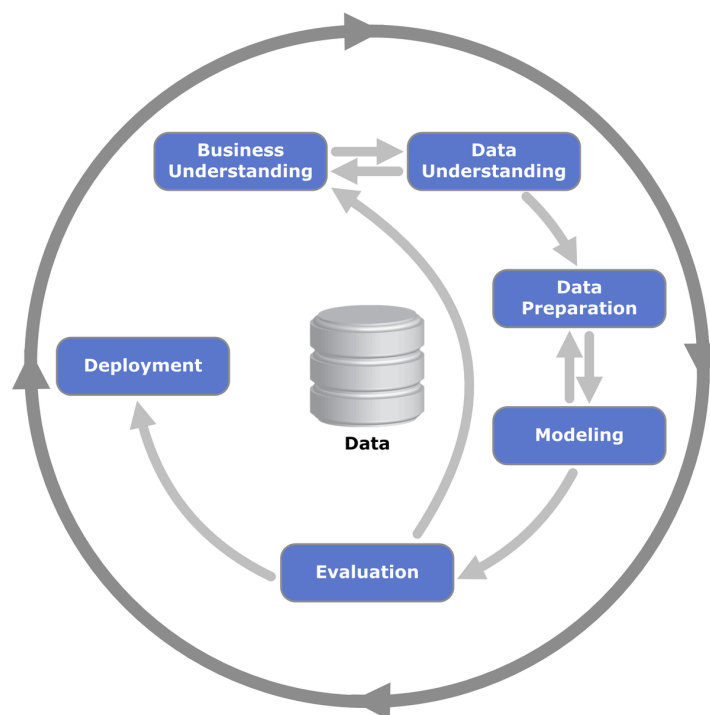
- Base de datos de Idealista: Esta base de datos contiene anuncios de ventas de propiedades en la ciudad de Madrid, España, durante el año 2018. Los datos otorgados por Idealista fueron previamente anonimizados* por Idealista debido a requerimientos legales introduciendo una pequeña perturbación aleatoria en el precio total y la posición exacta.
- Base de datos de Polígonos: Se utilizó para dividir el espacio de la ciudad de Madrid en zonas geográficas específicas, como barrios y áreas de interés dentro de Idealista.
- POIs (Puntos de Interés): Se recopilaron puntos de interés en la ciudad de Madrid, que son elementos clave dentro de esta metrópolis, abarcando desde ubicaciones como farmacias y restaurantes hasta equipamientos urbanos relevantes.
- Base de datos de OpenStreetMaps (OSM): Se incluyeron puntos de interés de OpenStreetMaps, como farmacias, restaurantes y otros equipamientos urbanos, para enriquecer aún más la información geográfica.
- Base de datos del INE (Instituto Nacional de Estadística): Esta fuente de datos contiene secciones censales del INE, así como datos del censo de población y viviendas

recopilados en 2011. Cada sección censal se identifica mediante un código de diez caracteres llamado CUSEC. A pesar de su antigüedad, este conjunto de datos proporciona información descriptiva de alto valor.

Utilizando estos recursos de datos mencionados, se creó un conjunto de datos de trabajo definitivo, el cual se detalla con mayor profundidad en el capítulo de ingeniería de variables de este documento.

Enfoque

Hemos definido como enfoque de este proyecto el abordaje del desafío planteado por Idealista siguiendo el ciclo de desarrollo de un modelo de inteligencia artificial estándar, el cual comprende las siguientes etapas: 1) Comprender el contexto del negocio, 2) explorar y entender a fondo los recursos de datos disponibles, 3) preparar los datos de manera minuciosa, 4) crear y ajustar el modelo, 5) evaluar y finalmente 6) implementar el modelo en un entorno de producción.



Nos planteamos como objetivo la generación de un modelo de predicción del valor de viviendas que alcance estándares de precisión equiparables o superiores a los modelos de referencia, priorizando el modelo que haya demostrado un rendimiento general óptimo, además de evaluar el rendimiento del modelo en términos de error y precisión en cada uno de los barrios de Madrid para estudiar la granularidad de dicho modelo y su precisión por zona.

Un aspecto fundamental de nuestro proyecto es garantizar la integridad y coherencia de las variables utilizadas en el entrenamiento del modelo puesto que el dataset dispone de datos reales y su evaluación será contrastada con casos de uso reales de la plataforma Idealista. Esto se logró a través de un proceso riguroso de limpieza y tratamiento de datos basados en modelos estadísticos y conocimiento de negocio.

Caso de uso propuesto

Según datos del mismo Idealista¹, Madrid ostenta una posición destacada en el panorama inmobiliario nacional y lidera la clasificación de ciudades con el mayor número de transacciones de compraventa. Un 25% de las propiedades se venden en menos de una semana, el 22% se comercializa en un plazo que oscila entre una semana y un mes, otro 25% requiere de uno a tres meses para su venta, un 22% demora de tres meses a un año en encontrar comprador, y un 5% precisa de más de un año para completar la transacción.

Con este panorama, decidimos abordar el desafío desde la perspectiva de los usuarios propietarios de inmuebles en Madrid que quieren estimar el valor de sus inmuebles para posteriormente anunciarlos en idealista o conectar con una agencia inmobiliaria que lleve la gestión comercial del mismo, agilizando una estimación ágil y significativa de sus propiedades con base en datos del mercado.

Como resultado final, planteamos la producción de un producto digital funcional ficticio desplegado en la nube. Este producto estaría al alcance de los usuarios de Idealista, permitiéndoles realizar predicciones sobre el valor de su inmueble en pocos minutos y acceder a información valiosa de los mismos simplemente proporcionando datos de sus propiedades. El valor predicho puede ser usada como un valor de referencia para contrastar con tasaciones hechas in situ, contrastarse con el valor de otros anuncios de inmuebles listados en las mismas zonas o simplemente para estimar el valor de las propiedades de los usuarios de manera ágil.

Para Idealista este producto digital representaría una oportunidad de activar y adquirir más anunciantes en la plataforma y referenciamientos a las inmobiliarias con quienes Idealista tiene convenios, y por tanto fortalecer sus líneas de negocio relacionadas con la tasación, el listamiento de propiedades en su plataforma y sus convenios comerciales con agencias inmobiliarias.

Marco Teórico

El marco teórico de este proyecto se basa en diversos conceptos fundamentales, entre los cuales se destaca el modelo de regresión hedónica en la investigación de viviendas y los modelos de Machine Learning con estructura de árboles. Además, se incorporan otras áreas de relevancia que contribuyen a la valoración precisa de las propiedades. A continuación se describen los cinco marcos dentro de los cuales se desarrolló este proyecto:

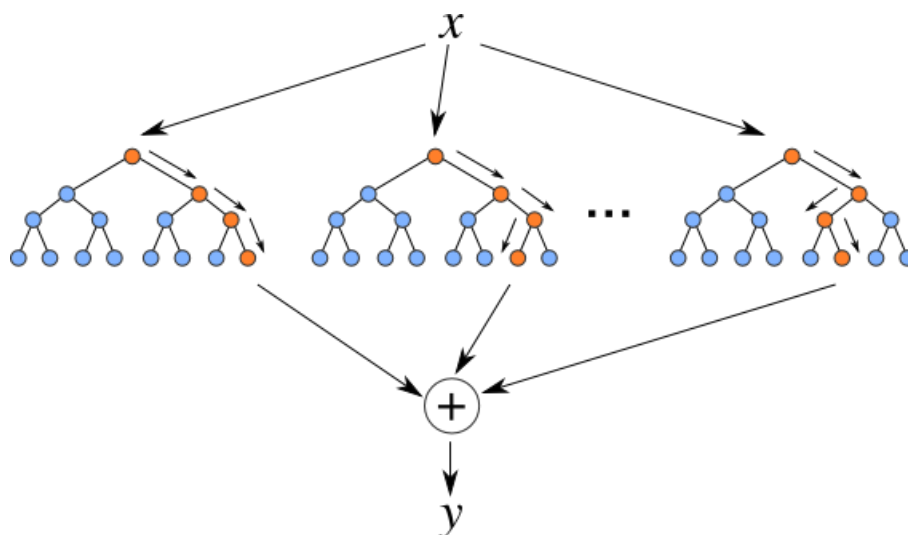
1. **Modelo de Regresión Hedónica en la Valoración de Viviendas:** La regresión hedónica se rige como un pilar esencial en la valoración de viviendas. Este enfoque permite desglosar el precio de una propiedad en función de sus características específicas. Esto conlleva a la identificación y cuantificación de cómo aspectos como el tamaño, la ubicación, el estado de la propiedad y otras características influyen en su precio de mercado.
2. **Importancia de la Ubicación y la Cercanía a Puntos de Interés:** Exploramos cómo la ubicación geográfica y la disponibilidad de servicios cercanos impactan en el precio de las

¹ Idealista – Cuánto se tarda en vender una vivienda en España en 2023
(<https://www.idealista.com/news/inmobiliario/vivienda/2023/03/03/803920-cuanto-se-tarda-en-vender-una-vivienda-en-espana-en-2023>)

viviendas, prestando atención a variables como la distancia al centro de la ciudad, a estaciones de metro y la Paseo La Castellana partiendo de la base de datos de ubicación de una propiedad y su proximidad a puntos de interés desempeñan un papel crucial en la valoración de viviendas.

