Proyecto Barcelona Rent Predict

Horacio Bleda Martínez

October 24, 2023

1 Informe del Proyecto de Predicción Precio Alquiler en Barcelona ciudad

1.1 Presentación de datos

Introducción:

En la ciudad de Barcelona, el mercado del alquiler ha sido durante mucho tiempo un tema candente y de creciente interés tanto para residentes como para inversores. En las últimas décadas, Barcelona ha experimentado variaciones significativas en los precios de alquiler, influenciadas por una amalgama de factores, incluyendo el turismo, la inversión extranjera y los cambios en las políticas de vivienda. A través de la lente de la ciencia de datos, este proyecto busca crear una herramienta que nos permita "predecir" cúal será el precio de este alquiler según las características de la vivienda y el barrio donde queremos vivir.

Los precios del alquiler en Barcelona no solo son un indicador crucial de la economía local, sino que también reflejan la accesibilidad y asequibilidad de la vida en la ciudad para sus habitantes. En un contexto donde la gentrificación y el desplazamiento de residentes a áreas más asequibles se están convirtiendo en cuestiones prevalentes, un análisis minucioso de los datos de alquiler se convierte en una herramienta indispensable para comprender y, eventualmente, abordar las problemáticas asociadas con el mercado de alquiler.

Este estudio tiene una misión clara: no sólo queremos adentrarnos en las oscilaciones de los precios de alquiler en Barcelona, sino que también vamos a construir un modelo que prediga estos precios. En lugar de quedarnos en la teoría, mostraremos en la práctica cómo este modelo podría ser usado en el día a día, desarrollando una situación real de puesta en servicio.

1.2 Características Generales

1.2.1 Tipología de Datos:

- **barrio, distrito**:
  - Tipo: Categóricos.
  - Descripción: Se refieren a ubicaciones geográficas y no tienen un orden inherente ni una cuantificación numérica.
- **price**:
  - Tipo: Numérico, continuo.
  - Descripción: Representan el precio de alquiler y pueden tomar cualquier valor en un rango.

- **size**:
  - Tipo: Numérico, continuo.
  - Descripción: Indican el Superficie de la propiedad en metros cuadrados.
- **rooms, bathrooms**:
  - Tipo: Numérico, discreto.
  - Descripción: Indican la cantidad de habitaciones y baños y toman valores enteros.

1.2.2 Sector:

- **Inmobiliario / Bienes Raíces**
  - Descripción: Los datos pertenecen al sector inmobiliario, específicamente al subsector de alquiler de propiedades.

1.2.3 Fuente:

La fuente final de datos para el modelo proviene de la integración de dos conjuntos de datos independientes. Inicialmente, obtuve 2450 registros a través de una API proporcionada por Idealista, que se redujeron a 1726 tras eliminar duplicados. Adicionalmente, utilicé un dataset de un proyecto alojado en GitHub (https://github.com/aayzaa/barcelona_rental_prices), con fecha de diciembre de 2020, que contenía las mismas variables, a excepción del barrio. Los precios en este segundo conjunto de datos se han actualizado a julio de 2023, con un incremento del casi 24%, según fuentes consultadas.

1.2.4 Contexto:

Los datos han sido recogidos en un contexto que busca explorar las características y precios de propiedades de alquiler en distintas zonas de Barcelona. La presencia de variables como el barrio y distrito indica la intención de analizar cómo la ubicación geográfica influencia el precio y otras características de las propiedades de alquiler.

1.2.5 Definición de las Variables

En el conjunto de datos destinado a prever el precio de alquiler de propiedades en Barcelona mediante un modelo de regresión lineal, se presentan variables críticas que teóricamente ejercen una influencia notable en la variable objetivo, el precio. La ubicación, representada por las variables "Barrio" y "Distrito", es fundamental, ya que las diferencias socioeconómicas y de demanda entre las zonas pueden manifestarse en discrepancias de precios significativas. "Superficie", "Habitaciones" y "Baños", respectivamente, se correlacionan con la utilidad y capacidad de la propiedad, afectando directamente al valor del alquiler. En el modelo regresivo lineal, estas variables se asocian de manera matemática con el precio a través de coeficientes, que se ajustarán para minimizar el error en las predicciones del alquiler, permitiendo, así, una comprensión más profunda de cómo cada factor contribuye al precio del alquiler y facilitando predicciones fundamentadas para futuras propiedades en el mercado barcelonés.

