Informe lab 8 DS

1. Exploración de datos

El análisis exploratorio comenzó con una inspección detallada del conjunto de datos de alquiler de viviendas en Brasil. Se analizaron las distribuciones de variables clave como el área, número de habitaciones, baños, espacios de parqueo y otros factores importantes que influyen en el precio del alquiler. Se identificaron patrones como la correlación positiva entre el área de la propiedad y el precio total de la renta. También se revisaron posibles valores atípicos, que indicaron la necesidad de manejarlos apropiadamente para no distorsionar el análisis.

Además, se visualizaron las distribuciones de las variables clave utilizando histogramas y gráficos de dispersión, lo que permitió una mejor comprensión de la estructura de los datos. Se identificó que algunas variables categóricas, como la ciudad y si las propiedades permiten animales o no, influyen en el precio del alquiler.

2. Preprocesamiento y limpieza de datos

Durante el preprocesamiento, se manejaron los valores faltantes mediante imputación con medianas para las variables numéricas y con la moda para las variables categóricas. Los outliers se analizaron cuidadosamente y se consideraron estrategias para suavizarlos sin perder información crítica.

Las variables categóricas se codificaron utilizando la técnica de **OneHotEncoder**, que permitió transformar valores como la ciudad, la opción de animales permitidos y si la vivienda estaba amueblada. Las variables numéricas se estandarizaron con el objetivo de normalizar las escalas y mejorar la capacidad predictiva de los modelos de aprendizaje automático.

3. Entrenamiento y selección de modelos Se implementaron tres modelos diferentes: **Regresión Lineal**, **Random Forest** y **Gradient Boosting**. Cada uno de ellos fue entrenado utilizando un pipeline que incluía el preprocesamiento de los datos y se evaluaron utilizando validación cruzada para asegurar la robustez de los resultados.

Las métricas principales utilizadas para la comparación de modelos fueron el **RMSE** (Root Mean Squared Error), el **MAE** (Mean Absolute Error), y el **R^2**. A continuación, los resultados obtenidos:

- **Linear Regression**:
  - RMSE: 481.06
  - MAE: 317.82
  - R^2: 0.9884
- **Random Forest**:
  - RMSE: 393.40
  - MAE: 151.27

    - R^2: 0.9922
- **Gradient Boosting**:
    - RMSE: 388.47
    - MAE: 216.15
    - R^2: 0.9924

El modelo de **Gradient Boosting** mostró el mejor rendimiento general con el RMSE más bajo y un alto valor de R^2, lo que indica que fue capaz de capturar de manera efectiva la relación entre las características de las viviendas y el precio del alquiler. Aunque el modelo de Random Forest también tuvo un rendimiento muy similar, Gradient Boosting se seleccionó como el modelo final.

4. Interpretación de resultados

La importancia de las características en el modelo seleccionado se evaluó utilizando el modelo de **Gradient Boosting**, que permitió observar que las variables más importantes para la predicción del precio de alquiler fueron el área, el número de habitaciones y el impuesto HOA .

Las fortalezas del modelo de Gradient Boosting incluyen su capacidad para manejar no linealidades en los datos, mientras que una debilidad potencial es su mayor complejidad y tiempo de entrenamiento en comparación con otros modelos como la regresión lineal.

Reflexión

Desafíos durante el proceso

Uno de los principales desafíos fue integrar el modelo de machine learning en una aplicación web interactiva utilizando **Streamlit**. Aunque Streamlit es una herramienta muy poderosa para crear aplicaciones de manera rápida y eficiente, presentaba algunos retos técnicos a la hora de manejar la validación de datos y la interacción en tiempo real con los usuarios. Asegurar que el modelo cargado estuviera correctamente preentrenado y que las predicciones se mostraran de manera clara fue un paso crítico en el desarrollo de la aplicación.

Aprendizajes significativos

El uso de Streamlit facilitó la creación de una interfaz interactiva sin tener que profundizar en herramientas más complejas como Flask o Django, lo que permitió una implementación rápida de la solución. La capacidad de Streamlit para mostrar gráficos y manejar la interacción con el usuario simplificó la tarea de comunicar los resultados del modelo de machine learning de manera efectiva y visual. Esta simplicidad, sin embargo, también introdujo algunas limitaciones que requieren considerar soluciones más robustas si se desea una mayor personalización o un despliegue más escalable.

Fortalezas y limitaciones

La principal fortaleza de **Streamlit** es su facilidad de uso. En pocos pasos, es posible crear una aplicación funcional que permita a los usuarios cargar datos, realizar predicciones y visualizar

los resultados. Además, su integración nativa con gráficos y visualizaciones lo convierte en una excelente herramienta para prototipos y pruebas rápidas.

No obstante, Streamlit tiene limitaciones en términos de escalabilidad y personalización avanzada. Aunque es ideal para laboratorios y pruebas, no es la mejor opción para aplicaciones de producción de gran escala que requieren mayor control sobre la arquitectura backend, manejo de bases de datos complejas o autenticación de usuarios. En esos casos, frameworks más robustos como Flask o Django son más adecuados. Además, el manejo de errores y la validación avanzada de datos en Streamlit puede ser más rudimentario en comparación con otras soluciones web más maduras.

Sugerencias para mejorar

Para futuras mejoras en el uso de Streamlit, sería beneficioso profundizar en la capacidad de personalizar los flujos de trabajo dentro de la aplicación, mejorando la validación de los datos ingresados por el usuario y añadiendo manejo de errores más detallado. Asimismo, explorar formas de integrar Streamlit con bases de datos en tiempo real y optimizar su desempeño para aplicaciones con mayor carga de usuarios sería una dirección a seguir.

En resumen, Streamlit es una herramienta poderosa y fácil de usar que permite crear aplicaciones web interactivas en poco tiempo. Sin embargo, para un entorno de producción más robusto o con necesidades complejas, puede ser necesario considerar otras alternativas que ofrezcan mayor flexibilidad y control.

Predicción de Precios de Alquiler de Viviendas en Brasil

Seleccione el modelo de Machine Learning

Linear Regression

Área (m^2)

50

Número de habitaciones

2

Número de baños

1

Número de espacios de parqueo

1

Impuesto HOA (R$)

500