3. **Análisis Estadístico e Ingeniería de Variables:** Se llevó a cabo un análisis estadístico de las variables relevantes para el proyecto como la correlación entre las variables y su influencia en el valor de las viviendas, utilizando métricas como el valor P (P-Value) para identificar los atributos que tienen una mayor relevancia en la valoración así como la exploración de la distribución de datos, identificación de valores atípicos y la evaluación de la normalidad. También se han tratado dichas variables para que puedan ser usadas en el modelo con métodos como target encoding, dummificación y ajuste logarítmico.
4. **Modelos de Machine Learning y Deep Learning:** Se incorporaron técnicas de machine learning y deep learning en el proyecto tales como KNN, XGboost, Random Forest y Linear Regression entre otras para encontrar el mejor rendimiento posible para la predicción del precio de las viviendas. Se dió énfasis a modelos de regresión basados en estructuras tipo árboles, como el Random Forest Regression, que han demostrado un rendimiento destacado en comparación con los otros métodos explorados.
5. **Modelo Random Forest Regression:** Random Forest Regression es una técnica de aprendizaje automático que combina múltiples modelos de regresión tipo árbol de decisión en un ensamblaje tipo Bagging, de dónde viene el nombre "bosque/forest", para realizar predicciones más precisas y robustas. Durante la ejecución de la fase de modelado, este tipo de modelo mostró un rendimiento superior en comparación a otros modelos como regresiones lineales y modelos de clustering tipo KNN (K-Nearest Neighbors). Como métrica de evaluación de este tipo de modelos es frecuente usar el RMSE (error cuadrático medio) porque penaliza los resultados en los que la desviación de la predicción es muy alta, MAPE (error medio en porcentaje) y Puntos por calidad del ajuste R^2 calculado como $\text{varianza error} / \text{varianza del conjunto de validación}$. Dichas métricas fueron empleadas durante la fase de evaluación y optimización.

Esquema gráfico de un modelo Random Forest



En conjunto, este marco teórico proporciona una sólida base conceptual que combina métodos tradicionales de valoración inmobiliaria con técnicas avanzadas de inteligencia artificial, con el fin de estimar de manera precisa el precio de las viviendas en el portal de Idealista.

Estado del Arte y referentes

Se han identificado y consultado diversos recursos que enriquecen la comprensión en el campo de la valoración de viviendas. Entre estos recursos se incluyen:

- El conjunto de datos "Boston House Prices-Advanced Regression Techniques"² que proporciona técnicas de regresión aplicadas a los precios de viviendas.
- El conjunto de datos "A geo-referenced micro-data set of real estate listings for Spain's three largest cities"³ desarrollado por el equipo de Idealista, que ofrece una visión detallada del mercado inmobiliario en las principales ciudades de España y enriquece el contexto local, así como el entendimiento de las variables del dataset del reto Idealista.
- Precedentes estadísticos de Hedonic Pricing Models⁴, que incluyen enfoques paramétricos y no paramétricos en la valoración de viviendas.
- Tendencias y avances recientes en el campo, como la aplicación de modelos de regresión espacial y el uso de técnicas de procesamiento de lenguaje natural (NLP) y aprendizaje profundo (LLM) en el análisis de descripciones de viviendas en anuncios en línea.

Estos recursos han enriquecido significativamente el enfoque en el proyecto y han contribuido a una comprensión más completa del tema, situando el proyecto en un contexto referenciado de la valoración inmobiliaria.

Capítulo 2: Análisis Exploratorio e Ingeniería de Variables

Estadísticas de Resumen

El conjunto de datos original consta de un total de 94,815 filas y 46 variables, lo que lo convierte en un recurso completo y con una baja presencia de valores nulos. En realidad, sólo tres variables presentan datos faltantes, siendo la más relevante de ellas "CONSTRUCTIONYEAR". No obstante, hemos identificado una solución para abordar esta carencia de datos al sustituirlos por los valores disponibles en "CADCONSTRUCTIONYEAR", que proviene de fuentes externas y es más confiable.

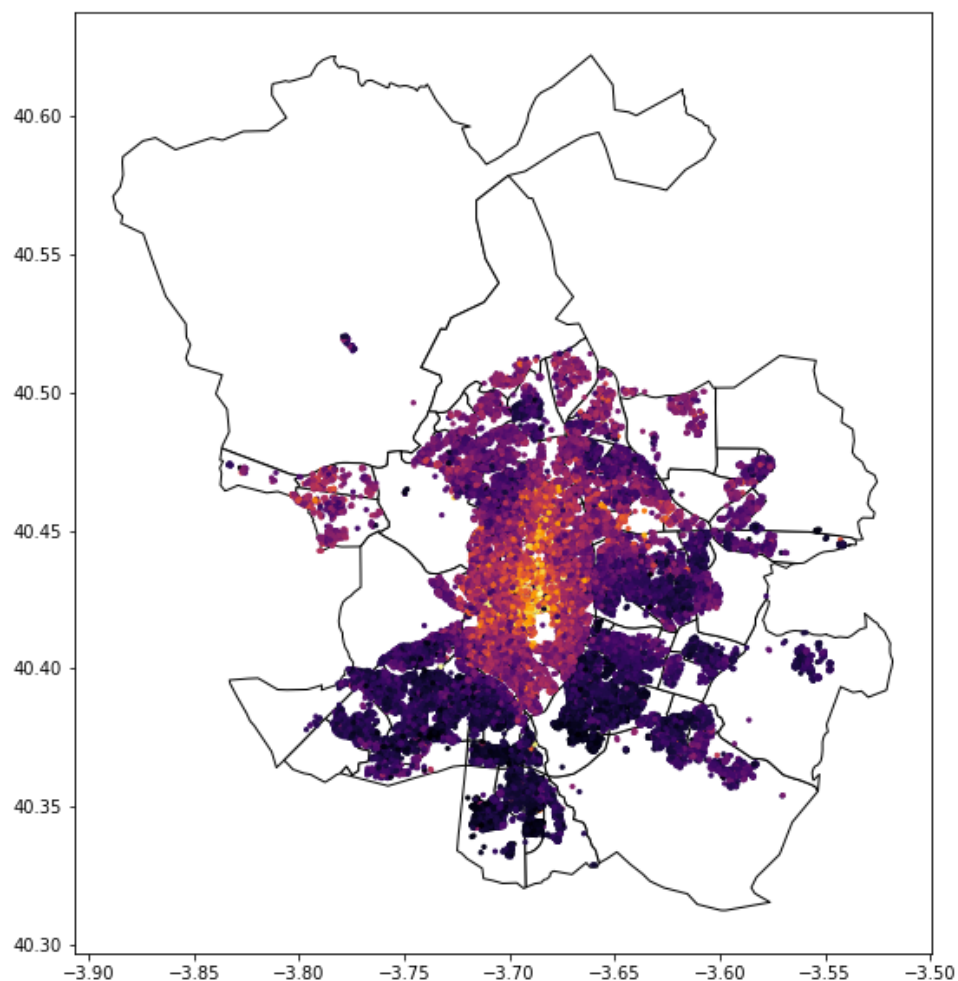
² Kaggle - House Prices - Advanced Regression Techniques
(<https://www.kaggle.com/competitions/house-prices-advanced-regression-techniques>)

³ Idealista (<https://drive.google.com/file/d/1QMZDAPFOCy95xLrBYQKSIYn4Kmb36Gi/view?usp=sharing>)

⁴ A review of hedonic pricing models in housing research
(https://www.researchgate.net/publication/287232776_A_review_of_hedonic_pricing_models_in_housing_research)

Después de resolver los valores nulos, hemos dirigido nuestra atención hacia la identificación de duplicados en el conjunto de datos. Utilizando el identificador único "ASSETID", hemos descubierto que inicialmente existen 75,804 resultados únicos bajo esta columna. Sin embargo, en consonancia con una decisión estratégica de negocio, hemos optado por conservar únicamente el registro más reciente, aplicando una ordenación de los activos de más reciente a más antiguo. Esto garantiza que nuestro conjunto de datos refleje la información más actualizada y relevante para nuestros análisis y objetivos.