1.2.6 Objetivos Clave del Proyecto

1. **Desarrollar un Modelo Predictivo para Estimar Precios de Alquiler:**
   - **Meta General:**

- Construir un modelo basado en regresión lineal que pueda anticipar los precios de alquiler en Barcelona, utilizando variables como barrio, Superficie y número de habitaciones y baños.

- **Evaluación del Modelo:**
  - Verificar la precisión del modelo comparando sus predicciones con precios de alquiler reales existentes y utilizar métricas relevantes para asegurar su fiabilidad.
- **Validación del Modelo:**
  - Confirmar que el modelo es robusto y fiable probándolo con un conjunto de datos independiente que no se haya utilizado durante su desarrollo.

2. **Demostración e Implementación del Modelo para Uso Público:**
   - **Aplicación Práctica con un Caso Real:**
     - **Intención:**
       * Utilizar un caso específico y real para demostrar cómo el modelo aplica sus cálculos y genera una predicción sobre el precio de alquiler.
     - **Análisis de Resultados:**
       * Examinar cómo la predicción se compara con el precio real (si está disponible), explorando y discutiendo la precisión y posibles desviaciones del modelo.
   - **Desarrollo de una Herramienta Interactiva para el Público:**
     - **Objetivo:**
       * Facilitar una herramienta en línea, basada en el modelo, que permita a los usuarios ingresar detalles de una propiedad y recibir una estimación del precio de alquiler.
     - **Facilidad de Uso:**
       * Asegurar que la herramienta sea intuitiva y que los resultados sean comprensibles para un amplio espectro de usuarios, sin requerir conocimientos técnicos previos.
     - **Acceso Público:**
       * Ofrecer la herramienta en línea para maximizar su accesibilidad y fomentar su uso por parte de una amplia audiencia, posibilitando así una interacción y retroalimentación valiosa de los usuarios.

En suma, este proyecto pretende no sólo desarrollar un modelo predictivo sólido sino también mostrar de manera transparente y educativa cómo este modelo se puede llevar desde una fase teórica y de desarrollo hasta una aplicación práctica y útil para el público general.

1.3 Desarrollo del Modelo

1.3.1 Descripción del Proceso de Desarrollo del Modelo

El desarrollo del modelo predictivo para los precios de alquiler en Barcelona se realizó siguiendo un proceso metódico y estructurado, desglosado a continuación en función del código proporcionado:

1. Selección de Variables Identificamos diversas variables como `barrio`, `distrito`, `precio`, `Superficie`, `habitaciones` y `baños` que se presumen influyentes en la predicción de los precios de alquiler, basándonos en el conocimiento del dominio y el contexto.

2. Técnica/Modelo de Machine Learning Utilizado Exploramos varios modelos de machine learning con la intención de comparar su desempeño y seleccionar el más óptimo. Los modelos examinados fueron: - **Regresión Lineal - Lasso - Árbol de Decisión - Support Vector Regression (SVR) - Random Forest**

3. Validación y Ajuste del Modelo Empleamos **validación cruzada** (`ShuffleSplit`) y **búsqueda de cuadrícula** (`GridSearchCV`) para afinar los parámetros de los modelos y asegurar robustez y consistencia en los resultados. - **ShuffleSplit**: Creó múltiples divisiones de datos en conjuntos de entrenamiento y prueba para ofrecer diferentes perspectivas de desempeño del modelo. - **GridSearchCV**: Exploró de manera exhaustiva las combinaciones de parámetros para cada modelo, asegurando que la configuración óptima fuera identificada para maximizar la precisión.

4. Elección del Modelo Final Basándonos en los resultados de la búsqueda de cuadrícula, **Random Forest** emergió como el modelo con el mejor rendimiento, ofreciendo un equilibrio entre sesgo y varianza y proporcionando un modelo robusto y generalizable para nuestras predicciones de precios de alquiler.

