Paso 1

Análisis Inicial, Conclusiones y Hallazgos Claves

El conjunto de datos contiene 21 columnas con información como el precio de las viviendas, el número de habitaciones, baños, superficie habitable (sqft_living), superficie del lote (sqft_lot), número de pisos, vistas, calificación (grade), entre otras.

Algunas columnas notables incluyen:

- price: precio de la vivienda.
- bedrooms, bathrooms: cantidad de habitaciones y baños.
- sqft_living, sqft_lot: área en pies cuadrados de la vivienda y el terreno.
- zipcode: código postal de la propiedad.
- yr_built, yr_renovated: año de construcción y renovación.

Aquí están los hallazgos clave del análisis exploratorio:

1. Valores faltantes:

No hay valores nulos en el conjunto de datos, por lo que no es necesario realizar imputación de datos.

2. Estadísticas descriptivas:

  - El precio promedio de las viviendas es de aproximadamente $540,000, con un máximo de $7.7 millones.
  - El número promedio de habitaciones es 3.37 y de baños es 2.11.
  - La superficie habitable promedio es de 2,080 sqft, con un máximo de 13,540 sqft.

3. Correlaciones con el precio:

  - sqft_living (superficie habitable) tiene la correlación más alta con el precio (0.70).
  - grade (calificación de la vivienda) tiene una correlación de 0.67.
  - sqft_above (área por encima del suelo) y sqft_living15 (área de viviendas cercanas) también muestran correlaciones significativas con el precio.

Aquí están las conclusiones del análisis exploratorio:

- Distribución de precios: La mayoría de las viviendas tienen precios por debajo de $1 millón, con algunos valores atípicos en el rango de millones.
- Tamaño de las viviendas: El tamaño (superficie habitable) muestra una correlación positiva con el precio. Viviendas más grandes tienden a ser más caras.
- Calidad (grade): Las viviendas con una calificación más alta también tienen precios más elevados, lo que indica que la calidad percibida de la vivienda es un factor clave en el precio.
- Ubicación: Viviendas frente al agua (waterfront) tienen una correlación positiva con el precio. Por otra parte, la ubicación en términos de latitud y longitud no tiene una correlación fuerte con el precio, pero la latitud muestra una leve tendencia (más al norte parece ser mejor).
- Renovación: Las casas renovadas muestran una ligera correlación positiva con el precio.

En el repositorio se encuenta una imagen llamada “Patrones Encontrados” corresponde a todo lo anteriormente mencionado puesto en gráficos, a continuación se encuentra la explicación respectiva de cada gráfico

1. **Distribución de precios**: El gráfico muestra que la mayoría de las viviendas tienen un precio por debajo de $1 millón, pero hay algunas viviendas de precios mucho más altos, lo que indica la presencia de outliers.
2. **Superficie habitable vs. Precio**: Hay una clara tendencia positiva. A medida que aumenta la superficie habitable (sqft_living), el precio también aumenta, aunque con algunos puntos dispersos.
3. **Calificación (grade) vs. Precio**: Las viviendas con calificaciones más altas tienden a tener precios más elevados. Esto refleja que la calidad percibida de la construcción influye en el valor.
4. **Propiedades frente al agua**: Las viviendas frente al agua tienen precios significativamente más altos que aquellas que no lo están, mostrando una ventaja clara en este tipo de localización.

Paso 2

Preprocesamiento de los Datos

El preprocesamiento de los datos comenzó con la limpieza y transformación de las variables clave. Se utilizó un conjunto de datos de precios de viviendas, que incluía características como el número de habitaciones, baños, y el tamaño de la vivienda en pies cuadrados. Uno de los principales pasos fue calcular una nueva variable: precio por metro cuadrado **(**price_per_sqft**),** dividiendo el precio total entre el área habitable de cada propiedad.

Luego, se calculó la media y la desviación estándar del precio por metro cuadrado, lo cual permitió definir un *umbral de atractivo* del 20% por encima de la media para identificar viviendas con un valor superior en función de su precio por área.

Finalmente, se dividió el conjunto de datos en variables independientes (características como área habitable, habitaciones, baños) y la variable dependiente (el precio por metro cuadrado ajustado al umbral de atractivo). Se realizó una normalización de las características para que todas estuvieran en la misma escala, usando un escalador estándar. Esto fue crucial para asegurar un rendimiento adecuado del modelo.

Paso 3

Selección de un modelo

Para ajustar el precio de las viviendas, se consideraron varios modelos antes de llegar a una decisión final. Inicialmente, se evaluaron métodos de regresión lineal y árboles de decisión como posibles opciones debido a su simplicidad y capacidad de explicar relaciones directas entre las variables independientes y el precio. Sin embargo, estos modelos no ofrecieron la flexibilidad necesaria para capturar las relaciones complejas entre las características de las propiedades y el precio por metro cuadrado.

A continuación, se intentó implementar un modelo utilizando Keras como una biblioteca de alto nivel para crear redes neuronales. Este enfoque prometía ser más adecuado, pero surgieron dificultades con la importación de las capas necesarias y la configuración del entorno. Debido a estos problemas técnicos y en búsqueda de una mayor flexibilidad, se decidió usar *TensorFlow* directamente, sin depender de las abstracciones de Keras.

Finalmente, se construyó un modelo de red neuronal utilizando *TensorFlow*. Este modelo consistió en una red densa con varias capas ocultas, incluyendo una capa inicial con 64 neuronas, seguida de una capa de 32 neuronas, ambas con la función de activación “ReLU”. La capa de salida fue diseñada para generar un valor continuo, prediciendo el porcentaje de atractivo basado en el precio por metro cuadrado.

El modelo fue entrenado y evaluado usando el conjunto de datos dividido en entrenamiento (80%) y prueba (20%), y se utilizó el error cuadrático medio (MSE) como métrica de evaluación. Aunque se consideraron los modelos de regresión y árbol de decisión, la red neuronal basada en TensorFlow demostró una mayor capacidad para capturar patrones complejos en los datos y, en consecuencia, fue seleccionada como la mejor opción.

Paso 4

Selección de viviendas atractivas para inversión

A partir del modelo entrenado, se utilizó la predicción del *porcentaje de atractivo* para identificar las viviendas más prometedoras en términos de inversión. El umbral de atractivo fue definido en función del precio por metro cuadrado, calculado previamente. Aquellas propiedades que superaban el umbral definido se consideraron como atractivas.

Partiendo de lo desarrollado y asumiendo que este es un estimador confiable del valor de mercado, se estableció un umbral de atractivo para identificar propiedades que presentaran un valor por encima del 20% del precio promedio por pie cuadrado. De esta manera, se consideran más atractivas para inversión aquellas propiedades cuyo valor proyectado supere significativamente la media del mercado.

Un ejemplo de una vivienda atractiva sería una propiedad con 3 habitaciones, 2 baños y un tamaño de 2,000 pies cuadrados, con un precio por pie cuadrado que excede el umbral de atractivo calculado. Este tipo de vivienda sería ideal para invertir, ya que tiene el potencial de ofrecer mayores retornos.

Finalmente, el modelo fue configurado para seleccionar las 10 viviendas más atractivas para inversión. Estas propiedades fueron extraídas del DataFrame original y almacenadas en un nuevo DataFrame denominado "*propiedades_Mejores_Invertir*", el cual facilita la visualización de las mejores oportunidades de inversión basadas en las predicciones del modelo. Este paso asegura que el sistema no solo identifique propiedades atractivas, sino que también las ordene y presente de manera estructurada.