Distribución de precios en Madrid



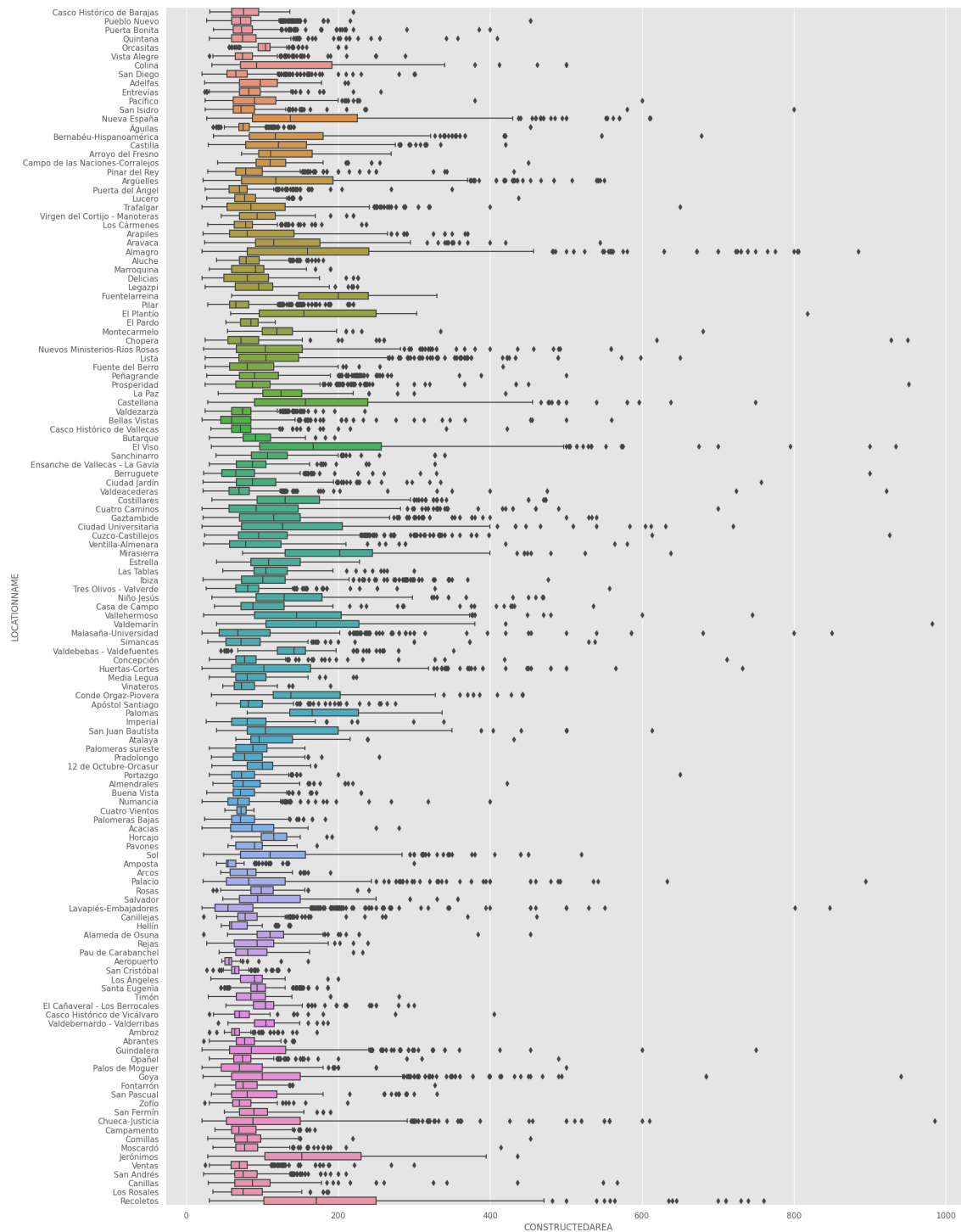
Comenzamos este proceso trazando visualmente todos los inmuebles contenidos en nuestro conjunto de datos, empleando sus coordenadas geográficas. Esto nos proporciona una primera perspectiva de la distribución espacial de los inmuebles.

Para obtener información agregada y contextualizada, fue necesario integrar nuestro conjunto de datos con otro conjunto que proporciona datos sobre los barrios de Madrid. Utilizamos la biblioteca Geopandas para llevar a cabo esta unión de datos. De esta manera, pudimos asignar cada inmueble a un barrio específico en Madrid, lo que nos permite situar cada propiedad dentro de un conjunto geográfico definido y cerrado.

En la imagen podemos ver cómo se vinculan las coordenadas de los inmuebles con los polígonos de los barrios en Madrid. Este paso es crucial para poder llevar a cabo análisis posteriores que dependen de la ubicación geográfica de los inmuebles en relación con su entorno.

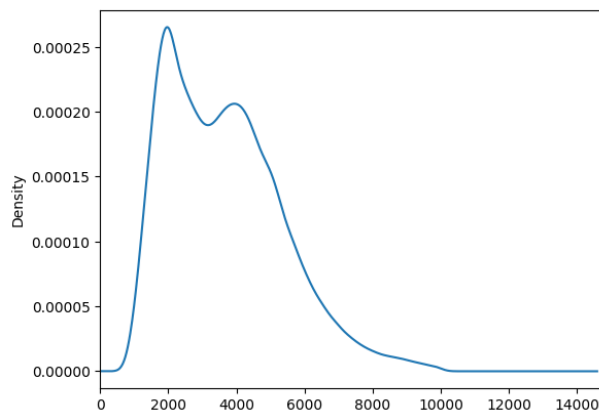
Una vez asignados los barrios podemos ver cómo se distribuyen los inmuebles dentro de cada zona según el área construida:

Distribución de área construida por barrio



La distribución sugiere que existen múltiples grupos o patrones dentro de los datos. Esto podría indicar la existencia de subconjuntos de datos con características distintas. Es crucial reconocer y tratar estas diferencias, ya que pueden influir significativamente en la eficacia del modelo.

Distribución de unitario

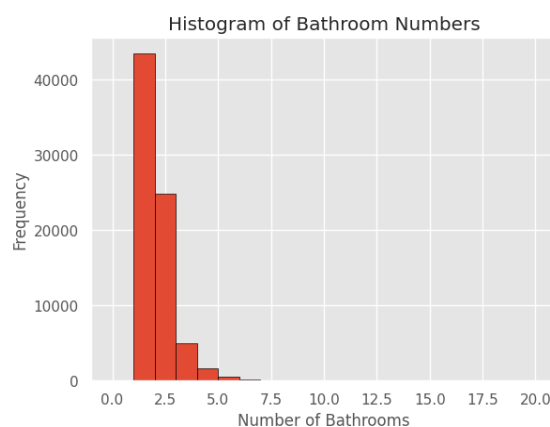


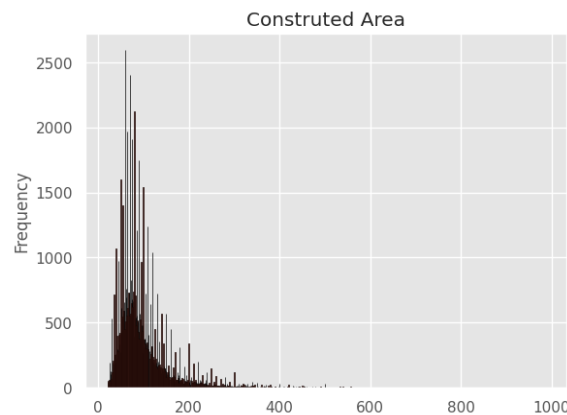
La asimetría en la distribución de valores significa que los datos no están igualmente distribuidos a ambos lados de la media. Esto puede afectar la capacidad del modelo para generalizar correctamente. Podría ser necesario aplicar transformaciones a los datos o considerar modelos más robustos que puedan manejar esta asimetría.

El amplio rango de valores en los datos indica que algunas variables pueden tener valores extremadamente altos o bajos. Esto puede llevar a desafíos en la escala de las variables y la interpretación del modelo. Normalizar o estandarizar las variables puede ser necesario para evitar que algunas características dominen sobre otras en el proceso de modelado.

Outliers y corrección de datos

En nuestro proceso de búsqueda de outliers en los datos de inmuebles, primero eliminamos aquellos registros que no cuentan con información sobre baños. Luego, realizamos un exhaustivo análisis para detectar cualquier incoherencia en la relación entre el número de baños y la superficie construida. A pesar de no encontrar situaciones aparentemente anómalas, seguimos evaluando cuidadosamente los datos para garantizar la integridad de nuestro conjunto de información.





Enfocándonos ahora en la relación entre el número de habitaciones y la superficie construida, hemos observado que esta relación exhibe una distribución asimétrica y extremadamente amplia. Para abordar esta situación, hemos implementado los siguientes pasos:

1. **Aplicamos una Regresión Lineal:** Utilizamos una regresión lineal para comprender mejor la relación entre el número de habitaciones y la superficie construida en el conjunto de datos.
2. **Análisis por Barrios:** Llevamos a cabo análisis específicos por barrios para comparar los inmuebles con su entorno. Esto nos permite evaluar cómo se comporta esta relación en diferentes áreas geográficas.
3. **Filtración de Datos no Coherentes:** Como parte de nuestra estrategia, identificamos y eliminamos aquellos registros que no muestran coherencia en la relación entre el área construida y el número de habitaciones dentro de cada barrio. Este proceso nos ayuda a mantener la integridad de nuestros datos y a garantizar que nuestras conclusiones sean sólidas y representativas.

En resumen, nuestro enfoque en esta etapa del proyecto se centra en comprender y optimizar la relación entre el número de habitaciones y la superficie construida, teniendo en cuenta la variabilidad geográfica de los inmuebles. Esto nos permitirá tomar decisiones más informadas y precisas en el proceso de análisis de datos.