Justificación de las Elecciones

- **Uso de Múltiples Modelos**: Permitió una perspectiva comparativa, asegurando que la elección del modelo estuviera basada en rendimiento observable y no en suposiciones.
- **Optimización de Parámetros**: Aseguró que el modelo estuviera bien afinado para ofrecer la mejor predicción posible con nuestro conjunto de datos.
- **Validación Cruzada**: Garantizó que el desempeño del modelo fuera robusto y consistente a través de diferentes conjuntos de datos, mitigando el riesgo de sobreajuste.

En resumen, este proceso no solo entregó un modelo predictivo robusto y validado, sino que también aseguró que las decisiones tomadas durante el desarrollo del modelo estuvieran fundamentadas y fueran transparentes, facilitando la interpretación y la confianza en las predicciones del modelo final.

1.3.2 Procesamiento y Limpieza de Datos

A continuación se describe el proceso seguido para la limpieza y preprocesamiento de los datos, etapa esencial para asegurar la calidad del modelo predictivo desarrollado posteriormente.

1. Creación de Nuevas Variables

- **Precio por m^2**: Se introdujo una nueva variable al dataset para representar el precio por metro cuadrado, lo cual provee una métrica normalizada de los precios.

```
df2 = df1.copy()
df2['precio_por_m2'] = (df2['precio']/df2['superficie']).round(2)
df2.head()
```

2. Eliminación de Outliers

- **Outliers en Precio por m^2**: Se eliminan los outliers de la variable `precio_por_m2` utilizando la media y desviación estándar, con un enfoque por barrio y distrito para preservar las características únicas de cada zona.

```
def remove_pps_outliers(df):
    df_out = pd.DataFrame()
    for (distrito, barrio), subdf in df.groupby(['distrito', 'barrio']):
        m = np.mean(subdf.precio_por_m2)
        st = np.std(subdf.precio_por_m2)
```

```
        reduced_df = subdf[(subdf.precio_por_m2>(m-st)) & (subdf.precio_por_m2<=(m+st))]
        df_out = pd.concat([df_out, reduced_df], ignore_index=True)
    return df_out

df3 = remove_pps_outliers(df2)
```

3. Manipulación de Variables

- **Análisis de Anomalías**: Se revisan y gestionan posibles anomalías o errores en los datos, por ejemplo, propiedades que presentan una cantidad de baños inusual respecto a las habitaciones.

```
df3[df3.baños>df3.habitaciones+2]
```

- **Eliminación de Variables**: Se eliminan variables que ya no son necesarias o que han sido transformadas, como `precio_por_m2`.

```
df5 = df4.drop(['precio_por_m2'], axis='columns')
```

4. Codificación One-Hot

- **Variables Dummy**: Se convierten las variables categóricas `barrio` y `distrito` en variables dummy para permitir una representación numérica en el modelo.

```
dummies_barrio = pd.get_dummies(df5.barrio)
dummies_distrito = pd.get_dummies(df5.distrito)
```

5. Limpieza Final de Variables

- **Depuración de Variables**: Se eliminan las variables categóricas originales y otras que ya no serán utilizadas en el modelo.

```
df7 = df6.drop('barrio', axis='columns')
df8 = df7.drop('distrito', axis='columns')
df9 = df8.drop('precio_range', axis='columns')
```

Este proceso de limpieza y preprocesamiento asegura que el modelo de machine learning opere sobre un dataset limpio y estructurado, incrementando la robustez y fiabilidad del modelo predictivo final.

1.4 Puesta en Marcha del Modelo

1.4.1 Descripción del Proceso de Implementación y Desafíos

La implementación del modelo de predicción se llevó a cabo en dos fases esenciales, cada una con sus desafíos y enfoques distintivos para garantizar un despliegue eficiente y operativo.

Fase 1: Implementación en Backend Local En la primera etapa, el modelo fue implementado en un **backend local** utilizando el servidor **Flask** de Python. Esta estrategia proporcionó una validación rápida y un desarrollo iterativo al permitir pruebas inmediatas y ajustes constantes sin la complejidad de desplegar en un entorno en la nube.