Como resultado de la búsqueda de outliers, el proceso de limpieza y la eliminación de duplicados en el conjunto de datos inicial que contenía 94,815 inmuebles, hemos obtenido un conjunto de datos definitivo que consta de 74,213 inmuebles. Este conjunto de datos depurado y optimizado será la entrada principal para la construcción y entrenamiento de nuestro modelo. La reducción del conjunto de datos a 74,213 inmuebles garantiza que estemos trabajando con datos de alta calidad y relevantes para nuestros objetivos analíticos y de modelado.

Target Encoding - Locationname (MEAN)

En el proyecto, se ha aplicado la técnica de "Target Encoding" a la variable categórica "Locationname". Esta técnica implica asignar a cada categoría de "Locationname" el valor medio del atributo objetivo "unitprice" correspondiente a esa categoría. La finalidad de esta técnica es representar de manera numérica la variable categórica "Locationname" de manera que se capturen

las relaciones entre la ubicación y el atributo objetivo "unitprice". Esta representación numérica resultante se utiliza en los modelos de machine learning para mejorar la capacidad predictiva, especialmente cuando la ubicación (barrio) es un factor importante en los resultados que se buscan predecir en el proyecto.

Próximos pasos en Ingeniería de variables

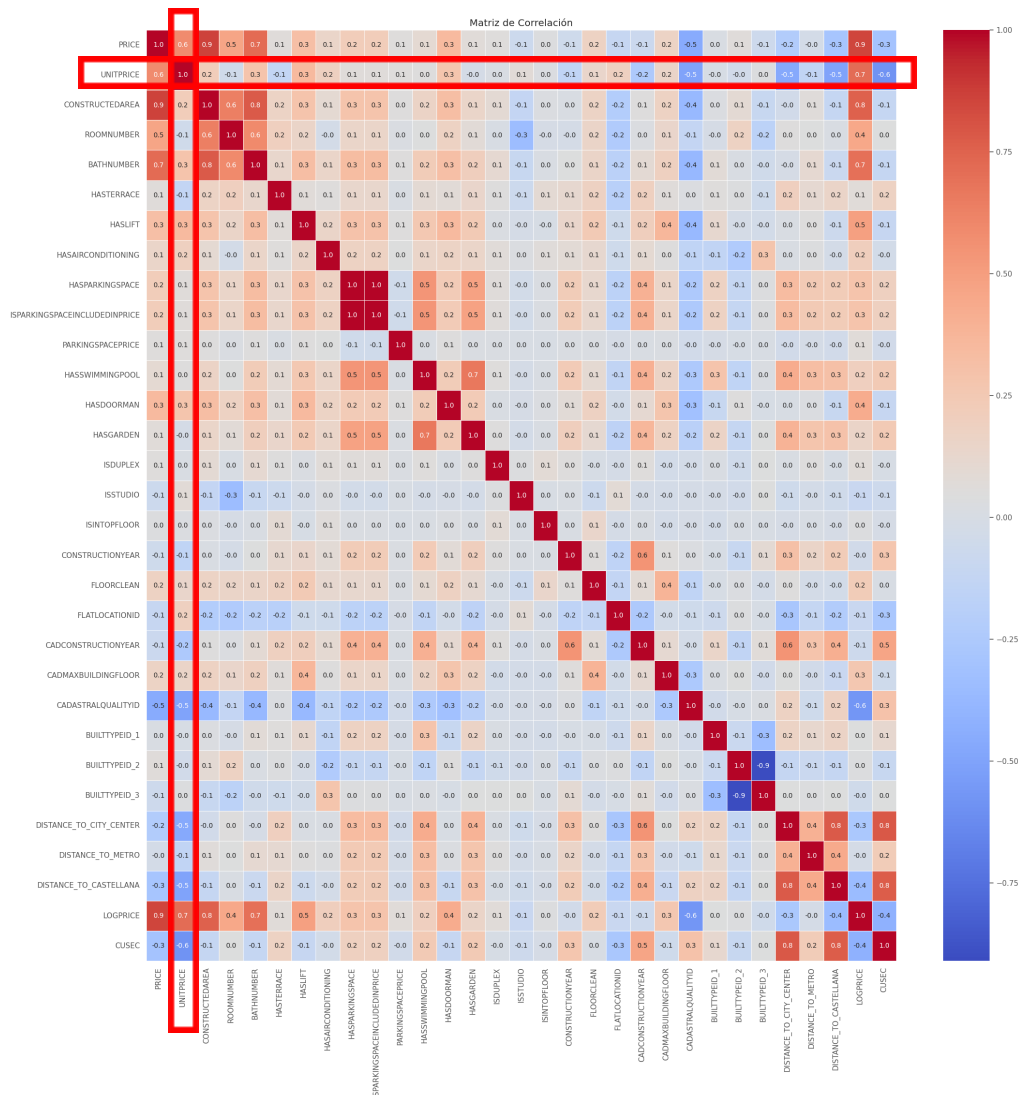
En las próximas iteraciones de la ingeniería de variables, estamos enfocados en seguir mejorando la calidad de nuestros datos y enriqueciendo nuestro conjunto de características. Aquí se resumen los pasos clave que vamos a tomar:

1. **Incorporación de Variables Espaciales:** Continuaremos asignando variables espaciales, como la "Sección Censal" y las "Zonas Homogéneas de Valor de Catastro". Estos datos nos proporcionarán información estadística sobre la población en cada área, así como detalles sobre compraventas, homogeneidad de viviendas y precios. Esto enriquecerá nuestro conjunto de datos y mejorará nuestra comprensión de los factores que afectan el valor de las propiedades.
2. **Ampliación de Regresiones Lineales:** Extendiendo el análisis de regresiones lineales más allá de "Superficie Construida", "Baños" y "Número de Habitaciones", trabajaremos en aplicar este enfoque a otras variables. El objetivo es obtener un dataset más limpio y confiable al comprender mejor las relaciones entre las distintas características.
3. **Análisis de Correlación e Ingeniería de Variables Avanzada:** Realizaremos un análisis de correlación más detallado para identificar relaciones y patrones complejos entre las variables. Además, seguiremos explorando oportunidades de ingeniería de variables más avanzada para mejorar la precisión de nuestros modelos.
4. **Prueba con Variables Descartadas:** Revisaremos las variables previamente descartadas para evaluar si alguna de ellas puede ser útil después de todo o si se pueden transformar de manera significativa para ser incluidas en el modelo.
5. **Homogeneidad de Zonas Urbanas:** Continuaremos investigando la homogeneidad de las zonas urbanas para comprender cómo afecta al mercado de propiedades y cómo podemos reflejar esto en nuestras características.

Estos pasos nos llevarán hacia un conjunto de datos más completo y sofisticado, lo que a su vez mejorará la precisión de nuestros modelos y nos permitirá tomar decisiones más informadas en el proyecto.

Matriz de Correlación

Una vez limpio el dataset aplicamos una matriz de distribución de manera que veamos cómo influye cada una de las variables en las demás. Teniendo en cuenta que nuestra variable objetivo será el valor unitario de los inmuebles:



La dispersión es muy grande, no hay variables que estén significativamente relacionadas con el valor unitario una vez obviadas aquellas variables directamente dependientes de éste como puede ser el precio o el precio logarítmico.

En resumen, para construir un modelo efectivo en estas condiciones, es esencial comprender y abordar la multimodalidad, la asimetría y el rango de valores. Se deben explorar técnicas de preprocesamiento de datos, selección de características y elección de modelos que sean adecuados para manejar estas características específicas de los datos. Además, es fundamental realizar un análisis exploratorio detallado para descubrir patrones y relaciones dentro de los datos que puedan guiar la construcción del modelo de manera efectiva.

Escala de las variables

De cara a igualar la escala de la variable objetivo y las de las variables de entrada del modelo decidimos aplicar la escala logarítmica tanto al valor unitario como al valor unitario medio de los barrios para que, a la hora de introducir las variables en el modelo sean más homogéneas.

Variables definitivas

Una vez eliminadas las variables redundantes con la variable objetivo se han elegido aquellas variables que, o bien tienen una correlación alta en la matriz y aquellas que por experiencia de negocio son relevantes a la hora de realizar un modelo de valoración inmobiliaria.

Se retuvieron las variables "ASSETID" y "LOCATIONNAME" con fines estadísticos posteriores.

#	Column	Non-Null	Count	Dtype
---	-----	-----	-----	-----
0	ASSETID	74213	non-null	object
1	CONSTRUCTEDAREA	74213	non-null	int64
2	ROOMNUMBER	74213	non-null	int64
3	BATHNUMBER	74213	non-null	int64
4	HASLIFT	74213	non-null	int64
5	ISDUPLEX	74213	non-null	int64
6	CADCONSTRUCTIONYEAR	74213	non-null	int64
7	BUILTTYPEID_1	74213	non-null	int64
8	BUILTTYPEID_2	74213	non-null	int64
9	BUILTTYPEID_3	74213	non-null	int64
10	DISTANCE_TO_CITY_CENTER	74213	non-null	float64
11	DISTANCE_TO_METRO	74213	non-null	float64
12	DISTANCE_TO_CASTELLANA	74213	non-null	float64
13	LOCATIONNAME	74213	non-null	object
14	UNIT_PRICE_LOG	74213	non-null	float64
15	LOCATION_MEAN_UNITPRICE_log	74213	non-null	float64

Capítulo 3: Modelado

Por motivos operacionales se ha decidido ir a probar el modelo que a priori mejor funcionará en este tipo de datos: el Random Forest. No por ello dejaremos de probar con otros dos métodos a modo de comprobación: Regresión Lineal y XGBoost

1. Random Forest Regression

Para mejorar el rendimiento de nuestro modelo Random Forest, llevaremos a cabo la búsqueda de hiperparámetros utilizando dos enfoques diferentes. Esto nos permitirá encontrar la configuración óptima que maximice la precisión de nuestras predicciones.

→ Random Forest cross validation K-Folds + Random Search

Descripción: En esta etapa de afinamiento de hiperparámetros, utilizamos el método de Búsqueda Aleatoria para encontrar la configuración óptima de hiperparámetros para nuestro modelo Random Forest con Validación Cruzada K-Folds. El proceso consiste en que el usuario establece un número de iteraciones para probar el modelo. En cada iteración, el algoritmo selecciona de manera aleatoria un conjunto de hiperparámetros de una cuadrícula predefinida (sin reemplazo) y calcula las métricas de pérdida correspondientes. Cuantas más iteraciones se realicen, mayores serán las probabilidades de encontrar los mejores hiperparámetros, pero esto también aumentará el tiempo de ejecución del algoritmo.

Ventajas:

- **Eficiencia Relativa:** Este método es más rápido en comparación con otros enfoques de búsqueda exhaustiva de hiperparámetros, lo que lo hace adecuado para ajustar modelos en un tiempo razonable.
- **Paralelización:** Permite la paralelización, lo que significa que se pueden probar varios conjuntos de hiperparámetros simultáneamente, acelerando aún más el proceso.
- **Exploración Aleatoria:** La aleatoriedad en la selección de hiperparámetros puede conducir a la exploración de un espacio más amplio, lo que aumenta las posibilidades de encontrar soluciones efectivas.

Desventajas:

- **Subóptimos con Pocas Iteraciones:** Con un número insuficiente de iteraciones, existe la posibilidad de que el algoritmo seleccione hiperparámetros subóptimos y no logre encontrar la mejor configuración.
- **Sin Garantía de Óptimo Global:** A diferencia de la búsqueda exhaustiva, no se garantiza que la Búsqueda Aleatoria encuentre la configuración de hiperparámetros óptima, ya que depende de la aleatoriedad en la selección.

Este enfoque combina la eficiencia de la Búsqueda Aleatoria con la validación cruzada K-Folds para ajustar eficazmente los hiperparámetros de nuestro modelo Random Forest, mejorando así su rendimiento en nuestro conjunto de datos específico.

Resultado:

```
{'n_estimators': 200,
 'min_samples_split': 2,
 'min_samples_leaf': 4,
 'max_features': 'sqrt',
 'max_depth': None,
 'bootstrap': False}
```

Métricas Generales

```
RMSE (Test): 0.19
MAE (Test): 0.13
R2 (Test): 0.85
MPE_media: 1.69%
```

→ Random forest cross validation K-Folds + Bayesian Search

Descripción: En esta etapa de modelado, implementamos una estrategia de búsqueda bayesiana para optimizar los hiperparámetros de nuestro modelo Random Forest. La búsqueda bayesiana utiliza información de iteraciones anteriores para tomar decisiones informadas sobre qué valores de hiperparámetros explorar a continuación. Esto acelera la convergencia hacia una solución óptima.

Ventajas:

- Aprendizaje de Iteraciones Anteriores: La búsqueda bayesiana se beneficia de las lecciones aprendidas en iteraciones anteriores, lo que permite una convergencia más rápida hacia soluciones óptimas.

Desventajas:

- Limitaciones en la Paralelización: En general, la búsqueda bayesiana no se beneficia de la paralelización, lo que puede aumentar significativamente el tiempo necesario para encontrar una solución.
- Riesgo de Máximo Local: Al igual que con todos los métodos bayesianos, existe el riesgo de converger en un máximo local en lugar de encontrar el máximo global, lo que podría limitar la calidad de la solución final.
- Esta combinación de Random Forest con Validación Cruzada K-Folds y Búsqueda Bayesiana nos permite obtener un modelo más preciso y eficiente, optimizando los hiperparámetros de Random Forest para nuestro conjunto de datos específico.

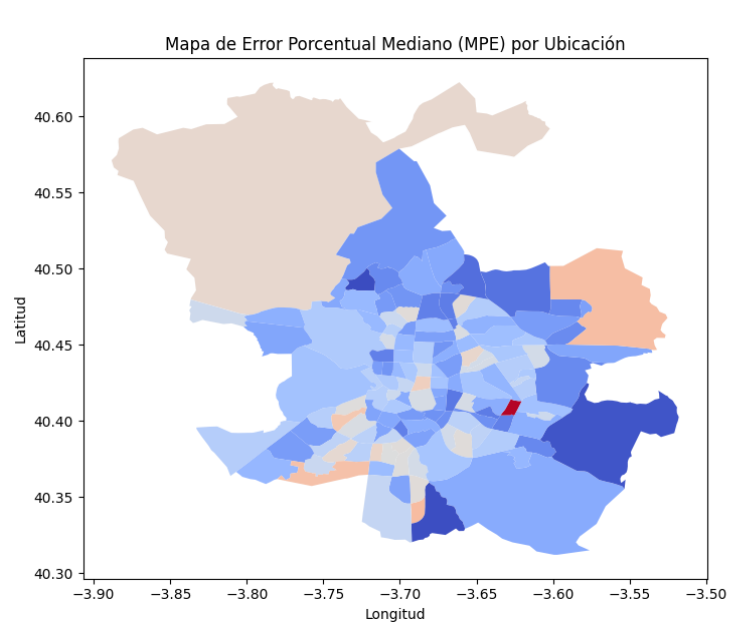
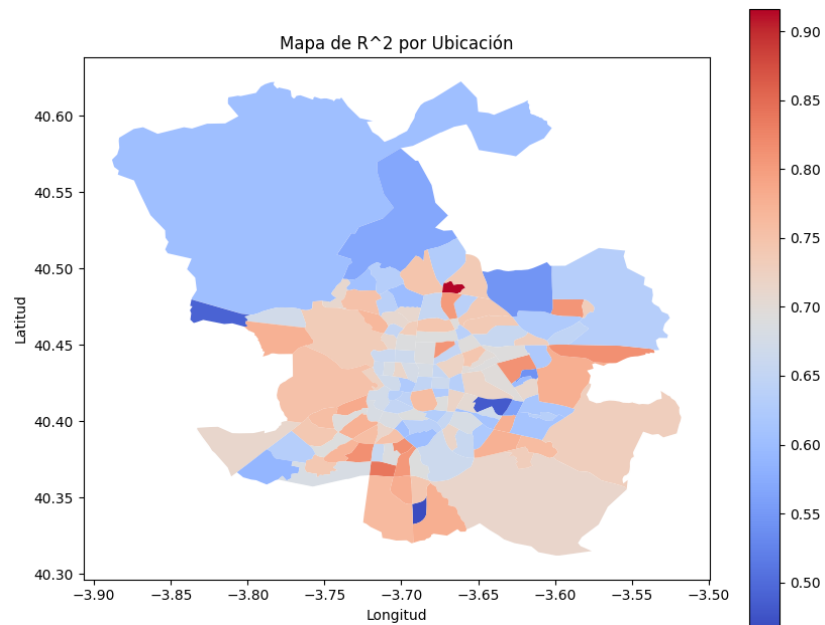
Resultado:

```
'bootstrap', False),
('max_depth', 34),
('max_features', 0.6857565539493793),
('min_samples_leaf', 4),
('min_samples_split', 2),
('n_estimators', 300)])
```


Métricas Generales

RMSE (Test): 0.19
MAE (Test): 0.13
R2 (Test): 0.85

Para comprobar la coherencia del modelo en cada uno de los barrios y dada la gran segmentación de los datos se hace una comprobación de las métricas particularizada para cada una de las zonas:



2. Regresión Lineal

La regresión lineal es una técnica fundamental en el campo de la estadística y el aprendizaje automático que se utiliza para comprender y modelar la relación entre unas variables independientes y una variable. Trata de establecer una relación lineal entre estas variables, lo que significa que se busca encontrar una línea recta que mejor se ajuste a los datos observados.