Desafíos y Soluciones:

- **Desafío:** Asegurar una comunicación fluida entre el modelo y el servidor Flask, y una transferencia de datos coherente entre frontend y backend.

- **Solución:** Implementar pruebas exhaustivas y corregir problemas de comunicación y transmisión de datos.

Fase 2: Migración a AWS EC2 Posteriormente, la implementación se trasladó a un servidor **EC2** en la nube de **AWS**, proporcionando una plataforma robusta y escalable, capaz de manejar una mayor carga de usuario y ofrecer una disponibilidad más constante.

Desafíos y Soluciones:

- **Desafío:** La configuración del servidor, seguridad del entorno y escalabilidad para manejar diversas cargas de usuarios en AWS EC2.
- **Solución:** Desplegar una estrategia de seguridad sólida, incluyendo grupos de seguridad y políticas de IAM, junto con herramientas de monitoreo para evaluar y ajustar el rendimiento del servidor.

La metodología bifásica permitió una validación y optimización iterativa del modelo, comenzando desde un entorno local controlado, y eventualmente expandiéndose a un servidor en la nube para aprovechar las ventajas de la computación en la nube y ofrecer el modelo a una audiencia más amplia de manera efectiva y segura. Es vital asegurar que cada fase de implementación y cada cambio en la configuración del entorno se pruebe de manera exhaustiva para asegurar que el modelo sigue cumpliendo con las expectativas y que los datos del usuario estén protegidos y seguros.

1.5 Conclusiones y Pasos Futuros

1.5.1 Reflexiones y Conclusiones

El camino recorrido a lo largo de este proyecto ha sido, sin duda, ilustrativo y desafiante en varios aspectos. La conclusión resultante es fundamentalmente positiva, identificando tanto los logros obtenidos como las lecciones aprendidas en el proceso.

1. **Implementación Exitosa:** A pesar de los retos, se ha logrado desplegar un modelo de machine learning, primero a nivel local y luego escalándolo a un entorno en la nube, lo cual en sí mismo es un logro significativo.

2. **Desafíos en la Nube:** Aunque se consultaron diversos tutoriales para la implementación en AWS EC2, esta fase no estuvo exenta de obstáculos. Los tutoriales, aunque informativos, a menudo no abarcan la totalidad de los problemas específicos que pueden surgir durante una implementación real, y algunas veces pueden estar desactualizados respecto a las versiones y características más recientes de las herramientas y plataformas.

3. **Importancia de Datos Actuales:** El proyecto subrayó la crítica importancia de disponer de un conjunto de datos completo, único y actualizado. Aunque el modelo ha sido desarrollado y desplegado exitosamente, es innegable que un dataset más robusto y actualizado habría potencialmente permitido desarrollar un modelo con una predicción más precisa y fiable.

4. **Servidor elegido en la nube:** Se debe tener en cuenta que este modelo se implementa dentro de un proyecto final donde la inversión en recursos ha sido nula, por lo que el servidor virtual contratado tiene posiblemente limitaciones de accesibilidad que no han podido ser todavía contrastadas. Es posible que durante la presentación seamos testigos de estas limitaciones.

1.5.2 Siguientes Pasos y Mejoras Futuras

1. **Mejora de la Base de Datos:** La calidad del modelo puede ser significativamente mejorada si en futuras iteraciones se dispone de una base de datos más amplia, diversa y actualizada.
2. **Documentación Detallada:** La creación de una documentación propia, detallada y específica para el despliegue en la nube que refleje las lecciones aprendidas y los desafíos superados durante este proyecto, facilitará futuras implementaciones y servirá como un recurso valioso para otros proyectos.
3. **Optimización del Modelo:** Considerar el feedback y los datos recabados durante la fase operativa del modelo para realizar ajustes y mejoras, refinando su precisión y confiabilidad.
4. **Seguridad y Confiabilidad:** Continuar mejorando los aspectos de seguridad y confiabilidad del modelo y de la infraestructura en la nube, asegurando que los datos del usuario se manejen de manera segura y que el modelo sea resistente a posibles fallos.