Métricas Generales:

```
MSE del modelo: 0.0548
R^2 del modelo: 0.7725
Media del Error Porcentual: 2.17%
```

3. XGBoost

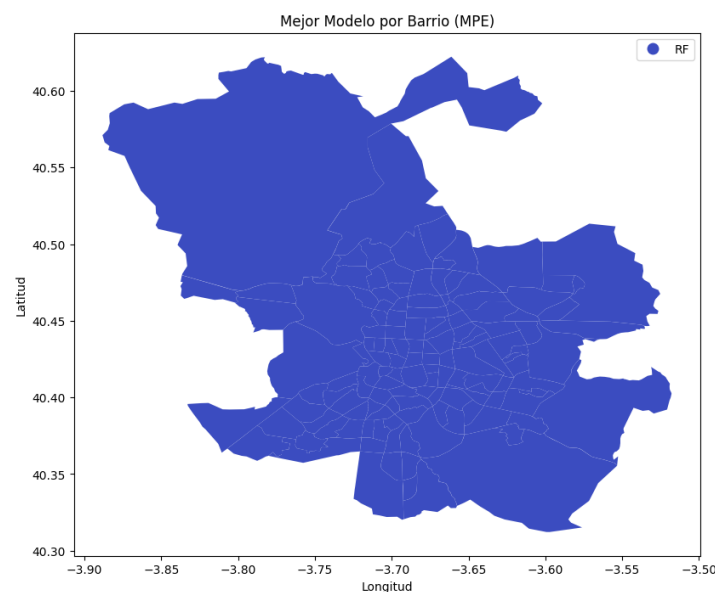
Para el modelo XGBoost se ha lanzado una Bayesian Search con el fin de localizar los hiperparámetros que mejor ajustan a este conjunto de datos, dándonos como resultado:

```
(['colsample_bytree', 0.8311563895216271),
('gamma', 1),
('learning_rate', 0.6020666067370067),
('max_depth', 40),
('min_child_weight', 6),
('n_estimators', 74),
```

Métricas Generales:

```
MAE (Test): 0.15
R2 (Test): 0.82
Rmse_test_xgb: 0.21
Media del Error Porcentual: 1.93%
```

Una vez lanzados los tres modelos y aunque aparentemente y en el aspecto general funciona mejor el Random Forest, se hace una comparativa a nivel barrio para ver qué modelo funciona mejor en cada una de las zonas



Selección final de modelo

Se ratifica el modelo de Random Forest como el más adecuado para la estimación del valor de las viviendas en todos los barrios, por tanto se lleva a producción.

Discusión

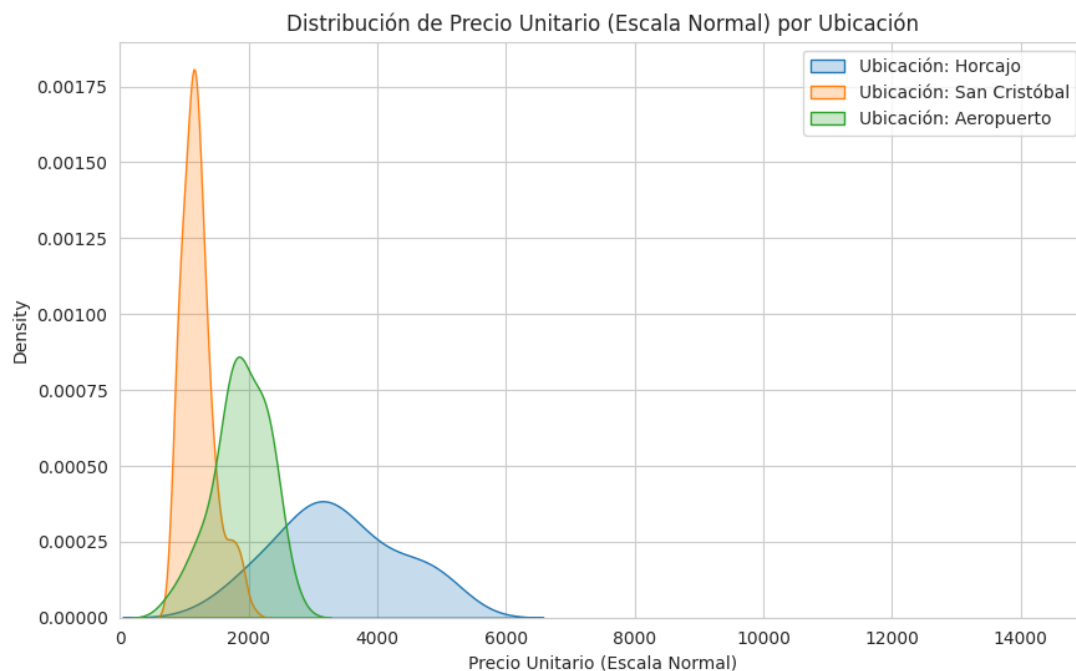
El modelo de Random Forest funciona bien en general, pero al entrar en la distribución por barrios vemos que hay zonas en las que el modelo se degrada bastante ([Ver notebook](#))

Ubicación: Horcajo
R²: 0.5632
MPE: 11.97%
Población de Muestras: 29

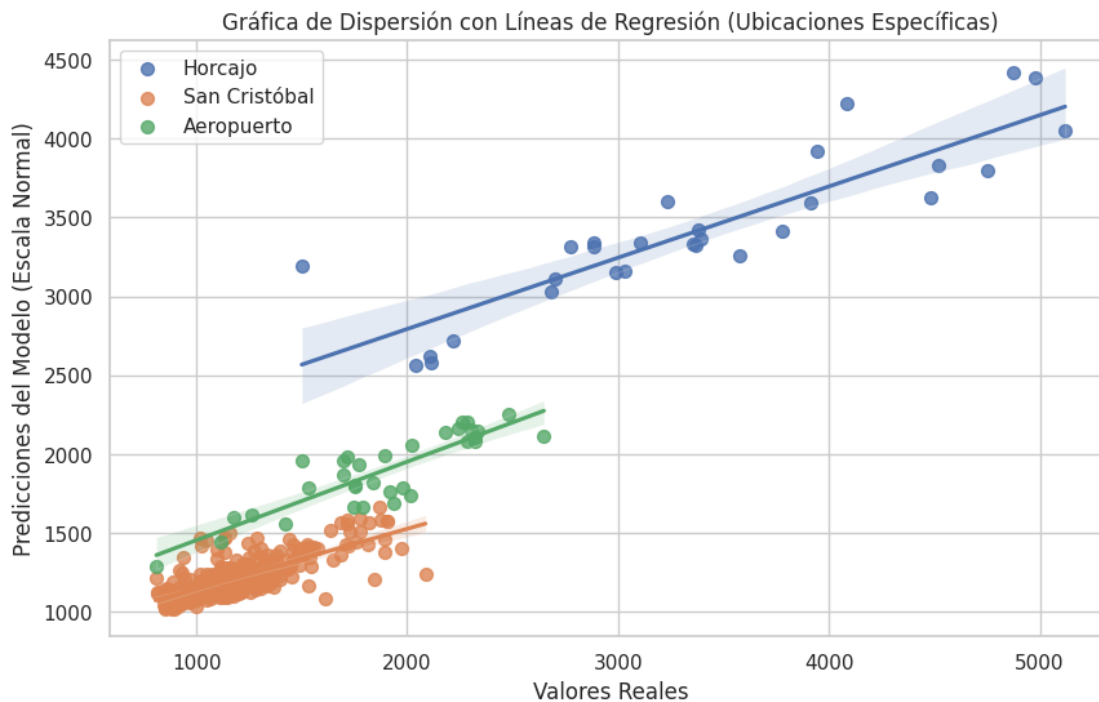
Ubicación: San Cristóbal
R²: 0.4670
MPE: 9.28%
Población de Muestras: 263

Ubicación: Aeropuerto
R²: 0.6320
MPE: 9.21%
Población de Muestras: 33

Y estas son sus distribuciones de Unitario:



Así como las gráficas de dispersión de valores reales frente a valores que predice el modelo.



Tenemos en común la pequeña población de las muestras, deberíamos pensar en la siguiente iteración hacer un modelo de gemelos urbanos de cara a aumentar la cantidad de muestras aunque sea en entornos diferentes, con características similares, además las nubes de punto son muy dispersas.

Vamos a intentar en la siguiente iteración hacer modelos específicos para cada uno de los barrios de manera que los hiperparámetros sean concretos para cada una de las zonificaciones y de esta manera combatir la disparidad de las muestras desde el entrenamiento

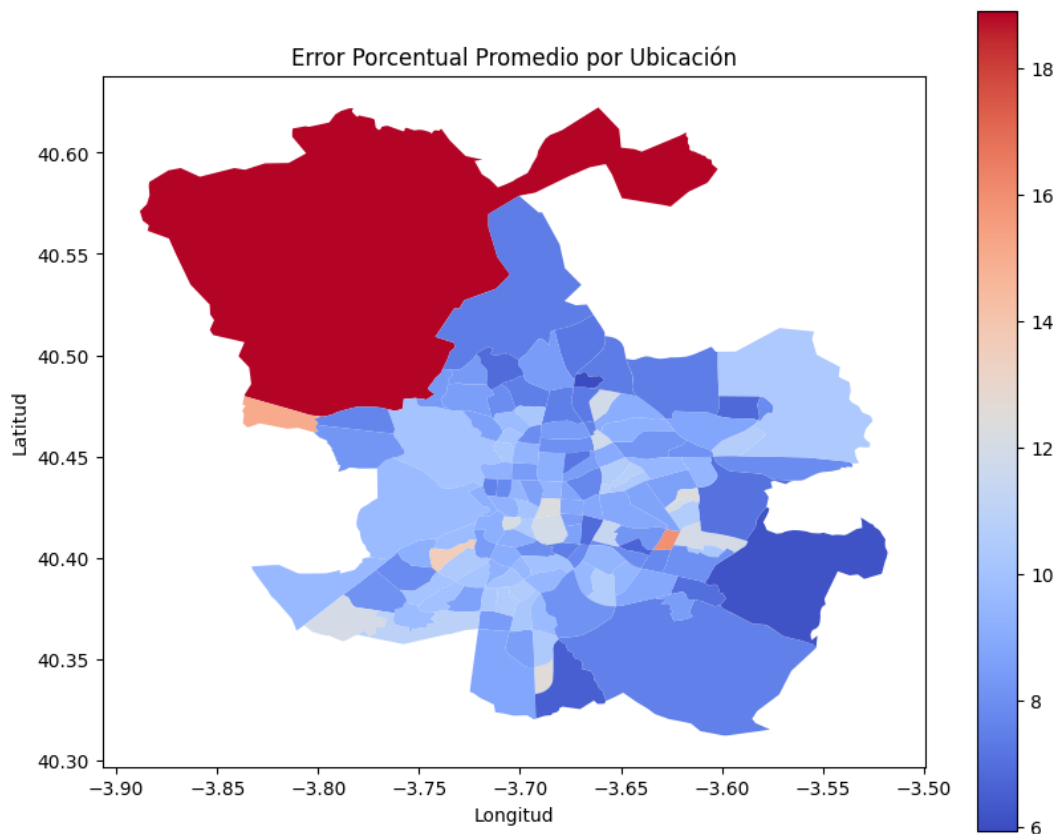
Modelado 2.0

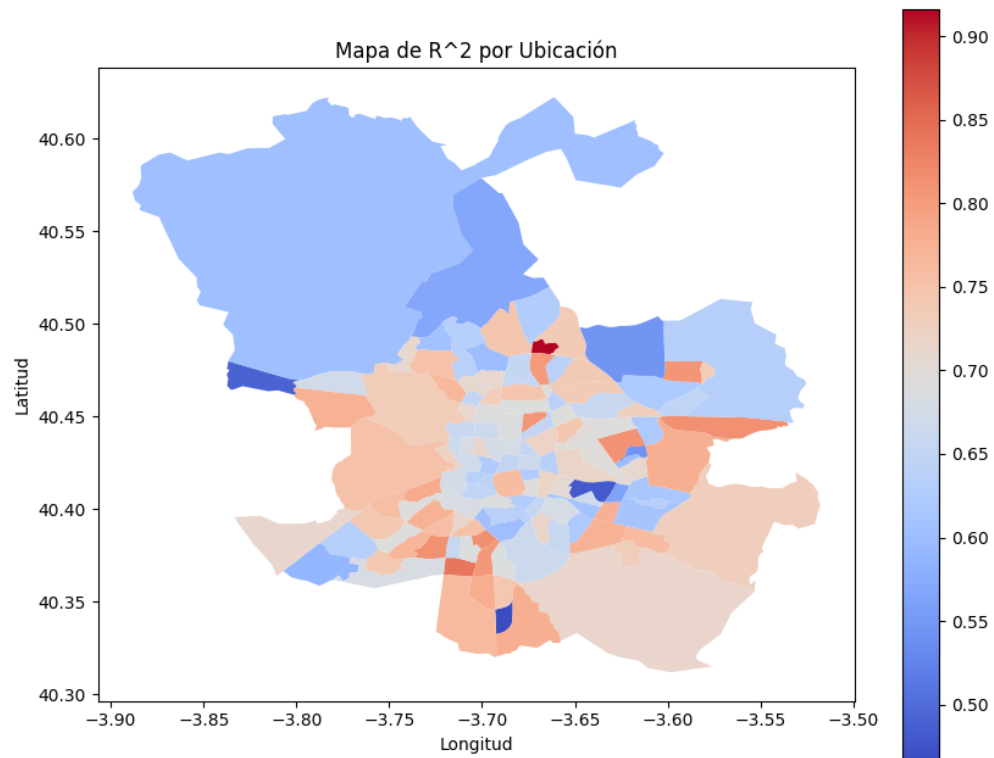
En el marco del proceso de mejora continua del producto, hemos dado inicio a la segunda iteración del modelado. En esta fase, estamos implementando enfoques específicos para cada uno de los barrios en los que opera nuestro sistema.

- Modelos Específicos para Cada Barrio:** Para llevar a cabo esta iteración, estamos generando modelos de manera individualizada para cada uno de los barrios que componen nuestra área de operación. Este enfoque nos permitirá adaptar mejor el sistema a las particularidades de cada ubicación y, en última instancia, mejorar la calidad de nuestras predicciones y recomendaciones.
- Bayesian Search para Cada Ubicación:** Como parte del proceso, hemos vuelto a iniciar el proceso de búsqueda bayesiana de hiperparámetros. Sin embargo, a diferencia de la iteración anterior, esta vez estamos realizando la búsqueda de hiperparámetros de manera independiente para cada uno de los barrios. Esta estrategia nos brinda la capacidad de afinar los modelos de manera más precisa y específica a las necesidades de cada ubicación.

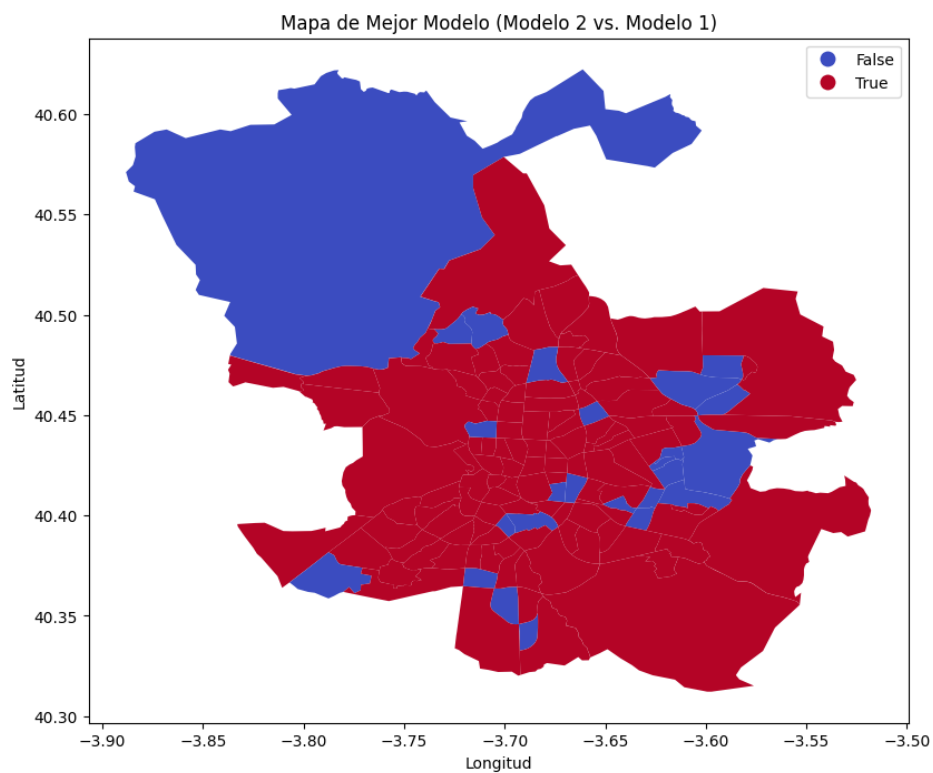
- **Entrenamiento con los Mejores Hiperparámetros por Barrio:** Una vez que hemos identificado los hiperparámetros óptimos para cada barrio mediante la búsqueda bayesiana, procedemos a entrenar los modelos utilizando estos ajustes personalizados. Esto asegura que cada modelo esté optimizado para su entorno específico y maximiza su capacidad predictiva.
- **Evaluación y Comparación de Métricas:** Finalmente, evaluamos el rendimiento de cada modelo en términos de métricas relevantes para cada barrio. Luego, comparamos los resultados obtenidos en esta iteración con los del modelo original basado en Random Forest. Esta comparación nos permitirá medir el impacto de nuestras mejoras y determinar si hemos logrado un avance significativo en la calidad de las predicciones.

Aquí tenemos los resultados de Error porcentual medio y R2 en cada uno de los barrios:





Haciendo un análisis a través del R^2 podemos determinar que en general el segundo modelo funciona mejor que el primero, pero en algunos casos el modelo original es más preciso, por lo que se tendrán que unir ambos modelos de cara a perfeccionar los resultados.



Capítulo 4: Productivización

AI Idealista - App

Después de completar la fase de Modelado y Evaluación, hemos desarrollado una aplicación web destinada a los usuarios de Idealista. Esta herramienta digital tiene como objetivo acercar el modelo desarrollado a los usuarios, permitiéndoles utilizarlo de manera efectiva. Los usuarios pueden proporcionar información sobre sus propiedades y ejecutar el modelo en tiempo real para obtener una estimación del valor de sus viviendas.

La aplicación cuenta con una barra lateral en la que los usuarios ingresan los detalles de su propiedad respondiendo a una serie de preguntas mediante widgets de entrada de la librería Streamlit. Estas preguntas incluyen información como el área en metros cuadrados, el número de habitaciones, el número de baños, la disponibilidad de ascensor, si es un dúplex, el año de construcción, si es una construcción nueva y otras características relevantes.


Estas variables proporcionadas por los usuarios son esenciales, ya que corresponden a 7 de las 10 variables necesarias para que el modelo realice una predicción precisa sobre el valor de la propiedad. Las tres variables restantes serán inferidas automáticamente por el código de la aplicación: la distancia al centro de la ciudad, la distancia al metro y la distancia a la Castellana G, así como el precio unitario promedio de la ubicación. Extenderemos este tema más adelante.

Una vez que el usuario ha completado la entrada de datos y hace clic en el botón "Predecir", la aplicación proporciona al usuario:

- Una estimación del precio total de la propiedad
- Una estimación del precio por metro cuadrado de la propiedad
- Un indicador de confianza en la predicción (Bajo, Medio, Alto).
- Una visualización de los datos ingresados.
- La geolocalización de la propiedad.
- Una lista de atributos adicionales que pueden influir en el precio de la propiedad, como la cédula de habitabilidad, la climatización, terraza o balcones, vistas, y luz natural.
- Información sobre la ventaja de anunciar la propiedad en la plataforma Idealista.
- Una breve explicación sobre el funcionamiento de la aplicación.

En resumen, el usuario puede obtener toda esta información en aproximadamente 2 minutos desde que ingresa a la aplicación hasta que recibe los resultados finales.

App AI-Idealista



Cuéntanos más sobre tu vivienda

- Ingresa la información sobre tu Inmueble y haz click en PREDICT
- Si no conoces un valor, no te preocupes, deja el valor que está por default

Datos Generales

¿Cuántos(m²) tiene tu inmueble?

150

¿Cuántas habitaciones tiene tu inmueble?

2

¿Cuántos baños tiene tu inmueble?

3

¿Tu finca tiene ascensor?

Si

¿Es Duplex?

No

Año de construcción de tu finca, si no sabes deja 1985

1500 1985 2023

Tipo de vivienda

¿Tu inmueble es una construcción nueva?

Si

¿Tu inmueble es de segunda mano sin remodelar?

No

¿Tu inmueble es de segunda mano en buenas condiciones?

No

Ubicación

Pick one

Abrantes

City

Madrid

Province

Madrid

Street

Gran via 84

Country

Spain

Predict

AI Idealista

La información de tu inmueble aparecerá aquí después de que hayas introducido los datos de tu inmueble

Estimación del precio total:

€ 318744

El valor del m2 de tu vivienda es:


€ 2125

Nivel de confianza:

Alto

Collected Data:

	CONSTRUCTEDAREA	ROOMNUMBER	BATHNUMBER	HASLIFT	ISDUPEX	CADCONSTRUCTIONYEAR
0	150	2	3	1	0	1,985



Estos son algunos atributos extra que pueden influenciar positiva o negativamente el precio de tu inmueble:

- El inmueble dispone de cédula de habitabilidad
- El inmueble tiene climatización
- El inmueble tiene terraza o balcones
- El inmueble tiene buena vista
- El inmueble tiene iluminación natural

Sube la vivienda al portal inmobiliario de idealista

Idealista es el mayor escaparate inmobiliario de España. Según datos de SimilarWeb en el mes de diciembre de 2022 el portal tuvo más de 43 millones de visitas, muy por encima de sus competidores. Si quieres que tu piso llegue a todos los compradores posibles, no puedes omitir este paso. Además, la agencia inmobiliaria que te acompañe en el proceso será la encargada de gestionar todas las visitas del inmueble y así, vender la vivienda en tiempo récord.

Acerca de esta app

Esta app ha sido desarrollada como trabajo final del MDS de The Valley por Alejandro López y Leonardo Velásquez bajo la dirección de David Rey, CDO de Idealista.

Para mas información [see the code.](#)

Inferencia automática de variables de distancia a puntos urbanos

Dado que el modelo requiere tres variables de distancia a puntos clave de interés en la ciudad de Madrid (1) Distancia a la Castellana, (2) Distancia al Metro y (3) Distancia al Centro, y dado que los usuarios rara vez tendrán acceso a estos valores, tuvimos como desafío el la inferencia de estas variables.

Al ser variables geográficas que dependen de redes de puntos y líneas de distancia, principalmente medidas de distancia euclidiana, desarrollar un modelo que calcula automáticamente estas variables utilizando la ubicación proporcionada por el usuario como entrada era un desafío que no podíamos abordar dentro del plazo previsto para el proyecto. Por lo tanto, ideamos una solución provisional para esta primera iteración de la aplicación. Esta solución implica el uso de una base de datos que contiene el valor promedio de estas variables por barrio. El script de Python consulta esta base de datos y utiliza el barrio indicado por el usuario como punto de referencia para proporcionar estas variables al modelo.

Además, aprovechamos esta misma base de datos para suministrar al modelo la variable *LOCATION_MEAN_UNIT_PRICE*, que representa el valor promedio del metro cuadrado en escala logarítmica por barrio. Esta variable también es relevante para el funcionamiento del modelo.

	LOCATIONNAME	LOCATION_MEAN_UNITPRICE_log	DISTANCE_TO_CITY_CENTER_M	DISTANCE_TO_METRO_M	DISTANCE_TO_CASTELLANA_M
0	12 de Octubre-Orcasur	7.435673	5.214738	0.478773	2.360633
1	Abrantes	7.534942	4.622853	0.276697	3.148402
2	Acacias	8.312837	1.655306	0.360338	0.944683
3	Adelfas	8.340179	3.271835	0.258131	1.907254
4	Aeropuerto	7.538134	11.237741	1.073053	9.424220

En futuras iteraciones del modelo, se plantea calcular dichas variables geográficas a través de la API de Google Maps.

Streamlit

Una vez que se entrenó y se identificó el mejor modelo Random Forest, se exportó en un archivo en formato (.pkl) utilizando la biblioteca Joblib. Esto permitió que el modelo fuera cargado y activado en una aplicación web.

Para desarrollar este producto digital, se utilizó la biblioteca Streamlit, una herramienta de código abierto que permite a los desarrolladores crear aplicaciones web interactivas de manera rápida y sencilla utilizando Python. Esto facilita que los usuarios pudieran utilizar la aplicación en tiempo real.

Se desarrolló un script en Python que se encarga de orquestar la lógica de la aplicación, incluyendo la carga de las dependencias, la base de datos de barrios, la carga del modelo y la lógica detrás de la aplicación Streamlit.

Puedes consultar el [código Python aquí](#).

Transparencia

Dado que esta aplicación está dirigida a usuarios y proporciona una predicción del valor de inmuebles, lo cual es un dato sensible que puede tener un impacto significativo en el usuario, es esencial presentar esta predicción junto con una estimación de confianza. Para abordar esto, hemos realizado una distribución por cuartiles del (R^2) del modelo por cada barrio. Luego, hemos definido una escala de tipo semáforo en la cual la predicción se considera de baja confianza si el R^2 se encuentra en el rango del 0% al 49% de la distribución, de confianza media si está en el rango del 50% al 74%, y de alta confianza si se sitúa entre el 75% y el 100%.

Esta escala se ha incorporado en la base de datos que el script de Python carga en la aplicación Streamlit al inicio, y la aplicación la consulta tomando como referencia el barrio proporcionado por el usuario. Esto asegura que las predicciones sean acompañadas de una indicación de confianza que ayudará a los usuarios a interpretar y utilizar la información de manera adecuada.

Productivización 2.0

Estas son una lista de posibles próximas iteraciones clave para mejorar y optimizar nuestra aplicación:

- **Integrar un llamado a la API de geolocalización de Google Maps:** Implementaremos la capacidad de llamar a una API de google maps para calcular automáticamente las distancias a puntos de interés en la ciudad, como la Castellana, el Metro y el Centro. Esto mejorará la precisión de nuestras predicciones y facilitará la experiencia del usuario.
- **Entregar un cálculo de intervalo de confianza al usuario:** Vamos a proporcionar al usuario un cálculo de intervalo de confianza junto con las predicciones. Esto permitirá que los usuarios comprendan mejor la fiabilidad de las estimaciones y tomen decisiones más informadas.
- **Monitorización y mantenimiento:** Estableceremos un sistema de monitorización continuo para supervisar el rendimiento de la aplicación y realizar mantenimiento proactivo. Esto garantizará que la aplicación funcione de manera óptima en todo momento.
- **Actualizar los modelos con bases de datos del 2022-3:** Mantendremos nuestros modelos actualizados mediante la incorporación de datos de propiedades actualizados del año 2022. Esto garantizará que nuestras predicciones reflejen con precisión las condiciones del mercado inmobiliario más recientes.
- **Reinforcement Learning:** Exploraremos oportunidades para implementar técnicas de aprendizaje por refuerzo (Reinforcement Learning) para mejorar la precisión del modelo y su capacidad de adaptación a cambios en el entorno.

Estos pasos nos ayudarán a mejorar y evolucionar nuestra aplicación, brindando a los usuarios una experiencia más confiable y valiosa en la búsqueda de información sobre el valor de las propiedades inmobiliarias.